Adaptivity in Online and Statistical Learning

Zakaria Mhammedi

A thesis submitted for the degree of
Doctor of Philosophy
The Australian National University

August 2021

Except where otherwise indicated, this thesis is my own original work. This thesis is based on five published articles and two additional manuscripts under review:

- **Mhammedi, Z.**, Williamson, R. C. (2018). *Constant Regret, Generalized Mixability, and Mirror Descent*. In **NeurIPS'18**.
- **Mhammedi, Z.**, Koolen, W. M., Van Erven, T. (2019). *Lipschitz Adaptivity with Multiple Learning Rates in Online Learning*. In **COLT'19**.
- **Mhammedi, Z.**, Grünwald, P. D., Guedj, B. (2019). *PAC-Bayes Un-Expected Bernstein Inequality*. In **NeurIPS'19**.
- **Mhammedi, Z.**, Guedj, B., and Williamson, R. C. (2020). *PAC-Bayesian Bound for the Conditional Value at Risk*. In **NeurIPS'20**.
- **Mhammedi, Z.**, Koolen, W. M. (2020). *Lipschitz and Comparator Norm Adaptivity in Online Learning*. In **COLT'20**.
- **Mhammedi, Z.**, Koolen, W. M. (2021). *Concentration Inequalities for CVaR with Near-optimal Quantile Level Dependence*. Under review.
- **Mhammedi, Z.**, Husain, H. (2021). *Risk-Monotonicity in Statistical Learning*. Under review.

I contributed to the majority of the writing and development of ideas of the above publications. Additional publications not included in this thesis are:

- Roy, S. K., **Mhammedi Z.**, Harandi, M. *Geometry Aware Constrained Optimization Techniques for Deep Learning*. In **CVPR'18**.
- **Mhammedi, Z.**, Foster, D. J., Simchowitz, M., Misra, D., Krishnamurthy, A., Rakhlin, A., Langford, J. 2020. *Learning the Linear Quadratic Regulator from Nonlinear Observations*. In **NeurIPS'20**.

Zakaria Mhammedi
24 August 2021

Acknowledgments

I am grateful to Bob Williamson for encouraging me to pursue a PhD in Theoretical Machine Learning despite my initial hesitation. Bob also introduced me to Peter Grünwald, who I visited and interned with at the Centrum Wiskunde Informatica (CWI). My time at CWI laid much of the foundation of my PhD work.

Through my interactions with Peter, I learned a lot about and developed a great appreciation for Statistical Learning Theory, which was instrumental to my PhD research.

During my visits to CWI, I also met Wouter Koolen who eventually took the role of my PhD advisor. I am thankful to Wouter for introducing me to Online Learning Theory, which became my main area of research. Working with Wouter was a unique experience that I thoroughly enjoyed and gained a lot from.

I thank both Peter and Wouter for being such great hosts each time I visited CWI. My thanks also go to:

- John Langford and Akshay Krishnamurthy for the opportunity to intern at Microsoft Research and introducing me to exciting bandit and RL problems.
- Hisham Husain for being a supportive peer during my PhD.
- Richard Nock for providing support to attend conferences.
- My family and friends for their encouragement and support.

Abstract

Many modern machine learning algorithms, though successful, are still based on heuristics. In a typical application, such heuristics may manifest in the choice of a specific Neural Network structure, its number of parameters, or the learning rate during training. Relying on these heuristics is not ideal from a computational perspective (often involving multiple runs of the algorithm), and can also lead to overfitting in some cases. This motivates the following question: for which machine learning tasks/settings do there exist efficient algorithms that automatically adapt to the best parameters? Characterizing the settings where this is the case and designing corresponding (parameter-free) algorithms within the online learning framework constitutes one of this thesis' primary goals. Towards this end, we develop algorithms for constrained and unconstrained online convex optimization that can automatically adapt to various parameters of interest such as the Lipschitz constant, the curvature of the sequence of losses, and the norm of the comparator. We also derive new performance lower-bounds characterizing the limits of adaptivity for algorithms in these settings. Part of systematizing the choice of machine learning methods also involves having "certificates" for the performance of algorithms. In the statistical learning setting, this translates to having (tight) generalization bounds. Adaptivity can manifest here through data-dependent bounds that become small whenever the problem is "easy". In this thesis, we provide such data-dependent bounds for the expected loss (the standard risk measure) and other risk measures. We also explore how such bounds can be used in the context of risk-monotonicity.

Contents

Chapter 1

Introduction

The goal of supervised machine learning is to build prediction models that generalize well on unseen data. However, many current machine learning models require the tuning of many parameters (e.g. step size) during the training phase, hindering their generalization performance in some cases. For example, training a Neural Network (NN) model using a gradient-based method typically involves tuning a "learning rate" parameter, which is often done using heuristics. Such heuristic-based methods often lack theoretical guarantees and can sometimes lead to over-fitting on the training data[1] (and thus poor generalization performance) [Feurer and Hutter, 2019]. Therefore, it is desirable to have machine learning algorithms that do not require such parameters in the first place and still perform competitively against models whose parameters are tuned to their theoretical optimal values given the problem at hand.

We have mentioned the learning rate as an example of a hyper-parameter needing tuning when training an NN model. There are other vital choices a practitioner needs to make in this case, such as the size and architecture of the NN. Choosing the best "model" has to be done carefully to not over-fit the available data. Generalization bounds, which control the deviation between the population and empirical error, can be used to select a model after observing the data. However, generic generalization bounds fail to account for key properties of candidate algorithms/models beyond their empirical performance on the observed data [Dziugaite et al., 2020]. A crucial property of algorithms known to be beneficial for generalization is *stability* [Bousquet and Elisseeff, 2002]. So it is desirable to have data-dependent generalization bounds that can automatically become small whenever an algorithm is stable. In this case, an algorithm may be considered stable if the empirical losses of the hypothesis it picks (after training) do not change much if a small number of data points are removed during training[2].

Perhaps the simplest framework under which one can study and characterize algorithms that do not require any hyper-parameter tuning is the online learning setting, which we discuss next.

[1] One can reduce the amount of overfitting using techniques such as cross-validation. However, such approaches still lacks solid theoretical guarantees.

[2] Here, we do not want to restrict ourselves to the standard notion of algorithmic stability studied by Bousquet and Elisseeff [2002].

1.1 Adaptivity in Online Learning

Due to its simple formulation, the online learning setting, where actions are taken sequentially, is an excellent starting point for studying adaptivity in machine learning. Algorithms developed for this setting have many applications and are behind various popular methods. A prominent application in machine learning is the stochastic optimization of non-convex objectives (relevant for training deep NN models); the well-known stochastic gradient descent is an example of an online learning algorithm used for this purpose. In general, many online learning algorithms can be used for stochastic optimization purposes via online-to-batch conversion techniques [Cesa-Bianchi et al., 2004; Shalev-Shwartz, 2012; Cutkosky, 2019a]. Recently, new reductions have been introduced that successfully leverage online learning algorithms for other ML applications such as differential privacy [Van der Hoeven, 2019a; Jun and Orabona, 2019b], online control [Agarwal et al., 2019b; Simchowitz, 2020; Foster and Simchowitz, 2020], and reinforcement learning [Cassel and Koren, 2020; Neu et al., 2017]. One attractive aspect of such reductions is that they transfer the performance and level of adaptivity of the online algorithms used. Thus, designing adaptive online learning algorithms has the potential to eliminate the need for heuristic hyper-parameter tuning in many machine learning applications.

Online learning can be performed with either full information or partial information (i.e. the bandit setting). In full information, the losses of all actions are revealed to the learner (the algorithm) at every prediction round, whereas in partial information, only the loss of the action played may be revealed. Many techniques have recently been introduced that reduce partial feedback settings to full information (see [Lykouris et al., 2018] for an overview). This thesis will be mainly concerned with the latter setting. A popular assumption in full information is that the observed losses are convex. This setting is known as Online Convex Optimization (OCO) [Hazan, 2016a], which subsumes the well-known experts' setting. Many machine learning applications can be reduced to OCO [Neu et al., 2017; Van der Hoeven, 2019a; Agarwal et al., 2019c; Simchowitz, 2020]. The non-convex online learning setting is naturally more challenging. Yet, techniques from the simpler OCO (or the experts' setting) have successfully been used to derive algorithms with provable guarantees for the non-convex setting [Agarwal et al., 2019a; Suggala and Netrapalli, 2020; Héliou et al., 2020]. For example, Agarwal et al. [2019a] showed that it is possible to achieve a sub-linear regret (sub-linear in the number of rounds) in non-convex online learning using Follow-the-Perturbed-Leader (an algorithm developed for the experts' setting [Kalai and Vempala, 2003]) and an offline optimization oracle. This online non-convex setting is outside the scope of this thesis.

The first part of this thesis will focus on the OCO setting, where we study the limits of adaptivity in both the bounded and unbounded cases.

Bounded online convex optimization. Bounded OCO is one of the most studied settings in online learning [Cesa-Bianchi and Lugosi, 2006; Hazan, 2016b]. This is a setting where at each round t, a learned (the algorithm) outputs a vector $\widehat{w}_t$ in

some bounded convex set $\mathcal{W}$, then the environment reveals a convex loss function $f_t : \mathcal{W} \to \mathbb{R}$. The output $\widehat{\boldsymbol{w}}_t$ can be any function on the past observed losses $(f_s)_{s<t}$. A prominent example is the online experts' setting, where the constrained set is the simplex. Many algorithms have been developed for bounded OCO, which achieve different degrees of adaptivity [Gaillard and and, 2014; Luo and Schapire, 2015; Van Erven and Koolen, 2016; Hazan et al., 2007]. The Online Gradient Descent (OGD) algorithm [Zinkevich, 2003] is an example of a popular OCO algorithm. For Lipschitz convex losses, OGD can achieve a sub-linear (in the number of rounds) regret; more precisely, the outputs $(\widehat{\boldsymbol{w}}_t)$ of the algorithm guarantee

$$\mathrm{Regret}_T(\boldsymbol{w}) := \sum_{t=1}^{T} (f_t(\widehat{\boldsymbol{w}}_t) - f_t(\boldsymbol{w})) \leq O(D\sqrt{T}), \quad \forall \boldsymbol{w} \in \mathcal{W}, \tag{1.1}$$

where $D := \sup_{\boldsymbol{w},\boldsymbol{w}' \in \mathcal{W}} \|\boldsymbol{w} - \boldsymbol{w}'\|$ is the "diameter" of the set $\mathcal{W}$. It is known that with proper tuning of the learning rate schedule of OGD, it is possible to achieve a small (logarithmic in the number of rounds) regret when the loss functions (f_t) are strongly convex [Hazan et al., 2007]. Online Newton Step (ONS) is another algorithm that can achieve a small (logarithmic in T) regret when the losses are exp-concave (a weaker condition than strong convexity) [Hazan et al., 2007]. Both OGD and ONS follow the framework of Online Mirror Descent (MD)[3]—a prominent algorithm in online learning—with different regularizers (see e.g. [Hazan, 2016b]). Interestingly, all these algorithms, including most MD instantiations, can be viewed as instances of the Exponential Weights Algorithm (EWA) with different priors and learning rates [Van der Hoeven et al., 2018].

Studying adaptivity for the above algorithms has often involved looking at different learning rate schedules or choices of regularizers (in the case of MD). Here adaptivity means achieving the smallest possible regret guarantee for the type of losses observed; for curved (strongly convex or exp-concave) losses, one would like to achieve a logarithmic regret (i.e. the best achievable regret for such losses). However, for algorithms such as OGD or ONS, the learning rate schedule must be set as a function of the parameters of curvature of the losses to achieve a logarithmic regret—these parameters may not be known in practice. Another parameter that one would like to adapt to is the "Lipschitz constant"; the maximum norm of the losses' gradients at the algorithm's iterates. Many algorithms require this parameter as input and may fail (not realize the desired regret) if the parameter is not a valid upper bound on the norm of the gradients [Van Erven et al., 2021]. Having a single algorithm that adapts to all these parameters (i.e. the Lipschitz constant and curvature parameters) simultaneously while achieving the optimal regret guarantee for the given type of losses is one of the contributions of this thesis (see Chapter 3).

In the experts' setting (a special case of OCO), it is possible to achieve a *constant* regret when the losses are *mixable*. This is achieved using the Aggregating Algo-

[3]Technically, the version of ONS presented in [Hazan et al., 2007] is framed as Follow-the-Regularized-Leader (FTRL). However, FTRL and MD may be seen as equivalent—see e.g. [McMahan, 2011].

rithm, which essentially outputs "exponential weights" using the observed losses [Vovk, 1998]. It turns out that a larger family of algorithms—referred to as generalized aggregating algorithms—can also achieve a constant regret [Reid et al., 2015]. This family consists of mirror descent algorithms with specific regularizers whose characterization is the subject of Chapter 2.

Unbounded online convex optimization. While the bounded OCO setting has by now been studied in some depth, the *unbounded* OCO setting—where the output set $\mathcal{W}$ may be unbounded—has only been explored relatively recently [McMahan and Streeter, 2010; Mcmahan and Streeter, 2012; Orabona, 2013; McMahan and Abernethy, 2013; Orabona, 2014a; McMahan and Orabona, 2014; Orabona and Pál, 2016a]. Optimal algorithms for this setting are known as parameter-free since they can compete against unbounded comparators $\boldsymbol{w}$ without any learning rate tuning. The typical regret bound achieved by parameter-free algorithms such as those in [Orabona, 2014b; McMahan and Orabona, 2014; Foster et al., 2017] is as follows:

$$\text{Regret}_T(\boldsymbol{w}) = \sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w})) \leq O\left(\|\boldsymbol{w}\|\sqrt{T\ln(1+\|\boldsymbol{w}\|T)}\right), \quad \forall \boldsymbol{w} \in \mathcal{W}. \quad (1.2)$$

Note that the parameter-free regret in (1.2) swaps the diameter D in the bounded case (see (1.1)) for the norm of the comparator $\|\boldsymbol{w}\|$. This comes at the price of logarithmic terms in $\boldsymbol{w}$ and T in the regret, which are unavoidable in the general unbounded case [McMahan and Orabona, 2014].

There are many natural settings with unbounded domains $\mathcal{W}$. One example is the problem of online learning of linear models [Kotlowski, 2017; Kempka et al., 2019a]. In this setting, it is impossible for OGD or other known classical algorithms designed for the bounded setting to achieve the optimal regret bound in (1.2) as the learning rate needs to be tuned as a function of the norm of the comparator $\boldsymbol{w}$, which is typically unknown in advance. On the other hand, parameter-free algorithms, which are designed for the unbounded setting, can be used to achieve the optimal regret in the case of online learning of linear models [Mhammedi and Koolen, 2020]. Parameter-free algorithms can confer benefits beyond the unbounded OCO case. For instance, they are useful in applications involving local differential privacy [Jun and Orabona, 2019a]. In this setting, a data provider presents a "sanitized" version of a data set to a learning algorithm along with some desired privacy level. As pointed out by Van der Hoeven [2019b], the desired privacy level may itself be considered as a privacy-sensitive feature. Van der Hoeven [2019b] showed that parameter-free algorithms can be used to achieve local differential privacy—in a way that standard algorithms such as OGD cannot—without requiring the privacy level as input.

What is also surprising about parameter-free algorithms is that even though they are designed for the unbounded setting, they can also be used in the bounded setting and even yield improved regret guarantees. There exist techniques that reduce the bounded OCO setting to the unbounded one, where the regret bound for the former becomes that of the unbounded algorithm used (up to constant factors) [Cutkosky

and Orabona, 2018; Cutkosky, 2020b]. What this means is that one can replace the diameter D typically present in the regret bounds for the bounded case (see (1.1)) by the norm of the comparator as in (1.2). This can be advantageous when the optimal comparator $\boldsymbol{w}$ (the one that minimizes $\sum_{t=1}^{T} f_t(\boldsymbol{w})$) has a small norm[4]. This fact has already been leveraged to produce improved regret bounds in the experts' setting—a special case of bounded OCO. One prominent example is the setting where experts have different loss ranges, and the goal is to achieve a regret bound against any given expert i that scales with the loss range of that expert. The algorithm of Bubeck et al. [2017] is the first to achieve this when the losses are positive. When the losses can take negative values, the sought guarantee is achieved by the algorithm of Foster et al. [2017]. However, this algorithm is inefficient as the per-round computational complexity is super-linear in the number of experts. The reduction due to Cutkosky and Orabona [2018] together with parameter-free algorithms lead to the first algorithm that achieves the desired multi-scale regret (up to log factors) with a linear run-time in the number of experts.

Parameter-free algorithms have other exciting applications. In stochastic optimization, it was shown that parameter-free algorithms could guarantee asymptotic convergence for Variationally Coherent Functions (VCFs) while automatically ensuring a near-optimal convergence rate for convex functions [Orabona and Pál, 2021]. VCFs are functions that include convex and non-convex functions such as quasi-convex, star-convex, and pseudo-convex functions.

It is worth noting one more useful property of parameter-free algorithms. As mentioned earlier, the regret bounds of these algorithms scale with the norm of the comparator as in (1.2) (instead of the diameter of the set in bounded OCO), which implies that the regret against the origin is bounded by a constant. This property in turn implies that one can aggregate the predictions of parameter-free algorithms effectively and at a very low cost. Suppose A and B are two parameter-free online algorithms that output vectors $(\widehat{\boldsymbol{w}}_t^{\mathsf{A}})$ and $(\widehat{\boldsymbol{w}}_t^{\mathsf{B}})$, respectively. Then, a third algorithm C that aggregates the predictions of A and B by addition; i.e. $\widehat{\boldsymbol{w}}_t^{\mathsf{C}} := \widehat{\boldsymbol{w}}_t^{\mathsf{A}} + \widehat{\boldsymbol{w}}_t^{\mathsf{B}}$, has a regret that is at most the minimum regret of algorithms A and B up to lower-order terms [Cutkosky, 2019b]. This aggregation property (which can be extended to more than two base algorithms) is very useful and has been leveraged to achieve optimal dynamic regret in OCO, where the comparator changes over time instead of being fixed [Cutkosky, 2020b].

One of the contributions of this thesis (see Chapter 4) are two parameter-free algorithms—FreeGrad and Matrix-FreeGrad—in OCO whose regret guarantees are

$$\mathrm{Regret}_T^{\mathsf{FreeGrad}}(\boldsymbol{w}) \leq O\left(\|\boldsymbol{w}\|\sqrt{\mathrm{tr}(V_T)\cdot\ln(1+\|\boldsymbol{w}\|\cdot\mathrm{tr}(V_T))}\right), \tag{1.3}$$

$$\mathrm{Regret}_T^{\mathsf{Matrix\text{-}FreeGrad}}(\boldsymbol{w}) \leq O\left(\sqrt{\boldsymbol{w}^\top V_T \boldsymbol{w}\cdot\ln(1+\boldsymbol{w}^\top V_T\boldsymbol{w}\cdot\det(V_T))}\right), \tag{1.4}$$

[4]We note, however, that the regret of parameter-free algorithms has a multiplicative logarithmic term in the horizon under the square-root—see (1.2). Thus, any conferred benefit in the bounded setting may eventually disappear as the horizon grows.

for all $\boldsymbol{w} \in \mathcal{W}$ and $V_T := I + \sum_{t=1}^{T} \nabla f_t(\widehat{\boldsymbol{w}}_t) \nabla f_t(\widehat{\boldsymbol{w}}_t)^\top$, for the algorithms' iterates $(\widehat{\boldsymbol{w}}_t)$. We note that the regret bound of FreeGrad in (1.3) is called *adaptive* as it replaces $O(T)$ in the generic parameter-free regret (1.2) by $\mathrm{tr}(V_T) = \sum_{t=1}^{T} \|\nabla f_t(\widehat{\boldsymbol{w}}_t)\|^2 \leq L^2 T$, where $L := \max_{t \in [T]} \|\nabla f_t(\widehat{\boldsymbol{w}}_t)\|$ is the Lipschitz constant. In many situations, $\mathrm{tr}(V_T)$ can be much smaller than $L^2 T$ [Srebro et al., 2010], in which case the adaptive regret of FreeGrad is superior to the generic one in (1.2).

Having the term $\mathrm{tr}(V_T)$ in the regret bound instead of $L^2 T$ (i.e. adaptive regret) allows for different types of adaptivity in various settings. For example, Orabona and Pál [2021] used parameter-free algorithms with an adaptive regret to guarantee almost-sure convergence for Variationally Coherent Functions. An adaptive regret can also be leveraged to achieve a logarithmic regret for strongly convex losses through existing reductions [Cutkosky and Orabona, 2018]. Improved online-to-batch conversion results for stochastic optimization are also enabled through adaptive regrets [Cutkosky, 2019a]. What is more, as we will see in Chapter 8, the guarantee of FreeGrad can be used beyond the online learning setting to prove a new concentration inequality for martingale difference sequences. The new inequality can be viewed as an empirical version of Freedman's inequality [Freedman, 1975].

FreeGrad also has the advantage of being *scale-free* in the sense that multiplying the losses by a constant $c > 0$ does not change the outputs of the algorithm—a desirable property in general [Orabona and Pál, 2016b]. We note that the price of being scale-free and parameter-free[5] comes in the form of an additional term in the regret that is independent of T and scales with $\|\boldsymbol{w}\|^3$. This term is, unfortunately, unavoidable in general if one insists on a $O(\sqrt{T})$ regret as we show via a new lower bound in Chapter 4. This result complements the lower bound due to Cutkosky and Boahen [2017] who showed that if one insists on a linear, up to log-factors, dependence in the comparator norm $\|\boldsymbol{w}\|$, then the regret must essentially grow exponentially with T. The parameter-free algorithm of Cutkosky and Orabona [2018] also achieves the adaptive regret in (1.2). However, the constants involved are worse than those of FreeGrad, and the algorithm is not scale-free.

Finally, we note that the regret bound of Matrix-FreeGrad can be interpreted as enabling adaptivity to the "directional variance". This type of regret has also recently been achieved by Cutkosky [2020a] using a completely different approach than ours. However, their algorithm is not scale-free, and the constants involved in the regret bound are worse than those of Matrix-FreeGrad. The main term $\boldsymbol{w}^\top V_T \boldsymbol{w}$ in the bound of Matrix-FreeGrad can be much smaller than the main term $\|\boldsymbol{w}\|^2 \mathrm{tr}(V_T)$ in the regret of FreeGrad [Cutkosky and Sarlós, 2019]. However, the regret bound of the former has the term $\ln \det(V_T)$, which can be as large as $d \ln T$, where d is the dimension of $\mathcal{W}$. For this reason, the bounds in (1.3) and (1.4) are not necessarily comparable in general. It is worth mentioning here that because both FreeGrad and Matrix-FreeGrad are parameter-free, it is possible (and very easy) to build a new algorithm based on them that achieves the minimum of the two regrets in (1.3) and 1.4. This can be done simply by adding the output vectors of the two algorithms and setting the resulting

[5] The existence of an algorithm that is simultaneously parameter-free and scale-free was posed as an open problem in [Orabona and Pál, 2016b].

vector as the output of the new algorithm; this follows from the aggregation property of parameter-free algorithms we mentioned earlier (see [Cutkosky, 2019b]).

In summary, for bounded OCO, we have developed an algorithm based on MetaGrad that can automatically adapt to the Lipschitz constant, the parameters of curvature of the loss functions, and the horizon while achieving the best possible regret guarantee (up to constant factors). In unbounded OCO, we have developed scale-free and parameter-free algorithms with state-of-the-art regret guarantees useful in many applications. We have also characterized the cost of being simultaneously scale-free and parameter-free. We now move to the statistical learning setting.

1.2 Adaptivity in Statistical Learning

Another prominent setting in machine learning where adaptivity is crucial is statistical learning. In contrast with the online setting, it is typical that the whole data be available at once in statistical learning. The goal is to choose a hypothesis in some set that optimizes the expectation of a loss function. Since the data-generating distribution (over which the expectation of the loss is taken) is typically unknown, one resorts to choosing a hypothesis that optimizes some performance measure on empirical samples (e.g. minimizing the cumulative empirical loss). However, to assess the generalization performance of the selected hypothesis beyond the observed instances, one needs bounds on the expected loss over the unknown data-generating distribution—also known as the *population risk*. Used for this purpose are bounds on the difference between empirical and population risks, which are available for various learning settings and come in different flavours [Reid, 2017].

Some of the most popular such bounds are those involving uniform convergence, which, as the name suggests, bound the difference between empirical and population risks for all hypotheses simultaneously [Bousquet et al., 2004]. Such bounds typically involve the Rademacher complexity of the hypothesis set. They are looser than some alternatives we discuss below since they do not account for the nature of the learning algorithm that chooses the hypotheses. Examples of bounds that *do* take into account the learning algorithm are those that rely on uniform stability [Bousquet and Elisseeff, 2002]; in this case, the bounds become smaller the more "stable" the algorithm is. Uniform stability is known to hold in some simple settings where the hypothesis set is typically embedded in an Euclidean space, and the loss is a convex function of the hypotheses (see [Bousquet and Elisseeff, 2002]). However, relying on stability limits the type of models and algorithms that one can use in practice. When considering larger models such as Neural Networks trained with gradient-based methods, the stability condition does not hold in general[6].

Data-dependent generalization bounds. To circumvent the limitations of classical uniform stability while still accounting for the nature of the algorithm that picks the

[6]There does exist techniques involving a careful choice of learning rate for gradient descent that ensure some degree of stability, see e.g. [Hardt et al., 2016].

hypothesis, one can aim for generalization bounds with data-dependent terms that automatically make the bound small when the algorithm is stable. It is sometimes also desirable to take advantage of any "easiness" of the statistical problem at hand. For example, when the data generating distribution has some favorable structure, it is desirable to have generalization bounds that adapt to this structure and become small. One of this thesis' contributions are such data-dependent bounds (see Chapter 5). These bounds are of PAC-Bayesian type and are known to be tighter than those based on uniform convergence (see [Guedj, 2019] for an overview of PAC-Bayesian bounds). The bounds we present in Chapter 5 automatically become small when the algorithm is stable or when the learning problem is "easy" as characterized by the Bernstein condition [Bartlett and Mendelson, 2006a].

Generalization bounds for alternative measures of risk. So far, we have only considered the *expected* risk as a performance measure in statistical learning. However, in many modern machine learning applications, the expected performance is not always the most suitable measure [Williamson and Menon, 2019; Chow et al., 2015; Ahmadi et al., 2021]. This is the case for applications where there are low-probability events that have severe consequences and are to be avoided. Take the medical field, for example, where the task is to select a vaccine among a set of candidates. In this case, the *average* efficacy and safety of the candidate vaccines is not an appropriate measure; one of the priorities, in this case, is likely to be avoiding any occurrence of severe side effects. However, if these side effects happen with a low enough probability, the average performance may not capture them effectively. Other applications involving such a trade-off between event severity and probability of occurrence include autonomous driving, risk of exposure to toxic compounds, etc. Given this limitation of the expected risk, alternative risk measures, such as Coherent Risk Measures (CRMs), are garnering more and more interest in the machine learning community. CRMs possess properties that make them desirable in many risk-sensitive applications (see e.g. [Lerasle et al., 2019; Agrawal et al., 2020a]). One prominent CRM is the Conditional Value at Risk (CVaR). The CVaR with parameter $\alpha \in [0, 1]$ of a random variable X is the expectation of X conditioned on it being greater than its $(1 - \alpha)$-quantile [Rockafellar, 1997]. CVaR is in a way special as essentially all coherent risk measures can be written in terms of it through the Kusuoka representation [Kusuoka, 2001].

In order to use CVaR in the context of learning and decision making, the ability to estimate the population CVaR from empirical samples is crucial. A starting point for this is through concentration inequalities—just as in the standard setting with the expectation. Concentration inequalities for CVaR have been developed and improved over the past decade in the works of Brown [2007]; Wang and Gao [2010]; Prashanth and Ghavamzadeh [2013]; Thomas and Learned-Miller [2019a]; Kolla et al. [2019a]. These inequalities can be used together with uniform convergence arguments to produce generalization bounds for CVaR relevant in the statistical learning setting. However, there are two shortcomings of proceeding in this fashion. As mentioned a few paragraphs earlier, uniform convergence bounds are typically not as

tight as their PAC-Bayesian counterparts. What is more, simply applying the classical PAC-Bayesian analysis due to McAllester [2003] to existing concentration inequalities for CVaR will also yield loose bounds; such bounds will be off by a "Jensen gap" [Mhammedi et al., 2020b]. The second shortcoming is that existing concentration inequalities for CVaR have a sub-optimal dependence in the quantile level α in many cases, making them loose in the first place. In this thesis, we will address the above issues by simultaneously improving the dependence in the quantile level α in the concentration bounds for CVaR—making the dependence optimal for many types of distributions—and developing tight PAC-Bayesian bounds for CVaR (see Chapters 6 and 7). We achieve these results by reducing the task of estimating CVaR to that of estimating a standard expectation from empirical means.

Risk monotonicity in statistical learning Risk monotonicity in statistical learning is another machine learning topic that gained popularity in recent years [Viering et al., 2019a,b; Viering and Loog, 2021]. Risk monotonicity is concerned with the curve of the expected risk of a given hypothesis (expected loss over test samples) as a function of the number of samples used to learn the hypothesis. It is tempting to think that increasing the sample size by, say one, will result in an updated hypothesis that decreases the risk in expectation. However, this does not seem to be true even for the most natural hypothesis, such as the Empirical Risk Minimizer (ERM). In fact, even though the risk of ERM typically converges to the minimum risk as the sample size grows, the intermediate behaviour can be somewhat arbitrary [Loog et al., 2019]. Within the empirical community, the non-monotonic behaviour of the risk has been witnessed through a phenomenon called *double descent* [Belkin et al., 2019; Mei and Montanari, 2019; Nakkiran, 2019; Nakkiran et al., 2020a; Derezinski et al., 2020; Chen et al., 2020; Nakkiran et al., 2020b], where the risk curve initially drops with the number of samples, then goes up and peaks before decreasing again. The sample size at which this peak occurs typically indicates the cross-over point where the model used switches from being over-parameterized to under-parameterized. However, over-parameterization versus under-parameterization is not the whole story. In fact, other scenarios lead to non-monotonic behaviour, such as when selecting the hypothesis based on a surrogate loss that is different from the one used for risk evaluation; doing this can lead to the *dipping* phenomenon [Loog and Duin, 2012; Loog, 2015], where the risk curve goes down to a minimum and increases after that, never reaching the minimum again.

It turns out that the peaking (double descent) and dipping phenomena do not fully characterize non-monotonic risk behaviour. Perhaps striking are the examples suggested by Loog et al. [2019] who looked at simple linear regression settings in one dimension with two instances and either square or absolute loss. They showed that for the absolute loss the risk curve can oscillate in a perpetual fashion, highlighting our current lack of understanding of generalization and confirming that non-monotonic risk behaviour is not limited to only dipping or peaking. Viering et al. [2019a] posed the open question of whether there exists a consistent algorithm that always has a monotonic risk curve. In this thesis, we answer this question in the posi-

tive by deriving the first risk monotonic algorithm in a general statistical learning setting with bounded losses (see Chapter 8). Our analysis shows that risk-monotonicity need not come at the price of a worse convergence rate to the optimal risk. In fact, we show that under general conditions, the risk of the new algorithm converges to the minimum risk at the standard Rademacher rate. One may also ask whether fast rates are achievable when the learning problem is "easy" as characterized by the Bernstein condition. By taking advantage of existing data-dependent bounds and analyses developed in Chapter 5, we show that optimal fast rates are achievable under the Bernstein condition while maintaining risk monotonicity. Finally, we study a stronger notion of monotonicity for martingale difference sequences. To develop an algorithm for this case, we derive a new data-dependent concentration inequality for martingales difference sequence. This new inequality can be viewed as an empirical version of Freedman's inequality [Freedman, 1975], or a version of the empirical Bernstein bound [Maurer and Pontil, 2009] that holds for martingale difference sequences. We derive this concentration inequality by building a new supermartingale based on the potential function of FreeGrad from Chapter 4.

1.3 Thesis Structure

This thesis is structured as a compilation of five published articles and an additional two manuscripts under review. Chapters 2, 3, and 4 are on adaptivity in online learning, and are based on the following publications:

- Chapter 2: Constant Regret, Generalized Mixability, and Mirror Descent.
  Mhammedi, Z., Williamson, R. C. (2018). **NeurIPS'18**.
- Chapter 3: Lipschitz Adaptivity with Multiple Learning Rates in Online Learning.
  Mhammedi, Z., Koolen, W. M., Van Erven, T. (2019). **COLT'19**.
- Chapter 4: Lipschitz and Comparator Norm Adaptivity in Online Learning.
  Mhammedi, Z., Koolen, W. M. (2020). In **COLT'20**.

Chapters 5, 6, 7, and 8 are on adaptivity in statistical learning, and are based on the following articles:

- Chapter 5: PAC-Bayes Un-Expected Bernstein Inequality.
  Mhammedi, Z., Grunwald, P. D., Guedj, B. (2019). **NeurIPS'19**.
- Chapter 6: PAC-Bayesian Bound for the Conditional Value at Risk.
  Mhammedi, Z., Guedj, B., and Williamson, R. C. (2020). **NeurIPS'20**
- Chapter 7: Concentration Inequalities for CVaR with Near-optimal Quantile Level Dependence.
  Mhammedi, Z., Koolen, W. M. (2021). Under review.
- Chapter 8: Risk-Monotonicity in Statistical Learning.
  Mhammedi, Z., Husain, H. (2021). Under review.

In Chapter 9, we conclude and point to some exciting future research directions.

Chapter 2

Constant Regret, Generalized Mixability, and Mirror Descent

This chapter considers the experts' setting (a special case of OCO), where we study achievable regret bounds for mixable losses. For such losses, it is a classical result that the aggregating algorithm achieves a constant regret against any given expert [Vovk, 1998]. Reid et al. [2015] introduced a generalized notion of mixability along with a generalized version of the classical aggregating algorithm. This generalized algorithm turns out to be mirror descent on the vector of experts' losses for a given choice of regularizer Φ. It was shown that when a loss is Φ-mixable (the generalized notion of mixability), a constant regret is achievable. However, characterizing when losses are mixable in this generalized sense was left as an open problem by Reid et al. [2015]. In this chapter, we give a complete characterization of the notion of generalized mixability. Surprisingly, we show that the Shannon entropy S is *fundamental* in the sense that if a loss is Φ-mixable for any entropy Φ, it must necessarily be S-mixable. What is more, the algorithm induced by the Shannon entropy (the aggregating algorithm) leads to the smallest worst-case regret among any other generalized aggregating algorithm.

Constant Regret, Generalized Mixability, and Mirror Descent

Zakaria Mhammedi
Research School of Computer Science
Australian National University and DATA61
zak.mhammedi@anu.edu.au

Robert C. Williamson
Research School of Computer Science
Australian National University and DATA61
bob.williamson@anu.edu.au

Abstract

We consider the setting of prediction with expert advice; a learner makes predictions by aggregating those of a group of experts. Under this setting, and for the right choice of loss function and "mixing" algorithm, it is possible for the learner to achieve a constant regret regardless of the number of prediction rounds. For example, a constant regret can be achieved for *mixable* losses using the *aggregating algorithm*. The *Generalized Aggregating Algorithm* (GAA) is a name for a family of algorithms parameterized by convex functions on simplices (entropies), which reduce to the aggregating algorithm when using the *Shannon entropy* S. For a given entropy Φ, losses for which a constant regret is possible using the GAA are called Φ-mixable. Which losses are Φ-mixable was previously left as an open question. We fully characterize Φ-mixability and answer other open questions posed by [13]. We show that the Shannon entropy S is fundamental in nature when it comes to mixability; any Φ-mixable loss is necessarily S-mixable, and the lowest worst-case regret of the GAA is achieved using the Shannon entropy. Finally, by leveraging the connection between the *mirror descent algorithm* and the update step of the GAA, we suggest a new *adaptive* generalized aggregating algorithm and analyze its performance in terms of the regret bound.

1 Introduction

Two fundamental problems in learning are how to aggregate information and under what circumstances can one learn fast. In this paper, we consider the problems jointly, extending the understanding and characterization of exponential mixing due to [20], who showed that not only does the "*aggregating algorithm*" learn quickly when the loss is suitably chosen, but that it is in fact a generalization of classical Bayesian updating, to which it reduces when the loss is log-loss [22]. We consider a general class of aggregating schemes, going beyond Vovk's exponential mixing, and provide a complete characterization of the mixing behavior for general losses and general mixing schemes parameterized by an arbitrary entropy function.

In the *game of prediction with expert advice* a *learner* predicts the outcome of a random variable (outcome of the *environment*) by aggregating the predictions of a pool of experts. At the end of each prediction round, the outcome of the environment is announced and the learner and experts suffer losses based on their predictions. We are interested in algorithms that the learner can use to "aggregate" the experts' predictions and minimize the *regret* at the end of the game. In this case, the regret is defined as the difference between the cumulative loss of the learner and that of the best expert in hindsight after T rounds.

The *Aggregating Algorithm* (AA) [20] achieves a constant regret — a precise notion of fast learning — for *mixable* losses; that is, the regret is bounded from above by a constant R_ℓ which depends only on the loss function ℓ and not on the number of rounds T. It is worth mentioning that mixability

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

is a weaker condition than exp-concavity, and contrary to the latter, mixability is an intrinsic, parametrization-independent notion [10].

Reid et al. [13] introduced the *Generalized Aggregating Algorithm* (GAA), going beyond the AA. The GAA is parameterized by the choice of a convex function Φ on the simplex (entropy) and reduces to the AA when Φ is the Shannon entropy. The GAA can achieve a constant regret for losses satisfying a certain condition called Φ-*mixability* (characterizing when losses are Φ-mixable was left as an open problem). This regret depends jointly on the *generalized mixability constant* η_ℓ^Φ — essentially the largest η such that ℓ is $(\frac{1}{\eta}\Phi)$-mixable — and the divergence $D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})$, where $\boldsymbol{q} \in \Delta_k$ is a prior distribution over k experts and $\boldsymbol{e}_\theta$ is the θth standard basis element of $\mathbb{R}^k$ [13]. At each prediction round, the GAA can be divided into two steps; a *substitution step* where the learner picks a prediction from a set specified by the Φ-mixability condition; and an *update step* where a new distribution $\boldsymbol{q}$ over experts is computed depending on their performance. Interestingly, this update step is exactly the *mirror descent algorithm* [17, 12] which minimizes the weighted loss of experts.

Contributions. We introduce the notion of a *support loss*; given a loss ℓ defined on any action space, there exists a proper loss $\underline{\ell}$ which shares the same Bayes risk as ℓ. When a loss is mixable, one can essentially work with a proper (support) loss instead — this will be the first stepping stone towards a characterization of (generalized) mixability.

The notion of Φ-mixable and the GAA were previously restricted to finite losses. We extend these to allow for the use of losses which can take infinite values (such as the $\log$-loss), and we show in this case that under the Φ-mixability condition a constant regret is achievable using the GAA.

For an entropy Φ and a loss ℓ, we derive a necessary and sufficient condition (Theorems 13 and 14) for ℓ to be Φ-mixable. In particular, if ℓ and Φ satisfy some regularity conditions, then ℓ is Φ-mixable if and only if $\eta_\ell\Phi - \mathrm{S}$ is convex on the simplex, where S is the Shannon entropy and η_ℓ is essentially the largest η such that ℓ is η-mixable [20, 19]. This implies that a loss ℓ is Φ-mixable only if it is η-mixable for some $\eta > 0$. This, combined with the fact that η-mixability is equivalently $(\frac{1}{\eta}\,\mathrm{S})$-mixability (Theorem 12), reflects one fundamental aspect of the Shannon entropy.

Then, we derive an explicit expression for the generalized mixability constant η_ℓ^Φ (Corollary 17), and thus for the regret bound of the GAA. This allows us to compare the regret bound R_ℓ^Φ of any entropy Φ with that of the Shannon entropy S. In this case, we show (Theorem 18) that $R_\ell^{\mathrm{S}} \leq R_\ell^\Phi$; that is, the GAA achieves the lowest worst-case regret when using the Shannon entropy — another result which reflects the fundamental nature of the Shannon entropy.

Finally, by leveraging the connection between the GAA and the mirror descent algorithm, we present a new algorithm — the *Adaptive Generalized Aggregating Algorithm* (AGAA). This algorithm consists of changing the entropy function at each prediction round similar to the *adaptive mirror descent algorithm* [17]. We analyze the performance of this algorithm in terms of its regret bound.

Layout. In §2, we give some background on loss functions and present new results (Theorem 4 and 5) based on the new notion of a *proper support loss*; we show that, as far as mixability is concerned, one can always work with a proper (support) loss instead of the original loss (which can be defined on an arbitrary action space). In §3, we introduce the notions of classical and generalized mixability and derive a characterization of Φ-mixability (Theorems 13 and 14). We then introduce our new algorithm — the AGAA — and analyze its performance. We conclude the paper by a general discussion and direction for future work. All proofs, except for that of Theorem 16, are deferred to Appendix C.

Notation. Let $m \in \mathbb{N}$. We denote $[m] \coloneqq \{1, \ldots, m\}$ and $\tilde{m} \coloneqq m - 1$. We write $\langle \cdot, \cdot \rangle$ for the standard inner product in Euclidean space. Let $\Delta_m \coloneqq \{\boldsymbol{p} \in [0, +\infty[^m \colon \langle \boldsymbol{p}, \boldsymbol{1}_m \rangle = 1\}$ be the *probability simplex* in $\mathbb{R}^m$, and let $\tilde{\Delta}_m \coloneqq \{\tilde{\boldsymbol{p}} \in [0, +\infty[^{\tilde{m}} \colon \langle \tilde{\boldsymbol{p}}, \boldsymbol{1}_{\tilde{m}} \rangle \leq 1\}$. We will extensively make use of the affine map $\amalg_m \colon \mathbb{R}^{\tilde{m}} \to \mathbb{R}^m$ defined by

$$\amalg_m(\boldsymbol{u}) \coloneqq [u_1, \ldots, u_{\tilde{m}}, 1 - \langle \boldsymbol{u}, \boldsymbol{1}_{\tilde{m}} \rangle]^\mathsf{T}. \tag{1}$$

We denote $\operatorname{int}\mathcal{C}$, $\operatorname{ri}\mathcal{C}$, and $\operatorname{rbd}\mathcal{C}$ the *interior*, *relative interior*, and *relative boundary* of a set $\mathcal{C} \in \mathbb{R}^m$, respectively [8]. The *sub-differential* of a function $f \colon \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ at $\boldsymbol{u} \in \mathbb{R}^m$ such that $f(\boldsymbol{u}) < +\infty$ is defined by ([8])

$$\partial f(\boldsymbol{u}) \coloneqq \{\boldsymbol{s}^* \in \mathbb{R}^m \colon f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \boldsymbol{s}^*, \boldsymbol{v} - \boldsymbol{u} \rangle, \forall \boldsymbol{v} \in \mathbb{R}^m\}. \tag{2}$$

Table 1 on page 9 provides a list of the main symbols used in this paper.

2 Loss Functions

In general, a loss function is a map $\ell\colon \mathcal{X} \times \mathcal{A} \to [0, +\infty]$ where $\mathcal{X}$ is an outcome set and $\mathcal{A}$ is an action set. In this paper, we only consider the case $\mathcal{X} = [n]$, i.e. finite outcome space. Overloading notation slightly, we define the mapping $\ell\colon \mathcal{A} \to [0, +\infty]^n$ by $[\ell(\boldsymbol{a})]_x = \ell(x, \boldsymbol{a}), \forall x \in [n]$ and denote $\ell_x(\cdot) \coloneqq [\ell(\cdot)]_x$. We further extend the new definition of ℓ to the set $\bigcup_{k\geq 1} \mathcal{A}^k$ such that for $x \in [n]$ and $A \coloneqq [\boldsymbol{a}_\theta]^\mathsf{T}_{1\leq\theta\leq k} \in \mathcal{A}^k$, $\ell_x(A) \coloneqq [\ell_x(\boldsymbol{a}_\theta)]^\mathsf{T}_{1\leq\theta\leq k} \in [0, +\infty]^k$. We define the *effective domain* of ℓ by $\operatorname{dom}\ell \coloneqq \{\boldsymbol{a} \in \mathcal{A}\colon \ell(\boldsymbol{a}) \in [0, +\infty[^n\}$, and the *loss surface* by $\mathcal{S}_\ell \coloneqq \{\ell(\boldsymbol{a})\colon \boldsymbol{a} \in \operatorname{dom}\ell\}$. We say that ℓ is *closed* if $\mathcal{S}_\ell$ is closed in $\mathbb{R}^n$. The *superprediction* set of ℓ is defined by $\mathscr{S}_\ell^\infty \coloneqq \{\ell(\boldsymbol{a}) + \boldsymbol{d} : (\boldsymbol{a}, \boldsymbol{d}) \in \mathcal{A} \times [0, +\infty[^n\}$. Let $\mathscr{S}_\ell \coloneqq \mathscr{S}_\ell^\infty \cap [0, +\infty[^n$ be its *finite* part.

Let $\boldsymbol{a}_0, \boldsymbol{a}_1 \in \mathcal{A}$. The prediction $\boldsymbol{a}_0$ is said to be *better* than $\boldsymbol{a}_1$ if the component-wise inequality $\ell(\boldsymbol{a}_0) \leq \ell(\boldsymbol{a}_1)$ holds and there exists some $x \in [n]$ such that $\ell_x(\boldsymbol{a}_0) < \ell_x(\boldsymbol{a}_1)$ [24]. A loss ℓ is *admissible* if for any $\boldsymbol{a} \in \mathcal{A}$ there are no better predictions.

For the rest of this paper (except for Theorem 4), we make the following assumption on losses;

Assumption 1. *ℓ is a closed, admissible loss such that* $\operatorname{dom}\ell \neq \varnothing$.

It is clear that there is no loss of generality in considering only admissible losses. The condition that ℓ is closed is a weaker version of the more common assumption that $\mathcal{A}$ is compact and that $\boldsymbol{a} \mapsto \ell(x, \boldsymbol{a})$ is continuous with respect to the extended topology of $[0, +\infty]$ for all $x \in [n]$ [9, 6]. In fact, we do not make any explicit topological assumptions on the set $\mathcal{A}$ ($\mathcal{A}$ is allowed to be open in our case). Our condition simply says that if a sequence of points on the loss surface converges in $[0, +\infty[^n$, then there exists an action in $\mathcal{A}$ whose image through the loss is equal to the limit. For example the 0-1 loss $\ell_{0\text{-}1}$ is closed, yet the map $\boldsymbol{p} \mapsto \ell_{0\text{-}1}(x, \boldsymbol{p})$ is not continuous on Δ_2, for $x \in \{0, 1\}$.

In this paragraph let $\mathcal{A}$ be the *n-simplex*, i.e. $\mathcal{A} = \Delta_n$. We define the *conditional risk* $L_\ell\colon \Delta_n \times \Delta_n \to \mathbb{R}$ by $L_\ell(\boldsymbol{p}, \boldsymbol{q}) = \mathbb{E}_{x\sim\boldsymbol{p}}[\ell_x(\boldsymbol{q})] = \langle \boldsymbol{p}, \ell(\boldsymbol{q})\rangle$ and the *Bayes risk* by $\underline{L}_\ell(\boldsymbol{p}) \coloneqq \inf_{\boldsymbol{q}\in\Delta_n} L_\ell(\boldsymbol{p}, \boldsymbol{q})$. In this case, the loss ℓ is *proper* if $\underline{L}_\ell(\boldsymbol{p}) = \langle \boldsymbol{p}, \ell(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \ell(\boldsymbol{q})\rangle$ for all $\boldsymbol{p} \neq \boldsymbol{q}$ in Δ_n (and *strictly proper* if the inequality is strict). For example, the *log*-loss $\ell_{\log}\colon \Delta_n \to [0, +\infty]^n$ is defined by $\ell_{\log}(\boldsymbol{p}) = -\log \boldsymbol{p}$, where the 'log' of a vector applies component-wise. One can easily check that $\ell_{\log}$ is strictly proper. We denote $\underline{L}_{\log}$ its Bayes risk.

The above definition of the Bayes risk is restricted to losses defined on the simplex. For a general loss $\ell\colon \mathcal{A} \to [0, +\infty]^n$, we use the following definition;

Definition 2 (Bayes Risk)**.** *Let $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a loss such that* $\operatorname{dom}\ell \neq \varnothing$. *The* Bayes risk *$\underline{L}_\ell\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ is defined by*

$$\forall \boldsymbol{u} \in \mathbb{R}^n, \quad \underline{L}_\ell(\boldsymbol{u}) \coloneqq \inf_{\boldsymbol{z}\in\mathscr{S}_\ell} \langle \boldsymbol{u}, \boldsymbol{z}\rangle . \tag{3}$$

The support function of a set $\mathcal{C} \subseteq \mathbb{R}^n$ is defined by $\sigma_{\mathcal{C}}(\boldsymbol{u}) \coloneqq \sup_{\boldsymbol{z}\in\mathcal{C}} \langle \boldsymbol{u}, \boldsymbol{z}\rangle$, $\boldsymbol{u} \in \mathbb{R}^n$, and thus it is easy to see that one can express the Bayes risk as $\underline{L}_\ell(\boldsymbol{u}) = -\sigma_{\mathscr{S}_\ell}(-\boldsymbol{u})$. Our definition of the Bayes risk is slightly different from previous ones ([9, 19, 6]) in two ways; 1) the Bayes risk is defined on all $\mathbb{R}^n$ instead of $[0, +\infty[^n$; and 2) the infimum is taken over the finite part of the superprediction set $\mathscr{S}_\ell^\infty$. The first point is a mere mathematical convenience and makes no practical difference since $\underline{L}_\ell(\boldsymbol{p}) = -\infty$ for all $\boldsymbol{p} \notin [0, +\infty[^n$. For the second point, swapping $\mathscr{S}_\ell$ for $\mathscr{S}_\ell^\infty$ in (3) does not change the value of $\underline{L}_\ell$ for mixable losses (see Appendix D). However, we chose to work with $\mathscr{S}_\ell$ — a subset of $\mathbb{R}^n$ — as it allows us to directly apply techniques from convex analysis.

Definition 3 (Support Loss)**.** *We call a map $\underline{\ell} : \Delta_n \to [0, +\infty]^n$ a* support loss *of ℓ if*

$$\forall \boldsymbol{p} \in \operatorname{ri}\Delta_n,\ \underline{\ell}(\boldsymbol{p}) \in \partial\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p});$$
$$\forall \boldsymbol{p} \in \operatorname{rbd}\Delta_n, \exists(\boldsymbol{p}_m) \subset \operatorname{ri}\Delta_n,\ \boldsymbol{p}_m \overset{m\to\infty}{\to} \boldsymbol{p} \ \textit{and}\ \underline{\ell}(\boldsymbol{p}_m) \overset{m\to\infty}{\to} \underline{\ell}(\boldsymbol{p}) \ \textit{component-wise,}$$

where $\partial\sigma_{\mathscr{S}_\ell}$ (see (2)*) is the sub-differential of the support function — $\sigma_{\mathscr{S}_\ell}$ — of the set $\mathscr{S}_\ell$.*

Theorem 4. *Any loss $\ell\colon \mathcal{A} \to [0, +\infty]^n$ such that* $\operatorname{dom}\ell \neq \varnothing$*, has a proper support loss $\underline{\ell}$ with the same Bayes risk, $\underline{L}_\ell$, as ℓ.*

Theorem 4 shows that regardless of the action space on which the loss is defined, there always exists a proper loss whose Bayes risk coincides with that of the original loss. This fact is useful in situations where the Bayes risk contains all the information one needs — such is the case for mixability. The next Theorem shows a stronger relationship between a loss and its corresponding support loss.

Theorem 5. *Let $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a loss and $\underline{\ell}$ be a proper support loss of ℓ. If the Bayes risk $\underline{L}_\ell$ is differentiable on $]0,+\infty[^n$, then $\underline{\ell}$ is uniquely defined on $\operatorname{ri}\Delta_n$ and*

$$\begin{aligned} &\forall \boldsymbol{p} \in \operatorname{dom}\underline{\ell}, \quad \exists \boldsymbol{a}_* \in \operatorname{dom}\ell, \qquad \ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p}), \\ &\forall \boldsymbol{a} \in \operatorname{dom}\ell, \quad \exists (\boldsymbol{p}_m) \subset \operatorname{ri}\Delta_n, \quad \underline{\ell}(\boldsymbol{p}_m) \overset{m\to\infty}{\to} \ell(\boldsymbol{a}) \ \textit{ component-wise.} \end{aligned}$$

Theorem 5 shows that when the Bayes risk is differentiable (a necessary condition for mixability — Theorem 12), the support loss is almost a reparametrization of the original loss, and in practice, it is enough to work with support losses instead. This will be crucial for characterizing Φ-mixability.

3 Mixability in the Game of Prediction with Expert Advice

We consider the setting of prediction with expert advice [20]; there a is pool of k experts, parameterized by $\theta \in [k]$, which make predictions $\boldsymbol{a}_\theta^t \in \mathcal{A}$ at each round t. In the same round, the learner predicts $\boldsymbol{a}_{\mathfrak{M}}^t := \mathfrak{M}\left(\boldsymbol{a}_{1:k}^t, (x^s, \boldsymbol{a}_{1:k}^s)_{1\le s<t}\right) \in \mathcal{A}$, where $\boldsymbol{a}_{1:k}^{\cdot} := [\boldsymbol{a}_\theta^{\cdot}]_{1\le\theta\le k}$, $(x^s) \subset [n]$ are outcomes of the environment, and $\mathfrak{M} : \mathcal{A}^k \times ([n] \times \mathcal{A}^k)^* \to \mathcal{A}$ is a *merging strategy* [19]. At the end of round t, x^t is announced and each expert θ [resp. learner] suffers a loss $\ell_{x^t}(\boldsymbol{a}_\theta)$ [resp. $\ell_{x^t}(\boldsymbol{a}_{\mathfrak{M}}^t)$], where $\ell\colon \mathcal{A} \to [0,+\infty]^n$. After $T > 0$ rounds, the cumulative loss of each expert θ [resp. learner] is given by $\operatorname{Loss}_\theta^\ell(T) := \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}_\theta^t)$ [resp. $\operatorname{Loss}_{\mathfrak{M}}^\ell(T) := \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}_{\mathfrak{M}}^t)$]. We say that $\mathfrak{M}$ achieves a *constant regret* if $\exists R > 0, \forall T > 0, \forall \theta \in [k], \operatorname{Loss}_{\mathfrak{M}}^\ell(T) \le \operatorname{Loss}_\theta^\ell(T) + R$. In what follows, this game setting will be referred to by $\mathfrak{G}_\ell^n(\mathcal{A}, k)$ and we only consider the case where $k \ge 2$.

3.1 The Aggregating Algorithm and η-mixability

Definition 6 (η-mixability)**.** *For $\eta > 0$, a loss $\ell\colon \mathcal{A} \to [0,+\infty]^n$ is said to be η-mixable, if $\forall \boldsymbol{q} \in \Delta_k$,*

$$\forall \boldsymbol{a}_{1:k} \in \mathcal{A}^k, \exists \boldsymbol{a}_* \in \mathcal{A}, \forall x \in [n], \quad \ell_x(\boldsymbol{a}_*) \le -\eta^{-1} \log \langle \boldsymbol{q}, \exp(-\eta \ell_x(\boldsymbol{a}_{1:k})) \rangle \,, \tag{4}$$

where the exp *applies component-wise. Letting $\mathfrak{H}_\ell := \{\eta > 0\colon \ell \text{ is } \eta\text{-mixable}\}$, we define the* mixability constant *of ℓ by $\eta_\ell := \sup \mathfrak{H}_\ell$ if $\mathfrak{H}_\ell \neq \varnothing$; and* 0 *otherwise. ℓ is said to be* mixable *if $\eta_\ell > 0$.*

If a loss ℓ is η-mixable for $\eta > 0$, the AA (Algorithm 1) achieves a constant regret in the $\mathfrak{G}_\ell^n(\mathcal{A}, k)$ game[20]. In Algorithm 1, the map $\mathfrak{S}_\ell\colon \mathscr{S}_\ell^\infty \to \mathcal{A}$ is a *substitution function* of the loss ℓ [20, 10]; that is, $\mathfrak{S}_\ell$ satisfies the component-wise inequality $\ell(\mathfrak{S}_\ell(\boldsymbol{s})) \le \boldsymbol{s}$, for all $\boldsymbol{s} \in \mathscr{S}_\ell^\infty$.

It was shown by Chernov et al. [6] that the η-mixability condition (4) is equivalent to the convexity of the *η-exponentiated* superprediction set of ℓ defined by $\exp(-\eta\mathscr{S}_\ell^\infty) := \{\exp(-\eta\boldsymbol{s})\colon \boldsymbol{s} \in \mathscr{S}_\ell^\infty\}$. Using this fact, van Erven et al. [19] showed that the mixability constant η_ℓ of a strictly proper loss $\ell\colon \Delta_n \to [0,+\infty[^n$, whose Bayes risk is twice continuously differentiable on $]0,+\infty[^n$, is equal to

$$\underline{\eta_\ell} := \inf_{\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n} \left(\lambda_{\max}([\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}}))\right)^{-1}, \tag{5}$$

where H is the Hessian operator and $\tilde{\underline{L}}_{\cdot} := \underline{L}_{\cdot} \circ \Pi_n$ (Π_n was defined in (1)). The next theorem extends this result by showing that the mixability constant η_ℓ of any loss ℓ is lower bounded by $\underline{\eta_\ell}$ in (5), as long as ℓ satisfies Assumption 1 and its Bayes risk is twice differentiable.

Theorem 7. *Let $\eta > 0$ and $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a loss. Suppose that $\operatorname{dom}\ell = \mathcal{A}$ and that $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$. If $\underline{\eta_\ell} > 0$ then ℓ is $\underline{\eta_\ell}$-mixable. In particular, $\eta_\ell \ge \underline{\eta_\ell}$.*

We later show that, under the same conditions as Theorem 7, we actually have $\eta_\ell = \underline{\eta_\ell}$ (Theorem 16) which indicates that the Bayes risk contains all the information necessary to characterize mixability.

Remark 8. *In practice, the requirement '$\operatorname{dom}\ell = \mathcal{A}$' is not necessarily a strict restriction to finite losses; it is often the case that a loss $\bar{\ell} : \bar{\mathcal{A}} \to [0,+\infty]^n$ only takes infinite values on the relative boundary of $\bar{\mathcal{A}}$ (such is the case for the* log*-loss defined on the simplex), and thus the restriction $\ell := \bar{\ell}|_{\mathcal{A}}$, where $\mathcal{A} = \operatorname{ri}\bar{\mathcal{A}}$, satisfies $\operatorname{dom}\ell = \mathcal{A}$. It follows trivially from the definition of mixability* (4) *that if ℓ is η-mixable and $\bar{\ell}$ is continuous with respect to the extended topology of $[0,+\infty]^n$ — a condition often satisfied — then $\bar{\ell}$ is also η-mixable.*

3.2 The Generalized Aggregating Algorithm and $(\eta, \Phi)-$mixability

A function $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ is an *entropy* if it is convex, its epigraph $\operatorname{epi}\Phi := \{(\boldsymbol{u}, h)\colon \Phi(\boldsymbol{u}) \leq h\}$ is closed in $\mathbb{R}^k \times \mathbb{R}$, and $\Delta_k \subseteq \operatorname{dom}\Phi := \{\boldsymbol{u} \in \mathbb{R}^k\colon \Phi(\boldsymbol{u}) < +\infty\}$. For example, the *Shannon entropy* is defined by $\mathrm{S}(\boldsymbol{q}) = +\infty$ if $\boldsymbol{q} \notin [0, +\infty[^k$, and

$$\forall \boldsymbol{q} \in [0, +\infty[^k, \quad \mathrm{S}(\boldsymbol{q}) = \sum_{i \in [k]\colon q_i \neq 0} q_i \log q_i, \tag{6}$$

The *divergence* generated by an entropy Φ is the map $D_\Phi\colon \mathbb{R}^n \times \operatorname{dom}\Phi \to [0, +\infty]$ defined by

$$D_\Phi(\boldsymbol{v}, \boldsymbol{u}) := \begin{cases} \Phi(\boldsymbol{v}) - \Phi(\boldsymbol{u}) - \Phi'(\boldsymbol{u}; \boldsymbol{v} - \boldsymbol{u}), & \text{if } \boldsymbol{v} \in \operatorname{dom}\Phi; \\ +\infty, & \text{otherwise.} \end{cases} \tag{7}$$

where $\Phi'(\boldsymbol{u}; \boldsymbol{v} - \boldsymbol{u}) := \lim_{\lambda \downarrow 0} [\Phi(\boldsymbol{u} + \lambda(\boldsymbol{v} - \boldsymbol{u})) - \Phi(\boldsymbol{u})]/\lambda$ (the limit exists since Φ is convex [15]).

Definition 9 (Φ-mixability)**.** *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy. A loss $\ell\colon \mathcal{A} \to [0, +\infty]^n$ is (η, Φ)-mixable for $\eta > 0$ if $\forall \boldsymbol{q} \in \Delta_k$, $\forall \boldsymbol{a}_{1:k} \in \mathcal{A}^k$, $\exists \boldsymbol{a}_* \in \mathcal{A}$, such that*

$$\forall x \in [n],\ \ell_x(\boldsymbol{a}_*) \leq \mathrm{Mix}^\eta_\Phi(\ell_x(\boldsymbol{a}_{1:k}), \boldsymbol{q}) := \inf_{\hat{\boldsymbol{q}} \in \Delta_k} \langle \hat{\boldsymbol{q}}, \ell_x(\boldsymbol{a}_{1:k}) \rangle + \eta^{-1} D_\Phi(\hat{\boldsymbol{q}}, \boldsymbol{q}). \tag{8}$$

When $\eta = 1$, we simply say that ℓ is Φ-mixable and we denote $\mathrm{Mix}_\Phi := \mathrm{Mix}^1_\Phi$. Letting $\mathfrak{H}^\Phi_\ell := \{\eta > 0\colon \ell$ is (η, Φ)-mixable$\}$, we define the generalized mixability constant *of (ℓ, Φ) by $\eta^\Phi_\ell := \sup \mathfrak{H}^\Phi_\ell$, if $\mathfrak{H}^\Phi_\ell \neq \varnothing$; and 0 otherwise.*

Reid et al. [13] introduced the GAA (see Algorithm 2) which uses an entropy function $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ and a substitution function $\mathfrak{S}_\ell$ (see previous section) to specify the learner's merging strategy $\mathfrak{M}$. It was shown that the GAA reduces to the AA when Φ is the Shannon entropy S. It was also shown that under some regularity conditions on Φ, the GAA achieves a constant regret in the $\mathfrak{G}^n_\ell(\mathcal{A}, k)$ game for any finite, (η, Φ)-mixable loss.

Our definition of Φ-mixability differs slightly from that of Reid et al. [13] — we use directional derivatives to define the divergence D_Φ. This distinction makes it possible to extend the GAA to losses which can take infinite values (such as the $\log$-loss defined on the simplex). We show, in this case, that a constant regret is still achievable under the (η, Φ)-mixability condition. Before presenting this result, we define the notion of Δ-*differentiability*; for $\mathfrak{l} \subseteq [k]$, let $\Delta_\mathfrak{l} := \{\boldsymbol{q} \in \Delta_k\colon q_\theta = 0, \forall \theta \notin \mathfrak{l}\}$. We say that an entropy Φ is Δ-differentiable if $\forall \mathfrak{l} \subseteq [k]$, $\forall \boldsymbol{u}, \boldsymbol{u}_0 \in \operatorname{ri}\Delta_\mathfrak{l}$, the map $\boldsymbol{z} \mapsto \Phi'(\boldsymbol{u}; \boldsymbol{z})$ is linear on $\mathcal{L}^0_\mathfrak{l} := \{\lambda(\boldsymbol{v} - \boldsymbol{u}_0)\colon (\lambda, \boldsymbol{v}) \in \mathbb{R} \times \Delta_\mathfrak{l}\}$.

Theorem 10. *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a Δ-differentiable entropy. Let $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a loss (not necessarily finite) such that $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$. If ℓ is (η, Φ)-mixable then the* GAA *achieves a constant regret in the $\mathfrak{G}^n_\ell(\mathcal{A}, k)$ game; for any sequence $(x^t, \boldsymbol{a}^t_{1:k})_{t=1}^T$,*

$$\mathrm{Loss}^\ell_{\mathrm{GAA}}(T) - \min_{\theta \in [k]} \mathrm{Loss}^\ell_\theta(T) \leq R^\Phi_\ell := \inf_{\boldsymbol{q} \in \Delta_k} \max_{\theta \in [k]} D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})/\eta^\Phi_\ell, \tag{9}$$

for initial distribution over experts $\boldsymbol{q}^0 = \operatorname{argmin}_{\boldsymbol{q} \in \Delta_k} \max_{\theta \in [k]} D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})$, where $\boldsymbol{e}_\theta$ is the θth basis element of $\mathbb{R}^k$, and any substitution function $\mathfrak{S}_\ell$.

Looking at Algorithm 2, it is clear that the GAA is divided into two steps; 1) a *substitution step* which consists of finding a prediction $\boldsymbol{a}_* \in \mathcal{A}$ satisfying the mixability condition (8) using a substitution function $\mathfrak{S}_\ell$; and 2) an *update step* where a new distribution over experts is computed. Except for the case of the AA with the $\log$-loss (which reduces to Bayesian updating [22]), there is not a unique choice of substitution function in general. An example of substitution function $\mathfrak{S}_\ell$ is the *inverse loss* [23]. Kamalaruban et al. [10] discuss other alternatives depending on the curvature of the Bayes risk. Although the choice of $\mathfrak{S}_\ell$ can affect the performance of the algorithm to some extent [10], the regret bound in (9) remains unchanged regardless of $\mathfrak{S}_\ell$. On the other hand, the update step is well defined and corresponds to a *mirror descent step* [13] (we later use this fact to suggest a new algorithm).

Algorithm 1: Aggregating Algorithm

input : $\boldsymbol{q}^0 \in \Delta_k$; $\eta > 0$; A η-mixable loss $\ell\colon \mathcal{A} \to [0, +\infty]^n$; A substitution function $\mathfrak{S}_\ell$.

output : Learner's predictions $(\boldsymbol{a}_*^t)$

for $t = 1$ **to** T **do**

  Observe $A^t = \boldsymbol{a}_{1:k}^t \in \mathcal{A}^k$;

  $\boldsymbol{a}_*^t \leftarrow \mathfrak{S}_\ell\left(-\frac{1}{\eta}\log\sum_{\theta\in[k]} q_\theta^{t-1} e^{-\eta\ell(\boldsymbol{a}_\theta^t)}\right)$;

  Observe outcome $x^t \in [n]$;

  $q_\theta^t \leftarrow \dfrac{q_\theta^{t-1}\exp(-\eta\ell_{x^t}(\boldsymbol{a}_\theta^t))}{\langle \boldsymbol{q}^{t-1}, \exp(-\eta\ell_{x^t}(A^t))\rangle}, \forall\theta \in [k]$;

end

Algorithm 2: Generalized Aggregating Algorithm

input : $\boldsymbol{q}^0 \in \Delta_k$; A Δ-differentiable entropy $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$; $\eta > 0$; A (η, Φ)-mixable loss $\ell\colon \mathcal{A} \to [0, +\infty]^n$; A substitution function $\mathfrak{S}_\ell$.

output : Learner's predictions $(\boldsymbol{a}_*^t)$

for $t = 1$ **to** T **do**

  Observe $A^t = \boldsymbol{a}_{1:k}^t \in \mathcal{A}^k$;

  $\boldsymbol{a}_*^t \leftarrow \mathfrak{S}_\ell\left(\left[\mathrm{Mix}_\Phi^\eta(\ell_x(A^t), \boldsymbol{q}^{t-1})\right]_{1\le x\le n}^\top\right)$;

  Observe outcome $x^t \in [n]$;

  $\boldsymbol{q}^t \leftarrow \operatorname*{argmin}_{\boldsymbol{\mu}\in\Delta_k} \langle \boldsymbol{\mu}, \ell_{x^t}(A^t)\rangle + \frac{1}{\eta} D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}^{t-1})$;

end

We conclude this subsection with two new and important results which will lead to a characterization of Φ-mixability. The first result shows that (η, S)-mixability is equivalent to η-mixability, and the second rules out losses and entropies for which Φ-mixability is not possible.

Theorem 11. *Let $\eta > 0$. A loss $\ell\colon \mathcal{A} \to [0, +\infty]^n$ is η-mixable if and only if ℓ is (η, S)-mixable.*

Proposition 12. *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ be an entropy and $\ell\colon \mathcal{A} \to [0, +\infty]^n$. If ℓ is Φ-mixable, then the Bayes risk satisfies $\underline{L}_\ell \in C^1(]0, +\infty[^n)$. If, additionally, $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$, then Φ must be strictly convex on Δ_k.*

It should be noted that since the Bayes risk of a loss ℓ must be differentiable for it to be Φ-mixable for some entropy Φ, Theorem 5 says that we can essentially work with a proper support loss $\underline{\ell}$ of ℓ. This will be crucial in the proof of the sufficient condition of Φ-mixability (Theorem 14).

3.3 A Characterization of Φ-Mixability

In this subsection, we first show that given an entropy $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ and a loss $\ell\colon \mathcal{A} \to [0, +\infty]^n$ satisfying certain regularity conditions, ℓ is Φ-mixable if and only if

$$\boxed{\underline{\eta_\ell}\Phi - \mathrm{S} \text{ is convex on } \Delta_k.} \tag{10}$$

Theorem 13. *Let $\eta > 0$, $\ell\colon \mathcal{A} \to [0, +\infty]^n$ a η-mixable loss, and $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ an entropy. If $\eta\Phi - \mathrm{S}$ is convex on Δ_k, then ℓ is Φ-mixable.*

The converse of Theorem 13 also holds under additional smoothness conditions on Φ and ℓ;

Theorem 14. *Let $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a loss such that $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$, and $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ an entropy such that $\tilde{\Phi} := \Phi \circ \Pi_k$ is twice differentiable on $\operatorname{int}\tilde{\Delta}_k$. Then ℓ is Φ-mixable only if $\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k.*

As consequence of Theorem 14, if a loss ℓ is not classically mixable, i.e. $\eta_\ell = 0$, it cannot be Φ-mixable for any entropy Φ. This is because $\eta_\ell\Phi - \mathrm{S} \overset{*}{=} \underline{\eta_\ell}\Phi - \mathrm{S} = -\mathrm{S}$ is not convex (where equality '*' is due to Theorem 7).

We need one more result before arriving at (10); Recall that the mixability constant η_ℓ is defined as the supremum of the set $\mathfrak{H}_\ell := \{\eta \ge 0\colon \ell \text{ is } \eta\text{-mixable}\}$. The next lemma essentially gives a sufficient condition for this supremum to be attained when $\mathfrak{H}_\ell$ is non-empty — in this case, ℓ is η_ℓ-mixable.

Lemma 15. *Let $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a loss. If $\operatorname{dom}\ell = \mathcal{A}$, then either $\mathfrak{H}_\ell = \varnothing$ or $\eta_\ell \in \mathfrak{H}_\ell$.*

Theorem 16. *Let ℓ and Φ be as in Theorem 14 with $\operatorname{dom}\ell = \mathcal{A}$. Then $\eta_\ell = \underline{\eta_\ell}$. Furthermore, ℓ is Φ-mixable if and only if $\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k.*

Proof. Suppose now that ℓ is mixable. By Lemma 15, it follows that ℓ is η_ℓ-mixable, and from Theorem 11, ℓ is $(\eta_\ell^{-1}\,\mathrm{S})$-mixable. Substituting Φ for $\eta_\ell^{-1}\,\mathrm{S}$ in Theorem 14 implies that $(\underline{\eta_\ell}/\eta_\ell - 1)\,\mathrm{S}$ is convex on $\operatorname{ri}\Delta_k$. Thus, $\eta_\ell \leq \underline{\eta_\ell}$, and since from Theorem 7 $\underline{\eta_\ell} \leq \eta_\ell$, we conclude that $\eta_\ell = \underline{\eta_\ell}$.

From Theorem 14, if ℓ is Φ-mixable then $\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k. Now suppose that $\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k. This implies that $\underline{\eta_\ell} > 0$, and thus from Theorem 7, ℓ is $\underline{\eta_\ell}$-mixable. Now since ℓ is $\underline{\eta_\ell}$-mixable and $\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k, Theorem 13 implies that ℓ is Φ-mixable. □

Note that the condition '$\operatorname{dom}\ell = \mathcal{A}$' is in practice not a restriction to finite losses — see Remark 8. Theorem 16 implies that under the regularity conditions of Theorem 14, the Bayes risk $\underline{L}_\ell$ [resp. $(\underline{L}_\ell, \Phi)$] contains all necessary information to characterize classical [resp. generalized] mixability.

Corollary 17 (The Generalized Mixability Constant)**.** *Let ℓ and Φ be as in Theorem 16. Then the generalized mixability constant (see Definition 9) is given by*

$$\eta_\ell^\Phi = \underline{\eta_\ell} \inf_{\tilde{\boldsymbol{q}} \in \operatorname{int}\tilde{\Delta}_k} \lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}), \tag{11}$$

where $\tilde{\Phi} := \Phi \circ \Pi_k, \tilde{\mathrm{S}} = \mathrm{S} \circ \Pi_k$, and Π_k is defined in (1).

Observe that when $\Phi = \mathrm{S}$, (11) reduces to $\eta_\ell^{\mathrm{S}} = \underline{\eta_\ell}$ as expected from Theorem 11 and Theorem 16.

3.4 The (In)dependence Between ℓ and Φ and the Fundamental Nature of S

So far, we showed that the Φ-mixability of losses satisfying Assumption 1 is characterized by the convexity of $\eta\Phi - \mathrm{S}$, where $\eta \in]0, \eta_\ell]$ (see Theorems 13 and 14). As a result, and contrary to what was conjectured previously [13], the generalized mixability condition does not induce a correspondence between losses and entropies; for a given loss ℓ, there is no particular entropy Φ^ℓ — specific to the choice of ℓ — which minimizes the regret of the GAA. Rather, the Shannon entropy S minimizes the regret regardless of the choice of ℓ (see Theorem 18 below). This reflects one fundamental aspect of the Shannon entropy.

Nevertheless, given a loss ℓ and entropy Φ, the curvature of the loss surface $\mathcal{S}_\ell$ determines the maximum 'learning rate' η_ℓ^Φ of the GAA; the curvature of $\mathcal{S}_\ell$ is linked to $\underline{\eta_\ell}$ through the Hessian of the Bayes risk (see Theorem 49 in Appendix H.2), which is in turn linked to η_ℓ^Φ through (11).

Given a loss ℓ, we now use the expression of η_ℓ^Φ in (11) to explicitly compare the regret bounds R_ℓ^Φ and R_ℓ^{S} achieved with the GAA (see (9)) using entropy Φ and the Shannon entropy S, respectively.

Theorem 18. *Let $\mathrm{S}, \Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$, where S is the Shannon entropy and Φ is an entropy such that $\tilde{\Phi} := \Phi \circ \Pi_k$ is twice differentiable on $\operatorname{int}\tilde{\Delta}_k$. A loss $\ell \colon \mathcal{A} \to [0, +\infty[^n$ with $\underline{L}_\ell$ twice differentiable on $]0, +\infty[^n$, is Φ-mixable only if $R_\ell^{\mathrm{S}} \leq R_\ell^\Phi$.*

Theorem 18 is consistent with Vovk's result [20, §5] which essentially states that the regret bound $R_\ell^{\mathrm{S}} = \eta_\ell^{-1}\log k$ is in general tight for η-mixable losses.

4 Adaptive Generalized Aggregating Algorithm

In this section, we take advantage of the similarity between the GAA's update step and the mirror descent algorithm (see Appendix E) to devise a modification to the GAA leading to improved regret bounds in certain cases. The GAA can be modified in (at least) two immediate ways; 1) changing the learning rate at each time step to speed-up convergence; and 2) changing the entropy, i.e. the regularizer Φ, at each time step — similar to the *adaptive* mirror descent algorithm [17, 12]. In the former case, one can use Corollary 17 to calculate the maximum 'learning rate' under the Φ-mixability constraint. Here, we focus on the second method; changing the entropy at each round. Algorithm 3 displays the modified GAA — which we call the *Adaptive Generalized Aggregating Algorithm* (AGAA) — in its most general form. In Algorithm 3, $\Phi^\star(\boldsymbol{z}) := \sup_{\boldsymbol{q}\in\Delta_k}\langle \boldsymbol{q}, \boldsymbol{z}\rangle - \Phi(\boldsymbol{q})$ is the *entropic dual* of Φ.

Given a (η, Φ)-mixable loss ℓ, we verify that Algorithm 3 is well defined; for simplicity, assume that $\operatorname{dom}\ell = \mathcal{A}$ and $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$. From the definition of an entropy, $|\Phi| < +\infty$ on Δ_k, and thus the entropic dual $\Phi_t^\star$ is defined and finite on all $\mathbb{R}^k$ (in particular at $\boldsymbol{\theta}^t$). On the

Algorithm 3: Adaptive Generalized Aggregating Algorithm (AGAA)

input : $\boldsymbol{\theta}^1 = \mathbf{0} \in \mathbb{R}^k$; A Δ-differentiable entropy $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$; $\eta > 0$; A (η, Φ)-mixable loss $\ell\colon \mathcal{A} \to [0, +\infty[^n$; A substitution function $\mathfrak{S}_\ell$; A protocol of choosing $\boldsymbol{\beta}^t$ at round t.

output : Learner's predictions $(\boldsymbol{a}_*^t)$

for $t = 1$ **to** T **do**

  Let $\Phi_t(\boldsymbol{w}) \coloneqq \Phi(\boldsymbol{w}) - \langle \boldsymbol{w}, \boldsymbol{\beta}^t - \boldsymbol{\theta}^t \rangle$; `// New entropy`

  Observe $A^t \coloneqq \boldsymbol{a}_{1:k}^t \in \mathcal{A}^k$; `// Experts' predictions`

  $\boldsymbol{a}_*^t \leftarrow \mathfrak{S}_\ell\left(\left[\mathrm{Mix}_{\Phi_t}^\eta(\ell_x(A^t), \nabla\Phi_t^\star(\boldsymbol{\theta}^t))\right]_{1\le x\le n}^\top\right)$; `// Learner's prediction`

  Observe $x^t \in [n]$ and pick some $\boldsymbol{v}^t \in \mathbb{R}^k$;

  $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta\ell_{x^t}(A^t)$;

end

other hand, from Proposition 12, Φ is strictly convex on Δ_k which implies that $\Phi^\star$ (and thus $\Phi_t^\star$) is differentiable on $\mathbb{R}^k$ (see e.g. [8, Thm. E.4.1.1]). It remains to check that ℓ is (η, Φ_t)-mixable. Since for $\eta > 0$, (η, Φ_t)-mixability is equivalent to $(\frac{1}{\eta}\Phi_t)$-mixability (by definition), Theorem 16 implies that ℓ is (η, Φ_t)-mixable if and only if $\underline{\eta_\ell}\eta^{-1}\Phi_t - \mathrm{S}$ is convex on Δ_k. This is in fact the case since Φ_t is an affine transformation of Φ, and we have assumed that ℓ is (η, Φ)-mixable.

In what follows, we focus on a particular instantiation of Algorithm 3 where we choose $\boldsymbol{\beta}^t \coloneqq -\eta\sum_{s=1}^{t-1}(\ell_{x^s}(A^s) + \boldsymbol{v}^s)$, for some (arbitrary for now) $(\boldsymbol{v}^s) \subset \mathbb{R}^k$. The $(\boldsymbol{v}^t)$ vectors act as correction terms in the update step of the AGAA. Using standard duality properties (see Appendix A), it is easy to show that the AGAA reduces to the GAA except for the update step where the new distribution over experts at round $t \in [T]$ is now given by

$$\boldsymbol{q}^t = \nabla\Phi^\star(\nabla\Phi(\boldsymbol{q}^{t-1}) - \eta\ell_{x^t}(A^t) - \eta\boldsymbol{v}^t).$$

Theorem 19. *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a Δ-differentiable entropy. Let $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a loss such that $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$. Let $\boldsymbol{\beta}^t = -\eta\sum_{s=1}^{t-1}(\ell_{x^s}(A^s) + \boldsymbol{v}^s)$, where $\boldsymbol{v}^s \in \mathbb{R}^k$ and $A^s \coloneqq \boldsymbol{a}_{1:k}^s \in \mathcal{A}^k$. If ℓ is (η, Φ)-mixable then for initial distribution $\boldsymbol{q}^0 = \mathrm{argmin}_{\boldsymbol{q}\in\Delta_k}\max_{\theta\in[k]} D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})$ and any sequence $(x^t, \boldsymbol{a}_{1:k}^t)_{t=1}^T$, the* AGAA *achieves the regret*

$$\forall\theta \in [k], \quad \mathrm{Loss}_{\mathrm{AGAA}}^\ell(T) - \mathrm{Loss}_\theta^\ell(T) \le R_\ell^\Phi + \Delta R_\theta(T), \tag{12}$$

where $\Delta R_\theta(T) \coloneqq \sum_{t=1}^{T-1}(v_\theta^t - \langle \boldsymbol{v}^t, \boldsymbol{q}^t \rangle)$.

Theorem 19 implies that if the sequence $(\boldsymbol{v}^t)$ is chosen such that $\Delta R_{\theta^*}(T)$ is negative for the best expert θ^* (in hindsight), then the regret bound '$R_\ell^\Phi + \Delta R_{\theta^*}(T)$' of the AGAA is lower than that of the GAA (see (9)), and ultimately that of the AA (when $\Phi = \mathrm{S}$). Unfortunately, due to Vovk's result [20, §5] there is no "universal" choice of $(\boldsymbol{v}^t)$ which guarantees that $\Delta R_{\theta^*}(T)$ is always negative. However, there are cases where this term is expected to be negative.

Consider a dataset where it is typical for the best experts (i.e., the θ^*'s) to perform poorly at some point during the game, as measured by their average loss, for example. Under such an assumption, choosing the correction vectors $\boldsymbol{v}^t$ to be negatively proportional to the average losses of experts, i.e. $\boldsymbol{v}^t \coloneqq -\frac{\alpha}{t}\sum_{s=1}^t \ell_{x^s}(A^s)$ (for small enough $\alpha > 0$), would be consistent with the idea of making $\Delta R_{\theta^*}(T)$ negative. To see this, suppose expert θ^* is performing poorly during the game (say at $t < T$), as measured by its instantaneous and average loss. At that point the distribution $\boldsymbol{q}^t$ would put more weight on experts performing better than θ^*, i.e. having a lower average loss. And since v_θ^t is negatively proportional to the average loss of expert θ, the quantity $v_{\theta^*}^t - \langle \boldsymbol{v}^t, \boldsymbol{q}^t \rangle$ would be negative — consistent with making $\Delta R_{\theta^*}(T) < 0$. On the other hand, if expert θ^* performs well during the game (say close to the best) then $v_{\theta^*}^t - \langle \boldsymbol{v}^t, \boldsymbol{q}^t \rangle \simeq 0$, since $\boldsymbol{q}^t$ would put comparable weights between θ^* and other experts (if any) with similar performance.

Example 1. *(A Negative Regret).* One can construct an example that illustrates the idea above. Consider the Brier game $\mathfrak{G}_{\ell_{\mathrm{Brier}}}^2(\Delta_2, 2)$; a probability game with 2 experts $\{\theta_1, \theta_2\}$, 2 outcomes $\{0, 1\}$, and where the loss ℓ_{Brier} is the *Brier* loss [21] (which is 1-mixable). Assume that; expert θ_1 consistently predicts $\Pr(x = 0) = 1/2$; expert θ_2 predicts $\Pr(x = 0) = 1/4$ during the first 50 rounds, then

switches to predicting $\Pr(x = 0) = 3/4$ thereafter; the outcome is always $x = 0$. A straightforward simulation using the AGAA with the Shannon entropy, Vovk's substitution function for the Brier loss [21], $\boldsymbol{\beta}^t$ as in Theorem 19 with $\boldsymbol{v}^t := -\frac{1}{8t}\sum_{s=1}^{t} \ell_{\mathrm{Brier}}(x^s, A^s)$, yields $R^{\Phi}_{\ell_{\mathrm{Brier}}} + \Delta R_{\theta^*}(T) \simeq -5$, $\forall T \geq 150$, where in this case $\theta^* = \theta_2$ is the best expert for $T \geq 150$. The learner then does *better* than the best expert. If we use the AA instead, the learner does worse than θ_2 by $\simeq R^{\mathrm{S}}_{\ell_{\mathrm{Brier}}} = \log 2$. $\square$

In real data, the situation described above — where the best expert does not necessarily perform optimally during the game — is typical, especially when the number of rounds T is large. We have tested the aggregating algorithms on real data as studied by Vovk [21]. We compared the performance of the AA with the AGAA, and found that the AGAA outperforms the AA, and in fact achieved a negative regret on two data sets. Details of the experiments are in Appendix J.

As pointed out earlier, there are situations where $\Delta R_{\theta^*}(T) \geq 0$ even for the choice of $(\boldsymbol{v}^t)$ in Example 1, and this could potentially lead to a large positive regret for the AGAA. There is an easy way to remove this risk at a small price; the outputs of the AGAA and the AA can themselves be considered as expert predictions. These predictions can in turn be passed to a new instance of the AA to yield a *meta prediction*. The resulting worst case regret is guaranteed not to exceed that of the original AA instance by more than $\eta^{-1}\log 2$ for an η-mixable loss. We test this idea in Appendix J.

5 Discussion and Future Work

In this work, we derived a characterization of Φ-mixability, which enables a better understanding of when a constant regret is achievable in the game of prediction with expert advice. Then, borrowing techniques from mirror descent, we proposed a new "adaptive" version of the generalized aggregating algorithm. We derived a regret bound for a specific instantiation of this algorithm and discussed certain situations where the algorithm is expected to perform well. We empirically demonstrated the performance of this algorithm on football game predictions (see Appendix J).

Vovk [20, §5] essentially showed that given an η-mixable loss there is no algorithm that can achieve a lower regret bound than $\eta^{-1}\log k$ on all sequences of outcomes. There is no contradiction in trying to design algorithms which perform well in expectation (maybe better than the AA) on "typical" data while keeping the worst case regret close to $\eta^{-1}\log k$. This was the motivation behind the AGAA. In future work, we will explore other choices for the correction vector $\boldsymbol{v}^t$ with the goal of lowering the (expected) bound in (12). In the present work, we did not study the possibility of varying the learning rate η. One might obtain better regret bounds using an adaptive learning rate as is the case with the mirror descent algorithm. Our Corollary 17 is useful in that it gives an upper bound on the maximal learning rate under the Φ-mixability constraint. Finally, although our Theorem 18 states that worst-case regret of the GAA is minimized when using the Shannon entropy, it would be interesting to study the dynamics of the AGAA with other entropies.

Table 1: A short list of the main symbols used in the paper

Symbol	Description
ℓ	A loss function defined on a set $\mathcal{A}$ and taking values in $[0, +\infty]^n$ (see Sec. 2)
$\mathscr{S}_\ell$	The finite part of the superprediction set of a loss ℓ (see Sec. 2)
$\underline{\ell}$	The support loss of a loss ℓ (see Def. 3)
$\underline{L}_\ell$	The Bayes risk corresponding to a loss ℓ (see Definition 2)
$\tilde{\underline{L}}_\ell$	The composition of the Bayes risk with an affine function; $\tilde{\underline{L}}_\ell := \underline{L}_\ell \circ \Pi_n$ (see (1))
S	The Shannon Entropy (see (6))
η_ℓ	The mixability constant of ℓ (see Def. 6) ; essentially the largest η s.t. ℓ is η-mixable.
$\underline{\eta}_\ell$	Essentially the largest η such that $\eta\underline{L}_\ell - \underline{L}_{\log}$ is convex (see (5) and [19])
η^{Φ}_ℓ	The generalized mixability constant (see Def. 9); the largest η s.t. ℓ is (η, Φ)-mixable.
$\mathfrak{S}_\ell$	A substitution function of a loss ℓ (see Sec. 3.1)
R^{Φ}_ℓ	The regret achieved by the GAA using entropy Φ (see (9) and Algorithm 2)

Acknowledgments

This work was supported by the Australian Research Council and DATA61.

References

[1] Ravi P Agarwal, Maria Meehan, and Donal O'Regan. *Fixed point theory and applications*, volume 141. Cambridge university press, 2001.

[2] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[3] Hardy, G.H. and Körner, T.W. *A Course of Pure Mathematics Centenary Edition*. Cambridge Mathematical Library, 2008.

[4] Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, second edition, 2011.

[5] Jonathan M. Borwein, Jon D. Vanderwerff, et al. *Convex functions: constructions, characterizations and counterexamples*, volume 109. Cambridge University Press Cambridge, 2010.

[6] Alexey Chernov, Yuri Kalnishkan, Fedor Zhdanov, and Vladimir Vovk. Supermartingales in prediction with expert advice. *Theoretical Computer Science*, 411(29-30):2647–2669, 2010.

[7] A. Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

[8] J-B. Hiriart-Urruty and C. Lemaréchal. Fundamentals of convex analysis., 2001.

[9] Yuri Kalnishkan, Volodya Vovk, and Michael V. Vyugin. Loss functions, complexities, and the legendre transformation. *Theoretical Computer Science*, 313(2):195–207, 2004.

[10] Parameswaran Kamalaruban, Robert Williamson, and Xinhua Zhang. Exp-concavity of proper composite losses. In *Conference on Learning Theory*, pages 1035–1065, 2015.

[11] Nelson Merentes and Kazimierz Nikodem. Remarks on strongly convex functions. *Aequationes mathematicae*, 80(1):193–199, 2010.

[12] Francesco Orabona, Koby Crammer, and Nicolò Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

[13] Mark D. Reid, Rafael M. Frongillo, Robert C. Williamson, and Nishant Mehta. Generalized mixability via entropic duality. In *Conference on Learning Theory*, pages 1501–1522, 2015.

[14] Joel W. Robbin and Dietmar A. Salamon. Introduction to differential geometry, eth, lecture notes. *ETH, Lecture Notes, preliminary version, January*, 2011.

[15] R. Tyrrell Rockafellar. Convex analysis. Princeton University Press, Princeton, NJ, 1997.

[16] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[17] Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, pages 1593–1601, 2014.

[18] John A. Thorpe. *Elementary Topics in Differential Geometry*. Springer Science & Business Media, 1994.

[19] Tim van Erven, Mark D. Reid, and Robert C. Williamson. Mixability is Bayes risk curvature relative to log loss. *Journal of Machine Learning Research*, 13:1639–1663, 2012.

[20] Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

[21] Vladimir Vovk and Fedor Zhdanov. Prediction with expert advice for the brier game. *Journal of Machine Learning Research*, 10(Nov):2445–2471, 2009.

[22] Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

[23] Robert C. Williamson. The geometry of losses. In *Conference on Learning Theory*, pages 1078–1108, 2014.

[24] Robert C. Williamson, Elodie Vernet, and Mark D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17:223:1–223:52, 2016.

Contents

A Notation and Preliminaries

For $n \in \mathbb{N}$, we define $\tilde{n} = n - 1$. We denote $[n] \coloneqq \{1, \dots, n\}$ the set of integers between 1 and n. Let $\langle \cdot, \cdot \rangle$ denote the standard inner product in $\mathbb{R}^n$ and $\|\cdot\|$ the corresponding norm. Let I_n and $\mathbf{1}_n$ denote the $n \times n$ identity matrix and the vector of all ones in $\mathbb{R}^n$. Let $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ denote the *standard basis* for $\mathbb{R}^n$. For a set $\mathfrak{l} \subsetneq \mathbb{N}$ and $\boldsymbol{r}_1, \dots, \boldsymbol{r}_n \in \mathbb{R}^k$, we denote $[\boldsymbol{r}_i]_{i \in \mathfrak{l}} \coloneqq [\boldsymbol{r}_{i_1}, \dots, \boldsymbol{r}_{i_k}] \in \mathbb{R}^{n \times k}$, where $\mathfrak{l} = \{i_1, \dots, i_k\}$ and $i_1 < \dots < i_k$. We denote its transpose by $[\boldsymbol{r}_i]_{i \in \mathfrak{l}}^\mathsf{T} \in \mathbb{R}^{k \times n}$. For two vectors $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$, we write $\boldsymbol{p} \leq \boldsymbol{q}$ [resp. $\boldsymbol{p} < \boldsymbol{q}$], if $\forall i \in [n], p_i \leq q_i$ [resp. $p_i < q_i$]. We also denote $\boldsymbol{p} \odot \boldsymbol{q} = [p_i q_i]_{1 \leq i \leq n}^\mathsf{T} \in \mathbb{R}^n$ the *Hadamard product* of $\boldsymbol{p}$ and $\boldsymbol{q}$. If $(\boldsymbol{c}_k)$ is a sequence of vectors in $\mathcal{C} \subseteq \mathbb{R}^n$, we simply write $(\boldsymbol{c}_k) \subset \mathcal{C}$. For a sequence $(\boldsymbol{v}_m) \subset \mathbb{R}^n$, we write $\boldsymbol{v}_m \overset{m \to \infty}{\to} \boldsymbol{v}$ or $\lim_{m \to \infty} \boldsymbol{v}_m = \boldsymbol{v}$, if $\forall i \in [n], \lim_{m \to \infty} [\boldsymbol{v}_m]_i = v_i$. For a square matrix $A \in \mathbb{R}^{n \times n}$, $\lambda_{\min}(A)$ [resp. $\lambda_{\max}(A)$] denotes its minimum [resp. maximum] eigenvalue. For $k \geq 1$, $\boldsymbol{u} \in [0, +\infty[^k$ and $\boldsymbol{w} \in \mathbb{R}^k$, we define $\log \boldsymbol{u} \coloneqq [\log u_i]_{1 \leq i \leq k}^\mathsf{T} \in \mathbb{R}^k$ and $\exp \boldsymbol{w} \coloneqq [\exp w_i]_{1 \leq i \leq k}^\mathsf{T} \in \mathbb{R}^k$.

Let $\Delta_n \coloneqq \{\boldsymbol{p} \in [0,1]^n : \langle \boldsymbol{p}, \mathbf{1}_n \rangle = 1\}$ be the *probability simplex* in $\mathbb{R}^n$. We also define $\tilde{\Delta}_n \coloneqq \{\tilde{\boldsymbol{p}} \in [0, +\infty[^{\tilde{n}} \colon \langle \tilde{\boldsymbol{p}}, \mathbf{1}_{\tilde{n}} \rangle \leq 1\}$. We will use the notations $\Delta_n^k \coloneqq (\Delta_n)^k$ and $\tilde{\Delta}_n^k \coloneqq (\tilde{\Delta}_n)^k$. For $\mathfrak{l} \subseteq [n]$, the set $\Delta_\mathfrak{l} = \{\boldsymbol{q} \in \Delta_n : q_i = 0, \forall i \in [n] \setminus \mathfrak{l}\}$ is a $|\mathfrak{l}|$-*face* of Δ_n. We denote $\Pi_\mathfrak{l}^n : \mathbb{R}^n \to \mathbb{R}^{|\mathfrak{l}|}$ the linear projection operator satisfying $\Pi_\mathfrak{l}^n \boldsymbol{u} = [u_i]_{i \in \mathfrak{l}}^\mathsf{T}$. If there is no ambiguity from the context, we may simply write $\Pi_\mathfrak{l}$ instead of $\Pi_\mathfrak{l}^n$. It is easy to verify that $\Pi_\mathfrak{l} \Pi_\mathfrak{l}^\mathsf{T} = I_{|\mathfrak{l}|}$ and that $\boldsymbol{q} \mapsto \Pi_\mathfrak{l} \boldsymbol{q}$ is a bijection from $\Delta_\mathfrak{l} \subseteq \Delta_n$ to $\Delta_{|\mathfrak{l}|}$. In the special case where $\mathfrak{l} = [\tilde{n}]$, we write $\Pi_n \coloneqq \Pi_{[\tilde{n}]}^n$ and we define the affine operator $\Pi_n : \mathbb{R}^{\tilde{n}} \to \mathbb{R}^n$ by $\Pi_n(\boldsymbol{u}) \coloneqq [u_1, \dots, u_{\tilde{n}}, 1 - \langle \boldsymbol{u}, \mathbf{1}_{\tilde{n}} \rangle]^\mathsf{T} = J_n \boldsymbol{u} + \boldsymbol{e}_n$, where $J_n \coloneqq \begin{bmatrix} I_{\tilde{n}} \\ -\mathbf{1}_{\tilde{n}}^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{n \times \tilde{n}}$.

For $\boldsymbol{u} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, we denote $\mathcal{H}_{\boldsymbol{u},c} \coloneqq \{\boldsymbol{y} \in \mathbb{R}^n : \langle \boldsymbol{y}, \boldsymbol{u} \rangle \leq c\}$ and $\mathcal{B}(\boldsymbol{u}, c) \coloneqq \{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{u} - \boldsymbol{v}\| \leq c\}$. $\mathcal{H}_{\boldsymbol{u},c}$ is a closed half space and $\mathcal{B}(\boldsymbol{u}, c)$ is the c-*ball* in $\mathbb{R}^n$ centered at $\boldsymbol{u}$. Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a non-empty set. We denote $\operatorname{int} \mathcal{C}$, $\operatorname{ri} \mathcal{C}$, $\operatorname{bd} \mathcal{C}$, and $\operatorname{rbd} \mathcal{C}$ the *interior*, *relative interior*, *boundary*, and *relative boundary* of a set $\mathcal{C} \in \mathbb{R}^n$, respectively [8]. We denote the *indicator function* of $\mathcal{C}$ by $\iota_\mathcal{C}$, where for $\boldsymbol{u} \in \mathcal{C}$, $\iota_\mathcal{C}(\boldsymbol{u}) = 0$, otherwise $\iota_\mathcal{C}(\boldsymbol{u}) = +\infty$. The *support function* of $\mathcal{C}$ is defined by

$$\sigma_\mathcal{C}(\boldsymbol{u}) \coloneqq \sup_{\boldsymbol{s} \in \mathcal{C}} \langle \boldsymbol{u}, \boldsymbol{s} \rangle \,, \ \boldsymbol{u} \in \mathbb{R}^n.$$

Let $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. We denote $\operatorname{dom} f \coloneqq \{\boldsymbol{u} \in \mathbb{R}^n : f(\boldsymbol{u}) < +\infty\}$ the *effective domain* of f. The function f is *proper* if $\operatorname{dom} f \neq \varnothing$. The function f is *convex* if $\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^n$ and $\lambda \in]0,1[$, $f(\lambda \boldsymbol{u} + (1-\lambda)\boldsymbol{v}) \leq \lambda f(\boldsymbol{u}) + (1-\lambda) f(\boldsymbol{v})$. When the latter inequality is strict for all $\boldsymbol{u} \neq \boldsymbol{v}$, f is *strictly convex*. When f is convex, it is *closed* if it is *lower semi-continuous*; that is, for all $\boldsymbol{u} \in \mathbb{R}^n$, $\liminf_{\boldsymbol{v} \to \boldsymbol{u}} f(\boldsymbol{v}) \geq f(\boldsymbol{u})$. The function f is said to be 1-*homogeneous* if $\forall (\boldsymbol{u}, \alpha) \in \mathbb{R}^n \times]0, +\infty[, f(\alpha \boldsymbol{u}) = \alpha f(\boldsymbol{u})$, and it is said to be 1-*coercive* if $\frac{f(\boldsymbol{u})}{\|\boldsymbol{u}\|} \to +\infty$ as $\|\boldsymbol{u}\| \to \infty$. Let f be proper. The *sub-differential* of f is defined by

$$\partial f(\boldsymbol{u}) \coloneqq \{\boldsymbol{s}^* \in \mathbb{R}^n : f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \langle \boldsymbol{s}^*, \boldsymbol{v} - \boldsymbol{u} \rangle \,, \forall \boldsymbol{v} \in \mathbb{R}^n\}.$$

Any element $\boldsymbol{s} \in \partial f(\boldsymbol{u})$ is called a *sub-gradient* of f at $\boldsymbol{u}$. We say that f is *directionally differentiable* if for all $(\boldsymbol{u}, \boldsymbol{v}) \in \operatorname{dom} f \times \mathbb{R}^n$ the limit $\lim_{t \downarrow 0} \frac{f(\boldsymbol{u} + t\boldsymbol{v}) - f(\boldsymbol{u})}{t}$ exists in $[-\infty, +\infty]$. In this case, we denote the limit by $f'(\boldsymbol{u}; \boldsymbol{v})$. When f is convex, it is directionally differentiable [15]. Let f be proper and directionally differentiable. The *divergence* generated by f is the map $D_f \colon \mathbb{R}^n \times \operatorname{dom} f \to [0, +\infty]$ defined by

$$D_f(\boldsymbol{v}, \boldsymbol{u}) \coloneqq \begin{cases} f(\boldsymbol{v}) - f(\boldsymbol{u}) - f'(\boldsymbol{u}; \boldsymbol{v} - \boldsymbol{u}), & \text{if } \boldsymbol{v} \in \operatorname{dom} f; \\ +\infty, & \text{otherwise.} \end{cases}$$

For $\mathfrak{l} \subset [n]$ and $f_\mathfrak{l} \coloneqq f \circ \Pi_\mathfrak{l}^\mathsf{T}$, it is easy to verify that $f_\mathfrak{l}'(\Pi_\mathfrak{l} \boldsymbol{p}; \Pi_\mathfrak{l} \boldsymbol{q} - \Pi_\mathfrak{l} \boldsymbol{p}) = f'(\boldsymbol{p}; \boldsymbol{q} - \boldsymbol{p}), \forall (\boldsymbol{p}, \boldsymbol{q}) \in \Delta_\mathfrak{l}$. In this case, it holds that $D_f(\boldsymbol{q}, \boldsymbol{p}) = D_{f_\mathfrak{l}}(\Pi_\mathfrak{l} \boldsymbol{q}, \Pi_\mathfrak{l} \boldsymbol{p})$. If f is differentiable [resp. twice differentiable] at $\boldsymbol{u} \in \operatorname{int} \operatorname{dom} f$, we denote $\nabla f(\boldsymbol{u}) \in \mathbb{R}^n$ [resp. $\mathsf{H} f(\boldsymbol{u}) \in \mathbb{R}^{n \times n}$] its *gradient* vector [resp. *Hessian* matrix] at $\boldsymbol{u}$. A vector-valued function $g \colon \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\boldsymbol{u}$ if for all $i \in [m]$, g_i is differentiable at $\boldsymbol{u}$. In this case, the *differential* of g at $\boldsymbol{u}$ is the linear operator $\mathsf{D} g(\boldsymbol{u}) : \mathbb{R}^n \to \mathbb{R}^m$ defined by $\mathsf{D} g(\boldsymbol{u}) \coloneqq [\nabla g_i(\boldsymbol{u})]_{1 \leq i \leq m}^\mathsf{T}$. If f has k continuous derivatives on a set $\Omega \subset \mathbb{R}^k$, we write $f \in C^k(\Omega)$.

We define $\tilde{f} : \mathbb{R}^{\tilde{n}} \to \mathbb{R} \cup \{+\infty\}$ by $\tilde{f} := f \circ \Pi_n + \iota_{\tilde{\Delta}_n}$. That is,

$$\tilde{f}(\tilde{\boldsymbol{u}}) := \begin{cases} f(J_n \tilde{\boldsymbol{u}} + \boldsymbol{e}_n), & \text{for } \tilde{\boldsymbol{u}} \in \tilde{\Delta}_n; \\ +\infty, & \text{for } \tilde{\boldsymbol{u}} \in \mathbb{R}^{n-1} \setminus \tilde{\Delta}_n. \end{cases} \tag{13}$$

If $\tilde{f}$ is directionally differentiable, then $f'(\boldsymbol{p}, \boldsymbol{q} - \boldsymbol{p}) = \tilde{f}'(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{p}})$, for $\boldsymbol{p}, \boldsymbol{q} \in \Delta_n$. If $\tilde{f}$ is differentiable at $\tilde{\boldsymbol{p}} = \Pi_n(\boldsymbol{p})$, then $\tilde{f}'(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{p}}) = \langle \nabla \tilde{f}(\tilde{\boldsymbol{p}}), \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{p}} \rangle$. If, additionally, f is differentiable at $\boldsymbol{p} \in \operatorname{ri} \Delta_k$, the chain rule yields $\nabla \tilde{f}(\tilde{\boldsymbol{p}}) = J_n^{\mathsf{T}} \nabla f(\boldsymbol{p})$. Since $J_n(\tilde{\boldsymbol{p}} - \tilde{\boldsymbol{q}}) = \Pi_n(\tilde{\boldsymbol{p}} - \tilde{\boldsymbol{q}}) = \boldsymbol{p} - \boldsymbol{q}$, it also follows that $\langle \tilde{\boldsymbol{p}} - \tilde{\boldsymbol{q}}, \nabla \tilde{f}(\tilde{\boldsymbol{p}}) \rangle = \langle \boldsymbol{p} - \boldsymbol{q}, \nabla f(\boldsymbol{p}) \rangle$.

The *Fenchel dual* of a (proper) function f is defined by $f^*(\boldsymbol{v}) := \sup_{\boldsymbol{u} \in \operatorname{dom} f} \langle \boldsymbol{u}, \boldsymbol{v} \rangle - f(\boldsymbol{u})$, and it is a closed, convex function on $\mathbb{R}^n$ [8]. The following proposition gives some useful properties of the Fenchel dual which will be used in several proofs.

Proposition 20 ([8])**.** *Let $f, h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. If f and h are proper and there are affine functions minorizing them on $\mathbb{R}^n$, then for all $\boldsymbol{v}_0 \in \mathbb{R}^n$*

$$\begin{array}{lll} (i) & g(\boldsymbol{u}) = f(\boldsymbol{u}) + r, \ \forall \boldsymbol{u} & \implies g^*(\boldsymbol{v}) = f^*(\boldsymbol{v}) - r, \ \forall \boldsymbol{v} \\ (ii) & g(\boldsymbol{u}) = f(\boldsymbol{u}) + \langle \boldsymbol{v}_0, \boldsymbol{u} \rangle, \ \forall \boldsymbol{u} & \implies g^*(\boldsymbol{v}) = f^*(\boldsymbol{v} - \boldsymbol{v}_0), \ \forall \boldsymbol{v} \\ (iii) & f \leq h & \implies f^* \geq h^*, \\ (iv) & \boldsymbol{s} \in \partial f^*(\boldsymbol{v}), \boldsymbol{v} \in \mathbb{R}^n & \implies f^*(\boldsymbol{v}) = \langle \boldsymbol{v}, \boldsymbol{s} \rangle - f(\boldsymbol{s}), \\ (v) & g(\boldsymbol{u}) = f(t\boldsymbol{u}), \ t > 0, \forall \boldsymbol{u} & \implies g^*(\boldsymbol{v}) = f^*(\boldsymbol{v}/t), \end{array}$$

A function $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ is an *entropy* if it is closed, convex, and $\Delta_k \subseteq \operatorname{dom} \Phi$. Its *entropic dual* $\Phi^\star : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ is defined by $\Phi^\star(\boldsymbol{z}) := \sup_{\boldsymbol{q} \in \Delta_k} \langle \boldsymbol{q}, \boldsymbol{z} \rangle - \Phi(\boldsymbol{q}), \boldsymbol{z} \in \mathbb{R}^k$. For the remainder of this paper, we consider entropies defined on $\mathbb{R}^k$, where $k \geq 2$.

Let $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy and $\Phi_\Delta := \Phi + \iota_{\Delta_k}$. In this case, $\Phi^\star = \Phi_\Delta^*$. It is clear that Φ_Δ is 1-coercive, and therefore, $\operatorname{dom} \Phi^\star = \operatorname{dom} \Phi_\Delta^* = \mathbb{R}^k$ [8, Prop. E.1.3.8]. The entropic dual of Φ can also be expressed using the Fenchel dual of $\tilde{\Phi} : \mathbb{R}^{k-1} \to \mathbb{R} \cup \{+\infty\}$ defined by (13) after substituting f by Φ and n by k. In fact,

$$\begin{aligned} \Phi^\star(\boldsymbol{z}) &= \sup_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \langle J_k \tilde{\boldsymbol{q}} + \boldsymbol{e}_k, \boldsymbol{z} \rangle - \Phi(J_k \tilde{\boldsymbol{q}} + \boldsymbol{e}_k), \\ &= \langle \boldsymbol{e}_k, \boldsymbol{z} \rangle + \sup_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \left\langle \tilde{\boldsymbol{q}}, J_k^{\mathsf{T}} \boldsymbol{z} \right\rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}), \\ &= \langle \boldsymbol{e}_k, \boldsymbol{z} \rangle + \tilde{\Phi}^*(J_k^{\mathsf{T}} \boldsymbol{z}), \end{aligned} \tag{14}$$

where (14) follows from the fact that $\operatorname{dom} \tilde{\Phi} = \tilde{\Delta}_k$. Note that when Φ is an entropy, $\tilde{\Phi}$ is a closed convex function on $\mathbb{R}^{k-1}$. Hence, it holds that $\tilde{\Phi}^{**} = \tilde{\Phi}$ [15].

The *Shannon entropy* by $\mathrm{S}(\boldsymbol{q}) := \sum_{i \in [k] : q_i \neq 0} q_i \log q_i$,[1] if $\boldsymbol{q} \in [0, +\infty[^k$; and $+\infty$ otherwise.

We will also make use of the following lemma.

Lemma 21 ([4])**.** *$\forall m \geq 1, \forall A, B \in \mathbb{R}^{m \times m}$, $\lambda_{\max}(AB) = \lambda_{\max}(BA)$ and $\lambda_{\min}(AB) = \lambda_{\min}(BA)$.*

B Technical Lemmas

This appendix presents technical lemmas which will be needed in various proofs of results from the main body of the paper.

For an open convex set Ω in $\mathbb{R}^n$ and $\alpha > 0$, a function $\phi \colon \Omega \to \mathbb{R}$ is said to be *α-strongly convex* if $\boldsymbol{u} \mapsto \phi(\boldsymbol{u}) - \alpha \|\boldsymbol{u}\|^2$ is convex on Ω [11]. The next lemma is a characterization of a generalization of α-strong convexity, where $\boldsymbol{u} \mapsto \|\boldsymbol{u}\|^2$ is replaced by any strictly convex function.

Lemma 22. *Let $\Omega \subseteq \mathbb{R}^n$ be an open convex set. Let $\phi, \psi \colon \Omega \to \mathbb{R}$ be twice differentiable.*

[1] The Shannon entropy is usually defined with a minus sign. However, it will be more convenient for us to work without it.

If ψ is strictly convex, then $\forall \boldsymbol{u} \in \Omega$, $\mathsf{H}\psi(\boldsymbol{u})$ is invertible, and for any $\alpha > 0$

$$\forall \boldsymbol{u} \in \Omega,\ \lambda_{\min}(\mathsf{H}\phi(\boldsymbol{u})(\mathsf{H}\psi(\boldsymbol{u}))^{-1}) \geq \alpha \iff \phi - \alpha\psi \text{ is convex}, \tag{15}$$

Furthermore, if $\alpha > 1$, then the left hand side of (15) *implies that $\phi - \psi$ is strictly convex.*

Proof. Suppose that $\inf_{\boldsymbol{u}\in\Omega} \lambda_{\min}(\mathsf{H}\phi(\boldsymbol{u})(\mathsf{H}\psi(\boldsymbol{u}))^{-1}) \geq \alpha$. Since g is strictly convex and twice differentiable on Ω, $\mathsf{H}\psi(\boldsymbol{u})$ is symmetric positive definite, and thus invertible. Therefore, there exists a symmetric positive definite matrix $G \in \mathbb{R}^{n\times n}$ such that $GG = \mathsf{H}\psi(\boldsymbol{u})$. Lemma 21 implies

$$\begin{array}{rcll}
 & \inf_{\boldsymbol{u}\in\Omega} \lambda_{\min}(\mathsf{H}\phi(\boldsymbol{u})(\mathsf{H}\psi(\boldsymbol{u}))^{-1}) & \geq \alpha, \\
\iff & \inf_{\boldsymbol{u}\in\Omega} \lambda_{\min}(G^{-1}\mathsf{H}\phi(\boldsymbol{u})G^{-1}) & \geq \alpha, \\
\iff & \forall \boldsymbol{u} \in \Omega, \forall \boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}, \frac{\boldsymbol{v}^\mathsf{T} G^{-1}(\mathsf{H}\phi(\boldsymbol{u}))G^{-1}\boldsymbol{v}}{\boldsymbol{v}^\mathsf{T}\boldsymbol{v}} & \geq \alpha, \\
\iff & \forall \boldsymbol{u} \in \Omega, \forall \boldsymbol{w} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}, \boldsymbol{w}^\mathsf{T}(\mathsf{H}\phi(\boldsymbol{u}))\boldsymbol{w} & \geq \alpha \boldsymbol{w}^\mathsf{T} GG\boldsymbol{w} = \boldsymbol{w}^\mathsf{T}(\alpha\mathsf{H}\psi(\boldsymbol{u}))\boldsymbol{w}, \\
\iff & \forall \boldsymbol{u} \in \Omega, \mathsf{H}\phi(\boldsymbol{u}) & \succeq \alpha\mathsf{H}\psi(\boldsymbol{u}), \\
\iff & \forall \boldsymbol{u} \in \Omega, \mathsf{H}(\phi - \alpha\psi)(\boldsymbol{u}) & \succeq 0,
\end{array}$$

where in the third and fifth lines we used the definition of minimum eigenvalue and performed the change of variable $\boldsymbol{w} = G^{-1}\boldsymbol{v}$, respectively. To conclude the proof of (15), note that the positive semi-definiteness of $\mathsf{H}(\phi - \alpha\psi)$ is equivalent to the convexity of $\phi - \alpha\psi$ [8, Thm B.4.3.1].

Finally, note that the equivalences established above still hold if we replace α, "$\geq$", and "$\succeq$" by 1, "$>$", and "$\succ$", respectively. The strict convexity of $\phi - \psi$ then follows from the positive definiteness of $\mathsf{H}(\phi - \psi)$ (ibid.). □

The following result due to [6] will be crucial to prove the convexity of the superprediction set (Theorem 48).

Lemma 23 ([6])**.** *Let $\Delta(\Omega)$ be the set of distributions over some set $\Omega \subseteq \mathbb{R}$. Let a function $Q : \Delta(\Omega) \times \Omega \to \mathbb{R}$ be such that $Q(\cdot, \omega)$ is continuous for all $\omega \in \Omega$. If for all $\boldsymbol{\pi} \in \Delta(\Omega)$ it holds that $\mathbb{E}_{\omega\sim\boldsymbol{\pi}} Q(\boldsymbol{\pi}, \omega) \leq r$, where $r \in \mathbb{R}$ is some constant, then*

$$\exists \boldsymbol{\pi} \in \Delta(\Omega), \forall \omega \in \Omega,\ Q(\boldsymbol{\pi}, \omega) \leq r.$$

Note that when Ω in the lemma above is $[n]$, $\Delta([n]) \equiv \Delta_n$.

The next crucial lemma is a slight modification of a result due to [6].

Lemma 24. *Let $f\colon \operatorname{ri}\Delta_n \times [n] \to \mathbb{R}$ be a continuous function in the first argument and such that $\forall(\boldsymbol{q}, x) \in \operatorname{ri}\Delta_n \times [n], -\infty < f(\boldsymbol{q}, x)$. Suppose that $\forall \boldsymbol{p} \in \operatorname{ri}\Delta_n, \mathbb{E}_{x\sim\boldsymbol{p}}[f(\boldsymbol{p}, x)] \leq 0$, then*

$$\forall \epsilon > 0, \exists \boldsymbol{p}_\epsilon \in \operatorname{ri}\Delta_n, \forall x \in [n], f(\boldsymbol{p}_\epsilon, x) \leq \epsilon.$$

Proof. Pick any $\delta > 0$ such that $\delta(n-1) < 1$, and $c_0 < 0$ such that $\forall(\boldsymbol{q}, x) \in \operatorname{ri}\Delta_n \times [n], c_0 \leq f(\boldsymbol{q}, x)$. We define $\Delta_n^\delta \coloneqq \{\boldsymbol{p} \in \Delta_n : \forall x \in [n], p_x \geq \delta\}$ and $g(\boldsymbol{q}, \boldsymbol{p}) \coloneqq \mathbb{E}_{x\sim\boldsymbol{q}}[f(\boldsymbol{p}, x)]$. For a fixed $\boldsymbol{q}$, $\boldsymbol{p} \mapsto g(\boldsymbol{q}, \boldsymbol{p})$ is continuous, since f is continuous in the first argument. For a fixed $\boldsymbol{p}$, $\boldsymbol{q} \mapsto g(\boldsymbol{q}, \boldsymbol{p})$ is linear, and thus concave. Since Δ_n^δ is convex and compact, g satisfies Ky Fan's minimax Theorem [1, Thm. 11.4], and therefore, there exists $\boldsymbol{p}^\delta \in \Delta_n^\delta$ such that

$$\forall \boldsymbol{q} \in \Delta_n^\delta,\ \ \mathbb{E}_{x\sim\boldsymbol{q}}[f(\boldsymbol{p}^\delta, x)] = g(\boldsymbol{q}, \boldsymbol{p}^\delta) \leq \sup_{\boldsymbol{\mu}\in\Delta_n^\delta} g(\boldsymbol{\mu}, \boldsymbol{\mu}) = \sup_{\boldsymbol{\mu}\in\Delta_n^\delta} \mathbb{E}_{x\sim\boldsymbol{\mu}}[f(\boldsymbol{\mu}, x)] \leq 0. \tag{16}$$

For $x_0 \in [n]$, let $\hat{\boldsymbol{q}} \in \Delta_n^\delta$ be such that $\hat{q}_{x_0} = 1 - \delta(n-1)$ and $\hat{q}_x = \delta$ for $x \neq x_0$ (this is a legitimate distribution since $\delta(n-1) < 1$ by construction). Substituting $\hat{\boldsymbol{q}}$ for $\boldsymbol{q}$ in (16) gives

$$\begin{array}{rrl}
 & (1-\delta(n-1))f(\boldsymbol{p}^\delta, x_0) + \delta\sum_{x\neq x_0} f(\boldsymbol{p}^\delta, x) & \leq 0, \\
\implies & (1-\delta(n-1))f(\boldsymbol{p}^\delta, x_0) & \leq -c_0\delta(n-1), \\
\implies & f(\boldsymbol{p}^\delta, x_0) & \leq [-c_0\delta(n-1)]/[1-\delta(n-1)].
\end{array}$$

Choosing $\delta^* \coloneqq \epsilon/[(-c_0+\epsilon)(n-1)]$, and $\boldsymbol{p}_\epsilon \coloneqq \boldsymbol{p}^{\delta^*}$ gives the desired result.

□

Lemma 25. *Let $f, g\colon I \to \mathbb{R}^n$, where $I \subseteq \mathbb{R}$ is an open interval containing 0. Suppose g [resp. f] is continuous [resp. differentiable] at 0. Then $t \mapsto \langle f(t), g(t)\rangle$ is differentiable at 0 if and only if $t \mapsto \langle f(0), g(t)\rangle$ is differentiable at 0, and we have*

$$\frac{d}{dt}\langle f(t), g(t)\rangle\Big|_{t=0} = \left\langle \frac{d}{dt}f(t)\Big|_{t=0}, g(0)\right\rangle + \frac{d}{dt}\langle f(0), g(t)\rangle\Big|_{t=0}.$$

Proof. We have

$$\begin{aligned}\frac{\langle f(t), g(t)\rangle - \langle f(0), g(0)\rangle}{t} &= \frac{\langle f(t), g(t)\rangle - \langle f(0), g(t)\rangle}{t} + \frac{\langle f(0), g(t)\rangle - \langle f(0), g(0)\rangle}{t},\\ &= \left\langle \frac{f(t)-f(0)}{t}, g(t)\right\rangle + \frac{\langle f(0), g(t)\rangle - \langle f(0), g(0)\rangle}{t}.\end{aligned}$$

But since g [resp. f] is continuous [resp. differentiable] at 0, the first term on the right hand side of the above equation converges to $\langle \frac{d}{dt}f(t)\big|_{t=0}, g(0)\rangle$ as $t \to 0$. Therefore, $\frac{1}{t}(\langle f(0), g(t)\rangle - \langle f(0), g(0)\rangle)$ admits a limit when $t \to 0$ if and only if $\frac{1}{t}(\langle f(t), g(t)\rangle - \langle f(0), g(0)\rangle)$ admits a limit when $t \to 0$. This shows that $t \mapsto \langle f(0), g(t)\rangle$ is differentiable at 0 if an only if $t \mapsto \langle f(t), g(t)\rangle$ is differentiable at 0, and in this case the above equation yields

$$\begin{aligned}\frac{d}{dt}\langle f(t), g(t)\rangle\Big|_{t=0} &= \lim_{t\to 0}\frac{\langle f(t), g(t)\rangle - \langle f(0), g(0)\rangle}{t},\\ &= \lim_{t\to 0}\left(\left\langle \frac{f(t)-f(0)}{t}, g(t)\right\rangle + \frac{\langle f(0), g(t)\rangle - \langle f(0), g(0)\rangle}{t}\right),\\ &= \left\langle \frac{d}{dt}f(t)\Big|_{t=0}, g(0)\right\rangle + \frac{d}{dt}\langle f(0), g(t)\rangle\Big|_{t=0}.\end{aligned}$$

□

Note that the differentiability of $t \mapsto \langle f(0), g(t)\rangle$ at 0 does not necessarily imply the differentiability of g at 0. Take for example $n = 3$, $f(t) = \mathbf{1}/3$ for $t \in]-1, 1[$, and

$$g(t) = \begin{cases} -t\boldsymbol{e}_1 + t\frac{\mathbf{1}}{3}, & \text{if } t \in]-1, 0[;\\ -t\frac{\mathbf{1}}{3} + t\boldsymbol{e}_2, & \text{if } t \in [0, 1[.\end{cases}$$

Thus, the function $t \mapsto \langle f(0), g(t)\rangle = 0$ is differentiable at 0 but g is not. The preceding Lemma will be particularly useful in settings where it is desired to compute the derivative $\frac{d}{dt}\langle f(0), g(t)\rangle|_{t=0}$ without any explicit assumptions on the differentiability of $g(t)$ at 0. For example, this will come up when computing $\frac{d}{dt}\langle \boldsymbol{p}, D\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\rangle|_{t=0}$, where $\boldsymbol{v} \in \mathbb{R}^{n-1}$ and $t \mapsto \tilde{\boldsymbol{\alpha}}^t$ is smooth curve on $\operatorname{int}\tilde{\Delta}_n$, with the only assumption that $\tilde{L}_\ell$ is twice differentiable at $\tilde{\boldsymbol{\alpha}}^0 \in \operatorname{int}\tilde{\Delta}_n$.

Lemma 26. *Let $\ell\colon \Delta_n \to [0, +\infty]^n$ be a proper loss. For any $\boldsymbol{p} \in \operatorname{ri}\Delta_n$, it holds that*

$$\ell \text{ is continuous at } \boldsymbol{p} \overset{(i)}{\iff} \underline{L}_\ell \text{ is differentiable at } \boldsymbol{p} \overset{(ii)}{\iff} \partial[-\underline{L}_\ell](\boldsymbol{p}) = \{\nabla\underline{L}_\ell(\boldsymbol{p})\} = \{\ell(\boldsymbol{p})\}.$$

$\overset{(i)}{\iff}$. This equivalence has been shown before by [24].

[$\overset{(ii)}{\iff}$] Since $\underline{L}_\ell(\boldsymbol{p}) = -\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})$, for all $\boldsymbol{p} \in \operatorname{ri}\Delta_n$, it follows that $\underline{L}_\ell$ is differentiable at $\boldsymbol{p}$ if and only if $\partial[-\underline{L}_\ell](\boldsymbol{p}) = \partial\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p}) = \{-\nabla\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})\} = \{\nabla\underline{L}_\ell(\boldsymbol{p})\}$ [8, Cor. D.2.1.4]. It remains to show that $\nabla\underline{L}_\ell(\boldsymbol{r}) = \ell(\boldsymbol{r})$ when $\underline{L}_\ell$ is differentiable at $\boldsymbol{r} \in \operatorname{ri}\Delta_n$. Let $\boldsymbol{\alpha}_x^t = \boldsymbol{r} + t\boldsymbol{e}_x$ and $\tilde{\boldsymbol{\alpha}}_x^t = \Pi_n(\boldsymbol{\alpha}_x^t)$, where $(\boldsymbol{e}_x)_{x\in[n]}$ is the standard basis of $\mathbb{R}^n$. For $x \in [n]$, the functions $f_x(t) \coloneqq \boldsymbol{\alpha}_x^t$ and $g_x(t) \coloneqq \tilde{\ell}(\tilde{\boldsymbol{\alpha}}_x^t)$ satisfy the conditions of Lemma 25. Therefore, $h_x(t) \coloneqq \langle f_x(0), g_x(t)\rangle = \langle \boldsymbol{r}, \tilde{\ell}(\tilde{\boldsymbol{\alpha}}_x^t)\rangle$ is differentiable at 0 and

$$\begin{aligned}\nabla\tilde{\underline{L}}(\boldsymbol{r})\boldsymbol{e}_x &= \frac{d}{dt}\tilde{\underline{L}}(\boldsymbol{\alpha}_x^t)\Big|_{t=0} = \frac{d}{dt}\langle f_x(t), g_x(t)\rangle\Big|_{t=0},\\ &= \left\langle \boldsymbol{e}_x, \tilde{\ell}(\tilde{\boldsymbol{r}})\right\rangle + \frac{d}{dt}h_x(t)\Big|_{t=0},\\ &= \tilde{\ell}_x(\tilde{\boldsymbol{r}}),\end{aligned}$$

where the last equality holds because h_x attains a minimum at 0 due to the properness of ℓ. The result being true for all $x \in [n]$ implies that $\nabla \underline{\tilde{L}}(\tilde{\boldsymbol{r}}) = \tilde{\ell}(\tilde{\boldsymbol{r}}) = \ell(\boldsymbol{r})$. □

The next Lemma is a restatement of earlier results due to [19]. Our proof is more concise due to our definition of the Bayes risk in terms of the support function of the superprediction set.

Lemma 27 ([19])**.** *Let $\ell\colon \Delta_n \to [0,+\infty]^n$ be a proper loss whose Bayes risk is twice differentiable on $]0,+\infty[^n$ and let $X_{\boldsymbol{p}} = I_{\tilde{n}} - \mathbf{1}_{\tilde{n}}\tilde{\boldsymbol{p}}^{\mathsf{T}}$. The following holds*

(i) $\forall \boldsymbol{p} \in \operatorname{ri}\Delta_n, \langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^{\mathsf{T}}$.

(ii) $\forall \tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$, $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) = \left[\begin{smallmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{smallmatrix}\right] \mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})$.

(iii) $\forall \tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$, $\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}) = -(X_{\boldsymbol{p}})^{-1}(\operatorname{diag}(\tilde{\boldsymbol{p}}))^{-1}$.

We show (i) and (ii). Let $\boldsymbol{p} \in \operatorname{ri}\Delta_n$ and $f(\tilde{\boldsymbol{q}}) := \langle \boldsymbol{p}, \tilde{\ell}(\tilde{\boldsymbol{q}})\rangle = \langle \boldsymbol{p}, \nabla \underline{L}_{\ell}(\boldsymbol{q})\rangle$, where the equality is due to Lemma 26. Since $\underline{L}_{\ell}$ is twice differentiable $]0,+\infty[^n$, f is differentiable on $\operatorname{int}\tilde{\Delta}_n$ and we have $\mathsf{D}f(\tilde{\boldsymbol{q}}) = \langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle$. Since ℓ is proper, f reaches a minimum at $\tilde{\boldsymbol{p}} \in \operatorname{int}\Delta_n$, and thus $\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^{\mathsf{T}}$ (this shows (i)). On the other hand, we have $\nabla \underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}}) = J_n^{\mathsf{T}} \nabla \underline{L}_{\ell}(\boldsymbol{p}) = J_n^{\mathsf{T}} \tilde{\ell}(\tilde{\boldsymbol{p}})$. By differentiating and using the chain the rule, we get $\mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}}) = [\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})]^{\mathsf{T}} J_n$. This means that $\forall i \in [\tilde{n}]$, $[\mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})]_{\bullet,i} = \nabla\tilde{\ell}_i(\tilde{\boldsymbol{p}}) - \nabla\tilde{\ell}_n(\tilde{\boldsymbol{p}})$, and thus $\sum_{i=1}^{\tilde{n}} p_i [\mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})]_{\bullet,i} = \sum_{i=1}^{\tilde{n}} p_i \nabla\tilde{\ell}_i(\tilde{\boldsymbol{p}}) - (1-p_n)\nabla\tilde{\ell}_n(\tilde{\boldsymbol{p}})$. On the other hand, it follows from point (i) of the lemma that $\sum_{i=1}^{n} p_i \nabla\tilde{\ell}_i(\tilde{\boldsymbol{p}}) = \mathbf{0}_{\tilde{n}}$. Therefore, $[\mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})]\tilde{\boldsymbol{p}} = -\nabla\tilde{\ell}_n(\tilde{\boldsymbol{p}})$ and, as a result, $\forall i \in [\tilde{n}]$, $[\mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})]_{\bullet,i} - [\mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})]\tilde{\boldsymbol{p}} = \nabla\tilde{\ell}_i(\tilde{\boldsymbol{p}})$. The last two equations can be combined as $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) = \left[\begin{smallmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{smallmatrix}\right] \mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})$.

[We show (iii)] It follows from (ii), since $\forall i \in [\tilde{n}]$, $\nabla[\tilde{\ell}_{\log}]_i(\tilde{\boldsymbol{p}}) = \frac{1}{p_i}\boldsymbol{e}_i$, for $\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$.

□

In the next lemma we state a new result for proper losses which will be crucial to prove a necessary condition for Φ-mixability (Theorem 14) — one of the main results of the paper.

Lemma 28. *Let $\ell\colon \Delta_n \to [0,+\infty]^n$ be a proper loss whose Bayes risk is twice differentiable on $]0,+\infty[^n$. For $\boldsymbol{v} \in \mathbb{R}^{n-1}$ and $\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$,*

$$\left\langle \boldsymbol{p}, (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}) \odot (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}) \right\rangle = -\boldsymbol{v}^{\mathsf{T}} \mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}}) [\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1} \mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \tag{17}$$

where $\boldsymbol{p} = \Pi_n(\tilde{\boldsymbol{p}})$ and $\underline{L}_{\log}$ is the Bayes risk of the log *loss.*

Furthermore, if $t \mapsto \tilde{\boldsymbol{\alpha}}^t$ is a smooth curve in $\operatorname{int}\tilde{\Delta}_n$ and satisfies $\tilde{\boldsymbol{\alpha}}^0 = \tilde{\boldsymbol{p}}$ and $\frac{d}{dt}\tilde{\boldsymbol{\alpha}}^t\big|_{t=0} = \boldsymbol{v}$, then $t \mapsto \langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\rangle$ is differentiable at 0 and we have

$$\frac{d}{dt}\left\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v} \right\rangle\Big|_{t=0} = -\boldsymbol{v}^{\mathsf{T}} \mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}. \tag{18}$$

Proof. We know from Lemma 27 that for $\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$, we have $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) = \left[\begin{smallmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{smallmatrix}\right] \mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})$, where $X_{\boldsymbol{p}} = I_{n-1} - \mathbf{1}_{n-1}\tilde{\boldsymbol{p}}^{\mathsf{T}}$. Thus, we can write

$$\begin{aligned}\left\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v} \odot \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v} \right\rangle &= \boldsymbol{v}^{\mathsf{T}} (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}))^{\mathsf{T}} \operatorname{diag}(\boldsymbol{p}) \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \\ &= \boldsymbol{v}^{\mathsf{T}} (\mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}}))^{\mathsf{T}} [X_{\boldsymbol{p}}^{\mathsf{T}},\ -\tilde{\boldsymbol{p}}] \operatorname{diag}(\boldsymbol{p}) \left[\begin{smallmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{smallmatrix}\right] \mathsf{H}\underline{\tilde{L}}_{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}. \end{aligned} \tag{19}$$

Observe that $[X_{\boldsymbol{p}}^\mathsf{T}, -\tilde{\boldsymbol{p}}]\operatorname{diag}(\boldsymbol{p}) = [I_{n-1} - \tilde{\boldsymbol{p}}\mathbf{1}_{n-1}^\mathsf{T}, -\tilde{\boldsymbol{p}}]\operatorname{diag}(\boldsymbol{p}) = [\operatorname{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^\mathsf{T},\ -\tilde{\boldsymbol{p}}p_n]$. Thus,

$$\begin{aligned}
[X_{\boldsymbol{p}}^\mathsf{T}, -\tilde{\boldsymbol{p}}]\operatorname{diag}(\boldsymbol{p})\begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T}\end{bmatrix} &= [\operatorname{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^\mathsf{T}, -\tilde{\boldsymbol{p}}p_n]\begin{bmatrix} I_{n-1} - \mathbf{1}_{n-1}\tilde{\boldsymbol{p}}^\mathsf{T} \\ -\tilde{\boldsymbol{p}}^\mathsf{T}\end{bmatrix}, \\
&= \operatorname{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^\mathsf{T} - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^\mathsf{T} + \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^\mathsf{T}(1-p_n) + p_n\tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^\mathsf{T}, \\
&= \operatorname{diag}(\tilde{\boldsymbol{p}}) - \tilde{\boldsymbol{p}}\tilde{\boldsymbol{p}}^\mathsf{T}, \\
&= \operatorname{diag}(\tilde{\boldsymbol{p}})\, X_{\boldsymbol{p}}, \\
&= -(\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1},
\end{aligned} \tag{20}$$

where the last equality is due to Lemma 27. The desired result follows by combining (19) and (20).

[We show (18)**]** Let $\tilde{\boldsymbol{p}} \in \operatorname{int}\tilde{\Delta}_n$, we define $\tilde{\boldsymbol{\alpha}}^t := \tilde{\boldsymbol{p}} + t\boldsymbol{v}$, $\boldsymbol{\alpha}^t := \Pi_n(\tilde{\boldsymbol{\alpha}}^t) = \boldsymbol{p} + tJ_n\boldsymbol{v}$, and $r(t) := \boldsymbol{\alpha}^t/\|\boldsymbol{\alpha}^t\|$, where $t \in \{s : \tilde{\boldsymbol{p}} + s\boldsymbol{v} \in \operatorname{int}\tilde{\Delta}_n\}$. Since $t \mapsto r(t)$ is differentiable at 0 and $t \mapsto \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}$ is continuous at 0, it follows from Lemma 22 that

$$\begin{aligned}
\frac{d}{dt}\left\langle r(0), \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\Big|_{t=0} &= \frac{d}{dt}\left\langle r(t), \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\Big|_{t=0} - \left\langle \dot{r}(0), \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \\
&= -\left\langle \dot{r}(0), \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle,
\end{aligned}$$

where the second equality holds since, according to Lemma 27, we have $\langle \boldsymbol{\alpha}^t, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\rangle = 0$. Since $r(0) = \boldsymbol{p}/\|\boldsymbol{p}\|$, $\dot{r}(0) = \|\boldsymbol{p}\|^{-1}(I_n - r(0)[r(0)]^\mathsf{T})J_n\boldsymbol{v}$, and $J_n = \begin{bmatrix} I_{n-1} \\ -\mathbf{1}_{n-1}^\mathsf{T}\end{bmatrix}$, we get

$$\begin{aligned}
\|\tilde{\boldsymbol{p}}\|\,\frac{d}{dt}\left\langle r(0), \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v}\right\rangle\Big|_{t=0} &= -\left\langle \left(I_n - r(0)[r(0)]^\mathsf{T}\right)J_n\boldsymbol{v}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \\
&= -\left\langle J_n\boldsymbol{v}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \qquad (21) \\
&= -\left\langle J_n\boldsymbol{v}, \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T}\end{bmatrix}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}\right\rangle, \\
&= -\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v},
\end{aligned}$$

where the passage to (21) is due to $r(0) = \boldsymbol{p}/\|\boldsymbol{p}\| \perp \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})$. In the last equality we used the fact that $J_n^\mathsf{T}\begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T}\end{bmatrix} = [I_{n-1}, -\mathbf{1}_{n-1}]\begin{bmatrix} I_{n-1} - \mathbf{1}_{n-1}\tilde{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T}\end{bmatrix} = I_{n-1}$. □

Proposition 29. *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ be an entropy and $\ell\colon \mathcal{A} \to [0,+\infty]^n$ a closed admissible loss. If ℓ is Φ-mixable, then $\forall \mathfrak{l} \subseteq [k]$ with $|\mathfrak{l}| > 1$, ℓ is $\Phi_\mathfrak{l}$-mixable and*

$$\forall \boldsymbol{q} \in \operatorname{rbd}\Delta_\mathfrak{l}, \forall \hat{\boldsymbol{q}} \in \operatorname{ri}\Delta_\mathfrak{l},\ \Phi'(\boldsymbol{q}; \hat{\boldsymbol{q}} - \boldsymbol{q}) = -\infty. \tag{22}$$

Given an entropy $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ and a loss $\ell\colon \mathcal{A} \to [0,+\infty]$, we define

$$\mathsf{m}_\Phi(x, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \boldsymbol{\mu}) := \langle \boldsymbol{\mu}, \ell_x(A)\rangle + D_\Phi(\boldsymbol{\mu}, \hat{\boldsymbol{q}}) - \ell_x(\boldsymbol{a}),$$

where $x \in [n]$, $A \in \mathcal{A}^k$, $\boldsymbol{a} \in \mathcal{A}$, and $\boldsymbol{q}, \hat{\boldsymbol{q}} \in \Delta_k$. Reid et al. [13] showed that ℓ is Φ mixable if and only if

$$\widehat{\mathsf{m}_\Phi} := \inf_{A\in\mathcal{A}^k, \hat{\boldsymbol{q}}\in\Delta_k}\ \sup_{\boldsymbol{a}_*\in\mathcal{A}}\ \inf_{\boldsymbol{\mu}\in\Delta_k, x\in[n]} \mathsf{m}_\Phi(x, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \boldsymbol{\mu}) \geq 0.$$

Proof of Proposition 29. **[We show that ℓ is $\Phi_\mathfrak{l}$-mixable]** Let $\mathfrak{l} \subseteq [k]$, with $|\mathfrak{l}| > 1$, $A \in \mathcal{A}^k$, and $\boldsymbol{q} \in \Delta_\mathfrak{l}$. Since ℓ is Φ-mixable, the following holds

$$\begin{aligned}
\exists \boldsymbol{a}_* \in \Delta_n, \forall x \in [n],\ \ell_x(\boldsymbol{a}_*) &\leq \inf_{\hat{\boldsymbol{q}}\in\Delta_k} \langle \hat{\boldsymbol{q}}, \ell_x(A)\rangle + D_\Phi(\hat{\boldsymbol{q}}, \boldsymbol{q}), \qquad (23) \\
&\leq \inf_{\hat{\boldsymbol{q}}\in\Delta_\mathfrak{l}} \langle \hat{\boldsymbol{q}}, \ell_x(A)\rangle + D_\Phi(\hat{\boldsymbol{q}}, \boldsymbol{q}), \qquad (24) \\
&= \inf_{\hat{\boldsymbol{q}}\in\Delta_\mathfrak{l}} \langle \Pi_\mathfrak{l}\hat{\boldsymbol{q}}, \Pi_\mathfrak{l}\ell_x(A)\rangle + D_{\Phi_\mathfrak{l}}(\Pi_\mathfrak{l}\hat{\boldsymbol{q}}, \Pi_\mathfrak{l}\boldsymbol{q}), \\
&= \inf_{\hat{\boldsymbol{\mu}}\in\Delta_{|\mathfrak{l}|}} \left\langle \hat{\boldsymbol{\mu}}, \ell_x(A\Pi_\mathfrak{l}^\mathsf{T})\right\rangle + D_{\Phi_\mathfrak{l}}(\hat{\boldsymbol{\mu}}, \Pi_\mathfrak{l}\boldsymbol{q}), \qquad (25)
\end{aligned}$$

where in (23) we used the fact that $\Phi_{\mathfrak{l}}(\Pi_{\mathfrak{l}}\boldsymbol{q}) = \Phi(\boldsymbol{q}), \forall \boldsymbol{q} \in \Delta_{\mathfrak{l}}$. Given that $A \mapsto A\Pi_{\mathfrak{l}}^{\mathsf{T}}$ [resp. $\boldsymbol{q} \mapsto \Pi_{\mathfrak{l}}\boldsymbol{q}$] is onto from $\mathcal{A}^k$ to $\mathcal{A}^{|\mathfrak{l}|}$ [resp. from $\Delta_{\mathfrak{l}}$ to $\Delta_{|\mathfrak{l}|}$], (25) implies that ℓ is $\Phi_{\mathfrak{l}}$-mixable.

[**We show** (22)] Suppose that there exists $\hat{\boldsymbol{q}} \in \operatorname{rbd}\Delta_k$ and $\boldsymbol{q} \in \operatorname{ri}\Delta_k$ such that $|\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})| < +\infty$. Let $f\colon [0, \epsilon] \to \mathbb{R}$ be defined by $f(\lambda) \coloneqq \Phi(\hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}}))$, where $\epsilon > 0$ is such that $\hat{\boldsymbol{q}} + \epsilon(\boldsymbol{q} - \hat{\boldsymbol{q}}) \in \operatorname{ri}\Delta_k$. The function f is closed and convex on $\operatorname{dom} f = [0, \epsilon]$ and $\lim_{\lambda\downarrow 0} \frac{f(\lambda) - f(0)}{\lambda} = f'(0; 1) = \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})$ which is finite by assumption. Using this and the fact that $\lambda f'(0; 1) = f'(0; \lambda)$, we have $\lim_{\lambda\downarrow 0} \lambda^{-1}(f(\lambda) - f(0) - f'(0; \lambda)) = 0$. Substituting f by its expression in terms of Φ in the latter equality gives

$$\lim_{\lambda\downarrow 0} \lambda^{-1} D_\Phi(\hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}}), \hat{\boldsymbol{q}}) = 0. \tag{26}$$

Let $\eta > 0$ and $\theta^* \in [k]$ be such that $\hat{q}_{\theta^*} = 0$. Suppose that ℓ is an admissible, Φ-mixable loss. The fact that ℓ is admissible implies that there exists $(x_0, x_1, \boldsymbol{a}_0, \boldsymbol{a}_1) \in [n] \times [n] \times \mathcal{A} \times \mathcal{A}$ such that [13]

$$\boldsymbol{a}_1 \in \operatorname{argmin}\{\ell_{x_0}(\boldsymbol{a}) : \ell_{x_1}(\boldsymbol{a}) = \inf_{\hat{\boldsymbol{a}} \in \mathcal{A}} \ell_{x_1}(\hat{\boldsymbol{a}})\} \text{ and } \inf_{\boldsymbol{a} \in \mathcal{A}} \ell_{x_0}(\boldsymbol{a}) = \ell_{x_0}(\boldsymbol{a_0}) < \ell_{x_0}(\boldsymbol{a}_1). \tag{27}$$

In particular, it holds that $\ell_{x_0}(\boldsymbol{a}_0) < \ell_{x_0}(\boldsymbol{a}_1)$. Fix $A \in \mathcal{A}^k$, such that $A_{\bullet,\theta^*} = \boldsymbol{a}_0$ and $A_{\bullet,\theta} = \boldsymbol{a}_1$ for $\theta \in [k] \setminus \{\theta^*\}$. Let

$$\boldsymbol{a}_* \coloneqq \operatorname*{argmax}_{\boldsymbol{a} \in \Delta_n} \inf_{\boldsymbol{\mu} \in \Delta_k, x \in [n]} \mathsf{m}_\Phi(x, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \boldsymbol{\mu}),$$

with $\hat{\boldsymbol{q}} \in \operatorname{rbd}\Delta_k$ as in (26). Note that $\boldsymbol{a}_*$ exists since ℓ is closed.

If $\boldsymbol{a}_*$ is such that $\ell_{x_1}(\boldsymbol{a}_*) > \ell_{x_1}(\boldsymbol{a}_1)$, then taking $\boldsymbol{\mu} = \hat{\boldsymbol{q}}$ puts all weights on experts predicting $\boldsymbol{a}_1$, while $D_\Phi(\boldsymbol{\mu}, \hat{\boldsymbol{q}}) = 0$. Therefore,

$$\widehat{\mathsf{m}_\Phi} \le \inf_{\boldsymbol{\mu} \in \Delta_k, x \in [n]} \mathsf{m}_\Phi(x, A, \boldsymbol{a}_*, \hat{\boldsymbol{q}}, \boldsymbol{\mu}) \le \mathsf{m}_\Phi(x_1, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \hat{\boldsymbol{q}}) < 0.$$

This contradicts the Φ-mixability of ℓ. Therefore, $\ell_{x_1}(\boldsymbol{a}_*) = \ell_{x_1}(\boldsymbol{a}_1)$, which by (27) implies $\ell_{x_0}(\boldsymbol{a}_*) \ge \ell_{x_0}(\boldsymbol{a}_1)$. For $\boldsymbol{q}^\lambda = \hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}})$, with $\boldsymbol{q} \in \operatorname{ri}\Delta_k$ as in (23) and $\lambda \in [0, \epsilon]$,

$$\begin{aligned}
\widehat{\mathsf{m}_\Phi} &\le \inf_{\boldsymbol{\mu} \in \Delta_k, x \in [n]} \mathsf{m}_\Phi(x, A, \boldsymbol{a}_*, \hat{\boldsymbol{q}}, \boldsymbol{\mu}), \\
&\le \mathsf{m}_\Phi(x_0, A, \boldsymbol{a}, \hat{\boldsymbol{q}}, \boldsymbol{q}^\lambda), \\
&= \langle \boldsymbol{q}^\lambda, \ell_{x_0}(A) \rangle + D_\Phi(\boldsymbol{q}^\lambda, \hat{\boldsymbol{q}}) - \ell_{x_0}(\boldsymbol{a}_*), \\
&= (1 - \lambda q_{\theta^*})\ell_{x_0}(\boldsymbol{a}_1) + \lambda q_{\theta^*}\ell_{x_0}(\boldsymbol{a}_0) + D_\Phi(\boldsymbol{q}^\lambda, \hat{\boldsymbol{q}}) - \ell_{x_0}(\boldsymbol{a}_*), \\
&\le (1 - \lambda q_{\theta^*})\ell_{x_0}(\boldsymbol{a}_*) + \lambda q_{\theta^*}\ell_{x_0}(\boldsymbol{a}_0) + D_\Phi(\boldsymbol{q}^\lambda, \hat{\boldsymbol{q}}) - \ell_{x_0}(\boldsymbol{a}_*), \\
&= \lambda q_{\theta^*}(\ell_{x_0}(\boldsymbol{a}_0) - \ell_{x_0}(\boldsymbol{a}_*)) + D_\Phi(\hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}}), \hat{\boldsymbol{q}}).
\end{aligned}$$

Since $q_{\theta^*} > 0$ ($\boldsymbol{q} \in \operatorname{ri}\Delta_k$) and $\ell_{x_0}(\boldsymbol{a}_0) < \ell_{x_0}(\boldsymbol{a}_1) \le \ell_{x_0}(\boldsymbol{a}_*)$, (23) implies that there exists $\lambda_* > 0$ small enough such that $\lambda_* q_{\theta^*}(\ell_{x_0}(\boldsymbol{a}_0) - \ell_{x_0}(\boldsymbol{a}_*)) + D_\Phi(\hat{\boldsymbol{q}} + \lambda_*(\boldsymbol{q} - \hat{\boldsymbol{q}}), \hat{\boldsymbol{q}}) < 0$. But this implies that $\widehat{\mathsf{m}_\Phi} < 0$ which contradicts the Φ-mixability of ℓ. Therefore, $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})$ is either equal to $+\infty$ or $-\infty$. The former case is not possible. In fact, since Φ is convex, it must have non-decreasing slopes; in particular, it holds that $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) \le \Phi(\boldsymbol{q} - \hat{\boldsymbol{q}}) - \Phi(\hat{\boldsymbol{q}})$. Since Φ is finite on Δ_k (by definition of an entropy), we have $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) < +\infty$. Therefore, we have just shown that

$$\forall \hat{\boldsymbol{q}} \in \operatorname{rbd}\Delta_k, \forall \boldsymbol{q} \in \operatorname{ri}\Delta_k,\ \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = -\infty. \tag{28}$$

Now suppose that $(\hat{\boldsymbol{q}}, \boldsymbol{q}) \in \operatorname{rbd}\Delta_{\mathfrak{l}} \times \operatorname{ri}\Delta_{\mathfrak{l}}$ for $\mathfrak{l} \subseteq [k]$, with $|\mathfrak{l}| > 1$. Note that in this case, we have $(\Phi_{\mathfrak{l}})'(\Pi_{\mathfrak{l}}\hat{\boldsymbol{q}}; \Pi_{\mathfrak{l}}(\boldsymbol{q} - \hat{\boldsymbol{q}})) = \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})$. We showed in the first step of this proof that under the assumptions of the proposition, ℓ must be $\Phi_{\mathfrak{l}}$-mixable. Therefore, repeating the steps above that lead to (28) for Φ, $\hat{\boldsymbol{q}}$, and $\boldsymbol{q}$ substituted by $\Phi_{\mathfrak{l}}$, $\Pi_{\mathfrak{l}}\boldsymbol{q} \in \operatorname{rbd}\Delta_{|\mathfrak{l}|}$, and $\Pi_{\mathfrak{l}}\boldsymbol{q} \in \operatorname{ri}\Delta_{|\mathfrak{l}|}$, we obtain $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = \Phi_{\mathfrak{l}}'(\Pi_{\mathfrak{l}}\hat{\boldsymbol{q}}; \Pi_{\mathfrak{l}}(\boldsymbol{q} - \hat{\boldsymbol{q}})) = -\infty$. This shows (22). □

Lemma 30. *For $\eta > 0$, $\mathrm{S}_\eta \coloneqq \eta^{-1}\,\mathrm{S}$ satisfies* (22) *for all $\mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$, where S is the Shannon entropy.*

Proof. Let $\mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$. Let $(\hat{\boldsymbol{q}}, \boldsymbol{q}) \in \operatorname{rbd} \Delta_{\mathfrak{l}} \times \operatorname{ri} \Delta_{\mathfrak{l}}$ and $\boldsymbol{q}^{\lambda} := \hat{\boldsymbol{q}} + \lambda(\boldsymbol{q} - \hat{\boldsymbol{q}})$, for $\lambda \in]0, 1[$. Let $\mathfrak{J} := \{j \in \mathfrak{l} : \hat{q}_j \neq 0\}$ and $\mathfrak{K} := \mathfrak{l} \setminus \mathfrak{J}$. We have

$$\begin{aligned}
\mathrm{S}(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) &= \lim_{\lambda \downarrow 0} \lambda^{-1} \left[\sum\nolimits_{\theta \in \mathfrak{l}} q_{\theta}^{\lambda} \log q_{\theta}^{\lambda} - \sum\nolimits_{\theta' \in \mathfrak{J}} \hat{q}_{\theta'} \log \hat{q}_{\theta'} \right], \\
&= \lim_{\lambda \downarrow 0} \lambda^{-1} \left[\sum\nolimits_{\theta \in \mathfrak{J}} (q_{\theta}^{\lambda} \log q_{\theta}^{\lambda} - \hat{q}_{\theta} \log \hat{q}_{\theta}) + \sum\nolimits_{\theta' \in \mathfrak{K}} q_{\theta'}^{\lambda} \log q_{\theta'}^{\lambda} \right]. \qquad (29)
\end{aligned}$$

Observe that the limit of either summation term inside the bracket in (29) is equal to zero. Thus, using l'Hopital's rule we get

$$\begin{aligned}
\mathrm{S}(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) &= \lim_{\lambda \downarrow 0} \left[\sum\nolimits_{\theta \in \mathfrak{J}} [(q_{\theta} - \hat{q}_{\theta}) \log q_{\theta}^{\lambda} + (q_{\theta} - \hat{q}_{\theta})] + \sum\nolimits_{\theta' \in \mathfrak{K}} [q_{\theta'} \log q_{\theta'}^{\lambda} + q_{\theta'}] \right], \\
&= \sum\nolimits_{\theta \in \mathfrak{J}} (q_{\theta} - \hat{q}_{\theta}) \log \hat{q}_{\theta} + \sum\nolimits_{\theta' \in \mathfrak{K}} q_{\theta'} \left[\lim_{\lambda \downarrow 0} \log q_{\theta'}^{\lambda} \right], \qquad (30)
\end{aligned}$$

where in (30) we used the fact that $\sum_{\theta \in \mathfrak{J}} (q_{\theta} - \hat{q}_{\theta}) + \sum_{\theta' \in \mathfrak{K}} q_{\theta'} = 0$. Since for all $\theta' \in \mathfrak{K}$, $\lim_{\lambda \downarrow 0} q_{\theta'}^{\lambda} = 0$, the right hand side of (25) is equal to $-\infty$. Therefore S satisfies (22). Since $\mathrm{S}_{\eta} = \eta^{-1} \mathrm{S}$, it is clear that S_{η} also satisfies (22).

□

Lemma 31. *Let $\Phi : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy satisfying* (22) *for all $\mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$. Then for all such $\mathfrak{l}$, it holds that*

$$\forall \boldsymbol{q} \in \Delta_{\mathfrak{l}}, \forall \boldsymbol{\mu} \in \Delta_k \setminus \Delta_{\mathfrak{l}},\ D_{\Phi}(\boldsymbol{\mu}, \boldsymbol{q}) = +\infty.$$

Proof. Let $\boldsymbol{\mu} \in \Delta_k \setminus \Delta_{\mathfrak{l}}$ and $\mathfrak{J} := \{\theta \in [k] : \mu_{\theta} \neq 0\} \cup \mathfrak{l}$. In this case, we have $\boldsymbol{q} \in \operatorname{rbd} \Delta_{\mathfrak{J}}$ and $\boldsymbol{q} + 2^{-1}(\boldsymbol{\mu} - \boldsymbol{q}) \in \operatorname{ri} \Delta_{\mathfrak{J}}$. Thus, since Φ satisfies (22) and $\Phi'(\boldsymbol{q}; \cdot)$ is 1-homogeneous [8, Prop. D.1.1.2], it follows that $2^{-1}\Phi'(\boldsymbol{q}; \boldsymbol{\mu} - \boldsymbol{q}) = \Phi'(\boldsymbol{q}; 2^{-1}(\boldsymbol{\mu} - \boldsymbol{q})) = -\infty$. Hence $D_{\Phi}(\boldsymbol{\mu}, \boldsymbol{q}) = +\infty$. □

Lemma 32. *Let $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy satisfying* (22) *for all $\mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$. If Φ satisfies* (22)*, then $\partial \tilde{\Phi}(\tilde{\boldsymbol{q}}) = \varnothing, \forall \tilde{\boldsymbol{q}} \in \operatorname{bd} \tilde{\Delta}_k$. Furthermore, $\forall \mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$,*

$$\forall \boldsymbol{d} \in \mathbb{R}^k, \forall \boldsymbol{q} \in \operatorname{ri} \Delta_{\mathfrak{l}},\ \operatorname{Mix}_{\Phi}(\boldsymbol{d}, \boldsymbol{q}) = \operatorname{Mix}_{\Phi_{\mathfrak{l}}}(\Pi_{\mathfrak{l}} \boldsymbol{d}, \Pi_{\mathfrak{l}} \boldsymbol{q}).$$

Proof. Let $\boldsymbol{\mu} \in \operatorname{rbd} \Delta_k$. Since Φ satisfies (22), it follows that $\forall \boldsymbol{q} \in \operatorname{ri} \Delta_k$, $\tilde{\Phi}(\tilde{\boldsymbol{\mu}}; \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{\mu}}) = \Phi'(\boldsymbol{\mu}; \boldsymbol{q} - \boldsymbol{\mu}) = -\infty$. Therefore, $\partial \tilde{\Phi}'(\tilde{\boldsymbol{\mu}}) = \varnothing$ [15, Thm. 23.4].

Let $\boldsymbol{d} \in \mathbb{R}^n$, $\mathfrak{l} \subseteq [k]$, with $|\mathfrak{l}| > 1$, and $\boldsymbol{q} \in \operatorname{ri} \Delta_{\mathfrak{l}}$. Then

$$\begin{aligned}
\operatorname{Mix}_{\Phi_{\mathfrak{l}}}(\Pi_{\mathfrak{l}} \boldsymbol{d}, \Pi_{\mathfrak{l}} \boldsymbol{q}) &= \inf_{\boldsymbol{\pi} \in \Delta_{|\mathfrak{l}|}} \langle \boldsymbol{\pi}, \Pi_{\mathfrak{l}} \boldsymbol{d} \rangle + D_{\Phi_{\mathfrak{l}}}(\boldsymbol{\pi}, \Pi_{\mathfrak{l}} \boldsymbol{q}), \\
&= \inf_{\boldsymbol{\mu} \in \Delta_{\mathfrak{l}}} \langle \boldsymbol{\mu}, \boldsymbol{d} \rangle + D_{\Phi}(\boldsymbol{\mu}, \boldsymbol{q}), \\
&\leq \inf_{\boldsymbol{\mu} \in \Delta_k} \langle \boldsymbol{\mu}, \boldsymbol{d} \rangle + D_{\Phi}(\boldsymbol{\mu}, \boldsymbol{q}), \qquad (31) \\
&= \operatorname{Mix}_{\Phi}(\boldsymbol{d}, \boldsymbol{q}).
\end{aligned}$$

To complete the proof, we need to show that (31) holds with equality. For this, it suffices to prove that $\forall \boldsymbol{\mu} \in \Delta_k \setminus \Delta_{\mathfrak{l}}$, $D_{\Phi}(\boldsymbol{\mu}, \boldsymbol{q}) = +\infty$. This follows from Corollary 31. □

Lemma 33. *Let $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy satisfying* (22) *for all $\mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$. Let $x \in [n], \boldsymbol{d} \in \mathbb{R}^k$, and $\boldsymbol{q} \in \Delta_k$. The infimum in*

$$\operatorname{Mix}_{\Phi}(\boldsymbol{d}, \boldsymbol{q}) = \inf_{\boldsymbol{\mu} \in \Delta_k} \langle \boldsymbol{\mu}, \boldsymbol{d} \rangle + D_{\Phi}(\boldsymbol{\mu}, \boldsymbol{q}) \qquad (32)$$

is attained at some $\boldsymbol{q}_* \in \Delta_k$. Furthermore, if $\boldsymbol{q} \in \operatorname{ri} \Delta_k$ and $\boldsymbol{q}_*$ is the infimum of (32) *then for any $\boldsymbol{s}_{\boldsymbol{q}}^* \in \operatorname{argmax}\{\langle \boldsymbol{s}, \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}} \rangle : \boldsymbol{s} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}})\}$, we have*

$$\tilde{\boldsymbol{q}}_* \in \partial \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^{\mathsf{T}} \boldsymbol{d}), \qquad (33)$$

$$\operatorname{Mix}_{\Phi}(\boldsymbol{d}, \boldsymbol{q}) = d_k + \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^*) - \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^{\mathsf{T}} \boldsymbol{d}). \qquad (34)$$

Proof. Let $\boldsymbol{q} \in \operatorname{ri} \Delta_k$. Since $\tilde{\boldsymbol{q}} \in \operatorname{int} \operatorname{dom} \tilde{\Phi} = \operatorname{int} \tilde{\Delta}_k$, the function $\tilde{\boldsymbol{\mu}} \mapsto -\tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}})$ is lower semicontinuous [15, Cor. 24.5.1]. Given that $\tilde{\boldsymbol{\mu}} \mapsto \langle \Pi_k(\tilde{\boldsymbol{\mu}}), \boldsymbol{d} \rangle + \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) - \tilde{\Phi}(\tilde{\boldsymbol{q}})$ is a closed convex function, it is also lower semi-continuous. Therefore, the function

$$\tilde{\boldsymbol{\mu}} \mapsto \langle \Pi_k(\tilde{\boldsymbol{\mu}}), \boldsymbol{d} \rangle + \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - \tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}})$$

is lower semicontinuous, and thus attains its minimum on the compact set $\tilde{\Delta}_k$ at some point $\tilde{\boldsymbol{q}}_*$. Using the fact that $D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}) = D_{\tilde{\Phi}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{q}})$, we get that

$$\boldsymbol{q}_* \coloneqq \Pi_k(\tilde{\boldsymbol{q}}_*) = \operatorname*{argmin}_{\boldsymbol{\mu} \in \Delta_k} \langle \boldsymbol{\mu}, \boldsymbol{d} \rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}). \tag{35}$$

If $\boldsymbol{q} \in \operatorname{rbd} \Delta_k$, then either $\boldsymbol{q}$ is a vertex of Δ_k or there exists $\mathfrak{l} \subsetneq [k]$ such that $\boldsymbol{q} \in \operatorname{ri} \Delta_{\mathfrak{l}}$. In the former case, it follows from (22) that $D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}) = +\infty$ for all $\boldsymbol{\mu} \in \Delta_k \setminus \{\boldsymbol{q}\}$, and thus the infimum of (32) is trivially attained at $\boldsymbol{\mu} = \boldsymbol{q}$. Now consider the alternative — $\boldsymbol{q} \in \Delta_{\mathfrak{l}}$ with $|\mathfrak{l}| > 1$. Using Corollary 31, we have $D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}) = +\infty$ for all $\boldsymbol{\mu} \in \Delta_k \setminus \Delta_{\mathfrak{l}}$. Therefore,

$$\begin{aligned} \operatorname{Mix}_\Phi(\boldsymbol{d}, \boldsymbol{q}) &= \inf_{\boldsymbol{\mu} \in \Delta_{\mathfrak{l}}} \langle \boldsymbol{\mu}, \boldsymbol{d} \rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}), \\ &= \inf_{\hat{\boldsymbol{\mu}} \in \Delta_{|\mathfrak{l}|}} \langle \hat{\boldsymbol{\mu}}, \Pi_{\mathfrak{l}} \boldsymbol{d} \rangle + D_{\Phi_{\mathfrak{l}}}(\hat{\boldsymbol{\mu}}, \Pi_{\mathfrak{l}} \boldsymbol{q}), \end{aligned} \tag{36}$$

where $\Phi_{\mathfrak{l}} \coloneqq \Phi \circ \Pi_{\mathfrak{l}}$. Since $\Pi_{\mathfrak{l}} \boldsymbol{q} \in \operatorname{ri} \Delta_{|\mathfrak{l}|}$, we can use the same argument as the previous paragraph with Φ and $\boldsymbol{q}$ replaced by $\Phi_{\mathfrak{l}}$ and $\Pi_{\mathfrak{l}} \boldsymbol{q}$, respectively, to show that the infimum in (36) is attained at some $\hat{\boldsymbol{q}}_* \in \Delta_{|\mathfrak{l}|}$. Thus, $\boldsymbol{q}_* \coloneqq \Pi_{\mathfrak{l}}^\intercal \hat{\boldsymbol{q}} \in \Delta_k$ attains the infimum in (32).

Now we show the second part of the lemma. Let $\boldsymbol{q} \in \operatorname{ri} \Delta_k$ and $\boldsymbol{q}_*$ be the infimum of (32). Since $\tilde{\Phi}$ is convex and $\tilde{\boldsymbol{q}} = \Pi_k(\boldsymbol{q}) \in \operatorname{int} \tilde{\Delta}_k = \operatorname{int} \operatorname{dom} \tilde{\Phi}$, we have $\partial \tilde{\Phi}(\tilde{\boldsymbol{q}}) \neq \varnothing$ [15, Thm. 23.4]. This means that there exists $\boldsymbol{s}_{\boldsymbol{q}}^* \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}})$ such that $\langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}} \rangle = \tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}})$ [8, p.166]. We will now show that $\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^\intercal \boldsymbol{d} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}}_*)$, which will imply that $\tilde{\boldsymbol{q}}_* \in \partial \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^\intercal \boldsymbol{d})$ (ibid., Cor. D.1.4.4). Let $\boldsymbol{q}_* = \operatorname{argmin}_{\boldsymbol{\mu} \in \Delta_k} \langle \boldsymbol{\mu}, \boldsymbol{d} \rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q})$. Thus, for all $\boldsymbol{\mu} \in \Delta_k$,

$$\begin{aligned} & \langle \boldsymbol{\mu}, \boldsymbol{d} \rangle + \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - \tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}) \geq \langle \boldsymbol{q}_*, \boldsymbol{d} \rangle + \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}} \rangle, \\ \implies \quad & \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) \geq \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) - \langle \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}_*, J_k^\intercal \boldsymbol{d} \rangle + \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{q}}_* \rangle + \Phi'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}), \\ \implies \quad & \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) \geq \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) - \langle \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}_*, J_k^\intercal \boldsymbol{d} \rangle + \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{q}}_* \rangle + \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}} \rangle, \\ \implies \quad & \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) \geq \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) + \langle \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}_*, \boldsymbol{s}_{\boldsymbol{q}}^* - J_k^\intercal \boldsymbol{d} \rangle, \end{aligned}$$

where in the second line we used the fact that $\forall \boldsymbol{q} \in \Delta_k, \langle \boldsymbol{q}, \boldsymbol{d} \rangle = \langle \tilde{\boldsymbol{q}}, J_k^\intercal \boldsymbol{d} \rangle + d_k$, and in third line we used the fact that $\forall \boldsymbol{s} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}}),\ \langle \boldsymbol{s}, \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}} \rangle \leq \tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}})$ (ibid.). This shows that $\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^\intercal \boldsymbol{d} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}}_*)$.

Substituting $\tilde{\Phi}'(\tilde{\boldsymbol{q}}; \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}})$ by $\langle \boldsymbol{s}_{\boldsymbol{q}}^*, \boldsymbol{q}_* - \boldsymbol{q} \rangle$ in the expression for $\operatorname{Mix}_\Phi(\boldsymbol{d}, \boldsymbol{q})$, we get

$$\begin{aligned} \operatorname{Mix}_\Phi(\boldsymbol{d}, \boldsymbol{q}) &= d_k + \langle \tilde{\boldsymbol{q}}_*, J_k^\intercal \boldsymbol{d} \rangle + \tilde{\Phi}(\tilde{\boldsymbol{q}}_*) - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}}_* - \tilde{\boldsymbol{q}} \rangle, \\ &= d_k + \langle \boldsymbol{s}_{\boldsymbol{q}}^*, \tilde{\boldsymbol{q}} \rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}) - [\langle \boldsymbol{s}_{\boldsymbol{q}}^* - J_k^\intercal \boldsymbol{d}, \tilde{\boldsymbol{q}}_* \rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}_*)], \\ &= d_k + \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^*) - \tilde{\Phi}^*(\boldsymbol{s}_{\boldsymbol{q}}^* - J_k^\intercal \boldsymbol{d}), \end{aligned}$$

where in the last line we used the fact that $\tilde{\Phi}$ is a closed convex function, and thus $\forall \tilde{\boldsymbol{q}} \in \tilde{\Delta}_k$, $\boldsymbol{s} \in \partial \tilde{\Phi}(\tilde{\boldsymbol{q}}) \implies \tilde{\Phi}^*(\boldsymbol{s}) = \langle \boldsymbol{s}, \tilde{\boldsymbol{q}} \rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}})$ (ibid., Cor. E.1.4.4).

□

Lemma 34. *Let $\boldsymbol{q} \in \Delta_k$. For any sequence $(\boldsymbol{d}_m)$ in $[0, +\infty[^k$ converging to $\boldsymbol{d} \in [0, +\infty]^k$ coordinate-wise and any entropy $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ satisfying* (22) *for $\mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$,*

$$\lim_{m \to \infty} \operatorname{Mix}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}) = \operatorname{Mix}_\Phi(\boldsymbol{d}, \boldsymbol{q}). \tag{37}$$

Proof of Lemma 34. Let $\boldsymbol{q} \in \Delta_k$ and $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy as in the statement of the Lemma. Let $(\boldsymbol{d}_m) \subset \mathbb{R}^k$ such that $\boldsymbol{d}_m \overset{m \to \infty}{\to} \boldsymbol{d} \in \mathbb{R}^k$. in $[0, +\infty[^k$. Let $\mathfrak{l} \coloneqq \{\theta \in [k] : d_\theta < +\infty\}$.

If $\mathfrak{l} = \varnothing$ then the result holds trivially since, on the one hand, $\mathrm{Mix}_\Phi(\boldsymbol{d}, \boldsymbol{q}) = +\infty$ and on the other hand $\mathrm{Mix}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}) \geq \min_{\theta \in [k]} d_{m,\theta} \overset{m\to\infty}{\to} +\infty$.

Assume now that $\mathfrak{l} \neq \varnothing$. Then

$$\begin{aligned}
\mathrm{Mix}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}) &= \inf_{\boldsymbol{\mu} \in \Delta_k} \langle \boldsymbol{\mu}, \boldsymbol{d}_m \rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}), && (38)\\
&\leq \inf_{\hat{\boldsymbol{\mu}} \in \Delta_\mathfrak{l}} \langle \hat{\boldsymbol{\mu}}, \boldsymbol{d} \rangle + D_\Phi(\hat{\boldsymbol{\mu}}, \boldsymbol{q}), && (39)\\
&< +\infty, && (40)
\end{aligned}$$

where the last inequality stems from the fact that $\Pi_\mathfrak{l} \boldsymbol{d}_m$ is a finite vector in $\mathbb{R}^{|\mathfrak{l}|}$. Therefore, (40) implies that the sequence $\alpha_m \coloneqq \mathrm{Mix}_\Phi(\boldsymbol{d}_m, \boldsymbol{q})$ is bounded. We will show that (α_m) converges in $\mathbb{R}$ and that its limit is exactly $\mathrm{Mix}_\Phi(\boldsymbol{d}, \boldsymbol{q})$. Let $(\hat{\alpha}_m)$ be any convergent subsequence of (α_m), and let $(\hat{\boldsymbol{d}}_m)$ be the corresponding subsequence of $(\boldsymbol{d}_m)$. Consider the infimum in (112) with $\boldsymbol{d}_m$ is replaced by $\hat{\boldsymbol{d}}_m$. From Lemma 33, this infimum is attained at some $\boldsymbol{q}_m \in \Delta_k$. Since Δ_k is compact, we may assume without loss of generality that $\boldsymbol{q}_m$ converges to some $\bar{\boldsymbol{q}} \in \Delta_k$. Observe that $\bar{\boldsymbol{q}}$ must be $\Delta_\mathfrak{l}$; suppose that $\exists \theta_* \in \bar{\mathfrak{l}}$ such that $\bar{q}_{\theta_*} > 0$. Then

$$\begin{aligned}
\hat{\alpha}_m &\geq \langle \boldsymbol{q}_m, \hat{\boldsymbol{d}}_m \rangle,\\
&\geq q_{m,\theta_*} \hat{d}_{m,\theta_*} \overset{m\to\infty}{\to} +\infty.
\end{aligned}$$

This would contradict the fact that α_m is bounded, and thus $\bar{\boldsymbol{q}} \in \Delta_\mathfrak{l}$. Using this, we get

$$\begin{aligned}
\mathrm{Mix}_\Phi(\hat{\boldsymbol{d}}_m, \boldsymbol{q}) &= \langle \boldsymbol{q}_m, \hat{\boldsymbol{d}}_m \rangle + D_\Phi(\boldsymbol{q}_m, \boldsymbol{q}),\\
&\geq \langle \Pi_\mathfrak{l} \boldsymbol{q}_m, \Pi_\mathfrak{l} \hat{\boldsymbol{d}}_m \rangle + D_\Phi(\boldsymbol{q}_m, \boldsymbol{q}),\\
&\overset{m\to\infty}{\to} \langle \Pi_\mathfrak{l} \bar{\boldsymbol{q}}, \Pi_\mathfrak{l} \boldsymbol{d} \rangle + D_\Phi(\bar{\boldsymbol{q}}, \boldsymbol{q}),\\
&= \langle \bar{\boldsymbol{q}}, \boldsymbol{d} \rangle + D_\Phi(\bar{\boldsymbol{q}}, \boldsymbol{q}), && (41)\\
&\geq \inf_{\hat{\boldsymbol{\mu}} \in \Delta_\mathfrak{l}} \langle \hat{\boldsymbol{\mu}}, \boldsymbol{d} \rangle + D_\Phi(\hat{\boldsymbol{\mu}}, \boldsymbol{q}). && (42)
\end{aligned}$$

where in (41) we use the fact that $\bar{\boldsymbol{q}} \in \Delta_\mathfrak{l}$. Combining (42) with (39) shows that $\hat{\alpha}_m$ converges to $\mathrm{Mix}_\Phi(\boldsymbol{d}, \boldsymbol{q}) = \inf_{\hat{\boldsymbol{\mu}} \in \Delta_\mathfrak{l}} \langle \hat{\boldsymbol{\mu}}, \boldsymbol{d} \rangle + D_\Phi(\hat{\boldsymbol{\mu}}, \boldsymbol{q})$. Since $(\hat{\alpha}_m)$ was any convergent subsequence of (α_m) (which is bounded), the result follows. □

C Proofs of Results in the Main Body

C.1 Proof of Theorem 4

Theorem 4 *Any loss $\ell \colon \mathcal{A} \to [0, +\infty]^n$ such that $\mathrm{dom}\, \ell \neq \varnothing$, has a proper support loss $\underline{\ell}$ with the same Bayes risk, $\underline{L}_\ell$, as ℓ.*

Proof. We will construct a proper support loss $\underline{\ell}$ of ℓ.

Let $\boldsymbol{p} \in \mathrm{ri}\, \Delta_n$ ($-\boldsymbol{p} \in \mathrm{int}\, \mathrm{dom}\, \sigma_{\mathscr{S}_\ell}$). Since the support function of a non-empty set is closed and convex, we have $\sigma^{**}_{\mathscr{S}_\ell} = \sigma_{\mathscr{S}_\ell}$ [8, Prop. C.2.1.2]. Pick any $\boldsymbol{v} \in \partial \sigma_{\mathscr{S}_\ell}(-\boldsymbol{p}) = \partial \sigma^{**}_{\mathscr{S}_\ell}(-\boldsymbol{p}) \neq \varnothing$. Since $\sigma^*_{\mathscr{S}_\ell} = \iota_{\mathscr{S}_\ell}$ [15], we can apply Proposition 20-(iv) with f replaced by $\sigma^*_{\mathscr{S}_\ell}$ to obtain $\langle -\boldsymbol{p}, \boldsymbol{v} \rangle = \sigma_{\mathscr{S}_\ell}(-\boldsymbol{p}) + \iota_{\mathscr{S}_\ell}(\boldsymbol{v})$. The fact that $\langle -\boldsymbol{p}, \boldsymbol{v} \rangle$ and $\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})$ are both finite implies that $\iota_{\mathscr{S}_\ell}(\boldsymbol{v}) = 0$. Therefore, $\boldsymbol{v} \in \mathscr{S}_\ell$ and $\langle \boldsymbol{p}, \boldsymbol{v} \rangle = -\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p}) = \underline{L}_\ell(\boldsymbol{p})$. Define $\underline{\ell}(\boldsymbol{p}) \coloneqq \boldsymbol{v} \in \mathscr{S}_\ell$.

Now let $\boldsymbol{p} \in \mathrm{rbd}\, \Delta_n$ and $\boldsymbol{q} \coloneqq \mathbf{1}_n / n$. Since the $\underline{L}_{\underline{\ell}}$ is a closed concave function and $\boldsymbol{q} \in \mathrm{int}\, \mathrm{dom}\, \underline{L}_{\underline{\ell}}$, it follows that $\underline{L}_{\underline{\ell}}(\boldsymbol{p} + m^{-1}(\boldsymbol{q} - \boldsymbol{p})) \overset{m\to\infty}{\to} \underline{L}_{\underline{\ell}}(\boldsymbol{p})$ [8, Prop. B.1.2.5]. Note that $\boldsymbol{q}_m \coloneqq \boldsymbol{p} + m^{-1}(\boldsymbol{q} - \boldsymbol{p}) \in \mathrm{ri}\, \Delta_n, \forall m \in \mathbb{N}$. Now let $v_{x,m} \coloneqq \underline{\ell}_x(\boldsymbol{q}_m)$, where $\underline{\ell}(\boldsymbol{q}_m)$ is as constructed in the previous paragraph. If $(v_{1,m})$ is bounded [resp. unbounded], we can extract a subsequence $(v_{1,\varphi_1(m)})$ which converges [resp. diverges to $+\infty$], where $\varphi_1 : \mathbb{N} \to \mathbb{N}$ is an increasing function. By repeating this process for $(v_{2,\varphi_1(m)})$ and so on, we can construct an increasing function $\varphi \coloneqq \varphi_n \circ \cdots \circ \varphi_1 : \mathbb{N} \to \mathbb{N}$, such that $\boldsymbol{v}_m \coloneqq [v_{x,\varphi(m)}]^\mathsf{T}_{x \in [n]}$ has a well defined (coordinate-wise) limit in $[0, +\infty]^n$. Define $\underline{\ell}(\boldsymbol{p}) \coloneqq \lim_{m\to\infty} \boldsymbol{v}_m$. By continuity of the inner product, we have

$$\begin{aligned}
\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p}) \rangle = \lim_{m\to\infty} \langle \boldsymbol{q}_{\varphi(m)}, \boldsymbol{v}_m \rangle &= \lim_{m\to\infty} \langle \boldsymbol{q}_{\varphi(m)}, \underline{\ell}(\boldsymbol{q}_{\varphi(m)}) \rangle,\\
&= \lim_{m\to\infty} \underline{L}_\ell(\boldsymbol{q}_{\varphi(m)}) = \underline{L}_\ell(\boldsymbol{p}).
\end{aligned}$$

By construction, $\forall m \in \mathbb{N}, \boldsymbol{p}_m \coloneqq \boldsymbol{q}_{\varphi(m)} \in \operatorname{ri}\Delta_n$ and $\underline{\ell}(\boldsymbol{p}_m) = \boldsymbol{v}_m \overset{m\to\infty}{\to} \underline{\ell}(\boldsymbol{p})$. Therefore, $\underline{\ell}$ is support loss of ℓ.

It remains to show that it is proper; that is $\forall \boldsymbol{p} \in \Delta_n, \forall \boldsymbol{q} \in \Delta_n$, $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{q})\rangle$. Let $\boldsymbol{q} \in \operatorname{ri}\Delta_n$. We just showed that $\forall \boldsymbol{p} \in \Delta_n, \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle = \underline{L}_\ell(\boldsymbol{p})$ and that $\underline{\ell}(\boldsymbol{q}) \in \mathscr{S}_\ell$. Using the fact that $\underline{L}_\ell(\boldsymbol{p}) = \inf_{\boldsymbol{z}\in\mathscr{S}_\ell}\langle \boldsymbol{p}, \boldsymbol{z}\rangle$, we obtain $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{q})\rangle$.

Now let $\boldsymbol{q} \in \operatorname{rbd}\Delta_k$. Since $\underline{\ell}$ is a support loss, we know that there exists a sequence $(\boldsymbol{q}_m) \subset \operatorname{ri}\Delta_n$ such that $\underline{\ell}(\boldsymbol{q}_m) \overset{m\to\infty}{\to} \underline{\ell}(\boldsymbol{q})$. But as we established in the previous paragraph, $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{q}_m)\rangle$. By passing to the limit $m \to \infty$, we obtain $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{q})\rangle$. Therefore $\underline{\ell}$ is a proper loss with Bayes risk $\underline{L}_\ell$. □

C.2 Proofs of Theorem 5 and Proposition 12

For a set $\mathcal{C}$, we denote $\operatorname{co}\mathcal{C}$ and $\overline{\operatorname{co}}\mathcal{C}$ its *convex hull* and *closed convex hull*, respectively.

Definition 35 ([8])**.** *Let $\mathcal{C}$ be non-empty convex set in $\mathbb{R}^n$. We say that $\boldsymbol{u} \in \mathcal{C}$ is an extreme point of $\mathcal{C}$ if there are no two different points $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ in $\mathcal{C}$ and $\lambda \in]0,1[$ such that $\boldsymbol{u} = \lambda\boldsymbol{u}_1 + (1-\lambda)\boldsymbol{u}_2$.*

We denote the set of extreme points of a set $\mathcal{C}$ by $\operatorname{ext}\mathcal{C}$.

Lemma 36. *Let $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a closed loss. Then $\operatorname{ext}\overline{\operatorname{co}}\mathscr{S}_\ell \subseteq \mathcal{S}_\ell$.*

Proof. Since $\operatorname{co}\mathscr{S}_\ell \subseteq \mathbb{R}^n$ is connected, $\operatorname{co}\mathscr{S}_\ell = \{\boldsymbol{v} + \sum_{k=1}^n \alpha_k\ell(\boldsymbol{a}_k)\colon (\boldsymbol{a}_{k\in[n]}, \boldsymbol{\alpha}, \boldsymbol{v}) \in \mathcal{A}^n \times \Delta_n \times [0,+\infty[^n\}$ [8, Prop. A.1.3.7].

We claim that $\overline{\operatorname{co}}\mathscr{S}_\ell = \operatorname{co}\mathscr{S}_\ell$. Let $(\boldsymbol{z}_m) \coloneqq (\boldsymbol{v}_m + \sum_{k=1}^n \alpha_{m,k}\ell(\boldsymbol{a}_{m,k}))$ be a convergent sequence in $[0,+\infty[^n$, where $(\boldsymbol{\alpha}_m)$, $([\boldsymbol{a}_{m,k}]_{k\in[n]})$ and $(\boldsymbol{v}_m)$ are sequences in Δ_n, $\mathcal{A}^n$, and $[0,+\infty[^n$, respectively. Since Δ_n is compact, we may assume, by extracting a subsequence if necessary, that $\boldsymbol{\alpha}_m \overset{m\to\infty}{\to} \boldsymbol{\alpha}^* \in \Delta_n$. Let $\mathcal{K} \coloneqq \{k \in [n] : \alpha_k^* \neq 0\}$. Since $\boldsymbol{z}_m$ converges, $([[\ell(\boldsymbol{a}_{m,k})]_{k\in\mathcal{K}}, \boldsymbol{v}_m])$ is a bounded sequence in $[0,+\infty[^{n|\mathcal{K}|+n}$. Since ℓ is closed, we may assume, by extraction a subsequence if necessary, that $\forall k \in \mathcal{K}$, $\ell(\boldsymbol{a}_{m,k}) \overset{m\to\infty}{\to} \ell(\boldsymbol{a}_k^*)$ and $\boldsymbol{v}_m \overset{m\to\infty}{\to} \boldsymbol{v}^*$, where $[\boldsymbol{a}_k^*]_{k\in\mathcal{K}} \in \mathcal{A}^{|\mathcal{K}|}$ and $\boldsymbol{v}^* \in [0,+\infty[^n$. Consequently,

$$\begin{aligned}\boldsymbol{v}^* + \sum_{k=1}^n \alpha_k^*\ell(\boldsymbol{a}_k^*) &= \lim_{m\to\infty}\left[\boldsymbol{v}_{m,k} + \sum_{k\in\mathcal{K}} \alpha_{m,k}\ell(\boldsymbol{a}_{m,k})\right],\\ &\leq \lim_{m\to\infty} \boldsymbol{z}_m,\end{aligned}$$

where the last inequality is coordinate-wise. Therefore, there exists $\boldsymbol{v}' \in [0,+\infty[^n$ such that $\lim_{m\to\infty}\boldsymbol{z}_m = \boldsymbol{v}' + \boldsymbol{v}^* + \sum_{k=1}^n \alpha_k^*\ell(\boldsymbol{a}_k^*) \in \operatorname{co}\mathscr{S}_\ell$. This shows that $\overline{\operatorname{co}}\mathscr{S}_\ell \subset \operatorname{co}\mathscr{S}_\ell$, and thus $\overline{\operatorname{co}}\mathscr{S}_\ell = \operatorname{co}\mathscr{S}_\ell$ which proves our first claim.

By definition of an extreme point, $\operatorname{ext}\overline{\operatorname{co}}\mathscr{S}_\ell \subseteq \overline{\operatorname{co}}\mathscr{S}_\ell$. Let $\boldsymbol{e} \in \operatorname{ext}\overline{\operatorname{co}}\mathscr{S}_\ell$ and $(\boldsymbol{a}_{k\in[n]}, \boldsymbol{\alpha}, \boldsymbol{v}) \in \mathcal{A}^n \times \Delta_n \times [0,+\infty[^n$ such that $\boldsymbol{e} = \sum_{k=1}^n \alpha_k\ell(\boldsymbol{a}_k) + \boldsymbol{v}$. If there exists $i, j \in [n]$ such that $\alpha_i\alpha_j \neq 0$ or $\alpha_i v_j \neq 0$ then $\boldsymbol{e}$ would violate the definition of an extreme point. Therefore, the only possible extreme points are of the form $\{\ell(\boldsymbol{a}) : \boldsymbol{a} \in \operatorname{dom}\ell)\} = \mathcal{S}_\ell$. □

Theorem 5 *Let $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a loss and $\underline{\ell}$ be a proper support loss of ℓ. If the Bayes risk $\underline{L}_\ell$ is differentiable on $]0,+\infty[^n$, then $\underline{\ell}$ is uniquely defined on $\operatorname{ri}\Delta_n$ and*

$$\begin{aligned}&\forall \boldsymbol{p} \in \operatorname{dom}\underline{\ell}, \quad \exists \boldsymbol{a}_* \in \operatorname{dom}\ell, \quad \ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p}),\\ &\forall \boldsymbol{a} \in \operatorname{dom}\ell, \quad \exists (\boldsymbol{p}_m) \subset \operatorname{ri}\Delta_n, \quad \underline{\ell}(\boldsymbol{p}_m) \overset{m\to\infty}{\to} \ell(\boldsymbol{a}) \ \textit{coordinate-wise.}\end{aligned}$$

Proof. Let $\boldsymbol{p} \in \operatorname{ri}\Delta_n$ and suppose that $\underline{L}_\ell$ is differentiable at $\boldsymbol{p}$. In this case, $\sigma_{\mathscr{S}_\ell}$ is differentiable at $-\boldsymbol{p}$, which implies [8, Cor. D.2.1.4]

$$\mathcal{F}(\boldsymbol{p}) \coloneqq \partial\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p}) = \{\nabla\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})\}. \tag{43}$$

On the other hand, the fact that $\sigma_{\mathscr{S}_\ell} = \sigma_{\overline{\operatorname{co}}\mathscr{S}_\ell}$ [8, Prop. C.2.2.1], implies $\mathcal{F}(\boldsymbol{p}) = \partial\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p}) = \partial\sigma_{\overline{\operatorname{co}}\mathscr{S}_\ell}(-\boldsymbol{p})$. The latter being an *exposed face* of $\overline{\operatorname{co}}\mathscr{S}_\ell$ implies that every extreme point of $\mathcal{F}(\boldsymbol{p})$

is also an extreme point of $\overline{\mathrm{co}}\mathscr{S}_\ell$ [8, Prop. A.2.3.7, Prop. A.2.4.3]. Therefore, from (43), $\underline{\ell}(\boldsymbol{p}) = \nabla\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})$ is the only extreme point of $\mathcal{F}(\boldsymbol{p}) \subset \overline{\mathrm{co}}\mathscr{S}_\ell$. From Lemma 36, there exists $\boldsymbol{a}_* \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p})$. In this paragraph, we showed the following

$$\forall \boldsymbol{p} \in \mathrm{ri}\,\Delta_n, \exists \boldsymbol{a}_* \in \mathrm{dom}\,\ell,\ \ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p}). \tag{44}$$

For the rest of this proof we will assume that $\underline{L}_\ell$ is differentiable on $]0,+\infty[^n$. Let $\boldsymbol{p} \in \mathrm{rbd}\,\Delta_n \cap \mathrm{dom}\,\underline{\ell}$. Since $\underline{\ell}$ is a support loss, there exists $(\boldsymbol{p}_m)$ in $\mathrm{ri}\,\Delta_n$ such that $(\underline{\ell}(\boldsymbol{p}_m))_m$ converges to $\underline{\ell}(\boldsymbol{p})$. From (44) it holds that $\forall \boldsymbol{p}_m \in \mathrm{ri}\,\Delta_n, \exists \boldsymbol{a}_m \in \mathcal{A}, \ell(\boldsymbol{a}_m) = \underline{\ell}(\boldsymbol{p}_m)$. Since $(\ell(\boldsymbol{a}_m))_m$ converges and ℓ is closed, there exists $\boldsymbol{a}_* \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_*) = \lim_{m\to\infty} \ell(\boldsymbol{a}_m) = \underline{\ell}(\boldsymbol{p})$.

Now let $\boldsymbol{a} \in \mathrm{dom}\,\ell$ and $f(\boldsymbol{p}, x) := \underline{\ell}_x(\boldsymbol{p}) - \ell_x(\boldsymbol{a})$. Since $\ell(\boldsymbol{a}) \in \mathscr{S}_\ell$ and $\underline{\ell}$ is proper, we have for all $\boldsymbol{p} \in \mathrm{ri}\,\Delta_n, \mathbb{E}_{x\sim\boldsymbol{p}}[f(\boldsymbol{p}, x)] \leq 0$ and $-\infty < f(\boldsymbol{p}, x), \forall x \in [n]$. Therefore, Lemma 24 implies that for all $m \in \mathbb{N} \setminus \{0\}$ there exists $\boldsymbol{p}_m \in \mathrm{ri}\,\Delta_n$, such that $\forall x \in [n], \underline{\ell}_x(\boldsymbol{p}_m) \leq \ell_x(\boldsymbol{a}) + 1/m$. On one hand, since $(\underline{\ell}(\boldsymbol{p}_m))$ is bounded (from the previous inequality), we may assume by extracting a subsequence if necessary, that $(\underline{\ell}(\boldsymbol{p}_m))_m$ converges. On the other hand, since $\boldsymbol{p}_m \in \mathrm{ri}\,\Delta_n$, (44) implies that there exists $\boldsymbol{a}_m \in \mathrm{dom}\,\ell$ such that $\underline{\ell}(\boldsymbol{p}_m) = \ell(\boldsymbol{a}_m)$. Since ℓ is closed and $(\ell(\boldsymbol{a}_m))_m$ converges, there exists $\boldsymbol{a}_* \in \mathcal{A}$, such that $\ell(\boldsymbol{a}_*) = \lim_{m\to\infty} \ell(\boldsymbol{a}_m) = \lim_{m\to\infty} \underline{\ell}(\boldsymbol{p}_m) \leq \ell(\boldsymbol{a})$. But since ℓ is admissible, the latter component-wise inequality implies that $\ell(\boldsymbol{a}_*) = \ell(\boldsymbol{a}) = \lim_{m\to\infty} \underline{\ell}(\boldsymbol{p})$. □

Lemma 37. *Let $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a loss satisfying Assumption 1. If $\underline{L}_\ell$ is not differentiable at $\boldsymbol{p}$ then there exist $\boldsymbol{a}_0, \boldsymbol{a}_1 \in \mathrm{dom}\,\ell$, such that $\ell(\boldsymbol{a}_0) \neq \ell(\boldsymbol{a}_1)$ and $\underline{L}_\ell(\boldsymbol{p}) = \langle \boldsymbol{p}, \ell(\boldsymbol{a}_0)\rangle = \langle \boldsymbol{p}, \ell(\boldsymbol{a}_1)\rangle$.*

Proof. Suppose $\underline{L}_\ell$ is not differentiable at $\boldsymbol{p} \in \mathrm{ri}\,\Delta_n$. Then from the definition of the Bayes risk, $\sigma_{\mathscr{S}_\ell}$ is not differentiable at $-\boldsymbol{p}$. This implies that $\mathcal{F}(\boldsymbol{p}) := \partial\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})$ has more than one element [8, Cor. D.2.1.4]. Since $\sigma_{\mathscr{S}_\ell} = \sigma_{\overline{\mathrm{co}}\mathscr{S}_\ell}$ (ibid.. Prop. C.2.2.1), $\mathcal{F}(\boldsymbol{p}) = \partial\sigma_{\overline{\mathrm{co}}\mathscr{S}_\ell}(-\boldsymbol{p})$ is a subset of $\overline{\mathrm{co}}\mathscr{S}_\ell$ and every extreme point of $\mathcal{F}(\boldsymbol{p})$ is also an extreme point of $\overline{\mathrm{co}}\mathscr{S}_\ell$ (ibid., Prop. A.2.3.7). Thus, from Lemma 36, we have $\mathrm{ext}\,\mathcal{F}(\boldsymbol{p}) \subset \mathcal{S}_\ell$. On the other hand, since $-\boldsymbol{p} \in \mathrm{int}\,\mathrm{dom}\,\sigma_{\mathscr{S}_\ell}$, $\mathcal{F}(\boldsymbol{p})$ is a compact, convex set [15, Thm. 23.4], and thus $\mathcal{F}(\boldsymbol{p}) = \mathrm{co}(\mathrm{ext}\,\mathcal{F}(\boldsymbol{p}))$ [8, Thm. A.2.3.4]. Hence, the fact that $\mathcal{F}(\boldsymbol{p})$ has more than one element implies there exists $\boldsymbol{a}_0, \boldsymbol{a}_1 \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_0), \ell(\boldsymbol{a}_1) \in \mathrm{ext}\,\mathcal{F}(\boldsymbol{p}) \subseteq \mathcal{F}(\boldsymbol{p})$ and $\ell(\boldsymbol{a}_0) \neq \ell(\boldsymbol{a}_1)$. Since $\mathcal{F}(\boldsymbol{p}) = \partial\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})$, Proposition 20-(iv) and the fact that $\sigma^*_{\mathscr{S}_\ell} = \iota_{\mathscr{S}_\ell}$ imply that $\underline{L}_\ell(\boldsymbol{p}) = \langle \boldsymbol{p}, \ell(\boldsymbol{a}_0)\rangle = \langle \boldsymbol{p}, \ell(\boldsymbol{a}_1)\rangle$. □

Proposition 12 *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be an entropy and $\ell\colon \mathcal{A} \to [0,+\infty]^n$. If ℓ is Φ-mixable, then the Bayes risk satisfies $\underline{L}_\ell \in C^1(]0,+\infty[^n)$. If, additionally, $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$, then Φ must be strictly convex on Δ_k.*

Proof. Let $\mathfrak{l} = \{1, 2\}$. Since ℓ is Φ-mixable, it must be $\Phi_\mathfrak{l}$-mixable, where $\Phi_\mathfrak{l} := \Phi_\mathfrak{l} \circ \Pi_\mathfrak{l}^\mathsf{T} : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ (Proposition 29). Let $\Psi := \Phi_\mathfrak{l}$.

For $w \in]0,+\infty[$ and $z \in \mathrm{int}\,\mathrm{dom}\,\tilde{\Psi}^* = \mathbb{R}$ (see appendix E), we define $(\tilde{\Psi}^*)'_\infty(w) := \lim_{t\to+\infty}[\tilde{\Psi}^*(z + tw) - \tilde{\Psi}^*(z)]/t$. The value of $(\tilde{\Psi}^*)'_\infty(w)$ does not depend on the choice of z, and it holds that $(\tilde{\Psi}^*)'_\infty(w) = \sigma_{\mathrm{dom}\,\tilde{\Psi}}(w)$ and $(\tilde{\Psi}^*)'_\infty(-w) = \sigma_{\mathrm{dom}\,\tilde{\Psi}}(-w)$ [8, Prop. C.1.2.2]. In our case, we have $\mathrm{dom}\,\tilde{\Psi} = [0, 1]$ (by definition of $\tilde{\Psi}$), which implies that $\sigma_{\mathrm{dom}\,\tilde{\Psi}}(1) = 1$ and $\sigma_{\mathrm{dom}\,\tilde{\Psi}}(-1) = 0$. Therefore, $(\tilde{\Psi}^*)'_\infty(1) + (\tilde{\Psi}^*)'_\infty(-1) = 1$. As a result $\tilde{\Psi}^*$ cannot be affine. For all $\delta > 0$, let $g_\delta\colon \mathbb{R} \times \{-1, 0, +1\} \to \mathbb{R}$ be defined by

$$g_\delta(s, u) := [\tilde{\Psi}^*(s + \delta(u + 1)/2) - \tilde{\Psi}^*(s + \delta(u - 1)/2)]/\delta.$$

Since $\tilde{\Psi}^*$ is convex it must have non-decreasing slopes (ibid., p.13). Combining this with the fact that $\tilde{\Psi}^*$ is not affine implies that

$$\exists s^*_\delta \in \mathbb{R},\ g_\delta(s^*_\delta, -1) < g_\delta(s^*_\delta, +1). \tag{45}$$

The fact that $\tilde{\Psi}^*$ has non-decreasing slopes also implies that

$$g_\delta(s^*_\delta, +1) = [\tilde{\Psi}^*(s^*_\delta + \delta) - \tilde{\Psi}^*(s^*_\delta)]/\delta \leq \lim_{t\to\infty}[\tilde{\Psi}^*(s^*_\delta + t) - \tilde{\Psi}^*(s^*_\delta)]/t = (\tilde{\Psi}^*)'_\infty(1) = 1.$$

Similarly, we have $0 = -(\tilde{\Psi}^*)'_\infty(-1) \leq g_\delta(s^*_\delta, -1)$. Let $\tilde{\mu} \in \partial\tilde{\Psi}^*(s^*_\delta)$. Since $\tilde{\Psi}$ is a closed convex function the following equivalence holds $\tilde{\mu} \in \partial\tilde{\Psi}^*(s^*_\delta) \iff s^*_\delta \in \partial\tilde{\Psi}(\tilde{\mu})$ (ibid., Cor. D.1.4.4).

Thus, if $\tilde{\mu} \in \{0,1\} = \operatorname{bd} \tilde{\Delta}_2$, then $\partial\tilde{\Psi}(\tilde{\mu}) \neq \varnothing$, which is not possible since ℓ is Ψ-mixable (Lemma 32).

[**We show** $\underline{L}_\ell \in C^1(]0,+\infty[^n)$] We will now show that $\underline{L}_\ell$ is continuously differentiable on $]0,+\infty[^n$. Since $\underline{L}_\ell$ is 1-homogeneous, it suffices to check the differentiability on $\operatorname{ri}\Delta_n$. Suppose $\underline{L}_\ell$ is not differentiable at $\boldsymbol{p} \in \operatorname{ri}\Delta_n$. From Lemma 37, there exists $\boldsymbol{a}_0, \boldsymbol{a}_1 \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_0), \ell(\boldsymbol{a}_1) \in \partial\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})$ and $\ell(\boldsymbol{a}_0) \neq \ell(\boldsymbol{a}_1)$. Let $A := [\boldsymbol{a}_0, \boldsymbol{a}_1] \in \mathbb{R}^{n\times 2}$, $\delta := \min\{|\ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)| : x \in [n], |\ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)| > 0\}$, and $s^*_\delta \in \mathbb{R}$ as in (45). We denote $g^- := g_\delta(s^*_\delta, -1)$ and $g^+ := g_\delta(s^*_\delta, +1) \in]0,1]$. Let $\tilde{\mu} \in \partial\tilde{\Psi}^*(s^*_\delta) \in \operatorname{int}\tilde{\Delta}_2$ and $\boldsymbol{\mu} = \amalg_2(\tilde{\mu}) \in \operatorname{ri}\Delta_2$. From the fact that ℓ is Ψ-mixable, $J_2^\mathsf{T}\ell_x(A) = \ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)$, and (8), there must exist $\boldsymbol{a}_* \in \mathcal{A}$ such that for all $x \in [n]$,

$$\begin{aligned}\ell_x(\boldsymbol{a}_*) &\leq \operatorname{Mix}_\Psi(\ell_x(A), \boldsymbol{\mu}),\\ &= \ell_x(a_1) + \tilde{\Psi}^*(s^*_\delta) - \tilde{\Psi}^*(s^*_\delta - \ell_x(\boldsymbol{a}_0) + \ell_x(\boldsymbol{a}_1)),\end{aligned}$$

and by letting sgn be the *sign* function

$$\leq \ell_x(a_1) + g_\delta(s^*_\delta, -\operatorname{sgn}[\ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)])[\ell_x(\boldsymbol{a}_0) - \ell_x(\boldsymbol{a}_1)], \tag{46}$$

where in (46) we used the fact that $\tilde{\Psi}^*$ has non-decreasing slopes and the definition of δ. When $\ell_x(\boldsymbol{a}_0) \leq \ell_x(\boldsymbol{a}_1)$, (46) becomes $\ell_x(\boldsymbol{a}_*) \leq (1-g^+)\ell_x(\boldsymbol{a}_1) + g^+\ell_x(\boldsymbol{a}_0)$. Otherwise, we have $\ell_x(\boldsymbol{a}_*) \leq (1-g^-)\ell_x(\boldsymbol{a}_1) + g^-\ell_x(\boldsymbol{a}_0) < (1-g^+)\ell_x(\boldsymbol{a}_1) + g^+\ell_x(\boldsymbol{a}_0)$. Since ℓ is admissible, there must exist at least one $x \in [n]$ such that $\ell_x(\boldsymbol{a}_0) > \ell_x(\boldsymbol{a}_1)$. Combining this with the fact that $p_x > 0, \forall x \in [n]$ ($\boldsymbol{p} \in \operatorname{ri}\Delta_n$), implies that $\langle \boldsymbol{p}, \ell(\boldsymbol{a}_*)\rangle < \langle \boldsymbol{p}, (1-g^+)\ell(\boldsymbol{a}_1) + g^+\ell(\boldsymbol{a}_0)\rangle = \underline{L}_\ell(\boldsymbol{p})$. This contradicts the fact that $\ell(\boldsymbol{a}_*) \in \mathscr{S}_\ell$. Therefore, $\underline{L}_\ell$ must be differentiable at $\boldsymbol{p}$. As argued earlier, this implies that $\underline{L}_\ell$ must be differentiable on $]0,+\infty[^n$. Combining this with the fact that $\underline{L}_\ell$ is concave on $]0,+\infty[^n$, implies that $\underline{L}_\ell$ is continuously differentiable on $]0,+\infty[^n$ (ibid., Rmk. D.6.2.6).

[**We show** $\tilde{\Phi}^* \in C^1(\mathbb{R}^{k-1})$] Suppose that $\tilde{\Phi}^*$ is not differentiable at some $\boldsymbol{s}^* \in \mathbb{R}^{k-1}$. Then there exists $\boldsymbol{d} \in \mathbb{R}^{k-1} \setminus \{\mathbf{0}_{\tilde{k}}\}$ such that $-(\tilde{\Phi}^*)'(\boldsymbol{s}^*; -\boldsymbol{d}) < (\tilde{\Phi}^*)'(\boldsymbol{s}^*; \boldsymbol{d})$. Since $\boldsymbol{s}^* \in \operatorname{int}\operatorname{dom}\tilde{\Phi}^*$, $(\tilde{\Phi}^*)'(\boldsymbol{s}^*, \cdot)$ is finite and convex [8, Prop. D.1.1.2], and thus it is continuous on $\operatorname{dom}\tilde{\Phi}^* = \mathbb{R}^{k-1}$ (ibid., Rmk. B.3.1.3). Consequently, there exists $\delta^* > 0$ such that

$$\forall \hat{\boldsymbol{d}} \in \mathbb{R}^{k-1}, \|\hat{\boldsymbol{d}} - \boldsymbol{d}\| \leq \delta^* \implies -(\tilde{\Phi}^*)'(\boldsymbol{s}^*; -\hat{\boldsymbol{d}}) < (\tilde{\Phi}^*)'(\boldsymbol{s}^*; \hat{\boldsymbol{d}}) \tag{47}$$

Let $g\colon \{-1,1\} \to \mathbb{R}$ be such that

$$g(u) := \sup_{\|\hat{\boldsymbol{d}} - \boldsymbol{d}\| \leq \delta^*} u \cdot (\tilde{\Phi}^*)'(\boldsymbol{s}^*; u\hat{\boldsymbol{d}}).$$

Note that since $\tilde{\Phi}^*$ has increasing slopes ($\tilde{\Phi}^*$ is convex), $g(1) \leq \sup_{\|\hat{\boldsymbol{d}} - \boldsymbol{d}\| \leq \delta^*} (\tilde{\Phi}^*)'_\infty(\hat{\boldsymbol{d}}) = \sup_{\|\hat{\boldsymbol{d}} - \boldsymbol{d}\| \leq \delta^*} \sigma_{\operatorname{dom}\tilde{\Phi}}(\hat{\boldsymbol{d}}) \leq 1$, where the last inequality holds because $\tilde{\Delta}_k \subset \mathcal{B}(\mathbf{0}_{\tilde{k}}, 1)$, and thus $\sigma_{\operatorname{dom}\tilde{\Phi}}(\hat{\boldsymbol{d}}) = \sigma_{\tilde{\Delta}_k}(\hat{\boldsymbol{d}}) \leq \sigma_{\mathcal{B}(\mathbf{0}_{\tilde{k}},1)}(\hat{\boldsymbol{d}}) = 1$. Let $\Delta g := g(1) - g(-1)$. From (47), it is clear that $\Delta g > 0$.

Suppose that $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$ and let $\underline{\ell}$ be a support loss of ℓ. By definition of a support loss, $\forall \boldsymbol{p} \in \operatorname{ri}\Delta_k, \tilde{\underline{\ell}}(\tilde{\boldsymbol{p}}) = \underline{\ell}(\boldsymbol{p}) = \nabla\underline{L}_\ell(\boldsymbol{p})$ (where $\tilde{\underline{\ell}} := \underline{\ell} \circ \amalg_n$). Thus, since $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$, $\tilde{\underline{\ell}}$ is differentiable on $\operatorname{int}\tilde{\Delta}_n$. Furthermore, $\underline{\ell}$ is continuous on $\operatorname{ri}\Delta_k$ given that $\underline{L}_\ell \in C^1(]0,+\infty[^n)$ as shown in the first part of this proof. We may assume without loss of generality that ℓ is not a constant function. Thus, from Theorem 5, $\underline{\ell}$ is not a constant function either. Consequently, the mean value theorem applied to $\underline{\ell}$ (see e.g. [16, Thm. 5.10]) between any two points in $\operatorname{ri}\Delta_n$ with distinct images under $\underline{\ell}$, implies that there exists $(\tilde{\boldsymbol{p}}_*, \boldsymbol{v}_*) \in \operatorname{int}\tilde{\Delta}_n \times \mathbb{R}^{n-1}$, such that $\mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}}_*)\boldsymbol{v}_* \neq \mathbf{0}_{\tilde{n}}$. For the rest of the proof let $(\tilde{\boldsymbol{p}}, \boldsymbol{v}) := (\tilde{\boldsymbol{p}}_*, \boldsymbol{v}_*)$ and define $\mathfrak{I} := \{x \in [n] : \mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} \neq 0\}$. From Lemma 27, we have $\langle \boldsymbol{p}, \mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^\mathsf{T}$, which implies that there exists $x \in \mathfrak{I}, \mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} > 0$. Thus, the set

$$\mathfrak{K} := \left\{x \in \mathfrak{I}\colon \mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} > 0\right\} \tag{48}$$

is non-empty. From this and the fact that $\boldsymbol{p} \in \operatorname{ri}\Delta_n$, it follows that

$$\sum_{x' \in \mathfrak{K}} p_{x'}\mathsf{D}\tilde{\underline{\ell}}_{x'}(\tilde{\boldsymbol{p}})\boldsymbol{v} > 0. \tag{49}$$

Let $\tilde{\boldsymbol{p}}^t := \tilde{\boldsymbol{p}} + t\boldsymbol{v}$. From Taylor's Theorem (see e.g. [3, §151]) applied to the function $t \mapsto \tilde{\underline{\ell}}(\tilde{\boldsymbol{p}}^t)$, there exists $\epsilon^* > 0$ and functions $\delta_x : [-\epsilon^*, \epsilon^*] \to \mathbb{R}^n$, $x \in [n]$, such that $\lim_{t\to 0} t^{-1}\delta_x(t) = 0$ and

$$\forall |t| \le \epsilon^*, \quad \underline{\ell}_x(\boldsymbol{p}^t) = \underline{\ell}_x(\boldsymbol{p}) + t\mathsf{D}\tilde{\underline{\ell}}_x(\boldsymbol{p})\boldsymbol{v} + \delta_x(t). \tag{50}$$

For $x \in [n]$, let $\boldsymbol{d}_x \in \mathbb{R}^{\tilde{k}}$ and suppose that $\|\boldsymbol{d}_x - \boldsymbol{d}\| \le \delta^*$ (we will define $\boldsymbol{d}_x$ explicitly later). By shrinking ϵ^* if necessary, we may assume that

$$\forall x \in \mathfrak{I}, \forall \theta \in [k], \forall |t| \le \epsilon^*, \quad d_\theta t^{-1}\delta_x(t) \le \frac{\delta^* \left|\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}\right|}{\sqrt{n}}, \tag{51}$$

$$\forall x \notin \mathfrak{I}, \quad \tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*\left(\boldsymbol{s}^* - \left[\delta_x\left(\epsilon^* \tfrac{d_\theta}{\|\boldsymbol{d}\|}\right)\right]_{\theta\in[\tilde{k}]}\right) \le \epsilon^* \tfrac{\Delta g}{4\|\boldsymbol{d}\|} \sum_{x'\in\mathfrak{K}} p_{x'}\mathsf{D}\tilde{\underline{\ell}}_{x'}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \tag{52}$$

$$\begin{aligned}\forall x \in [n], \quad \tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*\left(\boldsymbol{s}^* - \epsilon^* \tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}\boldsymbol{d}_x\right) &\le -(\tilde{\Phi}^*)'\left(\boldsymbol{s}^*; -\epsilon^* \tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}\boldsymbol{d}_x\right) \\ &\quad + \epsilon^* \tfrac{\Delta g}{4\|\boldsymbol{d}\|} \sum_{x'\in\mathfrak{K}} p_{x'}\mathsf{D}\tilde{\underline{\ell}}_{x'}(\tilde{\boldsymbol{p}})\boldsymbol{v},\end{aligned} \tag{53}$$

where (53) is satisfied for small enough ϵ^* because of (49) and the fact that

$$\tfrac{1}{\epsilon}\left(\tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*(\boldsymbol{s}^* - \epsilon\tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}\boldsymbol{d}_x\right) \underset{\epsilon\to 0}{\to} -(\tilde{\Phi}^*)'\left(\boldsymbol{s}^*; -\tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}\boldsymbol{d}_x\right),$$

and (52) is also satisfied for small enough ϵ^* because $\tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*\left(\boldsymbol{s}^* - \left[\delta_x\left(\epsilon\tfrac{d_\theta}{\|\boldsymbol{d}\|}\right)\right]_{\theta\in[\tilde{k}]}\right) = O\left(\max_{\{\theta\in[\tilde{k}]\}}\left|\delta_x\left(\epsilon\tfrac{d_\theta}{\|\boldsymbol{d}\|}\right)\right|\right) = o(\epsilon)$, where the first equality is due to the fact that $(\lambda, \boldsymbol{z}) \mapsto \frac{1}{\lambda}\left(\tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*(\boldsymbol{s}^* - \lambda\boldsymbol{z}\right)$ is uniformly bounded on compact subsets of $\mathbb{R}\times\mathbb{R}^{\tilde{k}}$ (by continuity of the directional derivative $(\tilde{\Phi}^*)'(\boldsymbol{s}^*;\cdot)$).

If $\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} \le 0$, then by the positive homogeneity of the directional derivative, the definition of the function g, and (53), we get

$$\tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*\left(\boldsymbol{s}^* - \epsilon^*\tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}\boldsymbol{d}_x\right) \le \epsilon^*\tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}g(1) + \epsilon^*\tfrac{\Delta g}{4\|\boldsymbol{d}\|}\sum_{x'\in\mathfrak{K}} p_{x'}\mathsf{D}\tilde{\underline{\ell}}_{x'}(\tilde{\boldsymbol{p}})\boldsymbol{v}. \tag{54}$$

On the other hand, if $\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} > 0$, then from the monotonicity of the slopes of $\tilde{\Phi}^*$, the positive homogeneity of the directional derivative, and the definition of the function g, it follows that

$$\begin{aligned}\tfrac{1}{\epsilon^*}\left(\tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*(\boldsymbol{s}^* - \epsilon^*\tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}\boldsymbol{d}_x\right) &\le -(\tilde{\Phi}^*)'\left(\boldsymbol{s}^*; -\tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}\boldsymbol{d}_x\right), \\ &= -\tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}(\tilde{\Phi}^*)'\left(\boldsymbol{s}^*; -\boldsymbol{d}_x\right), \\ &\le \tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}g(-1), \\ &= \tfrac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|}\left(-\Delta g + g(1)\right).\end{aligned} \tag{55}$$

Let $\lambda_\theta := \epsilon^*\frac{d_\theta}{\|\boldsymbol{d}\|}$, for $\theta \in [\tilde{k}]$. From Theorem 5, there exists $[\boldsymbol{a}_\theta]_{\theta\in[k]} \in \mathcal{A}^k$, such that

$$\ell(\boldsymbol{a}_k) = \underline{\ell}(\boldsymbol{p}) \quad \text{and} \quad \forall\theta \in [\tilde{k}],\ \ell(\boldsymbol{a}_\theta) = \underline{\ell}(\boldsymbol{p}^{\lambda_\theta}) = \underline{\ell}(\boldsymbol{p}) + \epsilon^*\tfrac{d_\theta}{\|\boldsymbol{d}\|}\mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})\boldsymbol{v} + \delta\left(\epsilon^*\tfrac{d_\theta}{\|\boldsymbol{d}\|}\right), \tag{56}$$

where $[\delta(\cdot)]_x := \delta_x(\cdot)$ for $x \in [n]$.

From the fact that ℓ is Φ-mixable, it follows that there exists $\boldsymbol{a}_* \in \mathcal{A}$ such that for all $x \in [n]$,

$$\ell_x(\boldsymbol{a}_*) \le \mathrm{Mix}_\Phi(\ell_x(\boldsymbol{a}_{1:k}), \boldsymbol{\mu}) = \ell_x(\boldsymbol{a}_k) + \tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*\left(\boldsymbol{s}^* - J_k^\mathsf{T}\ell_x(\boldsymbol{a}_{1:k})\right). \tag{57}$$

For $x \in [n]$, we now define $\boldsymbol{d}_x \in \mathbb{R}^{\tilde{k}}$ explicitly as

$$\forall\theta \in [\tilde{k}], \quad d_{x,\theta} := \begin{cases} d_\theta + \frac{\|\boldsymbol{d}\|}{\epsilon^*[\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}]}\delta_x\left(\epsilon^*\frac{d_\theta}{\|\boldsymbol{d}\|}\right), & \text{if } x \in \mathfrak{I} \\ d_\theta, & \text{otherwise.}\end{cases}$$

From (51), we have $\|\boldsymbol{d}_x - \boldsymbol{d}\| \leq \delta^*, \forall x \in [n]$. Furthermore, from (56) and the fact that for all $x \in [n]$, $J_k^\mathsf{T} \ell_x(\boldsymbol{a}_{1:k}) = [\ell_x(\boldsymbol{a}_\theta) - \ell_x(\boldsymbol{a}_k)]_{\theta \in [\tilde{k}]}$, we have

$$J_k^\mathsf{T} \ell_x(\boldsymbol{a}_{1:k}) = \begin{cases} \epsilon^* \frac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|} \boldsymbol{d}_x, & \text{if } x \in \mathfrak{I}; \\ \left[\delta_x \left(\epsilon^* \frac{d_\theta}{\|\boldsymbol{d}\|}\right)\right]_{\theta \in [\tilde{k}]}, & \text{otherwise.} \end{cases} \tag{58}$$

Using this, together with (54) and (55), we get $\forall x \in \mathfrak{I}$,

$$\begin{aligned} \tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*\left(\boldsymbol{s}^* - J_k^\mathsf{T} \ell_x(\boldsymbol{a}_{1:k})\right) &= \tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*\left(\boldsymbol{s}^* - \epsilon^* \frac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|} \boldsymbol{d}_x\right), \\ &\leq \epsilon^* \frac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|} g(1) - \epsilon^* \Delta g \frac{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\|\boldsymbol{d}\|} \mathbb{1}_{\{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} > 0\}} \\ &\quad + \epsilon^* \frac{\Delta g}{4\|\boldsymbol{d}\|} \mathbb{1}_{\{\mathsf{D}\tilde{\underline{\ell}}_x(\tilde{\boldsymbol{p}})\boldsymbol{v} \leq 0\}} \sum_{x' \in \mathfrak{K}} p_{x'} \mathsf{D}\tilde{\underline{\ell}}_{x'}(\tilde{\boldsymbol{p}})\boldsymbol{v}. \end{aligned} \tag{59}$$

Combining (57), (58), and (59) yields

$$\begin{aligned} \langle \boldsymbol{p}, \ell(\boldsymbol{a}_*)\rangle \leq \langle \boldsymbol{p}, \ell(\boldsymbol{a}_k)\rangle &+ \tfrac{\epsilon^*}{\|\boldsymbol{d}\|} \langle \boldsymbol{p}, \mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})\boldsymbol{v}\rangle g(1) - \tfrac{3\epsilon^* \Delta g}{4\|\boldsymbol{d}\|} \sum_{x' \in \mathfrak{K}} p_{x'} \mathsf{D}\tilde{\underline{\ell}}_{x'}(\tilde{\boldsymbol{p}})\boldsymbol{v} \\ &+ \sum_{x \notin \mathfrak{I}} p_x \left(\tilde{\Phi}^*(\boldsymbol{s}^*) - \tilde{\Phi}^*\left(\boldsymbol{s}^* - \left[\delta_x\left(\epsilon^* \tfrac{d_\theta}{\|\boldsymbol{d}\|}\right)\right]_{\theta \in [\tilde{k}]}\right)\right), \end{aligned}$$

using (52) and the fact that $\langle \boldsymbol{p}, \mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})\rangle = \boldsymbol{0}_{\tilde{n}}^\mathsf{T}$ (see Lemma 27), we get

$$\leq \langle \boldsymbol{p}, \ell(\boldsymbol{a}_k)\rangle - \tfrac{\epsilon^* \Delta g}{2\|\boldsymbol{d}\|} \sum_{x' \in \mathfrak{K}} p_{x'} \mathsf{D}\tilde{\underline{\ell}}_{x'}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \tag{60}$$

$$< \langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle, \tag{61}$$

where in (61) we used (49) and the fact that $\underline{\ell}(\boldsymbol{p}) = \ell(\boldsymbol{a}_k)$ (see (57)). Equation 61 shows that $\ell(\boldsymbol{a}^*) \notin \mathscr{S}_\ell$, which is a contradiction. □

C.3 Proof of Theorem 7

Theorem 7 *Let $\eta > 0$, and let $\ell \colon \mathcal{A} \to [0, +\infty]^n$ a loss. Suppose that $\operatorname{dom} \ell = \mathcal{A}$ and that $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$. If $\underline{\eta_\ell} > 0$ then ℓ is $\underline{\eta_\ell}$-mixable. In particular, $\eta_\ell \geq \underline{\eta_\ell}$.*

Proof. Let $\eta \coloneqq \underline{\eta_\ell}$. We will show that $\exp(-\eta \mathscr{S}_\ell)$ is convex, which will imply that ℓ is η-mixable [6].

Since $\underline{\eta_\ell} = \inf_{\tilde{\boldsymbol{p}} \in \operatorname{int} \tilde{\Delta}_n} (\lambda_{\max}([\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1} \mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})))^{-1} > 0$, $\eta \underline{L}_\ell - \underline{L}_{\log}$ is convex on $\operatorname{ri} \Delta_n$ [19, Thm. 10]. Let $\boldsymbol{p} \in \operatorname{ri} \Delta_n$ and define

$$\Lambda(\boldsymbol{r}) \coloneqq \underline{L}_{\log}(\boldsymbol{r}) + \langle \boldsymbol{r}, \eta \underline{\ell}(\boldsymbol{p}) - \ell_{\log}(\boldsymbol{p})\rangle, \ \boldsymbol{r} \in \operatorname{ri} \Delta_n.$$

Since Λ is equal to $\underline{L}_{\log}$ plus an affine function, it follows that $\eta \underline{L}_\ell - \Lambda$ is also convex on $\operatorname{ri} \Delta_n$. On the one hand, since $\underline{\ell}$ and $\ell_{\log}$ are proper losses, we have $\langle \boldsymbol{p}, \underline{\ell}(\boldsymbol{p})\rangle = \underline{L}_\ell(\boldsymbol{p})$ and $\langle \boldsymbol{p}, \ell_{\log}(\boldsymbol{p})\rangle = \underline{L}_{\log}(\boldsymbol{p})$ which implies that

$$\eta \underline{L}_\ell(\boldsymbol{p}) - \Lambda(\boldsymbol{p}) = 0. \tag{62}$$

On the other hand, since $\underline{L}_\ell$ and $\underline{L}_{\log}$ are differentiable we have $\underline{\ell}(\boldsymbol{p}) = \nabla \underline{L}_\ell(\boldsymbol{p})$ and $\nabla \underline{L}_{\log}(\boldsymbol{p}) = \ell_{\log}(\boldsymbol{p})$, which yields $\eta \nabla \underline{L}_\ell(\boldsymbol{p}) - \nabla \Lambda(\boldsymbol{p}) = \boldsymbol{0}_n$. This implies that $\eta \underline{L}_\ell - \Lambda$ attains a minimum at $\boldsymbol{p}$ [8, Thm. D.2.2.1]. Combining this fact with (62) gives $\eta \underline{L}_\ell(\boldsymbol{r}) \geq \Lambda(\boldsymbol{r}), \forall \boldsymbol{r} \in \operatorname{ri} \Delta_n$, or equivalently $-\eta \underline{L}_\ell \leq -\Lambda$. By Proposition 20-(iii), this implies

$$[-\eta \underline{L}_\ell]^* \geq [-\Lambda]^*. \tag{63}$$

Using Proposition 20-(ii), we get $[-\Lambda]^*(\boldsymbol{s}) = [-\underline{L}_{\log}]^*(\boldsymbol{s} - \ell_{\log}(\boldsymbol{p}) + \eta \underline{\ell}(\boldsymbol{p}))$ for $\boldsymbol{s} \in \mathbb{R}^n$. Since $-\eta \underline{L}_\ell(\boldsymbol{u}) = -\underline{L}_\ell(\eta \boldsymbol{u}) = \sigma_{\mathscr{S}_\ell}(-\eta \boldsymbol{u})$ and $\sigma_{\mathscr{S}_\ell}^* = \iota_{\mathscr{S}_\ell}$, Proposition 20-(v) implies $[-\eta \underline{L}_\ell]^*(\boldsymbol{s}) = \iota_{\mathscr{S}_\ell}(-\boldsymbol{s}/\eta)$. Similarly, we have $[-\underline{L}_{\log}]^*(\boldsymbol{s}) = \iota_{\mathscr{S}_{\log}}(-\boldsymbol{s})$. Therefore, (63) implies

$$\forall \boldsymbol{s} \in \mathbb{R}^n, \quad \iota_{\mathscr{S}_\ell}(-\boldsymbol{s}/\eta) \geq \iota_{\mathscr{S}_{\log}}(-\boldsymbol{s} + \ell_{\log}(\boldsymbol{p}) - \eta \underline{\ell}(\boldsymbol{p})).$$

This inequality implies that if $\boldsymbol{s} \in -\eta\mathscr{S}_\ell$, then $\boldsymbol{s} \in -\mathscr{S}_{\log} + \ell_{\log}(\boldsymbol{p}) - \eta\underline{\ell}(\boldsymbol{p})$. In particular, if $\boldsymbol{u} \in e^{-\eta\mathscr{S}_\ell}$ then

$$\boldsymbol{u} \in e^{-\mathscr{S}_{\log}+\ell_{\log}(\boldsymbol{p})-\eta\underline{\ell}(\boldsymbol{p})} \subseteq \mathcal{H}_{\tau(\boldsymbol{p}),1} = \{\boldsymbol{v} \in \mathbb{R}^n : \langle \boldsymbol{v}, \boldsymbol{p} \odot e^{\eta\underline{\ell}(\boldsymbol{p})}\rangle \leq 1\}. \tag{64}$$

To see the set inclusion in (64), consider $\boldsymbol{s} \in -\mathscr{S}_{\log} + \ell_{\log}(\boldsymbol{p}) - \eta\underline{\ell}(\boldsymbol{p})$, then by definition of the superprediction set $\mathscr{S}_{\log}$ there exists $\boldsymbol{r} \in \Delta_n$ and $\boldsymbol{v} \in [0, +\infty[^n$, such that $\boldsymbol{s} = \log \boldsymbol{r} - \log \boldsymbol{p} - \eta\underline{\ell}(\boldsymbol{p}) - \boldsymbol{v}$. Thus,

$$\langle e^{\boldsymbol{s}}, \boldsymbol{p} \odot e^{\eta\underline{\ell}(\boldsymbol{p})}\rangle = \langle \boldsymbol{r}, e^{-\boldsymbol{v}}\rangle \leq 1, \tag{65}$$

where the inequality is true because $\boldsymbol{r} \in \Delta_n$ and $\boldsymbol{v} \in [0, +\infty[^n$. The above argument shows that $e^{-\eta\mathscr{S}_\ell} \subseteq \mathcal{H}_{\tau(\boldsymbol{p}),1}$, where $\tau(\boldsymbol{p}) \coloneqq \boldsymbol{p} \odot e^{\eta\underline{\ell}(\boldsymbol{p})}$. Furthermore, $e^{-\eta\mathscr{S}_\ell} \subseteq \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap]0, +\infty[^n$, since all elements of $e^{-\eta\mathscr{S}_\ell}$ have non-negative, finite components. The latter set inclusion still holds for $\hat{\boldsymbol{p}} \in \operatorname{rbd} \Delta_n$. In fact, from the definition of a support loss, there exists a sequence $(\boldsymbol{p}_m)$ in $\operatorname{ri} \Delta_n$ converging to $\hat{\boldsymbol{p}}$ such that $\underline{\ell}(\boldsymbol{p}_m) \overset{m\to\infty}{\to} \underline{\ell}(\hat{\boldsymbol{p}})$. Equation 65 implies that for $\boldsymbol{u} \in e^{-\eta\mathscr{S}_\ell}$, $\langle \boldsymbol{u}, \boldsymbol{p}_m \odot e^{\eta\underline{\ell}(\boldsymbol{p}_m)}\rangle \leq 1$. Since the inner product is continuous, by passage to the limit, we obtain $\langle \boldsymbol{u}, \hat{\boldsymbol{p}} \odot e^{\eta\underline{\ell}(\hat{\boldsymbol{p}})}\rangle \leq 1$. Therefore,

$$e^{-\eta\mathscr{S}_\ell} \subseteq \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap]0, +\infty[^n. \tag{66}$$

Now suppose $\boldsymbol{u} \in \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap]0, +\infty[^n$; that is, for all $\boldsymbol{p} \in \Delta_n$,

$$\begin{aligned} 1 \geq \left\langle \boldsymbol{u}, \boldsymbol{p} \odot e^{\eta\underline{\ell}(\boldsymbol{p})}\right\rangle = \left\langle \boldsymbol{p}, \boldsymbol{u} \odot e^{\eta\underline{\ell}(\boldsymbol{p})}\right\rangle &= \left\langle \boldsymbol{p}, e^{\eta\underline{\ell}(\boldsymbol{p})+\log \boldsymbol{u}}\right\rangle, \\ &\geq e^{\langle \boldsymbol{p}, \eta\underline{\ell}(\boldsymbol{p})\rangle + \langle \boldsymbol{p}, \log \boldsymbol{u}\rangle}, \end{aligned} \tag{67}$$

where the first equality is obtained merely by expanding the expression of the inner product, and the second inequality is simply Jensen's Inequality. Since $\boldsymbol{u} \mapsto e^{\boldsymbol{u}}$ is strictly convex, the Jensen's inequality in (67) is strict unless $\exists (c, \boldsymbol{p}) \in \mathbb{R} \times \Delta_n$, such that

$$\eta\underline{\ell}(\boldsymbol{p}) + \log \boldsymbol{u} = c\mathbf{1}_n. \tag{68}$$

By substituting (68) into (67), we get $1 \geq \exp(c)$, and thus $c \leq 0$. Furthermore, (68) together with the fact that $\boldsymbol{u} \in]0, +\infty[^n$ imply that $\boldsymbol{p} \in \operatorname{dom} \underline{\ell}$, and thus there exists $\boldsymbol{a} \in \operatorname{dom} \ell$ such that $\ell(\boldsymbol{a}) = \underline{\ell}(\boldsymbol{p})$ (Theorem 5). Using this and rearranging (68), we get $\boldsymbol{u} = \exp(-\eta\ell(\boldsymbol{a}) + c\mathbf{1})$. Since $c \leq 0$, this means that $\boldsymbol{u} \in \exp(-\eta\mathscr{S}_\ell)$. Suppose now that (68) does not hold. In this case, (67) must be a strict inequality for all $\boldsymbol{p} \in \Delta_n$. By applying the $\log$ on both side of (67),

$$\forall \boldsymbol{p} \in \Delta_n, \eta\underline{L}_\ell(\boldsymbol{p}) + \langle \boldsymbol{p}, \log \boldsymbol{u}\rangle = \langle \boldsymbol{p}, \eta\underline{\ell}(\boldsymbol{p})\rangle + \langle \boldsymbol{p}, \log \boldsymbol{u}\rangle < 0. \tag{69}$$

Since $\boldsymbol{p} \mapsto \underline{L}_\ell(\boldsymbol{p}) = -\sigma_{\mathscr{S}_\ell}(-\boldsymbol{p})$ is a closed concave function, the map $g\colon \boldsymbol{p} \mapsto \eta\underline{L}_\ell(\boldsymbol{p}) + \langle \boldsymbol{p}, \log \boldsymbol{u}\rangle$ is also closed and concave, and thus upper semi-continuous. Since Δ_n is compact, the function g must attain its maximum in Δ_n. Due to (69) this maximum is negative; there exists $c_1 > 0$ such that

$$\forall \boldsymbol{p} \in \Delta_k, \langle \boldsymbol{p}, \eta\underline{\ell}(\boldsymbol{p})\rangle - \langle \boldsymbol{p}, -\log \boldsymbol{u}\rangle \leq -c_1. \tag{70}$$

Let $f(\boldsymbol{p}, x) \coloneqq \eta\underline{\ell}_x(\boldsymbol{p}) + \log u_x + c_1$, for $x \in [n]$. It follows from (70) that for all $\boldsymbol{p} \in \Delta_n$, $\mathbb{E}_{x\sim\boldsymbol{p}} f(\boldsymbol{p}, x) \leq 0$ and $\forall x \in [n], -\infty < f(\boldsymbol{p}, x)$. Thus, Lemma 25 applied to f with $\epsilon = c_1/2$, implies that there exists $\boldsymbol{p}_* \in \operatorname{ri} \Delta_n$, such that $\eta\underline{\ell}(\boldsymbol{p}_*) \leq -\log \boldsymbol{u} - c_1/2 \leq -\log \boldsymbol{u}$. From this inequality, $\boldsymbol{p}_* \in \operatorname{dom} \underline{\ell}$, and therefore, there exists $\boldsymbol{a}_* \in \operatorname{dom} \ell$ such that $\ell(\boldsymbol{a}_*) = \underline{\ell}(\boldsymbol{p}_*)$ (Theorem 5). This shows that $\eta\ell(\boldsymbol{a}_*) \leq -\log \boldsymbol{u}$, which implies that $\boldsymbol{u} \in \exp -\eta\mathscr{S}_\ell$. Therefore, $\bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap]0, +\infty[^n \subseteq e^{-\eta\mathscr{S}_\ell}$. Combining this with (66) shows that $e^{-\eta\mathscr{S}_\ell} = \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap]0, +\infty[^n$. Since $e^{-\eta\mathscr{S}_\ell}$ is the intersection of convex set, it is a itself convex set. Since $\operatorname{dom} \ell = \mathcal{A}$ by assumption, it follows that $\mathscr{S}_\ell = \mathscr{S}_\ell^\infty$, and thus $e^{-\eta\mathscr{S}_\ell^\infty}$ is convex. This last fact implies that ℓ is η-mixable [6]. □

C.4 Proof of Theorem 10

We start by the following characterization of Δ-differentiability (this was defined on page 5 of the main body of the paper).

Lemma 38. *Let $\Phi\colon \mathbb{R}^k \to \mathbb{R}\cup\{+\infty\}$ be an entropy. Then Φ is Δ-differentiable if and only if $\forall \mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$, $\tilde{\Phi}_{\mathfrak{l}} := \Phi \circ \Pi_k \circ [\Pi_{\mathfrak{l}}^{\tilde{k}}]^\top$ is differentiable on* $\operatorname{int} \tilde{\Delta}_{|\mathfrak{l}|}$.

Proof. This is a direct consequence of Proposition B.4.2.1 in [8], since 1) $\tilde{\Phi}_{\mathfrak{l}}$ is convex; and 2)

$$\begin{aligned}\tilde{\Phi}'_{\mathfrak{l}}(\tilde{\boldsymbol{u}};\tilde{\boldsymbol{v}}-\tilde{\boldsymbol{u}}) &= \tilde{\Phi}'([\Pi_{\mathfrak{l}}^{\tilde{k}}]^\top\tilde{\boldsymbol{u}};[\Pi_{\mathfrak{l}}^{\tilde{k}}]^\top(\boldsymbol{v}-\boldsymbol{u})),\\ &= \Phi'(\Pi_k[\Pi_{\mathfrak{l}}^{\tilde{k}}]^\top\tilde{\boldsymbol{u}};\Pi_k[\Pi_{\mathfrak{l}}^{\tilde{k}}]^\top(\boldsymbol{v}-\boldsymbol{u})),\end{aligned}$$

for all $\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}} \in \operatorname{int}\tilde{\Delta}_{|\mathfrak{l}|}$ and $\tilde{\Phi} := \Phi\circ\Pi_k$. □

Theorem 10 *Let $\Phi : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ be a Δ-differentiable entropy. Let $\ell : \mathcal{A} \to [0,+\infty]^n$ be a loss (not necessarily finite) such that $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$. If ℓ is (η,Φ)-mixable then the GAA achieves a constant regret in the $\mathfrak{G}_\ell^n(\mathcal{A},k)$ game; for any sequence $(x^t,\boldsymbol{a}_{1:k}^t)_{t=1}^T$,*

$$\operatorname{Loss}_{\mathrm{GAA}}^\ell(T) - \min_{\theta\in[k]}\operatorname{Loss}_\theta^\ell(T) \le R_\ell^\Phi := \inf_{\boldsymbol{q}\in\Delta_k}\max_{\theta\in[k]} D_\Phi(\boldsymbol{e}_\theta,\boldsymbol{q})/\eta_\ell^\Phi,$$

where $\boldsymbol{e}_\theta$ is the θth basis element of $\mathbb{R}^k$.

Proof. For all $\mathfrak{l} \subseteq [k]$ such that $|\mathfrak{l}| > 1$, let $\tilde{\Phi} := \Phi\circ\Pi_k$ and $\tilde{\Phi}_{\mathfrak{l}} := \tilde{\Phi}\circ[\Pi_{\mathfrak{l}}^{\tilde{k}}]^\top$. From Lemma 33 the infimum involved in the definition of the expert distribution $\boldsymbol{q}^t$ in Algorithm 2 is indeed attained. It remains to verify that this minimum is unique. This will become clear in what follows.

Let $\mathfrak{l}^0 = [k]$ and $\mathfrak{I}^t := \{\theta\in[k] : \ell_{x^t}(\boldsymbol{a}_\theta^t) < +\infty\}, t\in[T]$. For $t\in[T]$, we define the non-increasing sequence of subsets $(\mathfrak{l}^t)$ of $[k]$ defined by $\mathfrak{l}^t := \mathfrak{I}^t\cap\mathfrak{l}^{t-1}$. We show by induction that $\boldsymbol{q}^t\in\Delta_{\mathfrak{l}^t}$ and

$$\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\Pi_{\mathfrak{l}^t}^{\tilde{k}}\tilde{\boldsymbol{q}}^t) = \Pi_{\mathfrak{l}^t}^{\tilde{k}}\left(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0) - \sum_{s=1}^t J_k^\top\ell_{x^s}(A^s)\right), \tag{71}$$

where $A^s := [\boldsymbol{a}_\theta^s]\in\mathcal{A}^k$, $s\in\mathbb{N}$. Suppose that (71) holds true up to some $t\ge 1$. We will now show that it holds for $t+1$. To simplify expressions, we denote $\tilde{\boldsymbol{x}}_{\mathfrak{l}} := \Pi_{\mathfrak{l}}^{\tilde{k}}\tilde{\boldsymbol{x}}\in\mathbb{R}^{\mathfrak{l}}$ for $\tilde{\boldsymbol{x}}\in\mathbb{R}^{\tilde{k}}$, and $\boldsymbol{z}^t := \ell_{x^t}(A^t), t\in[T]$. From the definition of $\boldsymbol{q}^t$ in Algorithm 2, we have

$$\boldsymbol{q}^{t+1}\in\mathcal{M} := \operatorname*{Argmin}_{\boldsymbol{\mu}\in\Delta_k}\langle\boldsymbol{\mu},\boldsymbol{z}^{t+1}\rangle + D_\Phi(\boldsymbol{\mu},\boldsymbol{q}^t).$$

Using the definition of $\mathfrak{I}^{t+1}$,

$$\begin{aligned}\mathcal{M} &= \operatorname*{Argmin}_{\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}}}\langle\boldsymbol{\mu},\boldsymbol{z}^{t+1}\rangle + D_\Phi(\boldsymbol{\mu},\boldsymbol{q}^t),\\ &= \operatorname*{Argmin}_{\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}}}\langle\boldsymbol{\mu},\boldsymbol{z}^{t+1}\rangle + \tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^t}) - \tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}}^t) - \tilde{\Phi}'_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t;\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^t}-\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t).\end{aligned}$$

Now using the facts that $\boldsymbol{q}^t\in\Delta_{\mathfrak{l}^t}$, $\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}}\subseteq\Delta_{\mathfrak{l}^t}$, Φ is Δ-differentiable, and Lemma 38, we have

$$\mathcal{M} = \operatorname*{Argmin}_{\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}}}\langle\boldsymbol{\mu},\boldsymbol{z}^{t+1}\rangle + \tilde{\Phi}_{\mathfrak{l}^{t+1}}(\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}}) - \tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t) - \langle\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^t}-\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t,\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t)\rangle.$$

Using the facts that $\langle\boldsymbol{\mu},\boldsymbol{z}^{t+1}\rangle = z_k^{t+1} + \langle\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}},\Pi_{\mathfrak{l}^{t+1}}^{\tilde{k}}J_k^\top\boldsymbol{z}^{t+1}\rangle$, for $\tilde{\boldsymbol{\mu}}\in\tilde{\Delta}_{\mathfrak{l}^{t+1}}$, and $\langle\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^t},\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t)\rangle\mathcal{M} = \langle\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}},\Pi_{\mathfrak{l}^{t+1}}^{\mathfrak{l}^t}\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t)\rangle$ (since $\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}}$)

$$\begin{aligned}\mathcal{M} = \operatorname*{Argmin}_{\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}}}\langle\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}},-\Pi_{\mathfrak{l}^{t+1}}^{\mathfrak{l}^t}\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t) + \Pi_{\mathfrak{l}^{t+1}}^{\tilde{k}}J_k^\top\boldsymbol{z}^{t+1}\rangle + \tilde{\Phi}_{\mathfrak{l}^{t+1}}(\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}})\\ + \langle\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t,\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t)\rangle - \tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t),\end{aligned}$$

and since the last two terms are independent of $\boldsymbol{\mu}$,

$$\mathcal{M} = \operatorname*{Argmin}_{\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}}}\langle\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}},-\Pi_{\mathfrak{l}^{t+1}}^{\mathfrak{l}^t}\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t) + \Pi_{\mathfrak{l}^{t+1}}^{\tilde{k}}J_k^\top\boldsymbol{z}^{t+1}\rangle + \tilde{\Phi}_{\mathfrak{l}^{t+1}}(\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}}).$$

Now using Fenchel duality property in Proposition 20-(iv),

$$\mathcal{M} = \{\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}} : \Pi_{\mathfrak{l}^{t+1}}^{\tilde{k}}\circ\Pi_k(\boldsymbol{\mu}) = \tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}}\in\partial\tilde{\Phi}_{\mathfrak{l}^{t+1}}^*(\Pi_{\mathfrak{l}^{t+1}}^{\mathfrak{l}^t}\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}_{\mathfrak{l}^t}^t) - \Pi_{\mathfrak{l}^{t+1}}^{\tilde{k}}J_k^\top\boldsymbol{z}^{t+1})\}.$$

Finally, due to Lemma 29 and Proposition 12, $\tilde{\Phi}^*_{\mathfrak{l}^{t+1}}$ is differentiable on $\mathbb{R}^{|\mathfrak{l}^{t+1}|-1}$, and thus

$$\mathcal{M} = \{\Pi_k \circ [\Pi^{\tilde{k}}_{\mathfrak{l}^{t+1}}]^\mathsf{T} \circ \nabla\tilde{\Phi}^*_{\mathfrak{l}^{t+1}}(\Pi^{\mathfrak{l}^t}_{\mathfrak{l}^{t+1}}\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}^t_{\mathfrak{l}^t}) - \Pi^{\tilde{k}}_{\mathfrak{l}^{t+1}}J_k^\mathsf{T}\boldsymbol{z}^{t+1})\}. \tag{72}$$

From (72), we obtain

$$\nabla\tilde{\Phi}_{\mathfrak{l}^{t+1}}(\Pi^{\tilde{k}}_{\mathfrak{l}^{t+1}}\tilde{\boldsymbol{q}}^{t+1}) = \Pi^{\mathfrak{l}^t}_{\mathfrak{l}^{t+1}}\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}^t_{\mathfrak{l}^t}) - \Pi^{\tilde{k}}_{\mathfrak{l}^{t+1}}J_k^\mathsf{T}\boldsymbol{z}^{t+1}. \tag{73}$$

Thus using the induction assumption and the fact that $\Pi^{\mathfrak{l}^t}_{\mathfrak{l}^{t+1}}\Pi^{\tilde{k}}_{\mathfrak{l}^t} = \Pi^{\tilde{k}}_{\mathfrak{l}^{t+1}}$ (since $\mathfrak{l}^{t+1} \subseteq \mathfrak{l}^t$), the result follows, i.e. (71) is true for all $t \in [T]$. Furthermore, $\boldsymbol{q}^{t+1} \in \Delta_{\mathfrak{l}^{t+1}}$, since $\Pi^{\tilde{k}}_{\mathfrak{l}^{t+1}}\tilde{\boldsymbol{q}}^{t+1} \in \operatorname{dom}\tilde{\Phi}_{\mathfrak{l}^{t+1}} \subseteq \tilde{\Delta}_{|\mathfrak{l}^{t+1}|}$. Using the same arguments as above, one arrives at

$$\begin{aligned}\operatorname{Mix}_\Phi(\boldsymbol{q}^t, \boldsymbol{z}^{t+1}) = z_k^{t+1} + \inf_{\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^{t+1}}} &\langle\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}}, -\Pi^{\mathfrak{l}^t}_{\mathfrak{l}^{t+1}}\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}^t_{\mathfrak{l}^t}) + \Pi^{\tilde{k}}_{\mathfrak{l}^{t+1}}J_k^\mathsf{T}\boldsymbol{z}^{t+1}\rangle + \tilde{\Phi}_{\mathfrak{l}^{t+1}}(\tilde{\boldsymbol{\mu}}_{\mathfrak{l}^{t+1}}) \\ &+ \langle\tilde{\boldsymbol{q}}^t_{\mathfrak{l}^t}, \nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}^t_{\mathfrak{l}^t})\rangle - \tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}^t_{\mathfrak{l}^t}).\end{aligned}$$

Using the Fenchel duality property Proposition 20-(vi) and (72),

$$= z_k^{t+1} + \tilde{\Phi}^*_{\mathfrak{l}^t}(\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}^t_{\mathfrak{l}^t})) - \tilde{\Phi}^*_{\mathfrak{l}^{t+1}}(\Pi^{\mathfrak{l}^t}_{\mathfrak{l}^{t+1}}\nabla\tilde{\Phi}_{\mathfrak{l}^t}(\tilde{\boldsymbol{q}}^t_{\mathfrak{l}^t}) - \Pi^{\tilde{k}}_{\mathfrak{l}^{t+1}}J_k^\mathsf{T}\boldsymbol{z}^{t+1}). \tag{74}$$

On the other hand, Φ-mixability implies that there exists $\boldsymbol{a}^t_* \in \mathcal{A}^t$, such that for all $x^t \in [n]$,

$\forall t \in [T], \ell_{x^t}(\boldsymbol{a}^t_*) \leq \operatorname{Mix}_\Phi(\boldsymbol{q}^{t-1}, \boldsymbol{z}^t)$,

Summing this inequality for $t = 1, \ldots, T$ yields,

$$\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_*) \leq \sum_{t=1}^T \operatorname{Mix}_\Phi(\boldsymbol{q}^{t-1}, \boldsymbol{z}^t),$$

and thus using (74) and (73) yields

$$\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_*) \leq \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_k) + \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0)) - \tilde{\Phi}^*_{\mathfrak{l}^T}(\Pi^{\mathfrak{l}^{T-1}}_{\mathfrak{l}^T}\nabla\tilde{\Phi}_{\mathfrak{l}^{T-1}}(\tilde{\boldsymbol{q}}^{T-1}_{\mathfrak{l}^{T-1}}) - \Pi^{\tilde{k}}_{\mathfrak{l}^T}J_k^\mathsf{T}\boldsymbol{z}^T).$$

Finally, using (71) together with the fact that $\Pi^{\mathfrak{l}^{T-1}}_{\mathfrak{l}^T}\Pi^{\tilde{k}}_{\mathfrak{l}^{T-1}} = \Pi^{\tilde{k}}_{\mathfrak{l}^T}$

$$\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_*) \leq \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_k) + \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0)) - \tilde{\Phi}^*_{\mathfrak{l}^T}\left(\Pi^{\tilde{k}}_{\mathfrak{l}^T}\left(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0) - \sum_{t=1}^T J_k^\mathsf{T}\ell_{x^t}(A^t)\right)\right).$$

Using the definition of the Fenchel dual and Proposition 20-(vi) again, the above inequality becomes

$$\begin{aligned}\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_*) \leq &\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_k) + \langle\tilde{\boldsymbol{q}}^0, \nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0))\rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}^0) \\ &- \sup_{\boldsymbol{\pi}\in\Delta_{|\mathfrak{l}^T|}}\left[\left\langle\tilde{\boldsymbol{\pi}}, \Pi^{\tilde{k}}_{\mathfrak{l}^T}\left(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0) - \sum_{t=1}^T J_k^\mathsf{T}\ell_{x^t}(A^t)\right)\right\rangle - \tilde{\Phi}_{\mathfrak{l}^T}(\tilde{\boldsymbol{\pi}})\right], \\ = &\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_k) + \langle\tilde{\boldsymbol{q}}^0, \nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0))\rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}^0) \\ &+ \inf_{\boldsymbol{\mu}\in\Delta_{\mathfrak{l}^T}}\left[\left\langle\tilde{\boldsymbol{\mu}}, \sum_{t=1}^T J_k^\mathsf{T}\ell_{x^t}(A^t) - \nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0)\right\rangle + \tilde{\Phi}(\tilde{\boldsymbol{\mu}})\right].\end{aligned} \tag{75}$$

Using the fact that $\forall\theta \in [k] \setminus \mathfrak{l}^T$, $\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_\theta) = +\infty$ (by definition of $(\mathfrak{l}^t)$), the right hand side of (75) becomes

$$\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_k) + \langle\tilde{\boldsymbol{q}}^0, \nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0))\rangle - \tilde{\Phi}(\tilde{\boldsymbol{q}}^0) + \inf_{\boldsymbol{\mu}\in\Delta_k}\left[\left\langle\tilde{\boldsymbol{\mu}}, \sum_{t=1}^T J_k^\mathsf{T}\ell_{x^t}(A^t) - \nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0)\right\rangle + \tilde{\Phi}(\tilde{\boldsymbol{\mu}})\right].$$

Thus, we get

$$\begin{aligned}\forall\boldsymbol{\mu}\in\Delta_k,\ \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_*) \leq \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_k) + &\left\langle\tilde{\boldsymbol{\mu}}, \sum_{t=1}^T J_k^\mathsf{T}\ell_{x^t}(A^t)\right\rangle \\ &+ \tilde{\Phi}(\tilde{\boldsymbol{\mu}}) - \tilde{\Phi}(\tilde{\boldsymbol{q}}^0) - \langle\tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{q}}^0, \nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}^0)\rangle.\end{aligned}$$

Using the facts that $\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}_k^t) + \left\langle \tilde{\boldsymbol{\mu}}, \sum_{t=1}^T J_k^\intercal \ell_{x^t}(A^t) \right\rangle = \left\langle \boldsymbol{\mu}, \sum_{t=1}^T \ell_{x^t}(A^t) \right\rangle$ and the definition of the divergence,

$$\forall \boldsymbol{\mu} \in \Delta_k,\ \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}_*^t) \leq \left\langle \boldsymbol{\mu}, \sum_{t=1}^T \ell_{x^t}(A^t) \right\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}^0),$$

which for $\boldsymbol{\mu} = \boldsymbol{e}_\theta$ implies

$$\forall \theta \in [k],\ \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}_*^t) \leq \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}_\theta^t) + D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q}^0). \tag{76}$$

When instead of Φ-mixability, we have (η, Φ)-mixability, the last term in (76) becomes $\frac{D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q}^0)}{\eta}$ and the desired result follows.

□

C.5 Proof of Theorem 11

We require the following result:

Proposition 39. *For the Shannon entropy* S*, it holds that* $\tilde{\mathrm{S}}^*(\boldsymbol{v}) = \log(\langle \exp(\boldsymbol{v}), \mathbf{1}_{\tilde{k}} \rangle + 1), \forall \boldsymbol{v} \in \mathbb{R}^{k-1}$*, and* $\mathrm{S}^\star(\boldsymbol{z}) = \log\langle \exp(\boldsymbol{z}), \mathbf{1}_k \rangle, \forall \boldsymbol{z} \in \mathbb{R}^k$.

Proof. Given $\boldsymbol{v} \in \mathbb{R}^{k-1}$, we first derive the expression of the Fenchel dual $\tilde{\mathrm{S}}^*(\boldsymbol{v}) := \sup_{\tilde{\boldsymbol{q}} \in \tilde{\Delta}_k} \langle \tilde{\boldsymbol{q}}, \boldsymbol{v} \rangle - \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})$. Setting the gradient of $\tilde{\boldsymbol{q}} \mapsto \langle \tilde{\boldsymbol{q}}, \boldsymbol{v} \rangle - \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})$ to $\mathbf{0}_{\tilde{k}}$ gives $\boldsymbol{v} = \nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})$. For $\boldsymbol{q} \in]0, +\infty[^k$, we have $\nabla \mathrm{S}(\boldsymbol{q}) = \log \boldsymbol{q} + \mathbf{1}_k$, and from appendix A we know that $\nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}) = J_k^\intercal \nabla \mathrm{S}(\boldsymbol{q})$. Therefore,

$$\boldsymbol{v} = \nabla \tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}) \implies \boldsymbol{v} = J_k^\intercal \nabla \mathrm{S}(\boldsymbol{q}) \implies \boldsymbol{v} = \log \frac{\tilde{\boldsymbol{q}}}{q_k},$$

where the right most equality is equivalent to $\tilde{\boldsymbol{q}}/q_k = \exp(\boldsymbol{v})$. Since $\langle \tilde{\boldsymbol{q}}, \mathbf{1}_{\tilde{k}} \rangle = 1 - q_k$, we get $q_k = (\langle \exp(\boldsymbol{v}), \mathbf{1}_{\tilde{k}} \rangle + 1)^{-1}$. Therefore, the supremum in the definition of $\tilde{\mathrm{S}}^*(\boldsymbol{v})$ is attained at $\tilde{\boldsymbol{q}}_* = \exp(\boldsymbol{v})(\langle \exp(\boldsymbol{v}), \mathbf{1}_{\tilde{k}} \rangle + 1)^{-1}$. Hence $\tilde{\mathrm{S}}^*(\boldsymbol{v}) = \langle \tilde{\boldsymbol{q}}_*, \boldsymbol{v} \rangle - \langle \tilde{\boldsymbol{q}}_*, \log \tilde{\boldsymbol{q}}_* \rangle = \log(\langle \exp(\boldsymbol{v}), \mathbf{1}_{\tilde{k}} \rangle + 1)$. Finally, using (14) we get $\mathrm{S}^\star(\boldsymbol{z}) = \log\langle \exp(\boldsymbol{z}), \mathbf{1}_k \rangle$, for $\boldsymbol{z} \in \mathbb{R}^k$. □

Theorem 11 *Let* $\eta > 0$*. A loss* $\ell \colon \mathcal{A} \to [0, +\infty]^n$ *is* η*-mixable if and only if* ℓ *is* (η, S)*-mixable.*

Proof.

Claim 1. *For all* $\boldsymbol{q} \in \Delta_k$*,* $A := \boldsymbol{a}_{1:k} \in \mathbb{R}^k$*, and* $x \in [n]$

$$-\eta^{-1} \log \langle \exp(-\eta \ell_x(A)), \boldsymbol{q} \rangle = \mathrm{Mix}_{\mathrm{S}}^\eta(\ell_x(A), \boldsymbol{q}). \tag{77}$$

Let $\boldsymbol{q} \in \operatorname{ri} \Delta_k$. From Proposition 39, the Shannon entropy is such that $\mathrm{S}^\star$ is differentiable on $\mathbb{R}^k$, and thus it follows from Lemma 33 ((33)-(34)) that for any $\boldsymbol{d} \in [0, +\infty[^k$

$$\mathrm{Mix}_{\mathrm{S}}(\boldsymbol{d}, \boldsymbol{q}) = \mathrm{S}^\star(\nabla \mathrm{S}(\boldsymbol{q})) - \mathrm{S}^\star(\nabla \mathrm{S}(\boldsymbol{q}) - \boldsymbol{d}). \tag{78}$$

By definition of S, $\nabla \mathrm{S}(\boldsymbol{q}) = \log \boldsymbol{q} + \mathbf{1}_k$, and due to Proposition 39, $\mathrm{S}^\star(\boldsymbol{z}) = \log\langle \exp \boldsymbol{z}, \mathbf{1}_k \rangle$, $\boldsymbol{z} \in \mathbb{R}^k$. Therefore,

$$\nabla \mathrm{S}(\boldsymbol{q}) - \eta \boldsymbol{d} = \log(\exp(-\eta \boldsymbol{d}) \odot \boldsymbol{q}) + \mathbf{1}_k. \tag{79}$$

On the other hand, from [13] we also have

$$\mathrm{Mix}_{\mathrm{S}}^\eta(\boldsymbol{d}, \boldsymbol{q}) = \eta^{-1} \mathrm{Mix}_{\mathrm{S}}(\eta \boldsymbol{d}, \boldsymbol{q}),\ \eta > 0. \tag{80}$$

Combining (78)-(80), yields

$$-\eta^{-1} \log \langle \exp(-\eta \boldsymbol{d}, \boldsymbol{q} \rangle = \mathrm{Mix}_{\mathrm{S}}^\eta(\boldsymbol{d}, \boldsymbol{q}). \tag{81}$$

Suppose now that $\boldsymbol{q} \in \operatorname{ri}\Delta_{\mathfrak{l}}$ *for* $\mathfrak{l} \subseteq [k]$ *such that* $|\mathfrak{l}| > 1$. By repeating the argument above for $\mathrm{S}_{\mathfrak{l}} := \mathrm{S} \circ \Pi_{\mathfrak{l}}^{\mathsf{T}}$, we get

$$\begin{aligned}\forall \boldsymbol{d} \in [0,+\infty[^n,\ \mathrm{Mix}^{\eta}_{\mathrm{S}_{\mathfrak{l}}}(\Pi_{\mathfrak{l}}\boldsymbol{d}, \Pi_{\mathfrak{l}}\boldsymbol{q}) &= -\eta^{-1}\log\langle \exp(-\eta\Pi_{\mathfrak{l}}\boldsymbol{d}), \Pi_{\mathfrak{l}}\boldsymbol{q}\rangle,\\ &= -\eta^{-1}\log\langle \exp(-\eta\boldsymbol{d}), \boldsymbol{q}\rangle. \end{aligned} \tag{82}$$

Fix $x \in [n]$ and let $\hat{\boldsymbol{d}} := \ell_x(A) \in [0,+\infty]^k$. Let $(\hat{\boldsymbol{d}}_m) \subset [0,+\infty[^k$ be any sequence converging to $\hat{\boldsymbol{d}}$. Lemma 34, $\mathrm{Mix}^{\eta}_{\mathrm{S}}(\hat{\boldsymbol{d}}_m, \boldsymbol{q}) \overset{m\to\infty}{\to} \mathrm{Mix}^{\eta}_{\mathrm{S}}(\hat{\boldsymbol{d}}, \boldsymbol{q})$. Using this with (82) gives

$$\begin{aligned}-\eta^{-1}\log\langle \exp(-\eta\ell_x(A)), \boldsymbol{q}\rangle &= \lim_{m\to\infty} -\eta^{-1}\log\langle \exp(-\eta\hat{\boldsymbol{d}}_m), \boldsymbol{q}\rangle,\\ &= \lim_{m\to\infty} \mathrm{Mix}^{\eta}_{\mathrm{S}}(\hat{\boldsymbol{d}}_m, \boldsymbol{q}),\\ &= \mathrm{Mix}^{\eta}_{\mathrm{S}}(\hat{\boldsymbol{d}}, \boldsymbol{q}) = \mathrm{Mix}^{\eta}_{\mathrm{S}}(\ell_x(A), \boldsymbol{q}). \end{aligned} \tag{83}$$

It remains to check the case where $\boldsymbol{q}$ *is a vertex*; Without loss of generality assume that $\boldsymbol{q} = \boldsymbol{e}_1$ and let $\boldsymbol{\mu} \in \Delta_k \setminus \{\boldsymbol{e}_1\}$. Then there exists $\mathfrak{l}_* \subset [k]$, such that $(\boldsymbol{e}_1, \boldsymbol{\mu}) \in (\operatorname{rbd}\Delta_{\mathfrak{l}_*}) \times (\operatorname{ri}\Delta_{\mathfrak{l}_*})$ and by Lemma 30, $\mathrm{S}'(\boldsymbol{e}_1; \boldsymbol{\mu} - \boldsymbol{e}_1) = -\infty$. Therefore, $\forall \boldsymbol{q} \in \Delta_k \setminus \{\boldsymbol{e}_1\}, D_{\mathrm{S}_\eta}(\boldsymbol{q}, \boldsymbol{e}_1) = +\infty$, which implies

$$\begin{aligned}\forall x \in [n], \mathrm{Mix}^{\eta}_{\mathrm{S}}(\ell_x(A), \boldsymbol{e}_1) &= \inf_{\boldsymbol{q}\in\Delta_k} \langle \boldsymbol{q}, \ell_x(A)\rangle + D_{\mathrm{S}_\eta}(\boldsymbol{q}, \boldsymbol{e}_1),\\ &= \langle \boldsymbol{e}_1, \ell_x(A)\rangle + D_{\mathrm{S}_\eta}(\boldsymbol{e}_1, \boldsymbol{e}_1),\\ &= \langle \boldsymbol{e}_1, \ell_x(A)\rangle,\\ &= \ell_x(\boldsymbol{a}_1) = -\eta^{-1}\log\langle \exp(-\eta\ell_x(A)), \boldsymbol{e}_1\rangle. \end{aligned} \tag{84}$$

Combining (84) and (83) proves the claim in (77). The desired equivalence follows trivially from the definitions of η-mixability and (η, S)-mixability.

□

C.6 Proof of Theorem 13

We need the following lemma to show Theorem 13.

Lemma 40. *Let* Φ *be as in Theorem 13. Then* $\eta_\ell\Phi - \mathrm{S}$ *is convex on* Δ_k *only if* Φ *satisfies* (22).

Proof. Let $\hat{\boldsymbol{q}} \in \operatorname{rbd}\Delta_k$. Suppose that there exists $\boldsymbol{q} \in \operatorname{ri}\Delta_k$ such that $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) > -\infty$. Since Φ is convex, it must have non-decreasing slopes; in particular, it holds that $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) \le \Phi(\boldsymbol{q}) - \Phi(\hat{\boldsymbol{q}})$. Therefore, since Φ is finite on Δ_k (by definition of an entropy), we have $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) < +\infty$. Since by assumption $\eta_\ell\Phi - \mathrm{S}$ is convex and finite on the simplex, we can use the same argument to show that $[\eta_\ell\Phi - \mathrm{S}]'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = \eta_\ell\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) - \mathrm{S}'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) < +\infty$. This is a contradiction since $\mathrm{S}'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = -\infty$ (Lemma 30). Therefore, it must hold that $\Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}}) = -\infty$.

Suppose now that $(\hat{\boldsymbol{q}}, \boldsymbol{q}) \in (\operatorname{rbd}\Delta_{\mathfrak{l}}) \times (\operatorname{ri}\Delta_{\mathfrak{l}})$ for $\mathfrak{l} \subseteq [k]$, with $|\mathfrak{l}| > 1$. Let $\Phi_{\mathfrak{l}} := \Phi \circ \Pi_{\mathfrak{l}}^{\mathsf{T}}$ and $\mathrm{S}_{\mathfrak{l}} := \mathrm{S} \circ \Pi_{\mathfrak{l}}^{\mathsf{T}}$. Since $\eta_\ell\Phi - \mathrm{S}$ is convex on Δ_k and $\Pi_{\mathfrak{l}}$ is a linear function, $\eta_\ell\Phi_{\mathfrak{l}} - \mathrm{S}_{\mathfrak{l}}$ is convex on $\Delta_{|\mathfrak{l}|}$. Repeating the steps above for Φ and S substituted by $\Phi_{\mathfrak{l}}$ and $\mathrm{S}_{\mathfrak{l}}$, respectively, we get that $(\Phi_{\mathfrak{l}})'(\Pi_{\mathfrak{l}}\hat{\boldsymbol{q}}; \Pi_{\mathfrak{l}}\boldsymbol{q} - \Pi_{\mathfrak{l}}\hat{\boldsymbol{q}}) = -\infty$. Since $(\Phi_{\mathfrak{l}})'(\Pi_{\mathfrak{l}}\hat{\boldsymbol{q}}; \Pi_{\mathfrak{l}}\boldsymbol{q} - \Pi_{\mathfrak{l}}\hat{\boldsymbol{q}}) = \Phi'(\hat{\boldsymbol{q}}; \boldsymbol{q} - \hat{\boldsymbol{q}})$ the proof is completed. □

Theorem 13 *Let* $\eta > 0$, $\ell\colon \mathcal{A} \to [0,+\infty]^n$ *a* η*-mixable loss, and* $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ *an entropy. If* $\eta\Phi - \mathrm{S}$ *is convex on* Δ_k*, then* ℓ *is* Φ*-mixable.*

Proof. Assume $\eta_\ell\Phi - \mathrm{S}$ is convex on Δ_k. For this to hold, it is necessary that $\eta_\ell > 0$ since $-\mathrm{S}$ is strictly concave. Let $\eta := \eta_\ell$ and $\mathrm{S}_\eta := \eta^{-1}\mathrm{S}$. Then $\tilde{\mathrm{S}}_\eta = \eta^{-1}\tilde{\mathrm{S}}$ and $\tilde{\Phi} - \tilde{\mathrm{S}}_\eta = (\Phi - \mathrm{S}_\eta) \circ \Pi_k$ is convex on $\tilde{\Delta}_k$, since $\Phi - \mathrm{S}_\eta$ is convex on Δ_k and Π_k is affine.

Let $x \in [n], A := [\boldsymbol{a}_\theta]_{\theta\in[k]}$, and $\boldsymbol{q} \in \Delta_k$. Suppose that $\boldsymbol{q} \in \operatorname{ri}\Delta_k$ and let $\boldsymbol{s}^*_{\boldsymbol{q}} \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}})$ be as in Proposition 33. Note that if $\ell_x(\boldsymbol{a}_\theta) = +\infty, \forall\theta \in [k]$, then the Φ-mixability condition (8) is trivially satisfied. Suppose, without loss of generality, that $\ell_x(\boldsymbol{a}_k) < +\infty$. Let $(\boldsymbol{d}_m) \subset$

$[0,+\infty[^k$ be any sequence such that $\boldsymbol{d}_m \overset{m\to\infty}{\to} \boldsymbol{d} := \ell_x(A) \in [0,+\infty]^k$. From Lemmas 30 and 34, $\mathrm{Mix}_\Psi(\boldsymbol{d}_m, \boldsymbol{q}) \overset{m\to\infty}{\to} \mathrm{Mix}_\Psi(\boldsymbol{d}, \boldsymbol{q})$ for $\Psi \in \{\Phi, \mathrm{S}_\eta\}$.

Let $\tilde{\Upsilon}_{\boldsymbol{q}} : \mathbb{R}^{k-1} \to \mathbb{R} \cup \{+\infty\}$ be defined by

$$\tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{\mu}}) := \tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{\mu}}) + \langle \tilde{\boldsymbol{\mu}}, \boldsymbol{s}^*_{\boldsymbol{q}} - \nabla\tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})\rangle - \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) + \tilde{\mathrm{S}}^*_\eta(\nabla\tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})),$$

and it's Fenchel dual follows from Proposition 20 (i+ii):

$$\tilde{\Upsilon}^*_{\boldsymbol{q}}(\boldsymbol{v}) = \tilde{\mathrm{S}}^*_\eta(\boldsymbol{v} - \boldsymbol{s}^*_{\boldsymbol{q}} + \nabla\tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})) + \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) - \tilde{\mathrm{S}}^*_\eta(\nabla\tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})),$$

After substituting $\boldsymbol{v}$ by $\boldsymbol{s}^*_{\boldsymbol{q}} - J_k^\intercal \boldsymbol{d}$ in the expression of $\tilde{\Upsilon}^*_{\boldsymbol{q}}$ and rearranging, we get

$$\tilde{\mathrm{S}}^*_\eta(\nabla\tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})) - \tilde{\mathrm{S}}^*_\eta(\nabla\tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}}) - J_k^\intercal \boldsymbol{d}_m) = \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) - \tilde{\Upsilon}^*_{\boldsymbol{q}}(\boldsymbol{s}^*_{\boldsymbol{q}} - J_k^\intercal \boldsymbol{d}_m). \tag{85}$$

Since $\boldsymbol{s}^*_{\boldsymbol{q}} \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}})$ and $\tilde{\Phi}$ is a closed convex function, combining Proposition 20-(iv) and the fact that $\tilde{\Phi}^{**} = \tilde{\Phi}$ [8, Cor. E.1.3.6] yields $\langle \tilde{\boldsymbol{q}}, \boldsymbol{s}^*_{\boldsymbol{q}}\rangle - \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) = \tilde{\Phi}(\tilde{\boldsymbol{q}})$. Thus, after substituting $\tilde{\boldsymbol{\mu}}$ by $\tilde{\boldsymbol{q}}$ in the expression of $\tilde{\Upsilon}_{\boldsymbol{q}}$, we get

$$\tilde{\Phi}(\tilde{\boldsymbol{q}}) = \tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}}). \tag{86}$$

On the other hand, $\tilde{\Phi} - \tilde{\Upsilon}_{\boldsymbol{q}}$ is convex on $\tilde{\Delta}_k$, since $\tilde{\Upsilon}_{\boldsymbol{q}}$ is equal to $\tilde{\mathrm{S}}_\eta$ plus an affine function. Thus, $\partial[\tilde{\Phi} - \tilde{\Upsilon}_{\boldsymbol{q}}](\tilde{\boldsymbol{q}}) + \partial\tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}}) = \partial\tilde{\Phi}(\tilde{\boldsymbol{q}})$, since $\tilde{\Phi}$ and $\tilde{\Upsilon}_{\boldsymbol{q}}$ are both convex (ibid., Thm. D.4.1.1). Since $\tilde{\Upsilon}_{\boldsymbol{q}}$ is differentiable at $\tilde{\boldsymbol{q}}$, we have $\partial\tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}}) = \{\nabla\tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}})\} = \{\boldsymbol{s}^*_{\boldsymbol{q}}\}$. Furthermore, since $\boldsymbol{s}^*_{\boldsymbol{q}} \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}})$, then $\mathbf{0}_{\tilde{k}} \in \partial\tilde{\Phi}(\boldsymbol{q}) - \partial\tilde{\Upsilon}_{\boldsymbol{q}}(\tilde{\boldsymbol{q}}) = \partial[\tilde{\Phi} - \tilde{\Upsilon}_{\boldsymbol{q}}](\tilde{\boldsymbol{q}})$. Hence, $\tilde{\Phi} - \tilde{\Upsilon}_{\boldsymbol{q}}$ attains a minimum at $\tilde{\boldsymbol{q}}$ (ibid., Thm. D.2.2.1). Due to this and (86), $\tilde{\Phi} \geq \tilde{\Upsilon}_{\boldsymbol{q}}$, which implies that $\tilde{\Phi}^* \leq \tilde{\Upsilon}^*_{\boldsymbol{q}}$ (Proposition 20-(iii)). Using this in (85) gives for all $m \in \mathbb{N}$

$$\begin{aligned} \tilde{\mathrm{S}}^*_\eta(\nabla\tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}})) - \tilde{\mathrm{S}}^*_\eta(\nabla\tilde{\mathrm{S}}_\eta(\tilde{\boldsymbol{q}}) - J_k^\intercal \boldsymbol{d}_m) &\leq \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}}) - \tilde{\Phi}^*(\boldsymbol{s}^*_{\boldsymbol{q}} - J_k^\intercal \boldsymbol{d}_m), \\ \implies \mathrm{Mix}^\eta_{\mathrm{S}}(\boldsymbol{d}_m, \boldsymbol{q}) &\leq \mathrm{Mix}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}), \end{aligned}$$

where the implication is obtained by adding $[\boldsymbol{d}_m]_k$ on both sides of the first inequality and using Proposition 33.

Suppose now that $\boldsymbol{q} \in \mathrm{ri}\,\Delta_{\mathfrak{l}}$, with $|\mathfrak{l}| > 1$, and let $\Phi_{\mathfrak{l}} := \Phi \circ \Pi_{\mathfrak{l}}^\intercal$ and $\mathrm{S}_{\mathfrak{l}} := \mathrm{S} \circ \Pi_{\mathfrak{l}}^\intercal$. Note that since $\eta_\ell\Phi - \mathrm{S}$ is convex on Δ_k and $\Pi_{\mathfrak{l}}$ is a linear function, $\eta_\ell\Phi_{\mathfrak{l}} - \mathrm{S}_{\mathfrak{l}}$ is convex on $\Delta_{|\mathfrak{l}|}$. Repeating the steps above for Φ, S, $\boldsymbol{q}$, and A substituted by $\Phi_{\mathfrak{l}}$, $\mathrm{S}_{\mathfrak{l}}$, $\Pi_{\mathfrak{l}}\boldsymbol{q}$, and $A\Pi_{\mathfrak{l}}^\intercal$, respectively, yields

$$\begin{aligned} \mathrm{Mix}^\eta_{\mathrm{S}_{\mathfrak{l}}}(\Pi_{\mathfrak{l}}\boldsymbol{d}_m, \Pi_{\mathfrak{l}}\boldsymbol{q}) &\leq \mathrm{Mix}_{\Phi_{\mathfrak{l}}}(\Pi_{\mathfrak{l}}\boldsymbol{d}_m, \Pi_{\mathfrak{l}}\boldsymbol{q}), \\ \implies \mathrm{Mix}^\eta_{\mathrm{S}}(\boldsymbol{d}_m, \boldsymbol{q}) &\leq \mathrm{Mix}_\Phi(\boldsymbol{d}_m, \boldsymbol{q}), \\ \implies \mathrm{Mix}^\eta_{\mathrm{S}}(\ell_x(A), \boldsymbol{q}) &\leq \mathrm{Mix}_\Phi(\ell_x(A), \boldsymbol{q}), \end{aligned} \tag{87}$$

where the first implication follows from Lemma 32, since S_η and Φ both satisfy (22) (see Lemmas 30 and 40), and (87) is obtained by passage to the limit $m \to \infty$. Since $\eta = \eta_\ell > 0$, ℓ is η-mixable, which implies that ℓ is S_η-mixable (Theorem 11). Therefore, there exists $\boldsymbol{a}_* \in \mathcal{A}$, such that

$$\ell_x(\boldsymbol{a}_*) \leq \mathrm{Mix}^\eta_{\mathrm{S}}(\ell_x(A), \boldsymbol{q}) \leq \mathrm{Mix}_\Phi(\ell_x(A), \boldsymbol{q}). \tag{88}$$

To complete the proof (that is, to show that ℓ is Φ-mixable), it remains to consider the case where $\boldsymbol{q}$ is a vertex of Δ_k. Without loss of generality assume that $\boldsymbol{q} = \boldsymbol{e}_1$ and let $\boldsymbol{\mu} \in \Delta_k \setminus \{\boldsymbol{e}_1\}$. Thus, there exists $\mathfrak{l}_* \subseteq [k]$, with $|\mathfrak{l}_*| > 1$, such that $(\boldsymbol{e}_1, \boldsymbol{\mu}) \in (\mathrm{rbd}\,\Delta_{\mathfrak{l}_*}) \times (\mathrm{ri}\,\Delta_{\mathfrak{l}_*})$, and Lemma 40 implies that $\Phi'(\boldsymbol{e}_1; \boldsymbol{\mu} - \boldsymbol{e}_1) = -\infty$. Therefore, $\forall \boldsymbol{q} \in \Delta_k \setminus \{\boldsymbol{e}_1\}$, $D_\Phi(\boldsymbol{q}, \boldsymbol{e}_1) = +\infty$, which implies

$$\begin{aligned} \forall x \in [n], \mathrm{Mix}_\Phi(\ell_x(A), \boldsymbol{e}_1) &= \inf_{\boldsymbol{q}\in\Delta_k} \langle \boldsymbol{q}, \ell_x(A)\rangle + D_\Phi(\boldsymbol{q}, \boldsymbol{e}_1), \\ &= \langle \boldsymbol{e}_1, \ell_x(A)\rangle + D_\Phi(\boldsymbol{e}_1, \boldsymbol{e}_1) = \langle \boldsymbol{e}_1, \ell_x(A)\rangle, \\ &= \ell_x(\boldsymbol{a}_1). \end{aligned} \tag{89}$$

The Φ-mixability condition (8) is trivially satisfied in this case. Combining (88) and (89) shows that ℓ is Φ-mixable. □

C.7 Proof of Theorem 14

The following Lemma gives necessary regularity conditions on the entropy Φ under the assumptions of Theorem 14.

Lemma 41. *Let Φ and ℓ be as in Theorem 14. Then the following holds*

(i) $\tilde{\Phi}$ *is strictly concave on* $\operatorname{int}\tilde{\Delta}_k$.

(ii) $\tilde{\Phi}^*$ *is be continuously differentiable on* $\mathbb{R}^{k-1}$.

(iii) $\tilde{\Phi}^*$ *is twice differentiable on* $\mathbb{R}^{k-1}$ *and* $\forall \tilde{\boldsymbol{q}} \in \operatorname{int}\tilde{\Delta}_k, \mathsf{H}\tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})) = (\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}))^{-1}$.

(iv) For the Shannon entropy, we have $(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1} = \mathsf{H}\tilde{\mathrm{S}}^*(\nabla\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})) = \operatorname{diag}\tilde{\boldsymbol{q}} - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T}$.

Proof. Since ℓ is Φ-mixable and $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$, $\tilde{\Phi}^*$ is continously differentiable on $\mathbb{R}^{n-1}$ (Proposition 12). Therefore, $\tilde{\Phi}$ is strictly convex on $\operatorname{ri}\Delta_k$ [8, Thm. E.4.1.2].

The differentiability of $\tilde{\Phi}$ and $\tilde{\Phi}^*$ implies $\nabla\tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})) = \tilde{\boldsymbol{q}}$ (ibid.). Since $\tilde{\Phi}$ is twice differentiable on $\operatorname{int}\tilde{\Delta}_k$ (by assumption), the latter equation implies that $\tilde{\Phi}^*$ is twice differentiable on $\nabla\tilde{\Phi}(\operatorname{int}\tilde{\Delta}_k)$. Using the chain rule, we get $\mathsf{H}\tilde{\Phi}^*(\nabla\tilde{\Phi}(\boldsymbol{u}))\mathsf{H}\tilde{\Phi}(\boldsymbol{u}) = I_{\tilde{k}}$. Multiplying both sides of the equation by $(\mathsf{H}\tilde{\Phi}(\boldsymbol{u}))^{-1}$ from the right gives the expression in (iii). Note that $\mathsf{H}\tilde{\Phi}(\cdot)$ is in fact invertible on $\operatorname{int}\tilde{\Delta}_k$ since $\tilde{\Phi}$ is strictly convex on $\operatorname{int}\tilde{\Delta}_k$. It remains to show that $\nabla\tilde{\Phi}(\operatorname{int}\tilde{\Delta}_k) = \mathbb{R}^{k-1}$. This set equality follows from 1) $[\tilde{\boldsymbol{q}} \in \partial\tilde{\Phi}^*(\boldsymbol{s}) \iff \boldsymbol{s} \in \partial\tilde{\Phi}(\tilde{\boldsymbol{q}})]$ (ibid., Cor. E.1.4.4); 2) $\operatorname{dom}\tilde{\Phi}^* = \mathbb{R}^{k-1}$; and 3) $\forall\tilde{\boldsymbol{q}} \in \operatorname{bd}\tilde{\Delta}_k, \partial\tilde{\Phi}(\tilde{\boldsymbol{q}}) = \varnothing$ (Lemma 32).

For the Shannon entropy, we have $\tilde{\mathrm{S}}^*(\boldsymbol{v}) = \log(\langle\exp(\boldsymbol{v}), \mathbf{1}_{\tilde{\boldsymbol{k}}}\rangle + 1)$ (Proposition 39) and $\nabla\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}) = \log\frac{\tilde{\boldsymbol{q}}}{q_k}$, for $(\boldsymbol{v}, \tilde{\boldsymbol{q}}) \in \mathbb{R}^{k-1} \times \tilde{\Delta}_k$. Thus $(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1} = \mathsf{H}\tilde{\mathrm{S}}^*(\nabla\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})) = \operatorname{diag}\tilde{\boldsymbol{q}} - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T}$. □

To show Theorem 14, we analyze a particular parameterized curve defined in the next lemma.

Lemma 42. *Let $\ell\colon \Delta_n \to [0,+\infty]^n$ be a proper loss whose Bayes risk $\underline{L}_\ell$ is twice differentiable on $]0,+\infty[^n$, and let Φ be an entropy such that $\tilde{\Phi}$ and $\tilde{\Phi}^*$ are twice differentiable on $\operatorname{int}\tilde{\Delta}_k$ and $\mathbb{R}^{k-1}$, respectively. For $(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, V) \in \operatorname{int}\tilde{\Delta}_n \times \operatorname{int}\tilde{\Delta}_k \times \mathbb{R}^{\tilde{n}\times\tilde{k}}$, let $\beta : \mathbb{R} \to \mathbb{R}^n$ be the curve defined by*

$$\forall x \in [n], \quad \beta_x(t) = \tilde{\ell}_x(\tilde{\boldsymbol{p}}) + \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})) - \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}) - J_k^\mathsf{T}\tilde{\ell}_x(\tilde{P}^t)), \tag{90}$$

where $\tilde{P}^t = [\tilde{\boldsymbol{p}}\mathbf{1}_{\tilde{k}}^\mathsf{T} + tV,\ \tilde{\boldsymbol{p}}] \in \mathbb{R}^{\tilde{n}\times k}$ and $t \in \{s \in \mathbb{R} : \forall j \in [\tilde{k}],\ \tilde{\boldsymbol{p}} + sV_{\bullet,j} \in \operatorname{int}\tilde{\Delta}_n\}$. Then

$$\begin{aligned}
\beta(0) &= \tilde{\ell}(\tilde{\boldsymbol{p}}),\\
\dot{\beta}(0) &= \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V\tilde{\boldsymbol{q}},\\
\frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t)\right\rangle\Big|_{t=0} &= -\sum_{j=1}^{k-1} q_j V_{\bullet,j}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})V_{\bullet,j} - \operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\,\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}))^{-1}(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V)^\mathsf{T}).
\end{aligned} \tag{91}$$

Proof. Since $\tilde{P}^t = [\tilde{\boldsymbol{p}}\mathbf{1}_{\tilde{k}}^\mathsf{T} + tV,\ \tilde{\boldsymbol{p}}] \in \mathbb{R}^{\tilde{n}\times k}$, $\tilde{P}^0 = \tilde{\boldsymbol{p}}\mathbf{1}_k^\mathsf{T}$ and $\tilde{\ell}_x(\tilde{P}^0) = \tilde{\ell}_x(\tilde{\boldsymbol{p}})\mathbf{1}_k$. As a result, $J_k^\mathsf{T}\tilde{\ell}_x(\tilde{P}^0) = \mathbf{0}_{\tilde{k}}$, and thus $\beta_x(0) = \tilde{\ell}_x(\tilde{\boldsymbol{p}}) + \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})) - \tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}) - \mathbf{0}_{\tilde{k}}) = \tilde{\ell}_x(\tilde{\boldsymbol{p}})$. This shows that $\beta(0) = \tilde{\ell}(\tilde{\boldsymbol{p}})$. Let $\gamma_x(t) \coloneqq \nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}) - J_k^\mathsf{T}\tilde{\ell}_x(\tilde{P}^t)$. For $j \in [k-1]$,

$$\begin{aligned}
\frac{d}{dt}[\gamma_x(t)]_j &= \frac{d}{dt}\left([\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})]_j - [J_k^\mathsf{T}\tilde{\ell}_x(\tilde{P}^t)]_j\right),\\
&= -\frac{d}{dt}\left(\tilde{\ell}_x(\tilde{P}^t_{\bullet,j}) - \tilde{\ell}_x(\tilde{P}^t_{\bullet,k})\right),\\
&= -\frac{d}{dt}\left(\tilde{\ell}_x(\tilde{\boldsymbol{p}} + tV_{\bullet,j}) - \tilde{\ell}_x(\tilde{\boldsymbol{p}})\right), \qquad \left(\text{since } \frac{d}{dt}\ell_x(\tilde{P}^t_{\bullet,k}) = \frac{d}{dt}\tilde{\ell}_x(\tilde{\boldsymbol{p}}) = 0\right)\\
&= -\mathsf{D}\tilde{\ell}_x(\tilde{P}^t_{\bullet,j})V_{\bullet,j}.
\end{aligned}$$

From the definition of $\tilde{P}^t$, $\tilde{P}^0_{\bullet,j} = \tilde{\boldsymbol{p}}$, $\forall j \in [\tilde{k}]$, and therefore, $\dot{\gamma}_x(0) = -(\mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}})V)^\mathsf{T}$. By differentiating β_x in (90) and using the chain rule, $\dot{\beta}_x(t) = -(\dot{\gamma}_x(t))^\mathsf{T}\nabla\tilde{\Phi}^*(\gamma_x(t))$. By setting $t = 0$, $\dot{\beta}_x(0) = -(\dot{\gamma}_x(0))^\mathsf{T}\nabla\tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}})) = \mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}})V\tilde{\boldsymbol{q}}$. Thus, $\dot{\beta}(0) = \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V\tilde{\boldsymbol{q}}$. Furthermore,

$$
\begin{aligned}
\frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t)\right\rangle\Big|_{t=0} &= \frac{d}{dt}\sum_{x=1}^{n} p_x \left(\sum_{j=1}^{k-1} \mathsf{D}\tilde{\ell}_x(\tilde{P}^t_{\bullet,j})V_{\bullet,j}[\nabla\tilde{\Phi}^*(\gamma_x(t))]_j\right)\Bigg|_{t=0}, \\
&= \sum_{j=1}^{k-1}\frac{d}{dt}\left(\sum_{x=1}^{n} p_x \mathsf{D}\tilde{\ell}_x(\tilde{P}^t_{\bullet,j})V_{\bullet,j}[\nabla\tilde{\Phi}^*(\gamma_x(t))]_j\right)\Bigg|_{t=0}, \\
&= \sum_{j=1}^{k-1}\left(\frac{d}{dt}\left\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{P}^t_{\bullet,j})V_{\bullet,j}q_j\right\rangle\Big|_{t=0} + \sum_{x=1}^{n} p_x \mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}})V_{\bullet,j}\,\frac{d}{dt}[\nabla\tilde{\Phi}^*(\gamma_x(t))]_j\Big|_{t=0}\right), \\
&= -\sum_{j=1}^{k-1} q_j V_{\bullet,j}^\mathsf{T}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})V_{\bullet,j} - \sum_{x=1}^{n}\sum_{\substack{i=1\\ j=1}}^{k-1} p_x \mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}})V_{\bullet,j}[\mathsf{H}\tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}))]_{j,i}\mathsf{D}\tilde{\ell}_x(\tilde{\boldsymbol{p}})V_{\bullet,i}, \\
&= -\sum_{j=1}^{k-1} q_j V_{\bullet,j}^\mathsf{T}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})V_{\bullet,j} - \operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\,\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V\mathsf{H}\tilde{\Phi}^*(\nabla\tilde{\Phi}(\tilde{\boldsymbol{q}}))(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V)^\mathsf{T}), \\
&= -\sum_{j=1}^{k-1} q_j V_{\bullet,j}^\mathsf{T}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})V_{\bullet,j} - \operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\,\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}))^{-1}(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V)^\mathsf{T}),
\end{aligned}
$$

where in the third equality we used Lemma 25, in the fourth equality we used Lemma 28, and in the sixth equality we used Lemma 41-(iii).

□

In next lemma, we state a necessary condition for Φ-mixability in terms of the parameterized curve β defined in Lemma 42.

Lemma 43. *Let ℓ, Φ, and β be as in Lemma 42. If $\exists(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, V) \in \operatorname{int}\tilde{\Delta}_n \times \operatorname{int}\tilde{\Delta}_k \times \mathbb{R}^{\tilde{n}\times\tilde{k}}$ such that the curve $\gamma(t) := \tilde{\ell}(\tilde{\boldsymbol{p}} + tV\tilde{\boldsymbol{q}})$ satisfies $\frac{d}{dt}\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\rangle\Big|_{t=0} < 0$, then ℓ is not $\Phi-$mixable. In particular, $\exists P \in \operatorname{ri}\Delta_n^k$, such that $[\operatorname{Mix}_\Phi(\ell_x(P), \boldsymbol{q})]_{x\in[n]}^\mathsf{T}$ lies outside $\mathscr{S}_\ell$.*

Proof. First note that for any triplet $(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}, V) \in \operatorname{int}\tilde{\Delta}_n \times \operatorname{int}\tilde{\Delta}_k \times \mathbb{R}^{\tilde{n}\times\tilde{k}}$, the map $t \mapsto \left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle$ is differentiable at 0. This follows from Lemmas 25 and 42. Let $r(t) := \amalg_n(\tilde{\boldsymbol{p}} + tV\tilde{\boldsymbol{q}})$ and $\delta(t) := \langle r(t), \beta(t) - \gamma(t)\rangle$. Then

$$
\dot{\delta}(t) = \left\langle r(t), \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle + \langle V\tilde{\boldsymbol{q}}, \beta(t) - \gamma(t)\rangle\,.
$$

Since $t \mapsto \langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\rangle$ is differentiable at 0, it follows from Lemma 25 that $t \mapsto \dot{\delta}(t)$ is also differentiable at 0, and thus

$$
\begin{aligned}
\ddot{\delta}(0) &= \frac{d}{dt}\left\langle r(t), \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0} + \left\langle J_n V\tilde{\boldsymbol{q}}, \dot{\beta}(0) - \dot{\gamma}(0)\right\rangle, \\
&= \left\langle \dot{r}(0), \dot{\beta}(0) - \dot{\gamma}(0)\right\rangle + \frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0}, \qquad (92) \\
&= \left\langle J_n V\tilde{\boldsymbol{q}}, \dot{\beta}(0) - \dot{\gamma}(0)\right\rangle + \frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0}, \\
&= \frac{d}{dt}\left\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\right\rangle\Big|_{t=0} < 0, \qquad (93)
\end{aligned}
$$

where (92) and (93) hold because $\dot{\beta}(0) = \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})V\tilde{\boldsymbol{q}} = \dot{\gamma}(0)$ (see Lemma 42). According to Taylor's theorem (see e.g. [3, §151]), there exists $\epsilon > 0$ and $h : [-\epsilon, \epsilon] \to \mathbb{R}$ such that

$$\forall |t| \leq \epsilon,\ \delta(t) = \delta(0) + t\dot{\delta}(0) + \frac{t^2}{2}\ddot{\delta}(0) + h(t)t^2, \tag{94}$$

and $\lim_{t\to 0} h(t) = 0$. From Lemma 42, $\beta(0) = \gamma(0) = 0$ and $\dot{\beta}(0) = \dot{\gamma}(0)$. Therefore, $\delta(0) = \dot{\delta}(0) = 0$ and (94) becomes $\delta(t) = \frac{t^2}{2}\ddot{\delta}(0) + h(t)t^2$. Due to (93) and the fact that $\lim_{t\to 0} h(t) = 0$, we can choose $\epsilon_* > 0$ small enough such that $\delta(\epsilon_*) = \frac{\epsilon_*^2}{2}\ddot{\delta}(0) + h(\epsilon_*)\epsilon_*^2 < 0$. This means that $\langle \mathrm{II}_n(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}}), \beta(\epsilon_*)\rangle < \langle \mathrm{II}_n(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}}), \tilde{\ell}(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}})\rangle = \langle \mathrm{II}_n(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}}), \ell(\mathrm{II}_n(\tilde{\boldsymbol{p}} + \epsilon_* V\tilde{\boldsymbol{q}})\rangle$. Therefore, $\beta(\epsilon_*)$ must lie outside the superprediction set. Thus, the mixability condition (8) does not hold for $P^{\epsilon_*} = \mathrm{II}_n[\tilde{\boldsymbol{p}}\mathbf{1}_{\tilde{k}}^{\mathsf{T}} + \epsilon_* V,\ \tilde{\boldsymbol{p}}] \in \operatorname{ri}\Delta_n^k$. This completes the proof. □

Theorem 14 *Let $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a loss such that $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$, and $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ an entropy such that $\tilde{\Phi} := \Phi \circ \mathrm{II}_k$ is twice differentiable on $\operatorname{int}\tilde{\Delta}_k$. Then ℓ is Φ-mixable only if $\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k.*

Proof. We will prove the contrapositive; suppose that $\underline{\eta_\ell}\Phi - \mathrm{S}$ is not convex on Δ_k and we show that ℓ cannot be Φ-mixable. Note first that from Lemma 41-(iii), $\tilde{\Phi}^*$ is twice differentiable on $\mathbb{R}^{k-1}$. Thus Lemmas 42 and 43 apply. Let $\underline{\ell}$ be a proper support loss of ℓ and suppose that $\underline{\eta_\ell}\Phi - \mathrm{S}$ is not convex on Δ_k, This implies that $\underline{\eta_\ell}\tilde{\Phi} - \tilde{\mathrm{S}}$ is not convex on $\operatorname{int}\tilde{\Delta}_k$, and by Lemma 22 there exists $\tilde{\boldsymbol{q}}_* \in \operatorname{int}\tilde{\Delta}_k$, such that $1 > \underline{\eta_\ell}\lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}_*)(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}_*))^{-1})$. From this and the definition of $\underline{\eta_\ell}$, there exists $\tilde{\boldsymbol{p}}_* \in \operatorname{int}\tilde{\Delta}_n$ such that

$$1 > \frac{\lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}_*)(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}_*))^{-1})}{\lambda_{\max}([\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{\boldsymbol{p}}_*)]^{-1}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}}_*))} = \frac{\lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}_*)(\operatorname{diag}(\tilde{\boldsymbol{q}}_*) - \tilde{\boldsymbol{q}}_*\tilde{\boldsymbol{q}}_*^{\mathsf{T}}))}{\lambda_{\max}([\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{\boldsymbol{p}}_*)]^{-1}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}}_*))}, \tag{95}$$

where the equality is due to Lemma 41-(iv). For the rest of this proof let $(\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}}) = (\tilde{\boldsymbol{p}}^*, \tilde{\boldsymbol{q}}^*)$. By assumption, $\tilde{\underline{L}}_\ell$ twice differentiable and concave on $\operatorname{int}\tilde{\Delta}_n$, and thus $-\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})$ is symmetric positive semi-definite. Therefore, their exists a symmetric positive semi-definite matrix $\Lambda_{\boldsymbol{p}}$ such that $\Lambda_{\boldsymbol{p}}\Lambda_{\boldsymbol{p}} = -\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})$. From Lemma 41-(i), $\tilde{\Phi}$ is strictly convex on $\operatorname{int}\tilde{\Delta}_k$, and so there exists a symmetric positive definite matrix $K_{\boldsymbol{q}}$ such that $K_{\boldsymbol{q}}K_{\boldsymbol{q}} = \mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})$. Let $\boldsymbol{w} \in \mathbb{R}^{n-1}$ be the unit norm eigenvector of $[\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})$ associated with $\lambda_*^\ell := \lambda_{\max}([\mathsf{H}\tilde{\underline{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}}))$. Suppose that $c_\ell := \boldsymbol{w}^{\mathsf{T}}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} = 0$. Since $\boldsymbol{w}^{\mathsf{T}}\Lambda_{\boldsymbol{p}}\Lambda_{\boldsymbol{p}}\boldsymbol{w} = -c_\ell = 0$, it follows from the positive semi-definiteness of $\Lambda_{\boldsymbol{p}}$ that $\Lambda_{\boldsymbol{p}}\boldsymbol{w} = \mathbf{0}_{\tilde{n}}$, and thus $\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} = -\Lambda_{\boldsymbol{p}}\Lambda_{\boldsymbol{p}}\boldsymbol{w} = \mathbf{0}_{\tilde{n}}$. This implies that $\lambda_*^\ell = 0$, which is not possible due to (95). Therefore, $\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} \neq \mathbf{0}_{\tilde{n}}$. Furthermore, the negative semi-definiteness of $\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})$ implies that

$$c_\ell = \boldsymbol{w}^{\mathsf{T}}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} < 0. \tag{96}$$

Let $\boldsymbol{v} \in \mathbb{R}^{k-1}$ be the unit norm eigenvector of $K_{\boldsymbol{q}}(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^{\mathsf{T}})K_{\boldsymbol{q}}$ associated with $\lambda_*^\Phi := \lambda_{\min}(K_{\boldsymbol{q}}(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^{\mathsf{T}})K_{\boldsymbol{q}}) = \lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})(\operatorname{diag}(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^{\mathsf{T}}))$, where the equality is due to Lemma 21. Let $\hat{\boldsymbol{v}} := K_{\boldsymbol{q}}\boldsymbol{v}$.

We will show that for $V = \boldsymbol{w}\hat{\boldsymbol{v}}^{\mathsf{T}}$, the parametrized curve β defined in Lemma 42 satisfies $\frac{d}{dt}\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\rangle\Big|_{t=0} < 0$, where $\gamma(t) = \tilde{\underline{\ell}}(\tilde{\boldsymbol{p}} + tV\tilde{\boldsymbol{q}})$. According to Lemma 43 this would imply that there exists $P \in \operatorname{ri}\Delta_n^k$, such that $[\mathrm{Mix}_\Phi(\underline{\ell}_x(P), \boldsymbol{q})]_{x\in[n]}^{\mathsf{T}}$ lies outside $\mathscr{S}_\ell$. From Theorem 5, we know that there exists $A_* \in \mathcal{A}^k$, such that $\ell_x(A_*) = \underline{\ell}_x(P), \forall x \in [n]$. Therefore, $[\mathrm{Mix}_\Phi(\ell_x(A_*), \boldsymbol{q})]_{x\in[n]}^{\mathsf{T}} = [\mathrm{Mix}_\Phi(\underline{\ell}_x(P), \boldsymbol{q})]_{x\in[n]}^{\mathsf{T}} \notin \mathscr{S}_\ell$, and thus ℓ is not Φ-mixable.

From Lemma 42 (Equation 91) and the fact that $V_{\cdot,j} = \hat{v}_j\boldsymbol{w}$, for $j \in [\tilde{k}]$, we can write

$$\begin{aligned}\frac{d}{dt}\Big\langle \boldsymbol{p}, \dot{\beta}(t)\Big\rangle\Big|_{t=0} &= -\sum_{j=1}^{k-1} q_j\hat{v}_j^2\boldsymbol{w}^{\mathsf{T}}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} - \operatorname{tr}(\operatorname{diag}(\boldsymbol{p})\,\mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})V(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}}))^{-1}(\mathsf{D}\tilde{\underline{\ell}}(\tilde{\boldsymbol{p}})V)^{\mathsf{T}}),\\ &= -\langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}} \odot \hat{\boldsymbol{v}}\rangle\,\boldsymbol{w}^{\mathsf{T}}\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{w} - (\hat{\boldsymbol{v}}^{\mathsf{T}}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})\langle \boldsymbol{p}, [\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w}] \odot [(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w}]\rangle,\end{aligned}$$

where the second equality is obtained by noting that 1) $(\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})$ is a scalar quantity and can be factorized out; and 2) $\mathrm{tr}(\mathrm{diag}(\boldsymbol{p})\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w}(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w})^\mathsf{T}) = \langle \boldsymbol{p}, (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w}) \odot (\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{w})\rangle$.

On the other hand, from Lemma 28, $\frac{d}{dt}\langle \boldsymbol{p}, \dot{\gamma}(t)\rangle\big|_{t=0} = -\langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}}\rangle^2 \boldsymbol{w}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{q}})\boldsymbol{w}$. Using (17) and the definition of c_ℓ, we get

$$
\begin{aligned}
\frac{d}{dt}\Big\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\Big\rangle\Big|_{t=0} &= [-\langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}} \odot \hat{\boldsymbol{v}}\rangle + \langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}}\rangle^2]c_\ell + \\
&\qquad (\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})(\boldsymbol{w}^\mathsf{T}(\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))(\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\boldsymbol{p})\boldsymbol{w}), \\
&= -c_\ell[\langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}} \odot \hat{\boldsymbol{v}}\rangle - \langle \tilde{\boldsymbol{q}}, \hat{\boldsymbol{v}}\rangle^2 - \lambda_*^\ell(\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})], \\
&= -c_\ell[\hat{\boldsymbol{v}}^\mathsf{T}(\mathrm{diag}\,(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})\hat{\boldsymbol{v}} - \lambda_*^\ell(\hat{\boldsymbol{v}}^\mathsf{T}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q}))^{-1}\hat{\boldsymbol{v}})], \\
&= -c_\ell[\hat{\boldsymbol{v}}^\mathsf{T}(\mathrm{diag}\,(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})\hat{\boldsymbol{v}} - \lambda_*^\ell(\boldsymbol{v}^\mathsf{T}K_{\boldsymbol{q}}(K_{\boldsymbol{q}}K_{\boldsymbol{q}})^{-1}K_{\boldsymbol{q}}\boldsymbol{v})], \\
&= -c_\ell[\boldsymbol{v}^\mathsf{T}K_{\boldsymbol{q}}(\mathrm{diag}\,(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})K_{\boldsymbol{q}}\boldsymbol{v} - \lambda_*^\ell], \qquad (97) \\
&= -c_\ell[\lambda_*^\Phi - \lambda_*^\ell], \\
&= -c_\ell[\lambda_{\min}(\mathsf{H}\tilde{\Phi}(\boldsymbol{q})(\mathrm{diag}\,(\tilde{\boldsymbol{q}}) - \tilde{\boldsymbol{q}}\tilde{\boldsymbol{q}}^\mathsf{T})) - \lambda_{\max}(\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})(\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1})],
\end{aligned}
$$

where in (97) we used the fact that $\boldsymbol{v}^\mathsf{T}\boldsymbol{v} = 1$. The last equality combined with (95) and (96) shows that $\frac{d}{dt}\langle \boldsymbol{p}, \dot{\beta}(t) - \dot{\gamma}(t)\rangle\Big|_{t=0} < 0$, which completes the proof. □

C.8 Proof of Lemma 15

Lemma 15 *Let $\ell\colon \mathcal{A} \to [0, +\infty]^n$ be a loss. If $\mathrm{dom}\,\ell = \mathcal{A}$, then either $\mathfrak{H}_\ell = \varnothing$ or $\eta_\ell \in \mathfrak{H}_\ell$.*

Proof. Suppose $\mathfrak{H}_\ell \neq \varnothing$. Let $\boldsymbol{q} \in \Delta_k$, $A \coloneqq \boldsymbol{a}_{1:k} \in \mathcal{A}^k$. By definition of η_ℓ there exists $(\eta_m) \subset [0, +\infty[$ such that ℓ is η_m-mixable and $\eta_m \overset{m\to\infty}{\to} \eta_\ell$. Therefore, $\forall m \in \mathbb{N}$, $\exists \boldsymbol{a}_m \in \mathcal{A}$ such that

$$
\forall x \in [n],\ \ell_x(\boldsymbol{a}_m) \leq -\eta_m^{-1}\log\langle \boldsymbol{q}, \exp(-\eta_m(\ell_x(A))\rangle < +\infty, \qquad (98)
$$

where the right-most inequality follows from the fact $\mathrm{dom}\,\ell = \mathcal{A}$. Therefore, the sequence $(\ell(\boldsymbol{a}_m)) \subset [0, +\infty[^n$ is bounded, and thus admits a convergent subsequence. If we let $\boldsymbol{s}$ be the limit of this subsequence, then from (98) it follows that

$$
\forall x \in [n],\ \boldsymbol{s} \leq -\eta_\ell^{-1}\log\langle \boldsymbol{q}, \exp(-\eta_\ell(\ell_x(A))\rangle, \qquad (99)
$$

On the other hand, since ℓ is closed (by Assumption 1), it follows that there exists $\boldsymbol{a}_* \in \mathcal{A}$ such that $\ell(\boldsymbol{a}_*) = \boldsymbol{s}$. Combining this with (99) implies that ℓ is η_ℓ-mixable, and thus $\eta_\ell \in \mathfrak{H}_\ell$. □

C.9 Proof of Theorem 17

Theorem 17 *Let ℓ and Φ be as in Theorem 16. Then*

$$
\eta_\ell^\Phi = \underline{\eta_\ell} \inf_{\tilde{\boldsymbol{q}} \in \mathrm{int}\,\tilde{\Delta}_k} \lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}),
$$

Proof. From Theorem 16, ℓ is Φ_η-mixable if and only if $\underline{\eta_\ell}\Phi_\eta - \mathrm{S} = \eta^{-1}\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k. When this is the case, Lemma 22 implies that

$$
1 \leq \eta^{-1}\underline{\eta_\ell}(\inf_{\tilde{\boldsymbol{q}} \in \mathrm{int}\,\tilde{\Delta}_k} \lambda_{\min}[\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})[\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}})]^{-1}]), \qquad (100)
$$

where we used the facts that $\mathsf{H}(\eta^{-1}\underline{\eta_\ell}\tilde{\Phi}) = \eta^{-1}\underline{\eta_\ell}\mathsf{H}\tilde{\Phi}$, $\lambda_{\min}(\cdot)$ is linear, and $\eta^{-1}\underline{\eta_\ell}$ is independent of $\tilde{\boldsymbol{q}} \in \mathrm{int}\,\tilde{\Delta}_k$. Inequality 100 shows that the largest η such that ℓ is Φ_η-mixable is given by η_ℓ^Φ in (11). □

C.10 Proof of Theorem 18

Theorem 18 *Let* $\mathrm{S}, \Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$*, where* S *is the Shannon entropy and* Φ *is an entropy such that* $\tilde{\Phi} \coloneqq \Phi \circ \Pi_k$ *is twice differentiable on* $\operatorname{int} \tilde{\Delta}_k$*. A loss* $\ell \colon \mathcal{A} \to [0, +\infty[^n$*, with* $\underline{L}_\ell$ *twice differentiable on* $]0, +\infty[^n$*, is* Φ*-mixable only if* $R_\ell^{\mathrm{S}} \leq R_\ell^{\Phi}$*.*

Proof. Suppose ℓ is Φ-mixable. Then from Theorem 16, $\underline{\eta_\ell}\Phi - \mathrm{S}$ is convex on Δ_k, and thus $\underline{\eta_\ell} = \eta_\ell^{\mathrm{S}} > 0$ (Corollary 17). Furthermore, $\underline{\eta_\ell}\tilde{\Phi} - \tilde{\mathrm{S}} = [\underline{\eta_\ell}\Phi - \mathrm{S}] \circ \Pi_k$ is convex on $\operatorname{int} \tilde{\Delta}_k$, since Π_k is an affine function. It follows from Lemma 22 and Corollary 17 that

$$\eta_\ell^{\Phi} = \underline{\eta_\ell} \inf_{\tilde{\boldsymbol{q}} \in \operatorname{int} \tilde{\Delta}_k} \lambda_{\min}(\mathsf{H}\tilde{\Phi}(\tilde{\boldsymbol{q}})(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}) \geq 1 > 0.$$

Let $\boldsymbol{\mu} \in \operatorname{ri} \Delta_k$ and $\theta_* \coloneqq \operatorname{argmax}_\theta D_{\mathrm{S}}(\boldsymbol{e}_\theta, \boldsymbol{\mu})$. By definition of an entropy and the fact that the directional derivatives $\Phi'(\boldsymbol{\mu}; \cdot)$ and $\mathrm{S}'(\boldsymbol{\mu}; \cdot)$ are finite on Δ_k [8, Prop. D.1.1.2], it holds that $D_\Phi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}), D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) \in]0, +\infty[$. Therefore, there exists $\alpha > 0$ such that $\alpha^{-1} D_\Phi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) = D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})$. If we let $\Psi \coloneqq \alpha^{-1}\Phi$, we get

$$D_\Psi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) = D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}). \tag{101}$$

Let $d_\Psi(\tilde{\boldsymbol{q}}) \coloneqq \tilde{\Psi}(\tilde{\boldsymbol{q}}) - \tilde{\Psi}(\tilde{\boldsymbol{\mu}}) - \langle \tilde{\boldsymbol{q}} - \tilde{\boldsymbol{\mu}}, \nabla\tilde{\Psi}(\tilde{\boldsymbol{\mu}}) \rangle$. Observe that

$$d_\Psi(\tilde{\boldsymbol{q}}) = \Psi(\boldsymbol{q}) - \Psi(\boldsymbol{\mu}) - \langle \boldsymbol{q} - \boldsymbol{\mu}, \nabla\Psi(\boldsymbol{\mu}) \rangle = D_\Psi(\boldsymbol{q}, \boldsymbol{\mu}).$$

We define d_{S} similarly. Suppose that $\eta_\ell^{\Psi} > \eta_\ell^{\mathrm{S}} = \underline{\eta_\ell}$. Then, from Corollary 17, $\forall \tilde{\boldsymbol{q}} \in \operatorname{int} \tilde{\Delta}_k$, $\lambda_{\min}(\mathsf{H}\tilde{\Psi}(\tilde{\boldsymbol{q}})(\mathsf{H}\tilde{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}) > 1$. This implies that $\forall \tilde{\boldsymbol{q}} \in \operatorname{int} \tilde{\Delta}_k$, $\lambda_{\min}(\mathsf{H}d_\Psi(\tilde{\boldsymbol{q}})(\mathsf{H}d_{\mathrm{S}}(\tilde{\boldsymbol{q}}))^{-1}) > 1$, and from Lemma 22, $d_\Psi - d_{\mathrm{S}}$ must be strictly convex on $\operatorname{int} \tilde{\Delta}_k$. We also have $\nabla d_\Psi(\tilde{\boldsymbol{\mu}}) - \nabla d_{\mathrm{S}}(\tilde{\boldsymbol{\mu}}) = 0$ and $d_\Psi(\tilde{\boldsymbol{\mu}}) - d_{\mathrm{S}}(\tilde{\boldsymbol{\mu}}) = 0$. Therefore, $d_\Psi - d_{\mathrm{S}}$ attains a strict minimum at $\tilde{\boldsymbol{\mu}}$ (ibid., Thm. D.2.2.1); that is, $d_\Psi(\tilde{\boldsymbol{q}}) > d_{\mathrm{S}}(\tilde{\boldsymbol{q}})$, $\forall \tilde{\boldsymbol{q}} \in \tilde{\Delta}_k \setminus \{\tilde{\boldsymbol{\mu}}\}$. In particular, for $\tilde{\boldsymbol{q}} = \Pi_k(\boldsymbol{e}_{\theta_*})$, we get $D_\Psi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) = d_\Psi(\tilde{\boldsymbol{q}}) > d_{\mathrm{S}}(\tilde{\boldsymbol{q}}) = D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})$, which contradicts (101). Therefore, $\eta_\ell^{\Psi} \leq \eta_\ell^{\mathrm{S}}$, and thus

$$\begin{aligned} R_\ell^{\mathrm{S}}(\boldsymbol{\mu}) = \max_\theta D_{\mathrm{S}}(\boldsymbol{e}_\theta, \boldsymbol{\mu})/\eta_\ell^{\mathrm{S}} &= D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})/\eta_\ell^{\mathrm{S}}, \\ &\leq D_\Psi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})/\eta_\ell^{\Psi}, && (102) \\ &\leq \max_\theta D_\Psi(\boldsymbol{e}_\theta, \boldsymbol{\mu})/\eta_\ell^{\Psi}, \\ &= R_\ell^{\Psi}(\boldsymbol{\mu}), && (103) \end{aligned}$$

where (102) is due to $D_\Psi(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu}) = D_{\mathrm{S}}(\boldsymbol{e}_{\theta_*}, \boldsymbol{\mu})$ and $\eta_\ell^{\Psi} \leq \eta_\ell^{\mathrm{S}}$. Equation 103, implies that $R_\ell^{\mathrm{S}}(\boldsymbol{\mu}) \leq R_\ell^{\Phi}(\boldsymbol{\mu})$, since $R_\ell^{\Psi}(\boldsymbol{\mu}) = R_\ell^{\alpha\Phi}(\boldsymbol{\mu}) = R_\ell^{\Phi}(\boldsymbol{\mu})$ [13]. Therefore,

$$\forall \boldsymbol{\mu} \in \operatorname{ri} \Delta_k,\ R_\ell^{\mathrm{S}}(\boldsymbol{\mu}) \leq R_\ell^{\Phi}(\boldsymbol{\mu}). \tag{104}$$

It remains to consider the case where $\boldsymbol{\mu}$ is in the relative boundary of Δ_k. Let $\boldsymbol{\mu} \in \operatorname{rbd} \Delta_k$. There exists $\mathfrak{l}_0 \subsetneq [k]$ such that $\boldsymbol{\mu} \in \Delta_{\mathfrak{l}_0}$. Let $\theta^* \in [k] \setminus \mathfrak{l}_0$ and $\mathfrak{l} \coloneqq \mathfrak{l}_0 \cup \{\theta^*\}$. It holds that $\boldsymbol{\mu} \in \operatorname{rbd} \Delta_{\mathfrak{l}}$ and $\boldsymbol{\mu} + 2^{-1}(\boldsymbol{e}_{\theta^*} - \boldsymbol{\mu}) \in \operatorname{ri} \Delta_{\mathfrak{l}}$. Since ℓ is Φ-mixable, it follows from Proposition 29 and the 1-homogeneity of $\Phi'(\boldsymbol{\mu}; \cdot)$ [8, Prop. D.1.1.2] that

$$\Phi'(\boldsymbol{\mu}; \boldsymbol{e}_{\theta^*} - \boldsymbol{\mu}) = 2\Phi'(\boldsymbol{\mu}; [\boldsymbol{\mu} + 2^{-1}(\boldsymbol{e}_{\theta^*} - \boldsymbol{\mu})] - \boldsymbol{\mu}) = -\infty.$$

Hence,

$$\begin{aligned} R_\ell^{\Phi}(\boldsymbol{\mu}) &= \max_{\theta \in [k]} D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{\mu}), \\ &\geq D_\Phi(\boldsymbol{e}_{\theta^*}, \boldsymbol{\mu}) = \Phi(\boldsymbol{e}_{\theta^*}) - \Phi(\boldsymbol{\mu}) - \Phi'(\boldsymbol{\mu}; \boldsymbol{e}_{\theta^*} - \boldsymbol{\mu}) = +\infty. && (105) \end{aligned}$$

Inequality 105 also applies to S, since ℓ is $(\underline{\eta_\ell}^{-1}\,\mathrm{S})$-mixable. From (105) and (104), we conclude that $\forall \boldsymbol{\mu} \in \Delta_k,\ R_\ell^{\mathrm{S}}(\boldsymbol{\mu}) \leq R_\ell^{\Phi}(\boldsymbol{\mu})$. □

C.11 Proof of Theorem 19

Theorem 19 *Let $\Phi : \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ be a Δ-differentiable entropy. Let $\ell \colon \mathcal{A} \to [0, +\infty]^n$ be a loss such that $\underline{L}_\ell$ is twice differentiable on $]0, +\infty[^n$. Let $\boldsymbol{\beta}^t = -\eta \sum_{s=1}^{t-1} (\ell_{x^s}(A^s) + \boldsymbol{v}^s)$, where $\boldsymbol{v}^s \in \mathbb{R}^k$ and $A^s := \boldsymbol{a}^s_{1:k} \in \mathcal{A}^k$. If ℓ is (η, Φ)-mixable then for initial distribution $\boldsymbol{q}^0 = \operatorname{argmin}_{\boldsymbol{q} \in \Delta_k} \max_{\theta \in [k]} D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q})$ and any sequence $(x^t, \boldsymbol{a}^t_{1:k})_{t=1}^T$, the AGAA achieves the regret*

$$\forall \theta \in [k], \quad \mathrm{Loss}^\ell_{AGAA}(T) - \mathrm{Loss}^\ell_\theta(T) \le R^\Phi_\ell + \sum_{t=1}^{T-1} (v^t_\theta - \langle \boldsymbol{v}^t, \boldsymbol{q}^t \rangle).$$

Proof. Recall that $\Phi_t(\boldsymbol{w}) := \Phi(\boldsymbol{w}) - \langle \boldsymbol{w}, \boldsymbol{\beta}^t - \boldsymbol{\theta}^t \rangle$, where $\boldsymbol{\theta}^t = -\eta \sum_{s=1}^{t-1} \ell_{x^s}(A^s)$. From Theorem 16 and since Φ_t is equal to Φ plus an affine function, it is clear that if ℓ is (η, Φ)-mixable then ℓ is (η, Φ_t)-mixable. Thus, for all $(A^t, \boldsymbol{q}^{t-1}) \in \mathcal{A}^k \times \Delta_k$, there exists $\boldsymbol{a}^t_* \in \mathcal{A}$ such that for any outcome $x^t \in [n]$

$$\ell_{x^t}(\boldsymbol{a}^t_*) \le \eta^{-1}[\Phi^\star_t(\nabla \Phi_t(\boldsymbol{q}^{t-1})) - \Phi^\star_t(\nabla \Phi_t(\boldsymbol{q}^{t-1}) - \eta \ell_{x^t}(A^t))].$$

Summing over t from 1 to T, we get

$$\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_*) \le \eta^{-1}[\Phi^\star_1(\nabla \Phi_1(\boldsymbol{q}^0)) - \Phi^\star_T(\nabla \Phi_T(\boldsymbol{q}^{T-1}) - \eta \ell_{x^T}(A^T))] \tag{106}$$
$$+ \eta^{-1} \sum_{t=1}^{T-1} \left[\Phi^\star_{t+1}(\nabla \Phi_{t+1}(\boldsymbol{q}^t)) - \Phi^\star_t(\nabla \Phi_t(\boldsymbol{q}^{t-1}) - \eta \ell_{x^t}(A^t))\right].$$

ODue to the properties of the entropic dual [13] and the definition of Φ_t, the following holds for all $t \in [T]$ and $\boldsymbol{z}$ in $\mathbb{R}^k$,

$$\nabla \Phi_t(\boldsymbol{q}^{t-1}) = -\eta \sum_{s=1}^{t-1} \ell_{x^s}(A^s), \tag{107}$$

$$\Phi^\star_t(\boldsymbol{z}) = \Phi^\star(\boldsymbol{z} + \nabla \Phi(\boldsymbol{q}^{t-1}) + \eta \sum_{s=1}^{t-1} \ell_{x^s}(A^s)), \tag{108}$$

$$\nabla \Phi(\boldsymbol{q}^t) = \nabla \Phi(\boldsymbol{q}^{t-1}) - \eta \ell_{x^t}(A^t) - \eta \boldsymbol{v}^t. \tag{109}$$

Using (107)-(108), we get for all $0 \le t < T$, $\Phi^\star_{t+1}(\nabla \Phi_{t+1}(\boldsymbol{q}^t)) = \Phi^\star(\nabla \Phi(\boldsymbol{q}^t))$, and in particular $\Phi^\star_1(\nabla \Phi_1(\boldsymbol{q}^0)) = \Phi^\star(\nabla \Phi(\boldsymbol{q}^0))$. Similarly, using (107)-(109), gives $\Phi^\star_t(\nabla \Phi_t(\boldsymbol{q}^{t-1}) - \eta \ell_{x^t}(A^t)) = \Phi^\star(\nabla \Phi(\boldsymbol{q}^t) + \eta \boldsymbol{v}^t)$ for all $1 \le t \le T$. Substituting back into (106) yields

$$\sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_*) \le \eta^{-1}[\Phi^\star(\nabla \Phi(\boldsymbol{q}^0)) - \Phi^\star(\nabla \Phi(\boldsymbol{q}^T) + \eta \boldsymbol{v}^T)]$$
$$+ \eta^{-1} \sum_{t=1}^{T-1} \left[\Phi^\star(\nabla \Phi(\boldsymbol{q}^t)) - \Phi^\star(\nabla \Phi(\boldsymbol{q}^t) + \eta \boldsymbol{v}^t)\right], \tag{110}$$

To conclude the proof, we note that since Φ is convex it holds that

$$\Phi^\star(\nabla \Phi(\boldsymbol{q}^t)) - \Phi^\star(\nabla \Phi(\boldsymbol{q}^t) + \eta \boldsymbol{v}^t) \le -\eta \langle \boldsymbol{v}^t, \nabla \Phi^\star(\nabla \Phi(\boldsymbol{q}^t)) \rangle = -\eta \langle \boldsymbol{v}^t, \boldsymbol{q}^t \rangle, \tag{111}$$

which allows us to bound the sum on the right hand side of 110. To bound the rest of the terms, we use the fact that $\nabla \Phi(\boldsymbol{q}^T) = \nabla \Phi(\boldsymbol{q}^0) - \eta \sum_{t=1}^T (\ell_{x^t}(A^t) + \boldsymbol{v}^t)$, and thus by letting $\Phi_\eta := \eta^{-1} \Phi$,

$$\begin{aligned}
\eta^{-1}[\Phi^\star(\nabla \Phi(\boldsymbol{q}^0)) - \Phi^\star(\nabla \Phi(\boldsymbol{q}^T) + \eta \boldsymbol{v}^T)] &= \Phi^\star_\eta(\nabla \Phi_\eta(\boldsymbol{q}^0)) \\
&\quad - \Phi^\star_\eta\left(\nabla \Phi_\eta(\boldsymbol{q}^0) - \sum_{t=1}^T \ell_{x^t}(A^t) - \sum_{t=1}^{T-1} \boldsymbol{v}^t\right), \\
&= \inf_{q \in \Delta_k} \left\langle \boldsymbol{q}, \sum_{t=1}^T \ell_{x^t}(A^t) + \sum_{t=1}^{T-1} \boldsymbol{v}^t \right\rangle + \frac{D_\Phi(\boldsymbol{q}, \boldsymbol{q}^0)}{\eta}, \\
&\le \sum_{t=1}^T \ell_{x^t}(\boldsymbol{a}^t_\theta) + \sum_{t=1}^{T-1} v^t_\theta + \frac{D_\Phi(\boldsymbol{e}_\theta, \boldsymbol{q}^0)}{\eta}, \forall \theta \in [k].
\end{aligned}$$

Substituting this last inequality and (111) back into (110) yields the desired bound. □

D Defining the Bayes Risk Using the Superprediction Set

In this section, we argue that when a loss $\ell\colon \mathcal{A} \to [0,+\infty]^n$ is mixable, in the classical or generalized sense, it does not matter whether we define the Bayes risk $\underline{L}_\ell$ using the full superprediction set $\mathscr{S}_\ell^\infty$ or its finite part $\mathscr{S}_\ell$. Recall the definition of the Bayes risk;

Definition 2 *Let $\ell\colon \mathcal{A} \to [0,+\infty]^n$ be a loss such that $\operatorname{dom}\ell \neq \varnothing$. The* Bayes risk $\underline{L}_\ell : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ *is defined by*

$$\forall \boldsymbol{u} \in \mathbb{R}^n, \quad \underline{L}_\ell(\boldsymbol{u}) \coloneqq \inf_{\boldsymbol{z} \in \mathscr{S}_\ell} \langle \boldsymbol{u}, \boldsymbol{z} \rangle . \tag{112}$$

Note that the right hand side of (112) does not change if we substitute $\mathscr{S}_\ell$ for its closure — $\overline{\mathscr{S}_\ell}$ — with respect to $[0,+\infty]^n$. Thus, it suffices to show that $\mathscr{S}_\ell^\infty \subseteq \overline{\mathscr{S}_\ell}$ when the loss ℓ is mixable. We show this in Theorem 45, but first we give a characterization of the (finite part) of the superprediction set for a proper loss.

Proposition 44. *Let $\ell\colon \Delta_n \to [0,+\infty]^n$ be a proper loss. If $\underline{L}_\ell$ is differentiable on $]0,+\infty[^n$, then*

$$\overline{\mathscr{S}_\ell} \supseteq \mathcal{C}_\ell \coloneqq \{\boldsymbol{u} \in [0,+\infty]^n \colon \forall \boldsymbol{p} \in \Delta_n, \underline{L}_\ell(\boldsymbol{p}) \leq \langle \boldsymbol{p}, \boldsymbol{u} \rangle\}. \tag{113}$$

Proof. Let $\boldsymbol{v} \in \mathcal{C}_\ell \cap [0,+\infty[^n$. Let $f\colon \operatorname{ri}\Delta_n \times [n] \to \mathbb{R}$ be defined by $f(\boldsymbol{p}, x) \coloneqq \ell_x(\boldsymbol{p}) - v_x$. By the choice of $\boldsymbol{v}$, we have $\mathbb{E}_{x\sim\boldsymbol{p}} f(\boldsymbol{p}, x) = \langle \boldsymbol{p}, \ell(\boldsymbol{p}) \rangle - \langle \boldsymbol{p}, \boldsymbol{v} \rangle \leq 0$ for all $\boldsymbol{p} \in \Delta_n$. Since $\underline{L}_\ell$ is differentiable on $]0,+\infty[^n$, by assumption, ℓ is continuous on $\operatorname{ri}\Delta_n$, and thus f is continuous in the first argument. Since $\boldsymbol{v}$ has finite components, the map f satisfies all the conditions of Lemma 24. Therefore, there exists $(\boldsymbol{p}_m) \subset \operatorname{ri}\Delta_n$ such that

$$\forall m \in \mathbb{N}, \forall x \in [n],\ \ell_x(\boldsymbol{p}_m) \leq v_x + \frac{1}{m}. \tag{114}$$

Without loss of generality, we can assume by extracting a subsequence if necessary that $\ell(\boldsymbol{p}_m)$ converges to $\boldsymbol{s} \in [0,+\infty]^n$. By definition, we have $\boldsymbol{s} \in \overline{\mathscr{S}_\ell}$, and from (114) it follows that $\boldsymbol{s} \leq \boldsymbol{v}$ coordinate-wise. Thus, $\boldsymbol{v}$ is in $\overline{\mathscr{S}_\ell}$.

The above argument shows that $\mathcal{C}_\ell \cap [0,+\infty[^n \subseteq \overline{\mathscr{S}_\ell}$, and since $\overline{\mathscr{S}_\ell}$ is closed in $[0,+\infty]^n$ we have $\overline{\mathcal{C}} \subseteq \overline{\mathscr{S}_\ell}$, where $\overline{\mathcal{C}}$ is the closure of $\mathcal{C}_\ell \cap [0,+\infty[^n$ in $[0,+\infty]^n$. Now it suffice to show that $\mathcal{C}_\ell \subseteq \overline{\mathcal{C}}$ to complete the proof.

Let $\boldsymbol{u} \in \mathcal{C}_\ell$ and $\mathfrak{l} \coloneqq \{x \in [n] \colon u_x < +\infty\}$. Define $(\boldsymbol{u}_m) \subset [0,+\infty[^n$ by $u_{m,x} = u_x$ if $x \in \mathfrak{l}$; and m otherwise. Let $\boldsymbol{p} \in \Delta_n$. It follows that

$$\begin{aligned} \langle \boldsymbol{p}, \boldsymbol{u}_m \rangle &= \sum_{x' \in \mathfrak{l}} p_{x'} u_{m,x'} + \sum_{x \notin \mathfrak{l}} p_x u_{m,x}, \\ &= \sum_{x' \in \mathfrak{l}} p_{x'} u_{m,x'} + \sum_{x \notin \mathfrak{l}} p_x u_{m,x}, . \end{aligned} \tag{115}$$

Claim 2. $\forall \epsilon > 0, \exists m_\epsilon \geq 1, \forall \boldsymbol{p} \in \Delta_k, \underline{L}_\ell(\boldsymbol{p}) \leq \langle \boldsymbol{p}, \boldsymbol{u}_{m_\epsilon} \rangle - \epsilon.$

Suppose that Claim 2 is false. This means that there exists $\delta > 0$ such that

$$\forall \boldsymbol{m} \geq 1, \exists \boldsymbol{p}_m \in \Delta_n, \langle \boldsymbol{p}_m, \boldsymbol{u}_m \rangle - \delta < \underline{L}_\ell(\boldsymbol{p}_m). \tag{116}$$

We may assume, by extracting a subsequence if necessary (Δ_n is compact), that $(\boldsymbol{p}_m)$ converges to $\boldsymbol{p}_* \in \Delta_n$. Taking the limit $m \to \infty$ in (116) would lead to the contradiction '$\langle \boldsymbol{p}_*, \boldsymbol{u} \rangle < \underline{L}_\ell(\boldsymbol{p}_*)$', since from (115) we have $\lim_{m\to\infty} \langle \boldsymbol{p}_m, \boldsymbol{u}_m \rangle = \langle \boldsymbol{p}_*, \boldsymbol{u} \rangle$. Therefore, Claim 2 is true. For $\epsilon = \frac{1}{k}$ let $m_k \coloneqq m_\epsilon$ be as in Claim (2). The claim then implies that $\liminf_{k\to\infty} \langle \boldsymbol{p}, \boldsymbol{u}_{m_k} \rangle \geq \underline{L}_\ell(\boldsymbol{p})$ uniformly for $\boldsymbol{p} \in \Delta_k$. By the claim we also have that $\boldsymbol{u}_{m_k} \in \mathcal{C}_\ell \cap [0,+\infty[^n$ for all $k \in \mathbb{N}$, and by construction of $\boldsymbol{v}_m$, we have $\lim_{k\to\infty} \boldsymbol{u}_{m_k} = \boldsymbol{u}$. This shows that $\mathcal{C}_\ell \subseteq \overline{\mathcal{C}}$, which completes the proof.

□

Theorem 45. *Let $\ell : \mathcal{A} \to [0,+\infty]^n$ be a loss. If $\mathscr{S}_\ell^\infty \not\subseteq \overline{\mathscr{S}_\ell}$, then ℓ is not mixable.*

Proof. Suppose that ℓ is mixable and let $\underline{\ell}$ be a proper support loss of ℓ. From Proposition 12, $\underline{L}_\ell$ is differentiable on $]0,+\infty[^n$, and thus Theorem 5 implies that $\overline{\mathscr{S}_\ell} = \overline{\mathscr{S}_{\underline{\ell}}}$. Therefore, Lemma 44 implies that $\overline{\mathscr{S}_\ell} \supseteq \{\boldsymbol{u} \in [0,+\infty]^n \colon \forall \boldsymbol{p} \in \Delta_n, \underline{L}_\ell(\boldsymbol{p}) \leq \langle \boldsymbol{p}, \boldsymbol{u}\rangle\}$. Thus, if $\mathscr{S}_\ell^\infty \not\subseteq \overline{\mathscr{S}_\ell}$, there exists $\epsilon > 0$, $\boldsymbol{p}_\epsilon \in \Delta_k$, and $\boldsymbol{s} \in \mathscr{S}_\ell^\infty \setminus \overline{\mathscr{S}_\ell}$ such that

$$\langle \boldsymbol{p}_\epsilon, \boldsymbol{s}\rangle < \underline{L}_\ell(\boldsymbol{p}_\epsilon) - 2\epsilon. \tag{117}$$

Note that $\boldsymbol{p}_\epsilon$ cannot be in $\operatorname{ri}\Delta_n$; otherwise, (117) would imply that $\boldsymbol{s}$ has all finite components, and thus would be included in $\overline{\mathscr{S}_\ell}$, which is a contradiction. Assume from now on that $\boldsymbol{p}_\epsilon \in \operatorname{rbd}\Delta_n$. From the definition of the support loss, there exists a sequence $(\boldsymbol{p}_m) \subseteq \operatorname{ri}\Delta_n$ such that $\boldsymbol{p}_m \overset{m\to\infty}{\to} \boldsymbol{p}_\epsilon$ and $\underline{\ell}(\boldsymbol{p}_m) \overset{m\to\infty}{\to} \underline{\ell}(\boldsymbol{p}_\epsilon)$. Therefore, Theorem 5 implies that there exists $\boldsymbol{a}_\epsilon \in \mathcal{A}$ such that

$$\langle \boldsymbol{p}_\epsilon, \ell(\boldsymbol{a}_\epsilon)\rangle < \langle \boldsymbol{p}_\epsilon, \underline{\ell}(\boldsymbol{p}_\epsilon)\rangle + \epsilon. \tag{118}$$

To see this, note that since $(\boldsymbol{p}_m) \subset \operatorname{ri}\Delta_n \subseteq \operatorname{dom}\underline{\ell}$, Theorem 5 guarantees the existence of a sequence $(\boldsymbol{a}_m) \subset \mathcal{A}$ such that $\ell(\boldsymbol{a}_m) = \underline{\ell}(\boldsymbol{p}_m)$. On the other hand, for any $x \in [n]$ such that $\ell_x(\boldsymbol{p}_\epsilon) = +\infty$, we have $p_{\epsilon,x} = 0$ — otherwise, $\underline{L}_\ell(\boldsymbol{p}_\epsilon)$ would be infinite. It follows, by continuity of the inner product that $\langle \boldsymbol{p}_\epsilon, \ell(\boldsymbol{a}_m)\rangle \overset{m\to\infty}{\to} \langle \boldsymbol{p}_\epsilon, \underline{\ell}(\boldsymbol{p}_m)\rangle$, and thus it suffices to pick $\boldsymbol{a}_\epsilon$ equal to $\boldsymbol{a}_m$ for m large enough.

Now since ℓ is η-mixable, there exists $\eta > 0$ and $\boldsymbol{a}_* \in \mathcal{A}$ such that

$$\ell(\boldsymbol{a}_*) \leq -\eta^{-1}\log\left(\frac{1}{2}e^{-\eta \boldsymbol{s}} + \frac{1}{2}e^{-\eta\ell(\boldsymbol{a}_\epsilon)}\right),$$

and due to the convexity of $-\log$,

$$\leq \frac{1}{2}\boldsymbol{s} + \frac{1}{2}\ell(\boldsymbol{a}_\epsilon).$$

Using (117) and (118) yields

$$\langle \boldsymbol{p}_\epsilon, \ell(\boldsymbol{a}_*)\rangle \leq \underline{L}_\ell(\boldsymbol{p}_\epsilon) - \epsilon/2. \tag{119}$$

On the other hand, by definition of a proper support loss, $\langle \boldsymbol{p}_\epsilon, \underline{\ell}(\boldsymbol{p}_\epsilon)\rangle \leq \langle \boldsymbol{p}_\epsilon, \ell(\boldsymbol{a}_*)\rangle$. This combined with (119), lead to the contradiction $\langle \boldsymbol{p}_\epsilon, \underline{\ell}(\boldsymbol{p}_\epsilon)\rangle < \underline{L}_\ell(\boldsymbol{p}_\epsilon)$. □

E The Update Step of the GAA and the Mirror Descent Algorithm

In this section, we demonstrate that the update steps of the GAA and the Mirror Descent Algorithm are essentially the same (at least for finite losses) according to the definition of the MDA given by Beck and Teboulle [2];

Let $\ell \colon \mathcal{A} \to [0,+\infty[^n$ be a loss and $\Phi \colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ an entropy such that $\tilde{\Phi}$ is differentiable on $\operatorname{int}\tilde{\Delta}_k$. Let $\boldsymbol{q}^t$ be the update distribution of the GAA at round t and $\tilde{\boldsymbol{q}}^t = \Pi_k(\boldsymbol{q}^t)$. It follows from the definition of $\boldsymbol{q}^t$ (see Algorithm 2) that

$$\begin{aligned}
\tilde{\boldsymbol{q}}^t &= \operatorname*{argmin}_{\tilde{\boldsymbol{q}}\in\tilde{\Delta}_k} \left\langle \Pi_k(\tilde{\boldsymbol{q}}), \ell_{x^t}(A^t)\right\rangle + \eta^{-1}D_{\tilde{\Phi}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}^{t-1}),\\
&= \operatorname*{argmin}_{\tilde{\boldsymbol{q}}\in\tilde{\Delta}_k} \left\langle \tilde{\boldsymbol{q}}, J_k^\top \ell_{x^t}(A^t)\right\rangle + \eta^{-1}D_{\tilde{\Phi}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}^{t-1}),\\
&= \operatorname*{argmin}_{\tilde{\boldsymbol{q}}\in\tilde{\Delta}_k} \left\langle \tilde{\boldsymbol{q}}, \nabla l_t(\tilde{\boldsymbol{q}}^{t-1})\right\rangle + \eta^{-1}D_{\tilde{\Phi}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{q}}^{t-1}),
\end{aligned} \tag{120}$$

where $l_t(\tilde{\boldsymbol{\mu}}) \coloneqq \langle \Pi_k(\tilde{\boldsymbol{\mu}}), \ell_{x^t}(A^t)\rangle = \langle \boldsymbol{\mu}, \ell_{x^t}(A^t)\rangle$. Update (120) is, by definition [2], the MDA with the sequence of losses l_t on $\operatorname{int}\tilde{\Delta}_k$, 'distance' function $D_{\tilde{\Phi}}(\cdot,\cdot)$, and learning rate η. Therefore, the MDA is exactly the update step of the GAA.

F The Generalized Aggregating Algorithm Using the Shannon Entropy S

The purpose of this appendix is to show that the GAA reduces to the AA when the former uses the Shannon entropy. In this case, generalized and classical mixability are equivalent. In what follows, we make use of the following proposition which is proved in C.5.

Proposition 39 *For the Shannon entropy* S*, it holds that* $\tilde{\mathrm{S}}^*(\boldsymbol{v}) = \log(\langle \exp(\boldsymbol{v}), \mathbf{1}_{\tilde{k}}\rangle + 1), \forall \boldsymbol{v} \in \mathbb{R}^{k-1}$*, and* $\mathrm{S}^\star(\boldsymbol{z}) = \log\langle \exp(\boldsymbol{z}), \mathbf{1}_k\rangle, \forall \boldsymbol{z} \in \mathbb{R}^k$*.*

Let $\ell\colon \mathcal{A} \to [0, +\infty[^n$ be a loss and Φ be as in Proposition 33 and suppose that Φ and $\tilde{\Phi}^*$ are differentiable on $\operatorname{ri}\Delta_k$ and $\mathbb{R}^{k-1}$, respectively. It was shown in [13] that

$$\nabla\Phi^\star(\nabla\Phi(\boldsymbol{q}) - \ell_x(A)) = \operatorname*{argmin}_{\boldsymbol{\mu}\in\Delta_k} \langle \boldsymbol{\mu}, \ell_x(A)\rangle + D_\Phi(\boldsymbol{\mu}, \boldsymbol{q}), \tag{121}$$

$$\operatorname{Mix}_\Phi(\ell_x(A), \boldsymbol{q}) = \Phi^\star(\nabla\Phi(\boldsymbol{q})) - \Phi^\star(\nabla\Phi(\boldsymbol{q}) - \ell_x(A)). \tag{122}$$

Let $\boldsymbol{q} \in \operatorname{ri}\Delta_k$. By definition of S, $\nabla\,\mathrm{S}(\boldsymbol{q}) = \log \boldsymbol{q} + \mathbf{1}_k$, and due to Proposition 39, $\mathrm{S}^\star(\boldsymbol{z}) = \log\langle \exp \boldsymbol{z}, \mathbf{1}_k\rangle, \boldsymbol{z} \in \mathbb{R}^k$. Therefore, $\nabla\,\mathrm{S}(\boldsymbol{q}) - \eta\ell_x(A) = \log(\exp(-\eta\ell_x(A)) \odot \boldsymbol{q}) + \mathbf{1}_k$ and $\nabla\,\mathrm{S}^\star(\boldsymbol{z}) = \frac{\exp \boldsymbol{z}}{\langle \exp \boldsymbol{z}, \mathbf{1}_k\rangle}, \forall (x, A) \in [n] \times (\operatorname{dom}\ell)^k$. Thus,

$$\nabla\mathrm{S}^\star(\nabla\,\mathrm{S}(\boldsymbol{q}) - \eta\ell_x(A)) = \frac{\exp(-\eta\ell_x(A)) \odot \boldsymbol{q}}{\langle \exp(-\eta\ell_x(A)), \boldsymbol{q}\rangle}. \tag{123}$$

Let $\mathrm{S}_\eta \coloneqq \eta^{-1}\,\mathrm{S}$. Then $\nabla\,\mathrm{S} = \eta\nabla\,\mathrm{S}_\eta$ and $\forall \boldsymbol{z} \in \mathbb{R}^k, \nabla\,\mathrm{S}^\star_\eta(\boldsymbol{z}) = \nabla\,\mathrm{S}^\star(\eta\boldsymbol{z})$ [13].[2] Then the left hand side of (123) can be written as $\nabla\,\mathrm{S}^\star_\eta(\nabla\,\mathrm{S}_\eta(\boldsymbol{q}) - \ell_x(A))$. Using this fact, (121) and (123) show that the update distribution $\boldsymbol{q}^t$ of the GAA (Algorithm 2) coincides with that of the AA after substituting $\boldsymbol{q}, x$, and A by $\boldsymbol{q}^{t-1}, x^t$, and $A^t \coloneqq [\boldsymbol{a}_\theta]_{\theta\in[k]}$, respectively.

Now using the fact that $\operatorname{Mix}^\eta_{\mathrm{S}}(\ell_x(A), \boldsymbol{q}) = \eta^{-1}\operatorname{Mix}_{\mathrm{S}}(\eta\ell_x(A), \boldsymbol{q})$ [13] and (122), we get

$$\begin{aligned}\operatorname{Mix}^\eta_{\mathrm{S}}(\ell_x(A), \boldsymbol{q}) &= \eta^{-1}[\mathrm{S}^\star(\nabla\,\mathrm{S}(\boldsymbol{q})) - \mathrm{S}^\star(\nabla\,\mathrm{S}(\boldsymbol{q}) - \eta\ell_x(A))],\\ &= -\eta^{-1}\log\langle \exp(-\eta\ell_x(A)), \boldsymbol{q}\rangle.\end{aligned} \tag{124}$$

Equation 124 shows that the η-mixability condition is equivalent to the (η, S)-mixability condition for a finite loss. This remains true for losses taking infinite values — see the proof of Theorem 11 in Appendix C.5.

G Legendre Φ, but no Φ-mixable ℓ

In this appendix, we construct a *Legendre type* entropy [15] for which there are no Φ-mixable losses satisfying a weak condition (see below).

Let $\ell : \mathcal{A} \to [0, +\infty]^n$ be a loss satisfying condition 1. According to Alexandrov's Theorem, a concave function is twice differentiable almost everywhere (see e.g. [5, Thm. 6.7]). Now we give a version of Theorem 14 which does not assume the twice differentiability of the Bayes risk. The proof is almost identical to that of Theorem 14 with only minor modifications.

Theorem 46. *Let* $\Phi\colon \mathbb{R}^k \to \mathbb{R} \cup \{+\infty\}$ *be an entropy such that* $\tilde{\Phi}$ *is twice differentiable on* $\operatorname{int}\tilde{\Delta}_k$*, and* $\ell\colon \mathcal{A} \to [0, +\infty]^n$ *a loss satisfying Condition 1 and such that* $\exists(\tilde{\boldsymbol{p}}, \boldsymbol{v}) \in \mathcal{D} \times \mathbb{R}^{\tilde{n}}, \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v} \neq \mathbf{0}_{\tilde{n}}$*, where* $\mathcal{D} \subset \operatorname{int}\tilde{\Delta}_n$ *is a set of Lebesgue measure 1 where* $\underline{\tilde{L}}_\ell$ *is twice differentiable, and define*

$$\underline{\eta_\ell}^* \coloneqq \inf_{\tilde{\boldsymbol{p}}\in\mathcal{D}} (\lambda_{\max}([\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})]^{-1}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})))^{-1}. \tag{125}$$

Then ℓ *is* Φ*-mixable only if* $\underline{\eta_\ell}^*\Phi - \mathrm{S}$ *is convex on* Δ_k*.*

The new condition on the Bayes risk is much weaker than requiring $\underline{L}_\ell$ to be twice differentiable on $]0, +\infty[^n$. In the next example, we will show that there exists a Legendre type entropy for which there are no Φ-mixable losses satisfying the condition of Theorem 46.

Example 2. *Let* $\Phi : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ *be an entropy such that*

$$\forall q \in]0, 1[,\ \Phi(q, 1-q) = \tilde{\Phi}(q) = \int_{1/2}^q \log\left(\frac{\log(1-t)}{\log t}\right) dt.$$

[2] Reid et al. [13] showed the equality $\nabla\Phi^\star_\eta(\boldsymbol{u}) = \nabla\Phi^\star(\eta\boldsymbol{u}), \forall \boldsymbol{u} \in \operatorname{dom}\Phi^\star$, for any entropy differentiable on Δ_k - not just for the Shannon Entropy.

$\tilde{\Phi}$ is differentiable and strictly convex on the open set $(0,1)$. Furthermore, it satisfies (22) *which makes it a function of Legendre type [15, Lem. 26.2]. In fact,* (22) *is satisfied due to*

$$\left|\frac{d}{dq}\tilde{\Phi}(q)\right| = \left|\log\left(\frac{\log(1-q)}{\log q}\right)\right| \overset{q\to b}{\to} +\infty, \text{ where } b \in \{0,1\},$$
$$\frac{d^2}{dq^2}\tilde{\Phi}(q) = \frac{-1}{q\log q} + \frac{-1}{(1-q)\log(1-q)} > 0, \ \forall q \in]0,1[.$$

The Shannon entropy on Δ_2 is defined by $\mathrm{S}(q,1-q) = \tilde{\mathrm{S}}(q) = q\log q + (1-q)\log(1-q)$, for $q \in]0,1[$. Thus, $\frac{d^2}{dq^2}\tilde{\mathrm{S}}(q) = \frac{1}{q(1-q)}$.

Suppose now that there exists a Φ-mixable loss $\ell\colon \mathcal{A} \to [0,+\infty]^n$ satisfying condition 1 and such that $\exists(\tilde{\boldsymbol{p}},\boldsymbol{v}) \in \mathcal{D}\times\mathbb{R}^{\tilde{n}}, \mathsf{H}\tilde{\underline{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v} \neq \mathbf{0}_{\tilde{n}}$. Let $\underline{\eta_\ell}^*$ be as in (125)*. By definition, we have $\underline{\eta_\ell}^* < +\infty$, and thus*

$$\underline{\eta_\ell}^*\left[\frac{d^2}{dq^2}\tilde{\Phi}(q)\right]\left[\frac{d^2}{dq^2}\tilde{\mathrm{S}}(q)\right]^{-1} = \underline{\eta_\ell}^*\left(\frac{q-1}{\log q} + \frac{-q}{\log(1-q)}\right) \overset{q\to b}{\to} 0, \tag{126}$$

where $b \in \{0,1\}$. From Lemma 22, (126) *implies that $\underline{\eta_\ell}^*\Phi - \mathrm{S}$ is not convex on Δ_k, which is a contradiction according to Theorem 46.*

H Loss Surface and Superprediction Set

In this appendix, we derive an expression for the curvature of the image of a proper loss function. We will need the following lemma.

Lemma 47. *Let $\sigma : [0,+\infty[^n \to \mathbb{R}$ be a 1-homogeneous, twice differentiable function on $]0,+\infty[^n$. Then σ is concave on $]0,+\infty[^n$ if and only if $\tilde{\sigma} = \sigma \circ \Pi_n$ is concave on $\operatorname{int}\tilde{\Delta}_n$.*

Proof. The forward implication is immediate; if σ is concave on $]0,+\infty[^n$, then $\sigma\circ\Pi_k$ is concave on $\operatorname{int}\tilde{\Delta}_k$, since Π_k is an affine function.

Now assume that $\tilde{\sigma}$ is concave on $\operatorname{int}\tilde{\Delta}_k$. Let $\lambda \in [0,1]$ and $(\boldsymbol{p},\boldsymbol{q}) \in [0,+\infty[^n \times [0,+\infty[^n$. We need to show that

$$\lambda\sigma(\boldsymbol{p}) + (1-\lambda)\sigma(\boldsymbol{q}) \leq \sigma(\lambda\boldsymbol{p} + (1-\lambda)\boldsymbol{q}). \tag{127}$$

Note that if $\boldsymbol{p} = \mathbf{0}$ or $\boldsymbol{q} = \mathbf{0}$, (127) is trivially with equality due to the 1-homogeneity of σ. Now assume that $\boldsymbol{p}$ and $\boldsymbol{q}$ are non-zero and let $c \coloneqq \lambda\|\boldsymbol{p}\|_1 + (1-\lambda)\|\boldsymbol{q}\|_1$. For convenience, we also denote $\boldsymbol{p}_1 = \boldsymbol{p}/\|\boldsymbol{p}\|_1$ and $\boldsymbol{q}_1 = \boldsymbol{q}/\|\boldsymbol{q}\|_1$ which are both in Δ_n. It follows that

$$\begin{aligned}
\lambda\sigma(\boldsymbol{p}) + (1-\lambda)\sigma(\boldsymbol{q}) &= cM\left(\lambda\frac{\|\boldsymbol{p}\|_1}{c}\sigma(\boldsymbol{p}_1) + (1-\lambda)\frac{\|\boldsymbol{q}\|_1}{c}\sigma(\boldsymbol{q}_1)\right),\\
&= c\left(\lambda\frac{\|\boldsymbol{p}\|_1}{c}\tilde{\sigma}(\tilde{\boldsymbol{p}}_1) + (1-\lambda)\frac{\|\boldsymbol{q}\|_1}{c}\tilde{\sigma}(\tilde{\boldsymbol{q}}_1)\right),\\
&\leq c\tilde{\sigma}\left(\lambda\frac{\|\boldsymbol{p}\|_1}{c}\tilde{\boldsymbol{p}}_1 + (1-\lambda)\frac{\|\boldsymbol{q}\|_1}{c}\tilde{\boldsymbol{q}}_1\right),\\
&= c\sigma\left(\lambda\frac{\|\boldsymbol{p}\|_1}{c}\boldsymbol{p}_1 + (1-\lambda)\frac{\|\boldsymbol{q}\|_1}{c}\boldsymbol{q}_1\right),\\
&= \sigma(\lambda\boldsymbol{p} + (1-\lambda)\boldsymbol{q}),
\end{aligned}$$

where the first and last equalities are due the 1-homogeneity of σ and the inequality is due to $\tilde{\sigma}$ being concave on the $\operatorname{int}\tilde{\Delta}_n$. □

H.1 Convexity of the Superprediction Set

In the literature, many theoretical results involving loss functions relied on the fact that the superprediction set of a proper loss is convex [24, 7]. An earlier proof of this result by [24] was incomplete[3]. In the next theorem we restate this result.

Theorem 48. *If $\ell\colon \Delta_n \to [0,+\infty[^n$ is a continuous proper loss, then $\mathscr{S}_\ell = \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$. In particular, $\mathscr{S}_\ell$ is convex.*

$\mathscr{S}_\ell \subseteq \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$. : Let $\boldsymbol{v} \in \mathscr{S}_\ell$, $\boldsymbol{u} \in [0,+\infty[^n$, and $\boldsymbol{q} \in \Delta_n$ such that $\boldsymbol{v} = \ell(\boldsymbol{q}) + \boldsymbol{u}$. Since ℓ is proper then $\forall \boldsymbol{p} \in \Delta_n, \underline{L}_\ell(\boldsymbol{p}) = \langle \boldsymbol{p}, \ell(\boldsymbol{p})\rangle \leq \langle \boldsymbol{p}, \ell(\boldsymbol{q})\rangle \leq \langle \boldsymbol{p}, \ell(\boldsymbol{q})+\boldsymbol{u}\rangle = \langle \boldsymbol{p}, \boldsymbol{v}\rangle$. Therefore, $\boldsymbol{v} \in \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$.

[$\bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})} \subseteq \mathscr{S}_\ell$]: Let $\boldsymbol{v} \in \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$. Let $\Omega = [n]$, $\Delta(\Omega) = \Delta_n$, and $Q(\boldsymbol{p}, x) = \ell_x(\boldsymbol{p}) - v_x$ for all $(\boldsymbol{p}, x) \in \Delta_n \times [n]$. Since $\boldsymbol{v} \in \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{-\boldsymbol{p},-\underline{L}_\ell(\boldsymbol{p})}$, $\mathbb{E}_{x\sim\boldsymbol{p}} Q(\boldsymbol{p}, x) = \langle \boldsymbol{p}, \ell(\boldsymbol{p})\rangle - \langle \boldsymbol{p}, \boldsymbol{v}\rangle \leq 0$ for all $\boldsymbol{p} \in \Delta_n$. Lemma 23, implies that there exists $\boldsymbol{p}_* \in \Delta_n$ such that $Q(\boldsymbol{p}_*, x) = \ell_x(\boldsymbol{p}_*) - v_x \leq 0$, for all $x \in [n]$. This shows that $\boldsymbol{v} \in \mathscr{S}_\ell$. □

H.2 Curvature of the Loss Surface

The *normal curvature* of a $\tilde{n}$-manifold $\mathcal{S}$ [18] at a point $\boldsymbol{r} \in \mathcal{S}$ in the direction of $\boldsymbol{w} \in T_{\boldsymbol{r}}\mathcal{S}$, where $T_{\boldsymbol{r}}\mathcal{S}$ is the *tangent space* of $\mathcal{S}$ at $\boldsymbol{r} \in \mathcal{S}$, is defined by

$$\kappa(\boldsymbol{r}, \boldsymbol{w}) = \frac{\langle \boldsymbol{w}, \mathsf{DN}^{\mathcal{S}}(\boldsymbol{r})\boldsymbol{w}\rangle}{\langle \boldsymbol{w}, \boldsymbol{w}\rangle}, \tag{128}$$

where $\mathsf{N}^{\mathcal{S}}(\boldsymbol{r})$ is the normal vector to the surface at $\boldsymbol{r}$. The *minimum principal curvature* of $\mathcal{S}$ at $\boldsymbol{r}$ is expressed as $\underline{\kappa}(\boldsymbol{r}) \coloneqq \inf\{\kappa(\boldsymbol{r}, \boldsymbol{w}) : \boldsymbol{w} \in T_{\boldsymbol{r}}\mathcal{S} \cap \mathcal{B}(\boldsymbol{r}, 1)\}$.

In the next theorem, we establish a direct link between the curvature of a loss surface and the Hessian of the loss' Bayes risk.

Theorem 49. *Let $\ell\colon \operatorname{ri}\Delta_n \to [0,+\infty[^n$ be a loss whose Bayes risk is twice differentiable and strictly concave on $]0,+\infty[^n$. Let $\boldsymbol{p} \in \operatorname{ri}\Delta_n$, $X_{\boldsymbol{p}} \coloneqq I_{\tilde{n}} - \tilde{\boldsymbol{p}}\mathbf{1}_{\tilde{n}}^{\mathsf{T}}$, and $\boldsymbol{w} \in T_{\tilde{\ell}(\tilde{\boldsymbol{p}})}\mathcal{S}_\ell$. Then*

1. *$\exists \boldsymbol{v} \in \mathbb{R}^{n-1}$ such that $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v} = \boldsymbol{w}$.*
2. *$\mathcal{S}_\ell$ is a $\tilde{n}$-manifold.*
3. *The normal curvature of $\mathcal{S}_\ell$ at $\ell(\boldsymbol{p}) = \tilde{\ell}(\tilde{\boldsymbol{p}})$ in the direction $\boldsymbol{w}$ is given by*

$$\kappa_\ell(\ell(\boldsymbol{p}), \boldsymbol{w}) = \left\| \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^{\mathsf{T}} \end{bmatrix} (-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}} \boldsymbol{u} \right\|^{-1}, \tag{129}$$

where $\boldsymbol{u} = (-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}}\boldsymbol{v} / \|(-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}}\boldsymbol{v}\|$.

It becomes clear from (129) that smaller eigenvalues of $-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ will tend to make the loss surface more curved at $\ell(\boldsymbol{p})$, and vice versa.

Before proving Theorem 49, we first define parameterizations on manifolds.

Definition 50 (Local and Global Parameterization)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a $\tilde{n}$-manifold and $\mathcal{U}$ an open set in $\mathbb{R}^{\tilde{n}}$. The map $\varphi : \mathcal{U} \to \mathcal{S}$ is called a* local parameterization *of $\mathcal{S}$ if $\mathsf{D}\varphi(\boldsymbol{u}) : \mathbb{R}^{\tilde{n}} \to T_{\varphi(\boldsymbol{u})}\mathcal{S}$ is injective for all $\boldsymbol{u} \in \mathcal{U}$, where $T_{\varphi(\boldsymbol{u})}\mathcal{S}$ is the tangent space of $\mathcal{S}$ at $\varphi(\boldsymbol{u}) \in \mathcal{S}$. φ is called a* global *parameterization of $\mathcal{S}$ if it is, additionally, onto.*

Let φ be a global parameterization of $\mathcal{S}$ and $\mathsf{N}^\varphi \coloneqq \mathsf{N}^{\mathcal{S}} \circ \varphi$. By a direct application of the chain rule, (128) can be written as

$$\kappa(\varphi(\boldsymbol{u}), \boldsymbol{w}) = \frac{\langle \boldsymbol{w}, \mathsf{DN}^\varphi(\boldsymbol{u})\boldsymbol{v}\rangle}{\langle \boldsymbol{w}, \boldsymbol{w}\rangle}, \tag{130}$$

[3] It was claimed that if $\mathscr{S}_\ell$ is non-convex, there exists a point $\boldsymbol{s}_0$ on the loss surface $\mathcal{S}_\ell$ such that no hyperplane supports $\mathscr{S}_\ell$ at $\boldsymbol{s}_0$. The non-convexity of a set by itself is not sufficient to make such a claim; the continuity of the loss ℓ is required.

where $\boldsymbol{v}$ is such that $\mathsf{D}\varphi(\boldsymbol{u})\boldsymbol{v} = \boldsymbol{w}$. The existence of such a $\boldsymbol{v}$ is guaranteed by the fact that $\mathsf{D}\varphi$ is injective and $\dim \mathbb{R}^{\tilde{n}} = \dim T_{\varphi(\boldsymbol{u})}\mathcal{S} = \tilde{n}$.

Theorem 49. First we show that $\mathcal{S}_\ell$ is a $\tilde{n}$-manifold. Consider the map $\tilde{\ell} : \operatorname{int} \tilde{\Delta}_n \to \mathcal{S}_\ell$ and note that $\operatorname{int} \tilde{\Delta}_n$ is trivially a $\tilde{n}$-manifold. Due to the strict concavity of the Bayes risk, $\tilde{\ell}$ is injective [19] and from Lemmas 27 and 47, $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}) : \mathbb{R}^{\tilde{n}} \to T_{\tilde{\ell}(\tilde{\boldsymbol{p}})}\mathcal{S}_\ell$ is also injective. Therefore, $\tilde{\ell}$ is an *immersion* [14]. $\tilde{\ell}$ is also *proper* in the sense that the preimage of every compact subset of $\mathcal{S}_\ell$ is compact. Therefore, $\tilde{\ell}$ is a proper injective immersion, and thus it is an embedding from the $\tilde{n}$-manifold $\operatorname{int} \tilde{\Delta}_n$ to $\mathcal{S}_\ell$ (ibid.). Hence, $\mathcal{S}_\ell$ is a manifold.

Now we prove (129). The map $\tilde{\ell}$ is a global parameterization of $\mathcal{S}_\ell$. In fact, from Lemma 27, $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})$ has rank $\tilde{n}$, for all $\tilde{\boldsymbol{p}} \in \operatorname{int} \tilde{\Delta}_n$, which implies that $\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})$ is onto from $\mathbb{R}^{\tilde{n}}$ to $T_{\tilde{\ell}(\tilde{\boldsymbol{p}})}\mathcal{S}_\ell$. Therefore, given $\boldsymbol{w} \in T_{\tilde{\ell}(\tilde{\boldsymbol{p}})}\mathcal{S}_\ell$, there exists $\boldsymbol{v} \in \mathbb{R}^{\tilde{n}}$ such that $\boldsymbol{w} = \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}$. Furthermore, Lemma 27 implies that $\mathsf{N}^{\tilde{\ell}}(\tilde{\boldsymbol{p}}) = \boldsymbol{p}$, since $\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^\mathsf{T}$. Substituting $\mathsf{N}^{\tilde{\ell}}$ into (130) yields

$$
\begin{aligned}
\kappa_\ell(\tilde{\ell}(\tilde{\boldsymbol{p}}), \boldsymbol{w}) &= \frac{\boldsymbol{v}^\mathsf{T}(\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}}))^\mathsf{T} \begin{bmatrix} I_{\tilde{n}}, \\ \mathbf{1}_{\tilde{n}} \end{bmatrix} \boldsymbol{v}}{\left\langle \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v} \right\rangle}, \\
&= \frac{\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) \begin{bmatrix} X_{\boldsymbol{p}}^\mathsf{T}, & -\tilde{\boldsymbol{p}} \end{bmatrix} \begin{bmatrix} I_{\tilde{n}} \\ \mathbf{1}_{\tilde{n}} \end{bmatrix} \boldsymbol{v}}{\left\langle \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v} \right\rangle}, \\
&= \frac{\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}}{\boldsymbol{v}^\mathsf{T}\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) \begin{bmatrix} X_{\boldsymbol{p}}^\mathsf{T}, & -\tilde{\boldsymbol{p}} \end{bmatrix} \begin{bmatrix} X_{\boldsymbol{p}} \\ -\tilde{\boldsymbol{p}}^\mathsf{T} \end{bmatrix} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}}. \qquad (131)
\end{aligned}
$$

Setting $\boldsymbol{u} = (-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}}\boldsymbol{v}/\|(-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{\frac{1}{2}}\boldsymbol{v}\|$ in (131) gives the desired result. □

I Classical Mixability Revisited

In this appendix, we provide a more concise proof of the necessary and sufficient conditions for the convexity of the superprediction set [19].

Theorem 51. *Let $\ell \colon \Delta_n \to [0, +\infty[^n$ be a strictly proper loss whose Bayes risk is twice differentiable on $]0, +\infty[^n$. The following points are equivalent;*

(i) $\forall \tilde{\boldsymbol{p}} \in \operatorname{int} \tilde{\Delta}_n,\ \eta\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) \succeq \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})$.

(ii) $e^{-\eta\mathscr{S}_\ell} = \bigcap_{\boldsymbol{p}\in\Delta_n} \mathcal{H}_{\tau(\boldsymbol{p}),1} \cap [0, +\infty[^n$, *where* $\tau(\boldsymbol{p}) \coloneqq \boldsymbol{p} \odot e^{\eta\ell(\boldsymbol{p})}$.

(iii) $e^{-\eta\mathscr{S}_\ell}$ *is convex.*

Proof. We already showed (i) $\implies$ (ii) $\implies$ (iii) in the proof of Theorem 7.

We now show (iii) $\implies$ (i). Since $e^{-\eta\mathscr{S}_\ell}$ is convex, any point $\boldsymbol{s} \in \operatorname{bd} e^{-\eta\mathscr{S}_\ell}$ is supported by a hyperplane [8, Lem. A.4.2.1]. Since $\boldsymbol{u} \to e^{-\eta\boldsymbol{u}}$ is a homeomorphism, it maps boundaries to boundaries. From this and Lemma 36, $\operatorname{bd} e^{-\eta\mathscr{S}_\ell} = e^{-\eta\mathcal{S}_\ell}$. Thus, for $\boldsymbol{p} \in \operatorname{ri} \Delta_n$, there exists a unit-norm vector $\boldsymbol{u} \in \mathbb{R}^n$ such that for all $\boldsymbol{s} \in \mathscr{S}_\ell$ it either holds that $\langle \boldsymbol{u}, e^{-\eta\ell(\boldsymbol{p})}\rangle \leq \langle \boldsymbol{u}, e^{-\eta\boldsymbol{s}}\rangle$; or $\langle \boldsymbol{u}, e^{-\eta\ell(\boldsymbol{p})}\rangle \geq \langle \boldsymbol{u}, e^{-\eta\boldsymbol{s}}\rangle$. It is easy to see that it is the latter case that holds, since we can choose $\boldsymbol{s} = \ell(\boldsymbol{r}) + c\mathbf{1} \in \mathscr{S}_\ell$, for $\boldsymbol{r} \in \Delta_n$, and make $\langle \boldsymbol{u}, e^{-\eta\boldsymbol{s}}\rangle$ arbitrarily small by making $c \in \mathbb{R}$ large. Therefore, $\forall \boldsymbol{r} \in \operatorname{ri} \Delta_n, \langle \boldsymbol{u}, e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{p}})}\rangle = \langle \boldsymbol{u}, e^{-\eta\ell(\boldsymbol{p})}\rangle \geq \langle \boldsymbol{u}, e^{-\eta\ell(\boldsymbol{r})}\rangle = \langle \boldsymbol{u}, e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{r}})}\rangle$ and $\tilde{\boldsymbol{p}}$ is a critical point of the function $f(\tilde{\boldsymbol{r}}) \coloneqq \langle \boldsymbol{u}, e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{r}})}\rangle$ on $\operatorname{int} \tilde{\Delta}_n$. This implies that $\nabla f(\tilde{\boldsymbol{p}}) = \mathbf{0}_{\tilde{n}}$; that is, $-\eta\langle \boldsymbol{u}, \operatorname{diag}(e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{p}})})\mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = -\eta\langle \operatorname{diag}(e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{p}})})\boldsymbol{u}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\rangle = \mathbf{0}_{\tilde{n}}^\mathsf{T}$. From Lemma 27, there exists $\lambda \in \mathbb{R}$ such that $\operatorname{diag}(e^{-\eta\tilde{\ell}(\tilde{\boldsymbol{p}})})\boldsymbol{u} = \lambda\boldsymbol{p}$. Therefore, $\boldsymbol{u} = \lambda\boldsymbol{p} \odot e^{\eta\tilde{\ell}(\tilde{\boldsymbol{p}})}$, where $\lambda = \|\boldsymbol{p} \odot e^{\eta\tilde{\ell}(\tilde{\boldsymbol{p}})}\|^{-1}$,

since $\|\boldsymbol{u}\| = 1$. For $\boldsymbol{v} \in \mathbb{R}^{n-1}$, let $\tilde{\boldsymbol{\alpha}}^t := \tilde{\boldsymbol{p}} + t\boldsymbol{v}$, where $t \in \{s : \tilde{\boldsymbol{p}} + s\boldsymbol{v} \in \operatorname{int} \tilde{\Delta}_n\}$. Since f is twice differentiable and attains a maximum at $\tilde{\boldsymbol{p}}$,

$$\begin{aligned} 0 \geq \frac{1}{\lambda\eta} \frac{d^2}{dt^2} f \circ \tilde{\boldsymbol{\alpha}}^t \Big|_{t=0} &= \frac{1}{\lambda} \frac{d}{dt} \left\langle \boldsymbol{u}, \operatorname{diag} e^{-\eta \tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)} \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v} \right\rangle \Big|_{t=0}, \\ &= \frac{d}{dt} \left\langle \boldsymbol{p} \odot e^{\eta \tilde{\ell}(\tilde{\boldsymbol{p}})}, \operatorname{diag} e^{-\eta \tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)} \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{p}})\boldsymbol{v} \right\rangle \Big|_{t=0} + \frac{d}{dt} \left\langle \boldsymbol{p}, \mathsf{D}\tilde{\ell}(\tilde{\boldsymbol{\alpha}}^t)\boldsymbol{v} \right\rangle \Big|_{t=0}, \\ &= \eta \boldsymbol{v}^\mathsf{T} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) (\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v} - \boldsymbol{v}^\mathsf{T} \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}, \end{aligned} \tag{132}$$

where in the second equality we substituted $\boldsymbol{u}$ by $\lambda \boldsymbol{p} \odot e^{\eta \tilde{\ell}(\tilde{\boldsymbol{p}})}$ and in (132) we used (17) and (18) from Lemma 28. Note that by the assumptions on ℓ it follows that the Bayes risk $\underline{\tilde{L}}_\ell$ is strictly concave [19, Lemma 6] and $-\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ is symmetric negative-definite. In particular, $\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})$ is invertible. Setting $\hat{\boldsymbol{v}} := \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}})\boldsymbol{v}$ in (132) yields

$$0 \geq \eta \hat{\boldsymbol{v}} (\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1} \hat{\boldsymbol{v}} - \hat{\boldsymbol{v}} (\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{-1} \hat{\boldsymbol{v}}.$$

Since $\boldsymbol{v} \in \mathbb{R}^{n-1}$ was chosen arbitrarily, $(\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}))^{-1} \succeq \eta (\mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}}))^{-1}, \forall \tilde{\boldsymbol{p}} \in \operatorname{int} \tilde{\Delta}_n$. This is equivalent to the condition $\forall \tilde{\boldsymbol{p}} \in \operatorname{int} \tilde{\Delta}_n,\ \eta \mathsf{H}\underline{\tilde{L}}_\ell(\tilde{\boldsymbol{p}}) \succeq \mathsf{H}\underline{\tilde{L}}_{\log}(\tilde{\boldsymbol{p}})$. □

J An Experiment on Football Prediction Dataset

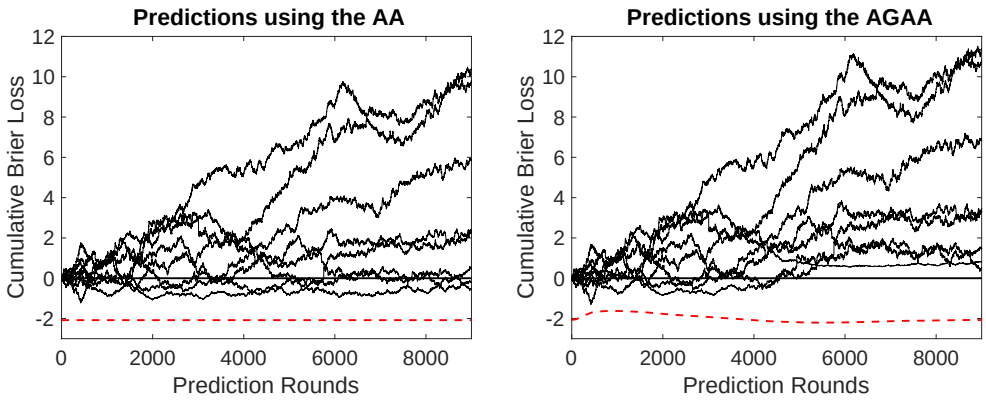

Figure 1: The figure corresponds to the 2005/2006, 2006/2007, 2007/2008, and 2008/2009 seasons. The solid lines represent, at each round t, the difference between the cumulative losses of the experts and that of the learner who uses either the AA (left) or the AGAA (right); that is, $\mathrm{Loss}_\theta^{\ell_{\mathrm{Brier}}}(t) - \mathrm{Loss}_{\mathfrak{M}}^{\ell_{\mathrm{Brier}}}(t)$, for $\mathfrak{M} \in \{\mathrm{AA}, \mathrm{AGAA}\}$. The red dashed lines represent the negative of the regret bound in (12) with respect to the best expert θ^*; that is, $-R^{\mathrm{S}}_{\ell_{\mathrm{Brier}}} - \Delta R_{\theta^*}(t)$ at each round t.

J.1 Testing the AGAA

To test the AGAA empirically, we used prediction data[4] from the British football leagues, including the Premier Leagues, Championships, Leagues 1-2, and Conferences. The first dataset contains predictions for the 2005/2006, 2006/2007, 2007/2008, and 2008/2009 seasons, matching the dataset used in [21]. The second dataset contains predictions for the 2009/2010, 2010/2011, 2011/2012, and 2012/2013 seasons. For this set, we considered predictions from 9 bookmakers; Bet365, Bet&Win, Blue Square, Gamebookers, Interwetten, Ladbrokes, Stan James, VC Bet, and William Hill.

On each dataset, we compared the performance of the AGAA with that of the AA using the Brier score (the Brier loss is 1-mixable). For the AGAA, we chose $\boldsymbol{\beta}^t$ according to Theorem 19 with $\boldsymbol{v}^t := -\frac{1}{2t} \sum_{s=1}^t \ell_{x^s}(A^s)$ and we set $\Phi = \mathrm{S}$, i.e. the Shannon entropy. The results in Figure 1 [resp. Figure 2] correspond to the seasons from 2005 to 2009 [resp. 2009 to 2013]. For fair comparison

[4] The data was collected from http://www.football-data.co.uk/.

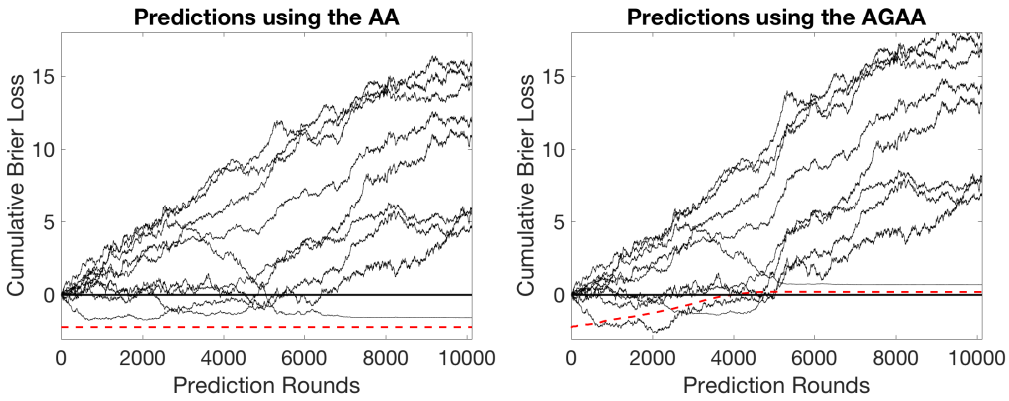

Figure 2: The figure corresponds to the 2009/2010, 2010/2011, 2011/2012, and 2012/2013 seasons The solid lines represent, at each round t, the difference between the cumulative losses of the experts and that of the learner who uses either the AA (left) or the AGAA (right); that is, $\mathrm{Loss}_{\theta}^{\ell_{\mathrm{Brier}}}(t) - \mathrm{Loss}_{\mathfrak{M}}^{\ell_{\mathrm{Brier}}}(t)$, for $\mathfrak{M} \in \{\mathrm{AA}, \mathrm{AGAA}\}$. The red dashed lines represent the negative of the regret bound in (12) with respect to the best expert θ^*; that is, $-R^{\mathrm{S}}_{\ell_{\mathrm{Brier}}} - \Delta R_{\theta^*}(t)$ at each round t.

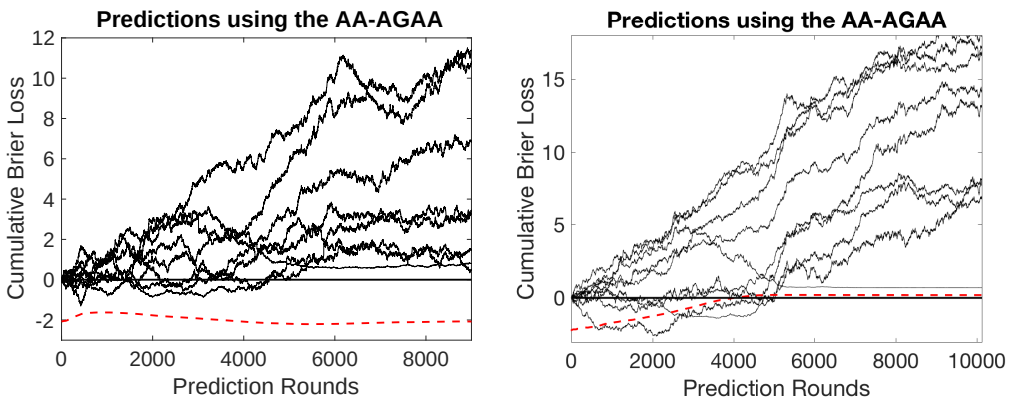

Figure 3: The figure on the left [resp. right] hand side corresponds to the football seasons from 2005 to 2009 [resp. 2009 to 2013]. The solid lines represent, at each round t, the difference between the cumulative losses of the experts and that of the learner using the AA-AGAA meta algorithm (refer to text); that is, $\mathrm{Loss}_{\theta}^{\ell_{\mathrm{Brier}}}(t) - \mathrm{Loss}_{\mathrm{AA\text{-}AGAA}}^{\ell_{\mathrm{Brier}}}(t)$. The red dashed lines represent the negative of the regret bound in (12) with respect to the best expert θ^*; that is, $-R^{\mathrm{S}}_{\ell_{\mathrm{Brier}}} - \Delta R_{\theta^*}(t)$ at each round t.

with the results of Vovk [21], we 1) used the same substitution function as [21]; 2) used the same method for converting odds to probabilities; and 3) sorted the data first by date then by league and then by name of the host team (For more detail see [21]).

In all figures the solid lines represent, at each round t, the difference /between the cumulative losses of the experts and that of the learners; that is, $\mathrm{Loss}_{\theta}^{\ell_{\mathrm{Brier}}}(t) - \mathrm{Loss}_{\mathfrak{M}}^{\ell_{\mathrm{Brier}}}(t)$, for $\mathfrak{M} = \mathrm{AA}, \mathrm{AGAA}$. The red dashed lines represent the negative of the regret bound in (12) with respect to the best expert θ^*; that is, $-R^{\mathrm{S}}_{\ell_{\mathrm{Brier}}} - \Delta R_{\theta^*}(t) = -R^{\mathrm{S}}_{\ell_{\mathrm{Brier}}} - \sum_{s=1}^{t-1}(v_{\theta}^{s} - \langle \boldsymbol{v}^s, \boldsymbol{q}^s \rangle)$ at each round t, where $(\boldsymbol{q}^s)$ are the distributions over experts.

From Figures 1 and 2 it can be seen that the learners using the AGAA perform better than the best expert (and better than the AA) at the end of the games.

J.2 Testing a AA-AGAA Meta-Learner

Consider the algorithm (referred to as AA-AGAA) that takes the outputs of the AGAA and the AA as in the previous section and aggregates them using the AA to yield a *meta prediction*. The worst case

regret of this algorithm is guaranteed not to exceed that of the original AA and AGAA by more than $\eta^{-1} \log 2$ for an η-mixable loss. Figure 3 shows the results for this algorithm for the same datasets as the previous section. The AA-AGAA still achieves a negative regret at the end of the game.

Chapter 3

Lipschitz Adaptivity with Multiple Learning Rates in Online Learning

This chapter considers the bounded OCO optimization setting, which generalizes the experts' setting from the previous chapter. In the *bounded* OCO setting, the seminal MetaGrad algorithm of Van Erven and Koolen [2016] can automatically adapt to different types of curvature of the observed loss functions. It is the first algorithm that adapts to the strong convexity and exp-concavity of the observed sequence of losses—achieving logarithmic regrets in these cases—without requiring knowledge of the curvature parameters. Algorithms such as online gradient descent, online newton step, and the exponential weights algorithm can also achieve a logarithmic regret in these cases. Still, they require knowledge of the parameters of curvature in advance [Hazan et al., 2007]. MetaGrad's guarantee shows that automatic curvature adaptivity is possible. However, MetaGrad (and many other algorithms) assumes that the losses have bounded gradients and require the corresponding bound as input—without it, the algorithm may completely fail [Van Erven et al., 2021]. On the other hand, there exist algorithms that can adapt to an unknown bound on the gradients, which we refer to as Lipschitz adaptivity. Thus, what remains unknown is whether curvature and Lipschitz adaptivity are possible simultaneously. In this chapter, we answer this question positively by studying novel gradient-clipping and restart schemes that are of independent interest in OCO. The restart scheme consists of reinitializing the algorithm whenever the ratio of the maximum norm of observed subgradient to the norm of the initial non-zero subgradient is too large (typically larger than the current number of rounds). Adopting this restart scheme solves issues related to Lipschitz adaptivity for many existing algorithms (e.g. those by Ross et al. [2013a]; Wintenberger [2017]; Kotłowski [2017]; Mhammedi et al. [2019b]; Kempka et al. [2019b]), making them *scale-free*. Here, scale-free means multiplying the losses by some positive constant does not change the algorithm's outputs—a naturally desirable property put forward by Cesa-Bianchi et al. [2007]; Orabona and Pál [2016b]. Carefully applying the restart and clipping schemes to MetaGrad leads to the *first* scale-free and curvature-adaptive algorithm for bounded OCO.

Proceedings of Machine Learning Research vol 99:1–22, 2019 32nd Annual Conference on Learning Theory

Lipschitz Adaptivity with Multiple Learning Rates in Online Learning

Zakaria Mhammedi ZAK.MHAMMEDI@ANU.EDU.AU
School of Engineering and Computer Science
The Australian National University and Data61
Wouter M. Koolen WMKOOLEN@CWI.NL
Centrum Wiskunde & Informatica
Amsterdam, the Netherlands
Tim van Erven TIM@TIMVANERVEN.NL
Statistics Department
Leiden University, the Netherlands

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We aim to design adaptive online learning algorithms that take advantage of any special structure that might be present in the learning task at hand, with as little manual tuning by the user as possible. A fundamental obstacle that comes up in the design of such adaptive algorithms is to calibrate a so-called step-size or learning rate hyperparameter depending on variance, gradient norms, etc. A recent technique promises to overcome this difficulty by maintaining multiple learning rates in parallel. This technique has been applied in the MetaGrad algorithm for online convex optimization and the Squint algorithm for prediction with expert advice. However, in both cases the user still has to provide in advance a Lipschitz hyperparameter that bounds the norm of the gradients. Although this hyperparameter is typically not available in advance, tuning it correctly is crucial: if it is set too small, the methods may fail completely; but if it is taken too large, performance deteriorates significantly. In the present work we remove this Lipschitz hyperparameter by designing new versions of MetaGrad and Squint that adapt to its optimal value automatically. We achieve this by dynamically updating the set of active learning rates. For MetaGrad, we further improve the computational efficiency of handling constraints on the domain of prediction, and we remove the need to specify the number of rounds in advance.

1. Introduction

We consider *online convex optimization* (OCO) of a sequence of convex functions $\ell_1, \dots, \ell_T$ over a given bounded convex domain, which become available one by one over the course of T rounds (Shalev-Shwartz, 2011; Hazan, 2016). Typically $\ell_t(\boldsymbol{u}) = \textsc{loss}(\boldsymbol{u}, \boldsymbol{x}_t, y_t)$ represents the *loss* of predicting with parameters $\boldsymbol{u}$ on the t-th data point $(\boldsymbol{x}_t, y_t)$ in a machine learning task. At the start of each round t, a learner has to predict the best parameters $\widehat{\boldsymbol{u}}_t$ for the function ℓ_t before finding out what ℓ_t is, and the goal is to minimize the *regret*, which is the difference in the sum of function values between the learner's predictions $\widehat{\boldsymbol{u}}_1, \dots, \widehat{\boldsymbol{u}}_T$ and the best fixed oracle parameters $\boldsymbol{u}$ that could have been chosen if all the functions had been given in advance. A special case of OCO is prediction with expert advice (Cesa-Bianchi and Lugosi, 2006), where the functions $\ell_t(\boldsymbol{u}) = \langle \boldsymbol{u}, \boldsymbol{l}_t \rangle$ are convex combinations of the losses $\boldsymbol{l}_t = (l_{t,1}, \dots, l_{t,K})$ of K expert predictors and the domain is the probability simplex.

Central results in these settings show that it is possible to control the regret with virtually no prior knowledge about the functions. For instance, knowing only a $\|\cdot\|_2$-upper-bound G on the gradients $\boldsymbol{g}_t = \nabla \ell_t(\widehat{\boldsymbol{u}}_t)$, the online gradient descent (OGD) algorithm guarantees $O(G\sqrt{T})$ regret by tuning its learning rate hyperparameter η_t proportional to $1/(G\sqrt{t})$ (Zinkevich, 2003), and in the case of prediction with expert advice the Hedge algorithm achieves regret $O(L\sqrt{T \ln K})$ knowing only an upper-bound L on the range $\max_k l_{t,k} - \min_k l_{t,k}$ of the expert losses (Freund and Schapire, 1997). Here G is the $\|\cdot\|_2$-Lipschitz constant of the learning task[1], and $L/2$ is the $\|\cdot\|_1$-Lipschitz constant over the probability simplex.

The above guarantees are tight if we make no further assumptions about the functions (ℓ_t) (Hazan, 2016; Cesa-Bianchi et al., 1997), but they can be significantly improved if the functions have additional special structure that makes the learning task easier. The literature on online learning explores multiple orthogonal dimensions in which tasks may be significantly easier in practice (see 'related work' below). Here, we focus on the following refined data-dependent regret guarantees, which are known to exploit multiple types of easiness at the same time:

$$\text{OCO:} \qquad O\left(\sqrt{V_T^{\boldsymbol{u}} d \log T}\right) \text{ for all } \boldsymbol{u}, \qquad \text{with } V_T^{\boldsymbol{u}} = \sum_{t=1}^{T} \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle^2, \tag{1}$$

$$\text{Experts:} \quad O\left(\sqrt{\mathbb{E}_{\rho(k)}[V_T^k]\,\mathrm{KL}(\rho\|\pi)}\right) \text{ for all } \rho, \qquad \text{with } V_T^k = \sum_{t=1}^{T} \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{e}_k, \boldsymbol{l}_t \rangle^2, \tag{2}$$

where d is the number of parameters and $\mathrm{KL}(\rho\|\pi) = \sum_{k=1}^{K} \rho(k) \ln \rho(k)/\pi(k)$ is the Kullback-Leibler divergence from a fixed prior distribution π over experts to any (data-dependent) comparator distribution ρ; for instance, ρ is allowed here to be a point-mass on the best expert k^* in hindsight, in which case we would have $\mathrm{KL}(\rho\|\pi) = -\ln \pi(k^*)$.

The OCO guarantee is achieved by the METAGRAD algorithm (Van Erven and Koolen, 2016), and implies regret that grows at most logarithmically in T both in case the losses are curved (exp-concave, strongly convex) and in the stochastic case whenever the losses are independent, identically distributed samples with variance controlled by a Bernstein condition (Koolen et al., 2016). The guarantee for the expert case is achieved by the SQUINT algorithm (Koolen and Van Erven, 2015; Koolen, 2015). It simultaneously exploits two types of structures: in many cases the V_T^k term is much smaller than L^2T (Gaillard et al., 2014; Koolen et al., 2016) and the so-called *quantile bound* $\mathrm{KL}(\rho\|\pi)$ is much smaller than the worst case $\ln K$ when multiple experts make good predictions (Chaudhuri et al., 2009; Chernov and Vovk, 2010). SQUINT and METAGRAD are both based on the same technique of tracking the empirical performance of *multiple learning rates* in parallel over quadratic approximations of the original losses. A computational difference though is that SQUINT is able to do this by a continuous integral that can be evaluated in closed form, whereas METAGRAD uses a discrete grid of learning rates.

Unfortunately, to achieve (1) and (2), both METAGRAD and SQUINT need knowledge of the Lipschitz constant (G or L, respectively). Overestimating G or L by a factor of $c > 1$ has the effect of reducing the effective amount of available data by the same factor c, but underestimating the Lipschitz constant is even worse since it can make the methods fail completely. In fact, the ability to adapt to G has been credited (Ward et al., 2018) as one of the main reasons for the practical

1. We slightly abuse terminology here, because the standard definition of a Lipschitz constant requires an upper-bound on the gradient norms for any parameters $\boldsymbol{u}$, not just for $\boldsymbol{u} = \widehat{\boldsymbol{u}}_t$, and may therefore be larger.

success of the AdaGrad algorithm (Duchi et al., 2011; McMahan and Streeter, 2010). Thus getting the Lipschitz constant right makes the difference between having practical algorithms and having promising theoretical results.

For OCO, an important first step towards combining Lipschitz adaptivity to G with regret bounds of the form (1) was taken by Cutkosky and Boahen (2017b), who aimed for (1) but had to settle for a weaker result with $G\sum_{t=1}^{T}\|\boldsymbol{g}_t\|_2\|\widehat{\boldsymbol{u}}_t-\boldsymbol{u}\|_2^2$ instead of $V_T^{\boldsymbol{u}}$. Although not sufficient to adapt to a Bernstein condition, they do provide a series of stochastic examples where their bound already leads to a fast $O(\ln^4 T)$ rates. For the expert setting, Wintenberger (2017) has made significant progress towards a version of (2) without the quantile bound improvement, but he is left with having to specify an initial guess L_{guess} for L that enters as $O(\ln\ln(L/L_{\text{guess}}))$ in his bound, which may yet be arbitrarily large when the initial guess is on the wrong scale.

Main Contributions. Our main contributions are that we complete the process began by Cutkosky and Boahen (2017b) and Wintenberger (2017) by showing that it is indeed possible to achieve (1) and (2) without prior knowledge of G or L. In fact, for the expert setting we are able to adapt to the tighter quantity $B \geq \max_k |\langle \widehat{\boldsymbol{u}}_t - \boldsymbol{e}_k, \boldsymbol{l}_t\rangle|$. We achieve these results by dynamically updating the set of active learning rates in METAGRAD and SQUINT depending on the observed Lipschitz constants. In both cases, we encounter a similar tuning issue as Wintenberger (2017), but we avoid the need to specify any initial guess using a new restarting scheme, which restarts the algorithm when the observed Lipschitz constant increases too much. Interestingly, the scheme and its analysis are different from the well-known doubling trick (Cesa-Bianchi and Lugosi, 2006), and the regret bound is dominated by the regret incurred over the last *two* epochs instead of just the last epoch. Adding up the regret bounds over the last two epochs leads to at most an extra $\sqrt{2}$ factor multiplying the final bound, and so this is the overhead we incur for Lipschitz adaptivity. In addition to these main results, we remove the need to specify the number of rounds T in advance for METAGRAD by adding learning rates as T gets larger, and we improve the computational efficiency of how it handles constraints on the domain of prediction: by a minor extension of the black-box reduction for projections of Cutkosky and Orabona (2018), we incur only the computational cost of projecting on the domain of interest in *Euclidean* distance. This should be contrasted with the usual projections in time-varying Mahalanobis distance for second-order methods like METAGRAD.

Related Work. We build on several lines of work that achieve subsets of Lipschitz, variance and quantile adaptivity. Lipschitz adaptivity in OCO is achieved by OGD with learning rate $\eta_t \propto 1/\sqrt{\sum_{s=1}^{t}\|\boldsymbol{g}_s\|_2^2}$, which leads to $O(\sqrt{\sum_{t=1}^{T}\|\boldsymbol{g}_t\|_2^2}) = O(G\sqrt{T})$ regret. This is the approach taken by AdaGrad (for each dimension separately) (Duchi et al., 2011; McMahan and Streeter, 2010). Lipschitz adaptive methods for prediction with expert advice (sometimes called scale-free) were obtained by Cesa-Bianchi et al. (2007) and De Rooij et al. (2014). These include a data-dependent variance term (though different from V_T^k in (2)), but no quantiles.

Dropping Lipschitz adaptivity, we find that bounds with V_T^k from (2) have previously been obtained by Gaillard et al. (2014) and Wintenberger (2014) without quantile bounds. Quantile adaptivity was achieved by Chaudhuri et al. (2009) and Chernov and Vovk (2010) without variance adaptivity, and with a slightly weaker notion of variance by Luo and Schapire (2015). In OCO, the analogue of quantile adaptivity is to adapt to the norm of $\boldsymbol{u}$, which has been achieved in various different ways, see for instance (McMahan and Abernethy, 2013; Cutkosky and Orabona, 2018).

Several other important (and related) criteria of easiness are actively considered in the literature. These include curvature of the loss functions, where earlier results achieve fast rates assuming that

the degree of curvature is known (Hazan et al., 2007), measured online (Bartlett et al., 2007; Do et al., 2009) or entirely unknown (Van Erven and Koolen, 2016; Cutkosky and Orabona, 2018). Fast rates are also possible for slowly-varying linear functions and, more generally, optimistically predictable gradient sequences (Hazan and Kale, 2010; Chiang et al., 2012; Rakhlin and Sridharan, 2013).

We view our results as a step towards developing algorithms that automatically adapt to multiple relevant measures of difficulty at the same time. It is not a given that such combinations are always possible. For example, Cutkosky and Boahen (2017a) show that Lipschitz adaptivity and adapting to the comparator complexity in OCO, although both achievable independently, cannot both be realized at the same time (at least not without further assumptions). A general framework to study which notions of task difficulty do combine into achievable bounds is provided by Foster et al. (2015). Foster et al. (2017) characterize the achievability of general data-dependent regret bounds for domains that are balls in general Banach spaces.

Outline. We add Lipschitz adaptivity to SQUINT for the expert setting in Section 2. Then, in Section 3, we do the same for METAGRAD in the OCO setting. The developments are analogous at a high level but differ in the details for computational reasons. We highlight the differences along the way. Section 3 further describes how to avoid specifying T in advance for METAGRAD. Then, in Section 4, we add efficient projections for METAGRAD, and finally Section 5 concludes with a discussion of directions for future work.

Problem Setting and Notation. In OCO, a learner repeatedly chooses actions $\widehat{\boldsymbol{u}}_t$ from a closed convex set $\mathcal{U} \subseteq \mathbb{R}^d$ during rounds $t = 1, \dots, T$, and suffers losses $\ell_t(\widehat{\boldsymbol{u}}_t)$, where $\ell_t : \mathcal{U} \to \mathbb{R}$ is a convex function. The learner's goal is to achieve small regret $R_T^{\boldsymbol{u}} = \sum_{t=1}^T \ell_t(\widehat{\boldsymbol{u}}_t) - \sum_{t=1}^T \ell_t(\boldsymbol{u})$ with respect to any comparator action $\boldsymbol{u} \in \mathcal{U}$, which measures the difference between the cumulative loss of the learner and the cumulative loss they could have achieved by playing the oracle action $\boldsymbol{u}$ from the start. A special case of OCO is prediction with expert advice, where $\ell_t(\boldsymbol{u}) = \langle \boldsymbol{u}, \boldsymbol{l}_t \rangle$ for $\boldsymbol{l}_t \in \mathbb{R}^K$ and the domain $\mathcal{U}$ is the probability simplex $\triangle_K = \{(u_1, \dots, u_K) : u_i \geq 0, \sum_i u_i = 1\}$. In this context we will further write $\boldsymbol{p}$ instead of $\boldsymbol{u}$ for the parameters to emphasize that they represent a probability distribution. We further define $[K] = \{1, \dots, K\}$.

2. An Adaptive Second-order Quantile Method for Experts

In this section, we present an extension of the SQUINT algorithm that adapts automatically to the loss range in the setting of prediction with expert advice.

Throughout this section, we denote the *instantaneous regret* of expert $k \in [K]$ in round t by $r_t^k \coloneqq \langle \widehat{\boldsymbol{p}}_t - \boldsymbol{e}_k, \boldsymbol{l}_t \rangle$, where $\widehat{\boldsymbol{p}}_t \in \triangle_K$ is the weight vector played by the algorithm and $\boldsymbol{l}_t \in \mathbb{R}^K$ is the observed loss vector. The cumulative regret with respect to expert k is given by $R_t^k \coloneqq \sum_{s=1}^t r_s^k$. The cumulative 'variance' with respect to expert k is measured by $V_t^k \coloneqq \sum_{s=1}^t v_s^k$ for $v_t^k \coloneqq (r_t^k)^2$. In the next subsection, we review the SQUINT algorithm.

2.1. The SQUINT Algorithm

We first describe the original SQUINT algorithm as introduced by Koolen and Van Erven (2015). Let π and γ be prior distributions with supports on $k \in [K]$ and $\eta \in]0, 1/2]$, respectively. After t

rounds, SQUINT outputs predictions

$$\widehat{\boldsymbol{p}}_{t+1} \propto \mathop{\mathbb{E}}_{\pi(k)\gamma(\eta)} \left[\eta e^{-\sum_{s=1}^{t} f_s(k,\eta)} \boldsymbol{e}_k \right], \tag{3}$$

where $f_t(k,\eta)$ are quadratic *surrogate losses* defined by

$$f_t(k,\eta) \coloneqq -\eta \langle \widehat{\boldsymbol{p}}_t - \boldsymbol{e}_k, \boldsymbol{l}_t \rangle + \eta^2 \langle \widehat{\boldsymbol{p}}_t - \boldsymbol{e}_k, \boldsymbol{l}_t \rangle^2.$$

Koolen and Van Erven (2015) propose to use the *improper prior* $\gamma(\eta) = 1/\eta$ which does not integrate to a finite value over its domain, but because of the weighting by η in (3) the predictions $\widehat{\boldsymbol{p}}_{t+1}$ are still well-defined. The benefit of the improper prior is that it allows calculating $\widehat{\boldsymbol{p}}_{t+1}$ in closed form (Koolen and Van Erven, 2015). It is also the natural candidate for Lipschitz adaptivity, as it is scale-invariant: the density of an interval only depends on the ratio of its endpoints, not on their location. For any distribution $\rho \in \triangle_K$, SQUINT achieves the following bound:

$$R_T^\rho = O\left(\sqrt{V_T^\rho \left(\mathrm{KL}(\rho \| \pi) + \ln \ln T \right)} \right),$$

where $R_T^\rho = \mathbb{E}_{\rho(k)}\left[R_T^k\right]$ and $V_T^\rho = \mathbb{E}_{\rho(k)}\left[V_T^k\right]$. This version of SQUINT assumes the loss range $\max_k l_{t,k} - \min_k l_{t,k}$ is at most 1, and can fail otherwise. In the next subsection, we present an extension of SQUINT which does not need to know the Lipschitz constant.

2.2. Lipschitz Adaptive SQUINT

We first design a version of SQUINT, called SQUINT+C, that still requires an initial estimate B of the Lipschitz constant. We then present SQUINT+L which tunes this parameter online. For now, we consider a fixed $B > 0$. In addition to this, the algorithm takes a prior distribution $\pi \in \triangle_K$. We denote the observed Lipschitz constant in round t at the algorithm's prediction $\widehat{\boldsymbol{p}}_t$ by $b_t \coloneqq \max_k |r_t^k| = \max_k |\langle \widehat{\boldsymbol{p}}_t - \boldsymbol{e}_k, \boldsymbol{l}_t \rangle|$, and denote its running maximum by $B_t \coloneqq B \vee \max_{s \le t} b_s$, with the convention that $B_0 = B$. We will also require a *clipped* version of the loss vector $\bar{\boldsymbol{l}}_t = \boldsymbol{l}_t \cdot B_{t-1}/B_t$, and denote by $\bar{r}_t^k = \langle \widehat{\boldsymbol{p}}_t - \boldsymbol{e}_k, \bar{\boldsymbol{l}}_t \rangle$ the *clipped instantaneous regret*; we will use that $|\bar{r}_t^k| \le B_{t-1}$. Following Cutkosky (2019), it suffices to control the regret for the clipped loss, because the cumulative difference is of the order of one round (*i.e.* a negligible lower-order constant):

$$R_T^k - \bar{R}_T^k \;\coloneqq\; \sum_{t=1}^{T} \left(r_t^k - \bar{r}_t^k \right) \;=\; \sum_{t=1}^{T} (B_t - B_{t-1}) \frac{r_t^k}{B_t} \;\le\; B_T - B_0. \tag{4}$$

This means we can focus on the regret for $\bar{\boldsymbol{l}}_t$, for which the range bound $|\bar{r}_t^k| \le B_{t-1}$ is available *ahead* of each round t. To motivate SQUINT+C, we define the potential function after T rounds by

$$\Phi_T \coloneqq \sum_k \pi_k \int_0^{\frac{1}{2B_{T-1}}} \frac{e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} - 1}{\eta} \, \mathrm{d}\eta \quad \text{where} \quad \bar{R}_T^k \coloneqq \sum_{t=1}^{T} \bar{r}_t^k \;\text{ and }\; \bar{V}_T^k \coloneqq \sum_{t=1}^{T} (\bar{r}_t^k)^2. \tag{5}$$

We also define $\Phi_0 = 0$ (due to the integrand being zero), even though it involves the meaningless B_{-1} in the upper limit. The algorithm is now derived from the desire of keeping this potential under control. As we will see in the analysis, this requirement uniquely forces the choice of weights

$$\widehat{p}_{T+1}^k \;\propto\; \pi_k \int_0^{\frac{1}{2B_T}} e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} \, \mathrm{d}\eta. \tag{6}$$

Algorithm 1 Restarts to make SQUINT+C or METAGRAD+C scale-free.

Require: ALG is either SQUINT+C or METAGRAD+C, taking as input parameter an initial scale B;

1: Play $\mathbf{0}$ for OCO or π for experts until the first time $t = \tau_1$ that $b_t \neq 0$;

2: Run ALG with input $B = B_{\tau_1}$ until the first time $t = \tau_2$ that $\dfrac{B_t}{B_{\tau_1}} > \displaystyle\sum_{s=1}^{t} \frac{b_s}{B_s}$;

3: Set $\tau_1 = \tau_2$ and goto line 2;

The predictions $\widehat{\boldsymbol{p}}_{t+1}$ take the same functional form as the original SQUINT, and can hence be evaluated in closed form (*i.e.* in terms of the Gaussian CDF). The regret analysis consists of two parts. First, we show that the algorithm keeps the potential small:

Lemma 1 *Given parameter $B > 0$,* SQUINT+C *ensures* $\Phi_T \leq \ln \frac{B_{T-1}}{B}$.

The next step of the argument is to show that a small potential Φ_T is useful. The argument here follows from (Koolen and Van Erven, 2015), specifically the version by Koolen (2015). We have:

Lemma 2 *For any comparator distribution $\rho \in \triangle_K$ the regret of* SQUINT+C *is at most*

$$\bar{R}_T^\rho \;\leq\; \sqrt{2\bar{V}_T^\rho}\left(1+\sqrt{2C_T^\rho}\right) + 5B_{T-1}\left(C_T^\rho + \ln 2\right), \quad \textit{where}$$

$$C_T^\rho \;\coloneqq\; \mathrm{KL}\left(\rho\|\pi\right) + \ln\left(\Phi_T + \frac{1}{2} + \ln\left(2 + \sum_{t=1}^{T-1}\frac{b_t}{B_t}\right)\right).$$

Keeping only the dominant terms, this reads $\bar{R}_T^\rho = O\left(\sqrt{\bar{V}_T^\rho\left(\mathrm{KL}\left(\rho\|\pi\right) + \ln\left(\Phi_T + \ln T\right)\right)}\right)$. Combining with (4), and Lemmas 1 and 2, we obtain a bound of the form

$$R_T^\rho \;=\; O\left(\sqrt{V_T^\rho\left(\mathrm{KL}\left(\rho\|\pi\right) + \ln\ln\frac{TB_{T-1}}{B}\right)} + 5B_T\left(\mathrm{KL}\left(\rho\|\pi\right) + \ln\ln\frac{TB_{T-1}}{B}\right)\right). \tag{7}$$

However, there does not seem to be any safe a-priori way to tune $B = B_0$. If we set it too small, the factor $\ln\ln(B_{T-1}/B)$ explodes. If we set it too large, with B much larger than the effective range of the data, then $B_T = B$ and the term outside the square-root on the RHS of (7) blows up. It does not appear possible to bypass this tuning dilemma directly within the current construction. Instead, we solve this problem using a new type of restarts that are different from the well-known doubling trick. For this, we present Algorithm 1, which applies to both SQUINT+C and METAGRAD+C (presented in the next section). It monitors a condition on the sequences (b_t) and (B_t) to trigger restarts.

Theorem 3 *Let* SQUINT+L *be the result of applying Algorithm 1 with* SQUINT+C *as* ALG. SQUINT+L *guarantees, for any comparator $\rho \in \triangle_K$,*

$$R_T^\rho \;\leq\; 2\sqrt{V_T^\rho}\left(1+\sqrt{2\Gamma_T^\rho}\right) + 10B_T\left(\Gamma_T^\rho + \ln 2\right) + 4B_T,$$

where $\Gamma_T^\rho \coloneqq \mathrm{KL}\left(\rho\|\pi\right) + \ln\left(\ln\left(\sum_{t=1}^{T-1} b_t/B_t\right) + \ln\left(2 + \sum_{t=1}^{T-1} b_t/B_t\right)\right) + 1/2.$

Note that Γ_T^ρ in Theorem 3 is equal to $\mathrm{KL}(\rho\|\pi) + O(\ln\ln T)$. Importantly, this theorem and Algorithm 1 do not depend on any initial guess B anymore. Instead, Algorithm 1 plays the starting parameters until the first time a non-zero loss is observed, and then monitors a data-dependent criterion that measures whether the loss range has increased by more than a factor that is roughly t, to decide when to trigger a restart. For most types of data, such large increases in the loss range should be rare after a few start-up rounds, so restarts should quickly stop occurring.

3. An Adaptive Method for Online Convex Optimization

We now present an extension of the METAGRAD algorithm which adapts automatically to the gradient norm in online convex optimization — we call this Lipschitz adaptive version METAGRAD+L. Recall that in the OCO setting, at each round t, the learner predicts a vector $\widehat{\boldsymbol{u}}_t$ in a closed convex set $\mathcal{U} \subset \mathbb{R}^d$, then suffers loss $\ell_t(\widehat{\boldsymbol{u}}_t)$, where $\ell_t : \mathcal{U} \to \mathbb{R}$ is a convex function. The goal of the learner is to minimize the regret $R_T^{\boldsymbol{u}} := \sum_{t=1}^T \ell_t(\widehat{\boldsymbol{u}}_t) - \sum_{t=1}^T \ell_t(\boldsymbol{u})$ with respect to the single best action $\boldsymbol{u} \in \mathcal{U}$ in hindsight. In this case, convexity of the losses implies that $\ell_t(\widehat{\boldsymbol{u}}_t) - \ell_t(\boldsymbol{u}) \leq \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle$, where $\boldsymbol{g}_t := \nabla \ell_t(\widehat{\boldsymbol{u}}_t)$, and so it suffices to control the *pseudo-regret* $\widetilde{R}_T^{\boldsymbol{u}} := \sum_{t=1}^T \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle$. We will assume that the set $\mathcal{U}$ is bounded, and denote its diameter by

$$D := \sup_{\boldsymbol{u},\boldsymbol{v} \in \mathcal{U}} \|\boldsymbol{u} - \boldsymbol{v}\|_2. \tag{8}$$

Without loss of generality, we will also assume that the set $\mathcal{U}$ is centered at $\mathbf{0}$. The proofs for this section are deferred to Appendix B. We now review the METAGRAD algorithm.

3.1. The METAGRAD Algorithm

The METAGRAD algorithm runs several sub-algorithms at each round: namely, a set of slave algorithms, which learn the best action in $\mathcal{U}$ given a learning rate η in some pre-defined grid $\mathcal{G}$, and a master algorithm, which learns the best learning rate. Through this, the METAGRAD algorithm controls the sum of *surrogate losses* $\sum_{t=1}^T f_t(\boldsymbol{u}, \eta)$ over all $\eta \in \mathcal{G}$ and $\boldsymbol{u} \in \mathcal{U}$ simultaneously, where

$$f_t(\boldsymbol{u}, \eta) := -\eta \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle + \eta^2 \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle^2, \tag{9}$$

and $\widehat{\boldsymbol{u}}_t$ is the master's prediction at round $t \in [T]$. Each slave algorithm takes as input a learning rate from a finite grid $\mathcal{G}$ (with $\lceil 1/2 \log_2 T \rceil$ points) in the form of a geometric progression and within the interval $[1/(5DG\sqrt{T}), 1/(5DG)]$, where G is an upper-bound on the norms of the gradients. In this case, G must be known in advance to construct the grid; in the proof of METAGRAD's regret bound, it is crucial for the learning rates to be in the right interval in order to invoke a certain Gaussian exp-concavity result due to Van Erven and Koolen (2016) for the surrogate losses in (9). In what follows, we let $\mathbf{S}_t := \sum_{s=1}^t \boldsymbol{g}_s \boldsymbol{g}_s^\intercal$, for $t \geq 0$.

Slaves' Predictions. Each slave $\eta \in \mathcal{G}$ starts with $\widehat{\boldsymbol{u}}_1^\eta = \mathbf{0} \in \mathcal{U}$, and at the end of round $t \geq 1$, it receives the master's prediction $\widehat{\boldsymbol{u}}_t$ and updates its own prediction in two steps:

$$\boldsymbol{u}_{t+1}^\eta := \widehat{\boldsymbol{u}}_t^\eta - \eta \boldsymbol{\Sigma}_{t+1}^\eta \boldsymbol{g}_t \left(1 + 2\eta \left(\widehat{\boldsymbol{u}}_t^\eta - \widehat{\boldsymbol{u}}_t\right)^\intercal \boldsymbol{g}_t\right), \text{ where } \boldsymbol{\Sigma}_{t+1}^\eta := \left(\tfrac{\mathbf{I}}{D^2} + 2\eta^2 \mathbf{S}_t\right)^{-1}, \tag{10}$$
$$\text{and } \widehat{\boldsymbol{u}}_{t+1}^\eta = \operatorname*{argmin}_{\boldsymbol{u} \in \mathcal{U}} \left(\boldsymbol{u}_{t+1}^\eta - \boldsymbol{u}\right)^\intercal \left(\boldsymbol{\Sigma}_{t+1}^\eta\right)^{-1} \left(\boldsymbol{u}_{t+1}^\eta - \boldsymbol{u}\right).$$

Master's Predictions. After receiving the slaves' predictions, $(\widehat{\boldsymbol{u}}_t^\eta)_{\eta\in\mathcal{G}}$, at round $t \geq 1$, the master algorithm aggregates them and outputs $\widehat{\boldsymbol{u}}_t \in \mathcal{U}$ according to:

$$\widehat{\boldsymbol{u}}_t := \frac{\sum_{\eta\in\mathcal{G}} \eta w_t^\eta \widehat{\boldsymbol{u}}_t^\eta}{\sum_{\eta\in\mathcal{G}} \eta w_t^\eta}; \quad w_t^\eta := e^{-\sum_{s=1}^{t-1} f_s(\widehat{\boldsymbol{u}}_s^\eta, \eta)}.$$

Van Erven and Koolen (2016) showed that METAGRAD has regret bounded by (1). In the next subsection, we present an extension of METAGRAD which does not require knowledge of either the horizon T or the Lipschitz constant (*i.e.* a bound on the norms of the gradients).

3.2. Lipschitz Adaptive METAGRAD

Similar to the SQUINT case, we first design a version of METAGRAD, called METAGRAD+C, which still requires an input $B > 0$ (in this case, B/D is the initial estimate of the Lipschitz bound). We then present METAGRAD+L which sets this parameter online. For now, we consider a fixed $B > 0$. We define $b_t := D\|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_2 = D\|\boldsymbol{g}_t\|_2$, for $t \geq 1$, and $b_0 := B$. We denote the running maximum of (b_t) by $B_t := \max_{0\leq s\leq t} b_s$. We will also require a *clipped* version of the gradient vector $\bar{\boldsymbol{g}}_t = \boldsymbol{g}_t \cdot B_{t-1}/B_t$, and denote by $\bar{r}_t^{\boldsymbol{u}} = \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \bar{\boldsymbol{g}}_t\rangle$ the *clipped instantaneous pseudo-regret* with respect to $\boldsymbol{u} \in \mathcal{U}$. In addition, it will be useful to define

$$\bar{f}_t(\boldsymbol{u}, \eta) := -\eta \bar{r}_t^{\boldsymbol{u}} + (\eta \bar{r}_t^{\boldsymbol{u}})^2 \quad \text{and} \quad \bar{\mathbf{S}}_t := \sum_{s=1}^{t} \bar{\boldsymbol{g}}_s \bar{\boldsymbol{g}}_s^\intercal. \tag{11}$$

Recall that in the original METAGRAD, the horizon T and the Lipschitz constant G were required to construct the grid of learning rates. We circumvent this by defining an infinite grid $\mathcal{G}$ in which, at any given round $t \geq 1$, only a finite number of (active) slaves — up to $\log_2 t$ many — output a non-zero prediction. Each slave η in this grid receives a prior weight $\pi(\eta) \in [0, 1]$, where $\sum_{\eta\in\mathcal{G}} \pi(\eta) = 1$. Given input $B > 0$ to METAGRAD+C, the grid $\mathcal{G}$ and the prior π are defined by

$$\mathcal{G} := \left\{ \eta_i := \frac{1}{5B2^i} : i \in \mathbb{N} \cup \{0\} \right\}; \quad \pi(\eta_i) := \frac{1}{(i+1)(i+2)}, \; i \in \mathbb{N} \cup \{0\}. \tag{12}$$

The subset of active slaves $\mathcal{A}_t$ at a round $t \geq 1$ is given by

$$\mathcal{A}_t := \left\{ \eta \in \mathcal{G} \cap \left[0, \tfrac{1}{5B_{t-1}}\right] : s_\eta < t \right\}, \text{ with } s_\eta := \min\left\{ t \geq 0 : \frac{1}{\eta} \leq D \sum_{s=1}^{t} \|\bar{\boldsymbol{g}}_s\|_2 + B_t \right\}. \tag{13}$$

We note that restricting the slaves (or learning rates) to the set $\mathcal{G}_t := \mathcal{G} \cap [0, 1/(5B_{t-1})]$ is similar in principle to clipping the upper integral range in the SQUINT+C case.

Slaves' Predictions. A slave $\eta \in \mathcal{G} \cap [0, 1/(5B_{t-1})]$ issues predictions $\widehat{\boldsymbol{u}}_t^\eta = \mathbf{0}$ in all rounds $t \leq s_\eta + 1$. From then on (*i.e.* at the end of round $t \geq s_\eta + 1$), it receives the master's prediction $\widehat{\boldsymbol{u}}_t$ as input and updates its own prediction in two steps:

$$\boldsymbol{u}_{t+1}^\eta := \widehat{\boldsymbol{u}}_t^\eta - \eta \boldsymbol{\Sigma}_{t+1}^\eta \bar{\boldsymbol{g}}_t \left(1 + 2\eta \left(\widehat{\boldsymbol{u}}_t^\eta - \widehat{\boldsymbol{u}}_t\right)^\intercal \bar{\boldsymbol{g}}_t\right), \text{ where } \boldsymbol{\Sigma}_{t+1}^\eta := \left(\tfrac{\mathbf{I}}{D^2} + 2\eta^2 \left(\bar{\mathbf{S}}_t - \bar{\mathbf{S}}_{s_\eta}\right)\right)^{-1},$$
$$\text{and } \widehat{\boldsymbol{u}}_{t+1}^\eta = \operatorname*{argmin}_{\boldsymbol{u}\in\mathcal{U}} \left(\boldsymbol{u}_{t+1}^\eta - \boldsymbol{u}\right)^\intercal \left(\boldsymbol{\Sigma}_{t+1}^\eta\right)^{-1} \left(\boldsymbol{u}_{t+1}^\eta - \boldsymbol{u}\right).$$

Master's Predictions. At each round $t \geq 1$, the master algorithm receives the slaves' predictions $(\widehat{\boldsymbol{u}}_t^\eta)_{t \in \mathcal{A}_t}$ and outputs

$$\widehat{\boldsymbol{u}}_t = \frac{\sum_{\eta \in \mathcal{A}_t} \eta w_t^\eta \widehat{\boldsymbol{u}}_t^\eta}{\sum_{\eta \in \mathcal{A}_t} \eta w_t^\eta}, \quad \text{where } w_t^\eta := \pi(\eta) e^{-\sum_{s=s_\eta+1}^{t-1} \bar{f}_s(\widehat{\boldsymbol{u}}_s^\eta, \eta)}. \tag{14}$$

Remark 4 (Number of Active Slaves) *At any round $t \geq 1$, the number of active slaves is at most $\lceil \log_2 t \rceil$. In fact, if $\eta \in \mathcal{A}_t$, then by definition $\eta \geq 1/(D \sum_{s=1}^{s_\eta} \|\boldsymbol{g}_s\|_2 + B_{s_\eta}) \geq 1/(tB_{t-1})$ (since $s_\eta \leq t-1$), and thus $\mathcal{A}_t \subset [1/(tB_{t-1}), 1/(5B_{t-1})]$. Since $\mathcal{A}_t$ is a grid in the form of a geometric progression with common ratio 2, there are at most $\lceil \log_2 t \rceil$ slaves in $\mathcal{A}_t$.*

To motivate METAGRAD+C, we define the potential function after $t \geq 0$ rounds by

$$\Phi_t := \pi(\mathcal{G}_t \setminus \mathcal{A}_t) + \sum_{\eta \in \mathcal{A}_t} \pi(\eta) e^{-\sum_{s=s_\eta+1}^{t} \bar{f}_s(\widehat{\boldsymbol{u}}_s^\eta, \eta)}, \quad \text{where } \mathcal{G}_t := \mathcal{G} \cap \left[0, \tfrac{1}{5B_{t-1}}\right]. \tag{15}$$

Let $\boldsymbol{u} \in \mathcal{U}$. Recall that the pseudo-regret is defined by $\tilde{R}_T^{\boldsymbol{u}} := \sum_{t=1}^T \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle$. We now defined its *clipped* version by $\bar{R}_T^{\boldsymbol{u}} := \sum_{t=1}^T \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \bar{\boldsymbol{g}}_t \rangle$. For $r_t^{\boldsymbol{u}} := \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle$, we have, similarily to (4),

$$\tilde{R}_T^{\boldsymbol{u}} - \bar{R}_T^{\boldsymbol{u}} = \sum_{t=1}^T (r_t^{\boldsymbol{u}} - \bar{r}_t^{\boldsymbol{u}}) \;=\; \sum_{t=1}^T (B_t - B_{t-1}) \frac{r_t^{\boldsymbol{u}}}{B_t} \;\leq\; B_T - B_0, \tag{16}$$

where the last inequality follows from the Cauchy-Schwarz inequality and the fact that $\mathcal{U}$ has diameter D, which together imply that $|r_t^{\boldsymbol{u}}| \leq B_t$. Using the inequality $e^{x-x^2} - 1 \leq x$, which holds for all $x \geq -1/2$, one can shown that the potential is a decreasing function of the number of rounds:

Lemma 5 METAGRAD+C *guarantees that $\Phi_T \leq \cdots \leq \Phi_0 = 1$, for all $T \in \mathbb{N}$.*

We now give an upper-bound on $\bar{R}_T^{\boldsymbol{u}}$ in terms of the clipped 'variance' $\bar{V}_T^{\boldsymbol{u}} := \sum_{t=1}^T (\bar{r}_t^{\boldsymbol{u}})^2$;

Theorem 6 *Given input $B > 0$, the clipped pseudo-regret for* METAGRAD+C *is bounded by*

$$\bar{R}_T^{\boldsymbol{u}} \leq 3\sqrt{\bar{V}_T^{\boldsymbol{u}} C_T} + 15 B_T C_T, \quad \textit{for any } \boldsymbol{u} \in \mathcal{U},$$

where $C_T := d \ln\left(1 + \frac{2\sum_{t=0}^{T-1} b_t^2}{25 d B_{T-1}^2}\right) + 2\ln\left(\log_2^+ \frac{\sqrt{\sum_{t=1}^T b_t^2}}{B} + 3\right) + 2$ *and* $\log_2^+ = 0 \vee \log_2$.

Remark 7 *For $\boldsymbol{u} \in \mathcal{U}$, we can relate the clipped pseudo-regret to the ordinary regret via $R_T^{\boldsymbol{u}} \leq \tilde{R}_T^{\boldsymbol{u}} \leq \bar{R}_T^{\boldsymbol{u}} + B_T$ (see (16)) and on the right-hand side we can also use that $\bar{V}_T^{\boldsymbol{u}} \leq V_T^{\boldsymbol{u}}$.*

An important aspect to note from Theorem 6 is that the ratio $\sqrt{\sum_{t=1}^T b_t^2}/B$, could in principle be arbitrarily large if the input B is too small compared to the actual norms of the gradients (for SQUINT it was the ratio B_{T-1}/B which was problematic). To resolve this issue, we use the same restart approach as in the SQUINT case:

Theorem 8 *Let* METAGRAD+L *be the result of applying Algorithm 1 to* METAGRAD+C. *Then the actual and linearised regrets for* METAGRAD+L *are both bounded by*

$$R_t^{\boldsymbol{u}} \;\leq\; \tilde{R}_T^{\boldsymbol{u}} \;\leq\; 3\sqrt{V_T^{\boldsymbol{u}}\Gamma_T} + 15B_T\Gamma_T + 4B_T \quad \textit{for all } \boldsymbol{u} \in \mathcal{U},$$

where $\Gamma_T \coloneqq 2d \ln\left(1 + \frac{2}{25d}\sum_{t=1}^T \frac{b_t^2}{B_t^2}\right) + 4\ln\left(\log_2^+ \sqrt{\sum_{t=1}^T (\sum_{s=1}^t \frac{b_s}{B_s})^2} + 3\right) + 4 = O(d \ln T)$.

Theorem 8 replaces the ratio $\sqrt{\sum_{t=1}^T b_t^2}/B$ appearing in the (clipped) pseudo-regret bound of METAGRAD+C by $\sigma_T \coloneqq \sqrt{\sum_{t=1}^T (\sum_{s=1}^t b_s/B_s)^2}$. The latter is independent of the input B and is always smaller than $T^{3/2}$; this is perfectly affordable since σ_T appears inside a $\ln\ln$. Our reason for including the linearised regret $\tilde{R}_T^{\boldsymbol{u}}$ in Theorem 8 is that a bound on it in terms of $V_T^{\boldsymbol{u}}$ is the precondition for fast rate results in individual-sequence settings based on curvature (Van Erven and Koolen, 2016) and in statistical settings under certain (Bernstein type) conditions (Koolen et al., 2016).

4. Efficient Implementation Through a Reduction to the Ball

Using METAGRAD (+C or +L), the computation of each slave prediction $\widehat{\boldsymbol{u}}_t^\eta$ requires a projection onto an arbitrary convex set $\mathcal{U}$ in *Mahalanobis distance*. Numerically, this typically requires $O(d^p)$ floating point operations (flops), for some $p \in \mathbb{N}$ which depends on the geometry of the set $\mathcal{U}$. Since p can be large in many applications, evaluating $\widehat{\boldsymbol{u}}_t^\eta$ for each slave η can become computationally prohibitive, especially when the number of slaves grows with T; for the METAGRAD versions discussed in this paper, there can be up to $\lceil \log_2 T \rceil$ slaves at round $T \geq 1$ (see Remark 4).

The goal of this section is to streamline these computations, which we will do in two steps. In Section 4.1, we will describe an efficient implementation of METAGRAD on the ball. The main idea here is that the Mahalanobis projections onto the ball, which are performed by the slaves, can reuse a common matrix decomposition. In Section 4.2, we will then obtain an algorithm for any bounded convex set $\mathcal{U}$ by applying the black-box reduction of Cutkosky and Orabona (2018) to METAGRAD on the ball enclosing $\mathcal{U}$. We show (Theorem 10) that the reduction also transports variance bounds. The techniques discussed here also apply to the versions of METAGRAD presented in the previous section. However, to simplify the presentation, we will only focus on the original METAGRAD. The proofs for this section are deferred to Appendix C.

4.1. Efficient Implementation of METAGRAD on the Ball

Suppose that $\mathcal{U}$ is the ball of diameter D: $\mathcal{U} = \mathcal{B}_D \coloneqq \left\{\boldsymbol{u} \in \mathbb{R}^d \colon \|\boldsymbol{u}\|_2 \leq D/2\right\}$. To compute the slave's prediction $\widehat{\boldsymbol{u}}_{t+1}^\eta$, the following quadratic program needs to be solved for each η:

$$\widehat{\boldsymbol{u}}_{t+1}^\eta = \operatorname*{argmin}_{\boldsymbol{u}\in\mathcal{U}} \left(\boldsymbol{u}_{t+1}^\eta - \boldsymbol{u}\right)^\intercal \left(\boldsymbol{\Sigma}_{t+1}^\eta\right)^{-1} \left(\boldsymbol{u}_{t+1}^\eta - \boldsymbol{u}\right), \tag{17}$$

where $\boldsymbol{u}_{t+1}^\eta$ (the unprojected prediction) and $\boldsymbol{\Sigma}_{t+1}^\eta = (\mathbf{I}/D^2 + 2\eta^2\mathbf{S}_t)^{-1}$ (the co-variance matrix) are defined in (10). Since $\mathcal{U}$ is a ball and $\boldsymbol{\Sigma}_{t+1}^\eta$ is symmetric positive-definite, (17) can be solved in $O(d^3)$ by performing a singular value decomposition of $\boldsymbol{\Sigma}_{t+1}^\eta$. Instead of doing this singular value decomposition separately for each η, we can be a little more efficient by doing a singular value decomposition of $\mathbf{S}_t$ once and then using the following lemma:

Algorithm 2 Reducing an OCO problem on $\mathcal{U} \subset R^d$ to one on a ball.

Require: A bounded convex set $\mathcal{U} \subset \mathbb{R}^d$ with diameter $D > 0$, a Lipschitz bound $G > 0$.
We write METAGRAD(D) for METAGRAD applied to the ball $\mathcal{B}_D$ enclosing $\mathcal{U}$.
for $t = 1$ **to** T **do**
Get $\widehat{\boldsymbol{u}}_t$ from METAGRAD(D) ; `//The initial input to METAGRAD is` $B = DG$.
Predict $\widehat{\boldsymbol{w}}_t = \Pi_{\mathcal{U}}(\widehat{\boldsymbol{u}}_t)$ and receive $\mathring{\boldsymbol{g}}_t = \nabla \ell_t(\widehat{\boldsymbol{w}}_t)$;
Set $\boldsymbol{g}_t \in \frac{1}{2}\left(\mathring{\boldsymbol{g}}_t + \|\mathring{\boldsymbol{g}}_t\| \partial \, \mathrm{d}_{\mathcal{U}}(\widehat{\boldsymbol{u}}_t)\right)$;
Send $\boldsymbol{g}_t$ to METAGRAD(D) ;
end for

Lemma 9 *Let $\boldsymbol{\Lambda}_t \coloneqq \operatorname{diag}((\lambda_t^i)_{i\in[d]})$ and $\mathbf{Q}_t$ be the matrices of eigenvalues and eigenvectors of $\mathbf{S}_t$, respectively, such that $\mathbf{Q}_t\mathbf{S}_t\mathbf{Q}_t^\intercal = \boldsymbol{\Lambda}_t$ and $\mathbf{Q}_t\mathbf{Q}_t^\intercal = \mathbf{I}$.*[2] *Then the solution of* (17) *is*

$$\widehat{\boldsymbol{u}}_{t+1}^\eta = \begin{cases} \boldsymbol{u}_{t+1}^\eta, & \text{if } \boldsymbol{u}_{t+1}^\eta \in \mathcal{U}, \\ \mathbf{Q}_t^\intercal (x_t^\eta \mathbf{I} + 2\eta^2 \boldsymbol{\Lambda}_t)^{-1} \mathbf{Q}_t \boldsymbol{v}_{t+1}^\eta, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{v}_{t+1}^\eta \coloneqq \left(\mathbf{I}/D^2 + 2\eta^2 \mathbf{S}_t\right) \boldsymbol{u}_{t+1}^\eta$ and the scalar x_t^η is the unique solution of

$$\rho_t^\eta(x) \coloneqq \sum_{i=1}^d \frac{\langle \boldsymbol{e}_i, \mathbf{Q}_t \boldsymbol{v}_{t+1}^\eta \rangle^2}{(x + 2\eta^2 \lambda_t^i)^2} = \frac{D^2}{4}. \tag{18}$$

Since ρ_t^η in (18) is strictly convex and decreasing, $\rho_t^\eta(x) = D^2/4$ can be solved using Newton's method in linear time.

A further improvement leverages the rank-one update $\mathbf{S}_t = \mathbf{S}_{t-1} + \boldsymbol{g}_t \boldsymbol{g}_t^\intercal$ to update $\boldsymbol{\Lambda}_{t-1}$ and $\mathbf{Q}_{t-1}$. It is possible to compute the new matrices $\boldsymbol{\Lambda}_t$ and $\mathbf{Q}_t$ in, respectively, $O(d^2)$ and $O(d^3)$ flops, where the latter cost for computing $\mathbf{Q}_t$ is only due to matrix multiplication (rather than a full singular value decomposition) (Bunch et al., 1978), and thus admits an efficient parallel implementation.

4.2. A Reduction to the Ball

In this subsection, we extend the black-box technique of Cutkosky and Orabona (2018) to reduce an OCO problem on an arbitrary bounded convex set $\mathcal{U} \subset \mathbb{R}^d$ to one on a ball, where the implementation of METAGRAD from the previous subsection can be applied.

Let D be the diameter of a closed bounded convex set $\mathcal{U} \subset \mathbb{R}^d$ as in (8), so that the ball $\mathcal{B}_D$ of radius $D/2$ encloses $\mathcal{U}$. As in the previous section, we again assume, without loss of generality, that $\mathcal{U}$ is centered at $\mathbf{0}$. For $\boldsymbol{u} \in \mathcal{U}$, we denote $\mathrm{d}_{\mathcal{U}}(\boldsymbol{u}) = \min_{\boldsymbol{w}\in\mathcal{U}} \|\boldsymbol{u} - \boldsymbol{w}\|_2$ the *distance function* from the set $\mathcal{U}$, and we define $\Pi_{\mathcal{U}}(\boldsymbol{u}) \coloneqq \{\boldsymbol{w} \in \mathcal{U} : \|\boldsymbol{w} - \boldsymbol{u}\|_2 = \mathrm{d}_{\mathcal{U}}(\boldsymbol{u})\}$. Algorithm 2 reduces the OCO problem on the set $\mathcal{U}$ to one on the ball $\mathcal{B}_D$, where the METAGRAD algorithm is used as a black-box to solve it. We note that Algorithm 2 (including its METAGRAD subroutine) only performs a single projection (applied to the output of METAGRAD) onto the set $\mathcal{U}$ in *Euclidean distance* — as opposed the *time-varying Mahalanobis distance* (17); the METAGRAD subroutine only performs projections onto the ball $\mathcal{B}_D$, which can be done efficiently as described in the previous subsection.

2. The existence of such a $\mathbf{Q}_t$ and $\boldsymbol{\Lambda}_t$ is guaranteed due to $\mathbf{S}_t$ being symmetric positive-definite.

In the next theorem, we assume that a Lipschitz bound $G > 0$ is known in advance[3], and we let $\mathring{R}_T^{\boldsymbol{u}} := \sum_{t=1}^T \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{u}, \mathring{\boldsymbol{g}}_t \rangle$ and $\mathring{V}_T^{\boldsymbol{u}} := \sum_{t=1}^T \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{u}, \mathring{\boldsymbol{g}}_t \rangle^2$ be the pseudo-regret and 'variance' corresponding to Algorithm 2. We now show that the (pseudo) regret guarantee of METAGRAD readily transfers to Algorithm 2 with almost no overhead:

Theorem 10 *Let $D > 0$, and suppose that the* METAGRAD*(D) subroutine of Algorithm 2 achieves a pseudo-regret bound of the form*

$$\tilde{R}_T^{\boldsymbol{u}} \leq \sqrt{V_T^{\boldsymbol{u}} \Gamma_T} + B\Gamma_T, \textit{ for all } \boldsymbol{u} \in \mathcal{B}_D,$$

where $\tilde{R}_t^{\boldsymbol{u}} := \sum_{t=1}^T \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle$, $V_t^{\boldsymbol{u}} := \sum_{t=1}^T \langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t \rangle^2$, and $\Gamma_T = O(d \ln(T/d))$. Then, Algorithm 2 guarantees:

$$\sum_{t=1}^T \left(\ell_t(\widehat{\boldsymbol{w}}_t) - \ell_t(\boldsymbol{u}) \right) \leq \mathring{R}_T^{\boldsymbol{u}} \leq \sqrt{\mathring{V}_T^{\boldsymbol{u}} \Gamma_T} + 4B\Gamma_T, \textit{ for all } \boldsymbol{u} \in \mathcal{U}.$$

From the standard black-box reduction of Cutkosky and Orabona (2018), we would obtain an unsatisfactory result in which $\mathring{V}_T^{\boldsymbol{u}}$ would be measured in terms of the fake gradients $\boldsymbol{g}_t$ that are supplied internally to METAGRAD(D) instead of the actual gradients $\mathring{\boldsymbol{g}}_t$. As this would not be sufficient to adapt to the easiness conditions described in the introduction, the proof of Theorem 10 involves an extra step to relate the variance term back to the actual gradients.

5. Conclusion

We present algorithms that adapt to the Lipschitz constant of the loss for OCO and experts, with hardly any overhead in terms of regret or computation compared to their previous counterparts that had to know the Lipschitz constant up-front. This fits into a larger picture of understanding which types of adaptivity are possible at which price in terms of additional regret and additional run time.

One surprising conclusion from our work is the following observation: for OCO, Cutkosky and Boahen (2017a) show that in general it is not possible to be adaptive to both the Lipschitz constant and the norm of the comparator $\|\boldsymbol{u}\|$ at the same time. Since the analogue of $\|\boldsymbol{u}\|$ in the expert setting is the complexity measure $\mathrm{KL}(\rho\|\pi)$, we might therefore conjecture that Lipschitz adaptivity would also be incompatible with a quantile regret bound in terms of $\mathrm{KL}(\rho\|\pi)$. However, our results show this conjecture to be false: for experts there is no conflict. This holds even in cases where the prior π is not uniform, and our results can easily be extended to a countably infinite number of experts where $\mathrm{KL}(\rho\|\pi)$ cannot even be uniformly bounded.

A final and very interesting question is when is it possible to exploit scenarios with large Lipschitz constants or loss ranges that occur only very infrequently. An example of this is found in statistical learning with heavy-tailed loss distributions. For such scenarios, martingale methods (related to our potential functions) suggest that it may be necessary to replace in $f_t(\boldsymbol{u}, \eta)$ the 'surrogate' negative quadratic term that our algorithms include in the exponent by another function appropriate for the specific distribution (Howard et al., 2018, Table 3). It is not currently clear what individual sequence analogues can be obtained.

3. If one uses METAGRAD+C or METAGRAD+L as the subroutine in Algorithm 2 instead of METAGRAD, then a Lipschitz bound need not be known in adavance; a version of Theorem 10 with different constants would still hold in this case.

Acknowledgments

We thank the anonymous reviewers for feedback that improved the presentation. Part of this work was performed while Zakaria Mhammedi was conducting an internship at the Centrum Wiskunde & Informatica (CWI). This work was also supported by the Australian Research Council and Data61.

References

Peter L. Bartlett, Elad Hazan, and Alexander Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 65–72, 2007.

James R. Bunch, Christopher P. Nielsen, and Danny C. Sorensen. Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik*, 31(1):31–48, 1978.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, 1997.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

Kamalika Chaudhuri, Yoav Freund, and Daniel Hsu. A parameter-free hedging algorithm. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pages 297–305, 2009.

Alexey V. Chernov and Vladimir Vovk. Prediction with advice of unknown number of experts. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 117–125, 2010.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Le, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Proc. of the 25th Annual Conference on Learning Theory (COLT)*, pages 6.1–6.20, 2012.

Ashok Cutkosky. Artificial constraints and lipschitz hints for unconstrained online learning. *arXiv preprint arXiv:1902.09013*, 2019.

Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Proceedings of the 30th Annual Conference on Learning Theory (COLT)*, pages 643–677, 2017a.

Ashok Cutkosky and Kwabena A. Boahen. Stochastic and adversarial online learning without hyperparameters. In *Advances in Neural Information Processing Systems*, pages 5059–5067, 2017b.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1493–1529, 2018.

Chuong B. Do, Quoc V. Le, and Chuan-Sheng Foo. Proximal regularization for online and batch learning. In *Proc. of the 26th Annual International Conference on Machine Learning (ICML)*, pages 257–264, 2009.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Tim van Erven and Wouter M. Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 3666–3674, 2016.

Dylan J Foster, Alexander Rakhlin, and Karthik Sridharan. Adaptive online learning. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3375–3383, 2015.

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Zigzag: A new approach to adaptive online learning. In *Proc. of the 2017 Annual Conference on Learning Theory (COLT)*, pages 876–924, 2017.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proc. of the 27th Annual Conference on Learning Theory (COLT)*, pages 176–196, 2014.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3–4):157–325, 2016.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2-3):165–188, 2010.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Steve Howard, Aaditya Ramdas, John McAuliffe, and Jasjeet Sekhon. Exponential line-crossing inequalities. *ArXiv e-prints*, August 2018.

Wouter M. Koolen. The relative entropy bound for Squint, August 2015. Adversarial Intelligence blog.

Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pages 1155–1175, 2015.

Wouter M. Koolen, Peter D. Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems*, pages 4457–4465, 2016.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: Adaptive NormalHedge. In *Proc. of The 28th Annual Conference on Learning Theory (COLT)*, pages 1286–1304, 2015.

Brendan McMahan and Jacob Abernethy. Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2724–2732, 2013.

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 244–256, 2010.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proc. of the 26th Annual Conference on Learning Theory (COLT)*, pages 993–1019, 2013.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Rachel Ward, Xiaoxia Wu, and Léon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *ArXiv:1806.01811 preprint*, 2018.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *ArXiv:1404.1356v2 preprint*, 2014.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106 (1):119–141, 2017.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th Annual International Conference on Machine Learning (ICML)*, pages 928–936, 2003.

Appendix A. Proofs of Section 2

Proof of Lemma 1 We proceed by induction on T. By definition $\Phi_0 = 0$. For $T \geq 0$, the definition (5) gives

$$\Phi_{T+1} = \underbrace{\sum_k \pi_k \int_0^{\frac{1}{2B_T}} \frac{e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} \left(e^{\eta \bar{r}_{T+1}^k - \eta^2 (\bar{r}_{T+1}^k)^2} - 1 \right)}{\eta} \, \mathrm{d}\eta}_{=:Q_1} + \underbrace{\sum_k \pi_k \int_0^{\frac{1}{2B_T}} \frac{e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} - 1}{\eta} \, \mathrm{d}\eta}_{=:Q_2} .$$

To control the first term Q_1, we apply the so-called 'prod bound' $e^{x-x^2} \leq 1 + x$ for $x \geq -1/2$ (Cesa-Bianchi et al., 2007) to $x = \eta \bar{r}_{T+1}^k$, which we may do as $\eta \bar{r}_{T+1} \geq -\frac{1}{2B_T} B_T$. Linearity and the definition of the weights (6), yield the following upper-bound on the term Q_1

$$\sum_k \pi_k \int_0^{\frac{1}{2B_T}} \frac{e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} \eta \bar{r}_{T+1}^k}{\eta} \, \mathrm{d}\eta = \left\langle \sum_k \pi_k \int_0^{\frac{1}{2B_T}} e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} \left(\widehat{\boldsymbol{p}}_{T+1} - \boldsymbol{e}_k \right) \mathrm{d}\eta, \bar{\boldsymbol{l}}_{T+1} \right\rangle = 0.$$

To control the second term Q_2, we extend the range of the integral to find

$$Q_2 \leq \sum_k \pi_k \int_0^{\frac{1}{2B_{T-1}}} \frac{e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} - 1}{\eta} \, \mathrm{d}\eta + \ln \frac{B_T}{B_{T-1}} = \Phi_T + \ln \frac{B_T}{B_{T-1}}.$$

■

Proof of Lemma 2 For any $\epsilon \in [0, 1/(2B_{T-1})]$, we may split the potential (5) as follows

$$\Phi_T \;=\; \underbrace{\sum_k \pi_k \int_0^{\epsilon} \frac{e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} - 1}{\eta} \,\mathrm{d}\eta}_{=:Q_1} + \underbrace{\sum_k \pi_k \int_{\epsilon}^{\frac{1}{2B_{T-1}}} \frac{e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k} - 1}{\eta} \,\mathrm{d}\eta}_{=:Q_2}.$$

For convenience, let us introduce $\bar{b}_t := \max_k |\bar{r}_t^k| = b_t \cdot B_{t-1}/B_t$ and abbreviate $\bar{S}_T := \sum_{t=1}^T \bar{b}_t$. To bound the left term Q_1 from below, we use $e^x - 1 \geq x$. Then combined with $\bar{R}_T^k \geq -\bar{S}_T$ and $\bar{V}_T^k \leq \sum_{t=1}^{T-1} \bar{b}_t^2 \leq B_{T-1}\bar{S}_T$ we find

$$Q_1 \;\geq\; \sum_k \pi_k \int_0^{\epsilon} \bar{R}_T^k - \eta \bar{V}_T^k \,\mathrm{d}\eta \;\geq\; -\left(\epsilon + \frac{\epsilon^2}{2} B_{T-1}\right) \bar{S}_T.$$

For the right term Q_2, we use KL duality to find

$$\begin{aligned} Q_2 \;&=\; \sum_k \pi_k \int_{\epsilon}^{\frac{1}{2B_{T-1}}} \frac{e^{\eta \bar{R}_T^k - \eta^2 \bar{V}_T^k}}{\eta} \,\mathrm{d}\eta + \ln\left(2B_{T-1}\epsilon\right), \\ &\geq\; e^{-\mathrm{KL}(\rho\|\pi)} \int_{\epsilon}^{\frac{1}{2B_{T-1}}} \frac{e^{\eta \bar{R}_T^\rho - \eta^2 \bar{V}_T^\rho}}{\eta} \,\mathrm{d}\eta + \ln\left(2B_{T-1}\epsilon\right). \end{aligned}$$

Way pick the admissible $\epsilon = 1/(2(\bar{S}_T + B_{T-1}))$ for which $\left(\epsilon + B_{T-1} \cdot \epsilon^2/2\right) \bar{S}_T \leq 1/2$ (as it is increasing in $\bar{S}_T \geq 0$ and decreasing in $B_{T-1} \geq 0$), and find

$$\Phi_T \;\geq\; e^{-\mathrm{KL}(\rho\|\pi)} \int_{\epsilon}^{\frac{1}{2B_{T-1}}} \frac{e^{\eta \bar{R}_T^\rho - \eta^2 \bar{V}_T^\rho}}{\eta} \,\mathrm{d}\eta - \frac{1}{2} - \ln\left(1 + \frac{\bar{S}_T}{B_{T-1}}\right),$$

which we may reorganise to

$$Q_3 := \ln \int_{\frac{1}{2(\bar{S}_T + B_{T-1})}}^{\frac{1}{2B_{T-1}}} \frac{e^{\eta \bar{R}_T^\rho - \eta^2 \bar{V}_T^\rho}}{\eta} \,\mathrm{d}\eta \;\leq\; \mathrm{KL}\left(\rho\|\pi\right) + \ln\left(\Phi_T + \frac{1}{2} + \ln\left(1 + \frac{\bar{S}_T}{B_{T-1}}\right)\right).$$

The argument to bound the integral in Q_3 splits in 3 cases. Let us abbreviate $R \equiv \bar{R}_T^\rho$ and $V \equiv \bar{V}_T^\rho$. Let $\hat{\eta} = \frac{R}{2V}$ be the maximiser of $\eta \to \eta R - \eta^2 V$.

1. First the important case, where $[\hat{\eta} - 1/\sqrt{2V}, \hat{\eta}] \subseteq [1/(2(\bar{S}_T + B_{T+1})), 1/(2B_{T-1})]$. Then

$$\begin{aligned} Q_3 \;&\geq\; \ln \int_{\hat{\eta} - \frac{1}{\sqrt{2V}}}^{\hat{\eta}} \frac{e^{\eta R - \eta^2 V}}{\eta} \,\mathrm{d}\eta \;\geq\; \ln \int_{\hat{\eta} - \frac{1}{\sqrt{2V}}}^{\hat{\eta}} \frac{e^{\left(\hat{\eta} - \frac{1}{\sqrt{2V}}\right) R - \left(\hat{\eta} - \frac{1}{\sqrt{2V}}\right)^2 V}}{\eta} \,\mathrm{d}\eta \\ &=\; \left(\hat{\eta} - \frac{1}{\sqrt{2V}}\right) R - \left(\hat{\eta} - \frac{1}{\sqrt{2V}}\right)^2 V + \ln\ln \frac{\hat{\eta}}{\hat{\eta} - \frac{1}{\sqrt{2V}}} \\ &=\; \frac{R^2}{4V} - \frac{1}{2} + \ln\ln \frac{1}{1 - \frac{\sqrt{2V}}{R}} \;\geq\; \frac{1}{2}\left(\frac{R}{\sqrt{2V}} - 1\right)^2 \end{aligned}$$

where the last inequality uses $\ln\ln(x/(x-1)) \geq 1-x$ for $x \geq 1$, which can be easily verified by a one-dimensional plot. We conclude

$$R \;\leq\; \sqrt{2V}\left(1+\sqrt{2\,\mathrm{KL}\left(\rho\|\pi\right)+2\ln\left(\Phi_T+\frac{1}{2}+\ln\left(1+\frac{\bar{S}_T}{B_{T-1}}\right)\right)}\right).$$

2. Then in the case where $\hat{\eta}-1/\sqrt{2V} < 1/\bar{S}_T$, we have

$$R \;<\; \sqrt{2V}+\frac{2V}{\bar{S}_T} \;\leq\; \sqrt{2V}+2B_{T-1},$$

and we are done again.

3. We come to the final case where $\hat{\eta} > 1/(2B_{T-1})$, meaning that $R > V/B_{T-1}$. Here we use that for any $u \in [1/(2(\bar{S}_T+B_{T-1})), 1/(2B_{T-1})]$

$$Q_3 \;\geq\; \ln\int_u^{\frac{1}{2B_{T-1}}} \frac{e^{uR-u^2V}}{\eta}\,\mathrm{d}\eta \;\geq\; uR(1-uB_{T-1})+\ln\ln\frac{1}{2uB_{T-1}},$$

and hence

$$R \leq \frac{Q_3-\ln\ln\frac{1}{2uB_{T-1}}}{u(1-uB_{T-1})}.$$

Picking the feasible $u=(5-\sqrt{5})/(10B_{T-1})$ and using $-\ln\ln(5/(5-\sqrt{5})) \leq \ln 2$ yields

$$R \;\leq\; 5B_{T-1}\left(\mathrm{KL}\left(\rho\|\pi\right)+\ln\left(\Phi_T+\frac{1}{2}+\ln\left(1+\frac{\bar{S}_T}{B_{T-1}}\right)\right)+\ln 2\right).$$

Finally, using the fact that

$$\frac{\bar{S}_T}{B_{T-1}} = \frac{1}{B_{T-1}}\sum_{t=1}^{T}\frac{B_{t-1}}{B_t}b_t \leq 1+\sum_{t=1}^{T-1}\frac{b_t}{B_t}$$

concludes the proof.

■

Proof of Theorem 3 The idea of the proof is to analyse the rounds in three parts, as shown in Figure 1.

For comparator $\rho \in \triangle_K$, $B>0$ and $\tau_1,\tau_2 \in \mathbb{N}$ such that $\tau_1<\tau_2$, we define the regret $R^{\rho}_{(\tau_1,\tau_2]}$ and variance $V^{\rho}_{(\tau_1,\tau_2]}$ of SQUINT+C started at round τ_1+1 (with input B_{τ_1}) and terminated after round τ_2 by

$$R^{\rho}_{(\tau_1,\tau_2]} := \sum_{t=\tau_1+1}^{\tau_2}\mathbb{E}_{\rho(k)}\left[r_t^k\right], \quad V^{\rho}_{(\tau_1,\tau_2]} := \sum_{t=\tau_1+1}^{\tau_2}\mathbb{E}_{\rho(k)}\left[(r_t^k)^2\right].$$

We also define

$$\Gamma^{\rho}_{(\tau_1,\tau_2]} \;:=\; \mathrm{KL}\left(\rho\|\pi\right)+\ln\left(\ln\sum_{t=1}^{\tau_2-1}\frac{b_t}{B_t}+\frac{1}{2}+\ln\left(2+\sum_{t=\tau_1+1}^{\tau_2-1}\frac{b_t}{B_t}\right)\right).$$

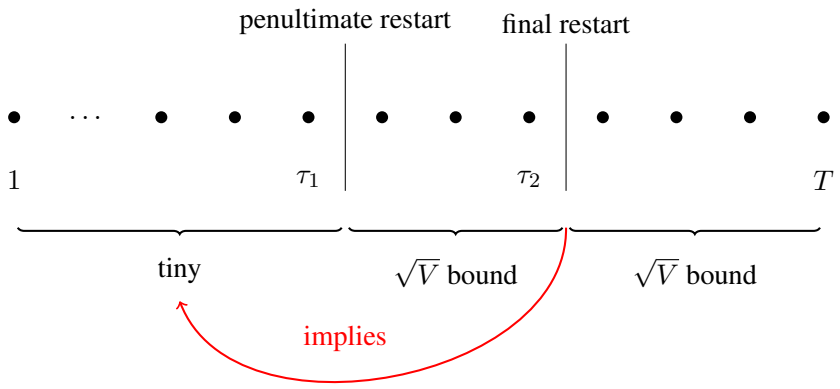

Figure 1: Regret bounding strategy; most general case

Lemma 11 *Let $\rho \in \triangle_K$ and $\tau_1, \tau_2 \in \mathbb{N}$ be such that $\tau_1 < \tau_2$. Suppose that $B_{\tau_2-1}/B_{\tau_1} \leq \sum_{t=1}^{\tau_2-1} b_t/B_t$ (this corresponds to the case where the restart condition in line 2 of Algorithm 1 is not triggered at the end of round $\tau_2 - 1$). Then, the regret $R^\rho_{(\tau_1,\tau_2]}$ of* SQUINT+C *satisfies:*

$$R^\rho_{(\tau_1,\tau_2]} \leq \sqrt{2V^\rho_{(\tau_1,\tau_2]}}\left(1+\sqrt{2\Gamma^\rho_{(\tau_1,\tau_2]}}\right) + 5B_{\tau_2}\left(\Gamma^\rho_{(\tau_1,\tau_2]} + \ln 2\right) + B_{\tau_2}. \tag{19}$$

Proof of Lemma 11 By the assumption that $B_{\tau_2-1}/B_{\tau_1} \leq \ln \sum_{t=1}^{\tau_2-1} b_t/B_t$ and Lemma 1, the potential function Φ_{τ_2} can be upper-bounded by

$$\Phi_{\tau_2} \leq \ln \frac{B_{\tau_2-1}}{B_{\tau_1}} \leq \ln \sum_{t=1}^{\tau_2-1} \frac{b_t}{B_t}.$$

Using this, together with Lemma 2 and (4), we get (19). ■

Assume without loss of generality that $b_1 \neq 0$. Then the regret of SQUINT+L in round $t = 1$ is bounded by $B_1 \leq B_T$, and SQUINT+C is started for the first time in round $t = 2$ with input $B = B_1$.

Now suppose first that the restart condition in line 2 of Algorithm 1 is never triggered, which means that $B_t/B_1 \leq \sum_{s=1}^t b_s/B_s$ for all rounds $t = 2, \ldots, T$. Then for any comparator distribution $\rho \in \triangle_K$, the result follows from Lemma 2 and the facts that $V^\rho_{(1:T]} \leq V^\rho_T$ and $\Gamma^\rho_{(1:T]} \leq \Gamma^\rho_T$.

Alternatively, suppose there is at least one restart. Then let $1 \leq \tau_1 < \tau_2 < T$ be such that $(\tau_1, \tau_2]$ and $(\tau_2, T]$ are the two intervals over which the last two runs of SQUINT+C occurred. We invoke Lemma 2 separately for both these intervals and use Lemma 11 to bound

$$\begin{aligned}
R^\rho_{(\tau_1,T]} &\leq \sqrt{2V^\rho_{(\tau_1,\tau_2]}}\left(1+\sqrt{2\Gamma^\rho_{(\tau_1,\tau_2]}}\right) + 5B_{\tau_2}\left(\Gamma^\rho_{(\tau_1,\tau_2]} + \ln 2\right) + B_{\tau_2} \\
&\quad + \sqrt{2V^\rho_{(\tau_2,T]}}\left(1+\sqrt{2\Gamma^\rho_{(\tau_2,T]}}\right) + 5B_T\left(\Gamma^\rho_{(\tau_2,T]} + \ln 2\right) + B_T, \\
&\leq 2\sqrt{V^\rho_{(\tau_1,T]}}\left(1+\sqrt{2\Gamma^\rho_{(\tau_1,T]}}\right) + 10B_T\left(\Gamma^\rho_{(\tau_1,T]} + \ln 2\right) + 2B_T, \qquad (20) \\
&\leq 2\sqrt{V^\rho_T}\left(1+\sqrt{2\Gamma^\rho_T}\right) + 10B_T\left(\Gamma^\rho_T + \ln 2\right) + 2B_T. \qquad (21)
\end{aligned}$$

where in (20) we used the fact that $\sqrt{x}+\sqrt{y} \le \sqrt{2x+2y}$. If there is exactly one restart, then (21) implies the desired result. If there are multiple restarts, then the proof is completed by bounding the contribution to the regret of all rounds $2, \dots, \tau_1$ by

$$R^{\boldsymbol{u}}_{(1,\tau_1]} \le \sum_{t=2}^{\tau_1} b_t \le B_{\tau_1} \sum_{t=1}^{\tau_1} \frac{b_t}{B_t} \le B_{\tau_1} \sum_{t=1}^{\tau_2} \frac{b_t}{B_t} < B_{\tau_2} \le B_T,$$

where the second to last inequality holds because there was a restart at the end of round $t = \tau_2$. Finally, by bounding the instantaneous regret from the first round by B_T, we obtain the desired result. ■

Appendix B. Proofs of Section 3

Proof of Lemma 5 Let $t \ge 1$. To simplify notation, we denote $\bar{r}^\eta_s := \langle \widehat{\boldsymbol{u}}_s - \widehat{\boldsymbol{u}}^\eta_s, \bar{\boldsymbol{g}}_s \rangle$, for $\boldsymbol{u} \in \mathcal{U}$ and $s \in \mathbb{N}$. By appealing to the prod-bound (*i.e.* $e^{x-x^2} - 1 \le x$, for $x \ge -1/2$), we have

$$\begin{aligned}
\Phi_{t+1} &= \pi(\mathcal{G}_{t+1} \setminus \mathcal{A}_{t+1}) + \sum_{\eta \in \mathcal{A}_{t+1}} w^\eta_{t+1} \left(e^{\eta \bar{r}^\eta_{t+1} - \eta (\bar{r}^\eta_{t+1})^2} - 1 \right) + \sum_{\eta \in \mathcal{A}_{t+1}} w^\eta_{t+1}, \\
&\le \pi(\mathcal{G}_{t+1} \setminus \mathcal{A}_{t+1}) + \sum_{\eta \in \mathcal{A}_{t+1}} w^\eta_{t+1} \eta \bar{r}^\eta_{t+1} + \sum_{\eta \in \mathcal{A}_{t+1}} w^\eta_{t+1}.
\end{aligned}$$

Now by (14)

$$\sum_{\eta \in \mathcal{A}_{t+1}} w^\eta_{t+1} \eta \bar{r}^\eta_{t+1} = \sum_{\eta \in \mathcal{A}_{t+1}} \eta w^\eta_{t+1} (\widehat{\boldsymbol{u}}_{t+1} - \widehat{\boldsymbol{u}}^\eta_{t+1})^\intercal \bar{\boldsymbol{g}}_t = 0.$$

Moreover, by definition of $\mathcal{G}_t$ and $\mathcal{A}_t$,

$$\begin{aligned}
&\pi(\mathcal{G}_{t+1} \setminus \mathcal{A}_{t+1}) + \sum_{\eta \in \mathcal{A}_{t+1}} w^\eta_{t+1} = \pi(\{\eta \in \mathcal{G}_{t+1} : s_\eta > t\}) + \sum_{\eta \in \mathcal{G}_{t+1} : s_\eta \le t} w^\eta_{t+1}, \\
&\le \pi(\{\eta \in \mathcal{G}_t : s_\eta > t\}) + \sum_{\eta \in \mathcal{G}_t : s_\eta \le t} w^\eta_{t+1} = \pi(\{\eta \in \mathcal{G}_t : s_\eta \ge t\}) + \sum_{\eta \in \mathcal{G}_t : s_\eta < t} w^\eta_{t+1}, \\
&= \pi(\mathcal{G}_t \setminus \mathcal{A}_t) + \sum_{\eta \in \mathcal{A}_t} w^\eta_{t+1} = \Phi_t.
\end{aligned}$$

Where we used that $w^\eta_{s_\eta+1} = \pi(\eta)$. Finally, as $\mathcal{A}_0 = \emptyset$ and $\mathcal{G}_0 = \mathcal{G}$, we find $\Phi_0 = \pi(\mathcal{G}) = 1$. ■

Proof of Theorem 6 Throughout this proof we will deal with slaves $\eta \in \mathcal{G}_T \setminus \mathcal{A}_T$ that are provisioned but not active yet by time T, and we will interpret their $s_\eta = T$ for uniform treatment, even though technically all we know from (13) is that $s_\eta \ge T$.

First due to Lemma 5, we have $\Phi_T \le 1$, where Φ_T is the potential defined in (15). Taking logarithms and rearranging, we find

$$\forall \eta \in \mathcal{G}_T, \quad -\sum_{t=s_\eta+1}^{T} \bar{f}_t(\widehat{\boldsymbol{u}}^\eta_t, \eta) \le -\ln \pi(\eta). \tag{22}$$

Moreover, every slave $\eta \in \mathcal{G}_T$ guarantees the following regret for the rounds $t = s_\eta + 1, \dots, T$ (see Van Erven and Koolen 2016, Lemma 5):

$$\sum_{t=s_\eta+1}^{T} \left(\bar{f}_t(\widehat{\boldsymbol{u}}_t^\eta, \eta) - \bar{f}_t(\boldsymbol{u}, \eta)\right) \le \ln\det\left(\mathbf{I} + 2\eta^2 D^2(\bar{\mathbf{S}}_T - \bar{\mathbf{S}}_{s_\eta})\right) + \tfrac{\|\boldsymbol{u}\|^2}{2D^2},$$
$$\le d\ln\left(1 + \tfrac{2D^2}{25dB_{T-1}^2}\operatorname{tr}\bar{\mathbf{S}}_T\right) + \tfrac{\|\boldsymbol{u}\|^2}{2D^2}, \tag{23}$$

where in (23) we used concavity of $\ln\det$, $\bar{\mathbf{S}}_{s_\eta} \succeq \mathbf{0}$, and the fact that $\eta \in \mathcal{G}_T \subset [0, 1/(5B_{T-1})]$. We then invert the 'wake up condition' (13) at time $s_\eta - 1$ to infer

$$-\sum_{t=1}^{s_\eta} \bar{f}_t(\boldsymbol{u}, \eta) \le \eta \sum_{t=1}^{s_\eta} \bar{r}_t^{\boldsymbol{u}} \le \frac{\sum_{t=1}^{s_\eta-1} \bar{r}_t^{\boldsymbol{u}} + \bar{r}_{s_\eta}^{\boldsymbol{u}}}{D\sum_{t=1}^{s_\eta-1}\|\bar{\boldsymbol{g}}_t\|_2 + B_{s_\eta-1}} \le 1. \tag{24}$$

Combining the bounds (22), (23), and (24), then dividing through by η, gives:

$$\forall \eta \in \mathcal{G}_T, \quad \bar{R}_T^{\boldsymbol{u}} \le \eta \bar{V}_T^{\boldsymbol{u}} + \tfrac{1}{\eta} C_T(\eta), \tag{25}$$

where $C_T(\eta) \coloneqq d\ln\left(1 + \frac{2D^2}{25dB_{T-1}^2}\operatorname{tr}\bar{\mathbf{S}}_T\right) - \ln\pi(\eta) + 2$.

Let C_T be as in the theorem statement and η_* be the estimator defined by $\eta_* \coloneqq \sqrt{C_T/\bar{V}_T^{\boldsymbol{u}}}$. Suppose that $\eta_* \le 1/(5B_{T-1})$. By construction of the grid $\mathcal{G}_T$, there exists $i \in \mathbb{N}$ such that

$$\hat{\eta} \coloneqq 2^{-i}/(5B_0) \in \mathcal{G}_T \ \text{ and } \ \hat{\eta} \in [\eta_*/2, \eta_*]\,. \tag{26}$$

Since $C_T \ge 1$, the estimator η_* can be lower-bounded by $1/\sqrt{\bar{V}_T^{\boldsymbol{u}}}$, and thus due to (26) we have $2^{-i}/(5B_0) \ge 1/\sqrt{4\bar{V}_T^{\boldsymbol{u}}}$. This implies that the prior weight on $\hat{\eta}$ satisfies

$$\frac{1}{\pi(\hat{\eta})} = (i+1)(i+2) \le \left(\log_2 \tfrac{2\sqrt{\bar{V}_T^{\boldsymbol{u}}}}{5B_0} + 1\right)\left(\log_2 \tfrac{2\sqrt{\bar{V}_T^{\boldsymbol{u}}}}{5B_0} + 2\right) \le \left(\log_2 \tfrac{\sqrt{\bar{V}_T^{\boldsymbol{u}}}}{B_0} + 3\right)^2. \tag{27}$$

Now from the fact that $1/\sqrt{\bar{V}_T^{\boldsymbol{u}}} \le \eta_* \le 1/(5B_{T-1}) \le 1/(5B_0)$, we have $\sqrt{\bar{V}_T^{\boldsymbol{u}}}/B_0 \ge 2$. This, combined with (27), implies that $C_T(\hat{\eta}) \le C_T$, where C_T is as in the theorem statement. Plugging $\eta = \hat{\eta}$ into (25) and using the fact that $\hat{\eta} \in [\eta_*/2, \eta_*]$, gives

$$\bar{R}_T^{\boldsymbol{u}} \le \hat{\eta}\bar{V}_T^{\boldsymbol{u}} + \tfrac{1}{\hat{\eta}} C_T(\hat{\eta}) \le \eta_*\bar{V}_T^{\boldsymbol{u}} + \tfrac{2}{\eta_*} C_T = 3\sqrt{\bar{V}_T^{\boldsymbol{u}} C_T}. \tag{28}$$

Now suppose that $\eta_* > 1/(5B_{T-1})$, and let $\hat{\eta} \coloneqq \max\mathcal{G}_T \ge 1/(10B_{T-1})$, where the last inequality follows by construction of $\mathcal{G}_T$. Note that in this case $\frac{1}{\pi(\hat{\eta})} \le (\log_2 \frac{2B_{T-1}}{B_0} + 1)(\log_2 \frac{2B_{T-1}}{B_0} + 2)$, and the inequality $C_T(\hat{\eta}) \le C_T$ still holds. Plugging $\eta = \hat{\eta}$ into (25) and using the assumption on η_*, *i.e.* $\eta_* > 1/(5B_{T-1})$, we obtain

$$\bar{R}_T^{\boldsymbol{u}} \le \hat{\eta}\bar{V}_T^{\boldsymbol{u}} + \tfrac{1}{\hat{\eta}} C_T(\hat{\eta}) \le \hat{\eta}\bar{V}_T^{\boldsymbol{u}} + \tfrac{1}{\hat{\eta}} C_T \le 15 B_T C_T. \tag{29}$$

By combining (28) and (29), we get the desired result. ■

Proof of Theorem 8 Assume without loss of generality that $b_1 \neq 0$. Then the regret of METAGRAD+L in round one is bounded by $B_1 \leq B_T$, and METAGRAD+C is started for the first time in round $t = 2$ with parameter $B = B_1$.

Let $V^{\boldsymbol{u}}_{(1:T]}$ and $C_{(1:T]}$ represent the quantities denoted by $V^{\boldsymbol{u}}_T$ and C_T in Theorem 6 but measured on rounds $2, \dots, T$. Now suppose first that the restart condition in line 2 of Algorithm 1 is never triggered, which means that

$$\frac{B_t}{B_1} \leq \sum_{s=1}^{t} \frac{b_s}{B_s}, \qquad \text{for all rounds } t = 2, \dots, T. \tag{30}$$

Then the result follows from Theorem 6, $V^{\boldsymbol{u}}_{(1:T]} \leq V^{\boldsymbol{u}}_T$, for all $\boldsymbol{u} \in \mathcal{U}$, and

$$\begin{aligned} C_{(1:T]} &= d \ln\left(1 + \frac{2}{25d} \frac{\sum_{t=1}^{T-1} b_t^2}{B_{T-1}^2}\right) + 2\ln\left(\log_2^+ \frac{\sqrt{\sum_{t=2}^{T} b_t^2}}{B_1} + 3\right) + 2, \\ &\leq d \ln\left(1 + \frac{2}{25d} \frac{\sum_{t=1}^{T-1} b_t^2}{B_{T-1}^2}\right) + 2\ln\left(\log_2^+ \sqrt{\sum_{t=2}^{T}\left(\sum_{s=1}^{t} \frac{b_s}{B_s}\right)^2} + 3\right) + 2, \qquad (31) \\ &\leq \Gamma_T, \end{aligned}$$

where in (31), we used (30). Alternatively, suppose there is at least one restart. Then let $1 \leq \tau_1 < \tau_2 < T$ be such that $(\tau_1, \tau_2]$ and $(\tau_2, T]$ are the two intervals over which the last two runs of METAGRAD+C occurred. We invoke Theorem 6 separately for both these intervals to bound

$$\begin{aligned} R^{\boldsymbol{u}}_{(\tau_1,T]} &\leq 3\sqrt{V^{\boldsymbol{u}}_{(\tau_1,\tau_2]} C_{(\tau_1,\tau_2]}} + 15 B_T C_{(\tau_1,\tau_2]} + B_{\tau_2} \\ &\quad + 3\sqrt{V^{\boldsymbol{u}}_{(\tau_2,T]} C_{(\tau_2,T]}} + 15 B_T C_{(\tau_2,T]} + B_T, \\ &\leq 3\sqrt{V^{\boldsymbol{u}}_{(\tau_1,\tau_2]} \Gamma_T/2} + 3\sqrt{V^{\boldsymbol{u}}_{(\tau_2,T]} \Gamma_T/2} + 15 B_T \Gamma_T + 2B_T, \\ &\leq 3\sqrt{V^{\boldsymbol{u}}_{(\tau_1,T]} \Gamma_T} + 15 B_T \Gamma_T + 2B_T, \qquad (32) \end{aligned}$$

where a subscript $(\tau_1, \tau_2]$ indicates a quantity measured only on rounds $\tau_1 + 1, \dots, \tau_2$ and the last inequality uses $\sqrt{x} + \sqrt{y} \leq \sqrt{2x + 2y}$. If there is exactly one restart, then (32) implies the desired result. If there are multiple restarts, then the proof is completed by bounding the contribution to the regret of all rounds $2, \dots, \tau_1$ by

$$R^{\boldsymbol{u}}_{(1,\tau_1]} \leq \sum_{t=2}^{\tau_1} b_t \leq B_{\tau_1} \sum_{t=1}^{\tau_1} \frac{b_t}{B_t} \leq B_{\tau_1} \sum_{t=1}^{\tau_2} \frac{b_t}{B_t} < B_{\tau_2} \leq B_T,$$

where the second to last inequality holds because there was a restart at $t = \tau_2$. Finally, by bounding the instantaneous regret from the first round by B_T, we obtain the desired result. ■

Appendix C. Proofs of Section 4

Proof of Lemma 9 We use the Lagrangian multiplier to solve (17). For this, let

$$\mathcal{L}(\boldsymbol{u},\mu) := \left(\boldsymbol{u}_{t+1}^{\eta} - \boldsymbol{u}\right)^{\intercal} \left(\boldsymbol{\Sigma}_{t+1}^{\eta}\right)^{-1} \left(\boldsymbol{u}_{t+1}^{\eta} - \boldsymbol{u}\right) + \mu(\boldsymbol{u}^{\intercal}\boldsymbol{u} - D^2).$$

Setting $\frac{\partial \mathcal{L}}{\partial \boldsymbol{u}} = 0$ implies that $2\left(\boldsymbol{\Sigma}_{t+1}^{\eta}\right)^{-1}\left(\boldsymbol{u} - \boldsymbol{u}_{t+1}^{\eta}\right) + 2\mu\boldsymbol{u} = 0$. After rearranging, this becomes

$$\begin{aligned}\boldsymbol{u} &= \left(\left(\mu + \tfrac{1}{D^2}\right)\mathbf{I} + 2\eta^2\mathbf{S}_t\right)^{-1}\left(\boldsymbol{\Sigma}_{t+1}^{\eta}\right)^{-1}\boldsymbol{u}_t^{\eta},\\ &= \mathbf{Q}_t^{\intercal}\left(x\mathbf{I} + 2\eta^2\boldsymbol{\Lambda}_t\right)^{-1}\mathbf{Q}_t\boldsymbol{v}_{t+1}^{\eta},\end{aligned}$$

where we set $x := \mu + 1/D^2$. The result follows after observing that $\boldsymbol{u}^{\intercal}\boldsymbol{u} = D^2/4 \iff \rho_t^{\eta}(x) = D^2/4$. ■

Proof of Theorem 10 Let $\mathring{R}_T^{\boldsymbol{u}} := \sum_{t=1}^{T}\langle \widehat{\boldsymbol{w}}_t - \boldsymbol{u}, \mathring{\boldsymbol{g}}_t\rangle$ and $\mathring{V}_T^{\boldsymbol{u}} := \sum_{t=1}^{T}\langle \widehat{\boldsymbol{w}}_t - \boldsymbol{u}, \mathring{\boldsymbol{g}}_t\rangle^2$ be the pseudo-regret and 'variance' of Algorithm 2. From our assumption on the pseudo-regret $\tilde{R}_T^{\boldsymbol{u}}$ of METAGRAD and the fact that $2\sqrt{x} = \inf_{\eta>0}\{\eta x + 1/\eta\}$, we have

$$\forall \boldsymbol{u} \in \mathcal{U} \subset \mathcal{B}_D, \forall \eta > 0, \quad \eta\tilde{R}_T^{\boldsymbol{u}} - \tfrac{\eta^2}{2}V_T^{\boldsymbol{u}} \le \tfrac{1}{2}\Gamma_T + \eta B\Gamma_T. \tag{33}$$

Now, as in the proof of (Cutkosky and Orabona, 2018, Theorem 3), we have

$$\langle \widehat{\boldsymbol{w}}_t - \boldsymbol{u}, \mathring{\boldsymbol{g}}_t\rangle \le 2\mathring{\ell}_t(\widehat{\boldsymbol{u}}_t) - 2\mathring{\ell}_t(\boldsymbol{u}), \tag{34}$$

where $\widehat{\boldsymbol{w}}_t = \Pi_{\mathcal{U}}(\widehat{\boldsymbol{u}}_t)$ is the prediction of Algorithm 2 at round t and $\mathring{\ell}_t$ is the function defined by $\mathring{\ell}_t(\boldsymbol{u}) := \frac{1}{2}\left(\langle \mathring{\boldsymbol{g}}_t, \boldsymbol{u}\rangle + \|\mathring{\boldsymbol{g}}_t\|\,\mathrm{d}_{\mathcal{U}}(\boldsymbol{u})\right)$. By convexity of $\mathring{\ell}_t$ and the fact that $\boldsymbol{g}_t \in \partial\mathring{\ell}_t(\widehat{\boldsymbol{u}}_t)$, we have

$$\langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t\rangle \ge \mathring{\ell}_t(\widehat{\boldsymbol{u}}_t) - \mathring{\ell}_t(\boldsymbol{u}) \ge \tfrac{1}{2}\langle \widehat{\boldsymbol{w}}_t - \boldsymbol{u}, \mathring{\boldsymbol{g}}_t\rangle, \quad \text{for } \boldsymbol{u} \in \mathcal{U}, \tag{35}$$

where the right-most inequality follows from (34). Since the function $x \mapsto x - x^2/2$ is strictly increasing on the interval $]-\infty, 1]$, (35) implies that for all $\eta \in]0, 1/B] =]0, 1/(DG)]$,

$$\tfrac{\eta}{2}\langle \widehat{\boldsymbol{w}}_t - \boldsymbol{u}, \mathring{\boldsymbol{g}}_t\rangle - \tfrac{\eta^2}{8}\langle \widehat{\boldsymbol{w}}_t - \boldsymbol{u}, \mathring{\boldsymbol{g}}_t\rangle^2 \le \eta\langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t\rangle - \tfrac{\eta^2}{2}\langle \widehat{\boldsymbol{u}}_t - \boldsymbol{u}, \boldsymbol{g}_t\rangle^2, \quad \text{for } \boldsymbol{u} \in \mathcal{U}.$$

Summing this over $t = 1, \ldots, T$ and using (33), we get for all $\eta \in]0, 1/B]$ and $\boldsymbol{u} \in \mathcal{U}$,

$$\begin{aligned}&\tfrac{1}{2}\mathring{R}_T^{\boldsymbol{u}} - \tfrac{\eta}{8}\mathring{V}_T^{\boldsymbol{u}} \le \tilde{R}_T^{\boldsymbol{u}} - \tfrac{\eta}{2}V_T^{\boldsymbol{u}} \le \tfrac{1}{2\eta}\Gamma_T + B\Gamma_T, \quad \text{and so}\\ &\mathring{R}_T^{\boldsymbol{u}} \le \tfrac{\eta}{4}\mathring{V}_T^{\boldsymbol{u}} + \tfrac{1}{\eta}\Gamma_T + 2B\Gamma_T.\end{aligned} \tag{36}$$

The 'unconstrained' $\eta \in [0, +\infty]$ which minimizes the RHS of (36) is given by $\eta_* := 2\sqrt{\Gamma_T/\mathring{V}_T^{\boldsymbol{u}}}$. We consider two cases: suppose first that $\eta_* \le 1/B$. For $\eta = \eta_*$, we have

$$\tfrac{\eta}{4}\mathring{V}_T^{\boldsymbol{u}} + \tfrac{1}{\eta}\Gamma_T = \sqrt{\mathring{V}_T^{\boldsymbol{u}}\Gamma_T}. \tag{37}$$

Now suppose that $\eta_* > 1/B$. For $\eta = 1/B$, we have

$$\tfrac{\eta}{4}\mathring{V}_T^{\boldsymbol{u}} + \tfrac{1}{\eta}\Gamma_T \le 2B\Gamma_T. \tag{38}$$

Combining (36)–(38) yields the desired bound. ■

Chapter 4

Lipschitz and Comparator-Norm Adaptivity in Online Learning

In this chapter, we shift our attention to *unbounded* OCO. Despite recent efforts in developing adaptive methods for the unbounded OCO setting, the resulting algorithms typically require additional information about the sequence of losses compared with (adaptive) algorithms for the bounded case. Namely, a bound on the norm of the gradients needs to be known in advance, which is not guaranteed to be available in practice. This motivated the COLT 2016 open problem by Orabona and Pál [2016b] who asked whether it is possible to compete against unbounded comparators in online linear optimization while being scale-free (and thus, not requiring a bound on the norm of the gradients). In this chapter, we derive the first algorithm—FreeGrad—that achieves precisely that at the cost of an additive penalty in the regret bound that depends on the cubed norm of the comparator. We complement this result with a new matching lower bound, thus fully characterizing the limits of adaptivity in this setting.

In addition to being scale-free and suited for the unbounded setting, FreeGrad also enjoys an *adaptive* regret bound where the main regret term depends on the sum of the squared norms of the observed gradients (similar to AdaGrad). This type of data-dependent bound has been useful in various applications; many existing reductions accept algorithms with such a regret bound to achieve different types of adaptivity [Cutkosky, 2019a]. The techniques introduced in this chapter also resolve some open questions relating to Lipschitz adaptivity in online learning. In particular, many algorithms in the literature that attempt to adapt to the scale of the losses have a term in their regret bounds which can, in principle, be arbitrarily large depending on the sequence of observed losses (leading to vacuous bounds) Ross et al. [2013b]; Wintenberger [2017]; Kotłowski [2017]; Mhammedi et al. [2019b]; Kempka et al. [2019b]. The restart scheme introduced in the previous chapter solves this issue for the bounded setting. In this chapter, we extend the restart trick and its analysis to the unbounded setting.

Proceedings of Machine Learning Research vol 125:1–30, 2020 33rd Annual Conference on Learning Theory

Lipschitz and Comparator-Norm Adaptivity in Online Learning

Zakaria Mhammedi ZAK.MHAMMEDI@ANU.EDU.AU
The Australian National University and Data61
Wouter M. Koolen WMKOOLEN@CWI.NL
Centrum Wiskunde & Informatica

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study Online Convex Optimization in the unbounded setting where neither predictions nor gradient are constrained. The goal is to simultaneously adapt to both the sequence of gradients and the comparator. We first develop parameter-free and scale-free algorithms for a simplified setting with hints. We present two versions: the first adapts to the squared norms of both comparator and gradients separately using $O(d)$ time per round, the second adapts to their squared inner products (which measure variance only in the comparator direction) in time $O(d^3)$ per round. We then generalize two prior reductions to the unbounded setting; one to not need hints, and a second to deal with the range ratio problem (which already arises in prior work). We discuss their optimality in light of prior and new lower bounds. We apply our methods to obtain sharper regret bounds for scale-invariant online prediction with linear models.

Keywords: Online Convex Optimization, Parameter-Free Online Learning, Scale-Invariant Online Algorithms

1. Introduction

We consider the setting of online convex optimization where the goal is to make sequential predictions to minimize a certain notion of *regret*. Specifically, at the beginning of each round $t \geqslant 1$, a *learner* predicts $\widehat{\boldsymbol{w}}_t$ in some convex set $\mathcal{W} \subseteq \mathbb{R}^d$ in dimension $d \in \mathbb{N}$. The *environment* then reveals a convex loss function $f_t \colon \mathcal{W} \to \mathbb{R}$, and the learner suffers loss $f_t(\widehat{\boldsymbol{w}}_t)$. The goal of the learner is to minimize the regret $\sum_{t=1}^T f_t(\widehat{\boldsymbol{w}}_t) - \sum_{t=1}^T f_t(\boldsymbol{w})$ after $T \geqslant 1$ rounds against any "comparator" prediction $\boldsymbol{w} \in \mathcal{W}$. Typically, an online learning algorithm outputs a vector $\widehat{\boldsymbol{w}}_t$, $t \geqslant 1$, based only on a sequence of observed sub-gradients $(\boldsymbol{g}_s)_{s<t}$, where $\boldsymbol{g}_s \in \partial f_s(\widehat{\boldsymbol{w}}_s)$, $s < t$. In this paper, we are interested in online algorithms which can guarantee a good regret bound (by a measure which we will make precise below) against any comparator vector $\boldsymbol{w} \in \mathcal{W}$, even when $\mathcal{W}$ is unbounded, and without prior knowledge of the maximum norm $L := \max_{t \leqslant T} \|\boldsymbol{g}_t\|$ of the observed sub-gradients. In what follows, we refer to L as the *Lipschitz constant*.

By assuming an upper-bound $D > 0$ on the norm of the desired comparator vector $\boldsymbol{w}$ in hindsight, there exist *Lipschitz-adaptive* algorithms that can achieve a sub-linear regret of order $LD\sqrt{T}$, without knowing L in advance. A Lipschitz-adaptive algorithm is also called *scale-free* (or scale-invariant) if its predictions do not change when the loss functions (f_t) are multiplied by a factor $c > 0$; in this case, its regret bound is expected to scale by the same factor c. When L is known in advance and $\mathcal{W} = \mathbb{R}^d$, there exists another type of algorithms, so-called *parameter-free*, which can achieve an $\widetilde{O}(\|\boldsymbol{w}\| L\sqrt{T})$ regret bound, where $\boldsymbol{w}$ is the desired comparator vector in hindsight (the notation $\widetilde{O}$

hides log-factors). Up to an additive lower-order term, this type of regret bound is also achievable for bounded $\mathcal{W}$ via the unconstrained-to-constrained reduction (Cutkosky, 2019).

The question of whether an algorithm can simultaneously be *scale-free* and *parameter-free* was posed as an open problem by Orabona and Pál (2016b). It was latter answered in the negative by Cutkosky and Boahen (2017). Nevertheless, Cutkosky (2019) recently presented algorithms which achieve an $\widetilde{O}(\|\boldsymbol{w}\|L\sqrt{T} + L\|\boldsymbol{w}\|^3)$ regret bound, without knowing either L or $\|\boldsymbol{w}\|$. This does not violate the earlier lower bound of Cutkosky and Boahen (2017), which insists on norm dependence $\widetilde{O}(\|\boldsymbol{w}\|)$.

Though Cutkosky (2019) designs algorithms that can to some extent adapt to both L and $\|\boldsymbol{w}\|$, their algorithms are still *not* scale-free. Multiplying (f_t), and as a result $(\boldsymbol{g}_t)$, by a positive factor $c > 0$ changes the outputs $(\widehat{\boldsymbol{w}}_t)$ of their algorithms, and their regret bounds scale by a factor c', not necessarily equal to c. Their algorithms depend on a parameter $\epsilon > 0$ which has to be specified in advance. This parameter appears in their regret bounds as an additive term and also in a logarithmic term of the form $\log(L^\alpha/\epsilon)$, for some $\alpha > 1$. As a result of this type of dependence on ϵ and the fact that $\alpha > 1$, there is no prior choice of ϵ which can make their regret bounds scale-invariant. What is more, without knowing L, there is also no "safe" choice of ϵ which can prevent the $\log(L^\alpha/\epsilon)$ term from becoming arbitrarily large relative to L (it suffices for ϵ to be small enough relative to the "unknown" L).

Contributions. Our main contribution is a new scale-free, parameter-free learning algorithm for OCO with regret at most $O(\|\boldsymbol{w}\|\sqrt{V_T \log(\|\boldsymbol{w}\|T)})$, for any comparator $\boldsymbol{w} \in \mathcal{W}$ in a bounded set $\mathcal{W}$, where $V_T := \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$. When the set $\mathcal{W}$ is unbounded, the algorithm achieves the same guarantee up to an additive $O(L\sqrt{\max_{t\leqslant T} B_t} + L\|\boldsymbol{w}\|^3)$, where $B_t := \sum_{s=1}^t \|\boldsymbol{g}_s\|/L_t$ and $L_t := \max_{s\leqslant t}\|\boldsymbol{g}_s\|$, for all $t \in [T]$. In the latter case, we also show a matching lower bound; when $\mathcal{W}$ is unbounded and without knowing L, any online learning algorithm which insists on an $\widetilde{O}(\sqrt{T})$ bound, has regret at least $\Omega(L\sqrt{B_T} + L\|\boldsymbol{w}\|^3)$. We also provide a second scale-invariant algorithm which replaces the leading $\|\boldsymbol{w}\|\sqrt{V_T}$ term in the regret bound of our first algorithm by $\sqrt{\boldsymbol{w}^\intercal \boldsymbol{V}_T \boldsymbol{w} \ln\det \boldsymbol{V}_T}$, where $\boldsymbol{V}_T := \sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\intercal$. Our starting point for designing our algorithms is a known potential function which we show to be controlled for a unique choice of output sequence $(\widehat{\boldsymbol{w}}_t)$.

As our main application, we show how our algorithms can be applied to learn linear models. The result is an online algorithm for learning linear models whose label predictions are invariant to coordinate-wise scaling of the input feature vectors. The regret bound of the algorithm is naturally also scale-invariant and improves on the bounds of existing state-of-the-art algorithms in this setting (Kotłowski, 2017; Kempka et al., 2019).

Related Work For an overview of Online Convex Optimization in the bounded setting, we refer to the textbook (Hazan, 2016). The unconstrained case was first studied by McMahan and Streeter (2010). A powerful methodology for the unbounded case is Coin Betting by Orabona and Pál (2016a). Even though not always visible, our potential functions are inspired by this style of thinking. We build our unbounded OCO learner by targeting a specific other constrained problem. We further employ several general reductions from the literature, including gradient clipping Cutkosky (2019), the constrained-to-unconstrained reduction Cutkosky and Orabona (2018), and the restart wrapper to pacify the final-vs-initial scale ratio appearing inside logarithms by Mhammedi et al. (2019). Our analysis is, at its core, proving a certain minimax result about sufficient-statistic-based potentials reminiscent of the Burkholder approach pioneered by Foster et al. (2017, 2018). Applications for scale-invariant learning in linear models were studied by Kempka et al. (2019).

For our multidimensional learner we took inspiration from the Gaussian Exp-concavity step in the analysis of the MetaGrad algorithm by Van Erven and Koolen (2016).

Outline In Section 2, we present the setting and notation, and formulate our goal. In Section 3, we present our main algorithms. In Section 4, we present new lower-bounds for algorithms which adapt to both the Lipschitz constant and the norm of the comparator. In Section 5, we apply our algorithms to online prediction with linear models.

2. Preliminaries

Our goal is to design scale-free algorithms that adapt to the Lipschitz constant L and comparator norm $\|\boldsymbol{w}\|$. We will first introduce the setting, then discuss existing reductions, and finally state what needs to be done to achieve our goal.

2.1. Setting and Notation

Let $\mathcal{W} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, be a convex set, and assume without loss of generality that $\mathbf{0} \in \mathcal{W}$. We allow the set $\mathcal{W}$ to be unbounded, and we define its (possibly infinite) diameter $D := \sup_{\boldsymbol{w},\boldsymbol{w}' \in \mathcal{W}} \|\boldsymbol{w} - \boldsymbol{w}'\| \in [0, +\infty]$. We consider the setting of Online Convex Optimization (OCO) where at the beginning of each round $t \geqslant 1$, the learner outputs a prediction $\widehat{\boldsymbol{w}}_t \in \mathcal{W}$, before observing a convex loss function $f_t : \mathcal{W} \to \mathbb{R}$, or an element of its sub-gradient $\boldsymbol{g}_t \in \partial f_t(\widehat{\boldsymbol{w}}_t)$ at $\widehat{\boldsymbol{w}}_t$. The goal of the learner is to minimize the regret after $T \geqslant 0$ rounds, which is given by

$$\sum_{t=1}^{T} f_t(\widehat{\boldsymbol{w}}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{w})$$

for any comparator vector $\boldsymbol{w} \in \mathcal{W}$. In this paper, we do not assume that T is known to the learner, and so we are after algorithms with so called *any-time* guarantees. By convexity, we have

$$\sum_{t=1}^{T} f_t(\widehat{\boldsymbol{w}}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{w}) \leqslant \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \widehat{\boldsymbol{w}}_t - \boldsymbol{w} \rangle, \quad \text{for all } \boldsymbol{w} \in \mathcal{W}, \tag{1}$$

and thus for the purpose of minimizing the regret, typical OCO algorithms minimize the RHS of (1), which is known as the *linearized regret*, by generating outputs $(\widehat{\boldsymbol{w}}_t)$ based on the sequence of observed sub-gradients $(\boldsymbol{g}_t)$. Likewise, we focus our attention exclusively on linear optimization.

Given a sequence of sub-gradients $(\boldsymbol{g}_t)$, it will be useful to define the running maximum gradient norm and the clipped sub-gradients

$$L_t := \max_{s \in [t]} \|\boldsymbol{g}_s\| \quad \text{and} \quad \bar{\boldsymbol{g}}_t := \boldsymbol{g}_t \cdot L_{t-1}/L_t,$$

for $t \geqslant 1$, with the convention that $L_0 = 0$. We also drop the subscript t from L_t when $t = T$, *i.e.* we write L for L_T.

We denote by $\mathrm{A}(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{t-1}; h_t)$ the output in round $t \geqslant 1$ of an algorithm A, which uses the observed sub-gradients so far and a *hint* $h_t \geqslant L_t$ on the upcoming sub-gradient $\boldsymbol{g}_t$. As per Section 1, we say that an algorithm is *scale-free* (or scale-invariant) if its predictions are invariant to any common positive scaling of the loss functions (f_t) and, if applicable, the hints.

Additional Notation. Given a closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$, we denote by $\Pi_{\mathcal{X}}(\boldsymbol{x})$ the Euclidean projection of a point $\boldsymbol{x} \in \mathbb{R}^d$ on the set $\mathcal{X}$; that is, $\Pi_{\mathcal{X}}(\boldsymbol{x}) \in \operatorname{argmin}_{\tilde{\boldsymbol{x}} \in \mathcal{X}} \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|$.

2.2. Helpful Reductions

The difficulty behind designing scale-free algorithms lies partially in the fact that L_t is unknown at the start of round t; before outputting $\widehat{\boldsymbol{w}}_t$. The following result due to Cutkosky (2019) quantifies the additional cost of proceeding with the plug-in estimate L_{t-1} for L_t:

Lemma 1 *Let* A *be an online algorithm which at the start of each round $t \geqslant 1$, has access to a hint $h_t \geqslant L_t$, and outputs $\mathrm{A}(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{t-1}; h_t) \in \mathcal{W}$, before observing $\boldsymbol{g}_t$. Suppose that* A *guarantees an upper-bound $R_T^{\mathrm{A}}(\boldsymbol{w})$ on its linearized regret for the sequence $(\boldsymbol{g}_t)$ and for all $\boldsymbol{w} \in \mathcal{W}, T \geqslant 1$. Then, algorithm* B *which at the start of each round $t \geqslant 1$ outputs $\widehat{\boldsymbol{w}}_t = \mathrm{A}(\bar{\boldsymbol{g}}_1, \ldots, \bar{\boldsymbol{g}}_{t-1}; L_{t-1})$, guarantees*

$$\sum_{t=1}^{T} \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{w}, \boldsymbol{g}_t \rangle \leqslant R_T^{\mathrm{A}}(\boldsymbol{w}) + \max_{t \in [T]} \|\widehat{\boldsymbol{w}}_t\| L_t + \|\boldsymbol{w}\| L, \quad \forall \boldsymbol{w} \in \mathcal{W}, T \geqslant 1. \tag{2}$$

First, we note that Lemma 1 is only really useful when $\mathcal{W}$ is bounded; otherwise, depending on algorithm A, the term $\max_{t \in [T]} L_t \|\widehat{\boldsymbol{w}}_t\|$ on the RHS of (2) could in principle be arbitrarily large even for fixed $\boldsymbol{w}$, L, and T. The moral of Lemma 1 is that as long as the set $\mathcal{W}$ is bounded, one does not really need to know L_t before outputting $\widehat{\boldsymbol{w}}_t$ to guarantee a "good" regret bound against any $\boldsymbol{w} \in \mathcal{W}$. For example, suppose that $\mathcal{W}$ has a bounded diameter D and algorithm A in Lemma 1 is such that $R_T^{\mathrm{A}}(\boldsymbol{w}) = \widetilde{O}(\|\boldsymbol{w}\| L \sqrt{T} + DL)$, for all $\boldsymbol{w} \in \mathcal{W}$. Then, from (2) and the fact that $\|\widehat{\boldsymbol{w}}_t\| \leqslant D$ (since $\widehat{\boldsymbol{w}}_t \in \mathcal{W}$), it is clear that algorithm B in Lemma 1 also guarantees the same regret bound $R_T^{\mathrm{A}}(\boldsymbol{w})$ up to an additive $2DL$, despite not having had the hints (h_t).

It is possible to extend the result of Lemma 1 so that the regret bound of algorithm B remains useful even in the case where $\mathcal{W}$ is unbounded. An approach suggested by Cutkosky (2019) is to restrict the outputs $(\widehat{\boldsymbol{w}}_t)$ of algorithm B to be in a non-decreasing sequence $(\mathcal{W}_t)$ of *bounded* convex subsets of $\mathcal{W}$. In this case, the diameters $(D_t) \subset \mathbb{R}$ of $(\mathcal{W}_t)$ need to be carefully chosen to achieve a desired regret bound. This approach, which essentially combines the idea of Lemma 1 and the unconstrained-to-constrained reduction due to Cutkosky and Orabona (2018), is formalized in the next lemma (essentially due to Cutkosky (2019)):

Lemma 2 *Let algorithm* A *be as in Lemma 1, and let $(\mathcal{W}_t)$ be a sequence of non-decreasing closed convex subsets of $\mathcal{W}$ with diameters $(D_t) \subset \mathbb{R}_{>0}$. Then, algorithm* B *which at the start of round $t \geqslant 1$ outputs $\widehat{\boldsymbol{w}}_t = \Pi_{\mathcal{W}_t}(\widetilde{\boldsymbol{w}}_t)$, where*

$$\widetilde{\boldsymbol{w}}_t := \mathrm{A}(\widetilde{\boldsymbol{g}}_1, \ldots, \widetilde{\boldsymbol{g}}_{t-1}; L_{t-1}) \quad \textit{and} \quad \widetilde{\boldsymbol{g}}_s := (\bar{\boldsymbol{g}}_s + \|\bar{\boldsymbol{g}}_s\| \cdot (\widetilde{\boldsymbol{w}}_s - \widehat{\boldsymbol{w}}_s)/\|\widetilde{\boldsymbol{w}}_s - \widehat{\boldsymbol{w}}_s\|)/2, \quad s < t,$$

guarantees, for all $\boldsymbol{w} \in \mathcal{W}$ and $T \geqslant 1$,

$$\sum_{t=1}^{T} \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{w}, \boldsymbol{g}_t \rangle \leqslant R_T^{\mathrm{A}}(\boldsymbol{w}) + \sum_{t=1}^{T} \|\boldsymbol{g}_t\| \cdot \|\boldsymbol{w} - \Pi_{\mathcal{W}_t}(\boldsymbol{w})\| + LD_T + L\|\boldsymbol{w}\|. \tag{3}$$

We see that compared to Lemma 1, the additional penalty that algorithm B incurs for restricting its predictions to the sets $\mathcal{W}_1, \ldots, \mathcal{W}_T \subseteq \mathcal{W}$ is the sum $\sum_{t=1}^{T} \|\boldsymbol{g}_t\| \cdot \|\boldsymbol{w} - \Pi_{\mathcal{W}_t}(\boldsymbol{w})\|$. The challenge is now in choosing the diameters (D_t) to control the trade-off between this sum and the term LD_T on

the RHS of (3). If T is known in advance, one could set $D_1 = \cdots = D_T = \sqrt{T}$, in which case the RHS of (3) is at most

$$R_T^{\mathsf{A}}(\boldsymbol{w}) + L(\|\boldsymbol{w}\|^3 + \|\boldsymbol{w}\|) + L\sqrt{T}. \tag{4}$$

We now instantiate the bound of Lemma 2 for another choice of (D_t) when T is unknown:

Corollary 3 *In the setting of Lemma 2, let $\mathcal{W}_t$ be the ball of diameter $D_t := \sqrt{\max_{s\leqslant t} B_s}$, $t \geqslant 1$, where $B_t := \sum_{s=1}^t \|\boldsymbol{g}_s\|/L_t$, and let $\mathcal{W} = \mathbb{R}^d$. Then the RHS of (3) is bounded from above by*

$$R_T^{\mathsf{A}}(\boldsymbol{w}) + L\|\boldsymbol{w}\|^3 + L\sqrt{\max_{t\in[T]} B_t} + L\|\boldsymbol{w}\|, \quad \forall \boldsymbol{w} \in \mathcal{W} = \mathbb{R}^d, T \geqslant 1. \tag{5}$$

We see that by the more careful choice of (D_t) in Corollary 3, one can replace the $L\sqrt{T}$ term in (4) by the smaller quantity $L\sqrt{\max_{t\in[T]} B_t}$; whether this can be improved further to $\sqrt{V_T}$, where $V_T = \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$, was raised as an open question by Cutkosky (2019). We will answer this in the negative in Theorem 14. We will also show in Theorem 15 below that, if one insists on a regret of order $\widetilde{O}(\sqrt{T})$, it is essentially not possible to improve on the penalty $L\|\boldsymbol{w}\|^3$ in (5).

2.3. Outlook

The conclusion that should be drawn from Lemmas 1 and 2 is the following; if one seeks an algorithm B with a regret bound of the form $\widetilde{O}(\|\boldsymbol{w}\|L\sqrt{T})$ up to some lower-order terms in T, without knowledge of L and regardless of whether $\mathcal{W}$ is bounded or not, it suffices to find an algorithm A which guarantees the sought type of regret whenever it has access to a sequence of hints (h_t) satisfying (as in Lemmas 1 and 2), $h_t \geqslant L_t$, for all $t \geqslant 1$. Thus, our first goal in the next section is to design a *scale-free* algorithm A which accesses such a sequence of hints and ensures that its linearized regret is bounded from above by:

$$O\left(\|\boldsymbol{w}\|\sqrt{V_T \ln(\|\boldsymbol{w}\|V_T)}\right), \quad \text{where } V_T := h_1^2 + \sum_{t=1}^T \|\boldsymbol{g}_t\|^2, \tag{6}$$

for all $\boldsymbol{w} \in \mathbb{R}^d, T \geqslant 0$, and $(\boldsymbol{g}_t) \subset \mathbb{R}^d$. We show an analogous "full-matrix" upgrade of order $\sqrt{\boldsymbol{w}^\intercal \boldsymbol{V}\boldsymbol{w} \ln(\boldsymbol{w}^\intercal \boldsymbol{V}\boldsymbol{w} \det \boldsymbol{V})}$, with $\boldsymbol{V} = \sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\intercal$. We note that if Algorithm A in Lemmas 1 and 2 is scale-free, then so is the corresponding Algorithm B.

If the desired set $\mathcal{W}$ has bounded diameter $D > 0$, then using the unconstrained-to-constrained reduction due to Cutkosky and Orabona (2018), it is straightforward to design a new algorithm based on A with regret also bounded by (6) up to an additive LD, for $\boldsymbol{w} \in \mathcal{W}$ (this is useful for Lemma 1).

Finally, we also note that algorithms which can access hints (h_t) such that $h_t \geqslant L_t$, for all $t \geqslant 1$, are of independent interest; in fact, it is the same algorithm A that we will use in Section 5 as a scale-invariant algorithm for learning linear models.

3. Scale-Free, Parameter-Free Algorithms for OCO

In light of the conclusions of Section 2, we will design new unconstrained scale-free algorithms which can access a sequence of hints (h_t) (as in Lemma 1) and guarantee a regret bound of the form given in (6). In this section, we will make the following assumption on the hints (h_t):

Assumption 1 *We assume that (i) (h_t) is a non-decreasing sequence; (ii) $h_t \geqslant L_t$, for all $t \geqslant 1$; and (iii) if the sub-gradients $(\boldsymbol{g}_s)$ are multiplied by a factor $c > 0$, then the hints (h_t) are multiplied by the same factor c.*

The third item of the assumption ensures that our algorithms are scale-free. We note that Assumption 1 is satisfied by the sequence of hints that Algorithm B constructs when invoking Algorithm A in Lemmas 1 and 2. To avoid an uninteresting case distinction, we will also make the following assumption, which is without loss of generality, since the regret is zero while $\boldsymbol{g}_t = \mathbf{0}$.

Assumption 2 *We assume that $L_1 = \|\boldsymbol{g}_1\| > 0$.*

3.1. FREEGRAD: An Adaptive Scale-Free Algorithm

In this subsection, we design a new algorithm based on a time-varying potential function, where the outputs of the algorithm are uniquely determined by the gradients of the potential function at its iterates—an approach used in the design of many existing algorithms (Cesa-Bianchi et al., 1997).

Let $t \geqslant 1$, let $(\boldsymbol{g}_s)_{s \leqslant t} \subset \mathbb{R}^d$ be a sequence of sub-gradients satisfying Assumption 2, and let (h_t) be a sequence of hints satisfying Assumption 1. Consider the following potential function:

$$\Phi_t := S_t + \frac{h_1^2}{\sqrt{V_t}} \cdot \exp\left(\frac{\|\mathbf{G}_t\|^2}{2V_t + 2h_t\|\mathbf{G}_t\|}\right), \quad t \geqslant 0, \tag{7}$$

$$\text{where} \quad S_t := \sum_{s=1}^{t} \langle \boldsymbol{g}_s, \widehat{\boldsymbol{w}}_s \rangle, \quad \mathbf{G}_t := \sum_{s=1}^{t} \boldsymbol{g}_s, \quad V_t := h_1^2 + \sum_{s=1}^{t} \|\boldsymbol{g}_s\|^2. \tag{8}$$

This potential function has appeared as a by-product in the analyses of previous algorithms such as the ones in (Cutkosky and Orabona, 2018; Cutkosky, 2019). The expression of Φ_t in (7) is interesting to us since it can be shown via the *regret-reward duality* (McMahan and Orabona, 2014) (as we do in the proof of Theorem 6 below) that any algorithm which outputs vectors $(\widehat{\boldsymbol{w}}_t)$ such that (Φ_t) is non-increasing for any sequence of sub-gradients $(\boldsymbol{g}_t)$, also guarantees a regret bound of the form (6). We will now design such an algorithm.

Definition 4 (FREEGRAD) *In round t, based on the sequence of past sub-gradients $(\boldsymbol{g}_s)_{s<t}$ and the available hint $h_t \geqslant L_t$, the* FREEGRAD *algorithm outputs the unconstrained iterate*

$$\widehat{\boldsymbol{w}}_t := -\mathbf{G}_{t-1} \cdot \frac{(2V_{t-1} + h_t\|\mathbf{G}_{t-1}\|) \cdot h_1^2}{2(V_{t-1} + h_t\|\mathbf{G}_{t-1}\|)^2 \sqrt{V_{t-1}}} \cdot \exp\left(\frac{\|\mathbf{G}_{t-1}\|^2}{2V_{t-1} + 2h_t\|\mathbf{G}_{t-1}\|}\right), \tag{9}$$

where $(\mathbf{G}_t)$ and (V_t) are as in (8).

The prediction (9) is forced by our design goal of decreasing potential $\Phi_t \leqslant \Phi_{t-1}$. To see why, observe that at zero $\boldsymbol{g}_t = \mathbf{0}$ we have $\Phi_t = \Phi_{t-1}$. The weights $\widehat{\boldsymbol{w}}_t$ cancel the derivative $\nabla_{\boldsymbol{g}_t}\Phi_t$ at $\boldsymbol{g}_t = \mathbf{0}$, ensuring there is no direction of linear increase (which would present a violation for tiny $\boldsymbol{g}_t$). Our main technical contribution in this subsection is to show that, in fact, with the choice of $(\widehat{\boldsymbol{w}}_t)_{t \geqslant 1}$ as in (9), the potential functions (Φ_t) are non-increasing for any sequence of sub-gradients $(\boldsymbol{g}_t)$:

Theorem 5 *For $(\widehat{\boldsymbol{w}}_t)$, and (Φ_t) as in (9), and (7), under Assumptions 1 and 2, we have:*

$$\Phi_T \leqslant \cdots \leqslant \Phi_0 = h_1, \quad \textit{for all } T \geqslant 1.$$

The proof of the theorem is postponed to Appendix A. Theorem 5 and the regret-reward duality (McMahan and Orabona, 2014) yield a regret bound for FREEGRAD. In fact, if $\Phi_T \leqslant \Phi_0$, then by the definition of Φ_T in (7), we have

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \widehat{\boldsymbol{w}}_t\rangle \leqslant \Phi_0 - \Psi_T(\mathbf{G}_T), \quad \text{where} \quad \Psi_T(\mathbf{G}) := \frac{h_1^2}{\sqrt{V_T}} \exp\left(\frac{\|\mathbf{G}\|^2}{2V_T + 2h_T\|\mathbf{G}\|}\right), \quad \mathbf{G} \in \mathbb{R}^d. \tag{10}$$

Now by Fenchel's inequality, we have $-\Psi_T(\mathbf{G}_T) \leqslant \langle \boldsymbol{w}, \mathbf{G}_T\rangle + \Psi_T^\star(-\boldsymbol{w})$, for all $\boldsymbol{w} \in \mathbb{R}^d$, where $\Psi_T^\star(\boldsymbol{w}) := \sup_{\boldsymbol{z}\in\mathbb{R}^d}\{\langle \boldsymbol{w}, \boldsymbol{z}\rangle - \Psi_T(\boldsymbol{z})\}$, $\boldsymbol{w} \in \mathbb{R}^d$, is the Fenchel dual of Ψ_T (Hiriart-Urruty and Lemaréchal, 2004). Combining this with (10), we obtain:

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \widehat{\boldsymbol{w}}_t\rangle \leqslant \inf_{\boldsymbol{w}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}\rangle + \Psi_T^\star(-\boldsymbol{w}) + \Phi_0\right\}. \tag{11}$$

Rearranging (11) for a given $\boldsymbol{w} \in \mathbb{R}^d$ leads to a regret bound of $\Psi_T^\star(-\boldsymbol{w}) + \Phi_0$. Further bounding this quantity using existing results due to Cutkosky and Orabona (2018); Cutkosky (2019); McMahan and Orabona (2014), leads to the following regret bound (the proof is in Appendix B.1):

Theorem 6 *Under Assumptions 1 and 2, for $(\widehat{\boldsymbol{w}}_t)$ as in (9), we have, with $\ln_+(\cdot) := 0 \vee \ln(\cdot)$,*

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \widehat{\boldsymbol{w}}_t - \boldsymbol{w}\rangle \leqslant \left[2\|\boldsymbol{w}\|\sqrt{V_T \ln_+\left(\frac{2\|\boldsymbol{w}\|V_T}{h_1^2}\right)}\right] \vee \left[4h_T\|\boldsymbol{w}\| \ln\left(\frac{4h_T\|\boldsymbol{w}\|\sqrt{V_T}}{h_1^2}\right)\right] + h_1,$$

for all $\boldsymbol{w} \in \mathcal{W} = \mathbb{R}^d, T \geqslant 1$.

Range-Ratio Problem. While the outputs $(\widehat{\boldsymbol{w}}_t)$ in (9) of FREEGRAD are scale-free for the sequence of hints (h_t) satisfying Assumption 1, there remains one serious issue; the fractions V_T/h_1^2 and h_T/h_1 inside the log-terms in the regret bound of Theorem 6 could in principle be arbitrarily large if h_1 is small enough relative to h_T. Such a problematic ratio has appeared in the regret bounds of many previous algorithms which attempt to adapt to the Lipschitz constant L (Ross et al., 2013; Wintenberger, 2017; Kotłowski, 2017; Mhammedi et al., 2019; Kempka et al., 2019).

When the output set $\mathcal{W}$ is bounded with diameter $D > 0$, this ratio can be dispensed of using a recently proposed restart trick due to Mhammedi et al. (2019), which restarts the algorithm whenever $L_t/L_1 > \sum_{s=1}^{t}\|\boldsymbol{g}_s\|/L_s$. The price to pay for this is merely an additive $O(LD)$ in the regret bound. However, this trick does not directly apply to our setting since in our case $\mathcal{W}$ may be unbounded. Fortunately, we are able to extend the analysis of the restart trick to the unbounded setting where a sequence of hints (h_t) satisfying Assumption 1 is available; the cost we incur in the regret bound is an additive lower-order $\widetilde{O}(\|\boldsymbol{w}\|L)$ term. Algorithm 1 displays our restart "wrapper", FREERANGE, which uses the outputs of FREEGRAD to guarantee the following regret bound (the proof is in Appendix B):

Theorem 7 *Let $(\widehat{\boldsymbol{w}}_t)$ be the outputs of* FREERANGE *(Algorithm 1). Then,*

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \widehat{\boldsymbol{w}}_t - \boldsymbol{w}\rangle \leqslant 2\|\boldsymbol{w}\|\sqrt{2V_T \ln_+\left(\|\boldsymbol{w}\|b_T\right)} + h_T \cdot (16\|\boldsymbol{w}\| \ln_+(2\|\boldsymbol{w}\|b_T) + 2\|\boldsymbol{w}\| + 3),$$

for all $\boldsymbol{w} \in \mathbb{R}^d, T \geqslant 1$, and $(\boldsymbol{g}_t) \subset \mathbb{R}^d$, where $b_T := 2\sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\frac{\|\boldsymbol{g}_s\|}{h_s} + 2\right)^2 \leqslant (T+1)^3$.

We next introduce our second algorithm, in which the variance is only measured in the comparator direction; the algorithm can be viewed as a "full-matrix" version of FREEGRAD.

Algorithm 1 FREERANGE: A Restart Wrapper for the Range-Ratio Problem (under Assumption 2).

Require: Hints (h_t) satisfying Assumption 1.

Set $\tau = 1$;
for $t = 1, 2, \ldots$ **do**
  Observe hint h_t;
  if $h_t/h_\tau > \sum_{s=1}^{t-1} \|\boldsymbol{g}_s\|/h_s + 2$ **then**
    Set $\tau = t$;
  end if
  Output $\widehat{\boldsymbol{w}}_t$ as in (9) with $(h_1, V_{t-1}, \mathbf{G}_{t-1})$ replaced by $(h_\tau, h_\tau^2 + \sum_{s=\tau}^{t-1} \|\boldsymbol{g}_s\|^2,\ \sum_{s=\tau}^{t-1} \boldsymbol{g}_s)$;
end for

3.2. MATRIX-FREEGRAD: Adapting to Directional Variance

Reflecting on the previous subsection, we see that the potential function that we ideally would like to use is $S_t + h_1 \exp\left(\frac{1}{2}\mathbf{G}_t^\intercal \boldsymbol{V}_t^{-1}\mathbf{G}_t - \frac{1}{2}\ln\det \boldsymbol{V}_t\right)$, $t \geqslant 1$, where $\boldsymbol{V}_t = \sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\intercal$. However, as we saw, this is a little too greedy even in one dimension, and we need to introduce some slack to make the potential controllable. In the previous subsection we did this by increasing the scalar denominator from V to $V + \|\mathbf{G}\|$, which acts as a barrier function restricting the norm of $\widehat{\boldsymbol{w}}_t$. In this section, we will instead employ a hard norm constraint. We will further need to include a fudge factor $\gamma > 1$ multiplying $\boldsymbol{V}$ to turn the above shape into a bona fide potential. To describe its effect, we define

$$\rho(\gamma) := \frac{1}{2\gamma}\left(\sqrt{(\gamma+1)^2 - 4e^{\frac{1}{2\gamma}-\frac{1}{2}}\gamma^{3/2}} + \gamma - 1\right), \quad \text{for } \gamma \geqslant 1. \tag{12}$$

The increasing function ρ satisfies $\lim_{\gamma\to 1}\rho(\gamma) = 0$, $\lim_{\gamma\to\infty}\rho(\gamma) = 1$, and $\rho(2) = 0.358649$.

The potential function of this section is parameterized by a *prod factor* $\gamma > 1$ (which we will set to some universal constant). We define

$$\Psi(\mathbf{G}, \boldsymbol{V}, h) := \frac{h_1 \exp\left(\inf_{\lambda\geqslant 0}\left\{\frac{1}{2}\mathbf{G}^\intercal\left(\gamma h_1^2\boldsymbol{I} + \gamma\boldsymbol{V} + \lambda\boldsymbol{I}\right)^{-1}\mathbf{G} + \frac{\lambda\rho(\gamma)^2}{2h^2}\right\}\right)}{\sqrt{\det\left(\boldsymbol{I} + \frac{1}{h_1^2}\boldsymbol{V}\right)}}, \tag{13}$$

where $\mathbf{G} \in \mathbb{R}^d$, $\boldsymbol{V} \in \mathbb{R}^{d\times d}$, and $h > 0$. We introduce the following algorithm to control Ψ.

Definition 8 (MATRIX-FREEGRAD) *In round t, given past sub-gradients $(\boldsymbol{g}_s)_{s<t}$ and a hint $h_t \geqslant L_T$, the* MATRIX-FREEGRAD *prediction is obtained from the gradient of Ψ in the first argument by*

$$\widehat{\boldsymbol{w}}_t \ := \ -\nabla^{(1,0,0)}\Psi(\mathbf{G}_{t-1}, \boldsymbol{V}_{t-1}, h_t), \tag{14}$$

where $\mathbf{G}_{t-1} = \sum_{s=1}^{t-1}\boldsymbol{g}_s$ and $\boldsymbol{V}_{t-1} := \sum_{s=1}^{t-1}\boldsymbol{g}_s\boldsymbol{g}_s^\intercal$.

We can compute $\widehat{\boldsymbol{w}}_t$ in $O(d^3)$ time per round by first computing an eigendecomposition of $\boldsymbol{V}_{t-1}$, followed by a one-dimensional binary search for the $\lambda_\star$ which achieves the inf in (13) with $(\mathbf{G}, \boldsymbol{V}, h) = (\mathbf{G}_{t-1}, \boldsymbol{V}_{t-1}, h_t)$. Then the output is given by

$$\widehat{\boldsymbol{w}}_t = -\Psi(\mathbf{G}_{t-1}, \boldsymbol{V}_{t-1}, h_t)\cdot\left(\gamma h_1^2\boldsymbol{I} + \gamma\boldsymbol{V}_{t-1} + \lambda_\star\boldsymbol{I}\right)^{-1}\mathbf{G}_{t-1}.$$

Our heavy-lifting step in the analysis is the following, which we prove in Appendix C:

Lemma 9 *For any vector $\boldsymbol{g}_t \in \mathbb{R}^d$ and $h_t > 0$ satisfying $\|\boldsymbol{g}_t\| \leqslant h_t$, the vector $\widehat{\boldsymbol{w}}_t$ in (14) ensures*

$$\boldsymbol{g}_t^\intercal \widehat{\boldsymbol{w}}_t \;\leqslant\; \Psi(\mathbf{G}_{t-1}, \boldsymbol{V}_{t-1}, h_t) - \Psi(\mathbf{G}_t, \boldsymbol{V}_t, h_t).$$

From here, we obtain our main result using telescoping and regret-reward duality:

Theorem 10 *Let $\boldsymbol{\Sigma}_T^{-1} := \gamma h_1^2 \boldsymbol{I} + \gamma \boldsymbol{V}_T$. For $(\widehat{\boldsymbol{w}}_t)$ as in (14), we have*

$$\sum_{t=1}^T \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{w}, \boldsymbol{g}_t \rangle \;\leqslant\; h_1 + \sqrt{Q_T^{\boldsymbol{w}} \ln_+ \left(\frac{\det\left(\gamma h_1^2 \boldsymbol{\Sigma}_T\right)^{-1}}{h_1^2} Q_T^{\boldsymbol{w}} \right)}, \quad \textit{for all } \boldsymbol{w} \in \mathbb{R}^d, \textit{ where}$$

$$Q_T^{\boldsymbol{w}} := \max\left\{ \boldsymbol{w}^\intercal \boldsymbol{\Sigma}_T^{-1} \boldsymbol{w}, \frac{1}{2}\left(\frac{h_T^2 \|\boldsymbol{w}\|^2}{\rho(\gamma)^2} \ln\left(\frac{\det\left(\gamma h_1^2 \boldsymbol{\Sigma}_T\right)^{-1}}{h_1^2} \frac{h_T^2 \|\boldsymbol{w}\|^2}{\rho(\gamma)^2} \right) + \boldsymbol{w}^\intercal \boldsymbol{\Sigma}_T^{-1} \boldsymbol{w} \right) \right\}.$$

Note in particular that the result is scale-free. Expanding the main case of the theorem (modest $\|\boldsymbol{w}\|$), we find regret bounded by

$$\sum_{t=1}^T \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{w}, \boldsymbol{g}_t \rangle \;\leqslant\; h_1 + h_1 \sqrt{\gamma \boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w} \ln_+ \left(\gamma \boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w} \det \boldsymbol{Q} \right)} \quad \text{where} \quad \boldsymbol{Q} = \boldsymbol{I} + \boldsymbol{V}_T / h_1^2.$$

This bound looks almost like an ideal upgrade of that in Theorem 6, though technically, the bounds are not really comparable since the $\ln\det \boldsymbol{Q}$ can be as large as $d \ln T$, potentially canceling the advantage of having $\boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}$ instead of $\|\boldsymbol{w}\|^2 \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$ inside the square-root. The matrix $\boldsymbol{Q}$ and hence any directional variance $\boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}$ is scale-invariant. The only fudge factor in the answer is the $\gamma > 1$. We currently cannot tolerate $\gamma = 1$, for then $\rho(\gamma) = 0$ so the lower-order term would explode. We note that a bound of the form given in the previous display, with the $\ln\det \boldsymbol{Q}$ replaced by the larger term $d \ln \operatorname{tr} \boldsymbol{Q}$, was achieved by a previous (not scale-free) algorithm due to Cutkosky and Orabona (2018).

Remark 11 *As Theorem 7 did in the previous subsection, our restarts method allows us to get rid of problematic scale ratios in the regret bound of Theorem 10; this can be achieved using* FREERANGE *with $(\widehat{\boldsymbol{w}}_t)$ set to be as in (14) instead of (9). The key idea behind the proof of Theorem 7 is to show that the regrets from all but the last two epochs add up to a lower-order term in the final regret bound. This still holds when $(\widehat{\boldsymbol{w}}_t)$ are the outputs of* MATRIX-FREEGRAD *instead of* FREEGRAD*, since by Theorem 10, the regret bound of* MATRIX-FREEGRAD *is of order at most d times the regret of* FREEGRAD *within any given epoch.*

As a final note about the algorithm, we may also develop a "one-dimensional" variant by replacing matrix inverse and determinant by their scalar analogues applied to $V_T = \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$. One effect of this is that the minimization in λ can be computed in closed form. The resulting potential and corresponding algorithm and regret bound are very close to those of Section 3.1.

Conclusion The algorithms designed in this section can now be used in the role of algorithm A in the reductions presented in Section 2.2. This will yield algorithms which achieve our goal; they adapt to the norm of the comparator and the Lipschitz constant and are completely scale-free, for both bounded and unbounded sets, without requiring hints. We now show that the penalties incurred by these reductions are not improvable.

4. Lower Bounds

As we saw in Corollary 3, given a base algorithm A, which takes a sequence of hints (h_t) such that $h_t \geqslant L_t$ for all $t \geqslant 1$, and which suffers regret $R_T^{\mathrm{A}}(\boldsymbol{w})$ against comparator $\boldsymbol{w} \in \mathcal{W}$, there exists an algorithm B for the setting without hints which suffers the same regret against $\boldsymbol{w}$ up to an additive penalty $L_T\|\boldsymbol{w}\|^3 + L_T\sqrt{\max_{t\in[T]} B_t}$, where $B_t = \sum_{s=1}^t \|\boldsymbol{g}_s\|/L_t$. In this section, we show that the penalty $L_T\|\boldsymbol{w}\|^3$ is not improvable if one insists on a regret bound of order $\widetilde{O}(\sqrt{T})$. We also show that it is not possible to replace the penalty $L_T\sqrt{\max_{t\in[T]} B_t}$ by the typically smaller quantity $\sqrt{V_T}$, where $V_T = \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$. Our starting point is the following lemma:

Lemma 12 *For all $t \geqslant 1$, past sub-gradients $(\boldsymbol{g}_s)_{s<t}$ and past and current outputs $(\widehat{\boldsymbol{w}}_s)_{s\leqslant t} \in \mathbb{R}^d$,*

$$\exists \boldsymbol{g}_t \in \mathbb{R}^d, \quad \sum_{s=1}^t \langle \boldsymbol{g}_s, \widehat{\boldsymbol{w}}_s\rangle \geqslant \|\widehat{\boldsymbol{w}}_t\| \cdot L_t/2, \quad \textit{where } L_t = \max_{s\leqslant t} \|\boldsymbol{g}_s\|.$$

Proof We want to find $\boldsymbol{g}_t$ such that $\langle \boldsymbol{g}_t, \widehat{\boldsymbol{w}}_t\rangle \geqslant \|\boldsymbol{w}_t\|L_t/2 - S_{t-1}$, where $S_{t-1} := \sum_{s=1}^{t-1}\langle \boldsymbol{g}_s, \widehat{\boldsymbol{w}}_s\rangle$. By restricting $\boldsymbol{g}_t$ to be aligned with $\widehat{\boldsymbol{w}}_t$, the problem reduces to finding $x = \|\boldsymbol{g}_t\|$ such that

$$x\|\widehat{\boldsymbol{w}}_t\| - |\|\widehat{\boldsymbol{w}}_t\| \cdot (L_{t-1} \vee x)/2 - S_{t-1}| \geqslant 0. \tag{15}$$

The LHS of (15) is a piece-wise linear function in x which goes to infinity as $x \to \infty$. Therefore, there exists a large enough $x \geqslant 0$ which satisfies (15). ■

Observe that if $\|\widehat{\boldsymbol{w}}_t\| \geqslant D_t > 0$, for $t \geqslant 1$, then by Lemma 12, there exists a sub-gradient $\boldsymbol{g}_t$ which makes the regret against $\boldsymbol{w} = \boldsymbol{0}$ at round t at least $D_t L/2$. This essentially means that if the sub-gradients $(\boldsymbol{g}_t)$ are unbounded, then the outputs $(\widehat{\boldsymbol{w}}_t)$ must be in a bounded set whose diameter will depend on the desired regret bound; if one insists on a regret of order $\widetilde{O}(\sqrt{T})$, then the outputs $\widehat{\boldsymbol{w}}_t, t \geqslant 1$, must be in a ball of radius at most $\widetilde{O}(\sqrt{T})$.

Cutkosky (2019) posed the question of whether there exists an algorithm which can guarantee a regret bound of order $\|\boldsymbol{w}\|\sqrt{V_T \ln(\|\boldsymbol{w}\|T)} + \sqrt{V_T \ln T} + L\|\boldsymbol{w}\|^3$, with $V_T = \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$, while adapting to both L and $\|\boldsymbol{w}\|$ (which essentially means replacing $L\sqrt{\max_{t\in[T]} B_t}$ in Corollary 3 by $\sqrt{V_T \ln T}$). Here, we ask the question whether $\|\boldsymbol{w}\|\sqrt{V_T \ln(\|\boldsymbol{w}\|T)} + \sqrt{V_T \ln T} + L\|\boldsymbol{w}\|^\nu$ is possible for any $\nu \geqslant 1$. If such an algorithm exists, then by Lemma 12, there exists a constant $b > 0$ such that its outputs $(\widehat{\boldsymbol{w}}_t)$ satisfy $\|\widehat{\boldsymbol{w}}_t\| \leqslant b\sqrt{V_t \ln t}/L_t$, for all $t \geqslant 1$. The next lemma, when instantiated with $\alpha = 2$, gives us a regret lower-bound on such algorithms (the proof is in Appendix D):

Lemma 13 *For all $b, c, \beta \geqslant 0$, $\nu \geqslant 1$, and $\alpha \in]1, 2]$, there exists $(\boldsymbol{g}_t) \in \mathbb{R}^d$, $T \geqslant 1$, and $\boldsymbol{w} \in \mathbb{R}^d$, such that for any sequence $(\widehat{\boldsymbol{w}}_t)$ satisfying $\|\widehat{\boldsymbol{w}}_t\| \leqslant b \cdot \sqrt{V_{\alpha,t} \ln(t)/L_t^\alpha}$, for all $t \in [T]$, where $V_{\alpha,t} := \sum_{s=1}^t \|\boldsymbol{g}_s\|^\alpha$, we have*

$$\sum_{t=1}^T \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{w}, \boldsymbol{g}_t\rangle \geqslant c \cdot \ln(1 + \|\boldsymbol{w}\|T)^\beta \cdot (L_T\|\boldsymbol{w}\|^\nu + L_T^{1-\alpha/2}(\|\boldsymbol{w}\| + 1)\sqrt{V_{\alpha,T} \ln T}).$$

By combining the results of Lemma 12 and 13, we have the following regret lower bound for algorithms with can adapt to both L and $\|\boldsymbol{w}\|$:

Theorem 14 *For any $\alpha \in]1,2]$, $c > 0$ and $\nu \geqslant 1$, there exists no algorithm that guarantees, up to multiplicative log-factors in $\|\boldsymbol{w}\|$ and T, a regret bound of the form $c \cdot (L_T\|\boldsymbol{w}\|^\nu + L_T^{1-\alpha/2}(\|\boldsymbol{w}\| + 1)\sqrt{V_{\alpha,T}\ln T})$, for all $T \geqslant 1$, $\boldsymbol{w} \in \mathbb{R}^d$, and $(\boldsymbol{g}_t) \subset \mathbb{R}^d$, where $V_{\alpha,T} := \sum_{t=1}^T \|\boldsymbol{g}_t\|^\alpha$.*

Proof By Lemma 12, the only candidate algorithms are those whose outputs $(\widehat{\boldsymbol{w}}_t)$ satisfy $\|\widehat{\boldsymbol{w}}_t\| \leqslant b\sqrt{V_{\alpha,t}\ln(t)/L_t^\alpha}$, for all $t \geqslant 1$, for some constant $b > 0$. By Lemma 13, no such algorithms can achieve the desired regret bound. ■

The regret lower bound in Theorem 14 does not apply to the case where $\alpha = 1$. In fact, thanks to Corollary 3 and our main algorithm in Section 3 (which can play the role of Algorithm A in Corollary 3), we know that there exists an algorithm B which guarantees a regret bound of order $\widetilde{O}(L_T\|\boldsymbol{w}\|^3 + \|\boldsymbol{w}\|\sqrt{V_T\ln(\|\boldsymbol{w}\|T)} + L_T\sqrt{\max_{t\in[T]} B_t})$, where $B_t = \sum_{s=1}^t \|\boldsymbol{g}_s\|/L_t$. Next we show that if one insists on a regret bound of order $\sqrt{B_T}$, or even $\sqrt{T}$ (up to log-factors), the exponent in $\|\boldsymbol{w}\|^3$ is unimprovable (the proof of Theorem 15 is in Appendix D.2).

Theorem 15 *For any $\nu \in [1,3[$ and $c > 0$, there exists no algorithm that guarantees, up to multiplicative log-factors in $\|\boldsymbol{w}\|$ and T, a regret bound of the form $c \cdot (L_T\|\boldsymbol{w}\|^\nu + L_T(\|\boldsymbol{w}\| + 1)\sqrt{T\ln T})$, for all $T \geqslant 1$, $\boldsymbol{w} \in \mathbb{R}^d$, and $(\boldsymbol{g}_t) \subset \mathbb{R}^d$.*

5. Application to Learning Linear Models with Online Algorithms

In this section, we consider the setting of online learning of linear models which is a special case of OCO. At the start of each round $t \geqslant 1$, a learner receives a feature vector $\boldsymbol{x}_t \in \mathcal{W} = \mathbb{R}^d$, then issues a prediction $\widehat{y}_t \in \mathbb{R}$ in the form of an inner product between $\boldsymbol{x}_t$ and a vector $\widehat{\boldsymbol{u}}_t \in \mathbb{R}^d$, *i.e.* $\widehat{y}_t = \widehat{\boldsymbol{u}}_t^\intercal \boldsymbol{x}_t$. The environment then reveals a label $y_t \in \mathbb{R}$ and the learner suffers loss $\ell(y_t, \widehat{y}_t)$, where $\ell\colon \mathbb{R}^2 \to \mathbb{R}$ is a fixed loss function which is convex and 1-Lipschitz in its second argument; this covers popular losses such as the logistic, hinge, absolute and Huberized squared loss. (Technically, the machinery developed so far and the reductions in Section 2.2 allow us to handle the non-Lipschitz case).

In the current setting, the regret is measured against the best fixed "linear model" $\boldsymbol{w} \in \mathbb{R}^d$ as

$$\textsc{Regret}_T(\boldsymbol{w}) := \sum_{t=1}^T \ell(y_t, \widehat{y}_t) - \sum_{t=1}^T \ell(y_t, \boldsymbol{w}^\intercal \boldsymbol{x}_t) \leqslant \sum_{t=1}^T \delta_t \langle \boldsymbol{x}_t, \widehat{\boldsymbol{u}}_t - \boldsymbol{w}\rangle, \tag{16}$$

where the last inequality holds for any sub-gradients $\delta_t \in \partial^{(0,1)}\ell(y_t, \widehat{y}_t)$, $t \geqslant 1$, due to the convexity of ℓ in its second argument, which in turn makes the function $f_t(\boldsymbol{w}) := \ell(y_t, \boldsymbol{w}^\intercal \boldsymbol{x}_t)$ convex for all $\boldsymbol{w} \in \mathcal{W} = \mathbb{R}^d$. Here, $\partial^{(0,1)}\ell$ denotes the sub-differential of ℓ with respect to its second argument. Thus, minimizing the regret in (16) fits into the OCO framework described in Section 2. In fact, we will show how our algorithms from Section 3 can be applied in this setting to yield scale-free, and even *rotation-free*, (all with respect to the feature vectors $(\boldsymbol{x}_t)$) algorithms for learning linear models. These algorithms can, without any prior knowledge on $\boldsymbol{w}$ or $(\boldsymbol{w}^\intercal \boldsymbol{x}_t)$, achieve regret bounds against any $\boldsymbol{w} \in \mathbb{R}^d$ matching (up to log-factors) that of OGD with optimally tuned learning rate.

As in Section 3, we focus on algorithms which make predictions based on observed sub-gradients $(\boldsymbol{g}_t)$; in this case, $\boldsymbol{g}_t = \boldsymbol{x}_t\delta_t \in \boldsymbol{x}_t \cdot \partial^{(0,1)}\ell(y_t, \widehat{y}_t) = \partial f_t(\widehat{\boldsymbol{u}}_t)$, $t \geqslant 1$, where $f_t(\boldsymbol{w}) = \ell(y_t, \boldsymbol{w}^\intercal \boldsymbol{x}_t)$. Since the loss ℓ is 1-Lipschitz, we have $|\delta_t| \leqslant 1$, for all $\delta_t \in \partial^{(0,1)}\ell(y_t, \widehat{y}_t)$ and $t \geqslant 1$, and so $\|\boldsymbol{g}_t\| \leqslant \|\boldsymbol{x}_t\|$. Since $\boldsymbol{x}_t$ is revealed at the beginning of round $t \geqslant 1$, the hint

$$h_t = \max_{s\leqslant t} \|\boldsymbol{x}_s\| \geqslant L_T = \max_{s\leqslant t} \|\boldsymbol{g}_s\| \tag{17}$$

is available ahead of outputting $\widehat{\boldsymbol{u}}_t$, and so our algorithms from Section 3 are well suited for this setting.

Improvement over Current Algorithms. We improve on current state-of-the-art algorithms in two ways; First, we provide a (coordinate-wise) scale-invariant algorithm which guarantees regret bound, against any $\boldsymbol{w} \in \mathbb{R}^d$, of order

$$\sum_{i=1}^{d} |w_i|\sqrt{V_{T,i}\ln(|w_i|\sqrt{V_{T,i}}T)} + |w_i|\ln_+(|w_i|\sqrt{V_{T,i}}T), \tag{18}$$

where $V_{T,i} := |x_{1,i}|^2 + \sum_{t=1}^{T}\delta_t^2|x_{t,i}|^2, i \in [d]$, which improves the regret bound of the current state-of-the-art scale-invariant algorithm SCLNOL$_1$ (Kempka et al., 2019) by a $\sqrt{\ln(\|\boldsymbol{w}\|T)}$ factor. Second, we provide an algorithm that is both scale and rotation invariant with respect to the input feature vectors $(\boldsymbol{x}_t)$ with a state-of-the-art regret bound; by scale and rotation invariance we mean that, if the sequence of feature vectors $(\boldsymbol{x}_t)$ is multiplied by $c\boldsymbol{O}$, where $c > 0$ and $\boldsymbol{O}$ is any special orthogonal matrix in $\mathbb{R}^{d\times d}$, the outputs $(\widehat{y}_t)$ of the algorithm remain unchanged. Arguably the closest algorithm to ours in the latter case is that of Kotłowski (2017) whose regret bound is essentially of order $\widetilde{O}(\sqrt{\boldsymbol{w}^\intercal \boldsymbol{S}_T \boldsymbol{w}})$ for any comparator $\boldsymbol{w} \in \mathbb{R}^d$, where $\boldsymbol{S}_T = \sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal$. However, in our case, instead of the matrix $\boldsymbol{S}_T$, we have $\boldsymbol{V}_T := \|\boldsymbol{x}_1\|^2\boldsymbol{I} + \sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t^\intercal\delta_t^2$, where $\delta_t \in \partial^{(0,1)}\ell(y_t, \widehat{y}_t), t \geqslant 1$, which can yield a much smaller bound for small (δ_t) (this typically happens when the algorithm starts to "converge").

A Scale-Invariant Algorithm. To design our first scale-invariant algorithm, we will use the outputs $(\widehat{\boldsymbol{w}}_t)$ of FREEGRAD in (9) with (h_t) as in (17), and a slight modification of FREERANGE (see Algorithm 2). This modification consists of first scaling the outputs $(\widehat{\boldsymbol{w}}_t)$ of FREEGRAD by the initial hint of the current epoch to make the predictions $(\widehat{y}_t)$ scale-invariant. By Theorem 20 below, the regret bound corresponding to such scaled outputs will have a lower-order term which, unlike in the regret bound of Theorem 6, does not depend on the initial hint. This breaks our current analysis of FREERANGE in the proof of Theorem 7 which we used to overcome the range-ratio problem. To solve this issue, we further scale the output $\widehat{\boldsymbol{w}}_t$ at round $t \geqslant 1$ by the sum $\sum_{s=1}^{\tau}\|\boldsymbol{x}_s\|/h_s$, where τ denotes the first index of the current epoch (see Algorithm 2). Due to this change, the proof of the next theorem differs slightly from that of Theorem 7.

First, we study the regret bound of Algorithm 2 in the case where $\mathcal{W} = \mathbb{R}$.

Theorem 16 *Let $d = 1$ and (h_t) be as in (17). If $(\widehat{u}_t)$ are the outputs of Algorithm 2, then for all $w \in \mathbb{R}; T \geqslant 1$; $(x_t, y_t) \subset \mathbb{R}^2$, s.t. $h_1 = |x_1| > 0$; and $\delta_t \in \partial^{(0,1)}\ell(y_t, x_t\widehat{u}_t)$, $t \in [T]$,*

$$\sum_{t=1}^{T}\delta_t x_t\cdot(\widehat{u}_t - w) \leqslant 2|w|\sqrt{V_T\ln_+(2|w|^2V_Tc_T)} + h_T|w|(14\ln_+(2|w|\sqrt{2V_Tc_T}) + 1) + 2 + \ln B_T, \tag{19}$$

where $V_T := |x_1|^2 + \sum_{t=1}^{T}\delta_t^2x_t^2$, $c_T := 2B_T^2\sum_{t=1}^{T}\left(\sum_{s=1}^{t}\frac{|x_s|}{h_s}\right)^2 \leqslant T^5$, and $B_T = \sum_{s=1}^{T}\frac{|x_s|}{h_s} \leqslant T$.

The proof of Theorem 16 is in Appendix E. If $(\widehat{u}_t)$ are the outputs of Algorithm 2 in the one-dimensional case, then by Theorem 16 and (16), the algorithm which, at each round $t \geqslant 1$, predicts

Algorithm 2 Modified FREERANGE for the setting of online learning of linear models.

Require: The hints (h_t) as in (17).

1: Set $\tau = 1$;
2: **for** $t = 1, 2, \ldots$ **do**
3: Observe hint h_t;
4: **if** $h_t/h_\tau > \sum_{s=1}^{t-1} \|\boldsymbol{x}_s\|/h_s + 1$ **then**
5: Set $\tau = t$;
6: **end if**
7: Output $\widehat{\boldsymbol{u}}_t = \widehat{\boldsymbol{w}}_t \cdot \left(h_\tau \cdot \sum_{s=1}^{\tau} \frac{\|\boldsymbol{x}_s\|}{h_s}\right)^{-1}$, where $\widehat{\boldsymbol{w}}_t$ is as in (9) with $(h_1, V_{t-1}, \mathbf{G}_{t-1})$ replaced by $(h_\tau, h_\tau^2 + \sum_{s=\tau}^{t-1} \|\boldsymbol{g}_s\|^2, \sum_{s=\tau}^{t-1} \boldsymbol{g}_s)$;
8: **end for**

$\widehat{y}_t = x_t \widehat{u}_t$ has regret bounded from above by the RHS of (19). Note also that the outputs $(\widehat{y}_t)$ are scale-invariant.

Now consider an algorithm A which at round $t \geqslant 1$ predicts $\widehat{y}_t = \sum_{i=1}^d x_{t,i}\widehat{u}_{t,i}$, where $(\widehat{u}_{t,i}), i \in [d]$, are the outputs of Algorithm 2 when applied to coordinate i; in this case, we will have a sequence of hints $(h_{t,i})$ for each coordinate i satisfying $h_{t,i} = \max_{s \leqslant t} |x_{t,i}|$, for all $t \geqslant 1$. Algorithm A is coordinate-wise scale-invariant, and due to (16) and Theorem 16, it guarantees a regret bound of the form (18). We note, however, that a factor d will appear multiplying the lower-order term $(2 + \ln B_T)$ in (19) (since the regret bounds for the different coordinates are added together). To avoid this, at the cost of a factor d appearing inside the logarithms in (18), it suffices to divide the outputs of algorithm A by d. To see why this works, see Theorem 20 in the appendix.

A Rotation-Invariant Algorithm. To obtain a rotation and scale-invariant online algorithm for learning linear models we will make use of the outputs of MATRIX-FREEGRAD instead of FREEGRAD. Let $(\widehat{y}_t)$ be the sequence of predictions defined by

$$\widehat{y}_t = \boldsymbol{x}_t^\intercal \widehat{\boldsymbol{w}}_t / h_1, \; t \geqslant 1, \tag{20}$$

with (h_t) as in (17) and where $\widehat{\boldsymbol{w}}_t$ are the predictions of a variant of MATRIX-FREEGRAD, where the leading h_1 in the potential (13) is replaced by 1 (we analyze this variant in Appendix C.1).

Theorem 17 *Let $\gamma > 0$ and (h_t) be as in (17). If $(\widehat{y}_t)$ are as in (20), then*

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \forall T \geqslant 1, \forall (\boldsymbol{g}_t) \subset \mathbb{R}^d, \;\; \text{REGRET}_T(\boldsymbol{w}) \;\leqslant\; 1 + \sqrt{Q_T^{\boldsymbol{w}} \ln_+ \left(\det\left(\gamma h_1^2 \boldsymbol{\Sigma}_T\right)^{-1} Q_T^{\boldsymbol{w}}\right)}, \;\; \textit{where}$$

$$Q_T^{\boldsymbol{w}} := \max\left\{\boldsymbol{w}^\intercal \boldsymbol{\Sigma}_T^{-1} \boldsymbol{w}, \frac{1}{2}\left(\frac{h_T \|\boldsymbol{w}\|^2}{\rho(\gamma)^2} \ln\left(\frac{h_T \|\boldsymbol{w}\|^2}{\rho(\gamma)^2} \det\left(\gamma h_1^2 \boldsymbol{\Sigma}_T\right)^{-1}\right) + \boldsymbol{w}^\intercal \boldsymbol{\Sigma}_T^{-1} \boldsymbol{w}\right)\right\},$$

and $\boldsymbol{\Sigma}_T^{-1} := \gamma h_1^2 \boldsymbol{I} + \gamma \sum_{t=1}^T \boldsymbol{g}_t \boldsymbol{g}_t^\intercal$.

Proof It suffices to use (16) and instantiate the regret bound in Theorem 21 with $(\epsilon, \sigma^{-2}) = (1, \gamma h_1^2)$. ■

The range-ratio problem manifests itself again in Theorem 17 through the term $\det(\gamma h_1^2 \boldsymbol{\Sigma}_T)^{-1}$. This can be solved using the outputs of Algorithm 2, where in Line 7, $\widehat{\boldsymbol{w}}_t$ is taken to be as in (20) (see Remark 11).

Acknowledgments

This work was also supported by the Australian Research Council and Data61.

References

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, 1997.

Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, pages 874–894, 2019.

Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Conference on Learning Theory*, pages 643–677, 2017.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1493–1529, 2018.

Tim van Erven and Wouter M. Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 3666–3674, 2016.

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Zigzag: A new approach to adaptive online learning. In *Proc. of the 2017 Annual Conference on Learning Theory (COLT)*, pages 876–924, 2017.

Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Online learning: Sufficient statistics and the Burkholder method. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 3028–3064, 2018.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3–4):157–325, 2016.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Science & Business Media, 2004.

Michal Kempka, Wojciech Kotłowski, and Manfred K. Warmuth. Adaptive scale-invariant online algorithms for learning linear models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3321–3330, 2019.

Wojciech Kotłowski. Scale-invariant unconstrained online learning. In *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, pages 412–433, 2017.

H. Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in Hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pages 1020–1039, 2014.

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 244–256, 2010.

Zakaria Mhammedi, Wouter M. Koolen, and Tim van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 2490–2511, 2019.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems*, pages 577–585, 2016a.

Francesco Orabona and Dávid Pál. Open problem: Parameter-free and scale-free online algorithms. In *Conference on Learning Theory*, pages 1659–1664, 2016b.

Stephane Ross, Paul Mineiro, and John Langford. Normalized online learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 537–545, 2013.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106 (1):119–141, 2017.

Appendix A. Proof of Theorem 5

The proof of Theorem 5 relies on the following key lemma:

Lemma 18 *For $G, g \in \mathbb{R}$ and, $V > 0$, define*

$$\Theta(G,V,g) := \frac{\sqrt{V}}{\sqrt{V+g^2}} \cdot \exp\left(\frac{(G+g)^2}{2V+2g^2+2|G+g|} - \frac{G^2}{2V+2|G|}\right) - \frac{gG(|G|+2V)}{2(|G|+V)^2} - 1.$$

It holds that $\Theta(G,V,g) \leqslant 0$, for all $G \in \mathbb{R}$, $V > 0$, and $g \in [-1,1]$.

Proof For notational simplicity we assume $G \geqslant 0$. Let us look at

$$\Gamma(G,V,g) := \frac{1}{2}\frac{(g+G)^2}{V+g^2+|G+g|} - \frac{1}{2}\frac{G^2}{G+V} - \ln\left(1+\frac{gG(G+2V)}{2(G+V)^2}\right) - \frac{1}{2}\ln\left(1+\frac{g^2}{V}\right).$$

Since $\ln$ is increasing, we have that $\Theta \leqslant 0$, if and only if, $\Gamma \leqslant 0$, and so we want to show $\Gamma \leqslant 0$ for all $V > 0, G \geqslant 0$, and $g \in [-1,1]$. Our approach will be to show that Γ is increasing in V. The result then follows from $\lim_{V\to\infty}\Gamma = 0$. It remains to study the derivative

$$\begin{aligned}\frac{\partial\Gamma}{\partial V} &= -\frac{1}{2}\frac{(g+G)^2}{(|g+G|+g^2+V)^2} + \frac{2gGV}{(G+V)(2(G+V)^2+gG(G+2V))} \\ &\quad + \frac{1}{2}\frac{G^2}{(G+V)^2} + \frac{1}{2}\frac{g^2}{g^2V+V^2}.\end{aligned}$$

Factoring this as a ratio of polynomials, we obtain:

$$\frac{\partial\Gamma}{\partial V} = \frac{\alpha_0+\alpha_1V+\alpha_2V^2+\alpha_3V^3+\alpha_4V^4+\alpha_5V^5}{2V\left(g^2+V\right)(G+V)^2\left((g+2)G^2+2(g+2)GV+2V^2\right)\left(|g+G|+g^2+V\right)^2},$$

where $\alpha_i, i \in [5]$, are polynomials in g and G whose explicit (yet gruesome) expressions are:

$$\begin{aligned}
\alpha_0 &= g^2(g+2)G^4\left(|g+G|+g^2\right)^2 \\
\alpha_1 &= g^2(g+2)G^3\left(2\left(g^2(G+4)+G\right)|g+G|+g^4(G+4)+2g^2(G+2)+8gG+4G^2\right) \\
\alpha_2 &= g^2G^2\begin{pmatrix}2\left(g^3(2G+9)+4g^2(G+3)+2gG(G+2)+4G(G+2)\right)|g+G| \\ +3g^2(3g+4)\left(g^2+1\right)-2(g+2)G^3 \\ +(g(g(3g+2)+2)+14)G^2+2g(g(g(g(g+2)+3)+15)+12)G\end{pmatrix} \\
\alpha_3 &= G\begin{pmatrix}2\left(6g^5+2g^4(G+4)+g^3G(4G+13)+4g^2G(2G+3)+gG^3+2G^3\right)|g+G| \\ +(g(g+2)(2g+13)+20)g^3G+2\left(g\left(3g^2+g-6\right)+10\right)g^2G^2 \\ +2(3g+4)\left(g^2+1\right)g^4-2(g+2)G^4-2(g(g+4)+2)gG^3\end{pmatrix} \\
\alpha_4 &= 2\begin{pmatrix}2\left(g^4+(5g+4)g^2G+(g+2)G^3+2(g+1)gG^2\right)|g+G|+g^6+g^4 \\ +(g(7g+4)+4)g^3G+((g-2)g+6)g^2G^2-2(g+2)G^4+(g(2g-3)-8)gG^3\end{pmatrix} \\
\alpha_5 &= 2\left(2|g+G|^3+g^4+4g^3G-6gG^2-2G^3\right)
\end{aligned}$$

Under our assumptions $V > 0$, $G > 0$ and $g \in [-1,1]$, the denominator of $\partial\Gamma/\partial V$ above is positive. Furthermore, its numerator, regarded as a polynomial in V, has exclusively positive

```
In[ ]:= $Assumptions = V > 0 && Element[G, Reals] && -1 ≤ g ≤ 1;
```

Let us call “gap” the difference between what we have and the upper bound we want to establish.

```
Φ[G_, V_] = -G^2/(2 (V + Abs[G])) + 1/2 Log[V];
```

```
In[ ]:= gap = -Log[1 - g D[Φ[G, V], G]] - (Φ[G + g, V + g^2] - Φ[G, V]) // FullSimplify;
```

We want to show that the gap is ≤ 0. Our approach will be to show that gap is increasing, so that we can then bound it by the limit

```
In[ ]:= Limit[gap, V → Infinity]
Out[ ]= 0
```

So why is gap increasing? Let us take the derivative

```
In[ ]:= dgap = D[gap, V];
```

and write it as a ratio of polynomials

```
In[ ]:= {num, den} = With[{rpoly = Factor[dgap]}, {Numerator[rpoly], Denominator[rpoly]} // Simplify];
```

Now the denominator is always positive

```
In[ ]:= den > 0 // FullSimplify
Out[ ]= True
```

We will show that the numerator is positive by showing that it is a polynomial with only positive coefficients. Here are the coefficients on the monomials V^i for i=0,1,...

```
In[ ]:= coefs = CoefficientList[num, V] // Simplify;
```

And all the coefficients are positive

```
In[ ]:= Map[FullSimplify[# ≥ 0] &, coefs]
Out[ ]= {True, True, True, True, True, True}
```

Figure 1: Mathematica notebook in support of Theorem 5.

coefficients $\alpha_i \geqslant 0$, as can be verified using computer algebra software (we used Mathematica's `FullSimplify`—see Figure 1). This implies that $\partial\Gamma/\partial V \geqslant 0$, for all $G \in \mathbb{R}$, $V > 0$, and $g \in [-1, 1]$, and so $\Gamma \leqslant \lim_{V\to\infty} \Gamma = 0$. ■

We are ourselves a bit disgruntled about the opacity of the above proof. On the one hand, it is just a tedious verification of an analytic statement about a function of three scalar variables, and one might expect that tighter statements require more sophisticated techniques (c.f. Kotłowski, 2017, Appendix F). It is quite plausible that positivity may be established in a somewhat more streamlined fashion using Sum-of-Squares techniques. Yet on the other hand, we were hoping to gain, from the proof, a deeper insight into the design of potential functions. Unfortunately this did not materialize. In particular, we still do not know how to address the multi-dimensional case of our Section 3.2 with

a similar potential. Controlling the intuitive upgrade of (7) where the exponential is replaced by

$$\exp\left(\sup_{\boldsymbol{u}} \frac{(\boldsymbol{u}^\intercal \mathbf{G})^2}{2(\boldsymbol{u}^\intercal \boldsymbol{V}\boldsymbol{u} + L\|\boldsymbol{u}\| \cdot |\boldsymbol{u}^\intercal \mathbf{G}|)} - \frac{1}{2}\ln\det \boldsymbol{V}\right)$$

is impossible, as witnessed by numerical counterexamples returned by random search, already in dimension 2.

We need one more result before we prove Theorem 5:

Lemma 19 *Let $G, s \in \mathbb{R}$ and $V, h \geqslant 0$. Then, the function*

$$g \mapsto \frac{1}{\sqrt{V+g^2}} \exp\left(\frac{g^2+2s+G^2}{2V+2g^2+2h\sqrt{g^2+2s+G^2}}\right),$$

is non-increasing on $\{g \geqslant 0 \mid g^2+2s+G^2 \geqslant 0\}$.

Proof It suffices to show that the function

$$\Xi(g) := \frac{g^2+2s+G^2}{2V+2g^2+2h\sqrt{g^2+2s+G^2}} - \frac{1}{2}\ln\left(V+g^2\right),$$

is non-increasing on $\{g \geqslant 0 \mid g^2+2s+G^2 \geqslant 0\}$. Evaluating the derivative of Ξ, we find that

$$\frac{\mathrm{d}\Xi}{\mathrm{d}g}(g) = -\frac{gN \cdot \left(2h^2N + 2VN + 2g^2N + 3g^2h + 3hV\right)}{2\left(g^2+V\right)\left(hN+g^2+V\right)^2}, \tag{21}$$

where $N := \sqrt{g^2+G^2+2s}$. The derivative in (21) is non-positive for all $V, h \geqslant 0$ and $g \geqslant 0$ such that $g^2+2s+G^2 \geqslant 0$. ■

Proof of Theorem 5. We will proceed by induction. By the fact that $\|\mathbf{G}_0\| = 0$ and the definition of the potential in (7), we have $\Phi_0 = h_1$. Now let $t \geqslant 0$, $h_1 > 0$, and $(S_t, V_t, h_t, \mathbf{G}_t) \in \mathbb{R} \times \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{>0} \times \mathbb{R}^d$. We will show that $\Phi_{t+1} - \Phi_t \leqslant 0$. First, note that for any $h_{t+1} \geqslant h_t$, we have

$$S_t + \frac{h_1^2}{\sqrt{V_t}} \cdot \exp\left(\frac{\|\mathbf{G}_t\|^2}{2V_t + 2h_{t+1}\|\mathbf{G}_t\|}\right) \leqslant \Phi_t = S_t + \frac{h_1^2}{\sqrt{V_t}} \cdot \exp\left(\frac{\|\mathbf{G}_t\|^2}{2V_t + 2h_t\|\mathbf{G}_t\|}\right).$$

Thus, for any $\boldsymbol{g}_{t+1} \in \mathbb{R}^d$ such that $\|\boldsymbol{g}_{t+1}\| \leqslant h_{t+1}$, and $(S, V, h, \mathbf{G}, \boldsymbol{g}) := (S_t, V_t, h_{t+1}, \mathbf{G}_t, \boldsymbol{g}_{t+1})$,

$$\begin{aligned}\Phi_{t+1} - \Phi_t \leqslant\ & \frac{h_1^2}{\sqrt{V+\|\boldsymbol{g}\|^2}} \cdot \exp\left(\frac{\|\boldsymbol{g}+\mathbf{G}\|^2}{2V+2\|\boldsymbol{g}\|^2+2h\|\boldsymbol{g}+\mathbf{G}\|}\right) \\ & - \left(1 + \frac{\langle \boldsymbol{g}, \mathbf{G}\rangle \cdot (2h\|\mathbf{G}\|+2V)}{2(h\|\mathbf{G}\|+V)^2}\right) \cdot \frac{h_1^2}{\sqrt{V}} \cdot \exp\left(\frac{\|\mathbf{G}\|^2}{2V+2h\|\mathbf{G}\|}\right).\end{aligned} \tag{22}$$

Let $\boldsymbol{g}_\star$ be the vector $\boldsymbol{g} \in \mathcal{B}_h$ which maximizes the RHS of (22), where $\mathcal{B}_h$ is the ball in $\mathbb{R}^d$ of radius h. Suppose that $\mathbf{G} \neq \mathbf{0}$, and let $\mathcal{H} := \{\boldsymbol{g} \in \mathbb{R}^d \mid \langle \boldsymbol{g}, \mathbf{G}\rangle = \langle \boldsymbol{g}_\star, \mathbf{G}\rangle\}$. Note that within the hyperplane $\mathcal{H}$, only the first term on the RHS of (22) varies. Since $\boldsymbol{g}_\star$ is the maximizer of the RHS of (22) within $\mathcal{B}_h$, instantiating Lemma 19 with $s := \langle \boldsymbol{g}_\star, \mathbf{G}\rangle$ and $G := \|\mathbf{G}\|$, implies that $\boldsymbol{g}_\star \in \operatorname{argmin}\{\|\boldsymbol{g}\| \mid \boldsymbol{g} \in \mathcal{H}\}$.

Adding this to the fact that $\mathcal{H}$ is a hyperplane orthogonal to $\mathbf{G}$ implies that $\boldsymbol{g}_\star$ and $\mathbf{G}$ must be aligned, *i.e.* there exists a $c_\star \in \mathbb{R}$ such that $\boldsymbol{g}_\star = c_\star \mathbf{G}/\|\mathbf{G}\|$. Therefore, we have $\|\boldsymbol{g}_\star + \mathbf{G}\| = |g_\star + G|$, where

$$g_\star := \begin{cases} c_\star, & \text{if } G > 0; \\ \|\boldsymbol{g}_\star\|, & \text{otherwise.} \end{cases}$$

Further, note that $|g_\star| \leqslant h$. Thus, the RHS of (22) is bounded from above by

$$\begin{aligned} \Delta := & \frac{h_1^2}{\sqrt{V + g_\star^2}} \cdot \exp\left(\frac{(g_\star + G)^2}{2V + 2g_\star^2 + 2h|g_\star + G|}\right) \\ & - \left(1 + \frac{g_\star G \cdot (2V + 2h|G|)}{2(V + h|G|)^2}\right) \cdot \frac{h_1^2}{\sqrt{V}} \cdot \exp\left(\frac{G^2}{2V + 2h|G|}\right). \end{aligned} \tag{23}$$

Note that Δ in (23) can be written in terms of the function Θ in Lemma 18 as:

$$\Delta = \frac{h_1^2}{\sqrt{V}} \cdot \exp\left(\frac{G^2}{2V + 2hG}\right) \cdot \Theta\left(\frac{G}{h}, \frac{V}{h^2}, \frac{g_\star}{h}\right).$$

Since $(G/h, V/h^2, g_\star/h) \in \mathbb{R} \times \mathbb{R}_{>0} \times [-1, 1]$, Lemma 18 implies that $\Theta(G/h, V/h^2, g_\star/h) \leqslant 0$, and so due to (23), we also have $\Delta \leqslant 0$. Since Δ is an upper-bound on the RHS of (22), it follows that $\Phi_{t+1} - \Phi_t \leqslant 0$ as desired. ■

Appendix B. Proofs of Section 3.1

B.1. Proof of Theorem 6

The proof of Theorem 6 follows from the next theorem by setting $\epsilon = 1$. Theorem 20 essentially gives the regret bound of FREEGRAD if its outputs $(\widehat{\boldsymbol{w}}_t)$ are scaled by a constant $\epsilon > 0$. This will be useful to us later.

Theorem 20 *Let $\epsilon > 0$, and $\widehat{\boldsymbol{u}}_t := \widehat{\boldsymbol{w}}_t/\epsilon$, for $(\widehat{\boldsymbol{w}}_t)$ as in (9). Then, under Assumptions 1 and 2:*

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \widehat{\boldsymbol{u}}_t - \boldsymbol{w} \rangle \leqslant \left[2\|\boldsymbol{w}\|\sqrt{V_T \ln_+\left(\frac{2\epsilon\|\boldsymbol{w}\|V_T}{h_1^2}\right)}\right] \vee \left[4h_T\|\boldsymbol{w}\| \ln\left(\frac{4h_T\epsilon\|\boldsymbol{w}\|\sqrt{V_T}}{h_1^2}\right)\right] + \frac{h_1}{\epsilon},$$

for all $\boldsymbol{w} \in \mathcal{W} = \mathbb{R}^d, T \geqslant 1$.

Proof Since the assumptions of Theorem 5 are satisfied, we have

$$\Phi_T = \sum_{t=1}^{T} \widehat{\boldsymbol{w}}_t^\intercal \boldsymbol{g}_t + \frac{h_1^2}{\sqrt{V_T}} \cdot \exp\left(\frac{\|\mathbf{G}_T\|^2}{2V_T + 2h_T\|\mathbf{G}_T\|}\right) \leqslant \Phi_0 = h_1, \tag{24}$$

Dividing both sides of (24) by $\epsilon > 0$ and rearranging yields

$$\sum_{t=1}^{T} \widehat{\boldsymbol{u}}_t^\intercal \boldsymbol{g}_t \leqslant \frac{h_1}{\epsilon} - \Theta_T(\mathbf{G}_T), \quad \text{where } \Theta_T(\mathbf{G}) := \frac{h_1^2}{\epsilon\sqrt{V_T}} \cdot \exp\left(\frac{\|\mathbf{G}\|^2}{2V_T + 2h_T\|\mathbf{G}\|}\right), \mathbf{G} \in \mathbb{R}^d,$$

By duality, we further have that

$$\sum_{t=1}^{T} \widehat{\boldsymbol{u}}_t^{\intercal} \boldsymbol{g}_t \leqslant \frac{h_1}{\epsilon} + \boldsymbol{w}^{\intercal} \mathbf{G}_T + \Theta_T^{\star}(-\boldsymbol{w}), \quad \text{for all } \boldsymbol{w} \in \mathbb{R}^d. \tag{25}$$

Since $\Theta_T(\mathbf{G}) = \Psi_T(\mathbf{G})/\epsilon$, for all $\mathbf{G} \in \mathbb{R}^d$, where Ψ_T is the function defined in (10), we have by the properties of the Fenchel dual (Hiriart-Urruty and Lemaréchal, 2004, Prop. 1.3.1) that

$$\Theta_T^{\star}(\boldsymbol{w}) = \Psi_T^{\star}(\epsilon \boldsymbol{w})/\epsilon, \quad \text{for all } \boldsymbol{w} \in \mathbb{R}^d. \tag{26}$$

We now bound $\Psi_T^{\star}(-\boldsymbol{w})$ from above, for $\boldsymbol{w} \in \mathbb{R}^d$. For this, note that $\Psi_T(\mathbf{G}) = \psi_T(\|\mathbf{G}\|/h_T)$, for $\mathbf{G} \in \mathbb{R}^d$, where

$$\psi_T(x) := \frac{h_1^2}{\sqrt{V_T}} \cdot \exp\left(\frac{x^2}{2V_T/h_T^2 + 2|x|}\right).$$

Thus, according to (McMahan and Orabona, 2014, Lemma 3) and the properties of duality (Hiriart-Urruty and Lemaréchal, 2004, Prop. E.1.3.1), we have

$$\Psi_T^{\star}(-\boldsymbol{w}) = \psi_T^{\star}(h_T \|\boldsymbol{w}\|). \tag{27}$$

On the other hand, (Cutkosky and Orabona, 2018, Lemma 18, 19) and (Orabona and Pál, 2016a, Lemma 18) provides the following upper-bound on $\psi_T^{\star}(u)$, $u \in \mathbb{R}$, using the Lambert function W (where $W(x)$ is defined as the principal solution to $W(x)e^{W(x)} = x$):

$$\psi_T^{\star}(u) \leqslant \Lambda_T(u) \vee \left(4u \cdot \ln\left(\frac{4u\sqrt{V_T}}{h_1^2}\right)\right), \tag{28}$$

$$\text{where} \quad \Lambda_T(y) := y\sqrt{\frac{2V_T}{h_T}} \cdot \left(\left(W\left(c_T^2 y^2\right)\right)^{1/2} - \left(W\left(c_T^2 y^2\right)\right)^{-1/2}\right), y \in \mathbb{R},$$

and $c_T := \sqrt{2}V_T/(h_T h_1^2)$. Using the fact that the Lambert function satisfies $(W(x))^{1/2} - (W(x))^{-1/2} \leqslant \sqrt{\ln_+ x}$, for all $x \geqslant 0$ (see Lemma 22), together with (28) and (27) implies that

$$\Psi_T^{\star}(-\boldsymbol{w}) \leqslant \left[2\|\boldsymbol{w}\|\sqrt{V_T \ln_+\left(\frac{2\|\boldsymbol{w}\|V_T}{h_1^2}\right)}\right] \vee \left[4h_T\|\boldsymbol{w}\| \ln\left(\frac{4h_T\|\boldsymbol{w}\|\sqrt{V_T}}{h_1^2}\right)\right],$$

for all $\boldsymbol{w} \in \mathbb{R}^d$. Combining this with (25) and (26) leads to the desired regret bound. ■

Proof of Theorem 6. Invoke Theorem 20 with $\epsilon = 1$. ■

B.2. Proof of Theorem 7

Proof Fix $\boldsymbol{w} \in \mathbb{R}^d$ and let $k \geqslant 1$ be the total number of epochs. We denote by $\tau_i \geqslant 1$ the start index of epoch $i \in [k]$. Further, for $\tau, \tau' \in \mathbb{N}$, we define $\tilde{\tau} := \tau - 1$ and $V_{\tau:\tau'} := h_\tau^2 + \sum_{s=\tau}^{\tilde{\tau}'} \|\boldsymbol{g}_s\|^2$ (note

how the upper index is exclusive). Recall that at epoch $i \in [k]$, the restart condition in Algorithm 1 is triggered at $t = \tau_{i+1} > \tau_i$ only if

$$\frac{h_t}{h_{\tau_i}} > \sum_{s=1}^{t-1} \frac{\|\boldsymbol{g}_s\|}{h_s} + 2 \geqslant \sum_{s=1}^{t} \frac{\|\boldsymbol{g}_s\|}{h_s}, \tag{29}$$

where the last inequality follows by Assumption 1. We note that (29) also implies that

$$h_{\tau_{i+1}} > 2h_{\tau_i}, \quad \text{for all } i \in [k]. \tag{30}$$

On the other hand, within epoch $i \in [k]$, $\frac{h_t}{h_{\tau_i}} \leqslant \sum_{s=1}^{t-1} \frac{\|\boldsymbol{g}_s\|}{h_s} + 2$, for all $\tau_i \leqslant t \leqslant \tilde{\tau}_{i+1}$, and thus

$$\frac{h_{\tilde{\tau}_{i+1}}}{h_{\tau_i}} \leqslant \frac{\sqrt{V_{\tau_i:\tau_{i+1}}}}{h_{\tau_i}} \leqslant \sqrt{\sum_{t=\tau_i}^{\tilde{\tau}_{i+1}} \left(\sum_{s=1}^{t-1} \frac{\|\boldsymbol{g}_s\|}{h_s} + 2 \right)^2} \leqslant \sqrt{\frac{b_T}{2}}, \tag{31}$$

where $b_T := 2\sum_{t=1}^{T}(\sum_{s=1}^{t-1} \|\boldsymbol{g}_s\|/h_s + 2)^2$. Therefore, by the regret bound of Theorem 6 and (31):

$$\sum_{s=\tau_i}^{\tilde{\tau}_{i+1}} \langle \boldsymbol{g}_s, \widehat{\boldsymbol{w}}_s - \boldsymbol{w} \rangle \leqslant 2\|\boldsymbol{w}\| \sqrt{V_{\tau_i:\tau_{i+1}} \ln_+ (\|\boldsymbol{w}\| b_T)} + (4\|\boldsymbol{w}\| \ln(2\|\boldsymbol{w}\| b_T) + 1) h_{\tau_{i+1}}, \quad i \in [k]. \tag{32}$$

Summing this inequality over $i = 1, \ldots, k-2$, we get:

$$\sum_{s=1}^{\tilde{\tau}_{k-1}} \langle \boldsymbol{g}_s, \widehat{\boldsymbol{w}}_s - \boldsymbol{w} \rangle \leqslant 2\|\boldsymbol{w}\| \sum_{i=1}^{k-2} \sqrt{V_{\tau_i:\tau_{i+1}} \ln_+ (\|\boldsymbol{w}\| b_T)} + \sum_{i=1}^{k-2} (4\|\boldsymbol{w}\| \ln(2\|\boldsymbol{w}\| b_T) + 1) h_{\tau_{i+1}}. \tag{33}$$

Now using (29) at $t = \tau_{i+2}$, we have for all $i \in [k]$,

$$\begin{aligned} V_{\tau_i:\tau_{i+1}} \leqslant h_{\tau_i}^2 + \sum_{s=1}^{\tilde{\tau}_{i+1}} \frac{\|\boldsymbol{g}_s\|^2}{h_{\tau_{i+1}}^2} \cdot h_{\tau_{i+1}}^2 &\leqslant h_{\tau_i}^2 + \sum_{s=1}^{\tilde{\tau}_{i+2}} \frac{\|\boldsymbol{g}_s\|^2}{h_s^2} \cdot h_{\tau_{i+1}}^2, \\ &\overset{(*)}{\leqslant} h_{\tau_i}^2 + \sum_{s=1}^{\tilde{\tau}_{i+2}} \frac{\|\boldsymbol{g}_s\|}{h_s} \cdot h_{\tau_{i+1}}^2, \\ &\overset{(29)}{\leqslant} h_{\tau_i}^2 + h_{\tau_{i+1}} h_{\tau_{i+2}} \leqslant 2h_{\tau_{i+2}}^2, \end{aligned} \tag{34}$$

where the inequality $(*)$ follows by Assumption 1. Now by (30), we also have

$$\sum_{i=1}^{k-2} h_{\tau_{i+1}} \leqslant \left(\sum_{i=1}^{k} \frac{1}{2^i} \right) h_{\tau_k} \leqslant h_{\tau_k}. \tag{35}$$

Thus, substituting (35) and (34) into (33), and using the fact that $h_{\tau_k} \leqslant h_T$, we get:

$$\begin{aligned} \sum_{s=1}^{\tilde{\tau}_{k-1}} \langle \boldsymbol{g}_s, \widehat{\boldsymbol{w}}_s - \boldsymbol{w} \rangle &\leqslant 4\|\boldsymbol{w}\| h_T \sqrt{2 \ln_+ (\|\boldsymbol{w}\| b_T)} + h_T \cdot (4\|\boldsymbol{w}\| \ln(2\|\boldsymbol{w}\| b_T) + 1), \\ &\leqslant h_T \cdot (8\|\boldsymbol{w}\| \ln_+(2\|\boldsymbol{w}\| b_T) + 2\|\boldsymbol{w}\| + 1), \end{aligned} \tag{36}$$

where in the last inequality, we used the fact that $\sqrt{2x} \leqslant x + 1/2$, for all $x \geqslant 0$. Now, summing (32) over the last two epochs, yields

$$\sum_{s=\tau_{k-1}}^{T} \langle \boldsymbol{g}_s, \widehat{\boldsymbol{w}}_s - \boldsymbol{w} \rangle \leqslant 2\|\boldsymbol{w}\|\sqrt{2V_T \ln_+ (\|\boldsymbol{w}\| b_T)} + 2h_T \cdot (4\|\boldsymbol{w}\| \ln_+(2\|\boldsymbol{w}\| b_T) + 1). \quad (37)$$

Adding (36) and (37) together leads to

$$\sum_{s=1}^{T} \langle \boldsymbol{g}_s, \widehat{\boldsymbol{w}}_s - \boldsymbol{w} \rangle \leqslant 2\|\boldsymbol{w}\|\sqrt{2V_T \ln_+ (\|\boldsymbol{w}\| b_T)} + (16\|\boldsymbol{w}\| \ln_+(2\|\boldsymbol{w}\| b_T) + 2\|\boldsymbol{w}\| + 3) h_T.$$

This concludes the proof. ■

Appendix C. Proofs for Section 3.2

In this section we work on a version of the potential function that does not have the tuning for Section 3.2 substituted in yet, so that we can prove the result necessary for Section 5 in one go. The potential is parameterized by a *prior variance* $\sigma^2 > 0$, *initial wealth* $\epsilon > 0$ and, as before *prod factor* $\gamma > 1$. It is defined by

$$\Psi(\mathbf{G}, \boldsymbol{V}, h) := \frac{\epsilon \exp\left(\inf_{\lambda \geqslant 0} \left\{ \frac{1}{2} \mathbf{G}^\intercal \left(\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V} + \lambda \boldsymbol{I}\right)^{-1} \mathbf{G} + \frac{\lambda \rho(\gamma)^2}{2h^2} \right\}\right)}{\sqrt{\det\left(\boldsymbol{I} + \sigma^2 \gamma \boldsymbol{V}\right)}}, \quad (38)$$

C.1. Proof of Lemma 9

We prove the claim in Lemma 9 for the more general potential (38). Let $\lambda_\star \geqslant 0$ be the minimizer in the problem $\Psi(\mathbf{G}_{t-1}, \boldsymbol{V}_{t-1}, h_t)$. With that notation, we see that $\widehat{\boldsymbol{w}}_t = -\Psi(\mathbf{G}_{t-1}, \boldsymbol{V}_{t-1}, h_t) \cdot \left(\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V} + \lambda_\star \boldsymbol{I}\right)^{-1} \mathbf{G}_{t-1}$. To prove the lemma, it suffices to prove the stronger statement obtained by picking the sub-optimal choice $\lambda = \lambda_\star$ for the problem $\Psi(\mathbf{G}_t, \boldsymbol{V}_t, h_t)$, and dividing by $\Psi(\mathbf{G}_{t-1}, \boldsymbol{V}_{t-1}, h_t) > 0$, *i.e.*

$$\begin{aligned} &- \boldsymbol{g}_t \cdot \left(\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V}_{t-1} + \lambda_\star \boldsymbol{I}\right)^{-1} \mathbf{G}_{t-1} \\ &\leqslant 1 - \frac{\exp\left(\frac{1}{2}\mathbf{G}_t^\intercal \left(\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V}_t + \lambda_\star \boldsymbol{I}\right)^{-1} \mathbf{G}_t + \frac{\lambda_\star \rho(\gamma)^2}{2h_t^2} - \frac{1}{2}\ln\det\left(\boldsymbol{I} + \sigma^2\gamma \boldsymbol{V}_t\right)\right)}{\exp\left(\frac{1}{2}\mathbf{G}_{t-1}^\intercal \left(\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V}_{t-1} + \lambda_\star \boldsymbol{I}\right)^{-1} \mathbf{G}_{t-1} + \frac{\lambda_\star \rho(\gamma)^2}{2h_t^2} - \frac{1}{2}\ln\det\left(\boldsymbol{I} + \sigma^2\gamma \boldsymbol{V}_{t-1}\right)\right)}. \end{aligned}$$

Let us abbreviate $\boldsymbol{\Sigma}^{-1} = \sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V}_{t-1} + \lambda_\star \boldsymbol{I}$. The matrix determinant lemma and monotonicity of matrix inverse give

$$\ln \frac{\det\left(\boldsymbol{I} + \sigma^2\gamma \boldsymbol{V}_t\right)}{\det\left(\boldsymbol{I} + \sigma^2\gamma \boldsymbol{V}_{t-1}\right)} = \ln\left(1 + \gamma \boldsymbol{g}_t^\intercal \left(\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V}_{t-1}\right)^{-1} \boldsymbol{g}_t\right) \geqslant \ln\left(1 + \gamma \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma} \boldsymbol{g}_t\right).$$

Then Sherman-Morrison gives

$$\mathbf{G}_t^\intercal \left(\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V}_t + \lambda_\star \boldsymbol{I}\right)^{-1} \mathbf{G}_t = \mathbf{G}_t^\intercal \boldsymbol{\Sigma} \mathbf{G}_t - \gamma \frac{(\boldsymbol{g}_t^\intercal \boldsymbol{\Sigma} \mathbf{G}_t)^2}{1 + \gamma \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma} \boldsymbol{g}_t}$$

and splitting off the last round $\mathbf{G}_t = \mathbf{G}_{t-1} + \boldsymbol{g}_t$ gives

$$\mathbf{G}_t^\intercal \left(\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V}_t + \lambda_\star \boldsymbol{I}\right)^{-1} \mathbf{G}_t \;=\; \mathbf{G}_{t-1}^\intercal \boldsymbol{\Sigma} \mathbf{G}_{t-1} + \frac{2\mathbf{G}_{t-1}^\intercal \boldsymbol{\Sigma}\boldsymbol{g}_t + \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\boldsymbol{g}_t - \gamma(\boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\mathbf{G}_{t-1})^2}{1+\gamma \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\boldsymbol{g}_t}.$$

All in all, it suffices to show

$$-\boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\mathbf{G}_{t-1} \;\leqslant\; 1 - \exp\left(\frac{2\mathbf{G}_{t-1}^\intercal \boldsymbol{\Sigma}\boldsymbol{g}_t + \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\boldsymbol{g}_t - \gamma(\boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\mathbf{G}_{t-1})^2}{2(1+\gamma \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\boldsymbol{g}_t)} - \frac{1}{2}\ln\left(1+\gamma \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\boldsymbol{g}_t\right)\right).$$

Introducing scalars $r = \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\mathbf{G}_{t-1}$ and $z = \boldsymbol{g}_t^\intercal \boldsymbol{\Sigma}\boldsymbol{g}_t$, this simplifies to

$$-r \;\leqslant\; 1 - \exp\left(\frac{2r + z - \gamma r^2}{2(1+\gamma z)} - \frac{1}{2}\ln\left(1+\gamma z\right)\right)$$

Being a square, $z \geqslant 0$ is positive. In addition, optimality of $\lambda_\star$ ensures that $\|\boldsymbol{\Sigma}\mathbf{G}_{t-1}\| = \frac{\rho(\gamma)}{h_t}$; this follows from the fact that $\frac{\mathrm{d}}{\mathrm{d}\lambda}\, \mathbf{G}_{t-1}^\intercal (\sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V} + \lambda \boldsymbol{I})^{-1}\mathbf{G}_{t-1}\big|_{\lambda=\lambda_\star} = \|\boldsymbol{\Sigma}\mathbf{G}_{t-1}\|^2$. In combination with $\|\boldsymbol{g}_t\| \leqslant h_t$, we find $|r| \leqslant \rho(\gamma) \leqslant 1$. The above requirement may hence be further reorganized to

$$2r - \gamma r^2 \;\leqslant\; -z + (1+\gamma z)\left(\ln\left(1+\gamma z\right) + 2\ln(1+r)\right).$$

The convex right hand side is minimized subject to $z \geqslant 0$ at

$$z \;=\; \max\left\{0, \frac{e^{\frac{1}{\gamma}-1-2\ln(1+r)} - 1}{\gamma}\right\}$$

so it remains to show

$$2r - \gamma r^2 \;\leqslant\; \begin{cases} \frac{1}{\gamma} - (1+r)^{-2} e^{\frac{1}{\gamma}-1}, & \text{if } \frac{1}{\gamma} - 1 \geqslant 2\ln(1+r); \\ 2\ln(1+r), & \text{otherwise.} \end{cases}$$

The function ρ in (12) is designed to satisfy the hardest case, where $r = -\rho(\gamma)$, with equality.

C.2. Proof of Theorem 10

We restate the claim for the potential (38) before tuning:

Theorem 21 (Theorem 10 rephrased) *Let $\boldsymbol{\Sigma}_T^{-1} := \sigma^{-2}\boldsymbol{I} + \gamma \boldsymbol{V}_T$. For $(\widehat{\boldsymbol{w}}_t)$ as in (14), we have*

$$\sum_{t=1}^{T} \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{w}, \boldsymbol{g}_t \rangle \;\leqslant\; \epsilon + \sqrt{Q_T^{\boldsymbol{w}} \ln_+\left(\frac{\det\left(\sigma^2 \boldsymbol{\Sigma}_T^{-1}\right)}{\epsilon^2} Q_T^{\boldsymbol{w}}\right)}, \quad \textit{for all } \boldsymbol{w} \in \mathbb{R}^d, \textit{ where}$$

$$Q_T^{\boldsymbol{w}} := \max\left\{\boldsymbol{w}^\intercal \boldsymbol{\Sigma}_T^{-1}\boldsymbol{w}, \frac{1}{2}\left(\frac{h_T^2\|\boldsymbol{w}\|^2}{\rho(\gamma)^2}\ln\left(\frac{\det\left(\sigma^2\boldsymbol{\Sigma}_T^{-1}\right)}{\epsilon^2}\frac{h_T^2\|\boldsymbol{w}\|^2}{\rho(\gamma)^2}\right) + \boldsymbol{w}^\intercal \boldsymbol{\Sigma}_T^{-1}\boldsymbol{w}\right)\right\}.$$

Using that $\Psi(\mathbf{G}, \boldsymbol{V}, h)$ is decreasing in h, we can telescope to obtain

$$\sum_{t=1}^{T} \boldsymbol{g}_t^\intercal \widehat{\boldsymbol{w}}_t \;\leqslant\; \Psi(\mathbf{0}, \mathbf{0}, h_1) - \Psi(\mathbf{G}_T, \boldsymbol{V}_T, h_T)$$

Using the definition reveals $\Psi(\mathbf{0}, \mathbf{0}, h_1) = \epsilon$, yielding

$$\sum_{t=1}^{T} \boldsymbol{g}_t^\intercal \widehat{\boldsymbol{w}}_t \;\leqslant\; \epsilon - \frac{\epsilon \exp\left(\inf_{\lambda \geqslant 0} \; \frac{1}{2} \mathbf{G}_T^\intercal \left(\boldsymbol{\Sigma}_T^{-1} + \lambda \boldsymbol{I}\right)^{-1} \mathbf{G}_T + \frac{\lambda \rho(\gamma)^2}{2 h_T^2}\right)}{\sqrt{\det\left(\sigma^2 \boldsymbol{\Sigma}_T^{-1}\right)}}. \tag{39}$$

To transform this into a regret bound, it remains to compute the convex conjugate of the RHS of (39) in $\mathbf{G}_T$. To this end, let

$$f(\mathbf{G}) \;=\; \exp\left(\inf_{\lambda \geqslant 0} \; \frac{1}{2} \mathbf{G}^\intercal \left(\boldsymbol{Q} + \lambda \boldsymbol{I}\right)^{-1} \mathbf{G} + \frac{\lambda Z}{2}\right).$$

The Fenchel dual of this function is

$$\begin{aligned}
f^\star(\boldsymbol{u}) \;&=\; \sup_{\mathbf{G}} \; \boldsymbol{w}^\intercal \mathbf{G} - \exp\left(\inf_{\lambda \geqslant 0} \; \frac{1}{2} \mathbf{G}^\intercal \left(\boldsymbol{Q} + \lambda \boldsymbol{I}\right)^{-1} \mathbf{G} + \frac{\lambda Z}{2}\right) \\
&=\; \sup_{\mathbf{G}, \lambda \geqslant 0} \; \boldsymbol{w}^\intercal \mathbf{G} - \exp\left(\frac{1}{2} \mathbf{G}^\intercal \left(\boldsymbol{Q} + \lambda \boldsymbol{I}\right)^{-1} \mathbf{G} + \frac{\lambda Z}{2}\right) \\
&=\; \sup_{\alpha, \lambda \geqslant 0} \; \alpha \boldsymbol{w}^\intercal \left(\boldsymbol{Q} + \lambda \boldsymbol{I}\right) \boldsymbol{w} - \exp\left(\frac{\alpha^2}{2} \boldsymbol{w}^\intercal \left(\boldsymbol{Q} + \lambda \boldsymbol{I}\right) \boldsymbol{w} + \frac{\lambda Z}{2}\right) \\
&=\; \sup_{\lambda \geqslant 0} \; \sqrt{\boldsymbol{w}^\intercal \left(\boldsymbol{Q} + \lambda \boldsymbol{I}\right) \boldsymbol{w}} X\left(\boldsymbol{w}^\intercal \left(\boldsymbol{Q} + \lambda \boldsymbol{I}\right) \boldsymbol{w} e^{-\lambda Z}\right),
\end{aligned}$$

where the model complexity is measured for $\theta \geqslant 0$ through the function $X(\theta) := \sup_\alpha \; \alpha - e^{\frac{\alpha^2}{2} - \frac{1}{2} \ln \theta}$. One can write $X(\theta) = W(\theta)^{1/2} - W(\theta)^{-1/2}$ in terms of the Lambert function W (where $W(x)$ is defined as the principal solution to $W(x) e^{W(x)} = x$). We will further use that $X(\theta)$ is increasing, and that it satisfies $X(\theta) \leqslant \sqrt{\ln_+ \theta}$ (see Lemma 22). Zero derivative of the above objective for λ occurs at the pleasantly explicit

$$\lambda \;=\; \frac{\ln \frac{\|\boldsymbol{w}\|^2}{Z}}{2Z} - \frac{\boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}}{2\|\boldsymbol{w}\|^2},$$

and hence the optimum for λ is either at that point or at zero, whichever is higher, with the crossover point at $\frac{\|\boldsymbol{w}\|^2}{Z} \ln \frac{\|\boldsymbol{w}\|^2}{Z} = \boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}$. Plugging that in, we find that

$$f^\star(\boldsymbol{w}) = \begin{cases} \sqrt{\frac{1}{2}\left(C + \boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}\right)} X\left(\frac{1}{2}\left(C + \boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}\right) e^{-\frac{\ln \frac{\|\boldsymbol{w}\|^2}{Z}}{2} + \frac{Z \boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}}{2\|\boldsymbol{w}\|^2}}\right), & \text{if } C \geqslant \boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}; \\ \sqrt{\boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}} X\left(\boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}\right), & \text{otherwise,} \end{cases}$$

where $C := \frac{\|\boldsymbol{w}\|^2}{Z} \ln \frac{\|\boldsymbol{w}\|^2}{Z}$. Using that $X(\theta)$ is increasing, we may drop the exponential in its argument in the first case, and obtain

$$f^\star(\boldsymbol{w}) \;\leqslant\; \sqrt{Q_T^{\boldsymbol{w}}} X(Q_T^{\boldsymbol{w}}) \quad \text{where} \quad Q_T^{\boldsymbol{w}} \;:=\; \max\left\{\boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}, \frac{1}{2}\left(\frac{\|\boldsymbol{w}\|^2}{Z} \ln \frac{\|\boldsymbol{w}\|^2}{Z} + \boldsymbol{w}^\intercal \boldsymbol{Q} \boldsymbol{w}\right)\right\}.$$

Note that this is a curious maximum between $\boldsymbol{w}^\intercal \boldsymbol{Q}\boldsymbol{w}$ (the larger for modest $\boldsymbol{w}$), and the *average* between that very same term and another quantity that grows super-linearly with $\|\boldsymbol{w}\|^2$ (so this is the winner for extreme $\boldsymbol{w}$).

Okay, now let's collect everything for the final result and undo the abbreviations. We have

$$\begin{aligned}\sum_{t=1}^{T} \boldsymbol{g}_t^\intercal \widehat{\boldsymbol{w}}_t &\leqslant \epsilon + \inf_{\boldsymbol{w}} \ \boldsymbol{w}^\intercal \mathbf{G}_T + \frac{\epsilon}{\sqrt{\det\left(\sigma^2 \boldsymbol{\Sigma}_T^{-1}\right)}} f^\star\left(-\frac{\sqrt{\det\left(\sigma^2 \boldsymbol{\Sigma}_T^{-1}\right)}}{\epsilon}\boldsymbol{w}\right), \\ &\leqslant \epsilon + \inf_{\boldsymbol{w}} \ \boldsymbol{w}^\intercal \mathbf{G}_T + \sqrt{Q_T^{\boldsymbol{w}}} X\left(\frac{\det\left(\sigma^2 \boldsymbol{\Sigma}_T^{-1}\right)}{\epsilon^2} Q_T^{\boldsymbol{w}}\right),\end{aligned}$$

where

$$Q_T^{\boldsymbol{w}} := \max\left\{\boldsymbol{w}^\intercal \boldsymbol{\Sigma}_T^{-1}\boldsymbol{w}, \frac{1}{2}\left(\frac{h_T^2\|\boldsymbol{w}\|^2}{\rho(\gamma)^2}\ln\left(\frac{\det\left(\sigma^2\boldsymbol{\Sigma}_T^{-1}\right)}{\epsilon^2}\frac{h_T^2\|\boldsymbol{w}\|^2}{\rho(\gamma)^2}\right) + \boldsymbol{w}^\intercal \boldsymbol{\Sigma}_T^{-1}\boldsymbol{w}\right)\right\}.$$

To complete the proof of Theorem 21, it remains to prove the following result.

Lemma 22 *For $\theta \geqslant 0$, define $X(\theta) := \sup_\alpha \ \alpha - e^{\frac{\alpha^2}{2} - \frac{1}{2}\ln\theta}$. Then $X(\theta) = (W(\theta))^{1/2} - (W(\theta))^{-1/2} = \sqrt{\ln\theta} + o(1)$.*

Proof The fact that $X(\theta) = (W(\theta))^{1/2} - (W(\theta))^{-1/2}$ follows from (Orabona and Pál, 2016a, Lemma 18). Recall that

$$\sup_x \ yx - e^x \ = \ y\ln y - y$$

Hence

$$\begin{aligned}X(\theta) &= \sup_\alpha \ \alpha - e^{\frac{\alpha^2}{2} - \frac{1}{2}\ln\theta} \\ &= \sup_\alpha \inf_\eta \ \alpha - \eta\left(\frac{\alpha^2}{2} - \frac{1}{2}\ln\theta\right) + \eta\ln\eta - \eta \\ &= \inf_\eta \ \frac{1}{2\eta} + \frac{\eta}{2}\ln\theta + \eta\ln\eta - \eta \\ &\leqslant \min\left\{\sqrt{\ln\theta} - \frac{1 + \frac{1}{2}\ln\ln\theta}{\sqrt{\ln\theta}}, \frac{\sqrt{\theta}}{2} - \frac{1}{\sqrt{\theta}}\right\} \\ &\leqslant \sqrt{\ln_+\theta}\end{aligned}$$

where we plugged in the sub-optimal choices $\eta = \frac{1}{\sqrt{\ln\theta}}$ (this requires $\theta \geqslant 1$) and $\eta = \frac{1}{\sqrt{\theta}}$. When we stick in $\eta = \frac{1}{\sqrt{\ln(e^{e^{-2}}+\theta)}}$ we find

$$X(\theta) \leqslant \frac{\ln(e^{e^{-2}} + \theta) + \ln\theta - \ln\left(\ln(e^{e^{-2}} + \theta)\right) - 2}{2\sqrt{\ln(e^{e^{-2}} + \theta)}} \leqslant \sqrt{\ln(e^{e^{-2}} + \theta)}$$

Note that $e^{e^{-2}} = 1.14492$. This is less than 2, the value of θ where $\sqrt{\theta}/2 - 1/\sqrt{\theta}$ becomes positive. ■

Appendix D. Proofs for Section 4

D.1. Proof of Lemma 13

Let $c, b, \beta \geqslant 0$, $\nu \geqslant 1$, $\alpha \in]1, 2]$, and $\gamma \in]-1, -\alpha^{-1}[$. We consider the 1-dimensional case (*i.e.* $d = 1$) and set $g_t = t^\gamma$, for all $t \geqslant 1$. Since $-1 < \gamma < -1/\alpha$, we have $L_t = L_1 = 1$, for all $t \geqslant 1$, and so the sequence $(\sqrt{V_{\alpha,t} \ln(t)/L_t^\alpha})$ is increasing. Further, there exists $p, q > 0$ such that,

$$\forall t \geqslant 1, \quad p\sqrt{\ln t} \leqslant \sqrt{V_{\alpha,t} \ln(t)/L_t^\alpha} = \sqrt{\ln t \sum_{s=1}^{t} s^{\alpha\gamma}} \leqslant q\sqrt{\ln t}. \tag{40}$$

Thus, given any sequence $(\widehat{w}_t) \in \mathbb{R}$ satisfying

$$|\widehat{w}_t| \leqslant b\sqrt{V_{\alpha,t} \ln(t)/L_t^\alpha}, \;\; t \geqslant 1,$$

we have, for $T \geqslant 1$ and $w = -2b\sqrt{V_{\alpha,T} \ln(T)/L_T^\alpha}$,

$$\begin{aligned}\sum_{t=1}^{T} g_t \cdot (\widehat{w}_t - w) &\geqslant b\sqrt{V_{\alpha,T} \ln(T)/L_T^\alpha} \cdot \sum_{t=1}^{T} g_t, \\ &\overset{(40)}{\geqslant} bp\sqrt{\ln T} \cdot \sum_{t=1}^{T} t^\gamma, \\ &\geqslant \frac{bp\sqrt{\ln T}}{\gamma + 1} \cdot ((T+1)^{\gamma+1} - 1). \end{aligned} \tag{41}$$

Now by the choice of w and (40), we have $|w| \leqslant 2bq\sqrt{\ln T}$, and so by (41),

$$\sum_{t=1}^{T} g_t \cdot (\widehat{w}_t - w) \geqslant L_T |w|^\nu \cdot \frac{p \cdot ((T+1)^{\gamma+1} - 1)}{(\gamma+1)(2q)^\nu b^{\nu-1} (\ln T)^{\nu/2-1/2}}. \tag{42}$$

Using again the fact that $|w| \leqslant 2bq\sqrt{\ln T}$ and (40), we have $L_T^{1-\alpha/2}(|w| + 1)\sqrt{V_{\alpha,T} \ln T} \leqslant 2bq^2 \ln T + q\sqrt{\ln T}$, and so due to (41), we have

$$\sum_{t=1}^{T} g_t \cdot (\widehat{w}_t - w) \geqslant L_T^{1-\alpha/2}(|w| + 1)\sqrt{V_T \ln T} \cdot \frac{(T+1)^{\gamma+1} - 1}{(\gamma+1)\left(\frac{2q^2}{p}\sqrt{\ln T} + \frac{q}{bp}\right)}. \tag{43}$$

Since $\gamma > -1$, the exists $T \geqslant 1$ such that

$$2c \cdot \ln(1 + |w|T)^\beta \leqslant \min\left(\frac{(T+1)^{\gamma+1} - 1}{(\gamma+1)\left(\frac{2q^2}{p}\sqrt{\ln T} + \frac{q}{bp}\right)}, \; \frac{p \cdot ((T+1)^{\gamma+1} - 1)}{(\gamma+1)(2q)^\nu b^{\nu-1} (\ln T)^{\nu/2-1/2}}\right),$$

and so for such a choice of T, (42) and (43) imply the desired result.

D.2. Proof of Theorem 15

We need the following lemma in the proof of Theorem 15:

Lemma 23 *For all $b, c, \beta \geqslant 0$ and $\nu \in [1, 3[$, there exists $(\boldsymbol{g}_t) \in \mathbb{R}^d$, $T \geqslant 1$, and $\boldsymbol{w} \in \mathbb{R}^d$, such that for any sequence $(\widehat{\boldsymbol{w}}_t)$ satisfying $\|\widehat{\boldsymbol{w}}_t\| \leqslant b \cdot \sqrt{t \ln t}$, for all $t \geqslant 1$, we have*

$$\sum_{t=1}^{T} \langle \widehat{\boldsymbol{w}}_t - \boldsymbol{w}, \boldsymbol{g}_t \rangle \geqslant c \cdot \ln(1 + \|\boldsymbol{w}\| T)^{\beta} \cdot (L_T \|\boldsymbol{w}\|^{\nu} + L_T(\|\boldsymbol{w}\| + 1)\sqrt{T \ln T}).$$

Proof Let $c, b, \beta, \geqslant 0$, $\nu \in [1, 3[$, and $\alpha \in]1, 2]$. We consider the 1-dimensional case (*i.e.* $d = 1$) and set $g_t = 1$, for all $t \geqslant 1$. In this case, we have $L_t = 1$, for all $t \geqslant 1$. Given any sequence $(\widehat{w}_t) \in \mathbb{R}$ satisfying

$$|\widehat{w}_t| \leqslant b\sqrt{t \ln t}, \quad t \geqslant 1, \tag{44}$$

we have, for $T \geqslant 1$ and $w = -2b\sqrt{T \ln T}$,

$$\begin{aligned} \sum_{t=1}^{T} g_t \cdot (\widehat{w}_t - w) &\overset{(44)}{\geqslant} b\sqrt{T \ln T} \cdot \sum_{t=1}^{T} g_t, \\ &= b\sqrt{T \ln T} \cdot T. \end{aligned} \tag{45}$$

Now since $|w| = 2b\sqrt{T \ln T}$, we have, by (45),

$$\sum_{t=1}^{T} g_t \cdot (\widehat{w}_t - w) \geqslant L_T |w|^{\nu} \cdot \frac{T^{3/2-\nu/2}}{2^{\nu} b^{\nu-1} (\ln T)^{\nu/2-1/2}}. \tag{46}$$

Using again the fact that $|w| = 2b\sqrt{T \ln T}$ and $L_T = 1$, we have $L_T(|w|+1)\sqrt{T \ln T} = 2bT \ln T + \sqrt{T \ln T}$, and so due to (45),

$$\sum_{t=1}^{T} g_t \cdot (\widehat{w}_t - w) \geqslant L_T(|w| + 1)\sqrt{T \ln T} \cdot \frac{T}{2\sqrt{T \ln T} + 1/b}. \tag{47}$$

Since $\nu \in [1, 3[$, the exists $T \geqslant 1$ such that

$$2c \cdot \ln(1 + |w| T)^{\beta} \leqslant \min\left(\frac{T}{2\sqrt{T \ln T} + 1/b}, \ \frac{T^{3/2-\nu/2}}{2^{\nu} b^{\nu-1} (\ln T)^{\nu/2-1/2}} \right),$$

and so for such a choice of T, (46) and (47) imply the desired result. ■

Proof of Theorem 15. By Lemma 12, the only candidate algorithms are the ones whose outputs $(\widehat{\boldsymbol{w}}_t)$ satisfy $\|\widehat{\boldsymbol{w}}_t\| \leqslant b\sqrt{t \ln t}$, for all $t \geqslant 1$, for some constant $b > 0$. By Lemma 23, no such algorithms can achieve the desired regret bound. ■

Appendix E. Proof of Section 5

Proof of Theorem 16. The proof is similar to that of Theorem 7 expect for some changes to account for the fact that the modified FREERANGE wrapper scales the outputs of FREEGRAD.

First, let us review some notation. Let $k \geqslant 1$ be the total number of epochs and denote by $\tau_i \geqslant 1$ the start index of epoch $i \in [k]$. Further, for $\tau, \tau' \in \mathbb{N}$, we define $\tilde{\tau} := \tau - 1$, $V_{\tau:\tau'} := |x_\tau|^2 + \sum_{s=\tau}^{\tilde{\tau}'} |g_s|^2$, and $B_\tau := \sum_{s=1}^{\tau} |x_s|/h_s$. In what follows, let $w \in \mathbb{R}$ be fixed.

Let $(\widehat{u}_t)$ be the outputs of algorithm 2 and $i \in [k]$. In this case, we have $\widehat{u}_t = \widehat{w}_t/(h_{\tau_i} B_{\tau_i})$, for all $t \in \{\tau_i, \dots, \tilde{\tau}_{i+1}\}$, and by Theorem 20, with $d = 1$, $\epsilon = h_{\tau_i} B_{\tau_i}$, and $g_t \in x_t \cdot \partial^{(0,1)} \ell(y_t, x_t \widehat{u}_t)$:

$$\begin{aligned}
\sum_{t=\tau_i}^{\tilde{\tau}_{i+1}} g_t \cdot (\widehat{u}_t - w) &\leqslant 2|w| \sqrt{V_{\tau_i:\tau_{i+1}} \ln_+ \left(\frac{2 h_{\tau_i} B_{\tau_i} |w| V_{\tau_i:\tau_{i+1}}}{h_{\tau_i}^2} \right)} \\
&\quad + 4 h_{\tau_{i+1}} |w| \ln \left(\frac{4 h_{\tilde{\tau}_{i+1}} h_{\tau_i} B_{\tau_i} |w| \sqrt{V_{\tau_i:\tau_{i+1}}}}{h_{\tau_i}^2} \right) + \frac{h_{\tau_i}}{h_{\tau_i} B_{\tau_i}}, \\
&\leqslant 2|w| \sqrt{V_{\tau_i:\tau_{i+1}} \ln_+ \left(\frac{2 B_{\tau_i} |w| V_{\tau_i:\tau_{i+1}}}{h_{\tau_i}} \right)} \\
&\quad + 4 h_{\tau_{i+1}} |w| \ln \left(\frac{4 h_{\tilde{\tau}_{i+1}} B_{\tau_i} |w| \sqrt{V_{\tau_i:\tau_{i+1}}}}{h_{\tau_i}} \right) + \frac{1}{B_{\tau_i}},
\end{aligned} \tag{48}$$

Recall that at epoch $i \in [k]$, the restart condition in Algorithm 2 is triggered at $t = \tau_{i+1} \geqslant \tau_i$ only if

$$\frac{h_t}{h_{\tau_i}} \overset{(*)}{>} \sum_{s=1}^{t-1} \frac{|x_s|}{h_s} + 1 = \sum_{s=1}^{t} \frac{|x_s|}{h_s}, \tag{49}$$

where the equality follows by the fact that when $(*)$ is satisfied for the first time, it must hold that $|x_t| = h_t$ (recall that the hints (h_t) satisfy (17)); in fact, we have,

$$h_{\tau_i} = |x_{\tau_i}|, \quad \text{for all } i \in [k]. \tag{50}$$

From (49), we get that

$$\frac{h_{\tilde{\tau}_{i+1}}}{h_{\tau_i}} \leqslant \frac{\sqrt{V_{\tau_i:\tau_{i+1}}}}{h_{\tau_i}} \leqslant \sqrt{\sum_{t=\tau_i}^{\tilde{\tau}_{i+1}} \left(\sum_{s=1}^{t} \frac{|x_s|}{h_s} \right)^2} \leqslant \sqrt{\frac{b_T}{2}}, \tag{51}$$

where $b_T := 2 \sum_{t=1}^{T} (\sum_{s=1}^{t} |x_s|/h_s)^2$. Plugging (51) into (48), and letting $c_T := B_T^2 b_T$, we get:

$$\begin{aligned}
\sum_{t=\tau_i}^{\tilde{\tau}_{i+1}} g_t \cdot (\widehat{u}_t - w) &\leqslant 2|w| \sqrt{V_{\tau_i:\tau_{i+1}} \ln_+ (|w| \sqrt{2 V_T c_T})} \\
&\quad + 4 h_{\tau_{i+1}} |w| \ln \left(2|w| \sqrt{2 V_T c_T} \right) + \frac{1}{B_{\tau_i}},
\end{aligned} \tag{52}$$

Summing this inequality over $i = 1, \dots, k-2$, we get:

$$\begin{aligned}
\sum_{t=1}^{\tilde{\tau}_{k-1}} g_t \cdot (\widehat{u}_t - w) &\leqslant 2|w| \sqrt{k \sum_{i=1}^{k-2} V_{\tau_i:\tau_{i+1}} \ln_+ \left(|w|\sqrt{2V_T c_T}\right)} \\
&\quad + \sum_{i=1}^{k-2} 4h_{\tau_{i+1}}|w| \ln\left(2|w|\sqrt{2V_T c_T}\right) + \sum_{i=1}^{k-2} \frac{1}{B_{\tau_i}}, , \\
&\leqslant 2|w| \sqrt{k \cdot \left(\sum_{s=1}^{\tau_{k-1}} |x_s|^2 + \sum_{i=1}^{k-2} h_{\tau_i}^2\right) \ln_+ \left(|w|\sqrt{2V_T c_T}\right)} \\
&\quad + \sum_{i=1}^{k-2} 4h_{\tau_{i+1}}|w| \ln\left(2|w|\sqrt{2V_T c_T}\right) + \sum_{i=1}^{k-2} \frac{1}{B_{\tau_i}}, \\
&\overset{(50)}{\leqslant} 2|w| \sqrt{2k \sum_{s=1}^{\tau_{k-1}} |x_s|^2 \ln_+ \left(|w|\sqrt{2V_T c_T}\right)} \\
&\quad + \sum_{i=1}^{k-2} 4h_{\tau_{i+1}}|w| \ln\left(2|w|\sqrt{2V_T c_T}\right) + \sum_{i=1}^{k-2} \frac{1}{B_{\tau_i}}. \qquad (53)
\end{aligned}$$

Using (49) again, we get that

$$\begin{aligned}
\sum_{s=1}^{\tau_{k-1}} |x_s|^2 = \sum_{s=1}^{\tau_{k-1}} \frac{|x_s|^2}{h_{\tau_{k-1}}^2} \cdot h_{\tau_{k-1}}^2 \leqslant \sum_{s=1}^{\tau_k} \frac{|x_s|^2}{h_s^2} \cdot h_{\tau_{k-1}}^2 \leqslant \sum_{s=1}^{\tau_k} \frac{|x_s|}{h_s} \cdot h_{\tau_{k-1}}^2 &\overset{(*)}{\leqslant} \left(\sum_{s=1}^{\tau_k} \frac{|x_s|}{h_s}\right)^2 \cdot \frac{h_{\tau_{k-1}}^2}{k}, \\
&\overset{(49)}{\leqslant} \frac{h_{\tau_k}^2}{k}, \qquad (54)
\end{aligned}$$

where the inequality $(*)$ follows by the fact that $\sum_{s=1}^{\tau_k} |x_s|/h_s \geqslant k$ due to (50). We also have

$$\sum_{i=1}^{k-2} h_{\tau_{i+1}} \leqslant \sum_{s=1}^{\tau_{k-1}} |x_s| = \sum_{s=1}^{\tau_{k-1}} \frac{|x_s|}{h_{\tau_{k-1}}} \cdot h_{\tau_{k-1}} \leqslant \sum_{s=1}^{\tau_k} \frac{|x_s|}{h_s} \cdot h_{\tau_{k-1}} \overset{(49)}{\leqslant} h_{\tau_k}. \qquad (55)$$

$$\sum_{i=1}^{k-2} \frac{1}{B_{\tau_i}} \overset{(50)}{=} \sum_{i=1}^{k-2} \frac{\frac{|x_{\tau_i}|}{h_{\tau_i}}}{B_{\tau_i}} = \sum_{i=1}^{k-2} \frac{\frac{|x_{\tau_i}|}{h_{\tau_i}}}{\sum_{s=1}^{\tau_i} \frac{|x_s|}{h_s}} \leqslant \sum_{t=1}^{\tau_{k-2}} \frac{\frac{|x_t|}{h_t}}{\sum_{s=1}^{t} \frac{|x_s|}{h_s}} \leqslant \ln B_T. \qquad (56)$$

Thus, substituting (56), (55), and (54) into (53), and using the fact that $h_{\tau_k} \leqslant h_T$, we get:

$$\begin{aligned}
\sum_{t=1}^{\tilde{\tau}_{k-1}} g_t \cdot (\widehat{u}_t - w) &\leqslant 2|w| h_T \sqrt{2 \ln_+(|w|\sqrt{2V_T c_T})} \\
&\quad + 4h_T|w| \ln\left(2|w|\sqrt{2V_T c_T}\right) + \ln B_T, \\
&\leqslant h_T|w| \cdot (6 \ln_+(2|w|\sqrt{2V_T c_T}) + 1) + \ln B_T, \qquad (57)
\end{aligned}$$

where in the last inequality, we used the fact that $\sqrt{2x} \leqslant x + 1/2$, for all $x \geqslant 0$. Now, summing (52) over the last two epochs, yields

$$\sum_{t=\tau_{k-1}}^{T} g_t \cdot (\widehat{u}_t - w) \leqslant 2|w|\sqrt{2V_T \ln_+(|w|\sqrt{2V_T c_T})} + 8h_T|w| \ln(2|w|\sqrt{2V_T c_T}) + 2. \quad (58)$$

Adding (57) and (58) together implies the desired result. ■

Chapter 5

PAC-Bayes Unexpected Bernstein Inequality

We now switch to the statistical learning setting and consider a notion of adaptivity through data-dependent bounds. While it is not completely clear whether existing generalization bounds can ever explain the often-witnessed (good) generalization performance of deep NNs [Nagarajan and Kolter, 2019], some recently proposed ideas seem promising. One of them involves deriving data-dependent PAC-Bayesian bounds that can automatically adapt to the easiness of the problem at hand, leading to non-vacuous bounds in many cases. In this chapter, we derive a data-dependent bound whose main term is reminiscent of the empirical loss variance, but unlike the latter, it converges to zero when the number of samples increases. This can lead to a much tighter generalization bound compared with the state-of-the-art as we show. The key tools used to derive our bound (which are of independent interest) are **I)** a new concentration inequality reminiscent of the empirical Bernstein inequality [Maurer and Pontil, 2009], and **II)** the idea of splitting the dataset in two and learning an *informed prior* from each half before combining them to produce the final bound.

PAC-Bayes Un-Expected Bernstein Inequality

Zakaria Mhammedi
The Australian National University and Data61
zak.mhammedi@anu.edu.au

Peter D. Grünwald
CWI and Leiden University
pdg@cwi.nl

Benjamin Guedj
Inria and University College London
benjamin.guedj@inria.fr

Abstract

We present a new PAC-Bayesian generalization bound. Standard bounds contain a $\sqrt{L_n \cdot \mathrm{KL}/n}$ complexity term which dominates unless L_n, the empirical error of the learning algorithm's randomized predictions, vanishes. We manage to replace L_n by a term which vanishes in many more situations, essentially whenever the employed learning algorithm is sufficiently stable on the dataset at hand. Our new bound consistently beats state-of-the-art bounds both on a toy example and on UCI datasets (with large enough n). Theoretically, unlike existing bounds, our new bound can be expected to converge to 0 faster whenever a Bernstein/Tsybakov condition holds, thus connecting PAC-Bayesian generalization and *excess risk* bounds—for the latter it has long been known that faster convergence can be obtained under Bernstein conditions. Our main technical tool is a new concentration inequality which is like Bernstein's but with X^2 taken outside its expectation.

1 Introduction

PAC-Bayesian generalization bounds [1, 8, 9, 17, 18, 20, 28, 29, 30] have recently obtained renewed interest within the context of deep neural networks [14, 34, 42]. In particular, Zhou et al. [42] and Dziugaite and Roy [14] showed that, by extending an idea due to Langford and Caruana [23], one can obtain nontrivial (but still not very strong) generalization bounds on real-world datasets such as MNIST and ImageNet. Since using alternative methods, nontrivial generalization bounds are even harder to get, there remains a strong interest in improved PAC-Bayesian bounds. In this paper, we provide a considerably improved bound whenever the employed learning algorithm is sufficiently *stable* on the given data.

Most standard bounds have an order $\sqrt{L_n \cdot \mathrm{COMP}_n/n}$ term on the right, where COMP_n represents model complexity in the form of a Kullback-Leibler divergence between a prior and a posterior, and L_n is the *posterior expected loss* on the training sample. The latter only vanishes if there is a sufficiently large neighborhood around the "center" of the posterior at which the training error is 0. In the two papers [14, 42] mentioned above, this is not the case. For example, the various deep net experiments reported by Dziugaite et al. [14, Table 1] with $n = 150000$ all have L_n around 0.03, so that $\sqrt{\mathrm{COMP}_n/n}$ is multiplied by a non-negligible $\sqrt{0.03} \approx 0.17$. Furthermore, they have COMP_n increasing substantially with n, making $\sqrt{L_n \cdot \mathrm{COMP}_n/n}$ converge to 0 at rate slower than $1/\sqrt{n}$.

In this paper, we provide a bound (Theorem 3) with L_n replaced by a second-order term V_n—a term which will go to 0 in many cases in which L_n does not. This can be viewed as an extension of an earlier second-order approach by Tolstikhin and Seldin [39] (TS from now on); they also replace L_n, but by a term that, while usually smaller than L_n, will tend to be larger than our V_n. Specifically, as

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

they write, in classification settings (our primary interest), their replacement is not much smaller than L_n itself. Instead our V_n can be very close to 0 in classification even when L_n is large. While the TS bound is based on an "empirical" Bernstein inequality due to [27][1], our bound is based on a different modification of Bernstein's moment inequality in which the occurrence of X^2 is taken outside of its expectation (see Lemma 13). We note that an empirical Bernstein inequality was introduced in [4, Theorem 1], and the name "Empirical Bernstein" was coined in [32].

The term V_n in our bound goes to 0—and our bound improves on existing bounds—whenever the employed learning algorithm is relatively stable on the given data; for example, if the predictor learned on an initial segment (say, 50%) of the dataset performs similarly (*i.e.* assigns similar losses to the same samples) to the predictor based on the full data. This improvement is reflected in our experiments where, except for very small sample sizes, we consistently outperform existing bounds both on a toy classification problem with label noise and on standard UCI datasets [13]. Of course, the importance of stability for generalization has been recognized before in landmark papers such as [7, 33, 38], and recently also in the context of PAC-Bayes bounds [35]. However, the data-dependent stability notion "V_n" occurring in our bound seems very different from any of the notions discussed in those papers.

Theoretically, a further contribution is that we connect our PAC-Bayesian generalization bound to *excess risk bounds*; we show that (Theorem 7) our generalization bound can be of comparable size to excess risk bounds up to an irreducible *complexity-free* term that is independent of model complexity. The excess risk bound that can be attained for any given problem depends both on the complexity of the set of predictors $\mathcal{H}$ and on the inherent "easiness" of the problem. The latter is often measured in terms of the exponent $\beta \in [0,1]$ of the *Bernstein condition* that holds for the given problem [6, 15, 19], which generalizes the exponent in the celebrated *Tsybakov margin condition* [5, 40]. The larger β, the faster the excess risk converges. In Section 5, we essentially show that the rate at which the $\sqrt{V_n \cdot \mathrm{COMP}_n / n}$ term goes to 0 can also be bounded by a quantity that gets smaller as β gets larger. In contrast, previous PAC-Bayesian bounds do not have such a property.

Contents. In Section 2, we introduce the problem setting and provide a first, simplified version of our main theorem. Section 3 gives our main bound. Experiments are presented in Section 4, followed by theoretical motivation in Section 5. The proof of our main bound is provided in Section 6, where we first present the convenient ESI language for expressing stochastic inequalities, and (our main tool) the unexpected Bernstein lemma (Lemma 13). The paper ends with an outlook for future work.

2 Problem Setting, Background, and Simplified Version of Our Bound

Setting and Notation. Let $Z_1, \dots, Z_n$ be i.i.d. random variables in some set $\mathcal{Z}$, with $Z_1 \sim D$. Let $\mathcal{H}$ be a hypothesis set and $\ell : \mathcal{H} \times \mathcal{Z} \to [0, b]$, $b > 0$, be a bounded loss function such that $\ell_h(Z) \coloneqq \ell(h, Z)$ denotes the loss that hypothesis h makes on Z. We call any such tuple $(D, \ell, \mathcal{H})$ a *learning problem*. For a given hypothesis $h \in \mathcal{H}$, we denote its *risk* (expected loss on a test sample of size 1) by $L(h) \coloneqq \mathbb{E}_{Z\sim D}[\ell_h(Z)]$ and its empirical error by $L_n(h) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell_h(Z_i)$. For any distribution P on $\mathcal{H}$, we write $L(P) \coloneqq \mathbb{E}_{h\sim P}[L(h)]$ and $L_n(P) \coloneqq \mathbb{E}_{h\sim P}[L_n(h)]$.

For any $m \in [n]$ and any variables $Z_1, \dots, Z_n$ in $\mathcal{Z}$, we denote $Z_{\le m} \coloneqq (Z_1, \dots, Z_m)$ and $Z_{<m} \coloneqq Z_{\le m-1}$, with the convention that $Z_{\le 0} = \varnothing$. Similarly, we denote $Z_{\ge m} \coloneqq (Z_m, \dots, Z_n)$ and $Z_{>m} \coloneqq Z_{\ge m+1}$, with the convention that $Z_{\ge n+1} = \varnothing$. As is customary in PAC-Bayesian works, a *learning algorithm* is a (computable) function $P : \bigcup_{i=1}^n \mathcal{Z}^i \to \mathcal{P}(\mathcal{H})$ that, upon observing input $Z_{\le n} \in \mathcal{Z}^n$, outputs a "posterior" distribution $P(Z_{\le n})(\cdot)$ on $\mathcal{H}$. The posterior could be a Gibbs or a generalized-Bayesian posterior but also other algorithms. When no confusion can arise, we will abbreviate $P(Z_{\le n})$ to P_n, and denote P_0 any "prior" distribution, *i.e.* a distribution on $\mathcal{H}$ which has to be specified in advance, before seeing the data; we will use the convention $P(\varnothing) = P_0$. Finally, we denote the Kullback-Leibler divergence between P_n and P_0 by $\mathrm{KL}(P_n \| P_0)$.

Comparing Bounds. Both existing state-of-the-art PAC-Bayes bounds and ours essentially take the following form; there exists constants $\mathcal{P}, \mathcal{A}, \mathcal{C} \ge 0$, and a function $\varepsilon_{\delta,n}$, logarithmic in $1/\delta$ and n, such

[1] An alternative form of empirical Bernstein inequality appears in [41], based on an inequality due to [11].

that for all $\delta \in]0,1[$, with probability at least $1-\delta$ over the sample $Z_1,\dots,Z_n$, it holds that,

$$L(P_n) - L_n(P_n) \le \mathcal{P}\cdot\sqrt{\frac{R_n\cdot(\mathrm{COMP}_n+\varepsilon_{\delta,n})}{n}} + \mathcal{A}\cdot\frac{\mathrm{COMP}_n+\varepsilon_{\delta,n}}{n} + \mathcal{C}\cdot\sqrt{\frac{R'_n\cdot\varepsilon_{\delta,n}}{n}}, \quad (1)$$

where $R_n, R'_n \ge 0$ are sample-dependent quantities which may differ from one bound to another. Existing classical bounds that after slight relaxations take on this form are due to Langford and Seeger [24, 37], Catoni [10], Maurer [26], and Tolstikhin and Seldin (TS) [39] (see the latter for a nice overview). In all these cases, $\mathrm{COMP}_n = \mathrm{KL}(P_n\|P_0)$, $R'_n = 0$, and—except for the TS bound—$R_n = L_n(P_n)$. For the TS bound, R_n is equal to the empirical loss variance. Our bound in Theorem 3 also fits (1) (after a relaxation), but with considerably different choices for COMP_n, R'_n, and R_n.

Of special relevance in our experiments is the bound due to Maurer [26], which as noted by TS [39] tightens the PAC-Bayes-kl inequality due to Seeger [36], and is one of the tightest known generalization bounds in the literature. It can be stated as follows: for $\delta \in]0,1[$, $n \ge 8$, and any learning algorithm P, with probability at least $1-\delta$,

$$\mathrm{kl}(L(P_n), L_n(P_n)) \le \frac{\mathrm{KL}(P_n\|P_0) + \ln\frac{2\sqrt{n}}{\delta}}{n}, \quad (2)$$

where kl is the binary Kullback-Leibler divergence. Applying the inequality $p \le q + \sqrt{2q\,\mathrm{kl}(p\|q)} + 2\,\mathrm{kl}(p\|q)$ to (2) yields a bound of the form (1) (see [39] for more details). Note also that using Pinsker's inequality together with (2) implies McAllester's classical PAC-Bayesian bound [28].

We now present a simplified version of our bound in Theorem 3 below as a corollary.

Corollary 1. *For any $1 \le m < n$ and any deterministic estimator $\hat{h} : \bigcup_{i=1}^n \mathcal{Z}^i \to \mathcal{H}$ (such as* ERM*), there exists $\mathcal{P}, \mathcal{A}, \mathcal{C} > 0$, such that (1) holds with probability at least $1-\delta$, with*

$$\mathrm{COMP}_n = \mathrm{KL}(P_n\|P(Z_{\le m})) + \mathrm{KL}(P_n\|P(Z_{>m})),$$

$$R'_n \coloneqq V'_n \coloneqq \frac{1}{n}\sum_{i=1}^m \ell_{\hat{h}(Z_{>m})}(Z_i)^2 + \frac{1}{n}\sum_{j=m+1}^n \ell_{\hat{h}(Z_{\le m})}(Z_j)^2, \quad (3)$$

$$R_n \coloneqq V_n \coloneqq \frac{1}{n}\mathbb{E}_{h\sim P_n}\left[\sum_{i=1}^m\left(\ell_h(Z_i) - \ell_{\hat{h}(Z_{>m})}(Z_i)\right)^2 + \sum_{j=m+1}^n\left(\ell_h(Z_j) - \ell_{\hat{h}(Z_{\le m})}(Z_j)\right)^2\right]. \quad (4)$$

Like in TS's and Catoni's bound, but unlike McAllester's and Maurer's, our $\varepsilon_{\delta,n}$ grows as $(\ln\ln n)/\delta$. Another difference is that our complexity term is a sum of two KL divergences, in which the prior (in this case $P(Z_{\le m})$ or $P(Z_{>m})$) is "informed"—when $m = n/2$, it is really the posterior based on half the sample. Our experiments confirm that this tends to be much smaller than $\mathrm{KL}(P_n\|P_0)$. While the idea to use part of the sample to create an informed prior is due to [2], we are the first to combine all parts (halves) into a single bound, which requires a novel technique. This technique can be applied to other existing bounds as well (see Section 3).

A larger difference between our bound and others is in the fact that we have $R_n = V_n$ instead of the typical empirical error $R_n = L_n(P_n)$. Only TS [39] have a R_n that is somewhat reminiscent of ours; in their case $R_n = \mathbb{E}_{h\sim P_n}[\sum_{i=1}^n(\ell_h(Z_i) - L_n(h))^2]/(n-1)$ is the empirical loss variance. The crucial difference to our V_n is that the empirical loss variance cannot be close to 0 unless a sizeable P_n-posterior region of h has empirical error almost constant on most data instances. For classification with 0-1 loss, this is a strong condition since the empirical loss variance is equal to $nL_n(P_n)(1-L_n(P_n))/(n-1)$, which is only close to 0 if $L_n(P_n)$ is itself close to 0 or 1. In contrast, our V_n can go to zero 0 even if the empirical error and variance do not, as long as the learning algorithm is sufficiently stable. This can be witnessed in our experiments in Section 4. In Section 5, we argue more formally that under a Bernstein condition, the $\sqrt{V_n\cdot\mathrm{COMP}_n/n}$ term in our bound can be much smaller than $\sqrt{\mathrm{COMP}_n/n}$. Note, finally, that the term V_n has a two-fold cross-validation flavor, but in contrast to a cross-validation error, for V_n to be small, it is sufficient that the losses are *similar*, not that they are small.

The price we pay for having $R_n = V_n$ in our bound is the right-most, irreducible remainder term in (1) of order at most $b/\sqrt{n}$. Note, however, that this term is decoupled from the complexity COMP_n, and thus it is not affected by COMP_n growing with the "size" of $\mathcal{H}$. The following lemma gives a tighter bound (tighter than the $b/\sqrt{n}$ just mentioned) on the irreducible term:

Lemma 2. *Suppose that the loss is bounded by 1* (i.e. $b = 1$) *and that n is even, and let $m = n/2$. For $\delta \in]0,1[$, R'_n as in (3), and any estimator $\hat{h} : \bigcup_{i=1}^n \mathcal{Z}^i \to \mathcal{H}$, we have, with probability at least $1-\delta$,*

$$\sqrt{\frac{R'_n}{n}} \leq \sqrt{\frac{2(L(\hat{h}(Z_{>m})) + L(\hat{h}(Z_{\leq m})))}{n}} + \frac{4\sqrt{\ln \frac{4}{\delta}}}{n}. \tag{5}$$

Behind the proof of the lemma is an application of Hoeffding's and the empirical Bernstein inequality [27] (see Section C). Note that in the realizable setting, the first term on the RHS of (5) can be of order $O(1/n)$ with the right choice of estimator $\hat{h}$ (*e.g.* ERM). In this case (still in the realizable setting), our irreducible term would go to zero at the same rate as other bounds which have $R_n = L_n(P_n)$.

3 Main Bound

We now present our main result in its most general form. Let $\vartheta(\eta) \coloneqq (-\ln(1-\eta) - \eta)/\eta^2$ and $c_\eta \coloneqq \eta \cdot \vartheta(\eta b)$, for $\eta \in]0, 1/b[$, where $b > 0$ is an upper-bound on the loss ℓ.

Theorem 3. [Main Theorem] *Let $Z_1, \ldots, Z_n$ be i.i.d. with $Z_1 \sim D$. Let $m \in [0..n]$ and π be any distribution with support on a finite or countable grid $\mathcal{G} \subset]0, 1/b[$. For any $\delta \in]0,1[$, and any learning algorithms $P, Q : \bigcup_{i=1}^n \mathcal{Z}^i \to \mathcal{P}(\mathcal{H})$, we have,*

$$L(P_n) \leq L_n(P_n) + \inf_{\eta \in \mathcal{G}} \left\{ c_\eta \cdot V_n + \frac{\mathrm{COMP}_n + 2\ln \frac{1}{\delta \cdot \pi(\eta)}}{\eta \cdot n} \right\} + \inf_{\nu \in \mathcal{G}} \left\{ c_\nu \cdot V'_n + \frac{\ln \frac{1}{\delta \cdot \pi(\nu)}}{\nu \cdot n} \right\}, \tag{6}$$

with probability at least $1-\delta$, where COMP_n, V'_n, and V_n are the random variables defined by:

$$\mathrm{COMP}_n \coloneqq \mathrm{KL}(P_n \| P(Z_{\leq m})) + \mathrm{KL}(P_n \| P(Z_{>m})), \tag{7}$$

$$V'_n \coloneqq \frac{1}{n} \sum_{i=1}^m \mathbb{E}_{h \sim Q(Z_{>i})} \left[\ell_h(Z_i)^2 \right] + \frac{1}{n} \sum_{j=m+1}^n \mathbb{E}_{h \sim Q(Z_{<j})} \left[\ell_h(Z_j)^2 \right],$$

$$V_n \coloneqq \frac{1}{n} \mathbb{E}_{h \sim P_n} \left[\sum_{i=1}^m \left(\ell_h(Z_i) - \mathbb{E}_{h' \sim Q(Z_{>i})} \left[\ell_{h'}(Z_i) \right] \right)^2 + \sum_{j=m+1}^n \left(\ell_h(Z_j) - \mathbb{E}_{h' \sim Q(Z_{<j})} \left[\ell_{h'}(Z_j) \right] \right)^2 \right].$$

While the result holds for all $0 \leq m \leq n$, in the remainder of this paper, we assume for simplicity that n is even and that $m = n/2$. We will also be using the grid $\mathcal{G}$ and distribution π defined by

$$\mathcal{G} \coloneqq \left\{ \tfrac{1}{2b}, \ldots, \tfrac{1}{2^K b} : K \coloneqq \left\lceil \log_2 \left(\tfrac{1}{2} \sqrt{\tfrac{n}{\ln \frac{1}{\delta}}} \right) \right\rceil \right\}, \quad \text{and} \quad \pi \equiv \text{uniform distribution over } \mathcal{G}. \tag{8}$$

Roughly speaking, this choice of $\mathcal{G}$ ensures that the infima in η and ν in (6) are attained within $[\min \mathcal{G}, \max \mathcal{G}]$. Using the relaxation $c_\eta \leq \eta/2 + \eta^2 11b/20$, for $\eta \leq 1/(2b)$, in (6) and tuning η and ν within the grid $\mathcal{G}$ defined in (8) leads to a bound of the form (1). Furthermore, we see that the expression of V_n in Corollary 1 now follows when Q is chosen such that, for $1 \leq i \leq m < j \leq n$, $Q(Z_{>i}) \equiv \delta(\hat{h}(Z_{>m}))$ and $Q(Z_{<j}) \equiv \delta(\hat{h}(Z_{\leq m}))$, for some deterministic estimator $\hat{h}$, where $\delta(h)(\cdot)$ denotes the Dirac distribution at $h \in \mathcal{H}$.

Online Estimators. It is clear that Theorem 3 is considerably more general than its Corollary 1; when predicting the j-th point Z_j, $j > m$, in the RHS sum of V_n, we could use a posterior $Q(Z_{<j}) \equiv \delta(\hat{h}(Z_{<j}))$ which does not only depend on $Z_1, \ldots, Z_m$, but also on part of the second sample, namely $Z_{m+1}, \ldots, Z_{j-1}$, and analogously when predicting Z_i, $i \leq m$, in the LHS sum of V_n. We can thus base our bound on a sum of errors achieved by *online estimators* $(\hat{h}(Z_{<j}))$ and $(\hat{h}(Z_{>i}))$ which converge to the final $\hat{h}(Z_{\leq n})$ based on the full data. Doing this would likely improve our bounds, but we did not try it in our experiments since it is computationally demanding.

Informed Priors. Other bounds can also be modified to make use of "informed priors" from each half of the data; in this case, the $\mathrm{KL}(P_n \| P_0)$ term in these bounds can be replaced by COMP_n defined in (7). As revealed by additional experiments in the Appendix H, doing this substantially improves the corresponding bounds when the learning algorithm is sufficiently stable. Here we show how this can be done for Maurer's bound in (2) (the details for other bounds are postponed to Appendix A).

Lemma 4. *Let $\delta \in]0,1[$ and $m \in [0..n]$. In the setting of Theorem 3, we have, with probability at least $1-\delta$,*

$$\mathrm{kl}(L(P_n), L_n(P_n)) \leq \frac{\mathrm{KL}(P_n \| P(Z_{\leq m})) + \mathrm{KL}(P_n \| P(Z_{>m})) + \ln \frac{4\sqrt{m(n-m)}}{\delta}}{n}. \tag{9}$$

Remark 5. *(Useful for Section 5 below) Though this may deteriorate the bound in practice, Theorem 3 allows choosing a learning algorithm P such that for $1 \leq m < n$, $P(Z_{\leq m}) \equiv P(Z_{>m}) \equiv P_0$ (i.e. no informed priors); this results in $\mathrm{COMP}_n = 2\mathrm{KL}(P_n \| P_0)$—the bound is otherwise unchanged.*

Biasing. The term V_n in our bound can be seen as the result of "biasing" the loss when evaluating the generalization error on each half of the sample. The TS bound, having a second order variance term, can be used in a way as to arrive at a bound like ours with the same V_n as in Corollary 1. The idea here is to apply the TS bound twice (once on each half of the sample) to the biased losses $\ell(h,\cdot) - \ell(\hat{h}(Z_{\leq m}),\cdot)$ and $\ell(h,\cdot) - \ell(\hat{h}(Z_{>m}),\cdot)$, then combine the results with a union bound. The details of this are postponed to Appendix B. Note however, that this trick will not lead to a bound with a V_n term as in Theorem 3, *i.e.* with the online posteriors $(Q(Z_{>i}))$ and $(Q(Z_{<j}))$ which get closer and closer to the final $Q(Z_{\leq m})$ based on the full sample.

4 Experiments

In this section, we experimentally compare our bound in Theorem 3 to that of TS [39], Catoni [9, Theorem 1.2.8] (with $\alpha = 2$), and Maurer in (2). For the latter, given $L_n(P_n) \in [0,1[$ and the RHS of (2), we solve for an upper bound of $L(P_n)$ by "inverting" the kl. We note that TS [39] do not claim that their bound is better than Maurer's in classification (in fact, they do better in other settings).

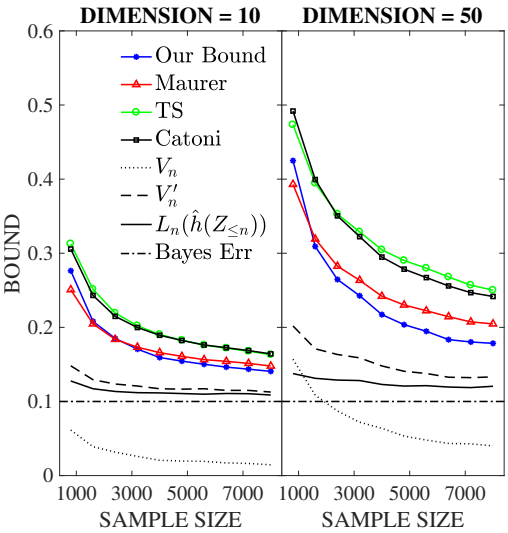

Figure 1: Results for the synthetic data.

UCI Dataset	d	Test err. of $\hat{h}$	Our bound	Maurer bound
Haberman (n=244)	3	0.272	0.521	**0.411**
Breast-C. (n=560)	9	0.068	0.185	**0.159**
TicTacToe (n=766)	27	0.046	**0.191**	0.216
Banknote (n=1098)	4	0.058	**0.125**	0.136
kr-vs-kp (n=2556)	73	0.044	**0.108**	0.165
Spambase (n=3680)	57	0.173	**0.293**	0.312
Mushroom (n=6500)	116	0.002	**0.018**	0.055
Adult (n=24130)	108	0.168	**0.195**	0.234

Table 1: Results for the UCI datasets.

Setting. We consider both synthetic and real-world datasets for binary classification, and we evaluate bounds using the 0-1 loss. In particular, the data space $\mathcal{Z}$ is $\mathcal{X} \times \mathcal{Y} \coloneqq \mathbb{R}^d \times \{0,1\}$, where $d \in \mathbb{N}$ is the dimension of the feature space. In this case, the hypothesis set $\mathcal{H}$ is also $\mathbb{R}^d$, and the error associated with $h \in \mathcal{H}$ on a sample $Z = (X,Y) \in \mathcal{X} \times \mathcal{Y}$ is given by $\ell_h(Z) = |Y - \mathbb{1}\{\phi(h^\top X) > 1/2\}|$, where $\phi(w) \coloneqq 1/(1+e^{-w}), w \in \mathbb{R}$. We learn our hypotheses using *regularized logistic regression*; given a sample $S = (Z_p, \ldots, Z_q)$, with $(p,q) \in \{(1,m),(m+1,n),(1,n)\}$ and $m = n/2$, we compute

$$\hat{h}(S) \coloneqq \arg\min_{h \in \mathcal{H}} \frac{\lambda \|h\|^2}{2} + \frac{1}{q-p+1} \sum_{i=p}^{q} Y_i \cdot \ln \phi(h^\top X_i) + (1-Y_i) \cdot \ln(1-\phi(h^\top X_i)). \tag{10}$$

For $Z_{\leq n} \in \mathcal{Z}^n$, and $1 \leq i \leq m < j \leq n$, we choose algorithm Q in Theorem 3 such that

$$Q(Z_{>i}) \equiv \delta\left(\hat{h}(Z_{>m})\right) \quad \text{and} \quad Q(Z_{<j}) \equiv \delta(\hat{h}(Z_{\leq m})).$$

Given a sample $S \neq \varnothing$, we set the "posterior" $P(S)$ to be a Gaussian centered at $\hat{h}(S)$ with variance $\sigma^2 > 0$; that is, $P(S) \equiv \mathcal{N}(\hat{h}(S), \sigma^2 I_d)$. The prior distribution is set to $P_0 \equiv \mathcal{N}(0, \sigma_0^2 I_d)$, for $\sigma_0 > 0$.

Parameters. We set $\delta = 0.05$. For all datasets, we use $\lambda = 0.01$, and (approximately) solve (10) using the BFGS algorithm. For each bound, we pick the $\sigma^2 \in \{1/2, \ldots, 1/2^J : J \coloneqq \lceil \log_2 n \rceil\}$ which minimizes it on the given data (with n instances). In order for the bounds to still hold with probability at least $1 - \delta$, we replace δ on the RHS of each bound by $\delta / \lceil \log_2 n \rceil$ (this follows from the application of a union bound). We choose the prior variance such that $\sigma_0^2 = 1/2$ (this was the best value on average for the bounds we compare against). We choose the grid $\mathcal{G}$ in Theorem 3 as in (8). Finally, we approximate Gaussian expectations using Monte Carlo sampling.

Synthetic data. We generate synthetic data for $d = \{10, 50\}$ and sample sizes between 800 and 8000. For a given sample size n, we 1) draw $X_1, \ldots, X_n$ [resp. $\epsilon_1, \ldots, \epsilon_n$] identically and independently from the multivariate-Gaussian distribution $\mathcal{N}(0, I_d)$ [resp. the Bernoulli distribution $\mathcal{B}(0.9)$]; and 2) we set $Y_i = \mathbb{1}\{\phi(h_*^\top X_i) > 1/2\} \cdot \epsilon_i$, for $i \in [n]$, where $h_* \in \mathbb{R}^d$ is the vector constructed from the first d digits of π. For example, if $d = 10$, then $h_* = (3, 1, 4, 1, 5, 9, 2, 6, 5, 3)^\top$. Figure 1 shows the results averaged over 10 independent runs for each sample size.

UCI datasets. For the second experiment, we use several UCI datasets. These are listed in Table 1 (where Breast-C. stands for Breast Cancer). We encode categorical variables in appropriate 0-1 vectors. This effectively increases the dimension of the input space (this is reported as d in Table 1). After removing any rows (*i.e.* instances) containing missing features and performing the encoding, the input data is scaled such that every column has values between -1 and 1. We used a 5-fold train-test split (n in Table 1 is the training set size), and the results in Table 1 are averages over 5 runs. We only compare with Maurer's bound since other bounds were worse than Maurer's and ours on all datasets.

Discussion. As the dimension d of the input space increases, the complexity $\mathrm{KL}(P_n \| P_0)$—and thus, all the PAC-Bayes bounds discussed in this paper—get larger. Our bound suffers less from this increase in d, since for a large enough sample size n, the term V_n is small enough (see Figure 1) to absorb any increase in the complexity. In fact, for large enough n, the irreducible (complexity-free) term involving V_n' in our bound becomes the dominant one. This, combined with the fact that for the 0-1 loss, $V_n' \approx L_n(P_n)$ for large enough n (see Figure 1), makes our bound tighter than others.

Adding a regularization term in the objective (10) is important as it stabilizes $\hat{h}(Z_{<m})$ and $\hat{h}(Z_{\geq m})$; a similar effect is achieved with methods like gradient descent as they essentially have a "built-in" regularization. For very small sample sizes, the regularization in (10) may not be enough to ensure that $\hat{h}(Z_{<m})$ and $\hat{h}(Z_{\geq m})$ are close to $\hat{h}(Z_{\leq n})$, in which case V_n need not be necessarily small. In particular, this is the case for the Haberman and the breast cancer datasets where the advantage of our bound is not fully leveraged, and Maurer's bound is smaller.

5 Theoretical Motivation of the Bound

In this section, we study the behavior of our bound (6) under a Bernstein condition:

Definition 6. [Bernstein Condition (BC)] *The learning problem $(D, \ell, \mathcal{H})$ satisfies the (β, B)-Bernstein condition, for $\beta \in [0, 1]$ and $B > 0$, if for all $h \in \mathcal{H}$,*

$$\mathbb{E}_{Z \sim D}\left[(\ell_h(Z) - \ell_{h_*}(Z))^2\right] \leq B \cdot \mathbb{E}_{Z \sim D}\left[\ell_h(Z) - \ell_{h_*}(Z)\right]^\beta,$$

where $h_* \in \arg\inf_{h \in \mathcal{H}} \mathbb{E}_{Z \sim D}\left[\ell_h(Z)\right]$ is a risk minimizer within the closer of $\mathcal{H}$.

The Bernstein condition [3, 5, 6, 15, 22] essentially characterizes the "easiness" of the learning problem; it implies that the variance in the excess loss random variable $\ell_h(Z) - \ell_{h_*}(Z)$ gets smaller the closer the risk of hypothesis $h \in \mathcal{H}$ gets to that of the risk minimizer h_*. For bounded loss functions, the BC with $\beta = 0$ always holds. The BC with $\beta = 1$ (the "easiest" learning setting) is also known as the *Massart noise condition* [25]; it holds in our experiment with synthetic data in Section 4, and also, *e.g.*, whenever $\mathcal{H}$ is convex and $h \mapsto \ell_h(z)$ is exp-concave, for all $z \in \mathcal{Z}$ [15, 31]. For more examples of learning settings where a BC holds see [22, Section 3].

Our aim in this section is to give an upper-bound on the infimum term involving V_n in (6), under a BC, in terms of the complexity COMP_n and the excess risks $\bar{L}(P_n)$, $\bar{L}(Q(Z_{>m}))$, and $\bar{L}(Q(Z_{\leq m}))$,

where for a distribution $P \in \mathcal{P}(\mathcal{H})$, the excess risk is defined by

$$\bar{L}(P) \coloneqq \mathbb{E}_{h\sim P}\left[\mathbb{E}_{Z\sim D}\left[\ell_h(Z)\right]\right] - \mathbb{E}_{Z\sim D}\left[\ell_{h_*}(Z)\right].$$

In the next theorem, we denote $Q_{\leq m} \coloneqq Q(Z_{\leq m})$ and $Q_{>m} \coloneqq Q(Z_{>m})$, for $m \in [n]$. To simplify the presentation further (and for consistency with Section 4), we assume that Q is chosen such that

$$Q(Z_{>i}) = Q_{>m}, \text{ for } 1 \leq i \leq m, \quad \text{and} \quad Q(Z_{<j}) = Q_{\leq m}, \text{ for } m < j \leq n. \tag{11}$$

Theorem 7. *Let $\mathcal{G}$ and π be as in (8), $\delta \in]0,1[$, and $\varepsilon_{\delta,n} = 2\ln\frac{1}{\delta\cdot\pi(\eta)} = 2\ln\frac{|\mathcal{G}|}{\delta}$, $\eta \in \mathcal{G}$. If the (β, B)-Bernstein condition holds with $\beta \in [0,1]$ and $B > 0$, then for any learning algorithms P and Q (with Q satisfying (11)), there exists a $\mathcal{C} > 0$, such that $\forall n \geq 1$ and $m = n/2$, with probability at least $1-\delta$,*

$$\frac{1}{\mathcal{C}}\cdot\inf_{\eta\in\mathcal{G}}\left\{c_\eta\cdot V_n + \frac{\mathrm{COMP}_n + \varepsilon_{\delta,n}}{\eta\cdot n}\right\} \leq \bar{L}(P_n) + \bar{L}(Q_{\leq m}) + \bar{L}(Q_{>m}) + \left(\frac{\mathrm{COMP}_n + \varepsilon_{\delta,n}}{n}\right)^{\frac{1}{2-\beta}} + \frac{\mathrm{COMP}_n + \varepsilon_{\delta,n}}{n}. \tag{12}$$

In addition to the "ESI" tools provided in Section 6 and Lemma 13, the proof of Theorem 7, presented in Appendix E, also uses an "ESI version" of the Bernstein condition due to [22].

First note that the only terms in our main bound (6), other than the infimum on the LHS of (12), are the empirical error $L_n(P_n)$ and a $\tilde{O}(1/\sqrt{n})$-complexity-free term which is typically smaller than $\sqrt{\mathrm{KL}(P_n\|P_0)/n}$ (*e.g.* when the dimension of $\mathcal{H}$ is large enough). The term $\sqrt{\mathrm{KL}(P_n\|P_0)/n}$ is often the dominating one in other PAC-Bayesian bounds when $\liminf_{n\to\infty} L_n(P_n) > 0$.

Now consider the remaining term in our main bound, which matches the infimum term on the LHS of (12), and let us choose algorithm P as per Remark 5, so that $\mathrm{COMP}_n = 2\mathrm{KL}(P_n\|P_0)$. Suppose that, with high probability (w.h.p.), $\mathrm{KL}(P_n\|P_0)/n$ converges to 0 for $n \to \infty$ (otherwise no PAC-Bayesian bound would converge to 0), then $(\mathrm{COMP}_n/n)^{1/(2-\beta)} + \mathrm{COMP}_n/n$—essentially the sum of the last two terms on the RHS of (12)—converges to 0 at a faster rate than $\sqrt{\mathrm{KL}(P_n\|P_0)/n}$ w.h.p. for $\beta > 0$, and at equal rate for $\beta = 0$. Thus, in light of Theorem 7, to argue that our bound can be better than others (still when $\liminf_{n\to\infty} L_n(P_n) > 0$), it remains to show that there exist algorithms P and Q for which the sum of the excess risks on the RHS of (12) is smaller than $\sqrt{\mathrm{KL}(P_n\|P_0)/n}$.

One choice of estimator with small excess risk is the *Empirical Risk Minimizer* (ERM). When $m = n/2$, if one chooses Q such that it outputs a Dirac around the ERM on a given sample, then under a BC with exponent β and for "parametric" $\mathcal{H}$ (such as the d-dimensional linear classifiers in Sec. 4), $\bar{L}(Q_{\leq m})$ and $\bar{L}(Q_{>m})$ are of order $\tilde{O}\left(n^{-1/(2-\beta)}\right)$ w.h.p. [3, 19]. However, setting $P_n \equiv \delta(\mathrm{ERM}(Z_{\leq n}))$ is not allowed, since otherwise $\mathrm{KL}(P_n\|P_0) = \infty$. Instead one can choose P_n to be the generalized-Bayes/Gibbs posterior. In this case too, under a BC with exponent β and for parametric $\mathcal{H}$, the excess risk is of order $\tilde{O}\left(n^{-1/(2-\beta)}\right)$ w.h.p. for clever choices of prior P_0 [3, 19].

6 Detailed Analysis

We start this section by presenting the convenient ESI notation and use it to present our main technical Lemma 13 (proofs of the ESI results are in Appendix D). We then continue with a proof of Theorem 3.

Definition 8. [ESI (*Exponential Stochastic Inequality, pronounce as:easy*) 19, 22] *Let $\eta > 0$, and X, Y be any two random variables with joint distribution D. We define*

$$X \trianglelefteq_\eta^D Y \iff X - Y \trianglelefteq_\eta^D 0 \iff \mathbb{E}_{(X,Y)\sim D}\left[e^{\eta(X-Y)}\right] \leq 1. \tag{13}$$

Definition 8 can be extended to the case where $\eta = \hat{\eta}$ is also a random variable, in which case the expectation in (13) needs to be replaced by the expectation over the joint distribution of $(X, Y, \hat{\eta})$. When no ambiguity can arise, we omit D from the ESI notation. Besides simplifying notation, ESIs are useful in that they simultaneously capture "with high probability" and "in expectation" results:

Proposition 9. [ESI Implications] *For fixed $\eta > 0$, if $X \trianglelefteq_\eta Y$ then $\mathbb{E}[X] \leq \mathbb{E}[Y]$. For both fixed and random $\hat{\eta}$, if $X \trianglelefteq_{\hat{\eta}} Y$, then $\forall\delta \in]0,1[$, $X \leq Y + \frac{\ln\frac{1}{\delta}}{\hat{\eta}}$, with probability at least $1-\delta$.*

In the next proposition, we present two results concerning transitivity and additive properties of ESI:

Proposition 10. [ESI Transitivity and Chain Rule] *(a) Let $Z_1, \dots, Z_n$ be any random variables on $\mathcal{Z}$ (not necessarily independent). If for some $(\gamma_i)_{i\in[n]} \in]0, +\infty[^n$, $Z_i \trianglelefteq_{\gamma_i} 0$, for all $i \in [n]$, then*

$$\sum_{i=1}^{n} Z_i \trianglelefteq_{\nu_n} 0, \quad \text{where } \nu_n \coloneqq \left(\sum_{i=1}^{n} \frac{1}{\gamma_i}\right)^{-1} \quad (\text{so if } \forall i \in [n], \gamma_i = \gamma > 0 \text{ then } \nu_n = \gamma/n). \tag{14}$$

(b) Suppose now that $Z_1, \dots, Z_n$ are i.i.d. and let $X : \mathcal{Z} \times \bigcup_{i=1}^{n} \mathcal{Z}^i \to \mathbb{R}$ be any real-valued function. If for some $\eta > 0$, $X(Z_i; z_{<i}) \trianglelefteq_\eta 0$, for all $i \in [n]$ and all $z_{<i} \in \mathcal{Z}^{i-1}$, then $\sum_{i=1}^{n} X(Z_i; Z_{<i}) \trianglelefteq_\eta 0$.

We now give a basic PAC-Bayesian result for the ESI context:

Proposition 11. [ESI PAC-Bayes] *Fix $\eta > 0$ and let $\{Y_h : h \in \mathcal{H}\}$ be any family of random variables such that for all $h \in \mathcal{H}$, $Y_h \trianglelefteq_\eta 0$. Let P_0 be any distribution on $\mathcal{H}$ and let $P : \bigcup_{i=1}^{n} \mathcal{Z}^i \to \mathcal{P}(\mathcal{H})$ be a learning algorithm. We have:*

$$\mathbb{E}_{h\sim P_n}[Y_h] \trianglelefteq_\eta \frac{\mathrm{KL}(P_n \| P_0)}{\eta}, \quad \text{where } P_n \coloneqq P(Z_{\leq n}).$$

In many applications (especially for our main result) it is desirable to work with a random (*i.e.* data-dependent) η in the ESI inequalities; one can tune η after seeing the data.

Proposition 12. [ESI from fixed to random η] *Let $\mathcal{G}$ be a countable subset of $]0, +\infty[$ and let π be a prior distribution over $\mathcal{G}$. Given a countable collection $\{Y_\eta : \eta \in \mathcal{G}\}$ of random variables satisfying $Y_\eta \trianglelefteq_\eta 0$, for all fixed $\eta \in \mathcal{G}$, we have, for arbitrary estimator $\hat{\eta}$ with support on $\mathcal{G}$,*

$$Y_{\hat{\eta}} \trianglelefteq_{\hat{\eta}} \frac{-\ln \pi(\hat{\eta})}{\hat{\eta}}. \tag{15}$$

The following key lemma, which is of independent interest, is central to our main result.

Lemma 13. [Key result: *un*-expected Bernstein] *Let $X \sim D$ be a random variable bounded from* above *by $b > 0$ almost surely, and let $\vartheta(u) \coloneqq (-\ln(1-u) - u)/u^2$. For all $0 < \eta < 1/b$, we have **(a)**:*

$$\mathbb{E}[X] - X \trianglelefteq_\eta^D c \cdot X^2, \quad \text{for all } c \geq \eta \cdot \vartheta(\eta b). \tag{16}$$

***(b)**: The result is tight; for every $c < \eta \cdot \vartheta(\eta b)$, there exists a distribution D so that (16) does not hold.*

Lemma 13 is reminiscent of the following slight variation of Bernstein's inequality [12]; let X be any random variable bounded from *below* by $-b$, and let $\kappa(x) \coloneqq (e^x - x - 1)/x^2$. For all $\eta > 0$, we have

$$\mathbb{E}[X] - X \trianglelefteq_\eta s \cdot \mathbb{E}[X^2], \quad \text{for all } s \geq \eta \cdot \kappa(\eta b). \tag{17}$$

Note that the un-expected Bernstein Lemma 13 has the X^2 lifted out of the expectation. In Appendix G, we prove (17) and compare it to standard versions of Bernstein. We also compare (16) to the related but distinct empirical Bernstein inequality due to [27, Theorem 4]. We now prove part (a) of Lemma 13, which follows easily from the proof of an existing result [16, 21]. Part (b) is novel; its proof is postponed to Appendix F.

Proof of Lemma 13-Part (a). [16] (see also [21]) showed in the proof of their lemma 4.1 that

$$\exp(\lambda\xi - \lambda^2\vartheta(\lambda)\xi^2) \leq 1 + \lambda\xi, \quad \text{for all } \lambda \in [0, 1[\text{ and } \xi \geq -1. \tag{18}$$

Letting $\eta = \lambda/b$ and $\xi = -X/b$, (18) becomes,

$$\exp(-\eta X - \eta^2\vartheta(\eta b)X^2) \leq 1 - \eta X, \quad \text{for all } \eta \in]0, 1/b[. \tag{19}$$

Taking expectation on both sides of (19) and using the fact that $1 - \eta\mathbb{E}[X] \leq \exp(-\eta\mathbb{E}[X])$ on the RHS of the resulting inequality, leads to (16). □

Proof of Theorem 3. Let $\eta \in]0, 1/b[$ and $c_\eta \coloneqq \eta \cdot \vartheta(\eta b)$. For $1 \leq i \leq m < j \leq n$, define

$$X_h(Z_i; z_{>i}) \coloneqq \ell_h(Z_i) - \mathbb{E}_{h'\sim Q(z_{>i})}[\ell_{h'}(Z_i)], \quad \text{for } z_{>i} \in \mathcal{Z}^{n-i},$$
$$\tilde{X}_h(Z_j; z_{<j}) \coloneqq \ell_h(Z_j) - \mathbb{E}_{h'\sim Q(z_{<j})}[\ell_{h'}(Z_j)], \quad \text{for } z_{<j} \in \mathcal{Z}^{j-1}.$$

Since ℓ is bounded from above by b, Lemma 13 implies that for all $h \in \mathcal{H}$ and $1 \le i \le m < j \le n$,

$$\forall z_{>i} \in \mathcal{Z}^{n-i}, \quad Y_h^\eta(Z_i; z_{>i}) \coloneqq \mathbb{E}_{Z_i' \sim D}\left[X_h(Z_i'; z_{>i})\right] - X_h(Z_i; z_{>i}) - c_\eta \cdot X_h(Z_i; z_{>i})^2 \trianglelefteq_\eta 0,$$
$$\forall z_{<j} \in \mathcal{Z}^{j-1}, \quad \tilde{Y}_h^\eta(Z_j; z_{<j}) \coloneqq \mathbb{F}_{Z_j' \sim D}[\tilde{X}_h(Z_j'; z_{<j})] - \tilde{X}_h(Z_j; z_{<j}) - c_\eta \cdot \tilde{X}_h(Z_j; z_{<j})^2 \trianglelefteq_\eta 0,$$

Since $Z_1, \ldots, Z_n$ are i.i.d. we can chain the ESIs above using Proposition 10-(b) to get:

$$S \coloneqq \sum_{i=1}^m Y_h^\eta(Z_i; Z_{>i}) \trianglelefteq_\eta 0, \qquad \tilde{S} \coloneqq \sum_{j=m+1}^n \tilde{Y}_h^\eta(Z_j; Z_{<j}) \trianglelefteq_\eta 0. \tag{20}$$

Applying PAC-Bayes (Proposition 11) to S and $\tilde{S}$ in (20) with priors $P(Z_{>m})$ and $P(Z_{\le m})$, respectively, and common posterior $P_n = P(Z_{\le n})$ on $\mathcal{H}$, we get, with $\mathrm{KL}_{>m} \coloneqq \mathrm{KL}(P_n \| P(Z_{>m}))$ and $\mathrm{KL}_{\le m} \coloneqq \mathrm{KL}(P_n \| P(Z_{\le m}))$:

$$\mathbb{E}_{h \sim P_n}\left[\sum_{i=1}^m Y_h^\eta(Z_i; Z_{>i})\right] - \frac{\mathrm{KL}_{>m}}{\eta} \trianglelefteq_\eta 0, \qquad \mathbb{E}_{h \sim P_n}\left[\sum_{j=m+1}^n \tilde{Y}_h^\eta(Z_j; Z_{<j})\right] - \frac{\mathrm{KL}_{\le m}}{\eta} \trianglelefteq_\eta 0.$$

We now apply Proposition 10-(a) to chain these two ESIs, which yields

$$\mathbb{E}_{h \sim P_n}\left[\sum_{i=1}^m Y_h^\eta(Z_i; Z_{>i}) + \sum_{j=m+1}^n \tilde{Y}_h^\eta(Z_j; Z_{<j})\right] \trianglelefteq_{\frac{\eta}{2}} \frac{\mathrm{KL}(P_n \| P(Z_{>m})) + \mathrm{KL}(P_n \| P(Z_{\le m}))}{\eta}.$$

With the prior π on $\mathcal{G}$, we have for any $\hat{\eta} = \hat{\eta}(Z_{\le n}) \in \mathcal{G} \subset [1/\sqrt{nb^2}, 1/b[$ (see Proposition 12),

$$\mathbb{E}_{h \sim P_n}\left[\sum_{i=1}^m Y_h^{\hat{\eta}}(Z_i; Z_{>i}) + \sum_{j=m+1}^n \tilde{Y}_h^{\hat{\eta}}(Z_j; Z_{<j})\right] \trianglelefteq_{\frac{\hat{\eta}}{2}} \frac{\mathrm{COMP}_n}{\hat{\eta}} - \frac{2 \ln \pi(\hat{\eta})}{\hat{\eta}}, \text{ i.e.,}$$
$$n \cdot (L(P_n) - L_n(P_n)) \quad \trianglelefteq_{\frac{\hat{\eta}}{2}} n \cdot c_{\hat{\eta}} \cdot V_n + \frac{\mathrm{COMP}_n + 2 \ln \frac{1}{\pi(\hat{\eta})}}{\hat{\eta}} +$$
$$\left[\sum_{i=1}^m \left(\mathbb{E}_{Z_i' \sim D}\left[\bar{\ell}_{Q_{>i}}(Z_i')\right] - \bar{\ell}_{Q_{>i}}(Z_i)\right) + \sum_{j=m+1}^n \left(\mathbb{E}_{Z_j' \sim D}\left[\bar{\ell}_{Q_{<j}}(Z_j')\right] - \bar{\ell}_{Q_{<j}}(Z_j)\right)\right], \tag{21}$$

where $\bar{\ell}_{Q_{>i}}(Z_i) \coloneqq \mathbb{E}_{h \sim Q(Z_{>i})}\left[\ell_h(Z_i)\right]$ and $\bar{\ell}_{Q_{<j}}(Z_j) \coloneqq \mathbb{E}_{h \sim Q(Z_{<j})}\left[\ell_h(Z_j)\right]$. Let U_n denote the quantity between the square brackets in (21). Using the un-expected Bernstein Lemma 13, together with Proposition 15, we get for any estimator $\hat{\nu}$ on $\mathcal{G}$:

$$U_n \trianglelefteq_{\hat{\nu}} c_{\hat{\nu}} \cdot \left(\sum_{i=1}^m \mathbb{E}_{h' \sim Q(Z_{>i})}\left[\ell_{h'}(Z_i)^2\right] + \sum_{j=m+1}^n \mathbb{E}_{h' \sim Q(Z_{<j})}\left[\ell_{h'}(Z_j)^2\right]\right) + \frac{\ln \frac{1}{\pi(\hat{\nu})}}{\hat{\nu}}. \tag{22}$$

By chaining (22) and (21) using Proposition 10-(a) and dividing by n, we get:

$$L(P_n) \trianglelefteq_{\frac{n\hat{\eta}\hat{\nu}}{\hat{\eta}+2\hat{\nu}}} L_n(P_n) + c_{\hat{\eta}} \cdot V_n + \frac{\mathrm{COMP}_n + 2 \ln \frac{1}{\pi(\hat{\eta})}}{\hat{\eta} \cdot n} + c_{\hat{\nu}} \cdot V_n' + \frac{\ln \frac{1}{\pi(\hat{\nu})}}{\hat{\nu} \cdot n}. \tag{23}$$

We now apply Proposition 9 to (23) to obtain the following inequality with probability at least $1 - \delta$:

$$L(P_n) \le L_n(P_n) + \left[c_{\hat{\eta}} \cdot V_n + \frac{\mathrm{COMP}_n + 2 \ln \frac{1}{\pi(\hat{\eta}) \cdot \delta}}{\hat{\eta} \cdot n}\right] + \left\{c_{\hat{\nu}} \cdot V_n' + \frac{\ln \frac{1}{\pi(\hat{\nu}) \cdot \delta}}{\hat{\nu} \cdot n}\right\}. \tag{24}$$

Inequality (6) follows after picking $\hat{\nu}$ and $\hat{\eta}$ to be, respectively, estimators which achieve the infimum over the closer of $\mathcal{G}$ of the quantities between braces and square brackets in (24). □

7 Conclusion and Future Work

The main goal of this paper was to introduce a new PAC-Bayesian bound based on a new proof technique; we also theoretically motivated the bound in terms of a Bernstein condition. The simple experiments we provided are to be considered as a basic sanity check—in future work, we plan to put the bound to real practical use by applying it to deep nets in the style of, *e.g.*, [42].

Acknowledgments

An anonymous referee made some highly informed remarks on our paper, which led us to substantially rewrite the paper and made us understand our own work much better. Part of this work was performed while Zakaria Mhammedi was interning at the Centrum Wiskunde & Informatica (CWI). This work was also supported by the Australian Research Council and Data61.

References

[1] Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

[2] Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in neural information processing systems*, pages 9–16, 2007.

[3] Jean-Yves Audibert. PAC-Bayesian statistical learning theory. *These de doctorat de l'Université Paris*, 6:29, 2004.

[4] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *International conference on algorithmic learning theory*, pages 150–165. Springer, 2007.

[5] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[6] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

[7] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[8] Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 2003.

[9] Olivier Catoni. *PAC-Bayesian Supervised Classification*. Lecture Notes-Monograph Series. IMS, 2007.

[10] Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Lecture Notes-Monograph Series. IMS, 2007.

[11] Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

[12] Nicolò Cesa-Bianchi and Gàbor Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, UK, 2006.

[13] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[14] Gintare K. Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*, 2017.

[15] Tim Van Erven, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

[16] Xiequan Fan, Ion Grama, Quansheng Liu, et al. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20, 2015.

[17] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 353–360. ACM, 2009.

[18] Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *The Journal of Machine Learning Research*, 16(1):787–860, 2015.

[19] Peter D. Grünwald and Nishant A. Mehta. Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *Journal of Machine Learning Research*, 2019. accepted pending minor revision. Available as arXiv preprint arXiv:1605.00252.

[20] Benjamin Guedj. A primer on PAC-Bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

[21] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

[22] Wouter M. Koolen, Peter D. Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems*, pages 4457–4465, 2016.

[23] John Langford and Rich Caruana. (Not) bounding the true error. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 809–816. MIT Press, 2002.

[24] John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

[25] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

[26] Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

[27] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings COLT 2009*, 2009.

[28] David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh ACM Conference on Computational Learning Theory (COLT' 98)*, pages 230–234. ACM Press, 1998.

[29] David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pages 164–171. ACM Press, 1999.

[30] David A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

[31] Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. In *Artificial Intelligence and Statistics*, pages 1085–1093, 2017.

[32] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679. ACM, 2008.

[33] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

[34] Behnam Neyshabur, Srinadh Bhojanapalli, David A. McAllester, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *ICLR*, 2018.

[35] Omar Rivasplata, Csaba Szepesvári, John S Shawe-Taylor, Emilio Parrado-Hernandez, and Shiliang Sun. Pac-bayes bounds for stable algorithms with instance-dependent priors. In *Advances in Neural Information Processing Systems*, pages 9214–9224, 2018.

[36] Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

[37] Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

[38] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[39] Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems*, pages 109–117, 2013.

[40] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[41] Olivier Wintenberger. Optimal learning with bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

[42] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *ICLR*, 2019.

A Informed Priors

Any bound of the form of (1) with $\mathrm{COMP}_n = \mathrm{KL}(P_n \| P_0)$ can be applied in a way as to replace this KL term by $\mathrm{KL}(P_n \| P(Z_{>m})) + \mathrm{KL}(P_n \| P(Z_{\leq m}))$, and thus making use of "informed priors". For this, it suffices to apply the bound on each part of the sample, *i.e.* $Z_{>m}$ and $Z_{\leq m}$, and then combine the resulting bounds with a union bound. In fact, suppose that (1) holds with $R_n = L_n(P_n)$ and $\mathcal{C} = 0$, and let $\delta \in]0,1[$. Applying the bound on the second part of the sample $Z_{>m}$ with prior $P(Z_{\leq m})$ and posterior P_n, we get, with probability at least $1-\delta$,

$$L(P_n) - L_{>m}(P_n) \leq \mathcal{P} \cdot \sqrt{\frac{L_{>m}(P_n) \cdot (\mathrm{KL}(P_n \| P(Z_{\leq m})) + \varepsilon_{\delta,n-m})}{n-m}} + \mathcal{A} \cdot \frac{\mathrm{KL}(P_n \| P(Z_{\leq m})) + \varepsilon_{\delta,n-m}}{n-m}, \tag{25}$$

where $L_{>m}(P_n) \coloneqq \frac{1}{n-m} \sum_{j=m+1}^{n} \mathbb{E}_{h \sim P_n}[\ell_h(Z_j)]$. Similarly, applying the bound on the first half of the sample $Z_{\leq m}$ with prior $P(Z_{>m})$ and posterior P_n, we get, with probability at least $1-\delta$,

$$L(P_n) - L_{\leq m}(P_n) \leq \mathcal{P} \cdot \sqrt{\frac{L_{\leq m}(P_n) \cdot (\mathrm{KL}(P_n \| P(Z_{>m})) + \varepsilon_{\delta,m})}{m}} + \mathcal{A} \cdot \frac{\mathrm{KL}(P_n \| P(Z_{>m})) + \varepsilon_{\delta,m}}{m}, \tag{26}$$

where $L_{\leq m}(P_n) \coloneqq \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{h \sim P_n}[\ell_h(Z_i)]$. Let $p \coloneqq m/n$ and $q \coloneqq (n-m)/n$ (note that $p+q=1$). Applying a union bound and adding $q \times$ (25) with $p \times$ (26), yields the bound

$$L(P_n) - L_n(P_n) \leq \mathcal{P} \cdot \sqrt{\frac{2L_n(P_n) \cdot (\mathrm{KL}(P_n \| P(Z_{>m})) + \mathrm{KL}(P_n \| P(Z_{\leq m})) + \bar{\varepsilon}_{\delta,n})}{n}} + \mathcal{A} \cdot \frac{\mathrm{KL}(P_n \| P(Z_{>m})) + \mathrm{KL}(P_n \| P(Z_{\leq m})) + \bar{\varepsilon}_{\delta,n}}{n}, \tag{27}$$

with probability at least $1-\delta$, where $\bar{\varepsilon}_{\delta,n} \coloneqq \varepsilon_{\delta/2,m} + \varepsilon_{\delta/2,n-m}$. To get to (27), we also used the fact that $\sqrt{x} + \sqrt{y} \leq \sqrt{2(x+y)}$, for all $x, y \in \mathbb{R}_{\geq 0}$.

The above trick does not directly apply to Maurer's bound in (2) (since the dependence on $L(P_n)$ is not linear). Instead, one can use the joint convexity of the binary Kullback-Leibler divergence kl in its two arguments as in the following proof of Lemma 4:

Proof of Lemma 4. Let $\delta \in]0,1[$. We can write $L_n(P_n)$ as

$$L_n(P_n) = \frac{p}{m} \sum_{i=1}^{m} \mathbb{E}_{h \sim P_n}[\ell_h(Z_i)] + \frac{q}{n-m} \sum_{j=m+1}^{n} \mathbb{E}_{h \sim P_n}[\ell_h(Z_j)],$$

where $p \coloneqq m/n$ and $q \coloneqq (n-m)/n$ (note that $p+q=1$). Let us denote

$$L_{\leq m}(P_n) \coloneqq \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{h \sim P_n}[\ell_h(Z_i)] \quad \text{and} \quad L_{>m}(P_n) \coloneqq \frac{1}{n-m} \sum_{j=m+1}^{n} \mathbb{E}_{h \sim P_n}[\ell_h(Z_j)].$$

By the joint convexity of the binary Kullback-Leibler divergence kl in its two arguments, we have

$$\begin{aligned} \mathrm{kl}(L(P_n) \| L_n(P_n)) &= \mathrm{kl}(pL(P_n) + qL(P_n) \| pL_{\leq m}(P_n) + qL_{>m}(P_n)), \\ &\leq p \cdot \mathrm{kl}(L(P_n) \| L_{\leq m}(P_n)) + q \cdot \mathrm{kl}(L(P_n) \| L_{>m}(P_n)), \\ &\leq p \cdot \frac{\mathrm{KL}(P_n \| P(Z_{>m})) + \ln \frac{4\sqrt{m}}{\delta}}{m}, \\ &\quad + q \cdot \frac{\mathrm{KL}(P_n \| P(Z_{\leq m})) + \ln \frac{4\sqrt{n-m}}{\delta}}{n-m}, \end{aligned} \tag{28}$$

with probability at least $1-\delta$, where the last inequality follows by Maurer's bound (2) and the union bound. Substituting the expressions of p and q in (28) yields the desired result. □

B Biasing

A PAC-Bayes bound similar to the one in our Corollary 1 can be obtained from the TS bound. For this, the TS bound must be applied twice, once on each part of the sample (*i.e.* $Z_{\leq m}$ and $Z_{>m}$) to *biased* losses. We demonstrate this in what follows.

Let $\hat{h} : \bigcup_{i=1}^{n} \mathcal{Z}^i \to \mathcal{H}$ be any estimator. The TS bound can be expressed in the form of (1) with $\text{COMP}_n = \text{KL}(P_n \| P_0)$, $\mathcal{C} = 0$, and $R_n = \mathbb{E}_{h\sim P_n}[\text{Var}_n[\ell_h(Z)]]$, where $\text{Var}_n[X]$ denotes the empirical variance. Applying the TS bound on the second part of the sample $Z_{>m}$ with prior P_0 and posterior P_n, and with the biased loss $\tilde{\ell}_h(Z) = \ell_h(Z) - \ell_{\hat{h}(Z_{\leq m})}(Z)$, gives

$$\tilde{L}(P_n) - \tilde{L}_{>m}(P_n) \leq \mathcal{P} \cdot \sqrt{\frac{\mathbb{E}_{h\sim P_n}[\text{Var}_{>m}[\tilde{\ell}_h(Z)]] \cdot (\text{KL}(P_n \| P_0) + \varepsilon_{\delta,n-m})}{n-m}} + \mathcal{A} \cdot \frac{\text{KL}(P_n \| P_0) + \varepsilon_{\delta,n-m}}{n-m}, \tag{29}$$

with probability at least $1-\delta$, where $\text{Var}_{>m}[X] \coloneqq \frac{1}{n-m}\sum_{i=m+1}^{n}\left(X_i - \frac{1}{n-m}\sum_{j=m+1}^{n} X_j\right)^2$, $\tilde{L}(P_n) \coloneqq \mathbb{E}_{h\sim P_n}[\mathbb{E}_{Z\sim D}[\tilde{\ell}_h(Z)]]$, and $\tilde{L}_{>m}(P_n) \coloneqq \frac{1}{n-m}\sum_{j=m+1}^{n}\mathbb{E}_{h\sim P_n}[\tilde{\ell}_h(Z_j)]$.

Doing the same on the first part of the sample $Z_{\leq m}$, but now with the loss $\check{\ell}_h(Z) \coloneqq \ell_h(Z) - \ell_{\hat{h}(Z_{>m})}(Z)$, yields

$$\check{L}(P_n) - \check{L}_{\leq m}(P_n) \leq \mathcal{P} \cdot \sqrt{\frac{\mathbb{E}_{h\sim P_n}[\text{Var}_{\leq m}[\check{\ell}_h(Z)]] \cdot (\text{KL}(P_n \| P_0) + \varepsilon_{\delta,m})}{m}} + \mathcal{A} \cdot \frac{\text{KL}(P_n \| P_0) + \varepsilon_{\delta,m}}{m}, \tag{30}$$

with probability at least $1-\delta$, where $\text{Var}_{\leq m}[X] \coloneqq \frac{1}{m}\sum_{i=1}^{m}\left(X_i - \frac{1}{m}\sum_{j=1}^{m} X_j\right)^2$, $\check{L}(P_n) \coloneqq \mathbb{E}_{h\sim P_n}[\mathbb{E}_{Z\sim D}[\check{\ell}_h(Z)]]$, and $\check{L}_{\leq m}(P_n) \coloneqq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{h\sim P_n}[\check{\ell}_h(Z_i)]$.

Two more applications of the TS bound with prior and posterior equal to P_0, yields,

$$L(\hat{h}(Z_{\leq m})) - L_{>m}(\hat{h}(Z_{\leq m})) \leq \mathcal{P} \cdot \sqrt{\frac{\text{Var}_{>m}[\ell_{\hat{h}(Z_{\leq m})}(Z)] \cdot \varepsilon_{\delta/2,n-m}}{n-m}} + \frac{\mathcal{A} \cdot \varepsilon_{\delta/2,n-m}}{n-m}, \text{ and} \tag{31}$$

$$L(\hat{h}(Z_{>m})) - L_{\leq m}(\hat{h}(Z_{>m})) \leq \mathcal{P} \cdot \sqrt{\frac{\text{Var}_{\leq m}[\ell_{\hat{h}(Z_{>m})}(Z)] \cdot \varepsilon_{\delta/2,m}}{m}} + \frac{\mathcal{A} \cdot \varepsilon_{\delta/2,m}}{m}, \tag{32}$$

with probability at least $1-\delta$, where

$$L_{\leq m}(\hat{h}(Z_{>m})) \coloneqq \frac{1}{m}\sum_{i=1}^{m}\ell_{\hat{h}(Z_{>m})}(Z_i) \quad \text{and} \quad L_{>m}(\hat{h}(Z_{\leq m})) \coloneqq \frac{1}{n-m}\sum_{j=m+1}^{n}\ell_{\hat{h}(Z_{\leq m})}(Z_j).$$

Let $p = m/n$ and $q = (n-m)/n$. Applying a union bound and combining (29)-(32), as

$$q \times ((29) + (31)) + p \times ((30) + (32)),$$

yields a bound of the form (1) with

$$R'_n = p \cdot \text{Var}_{\leq m}[\ell_{\hat{h}(Z_{>m})}(Z)] + q \cdot \text{Var}_{<m}[\ell_{\hat{h}(Z_{\leq m})}(Z)] \leq V'_n,$$

$$R_n = p \cdot \mathbb{E}_{h\sim P_n}[\text{Var}_{\leq m}[\check{\ell}_h(Z)]] + q \cdot \mathbb{E}_{h\sim P_n}[\text{Var}_{>m}[\tilde{\ell}_h(Z)]] \leq V_n,$$

where V'_n and V_n are as in Corollary 1.

A Direct Approach. Though the steps above lead to a bound similar to ours in Corollary 1, the constants involved may not be optimal. We now re-derive a modification of the TS bound with a V_n term like in Corollary 1, and with tighter constants. The proof techniques used here are the same as those used in the proof of Theorem 3. For $\eta \in]0, 1/b[$ (where $b > 0$ is an upper-bound on the loss ℓ) and $m \in [2..n]$, define

$$s_\eta \coloneqq \eta \cdot \kappa(\eta b), \quad \text{where } \kappa(\eta) \coloneqq (e^\eta - \eta - 1)/\eta^2,$$

$$\text{and} \quad \tilde{c}_\eta \coloneqq \frac{s_\eta m}{2m-2}\left(1 + \frac{\eta m}{2m-2}\right)^{-1}, \quad \lambda(\eta) \coloneqq \frac{\eta\beta(\eta)}{\eta + \beta(\eta)}, \quad \text{where } \beta(\eta) \coloneqq \eta + \frac{\eta^2 m^2}{2m-2}.$$

We assume that $n > 2$ is even in the next theorem. We remind the reader of the definitions

$$\mathrm{Var}_{\le m}[X] := \frac{1}{m}\sum_{i=1}^{m}\left(X_i - \frac{1}{m}\sum_{j=1}^{m} X_j\right)^2 \text{ and } \mathrm{Var}_{>m}[X] := \frac{1}{n-m}\sum_{i=m+1}^{n}\left(X_i - \frac{1}{n-m}\sum_{j=m+1}^{n} X_j\right)^2.$$

Theorem 14. [New PAC-Bayes Empirical Bernstein Bound] *Let $Z_1, \dots, Z_n$ be i.i.d. with $Z_1 \sim D$. Let $m = n/2 > 1$ and π be any distribution with support on a finite or countable grid $\mathcal{G} \subset]0, 1/b[$. For any $\delta \in]0, 1[$, learning algorithm $P : \bigcup_{i=1}^{n} \mathcal{Z}^i \to \mathcal{P}(\mathcal{H})$, and estimator $\hat{h} : \bigcup_{i=1}^{n} \mathcal{Z}^i \to \mathcal{H}$, we have,*

$$L(P_n) \le L_n(P_n) + \inf_{\eta\in\mathcal{G}}\left\{\tilde{c}_\eta \cdot G_n + \frac{\mathrm{COMP}_n + 2\ln\frac{1}{\delta\cdot\pi(\eta)}}{\lambda(\eta)\cdot n}\right\} + \inf_{\nu\in\mathcal{G}}\left\{\tilde{c}_\nu \cdot G'_n + \frac{\ln\frac{1}{\delta\cdot\pi(\nu)}}{\lambda(\nu)\cdot n}\right\},$$

with probability at least $1-\delta$, where COMP_n, G'_n, and G_n are the random variables defined by:

$$\mathrm{COMP}_n := \mathrm{KL}(P_n \| P(Z_{\le m})) + \mathrm{KL}(P_n \| P(Z_{>m})),$$
$$G'_n := \mathrm{Var}_{>m}\left[\ell_{\hat{h}(Z_{\le m})}(Z)\right] + \mathrm{Var}_{\le m}\left[\ell_{\hat{h}(Z_{>m})}(Z)\right],$$
$$G_n := \mathbb{E}_{h\sim P_n}\left[\mathrm{Var}_{>m}\left[\ell_h(Z) - \ell_{\hat{h}(Z_{\le m})}(Z)\right] + \mathrm{Var}_{\le m}\left[\ell_h(Z) - \ell_{\hat{h}(Z_{>m})}(Z)\right]\right].$$

Note that since $\mathrm{Var}_{\le m}(X) \le \sum_{i=1}^{m} X_i^2/m$ and $\mathrm{Var}_{>m}(X) \le \sum_{i=m+1}^{n} X_i^2/m$, we have

$$G_n \le V_n \quad \text{and} \quad G'_n \le V'_n,$$

where V_n and V'_n are defined in (4) and (3), respectively. However, one cannot directly compare G_n to the V_n defined in Theorem 3, since the latter uses "online" posteriors $(Q(Z_{>i}))$ and $Q(Z_{<j})$ which get closer and closer to the posterior $Q(Z_{\le n})$ based on the full sample.

To prove Theorem 14, we need the following self-bounding property of the empirical variance [27]:

$$m\,\mathrm{Var}[X] \trianglelefteq_\eta \frac{m^2}{m-1}\mathrm{Var}_m[X] - \frac{\eta m^2}{2m-2}\mathrm{Var}[X], \tag{33}$$

for any $\eta > 0$ and any bounded random variable X, where $\mathrm{Var}_m[X]$ is either $\mathrm{Var}_{>m}[X]$ or $\mathrm{Var}_{\le m}[X]$ (recall that $m = n/2$). Re-arranging (33) and dividing by $(1 + \eta m/(2m-2))$, leads to

$$m\,\mathrm{Var}[X] \trianglelefteq_{\beta(\eta)} \frac{m^2}{m-1}\cdot\left(1 + \frac{\eta m}{2m-2}\right)^{-1}\mathrm{Var}_m[X], \tag{34}$$
$$\text{where} \quad \beta(\eta) := \eta + \frac{\eta^2 m}{2m-2}.$$

Proof of Theorem 14. Let $\eta \in]0, 1/b[$ and $s_\eta := \eta\cdot\kappa(\eta b)$. We define

$$X_h(Z_i) := \ell_h(Z_i) - \ell_{\hat{h}(Z_{>m})}(Z_i), \quad \text{for } 1 \le i \le m,$$
$$\tilde{X}_h(Z_j) := \ell_h(Z_j) - \ell_{\hat{h}(Z_{\le m})}(Z_j), \quad \text{for } m < j \le n.$$

Since ℓ is bounded from above by b, the Bernstein inequality (17) applied to the zero-mean random variables $\mathbb{E}_{Z'_i\sim D}[X_h(Z'_i)] - X_h(Z_i), i \in [n]$, implies that for all $h \in \mathcal{H}$,

$$Y_h^\eta(Z_i) := \mathbb{E}_{Z'_i\sim D}[X_h(Z'_i)] - X_h(Z_i) - s_\eta\cdot\mathrm{Var}[X_h(Z)] \trianglelefteq_\eta 0, \quad \text{for } 1 \le i \le m,$$
$$\tilde{Y}_h^\eta(Z_j) := \mathbb{E}_{Z'_j\sim D}[\tilde{X}_h(Z'_j)] - \tilde{X}_h(Z_j) - s_\eta\cdot\mathrm{Var}[\tilde{X}_h(Z)] \trianglelefteq_\eta 0, \quad \text{for } m < j \le n.$$

Since $Z_1, \dots, Z_n$ are i.i.d. we can chain the ESIs above using Proposition 10-(b) to get:

$$S := \sum_{i=1}^{m} Y_h^\eta(Z_i) \trianglelefteq_\eta 0, \qquad \tilde{S} := \sum_{j=m+1}^{n} \tilde{Y}_h^\eta(Z_j) \trianglelefteq_\eta 0.$$

Chaining $S \trianglelefteq_\eta 0$ [resp. $\tilde{S} \trianglelefteq_\eta 0$] and (34) with $\mathrm{Var}_m \equiv \mathrm{Var}_{\le m}$ [resp. $\mathrm{Var}_m \equiv \mathrm{Var}_{>m}$] using Proposition 10-(a), yields,

$$W_h^\eta \trianglelefteq_{\frac{\eta\beta(\eta)}{\eta+\beta(\eta)}} 0 \quad \text{and} \quad \tilde{W}_h^\eta \trianglelefteq_{\frac{\eta\beta(\eta)}{\eta+\beta(\eta)}} 0, \quad \text{where} \tag{35}$$
$$W_h^\eta := \sum_{i=1}^{m}\left(\mathbb{E}_{Z'_i\sim D}[X_h(Z'_i)] - X_h(Z_i)\right) - \frac{s_\eta m^2}{m-1}\cdot\left(1 + \frac{\eta m}{2m-2}\right)^{-1}\mathrm{Var}_{\le m}[X_h(Z)],$$
$$\tilde{W}_h^\eta := \sum_{j=m+1}^{n}\left(\mathbb{E}_{Z'_j\sim D}[X_h(Z'_j)] - X_h(Z_j)\right) - \frac{s_\eta m^2}{m-1}\cdot\left(1 + \frac{\eta m}{2m-2}\right)^{-1}\mathrm{Var}_{>m}[X_h(Z)].$$

Let $\lambda(\eta) \coloneqq \eta\beta(\eta)/(\beta(\eta)+\eta)$. Applying PAC-Bayes (Proposition 11) to $W_h^\eta \trianglelefteq_{\lambda(\eta)} 0$ and $\tilde{W}_h^\eta \trianglelefteq_{\lambda(\eta)} 0$ in (35), with priors $P(Z_{>m})$ and $P(Z_{\leq m})$, respectively, and posterior $P_n = P(Z_{\leq n})$ on $\mathcal{H}$, we get:

$$\mathbb{E}_{h\sim P_n}\left[W_h^\eta\right] - \frac{\mathrm{KL}(P_n\|P(Z_{>m}))}{\lambda(\eta)} \trianglelefteq_{\lambda(\eta)} 0, \quad \mathbb{E}_{h\sim P_n}\left[\tilde{W}_h^\eta\right] - \frac{\mathrm{KL}(P_n\|P(Z_{\leq m}))}{\lambda(\eta)} \trianglelefteq_{\lambda(\eta)} 0.$$

We now apply Proposition 10-(a) to chain these two ESIs, which yields

$$\mathbb{E}_{h\sim P_n}\left[W_h^\eta + \tilde{W}_h^\eta\right] \trianglelefteq_{\frac{\lambda(\eta)}{2}} \frac{\mathrm{KL}(P_n\|P(Z_{>m})) + \mathrm{KL}(P_n\|P(Z_{\leq m}))}{\lambda(\eta)}.$$

With the discrete prior π on $\mathcal{G}$, we have for any $\hat{\eta} = \hat{\eta}(Z_{\leq n}) \in \mathcal{G} \subset 1/b \cdot [1/\sqrt{n}, 1[$ (see Proposition 12),

$$\mathbb{E}_{h\sim P_n}\left[W_h^\eta + \tilde{W}_h^\eta\right] \trianglelefteq_{\frac{\lambda(\hat{\eta})}{2}} \frac{\mathrm{COMP}_n}{\lambda(\hat{\eta})} - \frac{2\ln\pi(\hat{\eta})}{\lambda(\hat{\eta})}, \text{ i.e.,}$$

$$n\cdot(L(P_n) - L_n(P_n)) \quad \trianglelefteq_{\frac{\lambda(\hat{\eta})}{2}} n\cdot\tilde{c}_{\hat{\eta}}\cdot G_n + \frac{\mathrm{COMP}_n + 2\ln\frac{1}{\pi(\hat{\eta})}}{\lambda(\hat{\eta})} +$$

$$\left[\sum_{i=1}^m \left(\mathbb{E}_{Z_i'\sim D}\left[\ell_{\hat{h}_{>m}}(Z_i')\right] - \ell_{\hat{h}_{>m}}(Z_i)\right) + \sum_{j=m+1}^n \left(\mathbb{E}_{Z_j'\sim D}\left[\ell_{\hat{h}_{\leq m}}(Z_j')\right] - \ell_{\hat{h}_{\leq m}}(Z_j)\right)\right], \tag{36}$$

where $\hat{h}_{>m} \coloneqq \hat{h}(Z_{>m})$ and $\hat{h}_{\leq m} \coloneqq \hat{h}(Z_{\leq m})$. Let U_n denote the quantity between the square brackets in (36). Using the Bernstein inequality in (17) chained with (34), and Proposition 15, we get for any estimator $\hat{\nu}$ on $\mathcal{G}$:

$$U_n \trianglelefteq_{\lambda(\hat{\nu})} n\cdot\tilde{c}_{\hat{\nu}}\cdot\left(\mathrm{Var}_{\leq m}[\ell_{\hat{h}(Z_{>m})}(Z)] + \mathrm{Var}_{>m}[\ell_{\hat{h}(Z_{\leq m})}(Z)]\right) + \frac{\ln\frac{1}{\pi(\hat{\nu})}}{\lambda(\hat{\nu})}. \tag{37}$$

By chaining (36) and (37) using Proposition 10-(a), dividing by n, we get:

$$L(P_n) \trianglelefteq_{\frac{n\lambda(\hat{\eta})\lambda(\hat{\nu})}{\lambda(\hat{\eta})+2\lambda(\hat{\nu})}} L_n(P_n) + \tilde{c}_{\hat{\eta}}\cdot G_n + \frac{\mathrm{COMP}_n + 2\ln\frac{1}{\pi(\hat{\eta})}}{\lambda(\hat{\eta})\cdot n} + \tilde{c}_{\hat{\nu}}\cdot G_n' + \frac{\ln\frac{1}{\pi(\hat{\nu})}}{\lambda(\hat{\nu})\cdot n}. \tag{38}$$

We now apply Proposition 9 to (38) to obtain the following inequality with probability at least $1-\delta$:

$$L(P_n) \leq L_n(P_n) + \left[\tilde{c}_{\hat{\eta}}\cdot G_n + \frac{\mathrm{COMP}_n + 2\ln\frac{1}{\pi(\hat{\eta})\cdot\delta}}{\lambda(\hat{\eta})\cdot n}\right] + \left\{\tilde{c}_{\hat{\nu}}\cdot G_n' + \frac{\ln\frac{1}{\pi(\hat{\nu})\cdot\delta}}{\lambda(\hat{\nu})\cdot n}\right\}. \tag{39}$$

Inequality (6) follows after picking $\hat{\nu}$ and $\hat{\eta}$ to be, respectively, estimators which achieve the infimum over the closer of $\mathcal{G}$ of the quantities between braces and square brackets in (39). □

C Proof of Lemma 2

Proof. Throughout this proof, we denote $\hat{h}_{>m} \coloneqq \hat{h}(Z_{>m})$ and $\hat{h}_{\leq m} \coloneqq \hat{h}(Z_{\geq m})$. Let $\delta \in]0,1[$. Since the sample $Z_{\leq m}$ is independent of $Z_{>m}$, we have

$$\frac{2}{n}\sum_{i=1}^m \ell_{\hat{h}_{>m}}(Z_i)^2 = \mathrm{Var}_{\leq m}[\ell_{\hat{h}_{>m}}(Z)] + \left(\frac{1}{m}\sum_{i=1}^m \ell_{\hat{h}_{>m}}(Z_i)\right)^2. \tag{40}$$

On the other hand, from [27, Theorem 10], we have

$$\begin{aligned}\mathrm{Var}_{\leq m}[\ell_{\hat{h}_{>m}}(Z)] &\leq \frac{2(m-1)}{m}\mathrm{Var}[\ell_{\hat{h}_{>m}}(Z)] + \frac{8\ln\frac{1}{\delta}}{n},\\ &\overset{|\ell|\leq 1}{\leq} \frac{2(m-1)}{m}L(\hat{h}_{>m}) + \frac{8\ln\frac{1}{\delta}}{n},\end{aligned} \tag{41}$$

with probability at least $1-\delta$. By Hoeffding's inequality, we also have

$$\begin{aligned}\left(\frac{1}{m}\sum_{i=1}^m \ell_{\hat{h}_{>m}}(Z_i)\right)^2 &\leq 2L(\hat{h}_{>m})^2 + \frac{8\ln\frac{1}{\delta}}{n},\\ &\overset{|\ell|\leq 1}{\leq} 2L(\hat{h}_{>m}) + \frac{8\ln\frac{1}{\delta}}{n},\end{aligned} \tag{42}$$

with probability at least $1-\delta$. Combining (40), (41), and (42) together using a union bound, yields

$$\frac{2}{n}\sum_{i=1}^{m}\ell_{\hat{h}_{>m}}(Z_i)^2 \le \frac{4(n-1)}{n}L(\hat{h}_{>m}) + \frac{16\ln\frac{2}{\delta}}{n}, \tag{43}$$

with probability at least $1-\delta$. Applying the same argument on the second part of the sample $Z_{>m}$, yields

$$\frac{2}{n}\sum_{j=m+1}^{n}\ell_{\hat{h}_{\le m}}(Z_i)^2 \le \frac{4(n-1)}{n}L(\hat{h}_{\le m}) + \frac{16\ln\frac{2}{\delta}}{n}, \tag{44}$$

with probability at least $1-\delta$. Applying a union bound, and adding together (43) and (44) then dividing by 2, yields,

$$\begin{aligned} R'_n &\le \frac{2(n-1)}{n}\left(L(\hat{h}_{\le m}) + L(\hat{h}_{>m})\right) + \frac{16\ln\frac{4}{\delta}}{n}, \\ &\le 2\left(L(\hat{h}_{\le m}) + L(\hat{h}_{>m})\right) + \frac{16\ln\frac{4}{\delta}}{n}, \end{aligned} \tag{45}$$

with probability at least $1-\delta$. Diving (45) by n and applying the square-root yields the desired result. □

D Proofs for Section 6

Proof of Proposition 9. Let $Z = X - Y$. For fixed η, Jensen's inequality yields $\mathbb{E}[Z] \le 0$. For $\eta = \hat{\eta}$ that is either fixed or itself a random variable, applying Markov's inequality to the random variable $e^{-\hat{\eta}Z}$ yields $Z \le \frac{\ln\frac{1}{\delta}}{\hat{\eta}}$, with probability at least $1-\delta$, for any $\delta \in]0,1[$. □

Proof of Proposition 10. **[Part (a)]** Fix $(\gamma_i)_{i\in[n]} \in]0,+\infty[^n$, and let $\nu_j := \left(\sum_{i=1}^{j}\frac{1}{\gamma_i}\right)^{-1}$, for $j \in [n]$. We proceed by induction to show that $\forall j \in [n],\ \sum_{i=1}^{j} Z_i \trianglelefteq_{\nu_j} 0$. The result holds trivially for $j = 1$, since $\nu_1 = \gamma_1$. Suppose that

$$\sum_{i=1}^{j} Z_i \trianglelefteq_{\nu_j} 0, \tag{46}$$

for some $1 \le j < n$. We now show that (46) holds for $j+1$; we have,

$$\begin{aligned} \mathbb{E}\left[e^{\frac{\nu_j\gamma_{j+1}}{\nu_j+\gamma_{j+1}}\left(\sum_{i=1}^{j} Z_i + Z_{j+1}\right)}\right] &= \mathbb{E}\left[e^{\frac{\nu_j\gamma_{j+1}}{\nu_j+\gamma_{j+1}}\sum_{i=1}^{j} Z_i + \frac{\nu_j\gamma_{j+1}}{\nu_j+\gamma_{j+1}} Z_{j+1}}\right], \\ &\overset{\text{Jensen}}{\le} \frac{\gamma_{j+1}}{\nu_j+\gamma_{j+1}}\mathbb{E}\left[e^{\nu_j\sum_{i=1}^{j} Z_i}\right] + \frac{\nu_j}{\nu_j+\gamma_{j+1}}\mathbb{E}\left[e^{\gamma_{j+1}Z_{j+1}}\right], \\ &\overset{\text{using (46)}}{\le} 1. \end{aligned}$$

Thus the result holds for $j+1$, since $\nu_{j+1} = \frac{\nu_j\gamma_{j+1}}{\nu_j+\gamma_{j+1}}$. This establishes (14).

[Part (b)] This is a special case of [22, Lemma 6], who treat the general case with non-i.i.d. distributions. □

Proof of Proposition 11. Let $\rho(h) = (dP_n/dP_0)(h)$ be the density of $h \in \mathcal{H}$ relative to the prior measure P_0. We then have $\mathrm{KL}(P_n\|P_0) = \mathbb{E}_{h\sim P_n}[\ln\rho(h)]$. We can now write:

$$\begin{aligned} \mathbb{E}\left[e^{\eta\mathbb{E}_{h\sim P_n}[Y_h] - \mathrm{KL}(P_n\|P_0)}\right] &= \mathbb{E}\left[e^{\eta\mathbb{E}_{h\sim P_n}[Y_h - \ln\rho(h)]}\right], \\ &\le \mathbb{E}\left[\mathbb{E}_{h\sim P_n}\left[e^{\eta(Y_h - \ln\rho(h))}\right]\right], \quad \text{(Jensen's Inequality)} \\ &= \mathbb{E}\left[\mathbb{E}_{h\sim P_n}\left[\frac{dP_0}{dP_n}\cdot e^{\eta Y_h}\right]\right], \\ &= \mathbb{E}\left[\mathbb{E}_{h\sim P_0}\left[e^{\eta Y_h}\right]\right], \\ &= \mathbb{E}_{h\sim P_0}\left[\mathbb{E}\left[e^{\eta Y_h}\right]\right], \quad \text{(Tonelli's Theorem)} \\ &= 1, \end{aligned}$$

where the final equality follows from our assumption that $Y_h \trianglelefteq_\eta 0$, for all $h \in \mathcal{H}$. □

Proof of Proposition 12. Since $Y_\eta \trianglelefteq_\eta 0$, for $\eta \in \mathcal{G}$, we have in particular:

$$1 \geq \mathbb{E}\left[\sum_{\eta\in\mathcal{G}} \pi(\eta) e^{\eta Y_\eta}\right] \geq \mathbb{E}\left[\pi(\hat{\eta}) e^{\hat{\eta} Y_{\hat{\eta}}}\right], \tag{47}$$

where the right-most inequality follows from the fact that the expectation of a countable sum of positive random variable is greater than the expectation of a single element in the sum. Rearranging (47) gives (15). □

E Proof of Theorem 7

In what follows, for $h \in \mathcal{H}$, we denote $X_h(Z) \coloneqq \ell_h(Z) - \ell_{h_*}(Z)$ the excess loss random variable, where h_* is the risk minimizer within $\mathcal{H}$. Let

$$\rho(\eta) \coloneqq \frac{1}{\eta} \ln \mathbb{E}_{Z\sim D}\left[e^{-\eta X_h(Z)}\right]$$

be its *normalized cumulant generating function*. We need the following useful lemmas:

Lemma 15. **[22]** *Let $h \in \mathcal{H}$ and X_h be as above. Then, for all $\eta \geq 0$,*

$$\alpha_\eta \cdot X_h(Z)^2 - X_h(Z) \trianglelefteq_\eta \rho(2\eta) + \alpha_\eta \cdot \rho(2\eta)^2, \quad \textit{where } \alpha_\eta \coloneqq \frac{\eta}{1+\sqrt{1+4\eta^2}}.$$

Lemma 16. **[22]** *Let $b > 0$, and suppose that $X_h \in [-b, b]$ almost surely, for all $h \in \mathcal{H}$. If the (β, B)-Bernstein condition holds with $\beta \in [0,1]$ and $B > 0$, then*

$$\rho(\eta) \leq (B\eta)^{\frac{1}{1-\beta}}, \quad \textit{for all } \eta \in]0, 1/b].$$

Lemma 17. **[12]** *Let $b > 0$, and suppose that $X_h \in [-b, b]$ almost surely, for all $h \in \mathcal{H}$. Then*

$$\rho(\eta) \leq \frac{\eta b^2}{2}, \quad \textit{for all } \eta \in \mathbb{R}.$$

Proof of Theorem 7. First we apply the following inequality

$$(a-d)^2 \leq 2(a-c)^2 + 2(d-c)^2 \tag{48}$$

which holds for all $a, c, d \in \mathbb{R}$ to upper bound V_n. Let's focus on the first term in the expression of V_n, which we denote V_n^{left}: that is,

$$V_n^{\text{left}} \coloneqq \mathbb{E}_{h\sim P_n}\left[\frac{1}{n}\sum_{i=1}^m \left(\ell_h(Z_i) - \mathbb{E}_{h'\sim Q(Z_{>i})}\left[\ell_{h'}(Z_i)\right]\right)^2\right].$$

Letting $X_h(Z) \coloneqq \ell_h(Z) - \ell_{h_*}(Z)$ and applying (48) with $a = \ell_h(Z_i)$, $c = \ell_{h_*}(Z_i)$, and $d = \mathbb{E}_{h'\sim Q(Z_{>i})}\left[\ell_{h'}(Z_i)\right] \overset{*}{=} \mathbb{E}_{h'\sim Q(Z_{>m})}\left[\ell_{h'}(Z_i)\right]$ (where $\overset{*}{=}$ is due to our assumption on Q), we get:

$$\begin{aligned} V_n^{\text{left}} &\leq \mathbb{E}_{h\sim P_n}\left[\frac{2}{n}\sum_{i=1}^m X_h(Z_i)^2\right] + \frac{2}{n}\sum_{i=1}^m \left(\mathbb{E}_{h'\sim Q(Z_{>m})}\left[\ell_{h'}(Z_i)\right] - \ell_{h_*}(Z_i)\right)^2, \\ &\leq \mathbb{E}_{h\sim P_n}\left[\frac{2}{n}\sum_{i=1}^m X_h(Z_i)^2\right] + \mathbb{E}_{h\sim Q(Z_{>m})}\left[\frac{2}{n}\sum_{i=1}^m X_h(Z_i)^2\right]. \quad \text{(by Jensen's Inequality)} \end{aligned} \tag{49}$$

Let $i \in [m]$, $h \in \mathcal{H}$, and $\eta \in]0, 1/b[$. Under the (β, B)-Bernstein condition, Lemmas 15-17 imply,

$$\alpha_\eta \cdot X_h(Z_i)^2 \trianglelefteq_\eta X_h(Z_i) + \left(1 + \tfrac{b}{2}\right)(2B\eta)^{\frac{1}{1-\beta}}, \tag{50}$$

where $\alpha_\eta \coloneqq \eta/(1+\sqrt{1+4\eta^2})$. Now, due to the Bernstein inequality (17), we have

$$\begin{aligned} X_h(Z_i) &\trianglelefteq_\eta \mathbb{E}_{Z_i'\sim D}\left[X_h(Z_i')\right] + s_\eta \cdot \mathbb{E}_{Z_i'\sim D}\left[X_h(Z_i')^2\right], \quad \text{where } s_\eta \coloneqq \eta \cdot \kappa(\eta b), \\ &\trianglelefteq \mathbb{E}_{Z_i'\sim D}\left[X_h(Z_i')\right] + s_\eta \cdot \mathbb{E}_{Z_i'\sim D}\left[X_h(Z_i')\right]^\beta, \quad \text{(by the Bernstein condition)} \\ &\trianglelefteq_\eta 2\mathbb{E}_{Z_i'\sim D}\left[X_h(Z_i')\right] + a_\beta^{\frac{\beta}{1-\beta}} \cdot (s_\eta)^{\frac{1}{1-\beta}}, \quad \text{where } a_\beta \coloneqq (1-\beta)^{1-\beta}\beta^\beta. \end{aligned} \tag{51}$$

The last inequality follows by the fact that $z^\beta = a_\beta \cdot \inf_{\nu>0}\{z/\nu + \nu^{\frac{\beta}{1-\beta}}\}$, for $z \geq 0$ (in our case, we set $\nu = a_\beta \cdot s_\eta$ to get to (51)). By chaining (50) with (51) using Proposition 10-(a), we get:

$$\begin{aligned}\alpha_\eta \cdot X_h(Z_i)^2 &\trianglelefteq_{\frac{\eta}{2}} 2\mathbb{E}_{Z_i'\sim D}\left[X_h(Z_i')\right] + a_\beta^{\frac{\beta}{1-\beta}} \cdot (s_\eta)^{\frac{1}{1-\beta}} + \left(1+\tfrac{b}{2}\right)(2B\eta)^{\frac{1}{1-\beta}}. \\ &\trianglelefteq_{\frac{\eta}{2}} 2\mathbb{E}_{Z_i'\sim D}\left[X_h(Z_i')\right] + \mathcal{P}\cdot\eta^{\frac{1}{1-\beta}}, \text{ with } \mathcal{P} \coloneqq a_\beta^{\frac{\beta}{1-\beta}} + \left(1+\tfrac{b}{2}\right)(2B)^{\frac{1}{1-\beta}}, \end{aligned} \tag{52}$$

where in the last inequality we used $\kappa(1) \leq 1$. Since (52) holds for all $h \in \mathcal{H}$, it still holds in expectation over $\mathcal{H}$ with respect to the distribution $Q(Z_{>m})$ (recall that $i \leq m$);

$$\alpha_\eta \cdot \mathbb{E}_{h\sim Q(Z_{>m})}\left[X_h(Z_i)^2\right] \trianglelefteq_{\frac{\eta}{2}} 2\mathbb{E}_{h\sim Q(Z_{>m})}\left[\mathbb{E}_{Z_i'\sim D}\left[X_h(Z_i')\right]\right] + \mathcal{P}\cdot\eta^{\frac{1}{1-\beta}}. \tag{53}$$

Since the samples $Z_{\leq n}$ are i.i.d, we have $\mathbb{E}_{Z_i\sim D}\left[\ell_h(Z_i)\right] = \mathbb{E}_{Z_j\sim D}\left[\ell_h(Z_j)\right]$, for all $i, j \in [m]$. Thus, after summing (52) and (53), for $i = 1, \ldots, m$, using Proposition 10-(b) and dividing by n, we get

$$\frac{\alpha_\eta}{n}\sum_{i=1}^m X_h(Z_i)^2 \trianglelefteq_{\frac{n\cdot\eta}{2}} \mathbb{E}_{Z\sim D}\left[X_h(Z)\right] + \frac{\mathcal{P}}{2}\cdot\eta^{\frac{1}{1-\beta}}, \tag{54}$$

$$\mathbb{E}_{h\sim Q(Z_{>m})}\left[\frac{\alpha_\eta}{n}\sum_{i=1}^m X_h(Z_i)^2\right] \trianglelefteq_{\frac{n\cdot\eta}{2}} \mathbb{E}_{h\sim Q(Z_{>m})}\left[\mathbb{E}_{Z\sim D}\left[X_h(Z)\right]\right] + \frac{\mathcal{P}}{2}\cdot\eta^{\frac{1}{1-\beta}}. \quad (m = n/2) \tag{55}$$

Now we apply PAC-Bayes (Proposition 11) to (54), with prior $P(Z_{>m})$ and posterior P_n, and obtain:

$$\mathbb{E}_{h\sim P_n}\left[\frac{\alpha_\eta}{n}\sum_{i=1}^m X_h(Z_i)^2\right] \trianglelefteq_{\frac{n\cdot\eta}{2}} \mathbb{E}_{h\sim P_n}\left[\mathbb{E}_{Z\sim D}\left[X_h(Z)\right]\right] + \frac{\mathcal{P}}{2}\cdot\eta^{\frac{1}{1-\beta}} + \frac{2\mathrm{KL}(P_n\|P(Z_{>m}))}{\eta\cdot n}. \tag{56}$$

Note that the upper-bound on V_n^{left} in (49) is the sum of the left-hand sides of (55) and (56) divided by $\alpha_\eta/2$. From now on, we restrict η to the range $]0, 1/(2b)[$ and define

$$\mathcal{A}_\eta \coloneqq \tfrac{2c_\eta}{\alpha_\eta} \leq 2\vartheta\left(\tfrac{1}{2}\right)\cdot\left(1+\sqrt{1+\tfrac{1}{b^2}}\right) \eqqcolon \mathcal{A}, \qquad \eta \in \left]0, \tfrac{1}{2b}\right[.$$

Chaining (55) and (56) using Proposition 10-(a) and multiplying throughout by $\mathcal{A}_\eta$, yields

$$c_\eta \cdot V_n^{\text{left}} \trianglelefteq_{\frac{n\eta}{4\mathcal{A}_\eta}} \mathcal{A}\cdot\left(\bar{L}(P_n) + \bar{L}(Q(Z_{>m}))\right) + \mathcal{P}\mathcal{A}\eta^{\frac{1}{1-\beta}} + \frac{2\mathcal{A}\cdot\mathrm{KL}(P_n\|P(Z_{>m}))}{\eta\cdot n}. \tag{57}$$

By a symmetric argument, a version of (57), with $Q(Z_{>m})$ [resp. $P(Z_{>m})$] replaced by $Q(Z_{\leq m})$ [resp. $P(Z_{\leq m})$], holds for $V_n^{\text{right}} \coloneqq V_n - V_n^{\text{left}}$. Using Proposition 10-(a) again, to chain the ESI inequalities of $c_\eta \cdot V_n^{\text{left}}$ and $c_\eta \cdot V_n^{\text{right}}$, we obtain:

$$c_\eta \cdot V_n \trianglelefteq_{\frac{n\eta}{8\mathcal{A}_\eta}} \mathcal{A}\cdot\left(2\bar{L}(P_n) + \bar{L}(Q_{\leq m}) + \bar{L}(Q_{>m})\right) + 2\mathcal{P}\mathcal{A}\eta^{\frac{1}{1-\beta}} + \frac{2\mathcal{A}\cdot\mathrm{COMP}_n}{\eta\cdot n}, \tag{58}$$

where $Q_{>m} \coloneqq Q(Z_{>m})$ and $Q_{\leq m} \coloneqq Q(Z_{\leq m})$. Let $\delta \in]0, 1[$, and π and $\mathcal{G}$ be as in (8). Applying Proposition 12 to (58) to obtain the corresponding ESI inequality with a random estimator $\hat{\eta} = \hat{\eta}(Z_{\leq n})$ with support on $\mathcal{G}$, and then applying Proposition 9, we get, with probability at least $1-\delta$,

$$c_{\hat{\eta}} \cdot V_n \leq \mathcal{A}\cdot\left(2\bar{L}(P_n) + \bar{L}(Q_{\leq m}) + \bar{L}(Q_{>m})\right) + 2\mathcal{P}\mathcal{A}\hat{\eta}^{\frac{1}{1-\beta}} + \frac{2\mathcal{A}\cdot\mathrm{COMP}_n + 8\mathcal{A}\ln\frac{|\mathcal{G}|}{\delta}}{\hat{\eta}\cdot n}. \tag{59}$$

Now adding $(\mathrm{COMP}_n + \varepsilon_{\delta,n})/(\hat{\eta}\cdot n)$ on both sides of (59) and choosing the estimator $\hat{\eta}$ optimally in the closure of $\mathcal{G}$ yields the desired result. □

F Proof of Lemma 13

Proof. Part (a) of the lemma was shown in the main body of the paper[2]. Thus, we only prove part (b); we will show a slight extension, namely that for all $0 < u < 1$, for all $\beta > 0, u > 0$,

$$\sup_{\rho\leq u}\ \sup_{P:\mathbb{E}_P[X]=\rho, P(X\leq u)=1} \mathbb{E}_{X\sim P}\left[e^{\beta E[X]-X-cX^2}\right] > 1 \text{ if } 0 < c < \vartheta(u) \text{ or } \beta \neq 1.$$

[2]The proof was inspired by the proof of Theorem 4 in [21].

The statement of the lemma (16) follows as the special case for $\beta = 1$, by replacing X by ηX and setting u to $u := \eta b < 1$.

We prove this by considering the set of distributions satisfying the constraint $\mathbb{E}[X] = \rho$ that are supported on at most two points,

$$\mathcal{P}_{\underline{x},\rho,\bar{x},u} = \{P : P\{\underline{x}\} + P\{\bar{x}\} = 1; \mathbb{E}_P[X] = \rho, \underline{x} \le \bar{x} \le u\},$$

and showing that

$$\sup_{\rho \le u} \sup_{P \in \mathcal{P}_{\underline{x},\rho,\bar{x},u}} g_{c,\beta}(P), \quad \text{with} \quad g_{c,\beta}(P) := \mathbb{E}_{X\sim P}\left[e^{\beta\rho - X - cX^2}\right]$$

is larger than 1. We first show that , for any $\beta \neq 1$, we can choose such a P such that $\sup_{P \in \mathcal{P}_{\underline{x},\rho,\bar{x},u}} g_{c,\beta}(P) > 1$. To see this, write $g_{c,\beta}(P)$ as

$$p \cdot e^{-\underline{x}+\beta\rho - c\underline{x}^2} + (1-p)e^{-\bar{x}+\beta\rho - c\bar{x}^2}$$

with $\rho = \mathbb{E}_P[X]$. We need to maximize this over $\rho = p\underline{x} + (1-p)\bar{x}$, so that in the end, we want to maximize over $0 \le p \le 1, \underline{u} \le \underline{x} \le \bar{x} \le u$, the expression

$$p \cdot e^{-\underline{x}+\beta(p\underline{x}+(1-p)\bar{x}) - c\underline{x}^2} + (1-p)e^{-\bar{x}+\beta(p\underline{x}+(1-p)\bar{x}) - c\bar{x}^2}$$

Now we write $\underline{x} = \bar{x} - a$ for some $a \ge 0$. The expression becomes

$$p \cdot e^{-\beta pa + (\beta-1)\bar{x} + a - c(\bar{x}-a)^2} + (1-p) \cdot e^{-\beta pa + (\beta-1)\bar{x} - c\bar{x}^2}$$

which is equal to

$$f(p,a,\bar{x}) := e^{-c\bar{x}^2 - \beta pa + (\beta-1)\bar{x}}\left(pe^{a+2ca\bar{x}-ca^2} + 1 - p\right) \underset{\text{if } \beta = 1}{=} e^{c\bar{x}^2 - pa}\left(pe^{a+2ca\bar{x}-ca^2} + 1 - p\right),$$

where the dependency of f on c and β is suppressed in the notation. At $p = 1$ and $p = 0$, this simplifies to (using also $\underline{x}$ again)

$$f(1,a,\bar{x}) = e^{-c\bar{x}^2 - \beta a + (\beta-1)\bar{x}}\left(e^{a+2ca\bar{x}-ca^2}\right) = e^{-c\underline{x}^2 + (\beta-1)\underline{x}} \underset{\text{if } \beta = 1}{=} e^{-c\underline{x}^2}$$

$$f(0,a,\bar{x}) = e^{-c\bar{x}^2 + (\beta-1)\bar{x}} \underset{\text{if } \beta = 1}{=} e^{-c\bar{x}^2}.$$

If $\beta < 1$, we can choose $\underline{x} = \bar{x} - a$ negative yet very close to 0 making $f(1,a,\bar{x}) > 1$; if $\beta > 1$, we can choose $\bar{x}$ positive yet very close to 0 making $f(0,a,\bar{x}) > 1$. Thus, $\sup g_{c,\beta}(P)$ can be made larger than 1 by P satisfying the constraint if $\beta \neq 1$. This shows (F) for the case $\beta \neq 1$. Hence, from now on we restrict to the case $\beta = 1$; we will further restrict to $\underline{x}$ and $\bar{x}$ such that $\underline{x} \le 0 \le \bar{x}$ so $\bar{x} \le a$. We will determine the maximum over (F) for $a \ge \bar{x}$ and $0 \le p \le 1$, for each given $0 \le \bar{x} \le u$. The partial derivatives to p and a are:

$$\begin{aligned}
\frac{\partial}{\partial p} f(p,a,\bar{x}) &= e^{-c\bar{x}^2 - pa}\left(\left(e^{a+2ca\bar{x}-ca^2} - 1\right) - a \cdot \left(pe^{a+2ca\bar{x}-ca^2} + (1-p)\right)\right) \\
&= e^{-c\bar{x}^2 - pa}\left(e^{a+2ca\bar{x}-ca^2}(1-ap) - 1 - a + ap\right) \\
\frac{\partial}{\partial a} f(p,a,\bar{x}) &= -p \cdot e^{-c\bar{x}^2 - pa}\left(pe^{a+2ca\bar{x}-ca^2} + (1-p)\right) + \\
&\quad + e^{-c\bar{x}^2 - pa} \cdot p \cdot e^{a+2ca\bar{x}-ca^2} \cdot (1 + 2c\bar{x} - 2ca) \\
&= p(1-p) \cdot e^{-c\bar{x}^2 - pa} \cdot \left(-1 + e^{a+2ca\bar{x}-ca^2}(1 + 2c\frac{\bar{x}-a}{1-p})\right).
\end{aligned}$$

At $a = \bar{x}$ (*i.e.* $\underline{x} = 0$), $f(p,a,\bar{x})$ simplifies to

$$f(p,\bar{x},\bar{x}) = e^{-c\bar{x}^2 - p\bar{x}} \cdot (pe^{\bar{x}+c\bar{x}^2} + (1-p)) \text{ so } f(1,\bar{x},\bar{x}) = 1$$

and the partial derivative to p at $(p,a,\bar{x}) = (1,\bar{x},\bar{x})$ becomes

$$e^{-c\bar{x}^2 - \bar{x}}\left((e^{\bar{x}+c(\bar{x})^2} - 1) - \bar{x}e^{\bar{x}+c(\bar{x})^2}\right) = 1 - e^{-c\bar{x}^2 - \bar{x}} - \bar{x}.$$

If (F) is negative, we can take $a = \bar{x}$ and p slightly smaller than 1 to get $f(p,a,\bar{x}) > 1$. This happens if and only if c is smaller than

$$\frac{-\ln(1-\bar{x}) - \bar{x}}{\bar{x}^2} = \vartheta(\bar{x}).$$

Thus, by taking $\underline{x} = 0$ and $\bar{x} = a = u$, and p slightly smaller than 1 again, we get $f(p,a,\bar{x}) > 1$ if $c < \vartheta(u)$; this shows (F) for the case $\beta = 1$; the result is proved. □

G Comparison Between "Bernstein" Inequalities

Discussion and Proof of Our Version of Bernstein's Inequality (17). Standard versions of Bernstein's inequality (see [12], and [15, Lemma 5.6]) can also be brought in ESI notation. In particular, compared with our version they express the inequality in terms of the random variable $Y = -X$, which is then upper bounded by b; more importantly, they have the second moment rather than the variance on the right-hand side, resulting in a slightly worse multiplicative factor $\kappa(2\eta b)$ instead of our $\kappa(\eta b)$; the proof is a standard one (see [12, Lemma A.4]) with trivial modifications: let $U \coloneqq \eta X$ and $\bar{u} \coloneqq \eta b$. Since $\kappa(u)$ is nondecreasing in u and $U \le \bar{u}$, we have

$$\frac{e^U - U - 1}{U^2} \le \frac{e^{\bar{u}} - \bar{u} - 1}{\bar{u}^2},$$

and hence $e^U - U - 1 \le \kappa(\bar{u})U^2$. Taking expectation on both sides and using that $\ln \mathbb{E}[e^U] \le \mathbb{E}[U] - 1$, we get $\ln \mathbb{E}\left[e^U\right] - \mathbb{E}[U] \le \kappa(\bar{u})\mathbb{E}[U^2]$. The result follows by exponentiating, rearranging, and using the ESI definition.

Comparison Between Un-expected and Empirical Bernstein Inequalities. The proof of the following proposition demonstrates how the un-expected Bernstein inequality in Lemma 13 together with the standard Bernstein inequality (17) imply a version of the empirical Bernstein inequality in [27, Theorem 4] with slightly worse factors. However, the latter inequality cannot be used to derive our main result — we do really require our new inequality to show Theorem 3, since we need to "chain" it to work with samples of length n rather than 1 in a different way. In the next proposition, we will use the following grid $\mathcal{G}$ and distribution π,

$$\mathcal{G} \coloneqq \left\{\tfrac{1}{\nu}, \dots, \tfrac{1}{\nu^K} : K \coloneqq \left\lceil \log_\nu \left(\sqrt{\tfrac{n}{2\ln\frac{2}{\delta}}}\right)\right\rceil\right\}, \quad \text{and} \quad \pi = \text{uniform distribution over } \mathcal{G}. \tag{60}$$

for $\nu > 0$. To simplify the presentation, we will use $\nu = 2$ in the next proposition, albeit this may not be the optimal choice.

Proposition 18. *Let $\mathcal{G}$ be as in (60) with $\rho = 2$, and $Z, Z_1, \dots, Z_n$ be i.i.d random variables taking values in $[0,1]$. Then, for all $\delta \in]0,1[$, with probability at least $1-\delta$,*

$$\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^n Z_i \le \left(3\sqrt{\frac{\mathrm{Var}_n[Z]\cdot \ln\frac{2|\mathcal{G}|}{\delta}}{2n}} + \frac{11\ln\frac{2|\mathcal{G}|}{\delta}}{10n}\right) \vee \frac{11\ln\frac{2|\mathcal{G}|}{\delta}}{4n} + \frac{c_{1/2}\cdot\ln\frac{2}{\delta}}{2n},$$

where $\mathrm{Var}_n[Z] \coloneqq \frac{1}{n}\sum_{i=1}^n \left(Z_i - \frac{1}{n}\sum_{j=1}^n Z_j\right)^2$ is the empirical variance, $c_{1/2} \coloneqq \vartheta(1/2)/2$, and ϑ as in Lemma 13.

Proof. Let $\delta \in]0,1[$. Applying Lemma 13 to $X_i = Z_i - \mathbb{E}[Z]$, for $i \in [n]$, we get, for all $0 < \eta < 1/2$,

$$\mathbb{E}[Z] - Z_i \trianglelefteq_\eta c_\eta \cdot (Z_i - \mathbb{E}[Z])^2, \quad \text{where } c_\eta \coloneqq \eta\cdot\vartheta(\eta). \tag{61}$$

Applying Proposition 10-(b) to chain (61) for $i = 1, \dots, n$, then dividing by n yields

$$\begin{aligned}\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^n Z_i &\trianglelefteq_{n\eta} \frac{c_\eta}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z])^2, \\ &= c_\eta\cdot\mathrm{Var}_n[Z] + c_\eta\cdot\left(\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^n Z_i\right)^2,\end{aligned} \tag{62}$$

where the equality follows from the standard bias-variance decomposition. Let $\mathcal{G}$ and π be as in (60), and let $\hat{\eta} = \hat{\eta}(Z_{\le n})$ be any random estimator with support on $\mathcal{G}$. By Proposition 12, a version of (62) with η is replaced by $\hat{\eta}$ and $\ln(|\mathcal{G}|)/(n\hat{\eta})$ added to its RHS also holds. By applying Proposition 9 to this new inequality, we get, with probability at least $1-\delta$,

$$\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^n Z_i \le c_{\hat{\eta}}\cdot\mathrm{Var}_n[Z] + \frac{\ln\frac{|\mathcal{G}|}{\delta}}{n\cdot\hat{\eta}} + c_{\hat{\eta}}\cdot\left(\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^n Z_i\right)^2. \tag{63}$$

Now using Hoeffding's inequality [27, Theorem 3], we also have

$$\left(\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right)^2 \leq \frac{\ln\frac{1}{\delta}}{2n}, \tag{64}$$

with probability at least $1-\delta$. Thus, by combining (63) and (64) via the union bound, we get that, with probability at least $1-\delta$,

$$\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i \leq \left(c_{\hat{\eta}} \cdot \mathrm{Var}_n[Z] + \frac{\ln\frac{2|\mathcal{G}|}{\delta}}{n \cdot \hat{\eta}}\right) + \frac{c_{\hat{\eta}} \cdot \ln\frac{2}{\delta}}{2n}. \tag{65}$$

We now use the fact that for all $\eta \in]0, 1/2[$,

$$c_\eta = \eta \cdot \vartheta(\eta) \leq \frac{\eta}{2} + \frac{11\eta^2}{20}. \tag{66}$$

Let $\hat{\eta}_* \in [0, +\infty]$ be the un-constrained estimator defined by

$$\hat{\eta}_* := \sqrt{\frac{2\ln\frac{2|\mathcal{G}|}{\delta}}{\mathrm{Var}_n[Z] \cdot n}}.$$

Note that by our choice of $\mathcal{G}$ in (60), we always have $\hat{\eta}_* \geq \min \mathcal{G}$. Let $\hat{\eta} \in ([\hat{\eta}_*/2, \hat{\eta}_*] \cap \mathcal{G}) \neq \varnothing$, if $\hat{\eta}_* \leq 1$, and $\hat{\eta} = 1/2$, otherwise. In the first case (*i.e.* when $\hat{\eta}_* \leq 1$), substituting η for $\hat{\eta} \in ([\hat{\eta}_*/2, \hat{\eta}_*] \cap \mathcal{G})$ in the expression between brackets in (65), and using the fact that $\hat{\eta}_*/2 \leq \hat{\eta} \leq \hat{\eta}_*$ and (66), gives

$$c_{\hat{\eta}} \cdot \mathrm{Var}_n[Z] + \frac{\ln\frac{2|\mathcal{G}|}{\delta}}{\hat{\eta} \cdot n} \leq (1+2)\sqrt{\frac{\mathrm{Var}_n[Z] \cdot \ln\frac{2|\mathcal{G}|}{\delta}}{2n}} + \frac{11 \cdot \ln\frac{2|\mathcal{G}|}{\delta}}{10n}. \tag{67}$$

Now for the case where $\hat{\eta}_* \geq 1$, we substitute η for $\hat{\eta} = 1/2$ in the expression between brackets in (65), and use (66) and the fact that $1 \leq \hat{\eta}_* = \sqrt{2\ln(2|\mathcal{G}|/\delta)/(\mathrm{Var}_n[Z] \cdot n)}$, we get:

$$\begin{aligned} c_{\hat{\eta}} \cdot \mathrm{Var}_n[Z] + \frac{\ln\frac{2|\mathcal{G}|}{\delta}}{\hat{\eta} \cdot n} &\leq \left(\frac{\hat{\eta}}{2} + \frac{11\hat{\eta}^2}{20}\right) \cdot \mathrm{Var}_n[Z] + \frac{2 \cdot \ln\frac{2|\mathcal{G}|}{\delta}}{n}, \\ &\leq \left(\frac{\hat{\eta}}{2} + \frac{11\hat{\eta}^2}{20}\right) \cdot \frac{2\ln\frac{2|\mathcal{G}|}{\delta}}{n} + \frac{2 \cdot \ln\frac{2|\mathcal{G}|}{\delta}}{n}, \quad \text{(due to } \hat{\eta}_* \geq 1) \\ &= \frac{11\ln\frac{2|\mathcal{G}|}{\delta}}{4n}, \quad (\hat{\eta} = 1/2) \end{aligned} \tag{68}$$

Combining (65), with (67) and (68) yields the desired results. □

H Additional Experiments

H.1 Informed Priors

In this section, we run the same experiments as in Section 4 of the main body, except for the following changes

- For Maurer's bound, we use the version in our Lemma 4 with informed priors.
- For the TS and Catoni bounds, we build a prior from the first half of the data (*i.e.* we replace P_0 by $P(Z_{\leq m})$, where $m = n/2$) and use it to evaluate the bounds on the second half of the data. In this case, the "posterior" distribution is $P(Z_{>m})$, and thus the term $\mathrm{KL}(P_n \| P_0)$ is replaced by $\mathrm{KL}(P(Z_{>m}) \| P(Z_{\leq m}))$.

Recall that $P(Z_{>m}) \equiv \mathcal{N}(\hat{h}(Z_{>m}), \sigma^2 I_d)$, $P(Z_{\leq m}) \equiv \mathcal{N}(\hat{h}(Z_{\leq m}), \sigma^2 I_d)$, and $P(Z_{\leq n}) \equiv \mathcal{N}(\hat{h}(Z_{\leq n}), \sigma^2 I_d)$, where the variance σ^2 is learned from a geometric grid (see Section 4); our own bound is not affected by any of these changes. The results for the synthetic and UCI datasets are reported in Figure 2 and Table 2, respectively.

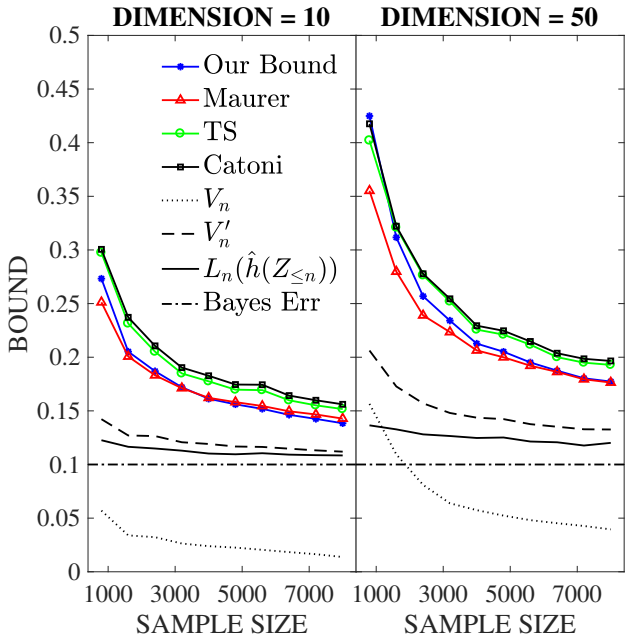

Figure 2: Results for the synthetic data with informed priors.

Dataset	n	d	Test error of $\hat{h}$	Our	Maurer	TS	Catoni
Haberman	244	3	0.272	0.52	0.459	0.501	0.55
Breast-C.	560	9	0.068	0.185	0.164	0.215	0.219
Tic-Tac-Toe	766	27	0.046	0.19	0.152	0.202	0.199
Bank-note	1098	4	0.058	0.125	0.117	0.136	0.143
kr-vs-kp	2556	73	0.044	0.107	0.102	0.123	0.127
Spam-base	3680	57	0.173	0.293	0.284	0.317	0.323
Mushroom	6500	116	0.002	0.018	0.016	0.023	0.024
Adult	24130	108	0.168	0.195	0.198	0.2	0.203

Table 2: Results for the UCI datasets.

Though our bound still performs better than Catoni's and TS, Maurer's bound in Lemma 4 tends to be slightly tighter than ours, especially when the sample size is small. We note, however, that the advantage of our bound has not been fully leveraged here; our bound in its full generality in Theorem 3 allows one to use "online posteriors" $(Q(Z_{>i}))$ and $(Q(Z_{<j}))$ in the V_n term which converge to the one based on the full sample, *i.e.* $Q(Z_{\leq n})$. We expect this to substantially improve our bound. However, we did not experiment with this due to computational reasons.

H.2 Maurer's Bound: Informed Versus Uninformed Priors

In this section, we compare the performance of Maurer's bound with and without informed priors (*i.e.* (2) and (9), respectively) on synthetic data in the same setting as Section 4. From Figure 3, we see that using informed priors as in Lemma 4 substantially improves Maurer's bound.

H.3 Varying the Bayes Error and Bayes Act

In this subsection, we run the same synthetic experiment as in Subsection (H.1) (*i.e.* using informed priors for all bounds), except for the following changes:

- We vary the Bayes error by varying the level of noise: we flip the labels with probability either 0.05, 0.1, or 0.2 (note that in Section 4 we flipped labels with probability 0.1).

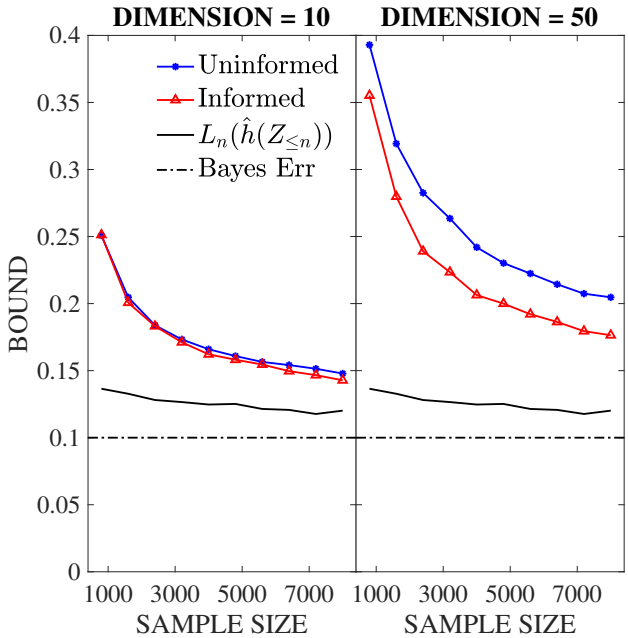

Figure 3: Results for the synthetic data: (Blue curve) Uninformed Maurer's bound (2); (Red curve) Informed Maurer's bound (9).

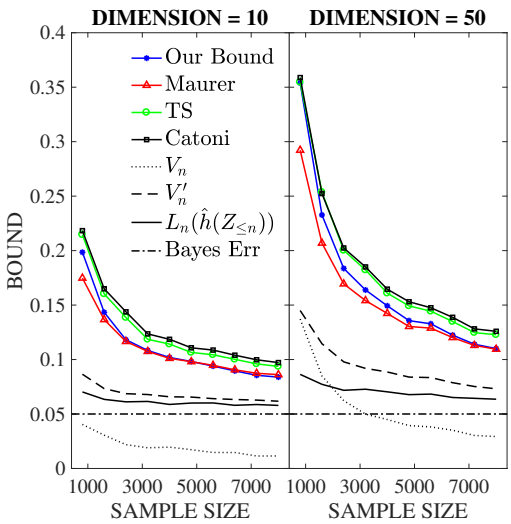

Figure 4: Results for the synthetic data with informed priors, randomly generated Bayes act, and Bayes error set to 0.05.

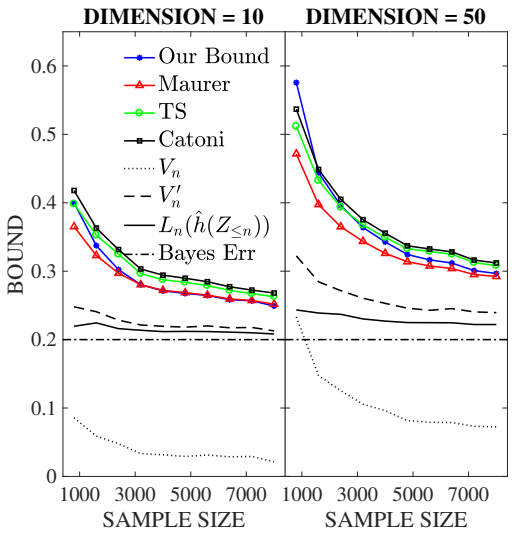

Figure 5: Results for the synthetic data with informed priors, randomly generated Bayes act, and Bayes error set to 0.2.

- In each case, we generate the synthetic data using a randomly generated h_* with coordinates uniformly sampled in the interval $[0, 1]$. The reported results in Figures 4-6 are averages over 10 runs for each tested sample size.

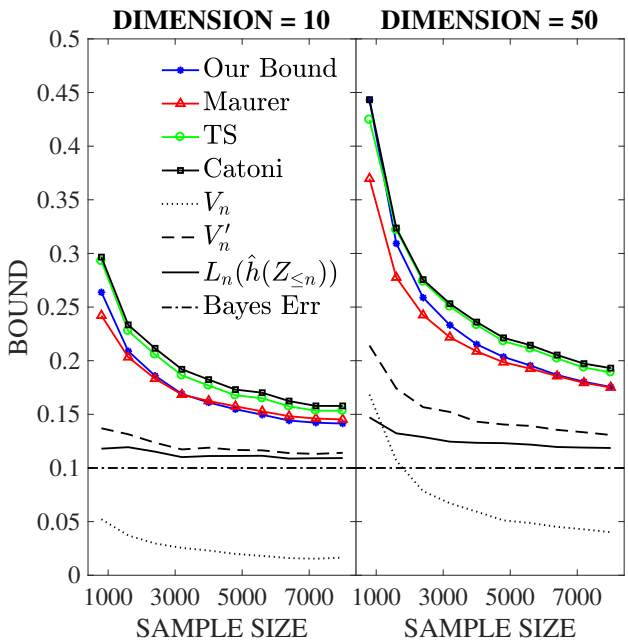

Figure 6: Results for the synthetic data with informed priors, randomly generated Bayes act, and Bayes error set to 0.1.

Chapter 6

PAC-Bayesian Bound for the Conditional Value at Risk

When talking about generalization bounds, it is natural to automatically think about bounding the difference between the *expected* risk and its empirical version. However, in some machine learning applications, the mean performance of an algorithm may not be the best objective. Such settings include prediction tasks where a mistake implies a disastrous consequence. If such mistakes happen with low enough probability, they may not be effectively captured by the mean performance. As a result, there is growing interest in working with alternative measures of risk which can better capture the "worst" events [Huo and Fu, 2017; Bhat and Prashanth, 2019; Williamson and Menon, 2019; Chen et al., 2009; Chow and Ghavamzadeh, 2014; Prashanth and Ghavamzadeh, 2013; Tamar et al., 2015; Pinto et al., 2017; Morimura et al., 2010; Takeda and Kanamori, 2009]. A prime candidate for this purpose is the Conditional Value at Risk (CVaR). $\mathrm{CVaR}_\alpha[X]$ measures the conditional expectation of a real random variable X conditioned on the event that X is greater than its $(1-\alpha)$-quantile. CVaR has recently been used in many risk-averse applications, including bandits, reinforcement learning, and fairness.

From a statistical learning perspective, it is desirable to have generalization bounds for algorithms when the objective is the CVaR of a loss. Generalization bounds for the expected risk are obtained via concentration inequalities. However, these inequalities cannot directly be used to estimate CVaR due to its non-linearity as a function of the data-generating distribution. For this reason, generalization bounds for CVaR are much harder to come by. In this chapter, we reduce the problem of estimating CVaR to that of estimating the standard expectation from empirical means. Thanks to this, we derived new tight concentration inequalities for CVaR and the *first* PAC-Bayesian bound when learning with a CVaR objective. The bound in question is data-dependent and becomes small whenever the empirical risk (empirical CVaR in this case) is small—a property only shared with state-of-the-art PAC-Bayesian bounds for the expected risk [Langford and Shawe-Taylor, 2003; Maurer, 2004].

PAC-Bayesian Bound for the Conditional Value at Risk

Zakaria Mhammedi
The Australian National University and Data61
zak.mhammedi@anu.edu.au

Benjamin Guedj
Inria and University College London
benjamin.guedj@inria.fr

Robert C. Williamson
bobwilliamsonoz@icloud.com

Abstract

Conditional Value at Risk (CVaR) is a family of "coherent risk measures" which generalize the traditional mathematical expectation. Widely used in mathematical finance, it is garnering increasing interest in machine learning, e.g., as an alternate approach to regularization, and as a means for ensuring fairness. This paper presents a generalization bound for learning algorithms that minimize the CVaR of the empirical loss. The bound is of PAC-Bayesian type and is guaranteed to be small when the empirical CVaR is small. We achieve this by reducing the problem of estimating CVaR to that of merely estimating an expectation. This then enables us, as a by-product, to obtain concentration inequalities for CVaR even when the random variable in question is unbounded.

1 Introduction

The goal in statistical learning is to learn hypotheses that generalize well, which is typically formalized by seeking to minimize the *expected* risk associated with a given loss function. In general, a loss function is a map $\ell\colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, where $\mathcal{X}$ is a feature space and $\mathcal{H}$ is an hypotheses space. In this case, the *expected risk* associated with a given hypothesis $h \in \mathcal{H}$ is given by $\mathrm{R}[\ell(h, X)] \coloneqq \mathrm{E}[\ell(h, X)]$. Since the data-generating distribution is typically unknown, the expected risk is approximated using observed i.i.d. samples $X_1, \ldots, X_n$ of X, and an hypothesis is then chosen to minimize the empirical risk $\widehat{\mathrm{R}}[\ell(h, X)] \coloneqq \sum_{i=1}^n \ell(h, X_i)/n$. When choosing an hypothesis $\hat{h}$ based on the empirical risk $\widehat{\mathrm{R}}$, one would like to know how close $\widehat{\mathrm{R}}[\ell(\hat{h}, X)]$ is to the actual risk $\mathrm{R}[\ell(\hat{h}, X)]$; only then can one infer something about the *generalization* property of the learned hypothesis $\hat{h}$.

Generalization bounds—in which the expected risk is bounded in terms of its empirical version up to some error—are at the heart of many machine learning problems. The main techniques leading to such bounds comprise uniform convergence arguments (often involving the Rademacher complexity of the set $\mathcal{H}$), algorithmic stability arguments (see e.g. [Bousquet and Elisseeff, 2002] and more recently the work from [Abou-Moustafa and Szepesvári, 2019, Bousquet et al., 2019, Celisse and Guedj, 2016]), and the PAC-Bayesian analysis for non-degenerate randomized estimators [McAllester, 2003]. Behind these techniques lies *concentration inequalities*, such as Chernoff's inequality (for the PAC-Bayesian analysis) and McDiarmid's inequality (for algorithmic stability and the uniform convergence analysis), which control the deviation between population and empirical averages [see Boucheron et al., 2003, 2013, McDiarmid, 1998, among others].

Standard concentration inequalities are well suited for learning problems where the goal is to minimize the expected risk $\mathrm{E}[\ell(h, X)]$. However, the expected risk—the mean performance of an algorithm—might fail to capture the underlying phenomenon of interest. For example, when

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

dealing with medical (responsitivity to a specific drug with grave side effects, etc.), environmental (such as pollution, exposure to toxic compounds, etc.), or sensitive engineering tasks (trajectory evaluation for autonomous vehicles, etc.), the mean performance is not necessarily the best objective to optimize as it will cover potentially disastrous mistakes (*e.g.*, a few extra centimeters when crossing another vehicle, a slightly too large dose of a lethal compound, etc.) while possibly improving on average. There is thus a growing interest to work with alternative measures of risk (other than the expectation) for which standard concentration inequalities do not apply directly. Of special interest are *coherent risk measures* [Artzner et al., 1999] which possess properties that make them desirable in mathematical finance and portfolio optimization [Allais, 1953, Ellsberg, 1961, Rockafellar et al., 2000], with a focus on optimizing for the worst outcomes rather than on average. Coherent risk measures have also been recently connected to fairness, and appear as a promising framework to control the fairness of an algorithm's solution [Williamson and Menon, 2019].

A popular coherent risk measure is the Conditional Value at Risk (CVaR; see Pflug, 2000); for $\alpha \in (0,1)$ and random variable Z, $\mathrm{CVaR}_\alpha[Z]$ measures the expectation of Z conditioned on the event that Z is greater than its $(1-\alpha)$-th quantile. CVaR has been shown to underlie the classical SVM [Takeda and Sugiyama, 2008], and has in general attracted a large interest in machine learning over the past two decades [Bhat and Prashanth, 2019, Chen et al., 2009, Chow and Ghavamzadeh, 2014, Huo and Fu, 2017, Morimura et al., 2010, Pinto et al., 2017, Prashanth and Ghavamzadeh, 2013, Takeda and Kanamori, 2009, Tamar et al., 2015, Williamson and Menon, 2019].

Various concentration inequalities have been derived for $\mathrm{CVaR}_\alpha[Z]$, under different assumptions on Z, which bound the difference between $\mathrm{CVaR}_\alpha[Z]$ and its standard estimator $\widehat{\mathrm{CVaR}}_\alpha[Z]$ with high probability [Bhat and Prashanth, 2019, Brown, 2007, Kolla et al., 2019, Prashanth and Ghavamzadeh, 2013, Wang and Gao, 2010]. However, none of these works extend their results to the statistical learning setting where the goal is to learn an hypothesis from data to minimize the conditional value at risk. In this paper, we fill this gap by presenting a sharp PAC-Bayesian generalization bound when the objective is to minimize the conditional value at risk.

Related Works. Deviation bounds for CVaR were first presented by Brown [2007]. However, their approach only applies to bounded continuous random variables, and their lower deviation bound has a sub-optimal dependence on the level α. Wang and Gao [2010] later refined their analysis to recover the "correct" dependence in α, albeit their technique still requires a two-sided bound on the random variable Z. Thomas and Learned-Miller [2019] derived new concentration inequalities for CVaR with a very sharp empirical performance, even though the dependence on α in their bound is sub-optimal. Further, they only require a one-sided bound on Z, without a continuity assumption.

Kolla et al. [2019] were the first to provide concentration bounds for CVaR when the random variable Z is unbounded, but is either sub-Gaussian or sub-exponential. Bhat and Prashanth [2019] used a bound on the Wasserstein distance between true and empirical cumulative distribution functions to substantially tighten the bounds of Kolla et al. [2019] when Z has finite exponential or kth-order moments; they also apply their results to other coherent risk measures. However, when instantiated with bounded random variables, their concentration inequalities have sub-optimal dependence in α.

On the statistical learning side, Duchi and Namkoong [2018] present generalization bounds for a class of coherent risk measures that technically includes CVaR. However, their bounds are based on uniform convergence arguments which lead to looser bounds compared with ours. Another bound based on a uniform convergence argument was also presented in a concurrent work by Curi et al. [2020].

Contributions. Given a learning algorithm that outputs a posterior distribution $\widehat{\rho}$ on $\mathcal{H}$, our main contribution is a *PAC-Bayesian generalization bound for the conditional value at risk*, where we bound the difference $\mathrm{CVaR}_\alpha[Z] - \widehat{\mathrm{CVaR}}_\alpha[Z]$, for $\alpha \in (0,1)$ and $Z \coloneqq \mathrm{E}_{h\sim\widehat{\rho}}[\ell(h,X)]$, by a term of order $\sqrt{\widehat{\mathrm{CVaR}}_\alpha[Z] \cdot \mathcal{K}_n/(n\alpha)}$, with $\mathcal{K}_n$ representing a complexity term which depends on $\mathcal{H}$. Due to the presence of $\widehat{\mathrm{CVaR}}_\alpha[Z]$ inside the square-root, our generalization bound has the desirable property that it becomes small whenever the empirical conditional value at risk is small. For the standard expected risk, only state-of-the-art PAC-Bayesian bounds share this property (see *e.g.* Catoni [2007], Langford and Shawe-Taylor [2003], Maurer [2004] or more recently in Mhammedi et al. [2019], Tolstikhin and Seldin [2013]). We refer to [Guedj, 2019] for a recent survey on PAC-Bayes.

As a by-product of our analysis, we derive a new way of obtaining *concentration bounds for the conditional value at risk* by reducing the problem to estimating expectations using empirical means. This reduction then makes it easy to obtain concentration bounds for $\mathrm{CVaR}_\alpha[Z]$ even when the random variable Z is unbounded (Z may be sub-Gaussian or sub-exponential). Our bounds have explicit constants and are sharper than existing ones due to Bhat and Prashanth [2019], Kolla et al. [2019] which deal with the unbounded case.

Outline. In Section 2, we define the conditional value at risk along with its standard estimator. In Section 3, we recall the statistical learning setting and present our PAC-Bayesian bound for CVaR. The proof of our main bound is in Section 4. In Section 5, we present a new way of deriving concentration bounds for CVaR which stems from our analysis in Section 4. Section 6 concludes and suggests future directions.

2 Preliminaries

Let $(\Omega, \mathcal{F}, P)$ be a probability space. For $p \in \mathbb{N}$, we denote by $\mathcal{L}^p(\Omega) \coloneqq \mathcal{L}^p(\Omega, \mathcal{F}, P)$ the space of p-integrable functions, and we let $\mathcal{M}_P(\Omega)$ be the set of probability measures on Ω which are absolutely continuous with respect to P. We reserve the notation E for the expectation under the reference measure P, although we sometimes write E_P for clarity. For random variables $Z_1, \dots, Z_n$, we denote $\widehat{P}_n \coloneqq \sum_{i=1}^n \delta_{Z_i}/n$ the empirical distribution, and we let $Z_{1:n} \coloneqq (Z_1, \dots, Z_n)$. Furthermore, we let $\pi \coloneqq (1, \dots, 1)^\top/n \in \mathbb{R}^n$ be the uniform distribution on the simplex. Finally, we use the notation $\widetilde{\mathcal{O}}$ to hide log-factors in the sample size n.

Coherent Risk Measures (CRM). A CRM [Artzner et al., 1999] is a functional $\mathrm{R}\colon \mathcal{L}^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ that is simultaneously, positive homogeneous, monotonic, translation equivariant, and sub-additive[1] (see Appendix B for a formal definition). For $\alpha \in (0,1)$ and a real random variable $Z \in \mathcal{L}_1(\Omega)$, the conditional value at risk $\mathrm{CVaR}_\alpha[Z]$ is a CRM and is defined as the mean of the random variable Z conditioned on the event that Z is greater than its $(1-\alpha)$-th quantile[2]. This is equivalent to the following expression, which is more convenient for our analysis:

$$\mathrm{CVaR}_\alpha[Z] = \mathrm{C}_\alpha[Z] \coloneqq \inf_{\mu \in \mathbb{R}} \left\{ \mu + \frac{\mathrm{E}[Z-\mu]_+}{\alpha} \right\}.$$

Key to our analysis is the *dual representation* of CRMs. It is known that any CRM $\mathrm{R}\colon \mathcal{L}^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ can be expressed as the *support function* of some closed convex set $\mathcal{Q} \subseteq \mathcal{L}^1(\Omega)$ [Rockafellar and Uryasev, 2013]; that is, for any real random variable $Z \in \mathcal{L}^1(\Omega)$, we have

$$\mathrm{R}[Z] = \sup_{q \in \mathcal{Q}} \mathrm{E}_P[Zq] = \sup_{q \in \mathcal{Q}} \int_\Omega Z(\omega) q(\omega) \mathrm{d}P(\omega). \quad \text{(dual representation)} \tag{1}$$

In this case, the set $\mathcal{Q}$ is called the *risk envelope* associated with the risk measure R. The risk envelope $\mathcal{Q}_\alpha$ of $\mathrm{CVaR}_\alpha[Z]$ is given by

$$\mathcal{Q}_\alpha \coloneqq \left\{ q \in \mathcal{L}^1(\Omega) \;\middle|\; \exists Q \in \mathcal{M}_P(\Omega),\ q = \frac{\mathrm{d}Q}{\mathrm{d}P} \le \frac{1}{\alpha} \right\}, \tag{2}$$

and so substituting $\mathcal{Q}_\alpha$ for $\mathcal{Q}$ in (1) yields $\mathrm{CVaR}_\alpha[Z]$.[3] Though the overall approach we take in this paper may be generalizable to other popular CRMs, (see Appendix B) we focus our attention on CVaR for which we derive new PAC-Bayesian and concentration bounds in terms of its natural estimator $\widehat{\mathrm{CVaR}}_\alpha[Z]$; given i.i.d. copies of $Z_1, \dots, Z_n$ of Z, we define

$$\widehat{\mathrm{CVaR}}_\alpha[Z] \coloneqq \widehat{\mathrm{C}}_\alpha[Z] \coloneqq \inf_{\mu \in \mathbb{R}} \left\{ \mu + \sum_{i=1}^n \frac{[Z_i - \mu]_+}{n\alpha} \right\}. \tag{3}$$

From now on, we write $\mathrm{C}_\alpha[Z]$ and $\widehat{\mathrm{C}}_\alpha[Z]$ for $\mathrm{CVaR}_\alpha[Z]$ and $\widehat{\mathrm{CVaR}}_\alpha[Z]$, respectively.

[1] These are precisely the properties which make coherent risk measures excellent candidates in some machine learning applications (see e.g. [Williamson and Menon, 2019] for an application to fairness)

[2] We use the convention in Brown [2007], Prashanth and Ghavamzadeh [2013], Wang and Gao [2010].

[3] The dual representation of CVaR was also leveraged in a concurrent work by Curi et al. [2020] for the purpose of efficiently optimizing CVaR objectives.

3 PAC-Bayesian Bound for the Conditional Value at Risk

In this section, we briefly describe the statistical learning setting, formulate our goal, and present our main results.

In the statistical learning setting, Z is a *loss* random variable which can be written as $Z = \ell(h, X)$, where $\ell\colon \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a loss function and $\mathcal{X}$ [resp. $\mathcal{H}$] is a feature [resp. hypotheses] space. The aim is to learn an hypothesis $\hat{h} = \hat{h}(X_{1:n}) \in \mathcal{H}$, or more generally a distribution $\widehat{\rho} = \widehat{\rho}(X_{1:n})$ over $\mathcal{H}$ (also referred to as randomized estimator), based on i.i.d. samples $X_1, \ldots, X_n$ of X which minimizes some measure of risk—typically, the expected risk $\mathrm{E}_P[\ell(\widehat{\rho}, X)]$, where $\ell(\widehat{\rho}, X) \coloneqq \mathrm{E}_{h\sim\widehat{\rho}}[\ell(h, X)]$.

Our work is motivated by the idea of replacing this expected risk by any coherent risk measure R. In particular, if $\mathcal{Q}$ is the risk envelope associated with R, then our quantity of interest is

$$\mathrm{R}[\ell(\widehat{\rho}, X)] \coloneqq \sup_{q\in\mathcal{Q}} \int_\Omega \ell(\widehat{\rho}, X(\omega))q(\omega)\mathrm{d}P(\omega).$$

Thus, given a consistent estimator $\widehat{\mathrm{R}}[\ell(\widehat{\rho}, X)]$ of $\mathrm{R}[\ell(\widehat{\rho}, X)]$ and some prior distribution ρ_0 on $\mathcal{H}$, our grand goal (which goes beyond the scope of this paper) is to bound the risk $\mathrm{R}[\ell(\widehat{\rho}, X)]$ as

$$\mathrm{R}[\ell(\widehat{\rho}, X)] \leq \widehat{\mathrm{R}}[\ell(\widehat{\rho}, X)] + \widetilde{\mathcal{O}}\left(\sqrt{\frac{\mathrm{KL}(\widehat{\rho}\|\rho_0)}{n}}\right), \tag{4}$$

with high probability. Based on (3), the consistent estimator we use for $\mathrm{C}_\alpha[\ell(\widehat{\rho}, X)]$ is

$$\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)] \coloneqq \inf_{\mu\in\mathbb{R}} \left\{\mu + \sum_{i=1}^n \frac{[\ell(\widehat{\rho}, X_i) - \mu]_+}{n\alpha}\right\}, \quad \alpha \in (0, 1). \tag{5}$$

This is in fact a consistent estimator (see e.g. [Duchi and Namkoong, 2018, Proposition 9]). As a first step towards the goal in (4), we derive a sharp PAC-Bayesian bound for the conditional value at risk, which we state now as our main theorem:

Theorem 1. *Let $\alpha \in (0, 1)$, $\delta \in (0, 1/2)$, $n \geq 2$, and $N \coloneqq \lceil \log_2(n/\alpha) \rceil$. Further, let ρ_0 be any distribution on an hypothesis set $\mathcal{H}$, $\ell\colon \mathcal{H} \times \mathcal{X} \to [0, 1]$ be a loss, and $X_1, \ldots, X_n$ be i.i.d. copies of X. Then, for any "posterior" distribution $\widehat{\rho} = \widehat{\rho}(X_{1:n})$ over $\mathcal{H}$, $\varepsilon_n \coloneqq \sqrt{\frac{2\ln(N/\delta)}{\alpha n}} + \frac{\ln(N/\delta)}{3\alpha n}$, and $\mathcal{K}_n \coloneqq \mathrm{KL}(\widehat{\rho}\|\rho_0) + \ln(N/\delta)$, we have, with probability at least $1 - 2\delta$.*

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]] \leq \widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)] + \sqrt{\frac{27\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)]\mathcal{K}_n}{5\alpha n}} + 2\varepsilon_n\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)] + \frac{27\mathcal{K}_n}{5n\alpha}. \tag{6}$$

Discussion of the bound. Although we present the bound for the bounded loss case, our result easily generalizes to the case where $\ell(h, X)$ is sub-Gaussian or sub-exponential, for all $h \in \mathcal{H}$. We discuss this in Section 5. Our second observation is that since $\mathrm{C}_\alpha[Z]$ is a coherent risk measure, it is convex in Z [Rockafellar and Uryasev, 2013], and so we can further bound the term $\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]]$ on the LHS of (6) from below by $\mathrm{C}_\alpha[\ell(\widehat{\rho}, X)] = \mathrm{C}_\alpha[\mathrm{E}_{h\sim\widehat{\rho}}[\ell(h, X)]]$. This shows that the type of guarantee we have in (6) is in general tighter than the one in (4).

Even though not explicitly done before, a PAC-Bayesian bound of the form (4) can be derived for a risk measure R using an existing technique due to McAllester [2003] as soon as, for any fixed hypothesis h, the difference $\mathrm{R}[\ell(h, X)] - \widehat{\mathrm{R}}[\ell(h, X)]$ is sub-exponential with a sufficiently fast tail decay as a function of n (see the proof of Theorem 1 in [McAllester, 2003]). While it has been shown that the difference $\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z]$ also satisfies this condition for bounded i.i.d. random variables $Z, Z_1, \ldots, Z_n$ (see *e.g.* Brown [2007], Wang and Gao [2010]), applying the technique of McAllester [2003] yields a bound on $\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h, X)]]$ (*i.e.* the LHS of (6)) of the form

$$\mathrm{E}_{h\sim\widehat{\rho}}[\widehat{\mathrm{C}}_\alpha[\ell(h, X)]] + \sqrt{\frac{\mathrm{KL}(\widehat{\rho}\|\rho_0) + \ln\frac{n}{\delta}}{\alpha n}}. \tag{7}$$

Such a bound is weaker than ours in two ways; (**I**) by Jensen's inequality the term $\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)]$ in our bound (defined in (5)) is always smaller than the term $\mathrm{E}_{h\sim\widehat{\rho}}[\widehat{\mathrm{C}}_\alpha[\ell(h, X)]]$ in (7); and (**II**) unlike in our bound in (6), the complexity term inside the square-root in (7) does not multiply the

empirical conditional value at risk $\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)]$. This means that our bound can be much smaller whenever $\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)]$ is small—this is to be expected in the statistical learning setting since $\widehat{\rho}$ will typically be picked by an algorithm to minimize the empirical value $\widehat{\mathrm{C}}_\alpha[\ell(\widehat{\rho}, X)]$. This type of improved PAC-Bayesian bound, where the empirical error appears multiplying the complexity term inside the square-root, has been derived for the expected risk in works such as [Catoni, 2007, Langford and Shawe-Taylor, 2003, Maurer, 2004, Seeger, 2002]; these are arguably the state-of-the-art generalization bounds.

A reduction to the expected risk. A key step in the proof of Theorem 1 is to show that for a real random variable Z (not necessarily bounded) and $\alpha \in (0,1)$, one can construct a function $g\colon \mathbb{R} \to \mathbb{R}$ such that the auxiliary variable $Y = g(Z)$ satisfies **(I)**

$$\mathrm{E}[Y] = \mathrm{E}[g(Z)] = \mathrm{C}_\alpha[Z]; \tag{8}$$

and **(II)** for i.i.d. copies $Z_{1:n}$ of Z, the i.i.d. random variables $Y_1 \coloneqq g(Z_1), \ldots, Y_n \coloneqq g(Z_n)$ satisfy

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \le \widehat{\mathrm{C}}_\alpha[Z](1+\epsilon_n), \quad \text{where } \epsilon_n = \widetilde{\mathcal{O}}(\alpha^{-1/2} n^{-1/2}), \tag{9}$$

with high probability. Thus, due to (8) and (9), bounding the difference

$$\mathrm{E}[Y] - \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{10}$$

is sufficient to obtain a concentration bound for CVaR. Since $Y_1, \ldots, Y_n$ are i.i.d., one can apply standard concentration inequalities, which are available whenever Y is sub-Gaussian or sub-exponential, to bound the difference in (10). Further, we show that whenever Z is sub-Gaussian or sub-exponential, then essentially so is Y. Thus, our method allows us to obtain concentration inequalities for $\widehat{\mathrm{C}}_\alpha[Z]$, even when Z is unbounded. We discuss this in Section 5.

4 Proof Sketch for Theorem 1

In this section, we present the key steps taken to prove the bound in Theorem 1. We organize the proof in three subsections. In Subsection 4.1, we introduce an auxiliary estimator $\widetilde{\mathrm{C}}_\alpha[Z]$ for $\mathrm{C}_\alpha[Z]$, $\alpha \in (0,1)$, which will be useful in our analysis; in particular, we bound this estimator in terms of $\widehat{\mathrm{C}}_\alpha[Z]$ (as in (9) above, but with the LHS replaced by $\widetilde{\mathrm{C}}_\alpha[Z]$). In Subsection 4.2, we introduce an auxiliary random variable Y whose expectation equals $\mathrm{C}_\alpha[Z]$ (as in (8)) and whose empirical mean is bounded from above by the estimator $\widetilde{\mathrm{C}}_\alpha[Z]$ introduced in Subsection 4.1—this enables the reduction described at the end of Section 3. In Subsection 4.3, we conclude the argument by applying the classical Donsker-Varadhan variational formula [Csiszár, 1975, Donsker and Varadhan, 1976].

4.1 An Auxiliary Estimator for CVaR

In this subsection, we introduce an auxiliary estimator $\widetilde{\mathrm{C}}_\alpha[Z]$ of $\mathrm{C}_\alpha[Z]$ and show that it is not much larger than $\widehat{\mathrm{C}}_\alpha[Z]$. For $\alpha, \delta \in (0,1)$, $n \in \mathbb{N}$, and $\pi \coloneqq (1,\ldots,1)^\top/n \in \mathbb{R}^n$, define:

$$\widetilde{\mathcal{Q}}_\alpha \coloneqq \left\{ \boldsymbol{q} \in [0, 1/\alpha]^n \;:\; |\mathrm{E}_{i\sim\pi}[q_i] - 1| \le \epsilon_n \right\}, \quad \text{where} \quad \epsilon_n \coloneqq \sqrt{\frac{2\ln\frac{1}{\delta}}{\alpha n}} + \frac{\ln\frac{1}{\delta}}{3\alpha n}. \tag{11}$$

Using the set $\widetilde{\mathcal{Q}}_\alpha$, and given i.i.d. copies $Z_1, \ldots, Z_n$ of Z, let

$$\widetilde{\mathrm{C}}_\alpha[Z] \coloneqq \sup_{\boldsymbol{q} \in \widetilde{\mathcal{Q}}_\alpha} \frac{1}{n}\sum_{i=1}^{n} Z_i q_i. \tag{12}$$

In the next lemma, we give a "variational formulation" of $\widetilde{\mathrm{C}}_\alpha[Z]$, which will be key in our results:

Lemma 2. *Let $\alpha, \delta \in (0,1)$, $n \in \mathbb{N}$, and $\widetilde{\mathrm{C}}_\alpha[Z]$ be as in (12). Then, for any $Z_1, \ldots, Z_n \in \mathbb{R}$,*

$$\widetilde{\mathrm{C}}_\alpha[Z] = \inf_{\mu \in \mathbb{R}} \left\{ \mu + |\mu|\epsilon_n + \frac{\mathrm{E}_{i\sim\pi}[Z_i - \mu]_+}{\alpha} \right\}, \quad \textit{where } \epsilon_n \textit{ as in (11).} \tag{13}$$

The proof of Lemma 2 (which is in Appendix A.1) is similar to that of the generalized Donsker-Varadhan variational formula considered in [Beck and Teboulle, 2003]. The "variational formulation" on the RHS of (13) reveals some similarity between $\widetilde{\mathrm{C}}_\alpha[Z]$ and the standard estimator $\widehat{\mathrm{C}}_\alpha[Z]$ defined in (3). In fact, thanks to Lemma 2, we have the following relationship between the two:

Lemma 3. *Let $\alpha, \delta \in (0,1)$, $n \in \mathbb{N}$, and $Z_1, \ldots, Z_n \in \mathbb{R}_{\geq 0}$. Further, let $Z_{(1)}, \ldots, Z_{(n)}$ be the decreasing order statistics of $Z_1, \ldots, Z_n$. Then, for ϵ_n as in (11), we have*

$$\widetilde{\mathrm{C}}_\alpha[Z] \leq \widehat{\mathrm{C}}_\alpha[Z] \cdot (1 + \epsilon_n); \tag{14}$$

and if $Z_1, \ldots, Z_n \in \mathbb{R}$ (not necessarily positive), then

$$\widetilde{\mathrm{C}}_\alpha[Z] \leq \widehat{\mathrm{C}}_\alpha[Z] + |Z_{(\lceil n\alpha \rceil)}| \cdot \epsilon_n. \tag{15}$$

The inequality in (15) will only be relevant to us in the case where Z maybe negative, which we deal with in Section 5 when we derive new concentration bounds for CVaR.

4.2 An Auxiliary Random Variable

In this subsection, we introduce a random variable Y which satisfies the properties in (8) and (9), where $Y_1, \ldots, Y_n$ are i.i.d. copies of Y (this is where we leverage the dual representation in (1)). This allows us to the reduce the problem of estimating CVaR to that of estimating an expectation.

Let $\mathcal{X}$ be an arbitrary set, and $f \colon \mathcal{X} \to \mathbb{R}$ be some fixed measurable function (we will later set f to a specific function depending on whether we want a new concentration inequality or a PAC-Bayesian bound for CVaR). Given a random variable X in $\mathcal{X}$ (arbitrary for now), we define

$$Z \coloneqq f(X) \tag{16}$$

and the auxiliary random variable:

$$Y \coloneqq Z \cdot \mathrm{E}[q_\star \mid X] = f(X) \cdot \mathrm{E}[q_\star \mid X], \quad \text{where} \quad q_\star \in \operatorname*{argmax}_{q \in \mathcal{Q}_\alpha} \mathrm{E}[Zq], \tag{17}$$

and $\mathcal{Q}_\alpha$ as in (2). In the next lemma, we show two crucial properties of the random variable Y—these will enable the reduction mentioned at the end of Section 3:

Lemma 4. *Let $\alpha, \delta \in (0,1)$ and $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathcal{X}$. Then, **(I)** the random variable Y in (17) and $Y_i \coloneqq Z_i \cdot \mathrm{E}[q_\star \mid X_i], i \in [n]$, where $Z_i \coloneqq f(X_i)$, are i.i.d. and satisfy $\mathrm{E}[Y] = \mathrm{E}[Y_i] = \mathrm{C}_\alpha[Z]$, for all $i \in [n]$; and **(II)** with probability at least $1 - \delta$,*

$$(\mathrm{E}[q_\star \mid X_1], \ldots, \mathrm{E}[q_\star \mid X_n])^\top \in \widetilde{\mathcal{Q}}_\alpha, \quad \textit{where } \widetilde{\mathcal{Q}}_\alpha \textit{ is as in } (11). \tag{18}$$

The random variable Y introduced in (17) will now be useful since due to (18) in Lemma 4, we have, for $\alpha, \delta \in (0,1)$; Z as in (16); and i.i.d. random variables $X, X_1, \ldots, X_n \in \mathcal{X}$,

$$P\left[\frac{1}{n}\sum_{i=1}^n Y_i \leq \widetilde{\mathrm{C}}_\alpha[Z]\right] \geq 1 - \delta, \quad \text{where } Y_i = Z_i \cdot \mathrm{E}[q_\star \mid X_i] \tag{19}$$

and $\widetilde{\mathrm{C}}_\alpha[Z]$ as in (12). We now present a concentration inequality for the random variable Y in (17); the proof, which can be found in Appendix A, is based on a version of the standard Bernstein's moment inequality [Cesa-Bianchi and Lugosi, 2006, Lemma A.5]:

Lemma 5. *Let $X, (X_i)_{i \in [n]}$ be i.i.d. random variables in $\mathcal{X}$. Further, let Y be as in (17), and $Y_i = f(X_i) \cdot \mathrm{E}[q_\star \mid X_i], i \in [n]$, with $q_\star$ as in (17). If $\{f(x) \mid x \in \mathcal{X}\} \subseteq [0,1]$, then for all $\eta \in [0, \alpha]$,*

$$\mathrm{E}\left[\exp\left(n\eta \cdot \left(\mathrm{E}_P[Y] - \frac{1}{n}\sum_{i=1}^n Y_i - \frac{\eta\kappa(\eta/\alpha)}{\alpha}\mathrm{C}_\alpha[Z]\right)\right)\right] \leq 1, \quad \textit{where } Z = f(X),$$

and $\kappa(x) \coloneqq (e^x - 1 - x)/x^2$, for $x \in \mathbb{R}$.

Lemma 5 will be our starting point for deriving the PAC-Bayesian bound in Theorem 1.

4.3 Exploiting the Donsker-Varadhan Formula

In this subsection, we instantiate the results of the previous subsections with $f(\cdot) \coloneqq \ell(\cdot, h)$, $h \in \mathcal{H}$, for some loss function $\ell\colon \mathcal{H} \times \mathcal{X} \to [0,1]$; in this case, the results of Lemmas 3 and 5 hold for

$$Z = Z_h \coloneqq \ell(h, X), \tag{20}$$

for any hypothesis $h \in \mathcal{H}$. Next, we will need the following result which follows from the classical Donsker-Varadhan variational formula [Csiszár, 1975, Donsker and Varadhan, 1976]:

Lemma 6. *Let $\delta \in (0,1)$, $\gamma > 0$ and ρ_0 be any fixed (prior) distribution over $\mathcal{H}$. Further, let $\{R_h : h \in \mathcal{H}\}$ be any family of random variables such that $\mathrm{E}[\exp(\gamma R_h)] \le 1$, for all $h \in \mathcal{H}$. Then, for any (posterior) distribution $\widehat{\rho}$ over $\mathcal{H}$, we have*

$$P\left[\mathrm{E}_{h\sim\widehat{\rho}}[R_h] \le \frac{\mathrm{KL}(\widehat{\rho}\|\rho_0) + \ln\frac{1}{\delta}}{\gamma}\right] \ge 1 - \delta.$$

In addition to Z_h in (20), define $Y_h \coloneqq \ell(h, X) \cdot \mathrm{E}[q_\star \mid X]$ and $Y_{h,i} \coloneqq \ell(h, X_i) \cdot \mathrm{E}[q_\star \mid X_i]$, for $i \in [n]$. Then, if we set $\gamma = \eta n$ and $R_h = \mathrm{E}_P[Y_h] - \sum_{i=1}^n Y_{h,i}/n - \eta\kappa(\eta/\alpha)\mathrm{C}_\alpha[Z_h]/\alpha$, Lemma 5 guarantees that $\mathrm{E}[\exp(\gamma R_h)] \le 1$, and so by Lemma 6 we get the following result:

Theorem 7. *Let $\alpha, \delta \in (0,1)$, and $\eta \in [0, \alpha]$. Further, let $X_1, \dots, X_n$ be i.i.d. random variables in $\mathcal{X}$. Then, for any randomized estimator $\widehat{\rho} = \widehat{\rho}(X_{1:n})$ over $\mathcal{H}$, we have, with $\widehat{Z} \coloneqq \mathrm{E}_{h\sim\widehat{\rho}}[\ell(h, X)]$,*

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]] \le \widehat{\mathrm{C}}_\alpha[\widehat{Z}](1+\epsilon_n) + \frac{\eta\kappa(\eta/\alpha)\,\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]]}{\alpha} + \frac{\mathrm{KL}(\widehat{\rho}\|\rho_0) + \ln\frac{1}{\delta}}{\eta n}, \tag{21}$$

with probability at least $1 - 2\delta$ on the samples $X_1, \dots, X_n$, where ϵ_n is defined in (11).

If we could optimize the RHS of (21) over $\eta \in [0, \alpha]$, this would lead to our desired bound in Theorem 1 (after some rearranging). However, this is not directly possible since the optimal η depends on the sample $X_{1:n}$ through the term $\mathrm{KL}(\widehat{\rho}\|\rho_0)$. The solution is to apply the result of Theorem 7 with a union bound, so that (21) holds for any estimator $\hat{\eta} = \hat{\eta}(X_{1:n})$ taking values in a carefully chosen grid $\mathcal{G}$; to derive our bound, we will use the grid $\mathcal{G} \coloneqq \left\{\alpha 2^{-1}, \dots, \alpha 2^{-N} \mid N \coloneqq \lceil 1/2 \log_2(n/\alpha) \rceil\right\}$. From this point, the proof of Theorem 1 is merely a mechanical exercise of rearranging (21) and optimizing $\hat{\eta}$ over $\mathcal{G}$, and so we postpone the details to Appendix A.

5 New Concentration Bounds for CVaR

In this section, we show how some of the results of the previous section can be used to reduce the problem of estimating $\mathrm{C}_\alpha[Z]$ to that of estimating a standard expectation. This will then enable us to easily obtain concentration inequalities for $\widehat{\mathrm{C}}_\alpha[Z]$ even when Z is unbounded. We note that previous works [Bhat and Prashanth, 2019, Kolla et al., 2019] used sophisticated techniques to deal with the unbounded case (sometimes achieving only sub-optimal rates), whereas we simply invoke existing concentration inequalities for empirical means thanks to our reduction.

The key results we will use are Lemmas 3 and 4, where we instantiate the latter with $\mathcal{X} = \mathbb{R}$ and $f \equiv \mathrm{id}$, in which case:

$$Y = Z \cdot \mathrm{E}[q_\star \mid Z], \quad \text{and } q_\star \in \operatorname*{argmax}_{q \in \mathcal{Q}_\alpha} \mathrm{E}[Zq]. \tag{22}$$

Together, these two lemmas imply that, for any $\alpha, \delta \in (0,1)$, i.i.d. random variables $Z_1, \dots, Z_n$,

$$\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] - \left|Z_{(\lceil n\alpha \rceil)}\right|\epsilon_n \le \mathrm{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i, \tag{23}$$

with probability at least $1 - \delta$, where ϵ_n is as in (11) and $Z_{(1)}, \dots, Z_{(n)}$ are the decreasing order statistics of $Z_1, \dots, Z_n \in \mathbb{R}$. Thus, getting a concentration inequality for $\widehat{\mathrm{C}}_\alpha[Z]$ can be reduced to getting one for the empirical mean $\sum_{i=1}^n Y_i/n$ of the i.i.d. random variables $Y_1, \dots, Y_n$. Next, we show that whenever Z is a sub-exponential [resp. sub-Gaussian] random variable, essentially so is Y. But first we define what this means:

Definition 8. *Let $\mathcal{I} \subseteq \mathbb{R}$, $b > 0$, and Z be a random variable such that, for some $\sigma > 0$,*

$$\mathrm{E}[\exp(\eta \cdot (Z - \mathrm{E}[Z]))] \leq \exp\left(\eta^2\sigma^2/2\right), \quad \forall \eta \in \mathcal{I},$$

Then, Z is (σ, b)-sub-exponential [resp. σ-sub-Gaussian] if $\mathcal{I} = (-1/b, 1/b)$ [resp. $\mathcal{I} = \mathbb{R}$].

Lemma 9. *Let $\sigma, b > 0$ and $\alpha \in (0,1)$. Let Z be a zero-mean real random variable and let Y be as in (22). If Z is (σ, b)-sub-exponential [resp. σ-sub-Gaussian], then*

$$\mathrm{E}[\exp(\eta Y)] \leq 2\exp(\eta^2\sigma^2/(2\alpha^2)), \quad \forall \eta \in (-\alpha/b, \alpha/b) \quad [\textit{resp. } \eta \in \mathbb{R}]. \tag{24}$$

Note that in Lemma 9 we have assumed that Z is a zero-mean random variable, and so we still need to do some work to derive a concentration inequality for $\widehat{\mathrm{C}}_\alpha[Z]$. In particular, we will use the fact that $\mathrm{C}_\alpha[Z - \mathrm{E}[Z]] = \mathrm{C}_\alpha[Z] - \mathrm{E}[Z]$ and $\widehat{\mathrm{C}}_\alpha[Z - \mathrm{E}[Z]] = \widehat{\mathrm{C}}_\alpha[Z] - \mathrm{E}[Z]$, which holds since C_α and $\widehat{\mathrm{C}}_\alpha$ are coherent risk measures, and thus translation equivariant (see Definition 14). We use this in the proof of the next theorem (which is in Appendix A):

Theorem 10. *Let $\sigma, b > 0$, $\alpha, \delta \in (0,1)$, and ϵ_n be as in (11). If Z is a σ-sub-Gaussian random variable, then with $\mathrm{G}[Z] \coloneqq \mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z]$ and $t_n \coloneqq \left|Z_{(\lceil n\alpha \rceil)} - \mathrm{E}[Z]\right| \cdot \epsilon_n$, we have*

$$P\left[\mathrm{G}[Z] \geq t + t_n\right] \leq \delta + 2\exp(-n\alpha^2 t^2/(2\sigma^2)), \quad \forall t \geq 0; \tag{25}$$

otherwise, if Z is (σ, b)-sub-exponential random variable, then

$$P\left[\mathrm{G}[Z] \geq t + t_n\right] \leq \delta + \begin{cases} 2\exp\left(-n\alpha^2 t^2/(2\sigma^2)\right), & \textit{if } 0 \leq t \leq \sigma^2/(b\alpha); \\ 2\exp\left(-n\alpha t/(2b)\right), & \textit{if } t > \sigma^2/(b\alpha). \end{cases}$$

We note that unlike the recent results due to Bhat and Prashanth [2019] which also deal with the unbounded case, the constants in our concentration inequalities in Theorem 10 are explicit.

When Z is a σ-sub-Gaussian random variable with $\sigma > 0$, an immediate consequence of Theorem 10 is that by setting $t = \sqrt{2\sigma^2 \ln(2/\delta)/(n\alpha^2)}$ in (25), we get that, with probability at least $1 - 2\delta$,

$$\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \leq \frac{\sigma}{\alpha}\sqrt{\frac{2\ln\frac{1}{\delta}}{n}} + \left|Z_{(\lceil n\alpha \rceil)} - \mathrm{E}[Z]\right| \cdot \left(\sqrt{\frac{2\ln\frac{1}{\delta}}{\alpha n}} + \frac{\ln\frac{1}{\delta}}{3\alpha n}\right). \tag{26}$$

A similar inequality holds for the sub-exponential case. We note that the term $|Z_{(\lceil n\alpha \rceil)} - \mathrm{E}[Z]|$ in (26) can be further bounded from above by

$$\frac{n\alpha}{\lfloor n\alpha \rfloor}\widehat{\mathrm{C}}_\alpha[Z] - \mathrm{E}_{\widehat{P}_n}[Z] + \left|\mathrm{E}[Z] - \mathrm{E}_{\widehat{P}_n}[Z]\right|. \tag{27}$$

This follows from the triangular inequality and facts that $\widehat{\mathrm{C}}_\alpha[Z] \geq \frac{1}{n\alpha}\sum_{i=1}^{\lfloor n\alpha \rfloor} Z_{(i)} \geq \frac{\lfloor n\alpha \rfloor}{n\alpha} Z_{(\lceil n\alpha \rceil)}$ (see *e.g.* Lemma 4.1 in Brown [2007]), and $\widehat{\mathrm{C}}_\alpha[Z] \geq \mathrm{E}_{\widehat{P}_n}[Z]$ [Ahmadi-Javid, 2012]. The remaining term $|\mathrm{E}_P[Z] - \mathrm{E}_{\widehat{P}_n}[Z]|$ in (27) which depends on the unknown P can be bounded from above using another concentration inequality.

Generalization bounds of the form (4) for unbounded but sub-Gaussian or sub-exponential $\ell(h, X)$, $h \in \mathcal{H}$, can be obtained using the PAC-Bayesian analysis of [McAllester, 2003, Theorem 1] and our concentration inequalities in Theorem 10. However, due to the fact that α is squared in the argument of the exponentials in these inequalities (which is also the case in the bounds of Bhat and Prashanth [2019], Kolla et al. [2019]) the generalization bounds obtained this way will have the α outside the square-root "complexity term"—unlike our bound in Theorem 1.

We conjecture that the dependence on α in the concentration bounds of Theorem 10 can be improved by swapping α^2 for α in the argument of the exponentials; in the sub-Gaussian case, this would move α inside the square-root on the RHS of (26). We know that this is at least possible for bounded random variables as shown in Brown [2007], Wang and Gao [2010]. We now recover this fact by presenting a new concentration inequality for $\widehat{\mathrm{C}}_\alpha[Z]$ when Z is bounded using the reduction described at the beginning of this section.

Theorem 11. *Let $\alpha, \delta \in (0,1)$, and $Z_{1:n}$ be i.i.d.* rvs *in $[0,1]$. Then, with probability at least $1 - 2\delta$,*

$$\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \leq \sqrt{\frac{12\mathrm{C}_\alpha[Z]\ln\frac{1}{\delta}}{5\alpha n}} \vee \frac{3\ln\frac{1}{\delta}}{\alpha n} + \mathrm{C}_\alpha[Z]\left(\sqrt{\frac{2\ln\frac{1}{\delta}}{\alpha n}} + \frac{\ln\frac{1}{\delta}}{3\alpha n}\right). \tag{28}$$

The proof is in Appendix A. The inequality in (28) essentially replaces the range of the random variable Z typically present under the square-root in other concentration bounds [Brown, 2007, Wang and Gao, 2010] by the smaller quantity $\mathrm{C}_\alpha[Z]$. The concentration bound (28) is not immediately useful for computational purposes since its RHS depends on $\mathrm{C}_\alpha[Z]$. However, it is possible to rearrange this bound so that only the empirical quantity $\widehat{\mathrm{C}}_\alpha[Z]$ appears on the RHS of (28) instead of $\mathrm{C}_\alpha[Z]$; we provide the means to do this in Lemma 13 in the appendix.

6 Conclusion and Future Work

In this paper, we derived a first PAC-Bayesian bound for CVaR by reducing the task of estimating CVaR to that of merely estimating an expectation (see Section 4). This reduction then made it easy to obtain concentration inequalities for CVaR (with explicit constants) even when the random variable in question is unbounded (see Section 5).

We note that the only steps in the proof of our main bound in Theorem 1 that are specific to CVaR are Lemmas 2 and 3, and so the question is whether our overall approach can be extended to other coherent risk measures to achieve (4).

In Appendix B, we discuss how our results may be extended to a rich class of coherent risk measures known as φ-entropic risk measures. These CRMs are often used in the context of robust optimization Namkoong and Duchi [2017], and are perfect candidates to consider next in the context of this paper.

Broader Impact

Coherent risk measures (including conditional value at risk) have been gaining significant traction in the machine learning community recently, as they allow for capturing in a much richer way the behaviour and performance of algorithms' outputs. This comes at the expense of a much harder theoretical analysis and such measures are not supported by as many guarantees than the traditional mean risk (expectation of the loss). We provide in this paper one of the few generalisation bounds for CVaR and we believe this will shed light on the advantages of using CVaR in machine learning. We intend our contributions to be of prime interest to theoreticians, but also to practitioners.

Acknowledgments and Disclosure of Funding

This work was supported by the Australian Research Council and Data61.

References

Karim T. Abou-Moustafa and Csaba Szepesvári. An exponential tail bound for lq stable learning rules. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA*, volume 98 of *Proceedings of Machine Learning Research*, pages 31–63. PMLR, 2019. URL http://proceedings.mlr.press/v98/abou-moustafa19a.html.

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, pages 503–546, 1953.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Sanjay P Bhat and LA Prashanth. Concentration of risk measures: A Wasserstein distance approach. In *Advances in Neural Information Processing Systems*, pages 11739–11748, 2019.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer School on Machine Learning*, pages 208–240. Springer, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *CoRR*, abs/1910.07833, 2019. URL http://arxiv.org/abs/1910.07833.

David B. Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Lecture Notes-Monograph Series. IMS, 2007.

Alain Celisse and Benjamin Guedj. Stability revisited: new generalisation bounds for the leave-one-out. *arXiv preprint arXiv:1608.06412*, 2016.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Youhua (Frank) Chen, Minghui Xu, and Zhe George Zhang. A Risk-Averse Newsvendor Model Under the CVaR Criterion. *Operations Research*, 57(4):1040–1044, 2009. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/25614814.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for CVaR Optimization in MDPs. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3509–3517. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5246-algorithms-for-cvar-optimization-in-mdps.pdf.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

Sebastian Curi, Kfir. Y. Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning, 2020.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time — III. *Communications on pure and applied Mathematics*, 29(4):389–461, 1976.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, pages 643–669, 1961.

Benjamin Guedj. A Primer on PAC-Bayesian Learning. *arXiv*, 2019. URL https://arxiv.org/abs/1901.05353.

Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society Open Science*, 4(11), 2017.

Ravi Kumar Kolla, LA Prashanth, Sanjay P Bhat, and Krishna Jagannathan. Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47(1): 16–20, 2019.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings COLT 2009*, 2009.

David A. McAllester. PAC-Bayesian Stochastic Model Selection. In *Machine Learning*, volume 51, pages 5–21, 2003.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes Un-Expected Bernstein Inequality. In *Advances in Neural Information Processing Systems*, pages 12180–12191, 2019.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 799–806, 2010.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pages 2971–2980, 2017.

Georg Ch. Pflug. *Some Remarks on the Value-at-Risk and the Conditional Value-at-Risk*, pages 272–281. Springer US, Boston, MA, 2000. ISBN 978-1-4757-3150-7. doi: 10.1007/978-1-4757-3150-7_15. URL https://doi.org/10.1007/978-1-4757-3150-7_15.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2817–2826. JMLR. org, 2017.

LA Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in neural information processing systems*, pages 252–260, 2013.

R Tyrrell Rockafellar and Stan Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1-2):33–53, 2013.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

Akiko Takeda and Takafumi Kanamori. A robust approach based on conditional value-at-risk measure to statistical learning problems. *European Journal of Operational Research*, 198(1): 287 – 296, 2009. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2008.07.027. URL http://www.sciencedirect.com/science/article/pii/S0377221708005614.

Akiko Takeda and Masashi Sugiyama. ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th international conference on Machine learning*, pages 1056–1063, 2008.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Philip Thomas and Erik Learned-Miller. Concentration inequalities for conditional value at risk. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6225–6233, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Ilya O. Tolstikhin and Yevgeny Seldin. Pac-bayes-empirical-bernstein inequality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 109–117, 2013. URL http://papers.nips.cc/paper/4903-pac-bayes-empirical-bernstein-inequality.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Ying Wang and Fuqing Gao. Deviation inequalities for an estimator of the conditional value-at-risk. *Operations Research Letters*, 38(3):236–239, 2010.

Robert Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797, 2019.

A Proofs

A.1 Proof of Lemma 2

Proof Let $\varphi(\cdot) \coloneqq \iota_{[0,1/\alpha]}(\cdot)$, where for a set $\mathcal{C} \subseteq \mathbb{R}$, $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$; and $+\infty$ otherwise. From (12), we have that $\widetilde{\mathrm{C}}_\alpha[Z]$ is equal to

$$\mathcal{P} \coloneqq \sup_{\boldsymbol{q}:|\mathrm{E}_{i\sim\pi}[q_i]-1|\le\epsilon_n} \mathrm{E}_{i\sim\pi}[Z_i q_i - \varphi(q_i)], \tag{29}$$

where we recall that $\pi = (1,\dots,1)^\top/n \in \mathbb{R}^n$. The Lagrangian dual $\mathcal{D}$ of (29) is given by

$$\begin{aligned}
\mathcal{D} &\coloneqq \inf_{\eta,\gamma\ge0}\left\{\eta-\gamma+(\eta+\gamma)\epsilon_n + \sup_{\boldsymbol{q}:0\le q_i\le 1/\alpha, i\in[n]}\{\mathrm{E}_{i\sim\pi}[(Z_i-\eta+\gamma)q_i-\varphi(q_i)]\}\right\},\\
&= \inf_{\eta,\gamma\ge0}\left\{\eta-\gamma+(\eta+\gamma)\epsilon_n + \mathrm{E}_{i\sim\pi}\left[\sup_{0\le x\le 1/\alpha}\{(Z_i-\eta+\gamma)x-\varphi(x)\}\right]\right\},\\
&= \inf_{\eta,\gamma\ge0}\{\eta-\gamma+(\eta+\gamma)\epsilon_n + \mathrm{E}_{i\sim\pi}[\varphi^\star(Z_i-\eta+\gamma)]\}, && (30)\\
&= \inf_{\mu\in\mathbb{R}}\{\mu+|\mu|\epsilon_n + \mathrm{E}_{i\sim\pi}[\varphi^\star(Z_i-\mu)]\}, && (31)
\end{aligned}$$

where (30) is due to $\{x \in \mathbb{R} \mid \varphi(x) < +\infty\} = [0, 1/\alpha]$, and (31) follows by setting $\mu \coloneqq \eta - \gamma$ and noting that the inf in (30) is always attained at a point $(\eta,\gamma) \in \mathbb{R}^2_{\ge0}$ satisfying $\eta\cdot\gamma = 0$, in which case $\eta+\gamma = |\mu|$; this is true because by the positivity of ϵ_n, if $\eta,\gamma > 0$, then $(\eta+\gamma)\epsilon_n$ can always be made smaller while keeping the difference $\eta-\gamma$ fixed. Finally, since the primal problem is feasible—$\boldsymbol{q} = \pi$ is a feasible solution—there is no duality gap (see the proof of [Beck and Teboulle, 2003, Theorem 4.2]), and thus the RHS of (31) is equal to $\mathcal{P}$ in (29). The proof is concluded by noting that the Fenchel dual of φ satisfies $\varphi^\star(z) = 0 \vee (z/\alpha)$, for all $z \in \mathbb{R}$. ■

A.2 Proof of Lemma 3

Proof Let $\widehat{\mu}$ be the argmin in $\mu \in \mathbb{R}$ of the RHS of (3). By Lemma 2, we have

$$\begin{aligned}
\widetilde{\mathrm{C}}_\alpha[Z] &= \inf_{\mu\in\mathbb{R}}\left\{\mu+|\mu|\epsilon_n + \frac{\mathrm{E}_{i\sim\pi}[Z_i-\mu]_+}{\alpha}\right\},\\
&\le \widehat{\mu}+|\widehat{\mu}|\epsilon_n + \frac{\mathrm{E}_{i\sim\pi}[Z_i-\widehat{\mu})]_+}{\alpha},\\
&= \widehat{\mathrm{C}}_\alpha[Z] + |\widehat{\mu}|\epsilon_n. \quad \text{(by definition of } \widehat{\mu}) && (32)
\end{aligned}$$

The inequality in (15) follows from (32) and the fact that $\widehat{\mu} = Z_{(\lceil n\alpha\rceil)}$ (see proof of [Brown, 2007, Proposition 4.1]).

Now we show (14) under the assumption that $Z_i \ge 0$, for all $i \in [n]$. Note that by definition $\widehat{\mathrm{C}}_\alpha[Z] = \widehat{\mu} + \frac{1}{\alpha}\mathrm{E}_{i\sim\pi}[Z_i-\widehat{\mu}]_+$, and so $\widehat{\mu} \le \widehat{\mathrm{C}}_\alpha[Z]$. Furthermore, since $\alpha \in (0,1)$ and $Z_i \ge 0$, for $i \in [n]$, the RHS of (3) is a decreasing function of μ on $]-\infty, 0]$, and thus $\widehat{\mu} \ge 0$ (since $\widehat{\mu}$ is the minimizer of (3)). Combining the fact that $0 \le \widehat{\mu} \le \widehat{\mathrm{C}}_\alpha[Z]$ with (32) completes the proof. ■

A.3 Proof of Lemma 4

Proof The first claim follows by the fact that $X_i, i \in [n]$, are i.i.d., and an application of the total expectation theorem. Now for the second claim, let $\Delta \coloneqq |\mathrm{E}_{\widehat{P}_n}[q_\star \mid X] - 1|$. Since $q_\star$ is a density, the total expectation theorem implies

$$\Delta = |\mathrm{E}_{\widehat{P}_n}[q_\star \mid X] - \mathrm{E}[\mathrm{E}[q_\star \mid X]]|,$$

and so by Bennett's inequality (see *e.g.* Theorem 3 in Maurer and Pontil [2009]) applied to the random variable $\mathrm{E}[q_\star \mid X]$, we get that, with probability at least $1-\delta$,

$$\begin{aligned}\Delta &\le \sqrt{\frac{2\mathrm{VAR}[\mathrm{E}[q_\star \mid X]]\ln\frac{1}{\delta}}{n}} + \frac{\|\mathrm{E}[q_\star \mid X]\|_\infty \ln\frac{1}{\delta}}{3n},\\ &\le \sqrt{\frac{2\,\mathrm{E}[\mathrm{E}[q_\star \mid X]^2]\ln\frac{1}{\delta}}{n}} + \frac{\|\mathrm{E}[q_\star \mid X]\|_\infty \ln\frac{1}{\delta}}{3n},\\ &\le \sqrt{\frac{2\|\mathrm{E}[q_\star \mid X]\|_\infty \ln\frac{1}{\delta}}{n}} + \frac{\|\mathrm{E}[q_\star \mid X]\|_\infty \ln\frac{1}{\delta}}{3n},\end{aligned}$$

where the last inequality follows by the fact that $\mathrm{E}[\mathrm{E}[q_\star \mid X]^2] \le \mathrm{E}[\mathrm{E}[q_\star \mid X]] \cdot \|\mathrm{E}[q_\star \mid X]\|_\infty = \|\mathrm{E}[q_\star \mid X]\|_\infty$, which holds since $\mathrm{E}[q_\star \mid X] \ge 0$ and $\mathrm{E}[\mathrm{E}[q_\star \mid X]] = \mathrm{E}[q_\star] = 1$. The proof is concluded by the facts that $\|\mathrm{E}[q_\star \mid X]\|_\infty \le \|q_\star\|_\infty$ (by Jensen's inequality); $\|q\|_\infty \le 1/\alpha$, for all $q \in \mathcal{Q}_\alpha$ by definition; and $q_\star \in \mathcal{Q}_\alpha$. ■

A.4 Proof of Lemma 5

We need the following lemma in the proof of Lemma 5:

Lemma 12. *Let $S, S_1, \ldots, S_n$ be i.i.d. random variable such that $S \in [0, B]$, $B > 0$. We have,*

$$\mathrm{E}_P\left[\exp\left(n\eta\,\mathrm{E}_P[S] - \eta\sum_{i=1}^n S_i - n\eta^2\kappa(\eta B)\cdot \mathrm{E}_P[S^2]\right)\right] \le 1, \tag{33}$$

for all $\eta \in [0, 1/B]$, where $\kappa(\eta) := (e^\eta - \eta - 1)/\eta^2$.

Proof The desired bound follows by the version of Bernstein's moment inequality in [Cesa-Bianchi and Lugosi, 2006, Lemma A.5] and [Mhammedi et al., 2019, Proposition 10-(b)]. ■

Proof of Lemma 5 By Lemma 4, the random variables $Y, Y_1, \ldots, Y_n$ are i.i.d., and so the result of Lemma 12 applies; this means that (33) holds for $(S, S_1, \ldots, S_n) = (Y, Y_1, \ldots, Y_n)$ and $B = b \ge \|Y\|_\infty$. Thus, to complete the proof it suffices to bound $\|Y\|_\infty$ and $\|Y\|_2^2 = \mathrm{E}[Y^2]$ from above. Starting with $\mathrm{E}[Y^2]$, and recalling that $Z = f(X) \in [0,1]$ by assumption, we have:

$$\begin{aligned}\mathrm{E}[Y^2] &= \mathrm{E}[Z^2 \cdot \mathrm{E}[q_\star \mid X]^2],\\ &\le \mathrm{E}[Z\cdot \mathrm{E}[q_\star \mid X]] \cdot \|Z \cdot \mathrm{E}[q_\star \mid X]\|_\infty, \text{(Hölder)}\\ &\le \mathrm{C}_\alpha[Z] \cdot \|Z\cdot \mathrm{E}[q_\star \mid X]\|_\infty, \qquad \text{(Lemma 4)}\\ &\le \mathrm{C}_\alpha[Z]/\alpha, \qquad (Z \le 1,\ q_\star \le 1/\alpha)\end{aligned}$$

where the fact that $q_\star \le 1/\alpha$ follows simply from $q_\star \in \mathcal{Q}_\alpha$ and the definition of $\mathcal{Q}_\alpha$. We also have

$$\begin{aligned}\|Y\|_\infty = \|Z\cdot \mathrm{E}[q_\star \mid X]\|_\infty &\le \|Z\|_\infty \cdot \|\mathrm{E}[q_\star \mid X]\|_\infty,\\ &\le \|q_\star\|_\infty, (Z \le 1 \text{ \& Jensen})\\ &\le 1/\alpha,\end{aligned}$$

again the last inequality follows from $q_\star \in \mathcal{Q}_\alpha$ and the definition of $\mathcal{Q}_\alpha$. ■

A.5 Proof of Theorem 7

Proof Let $h \in \mathcal{H}$ and $\alpha, \delta \in (0,1)$, and define

$$R_h := \mathrm{C}_\alpha[Z_h] - \frac{1}{n}\sum_{i=1}^n Y_i - \frac{\eta\kappa(\eta/\alpha)}{\alpha}\mathrm{C}_\alpha[Z_h], \tag{34}$$

where $Y_i := \ell(h, X_i)\cdot \mathrm{E}[q_\star \mid X_i], i \in [n]$, where $q_\star$ is as in (17) with Z as in (20). By Lemma 4, $\mathrm{C}_\alpha[Z_h] = \mathrm{E}_P[Y]$, where $Y := \ell(h, X)\cdot \mathrm{E}[q_\star \mid X]$. Thus, by Lemma 5 with $Z = Z_h$, we have $\mathrm{E}_P[\exp(n\eta R_h)] \le 1$. Applying Lemma 6 with R_h as in (34) and $\gamma = n\eta$, yields,

$$\begin{aligned}\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]] \le\ & \frac{1}{n}\sum_{i=1}^n \ell(\widehat{\rho}, X_i)\cdot \mathrm{E}[q_\star \mid X_i] + \frac{\eta\kappa(\eta/\alpha)\,\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]]}{\alpha}\\ &+ \frac{\mathrm{KL}(\widehat{\rho}\|\rho_0) + \ln\frac{1}{\delta}}{\eta n},\end{aligned} \tag{35}$$

with probability at least $1-\delta$. Now invoking Lemmas 3 and 4 (in particular (18)), yields

$$\frac{1}{n}\sum_{i=1}^{n}\ell(\widehat{\rho},X_i)\cdot \mathrm{E}[q_\star \mid X_i] \geq \widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot(1+\epsilon_n).$$

with probability at least $1-\delta$, where $\widehat{Z} \coloneqq \mathrm{E}_{h\sim\widehat{\rho}}[\ell(h,X)]$. Combining this with (35) via a union bound yields the desired bound. ■

A.6 Proof of Theorem 1

To prove Theorem 1, we will need the following lemma:

Lemma 13. *Let $R,\widehat{R},A,B>0$. If $R\leq \widehat{R}+\sqrt{RA}+B$, then*

$$R\leq \widehat{R}+\sqrt{\widehat{R}A}+2B+A.$$

Proof If $R\leq \widehat{R}+\sqrt{RA}+B$, then for all $\eta>0$,

$$R\leq \widehat{R}+\frac{\eta}{2}R+\frac{A}{2\eta}+B,$$

which after rearranging, becomes,

$$R\leq \frac{\widehat{R}}{1-\eta/2}+\frac{A}{2\eta\cdot(1-\eta/2)}+\frac{B}{1-\eta/2}, \qquad \text{for } \eta\notin\{0,2\}. \tag{36}$$

The minimizer of the RHS of (36) is given by

$$\eta=\eta_\star \coloneqq \frac{-A+\sqrt{A^2+4AB+4A\widehat{R}}}{2(B+\widehat{R})}.$$

Plugging this η into (36), yields,

$$\begin{aligned} R &\leq \widehat{R}+\frac{A}{2}+B+\frac{1}{2}\sqrt{4A\widehat{R}+A^2+4AB}, \\ &\leq \widehat{R}+A+2B+\sqrt{A\widehat{R}}, \end{aligned} \tag{37}$$

where (37) follows by the facts that $A^2+4AB\leq(A+2B)^2$ and $\sqrt{4\widehat{R}A+(A+2B)^2}\leq\sqrt{4\widehat{R}A}+A+2B$. ■

Proof of Theorem 1 Define the grid $\mathcal{G}$ by

$$\mathcal{G}\coloneqq\left\{2^{-1}\alpha,\ldots,2^{-N}\alpha \mid N\coloneqq\lceil 1/2\cdot\log_2\tfrac{n}{\alpha}\rceil\right\},$$

and let $\hat{\eta}=\hat{\eta}(Z_{1:n})\in\mathcal{G}$ be any estimator. Then, using the fact that $\kappa(x)\leq 3/5$, for all $x\leq 1/2$, and invoking Theorem 7 with a union bound over $\eta\in\mathcal{G}$, and $\varepsilon_n\coloneqq\sqrt{\frac{2\ln\frac{N}{\delta}}{\alpha n}}+\frac{\ln\frac{N}{\delta}}{3\alpha n}$, we get that

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]]-\widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot(1+\varepsilon_n)\leq\frac{\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta}}{\hat{\eta}n}+\frac{3\hat{\eta}}{5\alpha}\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]], \tag{38}$$

with probability at least $1-2\delta$, where we recall that $\widehat{Z}=\mathrm{E}_{h\sim\widehat{\rho}}[\ell(h,X)]$. Let $\hat{\eta}$ be an estimator which satisfies

$$\hat{\eta}\in[\eta_\star\wedge(\alpha/2),\ 2\eta_\star]\cap\mathcal{G}, \quad \text{where} \quad \eta_\star\coloneqq\sqrt{\frac{5\alpha\cdot(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{3n\,\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]]}} \tag{39}$$

is the unconstrained minimizer $\hat{\eta}$ of the RHS of (38). Since the loss ℓ has range in $[0,1]$, $\mathrm{KL}(\widehat{\rho}\|\rho_0)\geq 0$, and $(\delta,n)\in]0,1/2[\times[2,+\infty[$, we have $\eta_\star\geq\sqrt{\alpha/n}\geq\min\mathcal{G}$. This, with the fact that $\mathcal{G}$ is in the form of a geometric progression with common ratio 2 and $\max\mathcal{G}=\alpha/2$, ensures the existence (and in fact the uniqueness) of $\hat{\eta}$ satisfying (39).

Case 1. Suppose that $\eta_\star \le \alpha/2$. In this case, the estimator $\hat{\eta}$ in (39) satisfies $\eta_\star \le \hat{\eta} \le 2\eta_\star$. Plugging $\hat{\eta}$ into (38) yields

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]] - \widehat{\mathrm{C}}_\alpha[\widehat{Z}] \le 3\sqrt{\frac{3\,\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]]\cdot(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{5\alpha n}} + \widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot\varepsilon_n.$$

By applying Lemma 13 with $R = \mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]]$, $\widehat{R} = \widehat{\mathrm{C}}_\alpha[\widehat{Z}]$, $A = \frac{27(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{5\alpha n}$, and $B = \widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot\varepsilon_n$, we get

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]] - \widehat{\mathrm{C}}_\alpha[\widehat{Z}] \le \sqrt{\frac{27\widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{5\alpha n}} + 2\widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot\varepsilon_n + \frac{27(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{5n\alpha}. \tag{40}$$

Case 2. Suppose now that $\eta_\star > \alpha/2$. In this case, $\hat{\eta} = \alpha/2$. Plugging this into (38) and using the fact that $\eta_\star > \alpha/2$, yields

$$\mathrm{E}_{h\sim\widehat{\rho}}[\mathrm{C}_\alpha[\ell(h,X)]] - \widehat{\mathrm{C}}_\alpha[\widehat{Z}] \le \frac{4(\mathrm{KL}(\widehat{\rho}\|\rho_0)+\ln\frac{N}{\delta})}{\alpha n} + \widehat{\mathrm{C}}_\alpha[\widehat{Z}]\cdot\varepsilon_n. \tag{41}$$

Since $\widehat{\mathrm{C}}_\alpha[\widehat{Z}] \ge 0$ and $4 \le 27/5$, the RHS of (41) is less than the RHS of (40), which completes the proof. ■

A.7 Proof of Lemma 9

Proof Suppose that Z is (σ, b)-sub-exponential. Then,

$$\mathrm{E}[e^{\eta Z}] \le e^{\frac{\eta^2\sigma^2}{2}}, \quad \forall|\eta| \le 1/b. \tag{42}$$

Using that $\mathrm{E}[q_\star \mid Z] \le 1/\alpha$, we get

$$|\eta Y| \le |\eta Z|/\alpha, \quad \forall \eta \in \mathbb{R}, \tag{43}$$

and so, for all $|\eta| \le \alpha/b$, we have

$$\mathrm{E}[e^{\eta Y}] \le \mathrm{E}[e^{|\eta Y|}] \overset{(43)}{\le} \mathrm{E}[e^{\frac{\eta Z}{\alpha}}] + \mathrm{E}[e^{-\frac{\eta Z}{\alpha}}] \overset{(42)}{\le} 2e^{\frac{\eta^2\sigma^2}{2\alpha^2}}.$$

When Z is σ-sub-Gaussian case, the proof is the same, except that we replace b by 0. ■

A.8 Proof of Theorem 11

Proof Let $\mathcal{X} = [0,1]$ and $f \equiv \mathrm{id}$ be the identity map. By invoking Lemmas 3 and 5 with $Z = f(X) = X$; and using (19) (which is a consequence of Lemma 4), we get, for all $\eta \in [0,\alpha]$,

$$\mathrm{E}_P\left[\exp\left(n\eta\cdot\left(\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z](1+\epsilon_n) - \frac{\eta\kappa(\eta/\alpha)\mathrm{C}_\alpha[Z]}{\alpha}\right)\right)\right] \le 1, \tag{44}$$

with probability at least $1-\delta$, where ϵ_n is as in (11). By adding $\mathrm{C}_\alpha[Z]\cdot\epsilon_n$ to both sides of (44) and using the fact that $\kappa(x) \le 3/5$, for all $x \le 1/2$, we get, for all $\eta \in [0,\alpha/2]$,

$$\mathrm{E}_P\left[\exp\left(n\eta\cdot\left(\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] - \left(\frac{3\eta}{5\alpha}+\epsilon_n\right)\mathrm{C}_\alpha[Z]\right)\right)\right] \le 1, \tag{45}$$

with probability at least $1-\delta$. Let $W \coloneqq \mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] - \left(\frac{3\eta}{5\alpha}+\epsilon_n\right)\mathrm{C}_\alpha[Z]$, and note that by (45), we have

$$P[\mathrm{E}_P[\exp(n\eta W)] \le 1] \ge 1-\delta.$$

Let $\mathcal{E}$ be the event that $\mathrm{E}_P[\exp(n\eta W)] \le 1$. With this, we have, for any $\delta \in (0,1)$ and all $\eta \in [0, \alpha/2]$,

$$
\begin{aligned}
P\left[\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \ge \left(\frac{3\eta}{5\alpha} + \epsilon_n\right)\mathrm{C}_\alpha[Z] + \frac{\ln\frac{1}{\delta}}{\eta n}\right] &= P\left[e^{n\eta W} \ge \frac{1}{\delta}\right] \\
&= P\left[e^{n\eta W} \ge \frac{1}{\delta}\middle|\mathcal{E}\right]\cdot P[\mathcal{E}] \\
&\quad + P\left[e^{n\eta W} \ge \frac{1}{\delta}\middle|\mathcal{E}^{\mathrm{c}}\right]\cdot(1 - P[\mathcal{E}]), \\
&\le \delta\,\mathrm{E}[e^{n\eta W} \mid \mathcal{E}] + \delta, \qquad (46) \\
&\le 2\delta, \qquad \text{(by definition of } \mathcal{E}) \qquad (47)
\end{aligned}
$$

where (46) follows by Markov's inequality and (45). Now, we can re-express (47) as

$$
\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \le \left(\frac{3\eta}{5\alpha} + \epsilon_n\right)\mathrm{C}_\alpha[Z] + \frac{\ln\frac{1}{\delta}}{\eta n},
$$

with probability at least $1 - 2\delta$. By setting $\eta = \sqrt{\frac{5\alpha\ln\frac{1}{\delta}}{3n\mathrm{C}_\alpha[Z]}} \wedge \alpha/2$ (which does not depend on the samples), we get

$$
\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \le \epsilon_n \mathrm{C}_\alpha[Z] + \sqrt{\frac{12\mathrm{C}_\alpha[Z]\ln\frac{1}{\delta}}{5\alpha n}} \vee \frac{3\ln\frac{1}{\delta}}{\alpha n},
$$

with probability at least $1 - 2\delta$. ■

A.9 Proof of Theorem 10

Proof Let $\bar{Z} = Z - \mathrm{E}[Z]$. Suppose that Z is (σ, b)-sub-exponential. In this case, by Lemma 9 the random variable $Y \coloneqq \bar{Z}\cdot\mathrm{E}[q_\star \mid \bar{Z}]$ satisfies (24), and so by [Wainwright, 2019, Theorem 2.19], we have

$$
P\left[\mathrm{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i \ge t\right] \le \begin{cases} 2e^{-\frac{n\alpha^2 t^2}{2\sigma^2}}, & \text{if } 0 \le t \le \frac{\sigma^2}{b\alpha}; \\ 2e^{-\frac{n\alpha t}{2b}}, & \text{if } t > \frac{\sigma^2}{b\alpha}. \end{cases} \tag{48}
$$

For any real random variables A, B, and C, we have $[A \ge C] \implies [A \ge B \text{ or } B \ge C]$, and so $P[A \ge C] \le P[A \ge B] + P[B \ge C]$. Applying this with $A = \mathrm{C}_\alpha[\bar{Z}] - \widehat{\mathrm{C}}_\alpha[\bar{Z}] - |\bar{Z}_{(\lceil n\alpha\rceil)}|\epsilon_n$, $B = \mathrm{E}[Y] - \sum_{i=1}^n Y_i/n$, and $C = t \in \mathbb{R}$. we get:

$$
\begin{aligned}
P\left[\mathrm{C}_\alpha[\bar{Z}] - \widehat{\mathrm{C}}_\alpha[\bar{Z}] - |\bar{Z}_{(\lceil n\alpha\rceil)}|\cdot\epsilon_n \ge t\right] &\le P\left[\mathrm{C}_\alpha[\bar{Z}] - \widehat{\mathrm{C}}_\alpha[\bar{Z}] - |\bar{Z}_{(\lceil n\alpha\rceil)}|\cdot\epsilon_n \ge \mathrm{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i\right] \\
&\quad + P\left[\mathrm{E}[Y] - \frac{1}{n}\sum_{i=1}^n Y_i \ge t\right], \\
&\le \delta + \begin{cases} 2e^{-\frac{n\alpha^2 t^2}{2\sigma^2}}, & \text{if } 0 \le t \le \frac{\sigma^2}{b\alpha}; \\ 2e^{-\frac{n\alpha t}{2b}}, & \text{if } t > \frac{\sigma^2}{b\alpha}, \end{cases} \qquad (49)
\end{aligned}
$$

where the last inequality follows by (48) and the fact that (23) (with Z replaced by $\bar{Z}$) holds with probability at least $1 - \delta$. Since $\mathrm{C}_\alpha[Z]$ [resp. $\widehat{\mathrm{C}}_\alpha[Z]$] is a coherent risk measure, we have $\mathrm{C}_\alpha[\bar{Z}] = \mathrm{C}_\alpha[Z] - \mathrm{E}[Z]$ [resp. $\widehat{\mathrm{C}}_\alpha[\bar{Z}] = \widehat{\mathrm{C}}_\alpha[Z] - \mathrm{E}[Z]$], and so the LHS of (49) is equal to

$$
P\left[\mathrm{C}_\alpha[Z] - \widehat{\mathrm{C}}_\alpha[Z] \ge t + |\bar{Z}_{(\lceil n\alpha\rceil)}|\cdot\epsilon_n\right].
$$

This with the fact that $\bar{Z}_{(\lceil n\alpha\rceil)} = Z_{(\lceil n\alpha\rceil)} - \mathrm{E}[Z]$ completes the proof for the sub-exponential case.

When Z is σ-sub-Gaussian case, the proof is the same, except that we replace b by 0 and use the convention that $0/0 = +\infty$. ■

B Beyond CVaR

First, we give a formal definition of a coherent risk measure (CRM):

Definition 14. *We say that* $\mathrm{R}\colon \mathcal{L}^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ *is a coherent risk measure if, for any* $Z, Z' \in \mathcal{L}^1(\Omega)$ *and* $c \in \mathbb{R}$*, it satisfies the following axioms: (Positive Homogeneity)* $\mathrm{R}[\lambda Z] = \lambda \mathrm{R}[Z]$*, for all* $\lambda \in (0,1)$*; (Monotonicity)* $\mathrm{R}[Z] \leq \mathrm{R}[Z']$ *if* $Z \leq Z'$ *a.s.; (Translation Equivariance)* $\mathrm{R}[Z + c] = \mathrm{R}[Z] + c$*; (Sub-additivity)* $\mathrm{R}[Z + Z'] \leq \mathrm{R}[Z] + \mathrm{R}[Z']$.

It is known that the conditional value at risk is a member of a class of CRMs called φ-entropic risk measures Ahmadi-Javid [2012]. These CRMs are often used in the context of robust optimization Namkoong and Duchi [2017], and are perfect candidates to consider next in the context of this paper:

Definition 15. *Let* $\varphi\colon [0, +\infty[\to \mathbb{R} \cup \{+\infty\}$ *be a closed convex function such that* $\varphi(1) = 0$*. The* φ*-entropic risk measure with divergence level* c *is defined as*

$$\mathrm{ER}_\varphi^c[Z] \coloneqq \sup_{q \in \mathcal{Q}_\varphi^c} \mathrm{E}_P[Zq], \quad \textit{where}$$

$$\mathcal{Q}_\varphi^c \coloneqq \left\{ q \in \mathcal{L}^1(\Omega) \;\middle|\; \begin{array}{c} \exists Q \in \mathcal{M}_P(\Omega), q = \mathrm{d}Q/\mathrm{d}P, \\ \mathrm{D}_\varphi(Q \| P) \leq c \end{array} \right\},$$

and $\mathrm{D}_\varphi(Q \| P) \coloneqq \mathrm{E}_P[\varphi(q)]$ *is the* φ*-divergence between two distributions* Q *and* P*, where* $Q \ll P$ *and* $q = \frac{\mathrm{d}Q}{\mathrm{d}P}$.

As mentioned above, $\mathrm{CVaR}_\alpha[Z]$ is a φ-entropic risk measure; in fact, it is the φ-entropic risk measure at level $c = 0$ with $\varphi(\cdot) \coloneqq \iota_{[0,1/\alpha]}(\cdot)$, where for a set $\mathcal{C} \subseteq \mathbb{R}$, $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$; and $+\infty$ otherwise Ahmadi-Javid [2012].

The natural estimator $\widehat{\mathrm{ER}}_\varphi^c[Z]$ of $\mathrm{ER}_\varphi^c[Z]$ is defined by Ahmadi-Javid [2012]

$$\widehat{\mathrm{ER}}_\varphi^c[Z] = \inf_{\nu > 0, \mu \in \mathbb{R}} \left\{ \mu + \nu \, \mathrm{E}_{\widehat{P}_n} \left[\varphi^\star \left(\frac{Z - \mu}{\nu} - c \right) \right] \right\}.$$

Extending the results of Lemmas 2 and 3 comes down to finding an auxiliary estimator $\widetilde{\mathrm{ER}}_\varphi^c[Z]$ of $\mathrm{ER}_\varphi^c[Z]$ which satisfies (as in Lemma 3) $\widetilde{\mathrm{ER}}_\varphi^c[Z] \leq \widehat{\mathrm{ER}}_\varphi^c[Z] \cdot (1 + \epsilon_n)$, for some "small" ϵ_n, and

$$\frac{1}{n} \sum_{i=1}^n Z_i \cdot \mathrm{E}[q_\star \mid Z_i] \leq \widetilde{\mathrm{ER}}_\varphi^c[Z],$$

with high probability, where $q_\star \in \operatorname{argmin}_{q \in \mathcal{Q}_\varphi^c} \mathrm{E}[Zq]$. The similarities between the expressions of $\widehat{\mathrm{ER}}_\varphi^c[Z]$ and $\widehat{\mathrm{C}}_\alpha[Z]$ hint that it might be possible to find such an estimator by carefully constructing a set $\widetilde{\mathcal{Q}}_\varphi^c$ to play the role of the $\widetilde{\mathcal{Q}}_\alpha$ in Section 4. We leave such investigations for future work.

Chapter 7

Concentration Inequalities for CVaR with Near-optimal Quantile Level Dependence

In many machine learning applications, practitioners seek more and more to use measures of risk other than the expectation [Lerasle et al., 2019; Williamson and Menon, 2019; Laforgue et al., 2019; Khani et al., 2019]. This relatively recent shift is driven by the desire to guard against potentially catastrophic events which happen with low enough probability that the mean performance does not capture them. For example, in the medical field, the *average* efficacy and safety of a vaccine is not necessarily the most appropriate measure; in this case, avoiding severe side effects, even if they are relatively rare, is a priority. Alternative measures of risk are also often considered in financial applications such as risk-averse portfolio management. In such applications, decision making relies on the ability to accurately estimate the true underlying risk. This estimation task becomes challenging as soon as the risk measure is not the standard expectation which, unlike others, enjoys linearity. Thus, our work adds to the recent efforts in deriving concentration bounds for alternative risk measures.

One popular risk measure is the Conditional Value at Risk (CVaR). Given a quantile level α, and a random variable X, $\mathrm{CVaR}_\alpha[X]$ measures the expectation of X conditioned on X being greater than it's $(1-\alpha)$-quantile. The popularity of CVaR is partly due to the fact that it satisfies some desirable axioms that characterize a larger class of measures called *coherent risk measures* [Artzner et al., 1999]:

Definition 7.1. *A risk measure* $\mathbf{R}$ *is coherent if for any random variables* X, Y *it satisfies sub-additivity* ($\mathbf{R}[X+Y] \leq \mathbf{R}[X] + \mathbf{R}[Y]$)*; monotonicity* ($\mathbf{R}[Y] \leq \mathbf{R}[X]$ *for* $Y \leq X$ *a.s.*)*; positive-homogeneity* ($\mathbf{R}[\lambda X] = \lambda \mathbf{R}[X]$*, for* $\lambda > 0$)*; and translation equi-variance* ($\mathbf{R}[X - c] = \mathbf{R}[X] - c$)*, for any* $c \in \mathbb{R}$.

The conditional value at risk has the additional desirable property of being *law-invariant*, in the sense that if X and Y have the same probability law, then $\mathrm{CVaR}_\alpha[X] = \mathrm{CVaR}_\alpha[Y]$, for any $\alpha \in [0,1)$. What is more, it is known that any law-invariant coherent risk measure can be written in terms of CVaR. In fact, Kusuoka [2001]

essentially shows that for any law-invariant coherent risk measure $\mathbf{R}$, there exists a compact set $\mathcal{M}$ of measures on $[0,1]$ such that

$$\mathbf{R}[X] = \sup_{\mu \in \mathcal{M}} \int_0^1 \mathrm{CVaR}_\alpha[X] \mathrm{d}\mu(\alpha), \tag{7.1}$$

for any random variable X. This representation can be used to design new coherent risk measures simply by choosing the set $\mathcal{M}$. It also provides the possibility to transfer concentration inequalities obtained for CVaR to the larger class of law-invariant coherent risk measures—offering flexibility to practitioners in risk-sensitive applications.

Motivated by the desirable properties of CVaR and its ability to parameterize the family of law-invariant coherent measures, we will focus on deriving tight concentration bounds for an estimator of CVaR. We will pay particular attention to the dependence in α in our bounds—those with optimal dependence in α will be crucial if one uses them together with the Kusuoka representation in (7.1) (which involves an integral over α) to transfer concentration inequalities to the class of law-invariant coherent risk measures. Since $\mathrm{CVaR}_\alpha[X]$ is the expectation of X, conditioned on it being greater than its $(1-\alpha)$-quantile[1], one would expect its estimation from n i.i.d. samples $X_1, \ldots, X_n$ to discard about a $(1-\alpha)$ fraction of the samples (those that are below the $(1-\alpha)$-quantile). As a result, one would expect the concentration rate of any estimator of CVaR to be at best $O(1/\sqrt{\alpha n})$. Indeed, for bounded random variables, this rate was previously achieved by Brown [2007]; Wang and Gao [2010]. However, it is not clear if the techniques used in these works can be generalized beyond the case of bounded random variables. Current CVaR concentration inequality for the unbounded case (e.g. X is sub-Gaussian or sub-exponential) all have a sub-optimal dependence in the quantile parameter α, where the concentration rate becomes $1/(\alpha\sqrt{n})$ (we discuss this in more detail in the related work's paragraph below). This rate can be much worse (for small α) compared with the $1/\sqrt{\alpha n}$ rate achieved in the bounded case.

Besides seeking a concentration bound for CVaR with the tightest possible dependence in the quantile parameter α, we are also after so-called *time-uniform* bounds; these are bounds which hold uniformly for all sample sizes $n \geq 1$ simultaneously given a fixed confidence level δ. Such bounds are desirable in many machine learning application involving, for example, stopping rules in Bandits [Jamieson et al., 2014], or risk-monotonicity (see Chapter 8). Techniques used for the expectation to obtain time-uniform bounds would yield sub-optimal rates (at least in the dependence in α) when naively applied to CVaR. We address these challenges in this chapter.

Best dependence in α for a simple example. In Section 7.1.2, we work out illustrative examples in different settings to get an idea of the best dependence in α we can expect in the concentration rates. The following Bernoulli example illustrates the pitfall leading to confidence width $1/(\alpha\sqrt{n})$ and the optimality of width $1/\sqrt{\alpha n}$.

[1] We use the convention in [Brown, 2007; Wang and Gao, 2010; Prashanth and Ghavamzadeh, 2013].

For $p \in (0,1)$, a Bernoulli-p variable X has $\mathrm{CVaR}_\alpha[X] = (p/\alpha) \wedge 1$ at quantile level $\alpha \in [0,1]$. Given i.i.d. samples $X_1, \ldots, X_n$ the natural CVaR estimator $\widehat{\mathrm{C}}_n = (\hat{p}_n/\alpha) \wedge 1$ is obtained by plugging in the empirical mean $\hat{p}_n = \sum_{i=1}^n X_i/n$. To judge the quality of this estimator, we may invoke the Central Limit Theorem (CLT) to find that $\hat{p}_n/\alpha$ closely follows $\mathcal{N}(\frac{p}{\alpha}, \frac{p(1-p)}{n\alpha^2})$, leading to confidence intervals for it of order $1/(\alpha\sqrt{n})$, which is also tight. At first sight, one may suspect that these transfer to the estimator $\widehat{\mathrm{C}}_n$. Yet closer inspection of the case distinction in the definition of $\mathrm{CVaR}_\alpha[X]$ reveals that $\widehat{\mathrm{C}}_n$ concentrates *faster* (as a function of the quantile level α). The mean of $\widehat{\mathrm{C}}_n$ tends to $\mathrm{CVaR}_\alpha[X]$ in either case. When $p > \alpha$, its variance tends to zero exponentially fast. When $p < \alpha$, its variance tends to $\frac{p(1-p)}{n\alpha^2}$, (and at most that in the boundary case $p = \alpha$). In any case, the variance of the estimator is at most $1/(n\alpha)$, and confidence intervals for $\widehat{\mathrm{C}}_n$ can/should be of order $1/\sqrt{\alpha n}$.

Related work. A concentration bound for an estimator of CVaR was first presented by Brown [2007], who considered the case of bounded random variables and achieved the optimal dependence in α for the upper deviation, but not the lower deviation. Their work was followed up by that of Wang and Gao [2010] who also considered the bounded setting and provided the optimal dependence in the parameter α. When the random variable is bounded only from one side, the concentration inequalities due to Thomas and Learned-Miller [2019b] have a sharp empirical performance, though with a sub-optimal dependence on α.

The unbounded case was recently considered by e.g. [Kolla et al., 2019b; Prashanth et al., 2020] whose analysis relies on a concentration inequality for quantiles. Bhat and Prashanth [2019] also studied the unbounded setting, basing their analysis on the concentration of the empirical Cumulative Distribution Function (CDF) around the true CDF in Wasserstein distance. These works have all considered both light and heavy-tailed distributions. Mhammedi et al. [2020c] also considered the unbounded setting, albeit they restricted their analysis to the sub-Gaussian and sub-exponential cases. As far as we know, all existing concentration inequalities for estimators of CVaR in the sub-exponential or sub-Gaussian cases have a $1/(\alpha\sqrt{n})$ term. Mhammedi et al. [2020c] conjectured that this rate could be improved to $1/\sqrt{\alpha n}$ for the standard estimator of CVaR. In this chapter, we show that this is possible up to log-factors in $1/\alpha$.

In the context of statistical learning, Soma and Yoshida [2020]; Lee et al. [2020]; Curi et al. [2020] derived generalization bounds where the objective is the CVaR of a loss instead of the standard expectation. However, these bounds have a sub-optimal dependence in the quantile α. The first (PAC-Bayesian) generalization bound for CVaR with an optimal α dependence was presented by Mhammedi et al. [2020c] who, like us, also considered a reduction to estimating expectations from empirical means. However, their generalization inequality is restricted to bounded random variables. Finally, much work has been done in the context of best CVaR-arm identification in multi-armed bandits [Galichet et al., 2013; Kagrecha et al., 2019a,b; Tamkin et al., 2019; Torossian et al., 2019; Prashanth et al., 2020; Agrawal et al., 2020b; Baudry et al.,

2020], as well as for the regret minimisation problems Cassel et al. [2018].

Contributions. In this chapter, we improve almost uniformly on all the results just mentioned in terms of the dependence on the quantile parameter α in the bounds. We achieve this by deriving a new reduction (and extending an existing one) for estimating the conditional value at risk—making CVaR estimation as easy as estimating an expectation from empirical means. Using the classical Bennett's inequality in these reductions, we are able to achieve state-of-the-art concentration bounds for an estimator of CVaR. We recover and improve bounds that have the optimal dependence in α in the bounded setting, and we present the *first* bounds with a concentration rate of $1/\sqrt{\alpha n}$ (up to log factors in $1/\alpha$), which hold when the random variable of interest has light tails (e.g. sub-exponential). This provides a positive answer to a conjecture posed by [Mhammedi et al., 2020c]. We also tackle the case of heavy-tailed distributions, achieving the optimal rate where the random variable has a finite second moment. What is more, we present a new time-uniform Bernstein inequality (see Proposition 7.22) which, together with the reductions we derive, allows us to seamlessly obtain the first time-uniform concentration bounds for CVaR.

As an application, we show how our bounds can be applied in multi-arm bandits where the goal is to select an arm with the lowest CVaR. We consider both the fixed confidence and fixed budget settings and provide algorithms with state-of-the-art guarantees when the random variables are unbounded. Our time uniform bounds are crucial in the fixed confidence setting to achieve optimal sample complexity. We achieve the later with a slight extension of the lil' UCB algorithm by Jamieson et al. [2014]. For the fixed budget, we use our bounds together with the sequential halving algorithm to achieve state-of-the-art bounds.

Layout. In Section 7.1, we introduce the setting and notation. In Section 7.2, we start by presenting the reduction for CVaR estimation, then we state the bounds we get for the upper and lower deviation bounds in Subsections 7.2.1 and 7.2.2, respectively. Along the way we instantiate our results to popular settings such as when X is bounded or sub-exponential. In Section 7.3, we apply our results to the problem of multi-armed bandits in both the fixed confidence and fixed budget settings. Additional inequalities which are omitted from Section 7.2 are provided in Section 7.4. The proofs of the main results are postponed to Section 7.5. We conclude with a discussion in Section 7.6.

7.1 Preliminaries

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. We denote by $\mathcal{M}_{\mathbf{P}}(\Omega)$ the set of probability measures that are absolutely continuous with respect to $\mathbf{P}$, i.e. for any $\mathbf{Q} \in \mathcal{M}_{\mathbf{P}}(\Omega)$, we have $\mathbf{Q} \ll \mathbf{P}$. Given a random variable X, we denote by $\mathbf{P}_X$ its probability law, and

given samples $X_1, \ldots, X_n$ of X, we denote by $\widehat{\mathbf{P}}_n$ the empirical distribution defined by

$$\widehat{\mathbf{P}}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(x), \quad \text{for all } x \in \mathbb{R},$$

where δ_y is the Dirac delta at $y \in \mathbb{R}$. To simplify notation, we will use $X_{1:n} := (X_1, \ldots, X_n)$. The Conditional Value at Risk (CVaR) of a real random variable X at (upper) quantile level $\alpha \in (0,1)$ is

$$\mathrm{CVaR}_\alpha[X] := \mathbb{E}_{\mathbf{P}_X(x)}[X \mid X \geq x_\alpha], \quad \text{where} \quad x_\alpha := \inf\{x : \mathbf{P}_X([x, +\infty)) \leq \alpha\}. \tag{7.2}$$

For any probability measure μ over the reals and $Y \sim \mu$, we define $\mathrm{CVaR}_\alpha[\mu] := \mathrm{CVaR}_\alpha[Y]$ ($\mathrm{CVaR}_\alpha[\mu]$ is well defined since CVaR is law-invariant). In particular, we have $\mathrm{CVaR}_\alpha[\mathbf{P}_X] = \mathrm{CVaR}_\alpha[X]$. $\mathrm{CVaR}_\alpha[\mathbf{P}_X]$ and $\mathrm{CVaR}_\alpha[\widehat{\mathbf{P}}_n]$ admit the following useful formulations [Rockafellar and Uryasev, 2013]:

$$\mathrm{CVaR}_\alpha[\mathbf{P}_X] = \inf_{\mu \in \mathbb{R}} \left\{ \mu + \frac{\mathbb{E}[[X-\mu]_+]}{\alpha} \right\}, \tag{7.3a}$$

$$\text{and} \;\; \mathrm{CVaR}_\alpha[\widehat{\mathbf{P}}_n] = \inf_{\mu \in \mathbb{R}} \left\{ \mu + \frac{\sum_{i=1}^{n}[[X_i-\mu]_+]}{\alpha n} \right\}, \tag{7.3b}$$

where $[x]_+ := 0 \vee x$, for $x \in \mathbb{R}$. We note that the minimizer $\mu = x_\alpha$ is the quantile at upper level α. We also use the dual representation of CVaR [Rockafellar and Uryasev, 2013]:

$$\mathrm{CVaR}_\alpha[\mathbf{P}_X] = \sup_{Q \in \mathcal{Q}} \mathbb{E}[XQ], \;\; \mathcal{Q} := \left\{ Q = \frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}} \leq \frac{1}{\alpha}, \; \mathbf{Q} \in \mathcal{M}_{\mathbf{P}}(\Omega), \; \mathbb{E}[Q] = 1 \right\}, \tag{7.4a}$$

$$\mathrm{CVaR}_\alpha[\widehat{\mathbf{P}}_n] = \sup_{Q \in \widehat{\mathcal{Q}}_n} \frac{1}{n}\sum_{i=1}^{n} X_i Q_i, \quad \widehat{\mathcal{Q}}_n := \left\{ Q_{1:n} \in [0, 1/\alpha]^n, \; \frac{1}{n}\sum_{i=1}^{n} Q_i = 1 \right\}, \tag{7.4b}$$

where $\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P}$ represents the Radon Nikodym derivative. Here we remark that the optimizer Q satisfies $\mathbb{E}[Q \mid X = x] = \frac{1}{\alpha}\mathbb{I}\{x \geq x_\alpha\}$ for $x \neq x_\alpha$, and $\mathbb{E}[Q \mid X = x_\alpha] \in [0, 1/\alpha]$ (an intermediate value here is referred to as *atom splitting*). Throughout the rest of this chapter, we will consider a fixed upper quantile level $\alpha \in (0,1)$ and use the concise notation:

$$\mathrm{C}[X] := \mathrm{CVaR}_\alpha[\mathbf{P}_X] \quad \text{and} \quad \widehat{\mathrm{C}}_n[X] := \mathrm{CVaR}_\alpha\left[\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\right].$$

We now discuss some standard distributional assumptions we will make on the random variables X whose CVaR we seek to estimate.

7.1.1 Distributional Assumptions

Apart from the case where X is bounded, we will focus on the following two cases that include both light and heavy-tailed distributions: for some parameters $v, \lambda, p > 0$,

$$\mathbb{E}[e^{|X|^p/\lambda}] < v, \qquad \text{(uncentered light tail condition)}, \tag{7.5}$$

$$\text{or} \quad \mathbb{E}[|X|^p] < v, \qquad \text{(uncentered heavy tail condition)}. \tag{7.6}$$

For the heavy tailed case in (7.6) it only makes sense to consider $p > 1$. The condition in (7.5) covers both the sub-exponential and (uncentered) sub-Gaussian cases; the sub-exponential [resp. sub-Gaussian] case corresponds to the setting where $p = 1$ [resp. $p = 2$] in (7.5). Prashanth et al. [2020]; Bhat and Prashanth [2019] make similar assumptions with the difference that Bhat and Prashanth [2019] consider the centered case, where (7.6) hold with $|X - \mathbb{E}[X]|$ instead of $|X|$. In Section 7.6, we discuss how our approach can be extended to include this case. By an application of Markov's inequality, (7.5) and (7.6), imply respectively,

$$\mathbf{P}[|X| \geq t] \leq v e^{-\frac{t^p}{\lambda}}, \tag{7.7}$$

$$\text{and} \quad \mathbf{P}[|X| \geq t] \leq v t^{-p}, \tag{7.8}$$

for all $t > 0$. It will be more convenient to work with the tail probabilities from (7.7) and (7.8), instead of the moment inequalities in (7.5) and (7.6).

Remark 7.1. *Knowing the parameters v, p, λ is a common assumption in the heavy-tailed literature, where the uncentred assumption is also typical, see e.g. the trimmed mean estimator discussed by Bubeck et al. [2013]. Note that one should not think of the modelling assumption as a tuning knob: for the results to be meaningful the assumptions have to be* satisfied *for the distribution in question. For that reason one wants to pick the smallest class for which one can ensure that it contains the data-generating distribution. In applications, knowledge of p, v and λ can come either from modelling assumptions or from earlier estimates.*

In the following subsection, we work through some examples to build more intuition about CVaR for light and heavy-tailed distributions. The examples will also help us calibrate our expectations for what dependencies in α we can achieve in the concentration rates of our CVaR estimators.

7.1.2 Calibrating our Expectations about Empirical CVaR Concentration

In this section, we aim to develop intuition about the concentration behavior of the empirical CVaR estimator $\widehat{\mathrm{C}}_n[X]$ by investigating the standard Gaussian, Exponential and Pareto cases. To simplify matters, we take inspiration from the sandwich provided by the primal (7.3) and dual (7.4) formulation to study two empirical averages

with mean $\mathrm{CVaR}_\alpha[X]$ each, which closely constrain $\widehat{\mathrm{C}}_n[X]$, namely

$$L := \frac{1}{n}\sum_{i=1}^{n} \frac{X_i \mathbb{I}\{X_i \geq \mu_\alpha\}}{\alpha} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{X_i \mathbb{I}\{X_i \geq \hat{\mu}_\alpha\}}{\alpha} \leq \widehat{\mathrm{C}}_n[X]$$
$$\leq \mu_\alpha + \frac{\frac{1}{n}\sum_{i=1}^{n}[X_i - \mu_\alpha]_+}{\alpha} =: U,$$

where $\hat{\mu}_\alpha$ is the minimizer of the objective on the RHS of (7.3). The first and second sums are indeed approximately equal with high probability; this is at the core of the proof of our main Theorem 7.3 in the next section, and so we are happy to take it for granted here. Furthermore, $\mathbb{E}[L] = \mathbb{E}[U] = \mathrm{CVaR}_\alpha[X]$ by definition of CVaR. After this reduction to an i.i.d. problem, it remains to compute the variance of each term in the outermost sum, and invoke the CLT to get an idea for the correct confidence widths (or at least those attainable by methods invoking this sandwich internally). So let's compute. The standard exponential distribution has $(1-\alpha)$-quantile $\mu_\alpha = \ln(1/\alpha)$ and $\mathrm{CVaR}_\alpha[X] = 1 + \ln(1/\alpha)$. The variances are then equal to

$$\textsc{Var}[L] = \frac{2 + (1-\alpha)(\ln(1/\alpha)+2)\ln(1/\alpha)}{\alpha} - 1 \quad \text{and} \quad \textsc{Var}[U] = \frac{2}{\alpha} - 1. \tag{7.9}$$

The standard Gaussian distribution has $(1-\alpha)$-quantile $\mu_\alpha = \Phi^{-1}(1-\alpha)$ and $\mathrm{CVaR}_\alpha[X] = \frac{1}{2\pi\alpha}e^{-\frac{1}{2}\mu_\alpha^2}$. Exact expressions for the variance are unintelligible, but a computer-assisted symbolic analysis reveals that

$$\textsc{Var}[L] \sim \frac{2\ln(1/\alpha)}{\alpha} \qquad \text{and} \qquad \textsc{Var}[U] \sim \frac{1}{\alpha\ln(1/\alpha)}, \tag{7.10}$$

where $f(\alpha) \sim g(\alpha)$ denotes $\lim_{\alpha\to 0} f(\alpha)/g(\alpha) = 1$. Finally, the standard Pareto distribution (with density $px^{-1-p}\mathbb{I}\{x \geq 1\}$) has α quantile given by $\mu_\alpha = \alpha^{-1/p} > 1$ and $\mathrm{CVaR}_\alpha[X] = \frac{p\alpha^{-1/p}}{p-1}$ (requiring $p > 1$ to exist). We then find (this requires $p > 2$)

$$\textsc{Var}[L] = \frac{\frac{p}{p-2} + (1-\alpha)p^2}{\alpha^{1+\frac{2}{p}}(p-1)^2} \quad \text{and} \quad \textsc{Var}[U] = \frac{\frac{2}{p-2} + 2 - \alpha}{\alpha^{1+\frac{2}{p}}(p-1)^2}. \tag{7.11}$$

In the Gaussian and Exponential cases, the i.i.d. approximate sandwich $L \lesssim \widehat{\mathrm{C}}_n[X] \leq U$; the fact that $\mathbb{E}[L] = \mathbb{E}[U] = \mathrm{CVaR}_\alpha[X]$; and the variance approximations in (7.9) and (7.10) strongly suggest we should find confidence interval widths for $\mathrm{CVaR}_\alpha[X]$ of order $1/\sqrt{\alpha n}$ up to lower-order $\ln(1/\alpha)$ factors around $\widehat{\mathrm{C}}_n[X]$. We do indeed derive confidence intervals with such widths in the light tail case (see Section 7.2). In the Pareto case, (7.11) suggests confidence widths of order $\alpha^{-\frac{1}{2}-\frac{1}{p}}$, and ours will be of order α^{-1}, which match $\alpha^{-\frac{1}{2}-\frac{1}{p}}$ for $p = 2$.

In the next section, we start by presenting the main reductions we use to derive our new concentration inequalities for an estimator of CVaR. One of these reductions

leads to a Bernstein-like inequality for CVaR (Theorem 7.3), which is of independent interest. In Sections 7.2.1 and 7.2.2, we apply our reductions to obtain upper and lower deviations, respectively, under different distributional assumptions.

7.2 New Concentration Inequalities for CVaR

In this section, we derive new concentration inequalities for an estimator of the CVaR of a random variable X at quantile level $\alpha \in (0,1)$. We study the cases where X is bounded (in $[0,1]$) or unbounded with a light or heavy-tailed distribution as in (7.7) and (7.8), respectively, and derive state-of-the-art concentration inequalities with the optimal dependence in the quantile level α (up to a log-factor) in some of the cases (as suggested by our examples in Section 7.1.2). We will also derive so-called time-uniform concentration inequalities which bound the deviation of an estimator of the CVaR for all sample sizes simultaneously given a fixed confidence.

Behind our results are two simple reductions to estimating expectations from empirical samples. Reduction #1 [resp. #2] relies on the primal [resp. dual] representation of CVaR in (7.3) [resp. (7.4)], and is used to derive upper [resp. lower] deviation bounds. We now describe these reductions:

Reduction #1. The first reduction we use to derive our upper-deviation bounds relies on the "primal" representation of CVaR given in (7.3). To describe this reduction, let $B > 0$ and $Z, Z_1, \ldots, Z_n \in [-B, B]$ be i.i.d. random variables. Then, for any $\delta \in (0,1)$, and $\mu_\alpha \in \arg\min_{\mu \in \mathbb{R}} \{\mu + \mathbb{E}[Z - \mu]_+ / \alpha\}$, we have by (7.3) and Bennett's inequality (see e.g. [Maurer and Pontil, 2009, Theorem 3])

$$\widehat{\mathrm{C}}_n[Z] \overset{(*)}{\leq} \mu_\alpha + \frac{\sum_{i=1}^n [Z_i - \mu_\alpha]_+}{\alpha n} \overset{(**)}{\leq} \mathrm{C}[Z] + \sqrt{\frac{2\mathbb{E}[(\mu_\alpha + [Z - \mu_\alpha]_+/\alpha)^2] \ln \delta^{-1}}{n}} - \frac{B \ln \delta}{3\alpha n}, \tag{7.12}$$

with probability at least $1 - \delta$. The inequality $^{(*)}$ follows by the representation of $\widehat{\mathrm{C}}_n[Z]$ in (7.3), while $^{(**)}$ follows by applying the concentration inequality to the average of the i.i.d. random variables $\mu_\alpha + [Z_i - \mu_\alpha]_+/\alpha, i \in [n]$, which are bounded by B/α. The inequality in (7.12) implies that, as long as the random variable Z is bounded, bounding the expectation $\mathbb{E}[(\mu_\alpha + [Z - \mu_\alpha]_+/\alpha)^2]$ is sufficient to obtaining an upper deviation bound for $\widehat{\mathrm{C}}_n[Z]$. When dealing with an unbounded random variable X, we will resort to clipping, which will reduce the problem to the bounded case with Z being the new clipped random variable. However, this requires a careful choice of the clipping threshold to control the magnitude of the term $\mathbb{E}[(\mu_\alpha + [Z - \mu_\alpha]_+/\alpha)^2]$ in (7.12), on the one hand, and to control the difference between $\mathrm{C}[Z]$ and $\mathrm{C}[X]$, on the other—after all, we want $\mathrm{C}[X]$ and not $\mathrm{C}[Z]$ to appear on the RHS of our deviation bounds. We present the choice of thresholds in the next subsection. We note that the choice of Bennett's in the current reduction (to get to (7.12)) is crucial in obtaining the right dependencies in α in the final concentration inequalities. For example, when the random variable of interest is bounded, Hoeffding's inequality

would yield a $1/(\alpha\sqrt{n})$ rate instead of the optimal $1/\sqrt{\alpha n}$ rate we achieve using Bennett's inequality.

Finally, to obtain a time-uniform deviation bound, we will derive a version of (7.12) that holds for all sample sizes simultaneously given a fixed confidence level δ (see Proposition 7.22).

Reduction #2. Our second reduction for the lower deviation bounds (which is slightly more involved than the previous one) builds on a recent technique by Mhammedi et al. [2020c] that is based on the dual representations of CVaR in (7.4). To describe this technique, we introduce the implicit (unobserved) random variable Y defined by:

$$Y := X \cdot \mathbb{E}[Q_\star \mid X], \quad \text{where} \quad Q_\star \in \operatorname*{arg\,max}_{Q \in \mathcal{Q}} \mathbb{E}[XQ], \tag{7.13}$$

and $\mathcal{Q}$ is as in (7.4a). The reason that this is helpful is that the expectation of Y is the CVaR of X:

Lemma 7.1. *Let $\alpha \in (0,1)$ and $X, X_1, \ldots, X_n$ be i.i.d. random variables in $\mathbb{R}$ and $Q_\star$ be as in* (7.13)*. Then, the random variable Y in* (7.13) *and $Y_i := X_i \cdot \mathbb{E}[Q_\star \mid X_i], i \in [n]$, are i.i.d. and satisfy*

$$\mathbb{E}[Y] = \mathbb{E}[Y_i] = \mathrm{C}[X], \quad \textit{for all } i \in [n].$$

The result in Lemma 7.1 is crucial as it shows that the expectation of the random variable Y in (7.13) is exactly equal to the CVaR of X (this follows by the law of total expectation). This means that one could attempt to estimate $\mathrm{C}[X]$ using the average $\sum_{i=1}^n Y_i/n$. The challenge here is that (Y_i) are *implicit* random variables, since they depend on the unknown $Q_\star$ in (7.13). The trick to overcome this is to bound (with high probability) the average $\sum_{i=1}^n Y_i/n$ by a quantity that depends only on empirical samples. This is enabled by the next lemma due to Mhammedi et al. [2020c], where we use the following notation: for $X_{1:n} \in \mathbb{R}^n$ and $\epsilon > 0$, we define

$$\begin{aligned} \widehat{\mathrm{C}}_n[X;\epsilon] &:= \sup_{Q_{1:n} \in \widehat{\mathcal{Q}}_n(\epsilon)} \frac{1}{n}\sum_{i=1}^n X_i Q_i, \\ \text{where} \quad \widehat{\mathcal{Q}}_n(\epsilon) &:= \left\{ Q_{1:n} \in [0, 1/\alpha]^n, \left| \frac{1}{n}\sum_{i=1}^n Q_i - 1 \right| \le \epsilon \right\}. \end{aligned} \tag{7.14}$$

Lemma 7.2. *Let $\alpha \in (0,1)$, $\epsilon > 0$, and $n \in \mathbb{N}$. Then, for any $X_{1:n} \in \mathbb{R}^n$, and $\widehat{\mathrm{C}}_n[X;\epsilon]$ as in* (7.14)*,*

$$\widehat{\mathrm{C}}_n[X;\epsilon] = \inf_{\mu \in \mathbb{R}} \left\{ \mu + |\mu|\epsilon + \frac{\sum_{i=1}^n [X_i - \mu]_+}{\alpha n} \right\} \le \widehat{\mathrm{C}}_n[X] + |X_{(\lceil \alpha n \rceil)}| \cdot \epsilon, \tag{7.15}$$

where $X_{(1)}, \ldots, X_{(n)}$ are the decreasing order statistics of $X_1, \ldots, X_n$.

Note that from (7.4b), we have $\widehat{\mathrm{C}}_n[X;0] = \widehat{\mathrm{C}}_n[X]$. To see how Lemma 7.2 can be used to bound $\sum_{i=1}^n Y_i/n$ observe that, by definition of $\widehat{\mathrm{C}}_n[X;\cdot]$ in (7.14), we have for $Q_{\star,i} := \mathbb{E}[Q_\star \mid X_i]$,

$$\frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n X_i Q_{\star,i} \le \widehat{\mathrm{C}}_n[X;\epsilon], \quad \text{for any} \quad \epsilon \ge \left|\frac{1}{n}\sum_{i=1}^n Q_{\star,i} - 1\right|. \tag{7.16}$$

The next step is to leverage the fact that $\sum_{i=1}^n Q_{\star,i}/n$ concentrates around 1, and so by applying an appropriate concentration inequality (as we do in this chapter), we can pick an $\epsilon \le O(1/\sqrt{\alpha n})$—which does not depend on any implicit variables—for which (7.16) holds with high probability. By combining this with the right-most inequality in (7.15), we can bound $\sum_{i=1}^n Y_i/n$ from above by $\widehat{\mathrm{C}}_n[X] + |X_{(\lceil \alpha n \rceil)}| \cdot \epsilon_n$ with high probability.

What remains to do is to carefully select concentration inequalities to bound $\mathbb{E}[Y] - \sum_{i=1}^n Y_i/n$ and $|\sum_{i=1}^n Q_{\star,i}/n - 1|$ in a way to obtain optimal (or close to optimal) dependence in α in the final bound. For this, we apply Bennett's inequality and a new time-uniform Bernstein inequality (to obtain the time-uniform version of our bounds) to control the deviations of interest. The next problem is that the error terms in these bound will depend on the second moment of Y, which in turn depends on the unknown $Q_\star$ in (7.13). To overcome this challenge, we show that $\mathbb{E}[Y^2] \le \mathrm{C}[X^2]/\alpha$. Finally, for each of the distribution types we consider (those who satisfy (7.7) and (7.8)), we derive explicit bounds for $\mathrm{C}[X^2]$, leading to sharp dependencies in α in the final concentration inequalities (matching our expectations from Section 7.1.2). We now present our master theorem, which follows from the reduction just described—the result may be viewed as the CVaR version of the classical Bernstein inequality:

Theorem 7.3. *Let $\alpha, \delta \in (0,1)$, $B > 0$, $\rho > 1$, and $X, X_1, X_2, \cdots \in \mathbb{R}$ be i.i.d. random variables. Further, for $c := \sum_{k=2}^\infty 2/(k \ln^2(k))$, define $\phi_\rho(n) := c \ln_\rho(\rho n) \ln^2(\ln_\rho(\rho n))$;*

$$\epsilon_n := \sqrt{\frac{2\ln\delta^{-1}}{\alpha n}} + \frac{\ln\delta^{-1}}{3n}; \quad \textit{and} \quad \epsilon_n' := \sqrt{\frac{2\rho \ln\frac{\phi_\rho(n)}{\delta}}{\alpha n}} + \frac{2\ln\frac{\phi_\rho(n)}{\delta}}{3\alpha n}. \tag{7.17}$$

Then, (I) *for any* fixed $n \in \mathbb{N}$, *the random variable $Z := X \cdot \mathbb{I}\{|X| \le B\}$ satisfies,*

$$\mathbf{P}\left[\mathrm{C}[Z] - \widehat{\mathrm{C}}_n[Z;\epsilon_n] \le \sqrt{\frac{2\mathrm{C}[Z^2]\cdot\ln\delta^{-1}}{\alpha n}} + \frac{B\ln\delta^{-1}}{3\alpha n}\right] \ge 1 - 2\delta, \tag{7.18}$$

(II) *for any increasing $(B_n) \subset \mathbb{R}_{\ge 0}$ and $Z^{(n)} := X \cdot \mathbb{I}\{|X| \le B_n\}$,*

$$\mathbf{P}\left[\forall n \ge 1,\ \mathrm{C}[Z^{(n)}] - \widehat{\mathrm{C}}_n[Z^{(n)};\epsilon_n'] \le \sqrt{\frac{2\rho\mathrm{C}[(Z^{(n)})^2]\ln\frac{\phi_\rho(n)}{\delta}}{\alpha n}} + \frac{2B_n\ln\frac{\phi_\rho(n)}{\delta}}{3\alpha n}\right] \ge 1 - 2\delta. \tag{7.19}$$

We stress that (7.18) only holds for a sample size n chosen *prior* to seeing the data, whereas (8.16) holds for any n in *hindsight*. Hence, the latter is a time-uniform concentration inequality. The additional price we pay in (8.16) compared with (7.18) is a $O(\ln(\ln n)/\delta)$ term instead of $\ln(1/\delta)$. In Section 7.2.2, we will instantiate Theorem 7.3 under different assumptions on the distribution of X.

7.2.1 New Upper Deviation Bounds

In this subsection, we will use Reduction #1 described above to derive new upper deviation bounds for an estimator of CVaR. To apply the reduction, we need to **I**) relate $\mathrm{C}[Z]$ to $\mathrm{C}[X]$, where Z is a potentially clipped version of X, and **II**) bound the term $\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2]$ inside the square-root in (7.12).

Warm up. We start with the case of a bounded random variable $X \in [0,1]$. In this case, setting the clipping threshold to the trivial $B = 1$, we have $Z = X$, and so $\mathrm{C}[X] = \mathrm{C}[Z]$. It remains to bound $\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2]$. This task is relatively easy in the case since by definition of μ_α (recall that $\mu_\alpha \in \arg\min_{\mu\in\mathbb{R}}\{\mu + \mathbb{E}[Z-\mu]_+/\alpha\}$) and the fact that $X \in [0,1]$, we have that $\mu_\alpha \in [0,1]$. As a result, we get the following bound

$$\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2] \le \mathbb{E}[\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+]]/\alpha = \mathrm{C}[X]/\alpha,$$

where the equality follows by (7.3). This together with (7.12), and the time-uniform Bernstein inequality we derive in Proposition 7.22, yield the following deviation bounds:

Theorem 7.4. *Let $\alpha \in (0,1)$ and $X, X_1, \ldots, X_n$ be i.i.d. random variables in $[0,1]$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\widehat{\mathrm{C}}_n[X] \le \mathrm{C}[X] + \sqrt{\frac{2\mathrm{C}[X]\ln\delta^{-1}}{\alpha n}} + \frac{\ln\delta^{-1}}{3\alpha n}. \tag{7.20}$$

Further, for any $\rho > 1$, and $\phi_\rho(\cdot)$ as in Theorem 7.3, we have with probability at least $1-\delta$,

$$\forall n \ge 1, \quad \widehat{\mathrm{C}}_n[X] \le \mathrm{C}[X] + \sqrt{\frac{2\rho\mathrm{C}[X]\ln\frac{\phi_\rho(n)}{\delta}}{\alpha n}} + \frac{2\ln\frac{\phi_\rho(n)}{\delta}}{3\alpha n}. \tag{7.21}$$

We stress the ease with which we were able to derive these inequalities (in contrast with e.g. Brown [2007]; Wang and Gao [2010]). Compared to existing bounds, ours have the advantage that the main square-root error term has a $\mathrm{C}[X] \in [0,1]$. A similar bound to (7.20), though for the lower deviation, was recently presented by Mhammedi et al. [2020c]. The bounds in (7.20) and (7.21) can easily be rearranged (using Lemma A.1) so that the empirical CVaR appears inside the square-root error term (in place of the unknown $\mathrm{C}[X]$) (a time-uniform version of the corollary is postponed to Appendix 7.4).

Corollary 7.5. *Let $\alpha \in (0,1)$ and $X, X_1, \ldots, X_n$ be i.i.d. random variables in $[0,1]$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\widehat{\mathrm{C}}_n[X] \leq \mathrm{C}[X] + \sqrt{\frac{2\widehat{\mathrm{C}}_n[X]\ln\delta^{-1}}{\alpha n}} + \frac{8\ln\delta^{-1}}{3\alpha n}. \tag{7.22}$$

The unbounded case. We now consider the case of light and heavy-tailed distributions as characterized by (7.7) and (7.8), respectively. To derive similar deviations bounds as the ones in Theorem 7.4, for an unbounded random variable, we first need to relate the CVaR of a clipped version of X, which we denote by Z, to that of X, for an appropriately chosen clipping threshold B. Then, it remains to bound the expectation $\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2]$ in (7.16). The next two propositions answer to these needs. We now introduce a function that will be convenient in bounding the quantities of interest in (7.12) in the light-tailed case:

$$f_p(y) := 2\ln_+(v/y)^{1/p} + v\Gamma(1/p, \ln(v/y))/(yp), \quad \text{for } p, y > 0, \tag{7.23}$$

where $\Gamma(\cdot,\cdot)$ is the incomplete gamma function, and v is as in (7.5).

Proposition 7.6. *Let $v, y > 0$ and X be a real random variable. If X satisfies (7.7) with $p, v, \lambda > 0$, then for $Z := X \cdot \mathbb{I}\{|X| \leq (\lambda\ln(v/y))^{1/p}\}$ and $\mu_\alpha := \inf\{\mu \in \mathbb{R} \mid \mathbf{P}[Z \geq \mu]\}$, we have*

$$|\mathrm{C}[Z] - \mathrm{C}[X]| \leq y\lambda^{1/p} f_p(y)/\alpha, \text{ and } \mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2] \leq 4\lambda^{2/p} f_{p/2}(\alpha)/\alpha.$$

Furthermore, if X satisfies (7.8) with $p > 1$, then for $Z := X \cdot \mathbb{I}\{|X| \leq (v/y)^{1/p}\}$,

$$|\mathrm{C}[Z] - \mathrm{C}[X]| \leq \frac{kv^{1/p}y^{1-1/p}}{\alpha(p-1)}, \text{ and } \mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2 \leq 4v(v/y)^{2/p-1}/\alpha^2. \tag{7.24}$$

With this proposition and (7.12), we immediately get the following bound for the light-tailed case:

Theorem 7.7. *Let $\alpha, \delta \in (0,1)$, $\lambda, p, v > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. random variables such that X satisfies the tail probability inequality in (7.7). Further, let $f_p(\cdot)$ be as in (7.23). Then, for $n \geq 1$ and $Z := X \cdot \mathbb{I}\{X \leq (\lambda\ln(vn))^{1/p}\}$, we have, with probability at least $1-\delta$,*

$$\widehat{\mathrm{C}}_n[Z] - \mathrm{C}[X] \leq 2\lambda^{\frac{1}{p}}\sqrt{\frac{2f_{p/2}(\alpha)\ln\delta^{-1}}{\alpha n}} + \lambda^{\frac{1}{p}} \cdot \frac{f_p(1/n) - \frac{1}{3}\ln^{\frac{1}{p}}(vn)\ln\delta}{\alpha n}.$$

A time-uniform version of this inequality is stated in Theorem 7.25. A Taylor series expansion around $y = 0$ reveals that $f_p(y) = O(\mathrm{poly}(\ln(1/y)))$, for $p > 0$, and thus the deviation in Theorem 7.7 satisfies $\widehat{\mathrm{C}}_n[Z] - \mathrm{C}[X] \leq O(1/\sqrt{\alpha n})$ up to poly-log-factors in $1/\alpha$. Therefore, the bound in Theorem 7.7 has the optimal dependence on the quantile level α up to poly-log-factors (see Section 7.1.2). We achieve the same

dependence in α for the lower deviation in the next subsection. For a sub-exponential random variable X—one that satisfies (7.7) with $(v, p) = (1, 1)$—we have for $\rho \geq 1$,

$$f_{p/2}(\alpha) = 2(1 - 2\ln\alpha + \ln^2\alpha) \leq 2(1 - \ln\alpha)^2, \text{ and } f_p(1/(\rho n)) = 1 + 2\ln(\rho n). \tag{7.25}$$

This combined with Theorem 7.7 leads to the following corollary:

Corollary 7.8. *Let $\alpha, \delta \in (0, 1)$, $\lambda > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. λ-sub-exponential random variables, i.e. X satisfies* (7.7) *with $(v, p) = (1, 1)$. Then, for $Z := X \cdot \mathbb{I}\{X \leq \lambda \ln n\}$, we have*

$$\mathbf{P}\left[\widehat{\mathrm{C}}_n[Z] \leq \mathrm{C}[X] + 4\lambda \cdot (1 - \ln\alpha)\sqrt{\frac{\ln\delta^{-1}}{\alpha n}} + \lambda \cdot \frac{3 + 6\ln n - 2\ln n \ln\delta}{3\alpha n}\right] \geq 1 - \delta. \tag{7.26}$$

The time-uniform version of (7.26) is also easily obtained from Theorem 7.7 and (7.25) (see Corollary 7.21 for the statement). Using Proposition 7.6 and (7.12), we also immediately get the following bound for the heavy-tailed case (a time-uniform version is stated in Theorem 7.26):

Theorem 7.9. *Let $\alpha, \delta \in (0, 1)$, $v > 0$, $p > 1$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. random variables such that X satisfies the tail probability in* (7.8)*. Then, for $n \geq 1$ and $Z := X \cdot \mathbb{I}\{X \leq (-3pvn/\ln\delta)^{1/p}\}$, we have, with probability at least $1 - \delta$,*

$$\widehat{\mathrm{C}}_n[Z] - \mathrm{C}[X] \leq \frac{v^{\frac{1}{p}} p^{1+\frac{1}{p}}}{\alpha}\left(\frac{(1/3)^{1-\frac{1}{p}} + 2^{\frac{1}{2}} 3^{\frac{1}{p}-\frac{1}{2}}}{p-1}\right)\left(\frac{\ln\delta^{-1}}{n}\right)^{1-\frac{1}{p}}.$$

The dependence in α that we achieve in Theorem 7.9 is the best one we are aware of. [Prashanth et al., 2020, Theorem 4.1] may at first sight have the same dependence in α, but their multiplicative presence of the α-quantile μ_α carries another dependency, which may be polynomial (and not logarithmic). For example, a Pareto distribution with density $p\frac{\mathbb{I}\{x \geq 1\}}{x^{1+p}}$ has quantile $\mu_\alpha = \alpha^{-1/p}$ (see Section 7.1.2).

We now move to lower deviation bounds.

7.2.2 New Lower Deviation Bounds for CVaR

In this subsection, we use our second reduction—in particular, our Master Theorem 7.3—to derive new lower deviation bounds. To apply the theorem for our purposes, we need to **I**) relate $\mathrm{C}[Z]$ to $\mathrm{C}[X]$, where Z is a potentially clipped version of X, and **II**) bound the CVaR of Z^2. Proposition 7.6 in the previous section already takes care of **I**), and so it remains to bound $\mathrm{C}[Z^2]$.

Warm up. Suppose that the random variable X satisfies $X \in [0, 1]$. In this case, by setting $B \geq 1$ in Theorem 7.3, we get $\mathrm{C}[Z] = \mathrm{C}[X]$. What is more, since CVaR is

monotonic [Rockafellar and Uryasev, 2013], we have $\mathrm{C}[X^2] \leq \mathrm{C}[X]$. Finally, using Lemma 7.1, we have $\widehat{\mathrm{C}}_n[Z;\epsilon_n] = \widehat{\mathrm{C}}_n[X;\epsilon] \leq \widehat{\mathrm{C}}_n[X] + X_{(\lceil n\alpha \rceil)} \cdot \epsilon$, with high probability for ϵ equal to ϵ_n [resp. ϵ'_n] in (7.17). Plugging these facts into (7.18) and (8.16), we get the following corollary:

Corollary 7.10. *Let $\alpha \in (0,1)$ and $X, X_1, \ldots, X_n$ be i.i.d. random variables in $[0,1]$. Then, for any $\delta \in (0,1)$, with probability at least $1-2\delta$,*

$$\mathrm{C}[X] - \widehat{\mathrm{C}}_n[X] \leq \sqrt{\frac{2\mathrm{C}[X]\ln\delta^{-1}}{\alpha n}} + X_{(\lceil \alpha n \rceil)} \cdot \left(\sqrt{\frac{2\ln\delta^{-1}}{\alpha n}} + \frac{\ln\delta^{-1}}{3\alpha n} \right) + \frac{\ln\delta^{-1}}{3\alpha n}. \quad (7.27)$$

Further, for any $\rho > 1$, and ϵ'_n and $\phi_\rho(\cdot)$ as in Theorem 7.3, we have with probability at least $1-2\delta$,

$$\forall n \geq 1, \quad \mathrm{C}[X] - \widehat{\mathrm{C}}_n[X] \leq \sqrt{\frac{2\rho\mathrm{C}[X]\ln\frac{\phi_\rho(n)}{\delta}}{\alpha n}} + X_{(\lceil \alpha n \rceil)} \cdot \epsilon'_n + \frac{2\ln\frac{\phi_\rho(n)}{\delta}}{3\alpha n}.$$

The bounds in Corollary 7.10 and (7.21) are the first time-uniform bounds for CVaR with a $\ln\ln n$ under the square-root instead of a $\ln n$, and with the optimal dependence in α. We note also that since $X \in [0,1]$, we have $\mathrm{C}[X] \leq 1$ and $X_{(\lceil \alpha n \rceil)} \leq 1$, and so having $\mathrm{C}[X]$ and $X_{(\lceil n\alpha \rceil)}$ on the RHS of (7.27) only makes our inequality tighter than existing bounds which typical have the range of X (in this case 1) multiplying the error terms. We now present an empirical version of (7.27) which uses Lemma A.1 to rearrange the bound in Theorem 7.3 (we postpone the time-uniform variant to Appendix 7.4):

Corollary 7.11. *Let $\alpha, \delta \in (0,1)$, and $X_1, \ldots, X_n$ be i.i.d. random variables in $[0,1]$. Furthermore, let ϵ_n and $\widehat{\mathrm{C}}_n[\cdot;\cdot]$ be as in* (7.17) *and* (7.15)*, respectively. Then, with probability at least $1-2\delta$,*

$$\mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[X;\epsilon_n] + \sqrt{\frac{2\widehat{\mathrm{C}}_n[X;\epsilon_n]\cdot\ln\delta^{-1}}{\alpha n}} + \frac{8\ln\delta^{-1}}{3\alpha n}. \quad (7.28)$$

We note that, though $\widehat{\mathrm{C}}_n[X;\epsilon_n]$ can be evaluated from empirical samples using Lemma 7.1, one could also use the same lemma to bound $\widehat{\mathrm{C}}_n[X;\epsilon_n]$ from above by $\widehat{\mathrm{C}}_n[X] + |X_{(\lceil \alpha n \rceil)}| \cdot \epsilon_n \leq \widehat{\mathrm{C}}_n[X](1+\epsilon_n)$. Plugging this into (7.28) would yield a bound in terms of $\widehat{\mathrm{C}}_n[X]$ (similar to (7.22)).

The unbounded case. We now derive lower deviation bounds for a random variable that satisfies either (7.7) or (7.8). With Theorem 7.3 and Proposition 7.6 all that remains to do is to bound $\mathrm{C}[X^2]$:

Proposition 7.12. *Let $v, y > 0$ and f_p be as in* (7.23)*. If X satisfies* (7.7) *with $p, v, \lambda > 0$, then*

$$\mathrm{C}[X] \leq \lambda^{1/p} f_p(\alpha) \quad \textit{and} \quad \mathrm{C}[Z^2] \leq \mathrm{C}[X^2] \leq \lambda^{2/p} \cdot f_{p/2}(\alpha).$$

Furthermore, if X satisfies (7.8) *with $p > 1$, then for $Z := X \cdot \mathbb{I}\{|X| \le (v/y)^{1/p}\}$,*

$$\mathrm{C}[X] \le \frac{(2p-1)(v/\alpha)^{1/p}}{p-1}, \quad \textit{and} \quad \mathrm{C}[Z^2] \le v(v/y)^{2/p-1}/\alpha.$$

Using Proposition 7.12 [resp. 7.6] to bound $\mathrm{C}[Z^2]$ [resp. $|\mathrm{C}[Z] - \mathrm{C}[X]|$], and invoking our Master Theorem 7.3, we immediately get the following bound for the light-tailed case:

Theorem 7.13. *Let $\alpha, \delta \in (0,1)$, $\lambda, p, v > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. random variables such that X satisfies the tail probability in* (7.7)*. Further, let ϵ_n be as in Theorem 7.3, and $f_p(\cdot)$ as in* (7.23)*. Then, for $n \ge 1$ and $Z := X \cdot \mathbb{I}\{X \le (\lambda \ln(vn))^{1/p}\}$, we have, with probability at least $1 - 2\delta$,*

$$\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z] \le \lambda^{\frac{1}{p}} \sqrt{\frac{2 f_{p/2}(\alpha) \ln \delta^{-1}}{\alpha n}} + \lambda^{\frac{1}{p}} \cdot \frac{3 f_p(1/n) - \ln^{\frac{1}{p}}(vn) \ln \delta}{3\alpha n} + |Z_{(\lceil \alpha n \rceil)}| \cdot \epsilon_n. \tag{7.29}$$

A time-uniform version of this inequality is stated in Theorem 7.27. As discussed in the previous subsection, we have $f_p(y) = O(\mathrm{poly}(\ln(1/y)))$ for small y, and thus the deviations we present above satisfy $\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z] \le O(1/\sqrt{\alpha n})$ up to poly-log-factors in $1/\alpha$. Therefore, the bounds in Theorem 7.13 have the optimal dependence on the quantile level α up to poly-log-factors. For a sub-exponential random variable X, the explicit expressions of $f_{p/2}(\alpha)$ and $f_p(1/(\rho n))$, for $\rho \ge 1$, are given by (7.25). In this case, the inequality in (7.29) simplifies (see Corollary 7.20).

Using Proposition 7.12 [resp. 7.6] to bound $\mathrm{C}[Z^2]$ [resp. $|\mathrm{C}[Z] - \mathrm{C}[X]|$], and invoking our Master Theorem 7.3, we immediately get the following bound for the heavy-tailed case (the time-uniform version is stated in Theorem 7.14):

Theorem 7.14. *Let $\alpha, \delta \in (0,1)$, $v > 0$, $p > 2$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. random variables such that X satisfies the tail probability in* (7.8)*. Further, let ϵ_n be as in Theorem 7.3. Then, for $n \ge 1$ and $Z := X \cdot \mathbb{I}\{X \le (-3pvn/\ln \delta)^{1/p}\}$, we have, with probability at least $1 - 2\delta$,*

$$\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z] \le \frac{v^{\frac{1}{p}} p^{1+\frac{1}{p}}}{\alpha} \left(\frac{(1/3)^{1-\frac{1}{p}} + 2^{\frac{1}{2}} 3^{\frac{1}{p}-\frac{1}{2}}}{p-1} \right) \left(\frac{\ln \delta^{-1}}{n} \right)^{1-\frac{1}{p}} + |Z_{(\lceil \alpha n \rceil)}| \cdot \epsilon_n. \tag{7.30}$$

We remark that when the random variable X is supported on $\mathbb{R}_{\ge 0}$, then one can use the unclipped estimator $\widehat{\mathrm{C}}_n[X]$—which does not depend on the parameters of the distribution—instead of $\widehat{\mathrm{C}}_n[Z]$ in all the bounds above. In fact, when X is non-negative, we have $Z \le X$, and so by monotonicity of CVaR [Rockafellar and Uryasev, 2013], we have $\mathrm{C}[Z] \le \mathrm{C}[X]$. In this case, one can also replace $Z_{(\lceil n\alpha \rceil)}$ by $X_{(\lceil \alpha n \rceil)}$ on the RHS of the bounds, since $Z_{(\lceil n\alpha \rceil)} \le X_{(\lceil \alpha n \rceil)}$.

Remark 7.2. *The bounds we have presented in Theorem 7.13, Corollary 7.20, and Theorem 7.14 are of the form $\mathrm{C}[X] \le \widehat{\mathrm{C}}_n[Z] + \delta_n + |Z_{(\lceil \alpha n \rceil)}| \cdot \epsilon_n$, with high probability, for a $\delta_n > 0$*

which converges to zero as $n \to \infty$. For positive random variables, it is possible to rearrange this bound so that only distribution parameters appear on the right (and no empirical value such as $Z_{(\lceil \alpha n \rceil)}$). To achieve this, one can use the fact that $Z_{(\lceil n\alpha \rceil)} \leq \widehat{\mathrm{C}}_n[Z]$ (see e.g. proof of Proposition 4.1 in [Brown, 2007]), then rearrange the above inequality to obtain

$$\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z] \leq \frac{\epsilon_n}{1+\epsilon_n} \cdot \mathrm{C}[X] + \frac{\delta_n}{1+\epsilon_n} \leq \mathrm{C}[X] \cdot \epsilon_n + \delta_n. \tag{7.31}$$

Now $\mathrm{C}[X]$ can be bounded from above using Proposition 7.12 depending on its tail distribution.

In the next section, we apply some of our new concentration inequalities in the multi-armed bandit setting.

7.3 Applications

In this section, we show how the bounds we derived in Sections 7.2.1 and 7.2.2 can be used to achieve state-of-the-art performances in some multi-armed bandit problems.

7.3.1 Best Arm Identification, Fixed Confidence

We now consider the problem of best-arm identification when the performance measure is the CVaR instead of the mean. At each round t, the learner picks an arm $A_t \in [K]$, then observes a loss X_{t,A_t}, where for every $a \in [K]$, $X_a, X_{1,a}, X_{2,a}, \ldots$ are i.i.d. real random variables sampled from some fixed, but unknown distribution $\mathbf{P}_{X_a}$ supported on a bounded interval, which we take to be $[0,1]$ without loss of generality. To simplify notation, we will denote by $\mathbf{P}_a := \mathbf{P}_{X_a}$ for any arm a. Throughout this section, we will use the short-hand notation

$$\mathrm{C}_a := \mathrm{CVaR}_\alpha[\mathbf{P}_a], \tag{7.32}$$

for all $a \in [K]$, and we let $\widehat{\mathrm{C}}_{t,a}$ be any estimator of C_a given t i.i.d. samples. We assume without loss of generality that $\mathrm{C}_1 < \mathrm{C}_2 \cdots \leq \mathrm{C}_K$, which means that arm 1 has the lowest CVaR risk, and we define

$$\Delta_a := \mathrm{C}_a - \mathrm{C}_1. \quad \text{for all } 2 \leq a \leq K.$$

Given a round t, we denote by $N_a(t)$ the number of times arm $a \in [K]$ has been pulled.

Best CVaR arm identification with fixed confidence. We start by consider the setting of best CVaR-arm identification with a fixed confidence δ. In this setting, one would like an algorithm that identifies the best CVaR-arm with as few samples as possible. As far as we know, the best algorithm for this setting when dealing with means (instead of CVaR) is the lil' UCB algorithm due to Jamieson et al. [2014]. Their

algorithm relies on a time-uniform concentration inequality they provide for sub-Gaussian random variables. Extending their algorithm to the CVaR case requires a time-uniform concentration bound for CVaR, which we provided in Sections 7.2.1, 7.2.2, and Appendix 7.4. We now state a slightly more general version of the main result of Jamieson et al. [2014] that accounts for the differences between our concentration bounds and theirs:

Theorem 7.15. *Let $\alpha \in (0,1)$, $\nu, \rho > 1$, and $\tau \in \mathbb{N}$. Let $g : (0,1) \to \mathbb{R}_{>0}$ and for $a \in [K]$ define*

$$\mathcal{E}_a(y) := \left\{\forall t \geq \tau,\ |\mathrm{C}_a - \widehat{\mathrm{C}}_a| \leq \xi(t,y)\right\}, \text{ with } \xi(t,y) := \sqrt{\frac{\rho g(\alpha)}{t} \ln \frac{\ln(\rho t)}{y}},\ y \in (0,1). \tag{7.33}$$

Further, suppose that for some $s > 0$, $\mathbf{P}[\mathcal{E}_a(y)] \geq 1 - sy$, for all $y \in (0,1)$ and $a \in [K]$, and let $\delta \in (0, e^{-2} \wedge e^{-4}/s)$. Then, for

$$\kappa > \left(\ln \frac{2(\nu+1)}{\nu-1} + 2\right) \frac{7(1+\nu)^2}{2(\nu-1)^2}, \quad \text{and} \quad G_\alpha := \rho\nu^2(1+\nu^2)g(\alpha),$$

Algorithm 1 returns the best CVaR arm with probability $\geq 1 - (5s+1)\delta - (s\delta)^{1/8}$ after T rounds, where

$$T \leq 2 + (1+\kappa+\tau)K + (1+\kappa)G_\alpha \sum_{a=2}^{K} \frac{8\ln\left(\frac{s}{\delta}\right) + \ln\left(\frac{2}{\delta}\ln\left(\frac{e}{\delta} \vee \frac{\rho G_\alpha}{\delta \Delta_a^2}\right)\right)}{\Delta_a^2}. \tag{7.34}$$

For completeness, we include the proof of Theorem 7.15 in Appendix 7.5. We note that Theorem 7.15 requires an estimator of CVaR with a time-uniform concentration bound of the form (7.33) for some $\tau \geq 1$. The bound in (7.34) has a linear dependence in the parameter τ, and so τ must not be too large. For all the time-uniform bounds we present in this chapter, there exits a $\tau \leq O(1)$ such that (7.33) holds for a specific function g that depends on the parameters of the distribution of X. In fact, our bounds are already of the form (7.33) except for lower-order terms that become smaller than the main term $\xi(t,\delta)$ for some moderate $\tau \leq O(1)$. We now state a concentration bound of the form (7.33) for the sub-exponential case after getting rid of any lower-order terms. The steps in proof of the next theorem can easily be extended to the case where X satisfies the more general light tail condition in (7.5) (or the heavy tail condition in (7.6)).

Theorem 7.16. *Let $\alpha, \delta \in (0,1)$, $\lambda > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}_{\geq 0}$ be i.i.d. λ-sub-exponential random variables, i.e. X satisfies (7.7) with $(v,p) = (1,1)$. Further, let c be as in Theorem 7.3. Then, there exists a $\tau \leq O((\ln\alpha \ln\delta)/\alpha)$, such that for all $\rho > 1$ and $Z^{(n)} := X \cdot \mathbb{I}\{X \leq$*

Algorithm 1 lil UCB for CVaR MAB.

Require: Quantile level $\alpha \in (0,1)$
Confidence level $\delta \in (0,1)$.
Parameters $\nu, \kappa > 1$.
Initial exploration time τ.

for $a = 1, \ldots, K$ **do**
  Play arm a for τ rounds.
Set $t = \tau K + 1$, and $N_a(t) = \tau$, for all $a \in [K]$.
Define $\xi(s,\delta) = \sqrt{\frac{\rho g(\alpha)}{s} \ln \frac{\ln(\rho s)}{\delta}}$, for all $s \geq 1$.
while $N_a(t) < 1 + \kappa \sum_{a' \neq a} N_{a'}(t)$, for all $a \in [K]$ **do**
  Define $\widehat{\mathrm{C}}_{t,a} = \inf_{\mu > 0} \left\{ \mu + \sum_{s=1}^{t-1} \frac{\mathbb{I}\{A_s = a\} \cdot [X_{s,a} - \mu]_+}{\alpha t} \right\}$.
  Sample $A_t \in \arg\min_{a \in [K]} \left\{ \widehat{\mathrm{C}}_{N_a(t),a} - \nu \xi_a(t,\delta) \right\}$.
  Set $N_{A_t}(t+1) = N_{A_t}(t) + 1$ and $N_a(t+1) = N_a(t)$, for $a \neq A_t$.
  Set $t = t + 1$.

$\lambda \ln(\rho n)\}$, *we have*

$$\mathbf{P}\left[\forall n \geq \tau, \quad |\widehat{\mathrm{C}}_n[Z^{(n)}] - \mathrm{C}[X]| \leq 4\lambda(1 - \ln \alpha)\sqrt{2\frac{\rho \ln \frac{\ln(\rho n)}{\delta}}{\alpha n}}\right] \geq 1 - 2\delta\sqrt{c}/\ln \rho.$$

The theorem shows that the condition in (7.33) in Theorem 7.15 holds with $s = 2\sqrt{c}/\ln \rho$ and $g(\alpha) = 32(1 - \ln \alpha)^2/\alpha$. We note that the exact condition that τ needs to satisfy in Theorem 7.16 is in (7.98).

The combination of Theorems 7.15 and 7.16 implies that Algorithm 1 is the first algorithm for the fixed confidence CVaR-MAB setting whose sample complexity has the optimal dependence in the CVaR-risk gaps (Δ_a) and the quantile level α (up to log-factors in $1/\alpha$) for sub-exponential random variables (this is also true for the general light tail case in (7.5)).

7.3.2 Best Arm Identification, Fixed Budget

The point of this section is to show that our concentration inequalities for CVaR modularly slot into generic confidence-interval based algorithms. We follow the standard Sequential Halving algorithm template introduced by Karnin et al. [2013]. That is, we split the time horizon in $\ln_2 K$ equal-length phases, in which we sample the remaining arms uniformly, after which we eliminate the top half (by empirical CVaR) of the arms (note that for CVaR lower is better). The correct arm is output iff it is not eliminated in any phase. The best arm is eliminated in a given phase if it ends up in the top half, meaning that at least *half* of the remaining arms have to look better than it; a rare event.

We start our analysis from the following summary of Theorems 7.7 and 7.13:

Theorem 7.17. *Let $\alpha, \delta \in (0,1)$, $\lambda > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}_{\geq 0}$ be i.i.d. λ-sub-exponential random variables, i.e. X satisfies* (7.7) *with $(v,p) = (1,1)$. Then, for $Z := X \cdot \mathbb{I}\{X \leq \lambda \ln n\}$, and all $n \geq n_0$ where*

$$n_0 = \inf\left\{ n : \frac{2 + 3\ln(n/\alpha) - \ln n \ln \delta}{3\alpha n} \leq (1 - \ln \alpha)\sqrt{\frac{\ln \delta^{-1}}{\alpha n}} \right\},$$

we have with probability at least $1 - 3\delta$,

$$|\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z]| \leq 5\lambda \cdot (1 - \ln \alpha)\sqrt{\frac{\ln \delta^{-1}}{\alpha n}}.$$

The theorem is proved at the very end of the appendix. From here, we invert the probability bound to find that

$$\mathbf{P}\left\{ |\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z]| > \epsilon \right\} \;\leq\; \exp\left(-\frac{\alpha n \epsilon^2}{(5\lambda(1 + \ln(1/\alpha)))^2}(1 - o_n(1)) \right).$$

The classical Sequential Halving proof [Lattimore and Szepesvári, 2019, Theorem 33.10] then implies mutatis mutandis:

Corollary 7.18. *Consider a K-armed bandit with CVaR_α given by $\mathrm{C}_1 \leq \ldots \mathrm{C}_K$ (sorted here for convenience without loss of generality). The probability that Sequential Halving outputs the wrong arm is at most*

$$3\ln_2(K) \exp\left(\frac{-n\alpha(1 - o_n(1))}{(5\lambda(1 + \ln(1/\alpha)))^2 4\ln_2(K) \max_{i:\Delta_i > 0} \frac{i}{\Delta_i^2}} \right),$$

where $\Delta_i = \mathrm{C}_i - \mathrm{C}_1$ is the suboptimality gap of arm i.

Taking stock, we find that the overall dependence in α is, up to log factors, as if time were shrunk by a factor α, which indeed matches the intuition that, for every arm, the $1 - \alpha$ portion of samples below the α-quantile are not informative.

The new section provides some additional concentration inequalities omitted from Section 7.2.

7.4 Additional Inequalities

In this section, we include some additional concentration inequality for CVaR under different settings. We start by the time-uniform version of Corollary 7.11:

Corollary 7.19. *Let $\alpha, \delta \in (0,1)$, and $X_1, \ldots, X_n$ be i.i.d. random variables in $[0,1]$. Further, let $\widehat{\mathrm{C}}_n[\cdot;\cdot]$ be as in* (7.15)*, and ϵ_n' and $\phi_\rho(\cdot)$ be as in Theorem 7.3. Then, for $\rho > 1$,*

$$\mathbf{P}\left[\forall n \geq 1,\;\; \mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[X;\epsilon_n] + \sqrt{\frac{2\rho\widehat{\mathrm{C}}_n[X;\epsilon_n] \cdot \ln \frac{\phi_\rho(n)}{\delta}}{\alpha n}} + \frac{10 \ln \frac{\phi_\rho(n)}{\delta}}{3\alpha n} \right] \geq 1 - 2\delta. \tag{7.35}$$

For a sub-exponential random variable X—one that satisfies (7.7) with $(v,p) = (1,1)$—the explicit expressions of $f_{p/2}(\alpha)$ and $f_p(1/(\rho n))$, for $\rho \geq 1$, are given by (7.25). In this case, the inequality in (7.29) simplifies (see Corollary 7.20). Plugging these into (7.29) implies the following corollary:

Corollary 7.20. *Let $\alpha, \delta \in (0,1)$, $\lambda > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. λ-sub-exponential random variables, i.e. X satisfies (7.7) with $(v,p) = (1,1)$. Then, for ϵ_n as in Theorem 7.3 and $Z := X \cdot \mathbb{I}\{X \leq \lambda \ln n\}$, we have with probability at least $1 - 2\delta$,*

$$\mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[Z] + 2\lambda \cdot (1 - \ln \alpha)\sqrt{\frac{\ln \delta^{-1}}{\alpha n}} + \lambda \cdot \frac{3 + 6\ln n - \ln n \ln \delta}{3\alpha n} + |Z_{(\lceil n\alpha \rceil)}| \cdot \epsilon_n. \tag{7.36}$$

We now deal with the time-uniform upper and lower deviation bounds when X is sub-exponential, which is a special case of the finite exponential moment.

Corollary 7.21. *Let $\alpha, \delta \in (0,1)$, $\lambda > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. λ-sub-exponential random variables, i.e. X satisfies (7.7) with $(v,p) = (1,1)$. Further, let ϵ'_n and ϕ_ρ be as in Theorem 7.3. Then, for $\rho > 1$ and $Z^{(n)} := X \cdot \mathbb{I}\{X \leq \lambda \ln(\rho n)\}$, we have with probability at least $1 - 2\delta$, for all $n \geq 1$*

$$\begin{aligned}\mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[Z] + 2\lambda \cdot (1 - \ln \alpha)\sqrt{\frac{\rho \ln \frac{\phi_\rho(n)}{\delta}}{\alpha n}} + \lambda \cdot \frac{3 + 6\ln n + 2\ln n \ln \frac{\phi_\rho(n)}{\delta}}{3\alpha n} \\ + |Z^{(n)}_{(\lceil n\alpha \rceil)}| \cdot \epsilon'_n.\end{aligned} \tag{7.37}$$

Furthermore, we have, with probability at least $1 - \delta$, for all $n \geq 1$

$$\widehat{\mathrm{C}}_n[Z] \leq \mathrm{C}[X] + 2\lambda \cdot (1 - \ln \alpha)\sqrt{\frac{2\rho \ln \frac{\phi_\rho(n)}{\delta}}{\alpha n}} + \lambda \cdot \frac{3 + 6\ln n + 2\ln n \ln \frac{\phi_\rho(n)}{\delta}}{3\alpha n}.$$

7.5 Proofs

Proof of Lemma 7.2. Let $\varphi(\cdot) := \iota_{[0,1/\alpha]}(\cdot)$, where for a set $\mathcal{C} \subseteq \mathbb{R}$, $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$; and $+\infty$ otherwise. Further, let $\pi := (1, \ldots, 1)/n \in \Delta_n$. From (7.14), we have

$$\widehat{\mathrm{C}}_n[X;\epsilon] = \sup\left\{\mathbb{E}_{\pi(i)}[X_i Q_i - \varphi(Q_i)] : \left|\mathbb{E}_{\pi(i)}[Q_i] - 1\right| \leq \epsilon\right\}.$$

By introducing Lagrangian multipliers $\eta, y \geq 0$, we can write

$$\begin{aligned}
\widehat{\mathrm{C}}_n[X;\epsilon] &= \inf_{\eta,y\geq 0} \sup_{Q_{1:n}\in[0,1/\alpha]^n} \left\{ \begin{array}{c} \mathbb{E}_{\pi(i)}[X_i - \varphi(Q_i)] + \eta \cdot (1+\epsilon - \mathbb{E}_{\pi(i)}[Q_i]) \\ +y \cdot (\mathbb{E}_{\pi(i)}[Q_i] - 1 + \epsilon) \end{array} \right\}, \\
&= \inf_{\eta,y\geq 0} \left\{ \eta - y + (\eta + y)\epsilon + \sup_{Q_{1:n}\in[0,1/\alpha]^n} \left\{ \mathbb{E}_{\pi(i)}[(X_i - \eta + y)Q_i - \varphi(Q_i)] \right\} \right\}, \\
&= \inf_{\eta,y\geq 0} \left\{ \eta - y + (\eta + y)\epsilon + \mathbb{E}_{\pi(i)} \left[\sup_{0\leq x\leq 1/\alpha} \{(X_i - \eta + y)x - \varphi(x)\} \right] \right\}, \\
&= \inf_{\eta,y\geq 0} \left\{ \eta - y + (\eta + y)\epsilon + \mathbb{E}_{\pi(i)}[\varphi^\star(X_i - \eta + y)] \right\}, \qquad (7.38) \\
&= \inf_{\mu\in\mathbb{R}} \left\{ \mu + |\mu|\epsilon + \mathbb{E}_{\pi(i)}[\varphi^\star(X_i - \mu)] \right\}, \qquad (7.39)
\end{aligned}$$

where (7.38) is due to $\{x \in \mathbb{R} \mid \varphi(x) < +\infty\} = [0, 1/\alpha]$, and (7.39) follows by setting $\mu := \eta - y$ and noting that the inf in (7.38) is always attained at a point $(\eta, y) \in \mathbb{R}^2_{\geq 0}$ satisfying $\eta \cdot y = 0$, in which case $\eta + y = |\mu|$; this is true because by the positivity of ϵ, if $\eta, y > 0$, then $(\eta + y)\epsilon$ can always be made smaller while keeping the difference $\eta - y$ fixed. The proof is concluded by noting that the Fenchel dual of φ satisfies $\varphi^\star(z) = 0 \vee (z/\alpha)$, for all $z \in \mathbb{R}$. This establishes our first desired inequality:

$$\widehat{\mathrm{C}}_n[X;\epsilon] = \inf_{\mu\in\mathbb{R}} \left\{ \mu + |\mu|\epsilon + \frac{\sum_{i=1}^n [X_i - \mu]_+}{\alpha n} \right\}. \qquad (7.40)$$

Now, by letting $\widehat{\mu} \in \inf_{\mu>0}\{\mu + \sum_{i=1}^n [X_i - \mu]_+/(\alpha n)\}$, we have

$$\begin{aligned}
\widehat{\mathrm{C}}_n[X;\epsilon] &\leq \widehat{\mu} + |\widehat{\mu}|\epsilon + \frac{\sum_{i=1}^n [X_i - \widehat{\mu}]_+}{n\alpha}, \\
&= \widehat{\mathrm{C}}_n + |\widehat{\mu}|\epsilon, \qquad (7.41)
\end{aligned}$$

where the last inequality follows by definition of $\widehat{\mathrm{C}}_n$. The desired inequality follows by the fact that $\widehat{\mu} = X_{(\lceil n\alpha \rceil)}$—see e.g. proof of Proposition 4.1 in [Brown, 2007]. □

Proof of Proposition 7.12. Let $\mu_{\alpha,i} := \inf\{\mu \in \mathbb{R} \mid \mathbf{P}[X^i \geq \mu^i] \leq \alpha\}$, for $i \in [2]$. By definition of CVaR, we have, for $B > 0$

$$\begin{aligned}
\mathrm{C}[X] &= \mathbb{E}[X \mid X \geq \mu_1], \\
&= \frac{1}{\alpha}\mathbb{E}[X \cdot \mathbb{I}\{X \geq \mu_1\}], \\
&\leq \frac{1}{\alpha}\mathbb{E}[X \cdot \mathbb{I}\{\mu_1 \leq X \leq B\}] + \frac{1}{\alpha}\mathbb{E}[X \cdot \mathbb{I}\{X \geq B\}] \\
&\leq B + \frac{1}{\alpha}\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq B\}]. \qquad (7.42)
\end{aligned}$$

Setting B to $(\lambda \ln(v/\alpha))^{1/p}$ [resp. $(v/\alpha)^{1/p}$] when X satisfies (7.7) [resp. (7.8)] and invoking Lemma A.5 to bound $\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq B\}]/\alpha$ implies the desired bounds

on $\mathrm{C}[X]$. We now bound $\mathrm{C}[Z^2]$. We start with the light-tailed case; we have by monotonicity of CVaR [Rockafellar and Uryasev, 2013]

$$\begin{aligned}\mathrm{C}[Z^2] \leq \mathrm{C}[X^2] &= \mathbb{E}[X^2 \mid X^2 \geq \mu_{\alpha,2}^2], \\ &= \frac{1}{\alpha}\mathbb{E}[X^2 \cdot \mathbb{I}\{X^2 \geq \mu_{\alpha,2}^2\}]. \end{aligned} \tag{7.43}$$

Now the term on RHS of (7.43) can be bounded from above using lemma A.5 when X satisfies (7.7) leading to the desired result for the light-tailed case. For the heavy-tailed setting, i.e. for X satisfying (7.8) and $Z := X \cdot \mathbb{I}\{X \leq (v/\alpha)^{1/p}\}$, we have

$$\begin{aligned}\mathrm{C}[Z^2] &\leq \frac{1}{\alpha}\mathbb{E}[Z^2], && \text{(by Lemma A.3)} \\ &\leq \frac{1}{\alpha}(v/\alpha)^{2/p-1}\mathbb{E}[|Z|^p], && \text{(by Holder's inequality)} && (7.44) \\ &\leq \frac{1}{\alpha}(v/\alpha)^{2/p-1}\mathbb{E}[|X|^p], && && (7.45) \\ &\leq \frac{v}{\alpha}(v/\alpha)^{2/p-1}. \end{aligned}$$

This completes the proof. □

Proof of Proposition 7.6. Let $B > 0$. We have by the sub-additivity and monotonicity of CVaR:

$$\begin{aligned}\mathrm{C}[Z] &= \mathrm{C}[X - X \cdot \mathbb{I}\{X \geq B\}], \\ &\leq \mathrm{C}[X] + \mathrm{C}[-X \cdot \mathbb{I}\{X \geq B\}], && \text{(CVaR is sub-additive)} \\ &\leq \mathrm{C}[X] + \mathrm{C}[|X| \cdot \mathbb{I}\{X \geq B\}], && \text{(CVaR is monotonic)} \\ &\leq \mathrm{C}[X] + \mathrm{C}[|X| \cdot \mathbb{I}\{|X| \geq B\}], \\ &\leq \mathrm{C}[X] + \frac{1}{\alpha}\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq B\}]. && && (7.46) \end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathrm{C}[X] &= \mathrm{C}[X - X \cdot \mathbb{I}\{X \geq B\}], \\ &\geq \mathrm{C}[X] - \mathrm{C}[X \cdot \mathbb{I}\{X \geq B\}], && \text{(CVaR is sub-additive)} \\ &\geq \mathrm{C}[X] - \mathrm{C}[|X| \cdot \mathbb{I}\{X \geq B\}], && \text{(CVaR is monotonic)} \\ &\geq \mathrm{C}[X] - \mathrm{C}[|X| \cdot \mathbb{I}\{|X| \geq B\}], \\ &\geq \mathrm{C}[X] - \frac{1}{\alpha}\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq B\}]. && && (7.47) \end{aligned}$$

Thus, combining (7.46) and (7.47) imply

$$|\mathrm{C}[Z] - \mathrm{C}[X]| \leq \frac{1}{\alpha}\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq B\}].$$

Setting B to either $(\lambda \ln(v/y))^{1/p}$ or $(v/y)^{1/p}$, for $y > 0$, in this inequality and using

Lemma A.5 to bound the RHS, we obtain the desired bounds on $|\mathrm{C}[Z] - \mathrm{C}[X]|$.

We now prove the bounds on $\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2]$. Let $\mu_{\alpha,2} := \inf\{\mu \in \mathbb{R} \mid \mathbf{P}[|Z| \geq \mu] \leq \alpha\}$. By Lemma A.6, we have

$$\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2] \leq \frac{1}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{Z \geq \mu_\alpha\}] + \frac{3}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{|Z| \geq \mu_{\alpha,2}\}]. \tag{7.48}$$

Now since Z is a clipped version of X, both Z and $|Z|$ satisfy (7.7) whenever X satisfies (7.7). By this fact, Lemma A.5 and (7.48), we obtain the desired bound on $\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2]$ for the light-tailed case. We now move to the heavy-tailed case. By (7.48), we have

$$\begin{aligned}
\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z-\mu_\alpha]_+])^2] &\leq \frac{4}{\alpha^2}\mathbb{E}[Z^2], \\
&\leq \frac{4}{\alpha^2}(v/\alpha)^{2/p-1}\mathbb{E}[|Z|^p], \quad \text{(by Holder's inequality)} \qquad (7.49)\\
&\leq \frac{4}{\alpha^2}(v/\alpha)^{2/p-1}\mathbb{E}[|X|^p], \qquad (7.50)\\
&\leq v\frac{4}{\alpha^2}(v/\alpha)^{2/p-1}.
\end{aligned}$$

This completes the proof. □

7.5.1 Proof of Theorem 7.3

We start by presenting a time-uniform version of Bernstein inequality (see e.g. Maurer and Pontil [2009]):

Proposition 7.22. *Let $(B_n) \subset \mathbb{R}_{>0}$ be an increasing sequence, and $X, X_1, X_2, \cdots \in \mathbb{R}$ be i.i.d. random variables such that $\mathbb{E}[X^2] \leq V$. Then, for $Z := X \cdot \mathbb{I}\{|X| \leq B_n\}$, we have, for any $\delta \in (0,1)$ and $\rho > 0$, with probability at least $1-\delta$,*

$$\mathbf{P}\left[\exists n \geq 1, \left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z]\right| \geq \sqrt{\frac{2\rho V \ln(\phi_\rho(n)/\delta)}{n}} + \frac{2b\ln(\phi_\rho(n)/\delta)}{3n}\right] \leq \delta, \tag{7.51}$$

where $\phi_\rho(n) := \ln\left(c\ln_\rho(\rho n)\ln^2(\ln_\rho(\rho n))\right) = O(\ln\ln_\rho n)$ and $c := \sum_{k\geq 2}\frac{2}{k\ln^2(k)} \approx 4.214$.

Proof of Proposition 7.22. Let $\lambda > 0$ and $X^\lambda, X_1^\lambda, X_2^\lambda, \ldots$ be i.i.d. random variables such that $|X_i^\lambda| \leq B(\lambda)$, where $B : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is any increasing function. By [Howard et al., 2020, Corollary 1(c)], there exists a positive function $\mathfrak{s} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that

$$\mathbf{P}\left[\exists n \geq 1 : \left|\sum_{i=1}^n X_i^\lambda - n\mathbb{E}[X^\lambda]\right| \geq x + \mathfrak{s}\left(\frac{x}{\lambda}\right)\cdot(nV-\lambda)\right] \leq 2\exp\left(-\frac{x^2}{2\lambda + 2B(\lambda)x/3}\right),$$

which implies that for any $\delta' \in (0,1)$,

$$\mathbb{E}\left[\mathbb{I}\left\{\exists n \geq 1 : \begin{array}{l} \left|\sum_{i=1}^n X_i^\lambda - n\mathbb{E}[X^\lambda]\right| \geq \sqrt{2n\lambda \ln(2/\delta')} \\ \qquad + \frac{B(\lambda)\ln(2/\delta')}{3n} + \mathfrak{s}\left(\frac{x}{\lambda}\right) \cdot (nV - \lambda) \end{array}\right\}\right] \leq \delta'. \tag{7.52}$$

For $k \geq 2$ and $\rho \in (0,1]$, define $\lambda_k := V\rho^k$ and

$$\pi_k := \frac{1}{c' k \ln^2(k)}, \quad \text{where } c' := \sum_{k \geq 2} \frac{1}{k \ln^2(k)} \approx 2.107. \tag{7.53}$$

By setting $\lambda = \lambda_k$ and $\delta' = \delta_k :- \pi_k \delta$ in (7.52), we get that

$$\mathbb{E}\left[\mathbb{I}\left\{\exists n \geq 1 : \begin{array}{l} \left|\sum_{i=1}^n X_i^{\lambda_k} - n\mathbb{E}[X^{\lambda_k}]\right| \geq \sqrt{2n\lambda_k \ln(2/\delta_k)} \\ \qquad + \frac{B(\lambda_k)\ln(2/\delta_k)}{3n} + \mathfrak{s}\left(\frac{x}{\lambda_k}\right) \cdot (nV - \lambda_k) \end{array}\right\}\right] \leq \delta_k. \tag{7.54}$$

By summing over $k \in \mathbb{N}$, we get

$$\begin{aligned} \delta &\geq \mathbb{E}\left[\sum_{k \geq 1} \mathbb{I}\left\{\exists n \geq 1 : \begin{array}{l} \left|\sum_{i=1}^n X_i^{\lambda_k} - n\mathbb{E}[X^{\lambda_k}]\right| \geq \sqrt{2n\lambda_k \ln(2/\delta_k)} \\ \qquad + \frac{B(\lambda_k)\ln(2/\delta_k)}{3n} + \mathfrak{s}\left(\frac{x}{\lambda_k}\right) \cdot (nV - \lambda_k) \end{array}\right\}\right], \\ &\geq \mathbb{E}\left[\mathbb{I}\left\{\exists n, k \geq 1 : \begin{array}{l} \left|\sum_{i=1}^n X_i^{\lambda_k} - n\mathbb{E}[X^{\lambda_k}]\right| \geq \sqrt{2n\lambda_k \ln(2/\delta_k)} \\ \qquad + \frac{B(\lambda_k)\ln(2/\delta_k)}{3n} + \mathfrak{s}\left(\frac{x}{\lambda_k}\right) \cdot (nV - \lambda_k) \end{array}\right\}\right], \\ &= 1 - \mathbf{P}\left[\forall n, k \geq 1 : \begin{array}{l} \left|\sum_{i=1}^n X_i^{\lambda_k} - n\mathbb{E}[X^{\lambda_k}]\right| \leq \sqrt{2n\lambda_k \ln(2/\delta_k)} \\ \qquad + \frac{B(\lambda_k)\ln(2/\delta_k)}{3n} + \mathfrak{s}\left(\frac{x}{\lambda_k}\right) \cdot (nV - \lambda_k) \end{array}\right]. \end{aligned} \tag{7.55}$$

Now, let k_n be such that

$$\lambda_{k_n - 1} \leq nV \leq \lambda_{k_n}. \tag{7.56}$$

By (7.55) and the fact that $\lambda_{k_n} \geq nV$, we have that

$$\mathbf{P}\left[\forall n : \left|\sum_{i=1}^n X_i^{\lambda_{k_n}} - n\mathbb{E}[X^{\lambda_{k_n}}]\right| \leq \sqrt{2n\lambda_{k_n} \ln(2/\delta_{k_n})} + \frac{B(\lambda_{k_n})\ln(2/\delta_k)}{3n}\right] \geq 1 - \delta. \tag{7.57}$$

On the other hand, by (7.56), we have that

$$k_n \leq \ln_\rho(n) + 1 \quad \text{and} \quad \lambda_{k_n} \leq n\rho V.$$

Plugging this into (7.57), yields

$$\mathbf{P}\left[\forall n : \left|\sum_{i=1}^{n} X_i^{\lambda_{k_n}} - n\mathbb{E}[X^{\lambda_{k_n}}]\right| \leq \sqrt{2\rho n V \ln(\phi_\rho(n)/\delta)} + \frac{B(\lambda_{k_n})\ln(\phi_\rho(n)/\delta)}{3n}\right] \geq 1-\delta, \tag{7.58}$$

where $\phi_\rho(n) := 2c' \cdot (\ln_\rho(n)+1)\ln^2(\ln_\rho(n)+1)$. We obtain the desired result by applying this inequality with $Z^{(n)} = X^{\lambda_{k_n}} = X \cdot \mathbb{I}\{|X| \leq B_n\}$ and a function B satisfying $B^{-1}(B_n) = \lambda_{k_n}$ (there always exists an increasing function B satisfying this condition since (B_n) and (λ_{k_n}) are both increasing sequences). □

Using Proposition 7.22, we prove some useful facts on components of the random variable Y defined in (7.13):

Lemma 7.23. *Let $\alpha, \delta \in (0,1)$, $\rho > 1$, and $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathcal{X}$. Further, define*

$$\widehat{\mathcal{Q}}_n(\epsilon) := \left\{ Q_{1:n} \in [0, 1/\alpha]^n : \left|\frac{1}{n}\sum_{i=1}^{n} Q_i - 1\right| \leq \epsilon \right\}. \tag{7.59}$$

Then, for $Q_\star$ as in (7.13)*, and ϵ_n and ϵ'_n as in Theorem 7.3, we have*

$$\begin{aligned} &\mathbf{P}\left[(\mathbb{E}[Q_\star \mid X_1], \ldots, \mathbb{E}[Q_\star \mid X_n])^\top \in \widehat{\mathcal{Q}}_n(\epsilon_n)\right] \geq 1-\delta, \\ \textit{and}\quad &\mathbf{P}\left[\forall n \geq 1 : \ (\mathbb{E}[Q_\star \mid X_1], \ldots, \mathbb{E}[Q_\star \mid X_n])^\top \in \widehat{\mathcal{Q}}_n(\epsilon'_n)\right] \geq 1-\delta. \end{aligned} \tag{7.60}$$

Proof of Lemma 7.23. The first claim follows by the fact that $X_i, i \in [n]$, are i.i.d., and an application of the law of total expectation. Now for the second claim, let $\Delta := |\mathbb{E}_{\widehat{\mathbf{P}}_n}[Q_\star \mid X] - 1|$. Since $Q_\star$ is a density, the law of total expectation implies that

$$\Delta = |\mathbb{E}_{\widehat{\mathbf{P}}_n}[Q_\star \mid X] - \mathbb{E}[\mathbb{E}[Q_\star \mid X]]|, \tag{7.61}$$

and so by Bennett's inequality (see *e.g.* Theorem 3 in Maurer and Pontil [2009]) applied to the random variable $\mathbb{E}[q_\star \mid X]$, we get that, with probability at least $1-\delta$,

$$\begin{aligned} \Delta &\leq \sqrt{\frac{2\rho \textsc{Var}[\mathbb{E}[Q_\star \mid X]]\ln(1/\delta)}{n}} + \frac{\|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(1/\delta)}{3n}, \\ &\leq \sqrt{\frac{2\rho \mathbb{E}[\mathbb{E}[Q_\star \mid X]^2]\ln(1/\delta)}{n}} + \frac{\|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(\phi_\rho(n)/\delta)}{3n}, \\ &\leq \sqrt{\frac{2\rho \|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(1/\delta)}{n}} + \frac{\|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(1/\delta)}{3n}, && (7.62) \\ &\leq \sqrt{\frac{2\ln(1/\delta)}{\alpha n}} + \frac{\ln(1/\delta)}{3\alpha n}, && (7.63) \end{aligned}$$

where (7.62) follows by the fact that $\mathbb{E}[\mathbb{E}[Q_\star \mid X]^2] \leq \mathbb{E}[\mathbb{E}[Q_\star \mid X]] \cdot \|\mathbb{E}[Q_\star \mid X]\|_\infty = \|\mathbb{E}[Q_\star \mid X]\|_\infty$, which holds since $\mathbb{E}[Q_\star \mid X] \geq 0$ and $\mathbb{E}[\mathbb{E}[Q_\star \mid X]] = \mathbb{E}[Q_\star] = 1$. Fi-

nally, (7.63) follows by the facts that $\|\mathbb{E}[Q_\star \mid X]\|_\infty \leq \|Q_\star\|_\infty$ (by Jensen's inequality), and $\|Q_\infty\|_\infty \leq 1/\alpha$ by definition. This shows the first inequality we are after.

On the other hand, by Proposition 7.22 applied to the random variable $\mathbb{E}[Q_\star \mid X]$, we get that, with probability at least $1-\delta$,

$$\begin{aligned}
\forall n \geq 1, \quad \Delta &\leq \sqrt{\frac{2\rho \textsc{Var}[\mathbb{E}[Q_\star \mid X]] \ln(\phi_\rho(n)/\delta)}{n}} + \frac{2\|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(\phi_\rho(n)/\delta)}{3n}, \\
&\leq \sqrt{\frac{2\rho \mathbb{E}[\mathbb{E}[Q_\star \mid X]^2] \ln(\phi_\rho(n)/\delta)}{n}} + \frac{2\|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(\phi_\rho(n)/\delta)}{3n}, \\
&\leq \sqrt{\frac{2\rho \|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(\phi_\rho(n)/\delta)}{n}} + \frac{2\|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(\phi_\rho(n)/\delta)}{3n}, \quad (7.64) \\
&\leq \sqrt{\frac{2\rho \ln(\phi_\rho(n)/\delta)}{\alpha n}} + \frac{2\|\mathbb{E}[Q_\star \mid X]\|_\infty \ln(\phi_\rho(n)/\delta)}{3\alpha n}, \quad (7.65)
\end{aligned}$$

where (7.64) and (7.65) follow by the facts that $\mathbb{E}[\mathbb{E}[Q_\star \mid X]^2] \leq \|Q_\star\|_\infty \leq 1/\alpha$, as justified before. □

By Lemma 7.1 and dual formulation of CVaR in (7.4b) we have the following immediate Corollary:

Corollary 7.24. *Let $\alpha, \delta \in (0,1)$, $\rho > 1$, and $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathcal{X}$. Further, for Y as in* (7.13)*, and ϵ_n and ϵ'_n as in Theorem 7.3, we have*

$$\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^n Y_i \leq \widehat{\mathrm{C}}_n[X;\epsilon_n]\right] \geq 1-\delta \quad \textit{and} \quad \mathbf{P}\left[\forall i \geq 1,\ \frac{1}{n}\sum_{i=1}^n Y_i \leq \widehat{\mathrm{C}}_n[X;\epsilon'_n]\right] \geq 1-\delta. \tag{7.66}$$

Proof of Theorem 7.3. Let $Z := X \cdot \mathbb{I}\{|X| \leq B\}$ and

$$Q_\star \in \arg\max_{\mathcal{Q}} \mathbb{E}[XQ], \quad \text{where} \quad \mathcal{Q} := \left\{Q = \frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}} \leq \frac{1}{\alpha},\ \mathbf{Q} \in \mathcal{M}_{\mathbf{P}}(\Omega),\ \mathbb{E}[Q] = 1\right\}. \tag{7.67}$$

Let $U = Z \cdot \mathbb{E}[Q_\star \mid X], U_1 = Z_1 \cdot \mathbb{E}[Q_\star \mid X_1], U_2 = Z_2 \cdot \mathbb{E}[Q_\star \mid X_2], \ldots$ are i.i.d., and so by Bennett's inequality (see e.g. [Maurer and Pontil, 2009, Theorem 3]), we get that,

$$\mathbf{P}\left[\mathbb{E}[U] - \frac{1}{n}\sum_{i=1}^n U_i \leq \sqrt{\frac{2\mathbb{E}[U^2] \cdot \ln(1/\delta)}{n}} + \frac{B\ln(1/\delta)}{3\alpha n}\right] \leq 1-\delta. \tag{7.68}$$

There are three facts we need to arrive at the desired bound:

1. By the law of total expectation, we have $\mathbb{E}[U] = \mathrm{C}[Z]$.

2. We have:

$$\begin{aligned}
\mathbb{E}[U^2] &= \mathbb{E}[Z^2 \cdot \mathbb{E}[Q_\star \mid X]^2], \\
&\leq \mathbb{E}[\mathbb{E}[Z^2 \cdot Q_\star^2 \mid X]], && \text{(Jensen's inequality)} \\
&= \mathbb{E}[Z^2 \cdot Q_\star]/\alpha, && \text{(LTE and } Q_\star \leq 1/\alpha) \\
&\leq \sup_{Q \in \mathcal{Q}} \mathbb{E}[Z^2 \cdot Q]/\alpha, \\
&= \mathrm{C}[Z^2]/\alpha, && \text{(by definition of CVaR)} \\
&\leq \mathrm{C}[X^2]/\alpha,
\end{aligned}$$

where the last inequality follows by the fact that $0 \leq Z^2 \leq X^2$ and the monotonicity of CVaR.

3. Since $0 \leq Z \leq X$, we have

$$U_i \leq Y_i := X_i \cdot \mathbb{E}[Q_\star \mid X_i], \quad \forall i \in [n]. \tag{7.69}$$

4. By Corollary 7.24, we have, with probability at least $1 - \delta$,

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \leq \widehat{\mathrm{C}}_n[X; \epsilon_n], \tag{7.70}$$

where $\widehat{\mathrm{C}}_n[\cdot;\cdot]$ is as in (7.15).

Combining these four facts with (7.68), and a applying a union bound implies (7.18). We now show (8.16). For this, we use the time-uniform Bernstein inequality in Proposition 7.22 (instead of (7.68)), together with Corollary 7.24, the points 1-3 above, and a union bound. □

7.5.2 Proof of Theorem 7.7

We state and prove the following extension of Theorem 7.7:

Theorem 7.25. *Let $\alpha, \delta \in (0,1)$, $\lambda, p, v > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. random variables such that X satisfies the tail probability in* (7.7)*. Further, let $\phi_\rho(\cdot)$ be as in Theorem 7.3, and $f_p(\cdot)$ be as in* (7.23)*. Then, for $n \geq 1$ and $Z := X \cdot \mathbb{I}\{X \leq (\lambda \ln(vn))^{1/p}\}$, we have, with probability at least $1 - \delta$,*

$$\widehat{\mathrm{C}}_n[Z] - \mathrm{C}[X] \leq 2\lambda^{\frac{1}{p}} \sqrt{\frac{2 f_{p/2}(\alpha) \ln \delta^{-1}}{\alpha n}} + \lambda^{\frac{1}{p}} \cdot \frac{f_p(1/n) - \frac{1}{3} \ln^{\frac{1}{p}}(vn) \ln \delta}{\alpha n}.$$

Further, for $\rho > 1$ and $Z^{(n)} := X \cdot \mathbb{I}\{X \leq (\lambda \ln(v\rho n))^{1/p}\}$, we have, with probability at

least $1-\delta$,

$$\forall n \geq 1,\ \widehat{\mathrm{C}}_n[Z^{(n)}] - \mathrm{C}[X] \leq 2\lambda^{\frac{1}{p}} \sqrt{\frac{2\rho f_{p/2}(\alpha) \cdot \ln \frac{\phi_\rho(n)}{\delta}}{\alpha n}} + \lambda^{\frac{1}{p}} \cdot \frac{f_p\left(\frac{1}{\rho n}\right) + \frac{2}{3}\ln^{\frac{1}{p}}(v\rho n) \ln \frac{\phi_\rho(n)}{\delta}}{\alpha n}.$$

Proof. The proof follows by (7.12) and the bound on $|\mathrm{C}[Z] - \mathrm{C}[X]|$ [resp. $|\mathrm{C}[Z^{(n)}] - \mathrm{C}[X]|$] and $\mathbb{E}[(\mu_\alpha + \alpha^{-1}[Z-\mu_\alpha]_+)^2]$ [resp. $\mathbb{E}[(\mu_\alpha + \alpha^{-1}[Z^{(n)}-\mu_\alpha]_+)^2]$] obtained from Proposition 7.6. □

7.5.3 Proof of Theorem 7.9

We state and prove the following extension of Theorem 7.7:

Theorem 7.26. *Let $\alpha, \delta \in (0,1)$, $v > 0$, $p > 2$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. random variables such that X satisfies the tail probability in* (7.8)*. Further, let $\phi_\rho(\cdot)$ be as in Theorem 7.3. Then, for $n \geq 1$ and $Z := X \cdot \mathbb{I}\{X \leq (-3pvn/\ln\delta)^{1/p}\}$, we have, with probability at least $1-\delta$,*

$$\widehat{\mathrm{C}}_n[Z] - \mathrm{C}[X] \leq \frac{v^{\frac{1}{p}} p^{1+\frac{1}{p}}}{\alpha} \left(\frac{(1/3)^{1-\frac{1}{p}} + 2^{\frac{1}{2}} 3^{\frac{1}{p}-\frac{1}{2}}}{p-1} \right) \left(\frac{\ln \delta^{-1}}{n} \right)^{1-\frac{1}{p}}.$$

Further, for $\rho > 1$ and $Z^{(n)} := X \cdot \mathbb{I}\{X \leq (\frac{3\rho pvn}{2\ln(\phi_\rho(n)/\delta)})^{1/p}\}$, we have, with probability at least $1-\delta$,

$$\forall n \geq 1,\ \widehat{\mathrm{C}}_n[Z^{(n)}] - \mathrm{C}[X] \leq \frac{(\rho v)^{\frac{1}{p}} p^{1+\frac{1}{p}}}{\alpha} \left(\frac{(2/3)^{1-\frac{1}{p}} + \rho^{\frac{1}{2}} 2^{\frac{1}{2}} 3^{\frac{1}{p}-\frac{1}{2}}}{p-1} \right) \left(\frac{\ln \frac{\phi_\rho(n)}{\delta}}{n} \right)^{1-\frac{1}{p}}.$$

Proof. The proof follows by (7.12) and the bound on $|\mathrm{C}[Z] - \mathrm{C}[X]|$ [resp. $|\mathrm{C}[Z^{(n)}] - \mathrm{C}[X]|$] and $\mathbb{E}[(\mu_\alpha + \alpha^{-1}[Z-\mu_\alpha]_+)^2]$ [resp. $\mathbb{E}[(\mu_\alpha + \alpha^{-1}[Z^{(n)}-\mu_\alpha]_+)^2]$] obtained from Proposition 7.6. □

7.5.4 Proof of Theorem 7.13

We state and prove the following extension of Theorem 7.13:

Theorem 7.27. *Let $\alpha, \delta \in (0,1)$, $\lambda, p, v > 0$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. random variables such that X satisfies the tail probability in* (7.7)*. Further, let ϵ_n, ϵ_n' and $\phi_\rho(n)$ be as in Theorem 7.3, and $f_p(\cdot)$ as in* (7.23)*. Then, for $n \geq 1$ and $Z := X \cdot \mathbb{I}\{X \leq (\lambda \ln(vn))^{1/p}\}$, we have, with probability at least $1-2\delta$,*

$$\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z] \leq \lambda^{\frac{1}{p}} \sqrt{\frac{2 f_{p/2}(\alpha) \ln \delta^{-1}}{\alpha n}} + \lambda^{\frac{1}{p}} \cdot \frac{3 f_p(1/n) - \ln^{\frac{1}{p}}(vn) \ln \delta}{3\alpha n} + |Z_{(\lceil \alpha n \rceil)}| \cdot \epsilon_n.$$

Further, for $\rho > 1$ and $Z^{(n)} := X \cdot \mathbb{I}\{X \leq (\lambda \ln(v\rho n))^{1/p}\}$, we have, with probability at least $1 - 2\delta$, $\forall n \geq 1$,

$$\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z^{(n)}] \leq \lambda^{\frac{1}{p}} \sqrt{\frac{2\rho f_{p/2}(\alpha) \ln \frac{\phi_\rho(n)}{\delta}}{\alpha n}} + \lambda^{\frac{1}{p}} \cdot \frac{3 f_p\left(\frac{1}{\rho n}\right) + 2 \ln^{\frac{1}{p}}(vn) \ln \frac{\phi_\rho(n)}{\delta}}{3\alpha n} + |Z^{(n)}_{(\lceil \alpha n \rceil)}| \cdot \epsilon'_n.$$

Proof. The proof follows by Theorem 7.3 and 1) the bound on $\mathrm{C}[Z^2]$ [resp. $\mathrm{C}[(Z^{(n)})^2]$] from Proposition 7.12; and 2) the bound on $|\mathrm{C}[Z] - \mathrm{C}[X]|$ [resp. $|\mathrm{C}[Z^{(n)}] - \mathrm{C}[X]|$] from Proposition 7.6. □

7.5.5 Proof of Theorem 7.14

We state and prove the following extension of Theorem 7.14:

Theorem 7.28. *Let $\alpha, \delta \in (0,1)$, $v > 0$, $p > 2$, and $X, X_1, X_2 \cdots \in \mathbb{R}$ be i.i.d. random variables such that X satisfies the tail probability in* (7.8)*. Further, let ϵ_n, ϵ'_n and $\phi_\rho(n)$ be as in Theorem 7.3. Then, for $n \geq 1$ and $Z := X \cdot \mathbb{I}\{X \leq (-3pvn/\ln\delta)^{1/p}\}$, we have, with probability at least $1 - 2\delta$,*

$$\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z] \leq \frac{v^{\frac{1}{p}} p^{1+\frac{1}{p}}}{\alpha} \left(\frac{(1/3)^{1-\frac{1}{p}} + 2^{\frac{1}{2}} 3^{\frac{1}{p}-\frac{1}{2}}}{p-1} \right) \left(\frac{\ln \delta^{-1}}{n} \right)^{1-\frac{1}{p}} + |Z_{(\lceil \alpha n \rceil)}| \cdot \epsilon_n.$$

Further, for $\rho > 1$ and $Z^{(n)} := X \cdot \mathbb{I}\{X \leq (\frac{3\rho pvn}{2\ln(\phi_\rho(n)/\delta)})^{1/p}\}$, we have, with probability at least $1 - 2\delta$,

$$\forall n \geq 1, \;\; \mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z^{(n)}] \leq \frac{(\rho v)^{\frac{1}{p}} p^{1+\frac{1}{p}}}{\alpha} \left(\frac{(2/3)^{1-\frac{1}{p}} + \rho^{\frac{1}{2}} 2^{\frac{1}{2}} 3^{\frac{1}{p}-\frac{1}{2}}}{p-1} \right) \left(\frac{\ln \frac{\phi_\rho(n)}{\delta}}{n} \right)^{1-\frac{1}{p}} + |Z^{(n)}_{(\lceil \alpha n \rceil)}| \cdot \epsilon'_n.$$

Proof. The proof follows by Theorem 7.3 and 1) the bound on $\mathrm{C}[Z^2]$ [resp. $\mathrm{C}[(Z^{(n)})^2]$] from Proposition 7.12; and 2) the bound on $|\mathrm{C}[Z] - \mathrm{C}[X]|$ [resp. $|\mathrm{C}[Z^{(n)}] - \mathrm{C}[X]|$] from Proposition 7.6. □

7.5.6 Proof of Theorem 7.15

We closely follow the steps in the proof of [Jamieson et al., 2014, Theorem 2] with small modifications. We now introduce two useful inequalities:

Lemma 7.29. *For all $\rho > 1$, $t \geq 1$, $c > 0$, and $y \in (0,1]$, we have*

$$\frac{1}{t} \ln \frac{\ln(\rho t)}{y} \geq c \implies t \leq \frac{1}{c} \ln \frac{2 \ln\left(\frac{\rho}{cy}\right)}{y}, \tag{7.71}$$

and for all $s \geq e$, $\rho \in (1,2)$, and $0 < y \leq \delta \leq e^{-2}$, we have

$$\frac{1}{t}\ln\frac{\ln(\rho t)}{y} \geq \frac{c}{s}\ln\frac{\ln(\rho s)}{\delta} \implies t \leq \frac{s}{c\ln(1/\delta)}\ln\frac{2\ln\left(\frac{1}{cy}\right)}{y}. \tag{7.72}$$

Proof. Rearranging the LHS inequality in (7.71) yields,

$$t \leq \frac{1}{c}\ln\frac{\ln(\rho t)}{y}. \tag{7.73}$$

The LHS inequality in (7.71) also implies that

$$c \leq \frac{1}{t}\ln\frac{\ln(\rho t)}{y} \leq \frac{1}{ty}\ln(\rho t) \leq \frac{\rho}{y}\frac{\ln(\rho t)}{\rho t} \leq \frac{\sqrt{\rho}}{y\sqrt{t}}, \tag{7.74}$$

where the last inequality follows by the fact that $(\ln x)/x \leq 1/\sqrt{x}$, for all $x > 0$. Rearranging (7.74) yields $t \leq \frac{\rho}{c^2y^2}$. Plugging this into (7.73) implies (7.71).

We now show (7.72). Note that since $s \geq e$ and $\rho > 1$ the LHS inequality in (7.72) implies that

$$\frac{1}{t}\ln\frac{\ln(\rho t)}{y} \geq \frac{c}{s}\ln\frac{1}{\delta}.$$

Applying (7.71) to this inequality with c set to $c/s\ln(1/\delta)$ implies that

$$t \leq \frac{s}{c\ln(1/\delta)}\ln\frac{2\ln\left(\frac{\rho}{yc\ln(1/\delta)}\right)}{y}. \tag{7.75}$$

Now since $\delta \leq e^{-2}$, we have $\rho/(\ln(1/\delta)) \leq \rho/2 \leq 1$. Plugging this into (7.75) yields the desired result. □

Lemma 7.30. *Let g and τ be as in Theorem 7.15. If for some $s > 0$, $\mathbf{P}[\mathcal{E}_a(y)] \geq 1 - sy$, for all $y \in (0,1)$ and $a \in [K]$, then for all $\delta \in (0,1)$ and $G_\alpha := \rho\nu^2(\nu+1)^2 g(\alpha)$:*

$$\mathbf{P}\left[\forall t \geq \tau,\ \sum_{a=2}^{K} N_a(t) \leq K + G_\alpha \sum_{a=2}^{K} \frac{8\ln\left(\frac{s}{\delta}\right) + \ln\left(\frac{2}{\delta}\ln\left(\frac{e}{\delta} \vee \frac{\rho G_\alpha}{\delta\Delta_a^2}\right)\right)}{\Delta_a^2}\right] \geq 1 - (s+1)\delta. \tag{7.76}$$

Proof. Let $a \in \{2,\ldots,K\}$ and $\delta, y \in (0,1)$. Assume that $\mathcal{E}_1(\delta)$ and $\mathcal{E}_a(y)$ hold true and that $A_t = a$. In this case, we have for $t \geq \tau$,

$$\begin{aligned}
\mathrm{C}_a - \xi(N_a(t), y) - \nu\xi(N_a(t), \delta) &\leq \widehat{\mathrm{C}}_{N_a(t),a} - \nu\xi(N_a(t), \delta), \\
&\leq \widehat{\mathrm{C}}_{N_1(t),1} - \nu\xi(N_1(t), \delta), \quad \text{(because } A_t = a\text{)} \\
&\leq \mathrm{C}_1.
\end{aligned} \tag{7.77}$$

This implies that $\Delta_i \leq (1+\nu)\cdot\xi(N_a(t), y\wedge\delta)$. Thus, using (7.71) with $c = G_\alpha/\Delta_a^2$, we get that under the event $\mathcal{E}_1(\delta)\cap\mathcal{E}_a(y)\cap\{A_t = a\}$:

$$
\begin{aligned}
N_a(t) &\leq \frac{G_\alpha}{\Delta_a^2}\ln\frac{2\ln\left(\frac{\rho G_\alpha}{\Delta_a^2\cdot(y\wedge\delta)}\right)}{y\wedge\delta},\\
&\leq \frac{G_\alpha}{\alpha\Delta_a^2}\ln\frac{2\ln\left(\frac{\rho G_\alpha e}{\Delta_a^2\delta y}\right)}{y\delta},\\
&\leq \frac{G_\alpha}{\Delta_a^2}\ln\frac{2\ln\left(\left(e\vee\frac{\rho G_\alpha}{\Delta_a^2 y}\right)\frac{1}{\delta}\cdot\frac{e}{y}\right)}{y\delta},\\
&\leq \frac{G_\alpha}{\Delta_a^2}\left(\ln\frac{2}{y\delta}+\ln\left(\ln\left(\left(e\vee\frac{\rho G_\alpha}{\Delta_a^2 y}\right)\frac{1}{\delta}\right)+\ln\left(\frac{e}{y}\right)\right)\right),\\
&\leq \frac{G_\alpha}{\Delta_a^2}\left(\ln\frac{2}{y\delta}+\ln\ln\left(\left(e\vee\frac{\rho G_\alpha}{\Delta_a^2 y}\right)\frac{1}{\delta}\right)+\ln\ln\left(\frac{e}{y}\right)\right), \text{ (ln sub-additive in } \mathbb{R}_{\geq 1})\\
&= \frac{G_\alpha}{\Delta_a^2}\ln\frac{2\ln\left(\left(e\vee\frac{\rho G_\alpha}{\Delta_a^2}\right)\frac{1}{\delta}\right)}{\delta}+\frac{G_\alpha}{\Delta_a^2}\ln\frac{\ln(e/y)}{y},\\
&\leq \frac{G_\alpha}{\Delta_a^2}\ln\frac{2\ln\left(\left(e\vee\frac{\rho G_\alpha}{\Delta_a^2}\right)\frac{1}{\delta}\right)}{\delta}+\frac{2G_\alpha}{\Delta_a^2}\ln(1/y),
\end{aligned}
\tag{7.78}
$$

where the last inequality follows by the fact that $\ln(e/x)/x \leq 1/x^2$, for all $x\in(0,1)$. Define

$$
\nabla_a := \frac{G_\alpha}{\Delta_a^2}\ln\left(\frac{2}{\delta}\ln\left(\frac{e}{\delta}\vee\frac{\rho G_\alpha}{\delta\Delta_a^2}\right)\right).
$$

Since $N_a(t)$ only increases when arm a is pulled, (7.78) implies that for all $t\geq\tau$,

$$
N_a(t)\cdot\mathbb{I}\{\mathcal{E}_1(\delta)\cap\mathcal{E}_a(y)\} \leq 1+\nabla_a+\frac{2G_\alpha}{\Delta_a^2}\ln(1/y). \tag{7.79}
$$

Now define

$$
\Gamma_a := \sup\{y>0 : \mathcal{E}_a(y) \text{ holds}\}. \tag{7.80}
$$

By assumption, we have $\mathbf{P}[\Gamma_a \leq y] \geq 1 - ys$. Furthermore, (7.79) can be written as

$$
N_a(t)\cdot\mathbb{I}\{\mathcal{E}_1(\delta)\} \leq 1+\nabla_a+\frac{2G_\alpha}{\Delta_a^2}\ln(1/\Gamma_a).
$$

Using this inequality and the fact that $\mathbf{P}[\mathcal{E}_1(\delta)^c] \leq \delta s$ we have for any $x > 0$,

$$\begin{aligned}\mathbf{P}\left[\forall t \geq 1,\ \sum_{a=2}^{K} N_a(t) > x + \sum_{a=2}^{K}(\nabla_a + 1)\right] &\leq \delta s \\ &+ \mathbf{P}\left[\forall t \geq 1,\ \sum_{a=1}^{K} N_a(t) > x + \sum_{a=1}^{K}(\nabla_a + 1)\middle| \mathcal{E}_1(\delta)\right], \\ &\leq \delta s + \mathbf{P}\left[\sum_{a=1}^{K} \frac{2G_\alpha}{\Delta_a^2}\ln(1/\Gamma_a) > x\right]. \quad (7.81)\end{aligned}$$

Now define the random variable $Z_a := -\frac{2G_\alpha}{\Delta_a^2}\ln(s\Gamma_a)$, for $a \in [K]$. Since $\mathbf{P}[\Gamma_a < y] \leq ys$, we have $\mathbf{P}[Z_a > x] \leq \exp(-x/\lambda_a)$, where $\lambda_a := 2G_\alpha/\Delta_a^2$; that is Z_a is sub-exponential with parameter λ_a. Further, since $Z_1, \ldots, Z_K$ are independent, we have by standard techniques which bound the sum of sub-exponential random variables Jamieson et al. [2014]:

$$\mathbf{P}\left[\sum_{a=1}^{K} Z_a \geq z\right] \leq \exp\left(\frac{-z^2}{4\|\boldsymbol{\lambda}\|_2^2} \vee \frac{-z}{4\|\boldsymbol{\lambda}\|_\infty}\right) \leq \exp\left(\frac{-z^2}{4\|\boldsymbol{\lambda}\|_1^2} \vee \frac{-z}{4\|\boldsymbol{\lambda}\|_1}\right). \quad (7.82)$$

Combining (7.81) and (7.82) with $x = z + \|\lambda\|_1 \ln s$ and $z = 4\|\lambda\|_1 \ln(1/\delta)$ yields

$$\mathbf{P}\left[\sum_{a=2}^{K} N_a(t) > \sum_{a=2}^{K}\left(\frac{8\theta\ln(\delta^{-1}s)}{\alpha\Delta_a^2} + \nabla_a + 1\right)\right] \leq \delta \cdot (s+1). \quad (7.83)$$

□

Now to the final technical result:

Lemma 7.31. *Let $\nu > 1$ and $\rho \in (1,2)$. Further, suppose that for some $s > 0$, $\mathbf{P}[\mathcal{E}_a(y)] \geq 1 - sy$, for all $y \in (0,1)$ and $a \in [K]$. If*

$$\kappa \geq \frac{1}{1 - s\delta - \sqrt{-(s\delta)^{7/16}\ln(s\delta)}}\left(\frac{\ln\frac{2(\nu+1)}{\nu-1}}{\ln(1/\delta)} + 1\right)\frac{(1+\nu)^2}{(\nu-1)^2}, \quad (7.84)$$

for $\delta \in (0, e^{-2} \wedge e^{-4}/s)$, then for all $2 \leq a \leq K$ and τ as in Theorem 7.15, we have

$$\mathbf{P}\left[\forall a \in [K], \forall t \geq \tau : N_a(t) < 1 + \kappa \sum_{a' \neq a} N_{a'}(t)\right] \geq 1 - 4s\delta - (s\delta)^{1/8}.$$

Proof. Let $a > b$ and $y, \delta \in (0,1)$ with δ such that $\delta \leq e^{-2} \wedge e^{-4}/s$. Assuming that

the event $\mathcal{E}_a(y) \cap \mathcal{E}_b(\delta) \cap \{A_t = a\}$ holds, we have, for all $t \geq \tau$,

$$\begin{aligned} \mathrm{C}_a - \xi(N_a(t), y) - \nu\xi(N_a(t), \delta) &\leq \widehat{\mathrm{C}}_{N_a(t),a} - \nu\xi(N_a(t), \delta), \\ &\leq \widehat{\mathrm{C}}_{N_b(t),b} - \nu\xi(N_b(t), \delta), \\ &\leq \mathrm{C}_b - (\nu - 1)\xi(N_b(t), \delta). \end{aligned}$$

Since $\mathrm{C}_b \leq \mathrm{C}_a$ by assumption, the last inequality implies that

$$(1 + \nu)\xi(N_a(t), y \wedge \delta) \geq (\nu - 1)\xi(N_a(t), \delta).$$

Thus, using (7.72) with $c = (\nu - 1)^2/(\nu + 1)^2$, we get under the event $\mathcal{E}_a(y) \cap \mathcal{E}_b(\delta) \cap \{A_t = a\}$:

$$N_a(t) \leq \frac{(1 + \nu)^2}{(\nu - 1)^2 \ln(1/\delta)} \ln \left(\frac{2 \ln \left(\frac{(1+\nu)^2}{(\nu-1)^2 (y \wedge \delta)} \right)}{y \wedge \delta} \right) \cdot N_b(t).$$

Using this together with the fact that $N_a(t)$ only increases if the event $\{A_t = a\}$ holds, we get

$$(N_a(t) - 1) \cdot \mathbb{I}\{\mathcal{E}_a(y) \cap \mathcal{E}_b(\delta)\} \leq \frac{(1 + \nu)^2}{(\nu - 1)^2 \ln(1/\delta)} \ln \left(\frac{2 \ln \left(\frac{(1+\nu)^2}{(\nu-1)^2 (y \wedge \delta)} \right)}{y \wedge \delta} \right) \cdot N_b(t). \tag{7.85}$$

Now let $a > 1$ and $y = \delta^{a-1}$. Using the fact that $\ln(1/x)/x \leq 1/x^2$, for all $x \in (0, 1)$, we get

$$\begin{aligned} \frac{2 \ln \left(\frac{(1+\nu)^2}{(\nu-1)^2 (y \wedge \delta)} \right)}{y \wedge \delta} &\leq \frac{2(1 + \nu)^2}{(\nu - 1)^2 (y \wedge \delta)^2}, \\ &\leq \frac{4(1 + \nu)^2}{(\nu - 1)^2 \delta^{2(a-1)}}. \quad (\text{since } y = \delta^{a-1}) \end{aligned} \tag{7.86}$$

As a result, we have

$$\ln \frac{2 \ln \left(\frac{(1+\nu)^2}{(\nu-1)^2 (y \wedge \delta)} \right)}{y \wedge \delta} \leq 2(a - 1) \left(\ln \frac{2(1 + \nu)}{\nu - 1} + \ln(1/\delta) \right). \tag{7.87}$$

Using (7.87) together with (7.85) and the choice of $y = \delta^{a-1}$, we get that for $\mu > 0$

$$\mathbb{I}\{\mathcal{E}_a(\delta^{a-1})\} \cdot \frac{1}{a-1} \sum_{b=1}^{a-1} \mathbb{I}\{\mathcal{E}_b(\delta)\} > 1 - \mu \implies (1-\mu) \cdot (N_a(t) - 1) \leq \kappa' N_b(t), \tag{7.88}$$

$$\text{where} \quad \kappa' := \frac{2(\nu+1)^2}{(\nu-1)^2} \left(\frac{\ln \frac{2(\nu+1)}{\nu-1}}{\ln(1/\delta)} + 1 \right). \tag{7.89}$$

Thus, using the fact that $\mathbf{P}[\mathcal{E}_a(y)] \geq 1 - ys$ by assumption, we have for $\mu > 0$

$$\begin{aligned}
&\mathbf{P}\left[\exists (a,t) \in [2..K] \times [\tau, +\infty) : (1-\mu)(N_a(t) - 1) \geq \kappa' \sum_{b \neq a} N_b(t)\right] \\
&\quad \leq \mathbf{P}\left[\exists a \in [2..K] : \frac{\mathbb{I}\{\mathcal{E}_a(\delta^{a-1})\}}{a-1} \sum_{b=1}^{a-1} \mathbb{I}\{\mathcal{E}_b(\delta)\} \leq 1 - \mu\right], \\
&\quad \leq \sum_{a=1}^{K} \mathcal{E}_a(\delta^{a-1}) + \sum_{a=1}^{K} \mathbf{P}\left[\frac{1}{a-1} \sum_{b=1}^{a-1} \mathbb{I}\{\mathcal{E}_b(\delta)\} \leq 1 - \mu\right].
\end{aligned} \tag{7.90}$$

Now, by denoting $\delta_s := s\delta$ and using Hoeffding's inequality, we can bound the RHS of (7.90) by

$$\begin{aligned}
&\mathbf{P}\left[\exists a \in [2..K] : \frac{\mathbb{I}\{\mathcal{E}_a(\delta^{a-1})\}}{a-1} \sum_{b=1}^{a-1} \mathbb{I}\{\mathcal{E}_b(\delta)\} \leq 1 - \delta_s - (\mu - \delta_s)\right] \\
&\qquad \leq ((a-1)\delta_s) \wedge e^{-2(a-1)(\mu-\delta_s)^2}.
\end{aligned} \tag{7.91}$$

Thus, if we let $b_\star := \lceil \delta_s^{-7/16}/2 \rceil$ and set $\mu_\star = \delta_s + \sqrt{\delta_s^{7/16} \ln(1/\delta_s)}$, we obtain from (7.90) and (7.91) that

$$\begin{aligned}
&\mathbf{P}\left[\exists (a,t) \in [2..K] \times [\tau, +\infty) : (1-\mu_\star)(N_a(t) - 1) \geq \kappa' \sum_{b \neq a} N_b(t)\right] \\
&\quad \leq \sum_{a=1}^{K} \left(s\delta^{a-1} + ((a-1)\delta_s) \wedge \exp(-2(a-1)(\mu_\star - \delta_s)^2) \right), \\
&\quad \leq \frac{\delta_s}{1-\delta} + \delta_s b_\star^2 + \frac{e^{-2b_\star \delta_s^{7/16} \ln(1/\delta_s)}}{1 - e^{-2\delta_s^{7/16} \ln(1/\delta_s)}}, \\
&\quad \leq \frac{\delta_s}{1-\delta} + \delta_s b_\star^2 + \frac{\delta_s}{1 - e^{-2\delta_s^{7/16} \ln(1/\delta_s)}},
\end{aligned} \tag{7.92}$$

where the last inequality follows by the definition of $b_\star$. Now using the fact that

$e^x \leq 1 + x^2$, for all $x \leq 0$ and the assumption that $\delta_s \leq 1/e$, we get from (7.92) that

$$\begin{aligned}
p &:= \mathbf{P}\left[\exists (a,t) \in [2..K] \times [\tau, +\infty) : (1-\mu_\star)(N_a(t)-1) \geq \kappa' \sum_{b \neq a} N_b(t)\right], \\
&\leq \frac{\delta_s}{1-\delta} + \delta_s b_\star^2 + \frac{\delta_s}{4\delta_s^{7/8} \ln^2(1/\delta_s)}, \\
&\leq \frac{\delta_s}{1-\delta} + \delta_s b_\star^2 + \frac{\delta_s^{1/8}}{4}, \quad (\delta_s \leq e^{-1}) \\
&\leq 2\delta_s + \delta_s(\delta_s^{-7/8}/2 + 2) + \delta_s^{1/8}/4, \\
&\leq 4\delta_s + 3\delta_s^{1/8}/4, \\
&\leq 4\delta_s + \delta_s^{1/8}.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 7.15. Lemma 7.31 shows that the stopping condition is never triggered by sub-optimal arms with probability at least $1 - 4s\delta - \delta^{1/8}$. On the other hand, Lemma 7.30 shows that the total number of time sub-optimal arms are pulled is, with probability at least $1 - (s+1)\delta$, bounded as a function of the sub-optimality gaps, as long as κ satisfies (7.84). Under these two event, the stopping condition occurs after T such that

$$\begin{aligned}
T = \sum_{a=1}^{K} N_a(T) &\leq K\tau + 2 + (1+\kappa) \sum_{a=2}^{K} N_a(T), \\
&\leq 2 + (1+\kappa+\tau)K + (1+\kappa)G_\alpha \sum_{a=2}^{K} \frac{8\ln\left(\frac{s}{\delta}\right) + \ln\left(\frac{2}{\delta}\ln\left(\frac{e}{\delta} \vee \frac{\rho G_\alpha}{\delta \Delta_a^2}\right)\right)}{\Delta_a^2}.
\end{aligned}$$

The proof is completed by noting that

$$\begin{aligned}
\kappa &\geq \frac{1}{1 - s\delta - \sqrt{-(s\delta)^{7/16}\ln(s\delta)}} \left(\frac{\ln\frac{2(\nu+1)}{\nu-1}}{\ln(1/\delta)} + 1\right) \frac{(1+\nu)^2}{(\nu-1)^2}, \\
&\geq \left(\ln\frac{2(\nu+1)}{\nu-1} + 2\right) \frac{7(1+\nu)^2}{2(\nu-1)^2}.
\end{aligned} \tag{7.93}$$

□

7.5.7 Proof of Theorem 7.16

Proof. First, note that we have

$$\phi_\rho(n) := c\ln_\rho(\rho n)\ln^2(\ln_\rho(\rho n)) \leq \frac{c}{\ln^2\rho}\ln^2(\rho n). \tag{7.94}$$

Therefore, by Theorem 7.16 we have with probability $1 - 2\delta\sqrt{c}/\ln\rho$, for all $n \geq 1$

$$\mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[Z^{(n)}] + \delta_n + |Z^{(n)}_{(\lceil \alpha n \rceil)}| \cdot \epsilon'_n \leq \widehat{\mathrm{C}}_n[Z^{(n)}] + \delta_n + \widehat{\mathrm{C}}_n[Z^{(n)}] \cdot \epsilon'_n,$$

$$\text{where} \quad \delta_n = 2\lambda \cdot (1 - \ln\alpha)\sqrt{\frac{2\rho \ln \frac{\ln(\rho n)}{\delta}}{\alpha n}} + \lambda \frac{1 + 2\ln(\rho n) + \frac{4}{3}\ln(\rho n) \ln \frac{\ln(\rho n)}{\delta}}{\alpha n},$$

Now using (7.31), we get, with probability $1 - 2\delta\sqrt{c}/\ln\rho$, for all $n \geq 1$

$$\mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[Z^{(n)}] + \delta_n + \mathrm{C}[X] \cdot \epsilon'_n. \tag{7.95}$$

Plugging the bound on $\mathrm{C}[X]$ from Proposition 7.12 into (7.95) and using the fact that $f_1(\alpha)) = (1 - 2\ln(\alpha))$, we get, with probability $1 - 2\delta\sqrt{c}/\ln\rho$, for all $n \geq 1$,

$$\begin{aligned} \mathrm{C}[X] &\leq \widehat{\mathrm{C}}_n[Z^{(n)}] + \delta_n + \lambda(1 - 2\ln\alpha) \cdot \epsilon'_n, \\ &\leq \widehat{\mathrm{C}}_n[Z^{(n)}] + 4\lambda \cdot (1 - \ln\alpha)\sqrt{\frac{2\rho \ln \frac{\ln(\rho n)}{\delta}}{\alpha n}} + 2\lambda \frac{1 + \ln(\rho n/\alpha) + \frac{2}{3}\ln(\rho n) \ln \frac{\ln(\rho n)}{\delta}}{\alpha n}, \end{aligned} \tag{7.96}$$

where the last inequality holds for $n \geq \tau$, where $\tau \leq O((\ln\alpha \ln\delta)/\alpha)$. Now, by Corollary 7.21 and (7.94), we have with probability at least $1 - \delta\sqrt{c}/\ln\rho$, for all $n \geq 1$

$$\widehat{\mathrm{C}}_n[Z^{(n)}] \leq \mathrm{C}[X] + 2\lambda \cdot (1 - \ln\alpha)\sqrt{\frac{2\rho \ln \frac{\ln(\rho n)}{\delta}}{\alpha n}} + \lambda \cdot \frac{1 + 2\ln(\rho n) + \frac{4}{3}\ln(\rho n) \ln \frac{\ln(\rho n)}{\delta}}{\alpha n}. \tag{7.97}$$

Note that there exists $\tau \leq O((\ln\alpha \ln\delta)/\alpha)$, such that

$$\frac{2 + 2\ln(\rho n/\alpha) + \frac{4}{3}\ln(\rho n) \ln \frac{\ln(\rho n)}{\delta}}{\alpha n} \leq 4(1 - \ln\alpha)\sqrt{\frac{2\rho \ln \frac{\ln(\rho n)}{\delta}}{\alpha n}}, \quad \text{for all } n \geq \tau. \tag{7.98}$$

Combining this with (7.96) and (7.97), and applying a union bound, we get,

$$\mathbf{P}\left[\forall n \geq \tau, \quad |\widehat{\mathrm{C}}_n[Z^{(n)}] - \mathrm{C}[X]| \leq 4\sqrt{2}\lambda(1 - \ln\alpha)\sqrt{\frac{\rho \ln \frac{\ln(\rho n)}{\delta}}{\alpha n}}\right] \geq 1 - 2\delta\sqrt{c}/\ln\rho.$$

This completes the proof. □

7.5.8 Proof of Theorem 7.17

Proof. By Corollary 7.20, we have, with probability at least $1-2\delta$,

$$\mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[Z] + \delta_n + |Z_{(\lceil n\alpha \rceil)}| \cdot \epsilon_n,$$
$$\text{where} \quad \delta_n := 2\lambda \cdot (1-\ln\alpha)\sqrt{\frac{\ln\delta^{-1}}{\alpha n}} + \lambda \cdot \frac{3+6\ln n - \ln n \ln\delta}{3\alpha n}.$$

Now using the fact that the random variables are positive and (7.31), we get,

$$\mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[Z] + \delta_n + \mathrm{C}[X] \cdot \epsilon_n, \tag{7.99}$$

with probability at least $1-2\delta$. Plugging the bound on $\mathrm{C}[X]$ from Proposition 7.12 into (7.99) and using the fact that $f_1(\alpha)) = (1-2\ln(\alpha))$, we get, with probability at least $1-2\delta$,

$$\mathrm{C}[X] \leq \widehat{\mathrm{C}}_n[Z] + 4\lambda \cdot (1-\ln\alpha)\sqrt{\frac{\ln\delta^{-1}}{\alpha n}} + \lambda \cdot \frac{4+6\ln(n/\alpha) - 2\ln n \ln\delta}{3\alpha n}. \tag{7.100}$$

Now, by Corollary 7.8, we have, with probability at least $1-\delta$,

$$\widehat{\mathrm{C}}_n[Z] \leq \mathrm{C}[X] + 4\lambda \cdot (1-\ln\alpha)\sqrt{\frac{\ln\delta^{-1}}{\alpha n}} + \lambda \cdot \frac{3+6\ln n - \ln n \ln\delta}{3\alpha n}. \tag{7.101}$$

Therefore, for all $n \geq n_0$, where

$$n_0 = \sup\left\{ n : \lambda \cdot \frac{4+6\ln(n/\alpha) - 2\ln n \ln\delta}{3\alpha n} \leq 2\lambda \cdot (1-\ln\alpha)\sqrt{\frac{\ln\delta^{-1}}{\alpha n}} \right\},$$

we have from (7.100) and (7.101) and union bound, with probability at least $1-3\delta$,

$$|\mathrm{C}[X] - \widehat{\mathrm{C}}_n[Z]| \leq +5\lambda \cdot (1-\ln\alpha)\sqrt{\frac{\ln\delta^{-1}}{\alpha n}}.$$

This completes the proof. □

7.6 Discussion

Generalization bounds. As mentioned earlier, Mhammedi et al. [2020c] derived the first PAC-Bayesian generalization bound in the statistical learning setting with the optimal dependence in α. However, their results are restricted to bounded losses. The two-sided deviation bounds we presented in 7.2 can easily be combined with the McAllester's analysis (see e.g. [McAllester, 2003, Lemma 3]) to generate new PAC-Bayesian bounds for the unbounded losses (under a light or heavy-tailed distribution assumption) with a state-of-the-art dependence in α.

Limitations. One limitation of the concentration inequalities we present in this chapter for the unbounded setting lies in the moment assumptions we make in (7.5) and (7.6). Since we have not assumed the random variable X has zero mean, the assumptions in (7.5) and (7.6) are not translation invariant. For that matter, our estimators of CVaR are also not translation-invariant as they involve clipping. In some applications, such as in Bandits, it is desirable to have translation-invariant estimators. This can be achieved using the median of means Bubeck et al. [2013] instead of clipping. It is possible to apply the median of means in the context of our reduction (e.g. to control the deviation of the implicit random variable Y in (7.13)). While this would resolve the translation invariance issue, it would lead to a sub-optimal dependence in the parameter α in the unbounded settings we consider. Finding translation-invariant estimators for CVaR that have concentrate rates with the optimal dependence in α is an important future direction.

Chapter 8

Risk Monotonicity in Statistical Learning

Guarantees on the performance of machine learning algorithms are desirable, especially given the widespread deployment. A traditional performance guarantee often takes the form of a generalization bound, where the *expected* risk associated with hypotheses returned by an algorithm is bounded in terms of the corresponding empirical risk plus an additive error which typically converges to zero as the sample size increases. However, interpreting such bounds is not always straight forward and can be somewhat ambiguous. In particular, given that the error term in these bounds goes to zero, it is tempting to conclude that more data would monotonically decrease the expected risk of an algorithm such as the Empirical Risk Minimizer (ERM). However, this is not always the case; for example, Loog et al. [2019] showed that increasing the sample size by one, can sometimes make the test performance worse in expectation for commonly used algorithms such as ERM in popular settings including linear regression. This type of non-monotonic behaviour is still poorly understood and indeed not a desirable feature of an algorithm since it is expensive to acquire more data in many applications.

Non-monotonic behaviour of *risk curves* [Shalev-Shwartz and Ben-David, 2014]—the curve of the expected risk as a function of the sample size—has been observed in many previous works [Duin, 1995; Opper and Kinzel, 1996; Smola et al., 2000; Opper, 2001] (see also [Loog et al., 2019] for a nice account of the literature). At least two phenomena have been identified as being the cause behind such behaviour. The first one, coined *peaking* [Krämer, 2009; Duin, 2000], or *double descent* according to more recent literature [Belkin et al., 2018; Spigler et al., 2018; Belkin et al., 2019; Dereziński et al., 2019; Deng et al., 2019; Mei and Montanari, 2019; Nakkiran, 2019; Nakkiran et al., 2020a; Derezinski et al., 2020; Chen et al., 2020; Cheema and Sugiyama, 2020; d'Ascoli et al., 2020; Nakkiran et al., 2020b], is the phenomenon where the risk curve peaks at a certain sample size n. This sample size typically represents the crossover point from an over-parameterized to under-parameterized model. For example, when the number of data points is less than the number of parameters of a model (over-parameterized model), such as Neural Networks, the expected risk can typically increase until the number of data points exceeds the number of parameters

(under-parameterized model). The second phenomenon is known as *dipping* [Loog and Duin, 2012; Loog, 2015], where the risk curve reaches a minimum at a certain sample size n and increases after that—never reaching the minimum again even for very large n. This phenomenon typically happens when the algorithm is trained on a surrogate loss that differs from the one used to evaluate the risk [Ben-David et al., 2012].

It is becoming more apparent that the two phenomena just mentioned (double descent and dipping) do not fully characterize when non-monotonic risk behaviour occurs. Loog et al. [2019] showed that non-monotonic risk behaviour could happen outside these settings and formally prove that the risk curve of ERM is non-monotonic in linear regression with prevalent losses. The most striking aspect of their findings is that the risk curves in some of the cases they study can display a perpetual "oscillating" behaviour; there is no sample size beyond which the risk curve becomes monotone. In such cases, the risk's non-monotonicity cannot be attributed to the peaking/double descent phenomenon. Moreover, they rule out the dipping phenomenon by studying the ERM on the actual loss (not a surrogate loss).

The findings of Loog et al. [2019] stress our current lack of understanding of generalization. This was echoed more particularly by Viering et al. [2019a], who posed the following question as part of a COLT open problem:

How can we provably avoid non-monotonic behaviour?

Contributions. In this work, we answer the above question by presenting algorithms that are both consistent and risk-monotonic under weak assumptions. We study the guarantees of our algorithms in the standard statistical learning setting with bounded losses. For the first variant of our algorithm, we require a finite Rademacher complexity of the loss composed with the hypothesis class. Under this condition, we show that the algorithm is risk-monotonic and provide the rate of convergence of its excess-risk, which matches (up to log-factors) the optimal rate [Dudley, 1984; Talagrand, 1994; Boucheron et al., 2005; Zhivotovskiy and Hanneke, 2018] one would get without further assumptions on the loss function or the data-generating distribution[1]. We present a second variant of the algorithm that is also risk-monotonic and achieves faster excess-risk rates under the Bernstein condition [Bartlett and Mendelson, 2006b]. This shows that, as far as excess-risk rates are concerned, risk-monotonicity comes at virtually no price. Our results exceed the expectations expressed by Viering et al. [2019a] since these algorithms have monotonic risk curves for all sample sizes $n \geq 1$ as opposed to only for large enough n.

We go a step further by showing that risk-monotonicity in high probability (instead of in expectation) is also possible. In particular, we present an algorithm for this case which also achieves fast excess-risk rates under the Bernstein condition. In order to derive these results, we make use of the recent online convex optimization

[1] Additional assumptions may include mixability/exp-concavity of the loss [Van Erven et al., 2015; Mehta, 2017], and/or the Bernstein/Tsybakov condition on the data-generating distribution [Bousquet et al., 2004; Bartlett and Mendelson, 2006a].

algorithm FreeGrad [Mhammedi and Koolen, 2020] to prove a new time-uniform concentration inequality (of independent interest) for non-i.i.d. random variables with the same conditional mean. This inequality can be viewed as an empirical version of Freedman's inequality [Freedman, 1975] or an extension of the empirical Bernstein inequality [Maurer and Pontil, 2009] to non-i.i.d. random variables with the same conditional mean. Crucially, given a fixed confidence level, our concentration inequality holds for all sample sizes n, simultaneously (hence, the name time-uniform), which is what we need to achieve risk-monotonicity in high probability.

Approach overview. Given n samples, the key idea behind our approach is to iteratively generate a sequence of hypotheses $\hat{h}_1, \hat{h}_2, \ldots$ leading up to $\hat{h}_n$, where we only allow consecutive hypotheses, say $\hat{h}_{k-1}$ and $\hat{h}_k$ to differ if we can guarantee (with high enough confidence) that the risk associated with $\hat{h}_k$ is lower than that of $\hat{h}_{k-1}$. To test for this, we compare the empirical losses of $\hat{h}_{k-1}$ and $\hat{h}_k$, taking into account the potential gap between empirical and population expectations. Further, we provide a way of estimating this gap from samples using different types of concentration inequalities.

Related works. Much work has already been done in efforts to mitigate the non-monotonic behaviour of risk curves [Viering et al., 2019b; Nakkiran et al., 2020b; Loog et al., 2019]. For example, in the supervised learning setting with the zero-one loss, Ben-David et al. [2011] introduced the "memorize" algorithm that predicts the majority label on any test instance x that was observed during training; otherwise, a default label is predicted. Ben-David et al. [2011] showed that this algorithm is risk-monotonic. However, it is unclear how their result could generalize beyond the particular setting they considered. Risk-monotonic algorithms are also known for the case where the model is correctly specified (see Loog et al. [2019] for an overview); in this chapter, we do not make such an assumption.

Closer to our work is that of Viering et al. [2019b] who, like us, also used the idea of only updating the current predictor for sample size n if it has a lower risk than the predictor for sample size $n-1$. They determine whether this is the case by performing statistical tests on a validation set (or through cross-validation). They introduce algorithm wrappers that ensure that the risk curves of the final algorithms are monotonic with high probability. However, their results are specialized to the 0-1 loss, and they do not answer the question by Viering et al. [2019a] on the existence of learners that guarantee a monotonic risk in *expectation*.

Outline. In Section 8.1, we introduce the setting, notation, and relevant definitions. In Section 8.2, we present our risk-monotonic algorithms that answer to the question posed by Viering et al. [2019a]. Section 8.3 is dedicated to the proof of risk-monotonicity of one of our algorithms. In Section 8.4, we study risk-monotonicity in high-probability and present an algorithm for this case that is based on a new concentration inequality. In Section 8.5, we present an efficient version of our algorithm

for the case of convex losses. We conclude with a discussion in Section 8.11. All remaining proofs are differed to Sections 8.6, 8.7, 8.9, and 8.10.

8.1 Preliminaries

In this section, we present our notation and the relevant definitions needed for the rest of the chapter.

Setting and notation. Let $\mathcal{Z}$ [resp. $\mathcal{H}$] be an arbitrary feature [resp. hypothesis] space, and let $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$ be a bounded loss function. We denote by $\mathscr{P}(\mathcal{Z})$ the set of probability measures over $\mathcal{Z}$, and by $\mathscr{F}(\mathcal{Z}, \mathbb{R})$ the set of all bounded measurable functions from $\mathcal{Z}$ to $\mathbb{R}$. Data is represented by a random variable $Z \in \mathcal{Z}$ which we assume to be distributed according to an unknown distribution $P \in \mathscr{P}(\mathcal{Z})$. A *learning problem* is a tuple $(P, \ell, \mathcal{H})$. We assume throughout that $Z_1, Z_2, \ldots$ are i.i.d. copies of Z, and we denote by $\widehat{P}_n$ the empirical distribution defined by

$$\widehat{P}_n(\cdot) := \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}(\cdot), \quad n \in \mathbb{N},$$

where $\delta_z(\cdot)$ represents the Dirac distribution at $z \in \mathcal{Z}$. The empirical and population risks are

$$\widehat{L}_n(h) := \mathbb{E}_{\widehat{P}_n(Z)}[\ell(h, Z)] \quad \text{and} \quad L(h) := \mathbb{E}_{P(Z)}[\ell(h, Z)],$$

respectively, for $n \in \mathbb{N}$ and $h \in \mathcal{H}$. To simplify notation, we write $Z_{1:n} := (Z_1, \ldots, Z_n)$, for $n \in \mathbb{N}$, and we let $P^n := P \times \cdots \times P$ be the product distribution over $\mathcal{Z}^n$. In what follows, it will be useful to define the following function class:

$$\ell \circ \mathcal{H} := \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}.$$

We adopt standard non-asymptotic big-oh notation; for functions $f, g : \mathbb{N} \to \mathbb{R}_{\geq 0}$, we write $f \leq O(g)$ if there exists some universal constant $C > 0$ such that $f(n) \leq Cg(n)$, for all $n \in \mathbb{N}$. We also write $f \leq \widetilde{O}(g)$, if there exists a $C > 0$ such that $f(n)$ is less than $Cg(n)$ up to a multiplicative poly-log-factor in n, for all $n \in \mathbb{N}$. We now present a series of standard definitions we require:

Definition 8.1 (Rademacher Complexity)**.** *Let $Z_1, \ldots, Z_n$ be i.i.d. random variables in some set $\mathcal{Z}$. The Rademacher complexity of a function class $\mathcal{F} \subseteq \mathscr{F}(\mathcal{Z}, \mathbb{R})$ is defined by*

$$\mathfrak{R}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i)\right],$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables (i.e. $\mathbf{P}[\sigma_i = \pm 1] = 1/2$, $i \in [n]$).

Definition 8.2 (Consistency)**.** *An algorithm that for each sample size $n \in \mathbb{N}$ and any unknown distribution $P \in \mathscr{P}(\mathcal{Z})$ outputs a hypothesis $\hat{h}_n \in \mathcal{H}$ based on i.i.d. samples*

$Z_1, \ldots, Z_n \sim P$, *is* consistent *if the risk* $\mathbb{E}_{P(Z)}[\ell(\hat{h}_n, Z)]$ *converges to* $\mathcal{H}$*-optimal risk* $\inf_{h \in \mathcal{H}} \left\{ \mathbb{E}_{P(Z)}[\ell(h, Z)] \right\}$ *in probability as* $n \to \infty$.

Definition 8.3 (Risk-Monotonicity)**.** *An algorithm that for each sample size* $n \in \mathbb{N}$ *and any unknown* $P \in \mathscr{P}(\mathcal{Z})$ *outputs a hypothesis* $\hat{h}_n \in \mathcal{H}$ *based on i.i.d. samples* $Z_1, \ldots, Z_n \sim P$*, is risk-monotonic if*

$$\mathbb{E}_{P^n(Z_{1:n})}[L(\hat{h}_n)] \leq \mathbb{E}_{P^{n-1}(Z_{1:n-1})}[L(\hat{h}_{n-1})], \quad \textit{for all } n \in \mathbb{N}. \tag{8.1}$$

The notion of monotonicity we just defined corresponds to the strongest notion of monotonicity considered by Loog et al. [2019]; Viering et al. [2019a], which they refer to as global $\mathcal{Z}$-monotonicity.

8.2 Risk-Monotonic Algorithms

In this section, we present risk-monotonic algorithms for a bounded loss ℓ. Our first algorithm requires the Rademacher complexity of the function class $\ell \circ \mathcal{H}$ be finite, while for the second algorithm we assume that the hypothesis class $\mathcal{H}$ is finite and show that risk-monotonicity is possible while achieving fast excess-risk rates under the Bernstein condition; in subsection 8.2.2, we discuss how the finiteness-of-$\mathcal{H}$ assumption may be removed.

The two algorithms we present in this section are instantiations of Algorithm 2 with different choices of input sequence (δ_k). Naturally, both algorithms share the same underlying idea; given a sample size n and the task of generating a hypothesis $\hat{h}_n = \hat{h}_n(Z_{1:n})$, the approach we take is to generate a sequence of hypotheses $\hat{h}_1 = \hat{h}_1(Z_1), \hat{h}_2 = \hat{h}_2(Z_{1:2}), \ldots,$ leading up to $\hat{h}_n$, where we only allow two consecutive hypotheses, say $\hat{h}_{k-1}$ and $\hat{h}_k$, to differ if we can guarantee that the risk associated with $\hat{h}_k$ is smaller than that of $\hat{h}_{k-1}$. Doing this ensures that the hypotheses $(\hat{h}_k)$ have non-increasing risk as a function of k. To test whether a hypothesis $\tilde{h}$ has a smaller risk than $\hat{h}$ given n sample, we use the fact that

$$\left| \widehat{L}_n(\tilde{h}) - L(\tilde{h}) \right| \leq \epsilon_n \quad \text{and} \quad \left| \widehat{L}_n(\hat{h}) - L(\hat{h}) \right| \leq \epsilon_n, \tag{8.2}$$

with high probability, for some error ϵ_n that can be obtained through a concentration argument of the empirical risk. Using (8.2), we can be sure (with high probability) that the population risk of $\tilde{h}$ is less than that of $\hat{h}$ if $\widehat{L}_n(\tilde{h}) - \widehat{L}_n(\hat{h}) \leq -2\epsilon_n$. Therefore, we will essentially set $\delta_n = 2\epsilon_n$, for all $n \in \mathbb{N}$, in Algorithm 2.

In this section, we will apply two different concentration arguments to the empirical risk, leading to two different "gap" sequences (δ_k), and thus two different algorithm variants. The first algorithm relies on a uniform convergence argument (which is why we require a finite Rademacher complexity of $\ell \circ \mathcal{H}$). We show that this algorithm is risk-monotonic with an excess-risk rate matching the optimal rate one would get without imposing further constraints on the loss or the data-generating distribution. The second (risk-monotonic) variant of our algorithm relies on the empirical

Bernstein inequality [Maurer and Pontil, 2009]. The resulting expression of the gap sequence (δ_k) allows us to show a fast convergence rate of the excess-risk under the Bernstein condition. We now present the first variant of our algorithm:

Algorithm 2 Greedy Empirical Risk Minimization (GERM).

Require:
Samples $Z_1, \ldots, Z_n$.
An arbitrary initial hypothesis $\hat{h}_0 \in \mathcal{H}$.
A sequence (δ_k).

1: **for** $k = 1, \ldots, n$ **do**
2: Set $\tilde{h}_k \in \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{k} \ell(h, Z_i)$. `// ERM computation`
3: **if** $\frac{1}{k}\sum_{i=1}^{k} \ell(\tilde{h}_k, Z_i) - \frac{1}{k}\sum_{i=1}^{k} \ell(\hat{h}_{k-1}, Z_i) \leq -\delta_k$ **then**
4: Set $\hat{h}_k = \tilde{h}_k$.
5: **else**
6: Set $\hat{h}_k = \hat{h}_{k-1}$.
7: Return $\hat{h}_n$.

8.2.1 Greedy Empirical Risk Minimization via Uniform Convergence

For the variant of Algorithm 2 we consider in this subsection, we require the following assumption on the Rademacher complexity of the function class $\ell \circ \mathcal{H}$:

Assumption 8.1. The function class $\ell \circ \mathcal{H}$ satisfies $\mathfrak{R}_k(\ell \circ \mathcal{H}) < +\infty$, for all $k \in \mathbb{N}$.

Furthermore, for our algorithm to be consistent, we will also need that $\mathfrak{R}_k(\ell \circ \mathcal{H}) \to 0$ as $k \to \infty$. We will instantiate Algorithm 2 with the sequence (δ_k) given by:

$$\delta_k := 4\bar{\mathfrak{R}}_k + \sqrt{2\ln(2k)/k} + 2/k, \quad \text{for all } k \in \mathbb{N}, \tag{8.3}$$

where $(\bar{\mathfrak{R}}_k)$ are high-probability upper bounds on the Rademacher complexities $(\mathfrak{R}_k(\ell \circ \mathcal{H}))$; in particular, we require $(\bar{\mathfrak{R}}_k)$ to satisfy

$$\mathbf{P}\left[\bar{\mathfrak{R}}_k \geq \mathfrak{R}_k(\ell \circ \mathcal{H})\right] \geq 1 - 1/k, \quad \text{for all } k \in \mathbb{N}. \tag{8.4}$$

A candidate sequence $(\bar{\mathfrak{R}}_k)$ that satisfies (8.4), and which can be evaluated using samples, is given by

$$\bar{\mathfrak{R}}_k = 0 \vee \sup_{h \in \mathcal{H}} \left\{ \frac{1}{k}\sum_{i=1}^{k} \sigma_i \cdot \ell(h, Z_i) + \sqrt{\frac{2\ln(2k)}{k}} \right\}, \quad \text{for } k \in \mathbb{N}, \tag{8.5}$$

where $\sigma_1, \ldots, \sigma_k \in \{-1, +1\}$ are i.i.d. Rademacher random variables. The fact that the choice of $(\bar{\mathfrak{R}}_k)$ in (8.5) satisfies (8.4) follows directly from the following proposition

whose proof (see Section 8.6) is based on a standard application of McDiarmid's inequality:

Proposition 8.1. *For i.i.d. random variables $Z_{1:n} \in \mathcal{Z}$ [resp. Rademacher variables $\sigma_{1:n}$], we have,*

$$\forall \delta \in (0,1), \quad \mathbf{P}\left[\left|\mathfrak{R}_n(\ell \circ \mathcal{H}) - \sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \cdot \ell(h, Z_i)\right| \leq \sqrt{\frac{2\ln(2/\delta)}{n}}\right] \geq 1 - \delta.$$

We now present the first guarantees of GERM (the proofs are postponed to Section 8.3).

Theorem 8.2. *Algorithm 2 with (δ_k) as in* (8.3) *and $(\bar{\mathfrak{R}}_k)$ satisfying* (8.4) *is risk-monotonic according to Definition 8.3.*

As discussed above, a candidate choice for $\bar{\mathfrak{R}}_k$ that satisfies the condition of Theorem 8.2 is given in (8.5). We remark, however, that $(\bar{\mathfrak{R}}_k)$ as selected in (8.5) requires the optimization of an empirical objective over $\mathcal{H}$[2], and may be replaced by a more practical choice, given (8.4) is satisfied. For example, in some settings, a deterministic upper bound on $\mathfrak{R}_k(\ell \circ \mathcal{H})$ trivially satisfies (8.4) and if available, is another viable choice for $\bar{\mathfrak{R}}_k$. For example, in classification with the 0-1 loss, we have, by Sauer's Lemma (see e.g. [Bousquet et al., 2004, Lemma 1 and Page 198]), $\mathfrak{R}_k(\ell \circ \mathcal{H}) \leq 2\sqrt{\mathrm{VC}(\mathcal{H}) \cdot \ln(ek)/k}$, for $k \geq 1$. And so, any available bound on the VC-dimension of $\mathcal{H}$ can be used to bound the Rademacher complexity. As another example, consider a 1-Lipschitz loss and a hypothesis class consisting of L-layer feed-forward neural network, $L \geq 1$, with 1-Lipschitz activation and weight matrices with Frobenius norm bounded by B. In this case, the Rademacher complexity is bounded as $\mathfrak{R}_k \leq r\sqrt{2}(2B)^L/\sqrt{k}$, for all $k \geq 1$, where $r > 0$ is the maximum norm of the observed samples [Neyshabur et al., 2015].

We now show that our algorithm is also consistent under appropriate conditions:

Theorem 8.3. *Let $\hat{h}_n$ be the output of Algorithm 2 with (δ_k) as in* (8.3) *and $(\bar{\mathfrak{R}}_k)$ satisfying* (8.4)*. Then, for all n, with probability at least $1 - 2/n$,*

$$\mathbb{E}_{P(Z)}[\ell(\hat{h}_n, Z)] \leq \inf_{h \in \mathcal{H}} \left\{\mathbb{E}_{P(Z)}[\ell(h, Z)]\right\} + 12\bar{\mathfrak{R}}_n + 3\sqrt{\frac{2\ln(2n)}{n}} + \frac{2}{n}. \tag{8.6}$$

Theorem 8.3 implies that Algorithm 2, with the choices of (δ_k) and $(\bar{\mathfrak{R}}_k)$ as in the theorem statement, is consistent whenever the sequence $(\bar{\mathfrak{R}}_k)$ satisfies $\lim_{k\to\infty} \bar{\mathfrak{R}}_k = 0$ in probability. We also note that if $\bar{\mathfrak{R}}_k \leq O(\mathfrak{R}_k(\ell \circ \mathcal{H}) + (\ln(k)/k)^{1/2})$ with high probability (this is the case for $(\bar{\mathfrak{R}}_k)$ as in (8.5) by Proposition 8.1), then the rate achieved in (8.6) matches (up to log-factors) the optimal excess-risk rate[3] when no

[2]The objective in (8.5) can be non-convex even for a convex loss ℓ, and so its optimization can be NP-hard.

[3]Technically, the excess-risk lower bound presented in e.g. [Boucheron et al., 2005] is expressed in terms of the VC-dimension. Bounding the Rademacher complexity in terms of the VC-dimension using Sauer's Lemma, see e.g. [Bousquet et al., 2004, Lemma 1 and Page 198], establishes the optimality of the excess-risk rate we present.

additional assumptions are made about the loss or the data-generating distribution [Dudley, 1984; Talagrand, 1994; Boucheron et al., 2005; Zhivotovskiy and Hanneke, 2018]. We formalize this in the next corollary by presenting the excess-risk rate of Algorithm 2 with the particular choice of sequence $(\tilde{\mathfrak{R}}_k)$ in (8.5):

Corollary 8.4. *Algorithm 2 with* (δ_k) *as in* (8.3) *and* $(\tilde{\mathfrak{R}}_k)$ *as in* (8.5) *is risk-monotonic, and its output* $\hat{h}_n$ *satisfies, for all* n*, with probability at least* $1 - 2/n$*,*

$$\mathbb{E}_{P(Z)}[\ell(\hat{h}_n, Z)] \leq \inf_{h \in \mathcal{H}} \left\{ \mathbb{E}_{P(Z)}[\ell(h, Z)] \right\} + 12\mathfrak{R}_n(\ell \circ \mathcal{H}) + 4\sqrt{\frac{2\ln(2n)}{n}} + \frac{2}{n}.$$

The corollary follows directly from Theorems 8.2 and 8.3, and Proposition 8.1.

Remark 8.1. *The risk-monotonicity of Algorithm 2 in both variants we consider in this subsection and the next is ensured by the case distinction in Line 3. Therefore, the* ERM *hypothesis* $\tilde{h}_k$ *in Line 2 of the algorithm may be replaced by* any *other consistent hypothesis. The resulting algorithm will still be risk-monotonic and consistent. The corresponding proofs are easily extended to this case.*

Next, we present the second variant of our algorithm that relies on the empirical Bernstein inequality instead of uniform convergence to select the gap sequence (δ_k). Doing this allows for a better convergence rate of the excess-risk under the Bernstein condition.

8.2.2 Greedy Empirical Risk Minimization via Empirical Bernstein

In this subsection, we assume that the hypothesis set $|\mathcal{H}|$ is finite and instantiate Algorithm 2 with the sequence (δ_k) defined by (to see how this choice relates to the empirical Bernstein inequality see the proof of Lemma 8.16 in Section 8.9)

$$\delta_k := \sqrt{\frac{2\sum_{i=1}^{k}(\ell(\tilde{h}_k, Z_i) - \ell(\hat{h}_{k-1}, Z_i))^2 \ln(2k|\mathcal{H}|^2)}{(k-1)^2}} + \frac{5\ln(2k|\mathcal{H}|^2)}{k-1} + \frac{2}{k}, \tag{8.7}$$

$k \in \mathbb{N}$, where $(\tilde{h}_k)$ and $(\hat{h}_k)$ are as in Algorithm 2. One way to dispose of the finiteness assumption of the hypothesis class $\mathcal{H}$ is to consider a randomized hypothesis, instead of the ERM in Line 3 of Algorithm 2, and apply existing PAC-Bayesian bounds to determine the appropriate gap sequence (δ_k) (the PAC-Bayesian bounds presented in e.g. [Mhammedi et al., 2019a] lead to a gap sequence that is compatible with the analysis we carry for the results of this subsection). As noted in Remark 8.1, swapping the ERM for another predictor (randomized or not) need not affect risk-monotonicity[4]. We remark that one can also reduce the non-finite case to the finite one using empirical covers of $\mathcal{H}$ [Audibert, 2004a]. We will focus on the setting where $\mathcal{H}$ is finite in favor of exposition and highlighting the fast excess-risk rates under the Bernstein condition while maintaining risk-monotonicity.

[4]In the case of randomized predictors, the definition of risk-monotonicity in (8.1) and Line 3 of Algorithm 2 need be modified to incorporate expectations over the randomness of the predictors.

The choice of sequence (δ_k) in (8.7) will allow us to show a fast rate of convergence of the excess-risk (faster than the one obtained in Corollary 8.4) under the following Bernstein condition:

Definition 8.4. [Bernstein Condition] *The learning problem $(P, \ell, \mathcal{H})$ satisfies the (β, B)-Bernstein condition, for $\beta \in [0,1]$ and $B > 0$, if for all $h \in \mathcal{H}$,*

$$\mathbb{E}_{P(Z)}\left[(\ell(h,Z) - \ell(h_\star, Z))^2\right] \leq B \cdot \mathbb{E}_{P(Z)}\left[\ell(h,Z) - \ell(h_\star, Z)\right]^\beta,$$

where $h_* \in \arg\inf_{h\in\mathcal{H}} \mathbb{E}_{P(Z)}[\ell(h,Z)]$ is a risk minimizer within the closure of $\mathcal{H}$.

The Bernstein condition [Audibert, 2004b; Bartlett et al., 2006; Bartlett and Mendelson, 2006b; Erven et al., 2015; Koolen et al., 2016] essentially characterizes the easiness of the learning problem. In particular, it implies that the variance of the excess-loss random variable $\ell(h,Z) - \ell(h_\star, Z)$ vanishes when the risk associated with the hypothesis $h \in \mathcal{H}$ gets closer to the $\mathcal{H}$-optimal risk $L(h_\star)$. For bounded loss functions, the Bernstein condition with $\beta = 0$ always holds. The Bernstein condition with $\beta = 1$—also known as the *Massart noise condition* [Massart and Nédélec, 2006]—corresponds to the easiest learning setting. It holds, for example, when $\mathcal{H}$ is convex and $h \mapsto \ell(h,z)$ is exp-concave, for all $z \in \mathcal{Z}$ [Erven et al., 2015; Mehta, 2017] (for more examples of learning settings where a Bernstein condition holds see [Koolen et al., 2016, Section 3]). The case where $\beta \in (0,1)$ interpolates naturally between these two extreme cases, where intermediate excess-risk rates are achievable. We start by the statement of risk-monotonicity (the proof is postponed to Section 8.9):

Theorem 8.5. *Algorithm 2 with (δ_k) as in* (8.7) *is risk-monotonic according to Definition 8.3.*

The proof of the theorem is the same as that of Theorem 8.2 except for minor changes. Note that the algorithm does not require any Bernstein condition to ensure risk-monotonicity. We only use the condition in the next theorem to show intermediate excess-risk rates:

Theorem 8.6. *Let $B > 0$ and $\beta \in [0,1]$, and suppose that the (β, B)-Bernstein condition holds. Then, the output $\hat{h}_n$ of Algorithm 2 with (δ_k) as in* (8.7) *satisfies, $\forall n$, with probability at least $1 - 4/n$,*

$$\mathbb{E}_{P(Z)}[\ell(\hat{h}_n, Z)] \leq \inf_{h\in\mathcal{H}}\left\{\mathbb{E}_{P(Z)}[\ell(h,Z)]\right\} + O\left(\left(\frac{\ln(n|\mathcal{H}|)}{n}\right)^{\frac{1}{2-\beta}} + \frac{\ln(n|\mathcal{H}|)}{n}\right).$$

We note that the excess-risk rate achieved in Theorem 8.6 interpolates nicely between the fast $\widetilde{O}((\ln|\mathcal{H}|)/n)$ rate achieved when $\beta = 1$ (the easiest learning setting), and the worst-case rate $\widetilde{O}(\sqrt{(\ln|\mathcal{H}|)/n})$ achieved when $\beta = 0$ (this always holds for a bounded loss). Crucially, the bound in Theorem 8.6 is, in general, the best one can hope for up to log-factors (see e.g. [Koolen et al., 2016]). The results of Theorems 8.3 and 8.6 essentially show that, as far as excess-risk rates are concerned, risk-monotonicity can be achieved for free.

Before moving to the notion of risk-monotonicity with high probability in Section 8.4, we first prove one of the main results of the current section. We chose the proof of Theorem 8.2 as we find it most helpful in understanding how GERM achieves risk-monotonicity. The rest of the proofs are deferred to Sections 8.6, 8.7, 8.9, and 8.10.

8.3 Proof of Theorem 8.2

In the proofs of Theorems 8.2 and 8.3, we need the following standard uniform convergence result:

Theorem 8.7 (Bousquet et al. [2004]). *Let $n \in \mathbb{N}$ and $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables such that $Z \sim P$. Further, let $\mathcal{F} \subseteq \mathscr{F}(\mathcal{Z}, [0,1])$ be a class of functions bounded between 0 and 1. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{P(Z)}[f(Z)] - \mathbb{E}_{\widehat{P}_n(Z)}[f(Z)] \right| \leq 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

From this result and a union bound, we arrive at the following useful corollary:

Corollary 8.8. *Let $n \in \mathbb{N}$ and $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables such that $Z \sim P$. Further, let $\bar{\mathfrak{R}}_n > 0$ be such that $\mathbf{P}[\bar{\mathfrak{R}}_n \geq \mathfrak{R}_n(\ell \circ \mathcal{H})] \geq 1 - 1/n$, and define the event*

$$\mathcal{E}_n := \left\{ \left| L(h) - \widehat{L}_n(h) \right| \leq \epsilon_n, \text{ for all } h \in \mathcal{H} \right\}, \quad \text{where} \quad \epsilon_n := 2\bar{\mathfrak{R}}_n + \sqrt{\frac{\ln(2n)}{2n}}.$$

Then, $\mathbf{P}[\mathcal{E}_n] \geq 1 - 2/n$.

Proof of Theorem 8.2. First, note that by linearity of the expectation it suffices to show that

$$\mathbb{E}_{P^n(Z_{1:n})} \left[L(\hat{h}_n) - L(\hat{h}_{n-1}) \right] \leq 0.$$

Moving forward, we define $\Delta_n := L(\hat{h}_n) - L(\hat{h}_{n-1})$. Let ϵ_n and $\mathcal{E}_n$ be as in Corollary 8.8 with $\bar{\mathfrak{R}}_n$ being the high probability upper bound on $\mathfrak{R}_n(\ell \circ \mathcal{H})$ that Algorithm 2 has access to ($\bar{\mathfrak{R}}_n$ satisfies (8.4)). By Corollary 8.8, we have $\mathbf{P}[\mathcal{E}_n] \geq 1 - 2/n$. Now, by the law of the total expectation, we have

$$\begin{aligned} \mathbb{E}[\Delta_n] &= \mathbf{P}[\{\hat{h}_n \equiv \hat{h}_{n-1}\}] \cdot \mathbb{E}[\Delta_n \mid \{\hat{h}_n \equiv \hat{h}_{n-1}\}] \\ &\quad + \mathbf{P}[\{\hat{h}_n \not\equiv \hat{h}_{n-1}\}] \cdot \mathbb{E}[\Delta_n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\}], \\ &\leq \mathbf{P}[\{\hat{h}_n \equiv \hat{h}_{n-1}\}] \cdot \mathbb{E}[\Delta_n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\}], \end{aligned} \tag{8.8}$$

where the last inequality follows by the fact that if $\hat{h}_n \equiv \hat{h}_{n-1}$, then $\Delta_n = 0$. We note that the expectation in (8.8) is with respect to $P^n(Z_{1:n})$. Letting $p_n := \mathbf{P}[\{\hat{h}_n \equiv$

$\hat{h}_{n-1}\}] \in [0,1]$, we have from (8.8),

$$\begin{aligned}\mathbb{E}[\Delta_n] &\leq p_n \mathbf{P}[\mathcal{E}_n] \cdot \mathbb{E}[\Delta_n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\} \cap \mathcal{E}_n] + p_n \mathbf{P}[\mathcal{E}_n^c] \cdot \mathbb{E}[\Delta_n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\} \cap \mathcal{E}_n^c], \\ &\leq p_n \mathbb{E}[\Delta_n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\} \cap \mathcal{E}_n] + 2p_n/n, \qquad (8.9) \\ &= p_n \mathbb{E}[\Delta_n + 2/n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\} \cap \mathcal{E}_n], \qquad (8.10)\end{aligned}$$

where (8.9) follows by the fact that the loss ℓ takes values in $[0,1]$ and that $\mathbf{P}[\mathcal{E}_n^c] \leq 2/n$. Now, if $\hat{h}_n \not\equiv \hat{h}_{n-1}$, then by Line 3 of Algorithm 2, we have

$$\widehat{L}_n(\tilde{h}_n) \leq \widehat{L}_n(\hat{h}_{n-1}) - \delta_n, \qquad (8.11)$$

where we recall that δ_n is as in (8.3) (with $k = n$). Under the event $\mathcal{E}_n$, we have

$$L(\tilde{h}_n) - \epsilon_n \leq \widehat{L}_n(\tilde{h}_n) \quad \text{and} \quad \widehat{L}_n(\hat{h}_{n-1}) \leq L(\hat{h}_{n-1}) + \epsilon_n.$$

This, in combination with (8.11), implies that under the event $\mathcal{E}_n \cap \{\hat{h}_n \not\equiv \hat{h}_{n-1}\}$,

$$\Delta_n + 2/n = L(\tilde{h}_n) - L(\hat{h}_{n-1}) + 2/n \leq 2/n - \delta_n + 2\epsilon_n = 0,$$

where in the last equality we substituted the expression of δ_n in (8.3) (with $k = n$). As a result, we have

$$\mathbb{E}_{P^n(Z_{1:n})}[\Delta_n + 2/n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\} \cap \mathcal{E}_n] \leq 0. \qquad (8.12)$$

Combining (8.10) and (8.12) yields the desired result. □

8.4 Risk-Monotonicity in High Probability

So far, we have addressed the problem of risk monotonicity in expectation as posed by Viering et al. [2019a]. Another interesting yet realistic scenario [Viering et al., 2019b] is to study risk monotonicity for every realization of $(Z_1, Z_2, \dots)$ with high probability (instead of in-expectation).

Definition 8.5 (Strong Risk-Monotonicity)**.** *Let $\delta \in (0,1)$, and $Z, Z_1, Z_2, \dots$ be i.i.d. random variables such $Z \sim P$ for some unknown distribution P. An algorithm that for each $n \in \mathbb{N}$ outputs a hypothesis $\hat{h}_n \in \mathcal{H}$ based on $(Z_1, \dots, Z_n)$, is δ-strongly-risk-monotonic if there exists $n_0 \geq 1$*

$$\mathbf{P}\left[\forall n \geq n_0,\ \ L(\hat{h}_n) \leq L(\hat{h}_{n-1})\right] \geq 1 - \delta. \qquad (8.13)$$

The notion of strong risk monotonicity concerns individual risk curves; the curves of $\mathbb{E}[\ell(\hat{h}_n, Z)]$ (instead of $\mathbb{E}_{P^n(Z_{1:n})}\mathbb{E}[\ell(\hat{h}_n, Z)]$) as a function of n, for a given realization of $(Z_1, Z_2, \dots)$. As we will discuss below, a strongly risk-monotonic algorithm can easily be turned into a risk-monotonic algorithm as per Definition 8.3, albeit for $n \geq n_0$ in (8.1) instead of $n \in \mathbb{N}$.

We note that compared with the notion of risk monotonicity in Definition 8.3, strong risk-monotonicity allows for an $n_0 \geq 1$ beyond which monotonicity is achieved with high probability. The magnitude of n_0 will in general depend on the complexity of the class $\mathcal{H}$. Here, we make the simplifying assumption that the set $|\mathcal{H}|$ is finite, (as we did in Section 8.2.2) and we will show that (8.13) is achievable for n_0 as small as $\max\{n : 2\ln(\sqrt{n}|\mathcal{H}|^2/\delta) \geq n/2\}$ (this is essentially the smallest sample size for which the new concentration inequality we use is non-vacuous).

The notion δ-strong risk-monotonicity fits naturally with the following notion of consistency:

Definition 8.6 (δ-Consistency)**.** *Let $\delta \in (0,1)$, and $Z, Z_1, Z_2, \ldots$ be i.i.d. random variables such $Z \sim P$ for some unknown distribution P. An algorithm that for each $n \in \mathbb{N}$ outputs a hypothesis $\hat{h}_n \in \mathcal{H}$ based on $(Z_1, \ldots, Z_n)$, is* δ-consistent *if $\lim_{n\to\infty} \mathbb{E}_{P(Z)}[\ell(\hat{h}_n, Z)] = \inf_{h\in\mathcal{H}} \left\{\mathbb{E}_{P(Z)}[\ell(h, Z)]\right\}$ with probability at least $1-\delta$.*

For risk-monotonicity in expectation, we derived the performance gaps (δ_k) based on concentration inequalities (Rademacher and empirical Bernstein) to enforce $\mathbb{E}[L(\hat{h}_n)] \leq \mathbb{E}[L(\hat{h}_{n-1})]$. However, these inequalities only hold for a pre-specified sample size n and are, therefore, not strong enough to guarantee monotonicity for individual risk curves. To overcome this challenge, we present a new, so-called time-uniform, concentration inequality that holds for all sample sizes simultaneously. In the statement of this result, we will denote by $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}], t \geq 1$, where $\mathcal{F}_{t-1}$ is the sigma-algebra generated by random variables $X_1, \ldots, X_{t-1}$, with $\mathcal{F}_0 := \emptyset$.

Theorem 8.9. *Let $\gamma > 0$, $\delta \in (0,1)$, and $X, X_1, X_2, \cdots \in [0,1]$ be random variables such that $\mathbb{E}[X] = \mathbb{E}_{i-1}[X_i]$, $\forall i \geq 1$. Then, for $n_0 \geq \sup\{n : 2\ln\frac{\sqrt{n}}{\gamma\delta} \geq n\}$:*

$$\mathbf{P}\left[\forall n > n_0,\ \left|\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} X_i\right| \leq \left(\sqrt{(\gamma^2/n + \widehat{V}_n[X])\cdot\xi_n} + \gamma\xi_n\right)\frac{1}{1-\xi_n}\right] \geq 1-\delta,$$

$$\textit{where} \quad \widehat{V}_n[X] := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2 \quad \textit{and} \quad \xi_n := \frac{2\ln\frac{\sqrt{n}}{\gamma\delta}}{n}.$$

We note that the theorem holds for random variables that are not necessarily independent, or even identically distributed. Thus, the bound in Theorem 8.9 can be seen as an empirical (time-uniform) version of Freedman's inequality [Freedman, 1975], which is of independent interest[5]. Moreover, note that the empirical variance appears inside the square-root, enabling us to achieve fast rates under the Bernstein condition in a similar way as in Section 8.2.2. We prove Theorem 8.9 by constructing a new non-negative supermartingale based of the potential function of the recent FreeGrad algorithm [Mhammedi and Koolen, 2020] for online convex optimization (more on this in Section 8.7). Our proof technique is very similar to the one introduced in [Jun

[5] A concentration inequality reminiscent of the one in Theorem 8.9 appeared in Howard et al. [2018]. However, their bound has a V_n under the square-root that is not equal to the empirical variance.

and Orabona, 2019b]. The only difference is that we use the specific shape of FreeGrad's potential function to build our supermartingale. Finally, we note that the $\ln n$ factors present in our new concentration inequality (thanks the ξ_n terms) can be improved to $\ln\ln n$ by carefully 'mixing' the FreeGrad supermartingal over different values of $\gamma > 0$ (see also discussion in [Jun and Orabona, 2019b, Section 7.2]).

Choice of sequence (δ_k). For $\delta \in (0,1)$ and $k \geq 1$, we define

$$\xi_k := \frac{2}{k} \ln\left(\frac{|\mathcal{H}|^2\sqrt{k}}{\delta}\right), \quad \text{and} \quad n_0 := \max\{n : \xi_n \geq 1/2\}. \tag{8.14}$$

With this, and $(\tilde{h}_k)$ and $(\hat{h}_{k-1})$ as in Algorithm 2, we let

$$\delta_k := \begin{cases} 4\xi_k + 2\sqrt{\sum_{i=1}^k (\ell(\tilde{h}_k, Z_i) - \ell(\hat{h}_{k-1}, Z_i))^2 \cdot \xi_k/k}, & \text{if } k > n_0; \\ 0, & \text{otherwise.} \end{cases} \tag{8.15}$$

This choice of δ_k is related to the new concentration inequality in Theorem 8.9, which we show explicitly in the proof of Lemma 8.19 in Section 8.10. As we shall see shortly, this choice of δ_k will not only allow us to ensure monotonicity of individual risk curves but will also enable us to achieve fast rates under the Bernstein condition (Definition 8.4). We now present the first guarantee of our Algorithm 2 with the choice of (δ_k) as in (8.15):

Theorem 8.10. *For any $\delta \in (0,1)$, Algorithm 2 with (δ_k) as in* (8.15) *is δ-strongly risk-monotonic according to Definition 8.5.*

By modifying the proof of Theorem 8.2 slightly, it is easy to show that by setting $\delta = 1/k$ in (8.14) and adding $1/k$ in the definition of δ_k in (8.15), the resulting Algorithm 2 is risk-monotonic according to Definition 8.3, albeit only for $n > \max\{k : 2\ln(|\mathcal{H}|^2 k^{3/2}) \geq k\}$, instead of all $n \in \mathbb{N}$. This shows that, indeed, the notion of risk-monotonicity we introduced in Definition 8.6, is in a way stronger than the one we considered in the previous section. We now bound the excess risk rate of the new algorithm:

Theorem 8.11. *Let $\delta \in (0,1)$ and n_0 be as in* (8.14)*. Further, let $B > 0$ and $\beta \in [0,1]$, and suppose that the (β, B)-Bernstein condition holds. Then, the output $\hat{h}_n$ of Algorithm 2 with (δ_k) as in* (8.15) *satisfies, with probability at least $1-\delta$, for all $n \geq n_0$,*

$$\mathbb{E}_{P(Z)}[\ell(\hat{h}_n, Z)] \leq \inf_{h \in \mathcal{H}} \left\{\mathbb{E}_{P(Z)}[\ell(h, Z)]\right\} + O\left(\max_{\gamma \in \{\beta, 1\}} \left(\frac{\ln(n|\mathcal{H}|/\delta)}{n}\right)^{\frac{1}{2-\gamma}}\right). \tag{8.16}$$

It is striking that one can achieve strong risk-monotonicity without compromising fast excess risk rates under the Bernstein condition. We note that the Bernstein condition (for $\beta > 0$) is not needed for the instantiation of Algorithm 2 in Theorem

8.11 to be δ-consistent (Definition 8.6). Indeed, since ℓ is bounded, the Bernstein condition is satisfied for $\beta = 0$. Thus, substituting this into (8.16) and taking the limit as $n \to \infty$ implies the required condition for δ-consistency.

In the next section, we present an efficient version of GERM for the case of convex losses.

8.5 An Efficient Algorithm for Convex Losses

In this subsection, we present an efficient risk-monotonic algorithm for the case where the function $h \mapsto \ell(h, z)$ is convex, for all $z \in \mathcal{Z}$. In contrast with GERM (Algorithm 2), the algorithm we present here (Algorithm 3) does not require n-ERM computations.

For simplicity of the exposition, we assume in the subsection that $\mathcal{H}$ in a convex bounded subset of $\mathbb{R}^d$, albeit much of the techniques we present here also work if $\mathcal{H}$ is an unbounded subset of a Banach space. We will denote by $D := \sup_{h,h' \in \mathcal{H}} \|h - h'\|$ the *diameter* of the set $\mathcal{H}$, where $\|\cdot\|$ represents the Euclidean norm. Moving forward, we require the notions of *sub-differential* set and *sub-gradients*; for any $z \in \mathcal{Z}$, the sub-differential of a function $h \mapsto \ell(h, z)$ at $\tilde{h} \in \mathcal{H}$ is defined by

$$\partial_h \ell(\tilde{h}, z) := \left\{ g \in \mathbb{R}^d : \ell(h, z) \geq \ell(\tilde{h}, z) + g^\top (h - \tilde{h}), \forall h \in \mathcal{H} \right\}. \tag{8.17}$$

Any element $g \in \partial_h \ell(\tilde{h}, z)$ is a sub-gradient of $h \mapsto \ell(h, z)$ at $\tilde{h}$. We assume that the loss ℓ is G-Lipschitz in the first argument; that is

$$\|g\| \leq G, \quad \text{for all } g \in \partial_h \ell(\tilde{h}, z).$$

Let $\mathbf{A}$ be an online learning algorithm which operates in rounds; at each round i, $\mathbf{A}$ outputs $\tilde{h}_i \in \mathbb{R}^d$ then receives a vector $g_i \in \mathbb{R}^d$. We assume that for any sequence $(g_i) \subset \mathbb{R}^d$ such that $\forall i, \|g_i\| \leq G$, and some $G > 0$, the outputs $(\tilde{h}_i)$ of $\mathbf{A}$ satisfy

$$\sum_{i=1}^{n} g_i^\top (\tilde{h}_i - h) \leq 2GD\sqrt{n}, \quad \text{for all } h \in \mathcal{H}, \tag{8.18}$$

The LHS of (8.18) is the *regret* of Algorithm $\mathbf{A}$ against the *comparator* $h \in \mathcal{H}$. An example of algorithm $\mathbf{A}$ which satisfies the regret bound in (8.18) is the Online Gradient Descent (OGD) algorithm (see e.g. Hazan [2016b]). Using the proof steps of Theorems 8.3 and (8.1), and standard online-to-batch conversion techniques (see e.g. Cutkosky [2019a]), we get the following result (which we state without proof):

Claim 8.12. *Let $\mathcal{H} \subset \mathbb{R}^d$ be a bounded convex set with diameter $D > 0$, and suppose that ℓ is convex and G-Lipschitz in the first argument. Then, given*

- *a sequence $(\bar{\mathfrak{R}}_k) \subset \mathbb{R}_{\geq 0}$ satisfying $\mathbf{P}[\bar{\mathfrak{R}}_k \geq \mathfrak{R}_k(\ell \circ \mathcal{H})] \geq 1 - 1/k$, for all $k \in \mathbb{N}$, and*
- *an algorithm $\mathbf{A}$ with a regret bound as in* (8.18),

Algorithm 3 is risk-monotonic and its output $\hat{h}_n$ satisfies $\forall n \geq 1$, with probability at least $1 - 2/n$,

$$\mathbb{E}_{P(Z)}[\ell(\hat{h}_n, Z)] \leq \inf_{h \in \mathcal{H}} \left\{ \mathbb{E}_{P(Z)}[\ell(h, Z)] \right\} + 12\bar{\mathfrak{R}}_n + 3\sqrt{\frac{2\ln(2n)}{n}} + \frac{2}{n}. \tag{8.19}$$

Algorithm 3 An efficient risk-monotone algorithm for convex losses.

Require:
Samples $Z_1, \ldots, Z_n$.
An arbitrary initial hypothesis $\hat{h}_0$ in $\mathcal{H}$
An online learning algorithm A operating on $\mathcal{H}$.
A sequence $(\bar{\mathfrak{R}}_n)$.

Set $\breve{h}_0 = \tilde{h}_0 = \hat{h}_0$
Set $S_0 = 0$
Send 0 to A
for $k = 1, \ldots, n$ **do**
  Get $\breve{h}_k$ from A
  Set $\tilde{h}_k = ((k-1)\tilde{h}_{k-1} + \breve{h}_k)/k$.
  Send $g_k \in \partial_h \ell(\tilde{h}_k, Z_k)$ to A.
  Set $\delta_k = 2\bar{\mathfrak{R}}_k + \frac{3}{2}\sqrt{2\ln(2k)/k} + 3/k$.
  if $\frac{1}{k}\sum_{i=1}^{k} \ell(\tilde{h}_k, Z_i) - \frac{1}{k}\sum_{i=1}^{k} \ell(\hat{h}_{k-1}, Z_i) \leq -\delta_k$ **then**
    Set $\hat{h}_k = \tilde{h}_k$
  else
    Set $\hat{h}_k = \hat{h}_{k-1}$
Return $\hat{h}_n$.

As we did in Corollary 8.4 in the previous subsection, if we instantiate Algorithm 3 with $(\bar{\mathfrak{R}}_k)$ as in (8.5), we can essentially replace the term $\bar{\mathfrak{R}}_n$ on the RHS of (8.19) by $\mathfrak{R}_n(\ell \circ \mathcal{H})$; in this case, Algorithm 3 achieves the standard Rademacher risk rate when no additional assumptions on the learning problem are made.

The next sections are dedicated to the proofs of the results of this chapter. We start by the proof of Proposition 8.1.

8.6 Estimating the Rademacher Complexity

In this section, we derive an estimator of the Rademacher complexity of a class $\mathcal{F}$ consisting of functions taking values in the interval $[0, 1]$. We will be using McDiarmid's inequality:

Theorem 8.13 (McDiarmid [1989]). *Let $c > 0$ and $X_1, \ldots, X_n$ be independent random*

variables taking values in a set A, and assume that $f : \mathcal{A}^n \to \mathbb{R}$ satisfies,

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{A}} \left| f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n) \right| \leq c, \quad \forall i \in [n]. \tag{8.20}$$

Then, for every $t > 0$, $\mathbf{P}[|f(X_1,\ldots,X_n) - \mathbb{E}[f(X_1,\ldots,X_n)]| \geq t] \leq 2\exp(-2t^2/(nc^2))$.

Lemma 8.14. *For $\mathcal{A} := \{\pm 1\} \times \mathcal{Z}$, the function $f : \mathcal{A}^n \to \mathbb{R}$ defined by*

$$f((\sigma_1,z_1),\ldots,(\sigma_n,z_n)) := \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i \cdot \ell(h,z_i),$$

satisfies (8.20) *with $c = 2/n$ and $x_i := (\sigma_i, z_i)$, $i \in [n]$.*

Proof. For any $i \in [n]$ and $x_1 = (\sigma_1,z_1),\ldots,x_n = (\sigma_n,z_n), x_i' = (\sigma_i',z_i') \in \mathcal{A}$, we have

$$\begin{aligned}\left| f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n) \right| &\leq \sup_{h\in\mathcal{H}} \frac{1}{n} \left| \sigma_i \cdot \ell(h,z_i) - \sigma_i' \cdot \ell(h,z_i') \right|, \\ &\leq \frac{2}{n},\end{aligned}$$

where the last inequality follows by the fact that the loss ℓ takes values in $[0,1]$ and $\sigma_i \in \{\pm 1\}$, for all $i \in [n]$. This completes the proof. □

Proof of Proposition 8.1. By definition of the Rademacher inequality, we have

$$\mathfrak{R}_n(\ell \circ \mathcal{H}) = \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i \cdot \ell(h,Z_i)\right]. \tag{8.21}$$

Thus, the desired inequality follow directly by McDiarmid's inequality (Theorem 8.13) and Lemma 8.14. □

8.7 Proof of the New Concentration Bound (Theorem 8.9)

To prove Theorem 8.9, we will construct a non-negative supermartingale with the help of the recent FreeGrad algorithm [Mhammedi and Koolen, 2020]. Our proof technique is very similar to the one introduced in [Jun and Orabona, 2019b], except that we use the specific shape of FreeGrad's potential function to build our supermartingale.

For $\gamma > 0$, $n \in \mathbb{N}$, and random variables $Y, Y_1, Y_2, \ldots$, we define

$$M_n := \frac{\gamma^2}{\sqrt{V_n}} \cdot \exp\left(\frac{|S_n|^2}{2V_n + 2|S_n|}\right), \quad \text{where} \quad \begin{cases} S_n := \sum_{i=1}^n Y_i; \\ V_n := \gamma^2 + \sum_{i=1}^n Y_i^2. \end{cases} \tag{8.22}$$

In what follows, we will denote by $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}], t \geq 1$, where $\mathcal{F}_{t-1}$ is the sigma-algebra generated by the random variables $Y_1,\ldots,Y_{t-1}$.

Proposition 8.15. *For $\gamma > 0$, and any random variables $Y, Y_1, Y_2, \dots \in [-1, 1]$ satisfying $\mathbb{E}_{i-1}[Y_i] = \mathbb{E}[Y] = 0$, for all $i \in [n]$, the sequence (M_n) defined in* (8.22) *is a non-negative supermartingale; that is, $M_n \geq 0$, for all $n \geq 1$, and*

$$\mathbb{E}_n[M_{n+1}] \leq M_n, \quad \textit{for all } n \geq 1.$$

Before we present the proof of this theorem, we first give a short background on FreeGrad (the reader may also refer to Chapter 4):

FreeGrad. FreeGrad is an algorithm for unconstrained online convex optimization—it is a so-called parameter-free algorithm. The algorithm operates in rounds, where at each round t, FreeGrad outputs $\widehat{\boldsymbol{w}}_t$ in some convex set $\mathcal{W}$, say $\mathbb{R}^d$, then observes a vector $\boldsymbol{g}_t \in \mathbb{R}^d$, typically the sub-gradient of a loss function at round t. The algorithm guarantees a regret bound of the form $\sum_{t=1}^T \boldsymbol{g}_t^\top (\widehat{\boldsymbol{w}}_t - \boldsymbol{w}) \leq \widetilde{O}(\|\boldsymbol{w}\| \sqrt{V_T})$, for all $\boldsymbol{w} \in \mathcal{W}$, where $V_T := \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$. What is more, FreeGrad's outputs $(\widehat{\boldsymbol{w}}_t)$ ensure that a certain potential function—whose form is reminiscent of (8.22) with Y_i [resp. $|\cdot|$] replaced by $\boldsymbol{g}_i$ [resp. $\|\cdot\|$]—is non-increasing (see [Mhammedi and Koolen, 2020, Theorem 5]). In the proof of Proposition 8.15, we will reason about the outputs of FreeGrad in one dimension (i.e. $d = 1$) in response to the input $(g_t) \equiv (Y_t)$.

One way to prove Proposition 8.15 is to show that FreeGrad is a betting algorithm that bets fractions smaller than one of its current wealth at each round. In this case, Proposition 8.15 would follow from existing results due to, for example, Jun and Orabona [2019b]. However, for the sake of simplicity, we decided to present a proof that does not explicitly refer to bets.

Proof of Proposition 8.15. By [Mhammedi and Koolen, 2020, Theorem 5], FreeGrad's outputs $(\widehat{w}_i)$ in response to (Y_i) and initial scale $\gamma > 0$[6] guarantee,

$$\widehat{w}_{n+1} \cdot Y_{n+1} + M_{n+1} \leq M_n, \quad \text{for all } n \in \mathbb{N}, \tag{8.23}$$

where (M_n) are as in (8.22). Re-arranging this inequality, yields

$$\mathbb{E}_n[M_{n+1} - M_n] \leq -\mathbb{E}_n[\widehat{w}_{n+1} \cdot Y_{n+1}] = -\widehat{w}_{n+1} \cdot \mathbb{E}_n[Y_{n+1}] = 0,$$

where the penultimate equality follows by the fact that $\widehat{w}_{n+1}$ is a deterministic function of the history up to round n, inclusive, and the last equality follows by the assumption that $\mathbb{E}_n[Y_{n+1}] = 0$. □

Proof of Theorem 8.9. Let $\gamma > 0$ and $\delta \in (0, 1)$. By Proposition 8.15 and Ville's inequality (a generalization of Markov's inequality for supermartingales—see Lemma B.6), we have

$$\mathbf{P}[\exists n \geq 1,\ M_n \geq \gamma/\delta] \leq \delta \mathbb{E}[M_0]/\gamma = \delta. \tag{8.24}$$

[6] Technically, FreeGrad also requires a sequence of hints (h_t) that provides upper bounds on $(|Y_t|)$. Since $Y_i \in [-1, 1]$, these hints can all be set to $\gamma > 0$.

Further, for $Y_i := X_i - \mathbb{E}[X]$ and $\Delta_n := |\sum_{i=1}^n Y_i/n|$, $\xi_n := \frac{2}{n}\ln\frac{\sqrt{n}}{\gamma\delta}$, and the notation in (8.22), we have

$$\begin{aligned}
\Delta_n &\geq \left(\sqrt{(\gamma^2/n + \widehat{V}_n[X])\cdot\xi_n} + \gamma\xi_n\right)\frac{1}{1-\xi_n} \\
&\implies \frac{\Delta_n^2}{\gamma^2/n + \widehat{V}_n[X] + \Delta_n^2 + \Delta_n} \geq \xi_n, \\
&\implies \frac{\Delta_n^2}{\gamma^2/n + \widehat{V}_n[X] + \Delta_n^2 + \Delta_n} \geq \frac{2\ln\frac{\sqrt{V_n}}{\gamma\delta}}{n}, && (8.25) \\
&\implies \frac{|S_n|^2}{2V_n + 2|S_n|} \geq \ln\frac{\sqrt{V_n}}{\gamma\delta}, && (8.26) \\
&\implies M_n \geq \gamma/\delta. && (8.27)
\end{aligned}$$

where (8.25) follows by the fact that $|X_i| \leq 1$, and (8.26) follows by the the bias variance decomposition: we have $V_n/n = \gamma^2/n + \widehat{V}_n[X] + \Delta_n^2$. Thus, (8.27) implies that

$$\begin{aligned}
\delta &\geq \mathbf{P}[\exists n \geq n_0,\ M_n \geq \gamma/\delta] \\
&\geq \mathbf{P}\left[\exists n \geq n_0,\ \Delta_n \geq \left(\sqrt{(\gamma^2/n + \widehat{V}_n[X])\cdot\xi_n} + \gamma\xi_n\right)\frac{1}{1-\xi_n}\right] \leq, && (8.28)
\end{aligned}$$

where the first inequality follows by (8.24). □

8.8 Proof of Theorem 8.3

Proof. Let ϵ_n and $\mathcal{E}_n$ be as in Corollary 8.8 with $\bar{\mathfrak{R}}_n$ being the high probability upper bound on $\mathfrak{R}_n(\ell \circ \mathcal{H})$ that Algorithm 2 has access to ($\bar{\mathfrak{R}}_n$ satisfies (8.4)). By Corollary 8.8, we have $\mathbf{P}[\mathcal{E}_n] \geq 1 - 2/n$. For the rest of this proof, we will condition on the event $\mathcal{E}_n$.

By definition of $\tilde{h}_n$ in Algorithm 2 and Corollary 8.8, we have

$$\begin{aligned}
L(\tilde{h}_n) &\leq \widehat{L}_n(\tilde{h}_n) + \epsilon_n, && \text{(under the event } \mathcal{E}_n) \\
&= \inf_{h\in\mathcal{H}} \widehat{L}_n(h) + \epsilon_n, && (\tilde{h}_n \text{ is the ERM}) \\
&\leq \inf_{h\in\mathcal{H}} L(h) + 2\epsilon_n. && \text{(under the event } \mathcal{E}_n) && (8.29)
\end{aligned}$$

We now consider two cases pertaining to the condition in Line 3 of Algorithm 2:

Case 1. Suppose that the condition in Line 3 of Algorithm 2 is satisfied for $k = n$. In this case,

$$\begin{aligned} L(\hat{h}_n)] &= L(\tilde{h}_n), \\ &\leq \inf_{h \in \mathcal{H}} L(h) + 2\epsilon_n. \qquad \text{(by (8.29))} \end{aligned} \tag{8.30}$$

Case 2. Suppose the condition in Line 3 does not hold for $k = n$. This means $\hat{h}_n \equiv \hat{h}_{n-1}$, and so

$$\begin{aligned} L(\tilde{h}_n) &\geq \widehat{L}_n(\tilde{h}_n) - \epsilon_n, && \text{(under the event } \mathcal{E}_n) \\ &> \widehat{L}_n(\hat{h}_n) - \delta_n - \epsilon_n, && \text{(condition in Line 3 is false)} \\ &\geq L(\hat{h}_n) - \delta_n - 2\epsilon_n. && \text{(under the event } \mathcal{E}_n) \end{aligned} \tag{8.31}$$

Thus, by combining (8.29) and (8.31), we get

$$L(\hat{h}_n) \leq \inf_{h \in \mathcal{H}} L(h) + 4\epsilon_n + \delta_n. \tag{8.32}$$

From (8.30) and (8.32), we conclude that with probability $\mathbf{P}[\mathcal{E}_n] \geq 1 - 2/n$:

$$\mathbb{E}_{P(Z)}[\ell(\hat{h}_n, Z)] \leq \inf_{h \in \mathcal{H}} \left\{ \mathbb{E}_{P(Z)}[\ell(h, Z)] \right\} + 4\epsilon_n + \delta_n,$$

which implies the desired result. □

8.9 Proofs of Theorems 8.5 and 8.6

We start by presenting a sequence of intermediate results needed in the proofs of Theorems 8.5 and 8.6.

8.9.1 Intermediate Results

The proofs of all the results in this subsection are postponed to Subsection 8.9.3. We begin by a result pertaining to the concentration of the empirical risk using the empirical Bernstein inequality [Maurer and Pontil, 2009]:

Lemma 8.16. *Let $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$ and suppose that $\mathcal{H}$ is a finite set. Further, let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables such that $Z \sim P$. Then, for all $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\begin{aligned} L(h) - L(h') \leq \widehat{L}_n(h) - \widehat{L}_n(h') &+ \sqrt{\frac{2\sum_{i=1}^n (\ell(h, Z_i) - \ell(h', Z_i))^2 \ln(2|\mathcal{H}|^2/\delta)}{(n-1)^2}} \\ &+ \frac{5\ln(2|\mathcal{H}|^2/\delta)}{n-1}, \end{aligned} \tag{8.33}$$

for all $h, h' \in \mathcal{H}$.

The next lemma provides a way of bounding the square-root term in (8.33) under the Bernstein condition:

Lemma 8.17. *Let $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$, where $\mathcal{H}$ is a finite set. Further, let $\beta \in [0,1]$, $B > 0$, $n \in \mathbb{N}$, and suppose that the (β, B)-Bernstein condition holds. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sqrt{\frac{2^4 \sum_{i=1}^n (\ell(h, Z_i) - \ell(h_\star, Z_i))^2 \ln(2|\mathcal{H}|^2/\delta)}{(n-1)^2}} \leq \frac{L(h) - L(h_\star)}{2} + O\left(\max_{\beta' \in \{1,\beta\}} \left(\frac{\ln(|\mathcal{H}|/\delta)}{n}\right)^{\frac{1}{2-\beta'}}\right), \tag{8.34}$$

for all $h \in \mathcal{H}$, where $h_\star \in \arg\inf_{h \in \mathcal{H}} L(h)$.

Using the previous two lemmas, we derive the excess-risk rate of ERM under the Bernstein condition:

Lemma 8.18. *Let $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$, where $\mathcal{H}$ is a finite set. Further, let $\beta \in [0,1]$, $B > 0$, and suppose that the (β, B)-Bernstein condition holds. Then, the* ERM *$\tilde{h}_n \in \arg\inf_{h \in \mathcal{H}} \mathbb{E}_{\widehat{P}_n(Z)}[\ell(h, Z)]$ satisfies, with probability at least $1 - 2/n$,*

$$L(\tilde{h}_n) - L(h_\star) \leq O\left(\left(\frac{\ln(n|\mathcal{H}|)}{n}\right)^{\frac{1}{2-\beta}} + \frac{\ln(n|\mathcal{H}|)}{n}\right).$$

8.9.2 Proofs of Theorems 8.5 and 8.6

In what follows, it will be useful to define the event

$$\mathcal{E}_n := \left\{L(\tilde{h}_n) - L(\hat{h}_{n-1}) \leq \widehat{L}_n(\tilde{h}_n) - \widehat{L}_n(\hat{h}_{n-1}) + \epsilon_n\right\}, \; n \in \mathbb{N}, \tag{8.35}$$

$$\text{where} \quad \epsilon_n := \sqrt{\frac{2\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(\hat{h}_{n-1}, Z_i))^2 \ln(2n|\mathcal{H}|^2)}{(n-1)^2}} + \frac{5\ln(2n|\mathcal{H}|^2)}{n-1}, \tag{8.36}$$

and $(\tilde{h}_k)$ and $(\hat{h}_k)$ are as in Algorithm 2 with the choice of (δ_k) in (8.7). We note that by Lemma 8.16, we have $\mathbf{P}[\mathcal{E}_n] \geq 1 - 1/n \geq 1 - 2/n$, for all n. We begin by the proof of risk-monotonicity:

Proof of Theorem 8.5. Let $\Delta_n := L(\hat{h}_n) - L(\hat{h}_{n-1})$. Using the definitions of $\mathcal{E}_n, \epsilon_n$, and δ_n as in (8.35), (8.36), and (8.7) (with $k = n$), respectively, and following exactly the same steps as in the proof of Theorem 8.2, we arrive at

$$\mathbb{E}[\Delta_n] \leq \mathbb{E}[\Delta_n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\} \cap \mathcal{E}_n] + 2/n. \tag{8.37}$$

Now, if $\hat{h}_n \not\equiv \hat{h}_{n-1}$, then by Line 3 of Algorithm 2, we have

$$\widehat{L}_n(\tilde{h}_n) \leq \widehat{L}_n(\hat{h}_{n-1}) - \delta_n, \tag{8.38}$$

Under the event $\mathcal{E}_n$, we have

$$L(\tilde{h}_n) - L(\hat{h}_{n-1}) \leq \widehat{L}_n(\tilde{h}_n) - \widehat{L}_n(\hat{h}_{n-1}) + \epsilon_n.$$

This, in combination with (8.38), implies that under the event $\mathcal{E}_n \cap \{\hat{h}_n \not\equiv \hat{h}_{n-1}\}$,

$$\begin{aligned}\Delta_n + 2/n &= L(\tilde{h}_n) - L(\hat{h}_{n-1}) + 2/n, \\ &\leq 2/n - \delta_n + \epsilon_n, \\ &= 0,\end{aligned}$$

where in the last equality we substituted the expression of δ_n in (8.7) (with $k = n$). As a result,

$$\mathbb{E}_{P^n(Z_{1:n})}[\Delta_n + 2/n \mid \{\hat{h}_n \not\equiv \hat{h}_{n-1}\} \cap \mathcal{E}_n] \leq 0. \tag{8.39}$$

Combining (8.37) and (8.39) yields the desired result.

□

Proof of Theorem 8.6. Let ϵ_n and δ_n be as in (8.36) and (8.7) (with $k = n$), respectively. We consider two cases pertaining to the condition in Line 3 of Algorithm 2:

Case 1. Suppose that the condition in Line 3 of Algorithm 2 is satisfied for $k = n$. In this case, we have, by Lemma 8.18,

$$L(\hat{h}_n) - L(h_\star) = L(\tilde{h}_n) - L(h_\star) \leq O\left(\left(\frac{\ln(n|\mathcal{H}|)}{n}\right)^{\frac{1}{2-\beta}} + \frac{\ln(n|\mathcal{H}|)}{n}\right), \tag{8.40}$$

with probability at least $1 - 2/n$.

Case 2. Now suppose the condition in Line 3 does not hold for $k = n$. This means that $\hat{h}_n \equiv \hat{h}_{n-1}$, and so

$$\widehat{L}_n(\hat{h}_n) - \widehat{L}_n(\tilde{h}_n) \leq \delta_n. \tag{8.41}$$

Using this and Lemma 8.16, we have, with probability at least $1 - 1/n$,

$$\begin{aligned}
L(\hat{h}_n) &= L(\tilde{h}_n) + (L(\hat{h}_n) - L(\tilde{h}_n)), \\
&\leq L(\tilde{h}_n) + \widehat{L}_n(\hat{h}_n) - \widehat{L}_n(\tilde{h}_n) + \epsilon_n, \qquad \text{(Lemma 8.16)} \\
&\leq L(\tilde{h}_n) + \delta_n + \epsilon_n, \qquad \text{(by (8.41))} \\
&\overset{\hat{h}_n \equiv \hat{h}_{n-1}}{=} L(\tilde{h}_n) + \sqrt{\frac{8\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(\hat{h}_n, Z_i))^2 \ln(2n|\mathcal{H}|^2)}{(n-1)^2}} \\
&\quad + \frac{10\ln(2n|\mathcal{H}|^2)}{n-1} + \frac{2}{n}, \\
&\leq L(\tilde{h}_n) + \sqrt{\frac{16\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(h_\star, Z_i))^2 \ln(2n|\mathcal{H}|^2)}{(n-1)^2}} + \frac{10\ln(2n|\mathcal{H}|^2)}{n-1} + \frac{2}{n} \\
&\quad + \sqrt{\frac{16\sum_{i=1}^n (\ell(\hat{h}_n, Z_i) - \ell(h_\star, Z_i))^2 \ln(2n|\mathcal{H}|^2)}{(n-1)^2}},
\end{aligned} \tag{8.42}$$

where to obtain the last inequality, we used the fact that $(a-c)^2 \leq 2(a-b)^2 + 2(b-c)^2$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b, c \in \mathbb{R}_{\geq 0}$. Now, by (8.42), Lemma 8.17, and a union bound, we obtain, with probability at least $1 - 2/n$,

$$\begin{aligned}
L(\hat{h}_n) - L(h_\star) \leq L(\tilde{h}_n) - L(h_\star) + \frac{L(\tilde{h}_n) - L(h_\star)}{2} + \frac{L(\hat{h}_n) - L(h_\star)}{2} + \frac{2}{n} \\
+ O\left(\max_{\beta' \in \{1, \beta\}} \left(\frac{\ln(n|\mathcal{H}|)}{n}\right)^{\frac{1}{2-\beta'}}\right) + \frac{10\ln(2n|\mathcal{H}|^2)}{n-1},
\end{aligned}$$

which, after re-arranging, becomes

$$\begin{aligned}
\frac{L(\hat{h}_n) - L(h_\star)}{2} \leq \frac{3(L(\tilde{h}_n) - L(h_\star))}{2} + O\left(\max_{\beta' \in \{1, \beta\}} \left(\frac{\ln(n|\mathcal{H}|)}{n}\right)^{\frac{1}{2-\beta'}}\right) \\
+ \frac{10\ln(2n|\mathcal{H}|^2)}{n-1} + \frac{2}{n}.
\end{aligned} \tag{8.43}$$

Combining (8.43) with Lemma 8.18, and applying a union bound, we get, with probability at least $1 - 4/n$,

$$L(\hat{h}_n) - L(h_\star) \leq O\left(\left(\frac{\ln(n|\mathcal{H}|)}{n}\right)^{\frac{1}{2-\beta}} + \frac{\ln(n|\mathcal{H}|)}{n}\right). \tag{8.44}$$

The combination of (8.40) and (8.44) lead to the desired result. □

8.9.3 Proofs of Intermediate Results

Proof of Lemma 8.16. The proof follows by the empirical Bernstein inequality [Maurer and Pontil, 2009, Corollary 5] with the function $f : (\mathcal{H} \times \mathcal{H}) \times \mathcal{Z} \to [0, 1]$ defined

by

$$f((h,h'),z) = \big(\ell(h,z) - \ell(h',z) + 1\big)/2.$$

In particular, [Maurer and Pontil, 2009, Corollary 5] implies that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathbb{E}_{P(Z)}[f((h,h'),Z)] \le \mathbb{E}_{\widehat{P}_n(Z)}[f((h,h'),Z)] + \sqrt{\frac{2\widehat{V}_n \cdot \ln(2|\mathcal{H}|^2/\delta)}{n-1}} + \frac{7\ln(2|\mathcal{H}|^2/\delta)}{3(n-1)}, \tag{8.45}$$

for all $h, h' \in \mathcal{H}$, where $\widehat{V}_n$ is the sample variance:

$$\begin{aligned}
\widehat{V}_n &:= \frac{1}{n(n-1)} \sum_{1\le i<j\le n} (f((h,h'),Z_i) - f((h,h'),Z_j))^2, \\
&= \frac{1}{4n(n-1)} \sum_{1\le i<j\le n} (\ell(h,Z_i) - \ell(h',Z_i) - \ell(h,Z_j) + \ell(h',Z_j))^2, \\
&= \frac{n}{4(n-1)} \mathbb{E}_{\widehat{P}_n(Z)} \left[\Big(\ell(h,Z) - \ell(h',Z) - \mathbb{E}_{\widehat{P}_n(Z')}[\ell(h,Z') - \ell(h',Z')] \Big)^2 \right], \\
&= \frac{n}{4(n-1)} \left(\mathbb{E}_{\widehat{P}_n(Z)} \left[\big(\ell(h,Z) - \ell(h',Z)\big)^2 \right] - \mathbb{E}_{\widehat{P}_n(Z')}[\ell(h,Z') - \ell(h',Z')]^2 \right), \\
&\le \frac{n}{4(n-1)} \mathbb{E}_{\widehat{P}_n(Z)} \left[\big(\ell(h,Z) - \ell(h',Z)\big)^2 \right].
\end{aligned} \tag{8.46}$$

Plugging (8.46) into (8.45) and multiplying the resulting inequality by 2, leads to the desired inequality. □

Proof of Lemma 8.18. To simplify notation, we define

$$\epsilon'_n := \sqrt{\frac{2\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(h_\star, Z_i))^2 \ln(2n|\mathcal{H}|^2)}{(n-1)^2}} + \frac{5\ln(2n|\mathcal{H}|^2)}{n-1}, \quad n \in \mathbb{N}.$$

By Lemma 8.16, we have, with probability at least $1 - 1/n$,

$$\begin{aligned}
L(\tilde{h}_n) &= L(h_\star) + (L(\tilde{h}_n) - L(h_\star)), \\
&\le L(h_\star) + \widehat{L}_n(\tilde{h}_n) - \widehat{L}_n(h_\star) + \epsilon'_n, && \text{(Lemma 8.16)} \\
&\le L(h_\star) + \epsilon'_n, && (\tilde{h}_n \text{ is the ERM}) \\
&= L(h_\star) + \sqrt{\frac{2\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(h_\star, Z_i))^2 \ln(2n|\mathcal{H}|^2)}{(n-1)^2}} + \frac{5\ln(2n|\mathcal{H}|^2)}{n-1}.
\end{aligned} \tag{8.47}$$

By applying Lemma 8.17 to bound the middle term on the RHS of (8.47), we get with

probability at least $1 - 2/n$,

$$L(\tilde{h}_n) = L(h_\star) + \frac{L(\tilde{h}_n) - L(h_\star)}{2} + O\left(\max_{\beta' \in \{1,\beta\}} \left(\frac{\ln(n|\mathcal{H}|)}{n}\right)^{\frac{1}{2-\beta'}}\right) + \frac{5\ln(2n|\mathcal{H}|^2)}{n-1},$$
$$= L(h_\star) + \frac{L(\tilde{h}_n) - L(h_\star)}{2} + O\left(\max_{\beta' \in \{1,\beta\}} \left(\frac{\ln(n|\mathcal{H}|)}{n}\right)^{\frac{1}{2-\beta'}}\right). \tag{8.48}$$

After rearranging (8.48), we obtain the desired result. □

Proof of Lemma 8.17. Let $c := 2^4$. We use the fact that $\sqrt{xy} \leq (\nu x + y/\nu)/2$, for all $\nu > 0$, and apply it to the LHS of (8.34) with

$$\nu = \frac{\eta \cdot (n-1)}{8n}, \quad x = \frac{1}{n-1}\sum_{i=1}^{n}(\ell(\tilde{h}_n, Z_i) - \ell(\hat{h}_{n-1}, Z_i))^2, \quad \text{and} \quad y = \frac{c\ln(2|\mathcal{H}|^2/\delta)}{n-1},$$

which leads to, for all $\eta > 0$,

$$r_{n,\delta}(h) := \sqrt{\frac{c\sum_{i=1}^{n}(\ell(h, Z_i) - \ell(h_\star, Z_i))^2 \ln(2|\mathcal{H}|^2/\delta)}{(n-1)^2}},$$
$$\leq \frac{\eta}{16n}\sum_{i=1}^{n}(\ell(h, Z_i) - \ell(h_\star, Z_i))^2 + \frac{4nc\ln(2|\mathcal{H}|^2/\delta)}{(n-1)^2\eta},$$
$$\leq \frac{\eta}{16n}\sum_{i=1}^{n}(\ell(h, Z_i) - \ell(h_\star, Z_i))^2 + \frac{8c\ln(2|\mathcal{H}|^2/\delta)}{(n-1)\eta}, \tag{8.49}$$

where the last inequality follows by the fact that $n \leq 2(n-1)$, for all $n > 1$. Let

$$C_\beta := \left((1-\beta)^{1-\beta}\beta^\beta\right)^{\frac{\beta}{1-\beta}} + 3/2(2B)^{\frac{1}{1-\beta}}.$$

By combining (8.49) and Lemma B.7, we get, for any $\delta \in (0,1)$ and $\eta \in [0, 1/2]$, with probability at least $1 - \delta$,

$$\forall h \in \mathcal{H}, \;\; r_{n,\delta}(h) \leq (L(h) - L(h_\star))/2 + C_\beta \cdot \eta^{\frac{1}{1-\beta}}/4 + \frac{\ln(|\mathcal{H}|/\delta)}{2n\eta} + \frac{8c\ln(2|\mathcal{H}|^2/\delta)}{(n-1)\eta},$$
$$\leq (L(h) - L(h_\star))/2 + C_\beta \cdot \eta^{\frac{1}{1-\beta}}/4 + \frac{(16c + 1/2)\ln(2|\mathcal{H}|^2/\delta)}{n \cdot \eta}, \tag{8.50}$$

where the last inequality follows by the fact that $n \leq 2(n-1)$ and $|\mathcal{H}| \geq 1$. Now, minimizing the RHS of (8.50) over $\eta \in (0, 1/2)$ and invoking Lemma B.8, we get, for

any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\begin{aligned} r_{n,\delta}(h) &\leq \frac{L(h)-L(h_\star)}{2} + \frac{C_\beta \cdot (3-2\beta)}{4(1-\beta)} \left(\frac{4(1-\beta)(16c+1/2)\ln(2|\mathcal{H}|^2/\delta)}{C_\beta \cdot n} \right)^{\frac{1}{2-\beta}} \\ &\quad + \frac{2(16c+1/2)\ln(2|\mathcal{H}|^2/\delta)}{n}, \\ &\leq \frac{L(h)-L(h_\star)}{2} + \frac{C_\beta^{\frac{1-\beta}{2-\beta}} \cdot (3-2\beta)}{4(1-\beta)} \left(\frac{4(1-\beta)(16c+1/2)\ln(2|\mathcal{H}|^2/\delta)}{n} \right)^{\frac{1}{2-\beta}} \\ &\quad + \frac{2(16c+1/2)\ln(2|\mathcal{H}|^2/\delta)}{n}, \end{aligned} \tag{8.51}$$

for all $h \in \mathcal{H}$. Combining (8.51) with the fact that $\beta \mapsto C_\beta^{\frac{1-\beta}{2-\beta}}$ is bounded in $[0,1)$, we get the desired result. □

8.10 Proofs of Theorems 8.10 and 8.11

We start by presenting a sequence of intermediate results needed in the proofs of Theorems 8.5 and 8.6.

8.10.1 Intermediate Results

The proofs of all the results in this subsection are postponed to Subsection 8.10.3. We now present a bound on the risk difference $L(h) - L(h')$, for any $h, h' \in \mathcal{H}$, using our new time-uniform empirical Bernstein inequality in Theorem 8.9. For this, we recall the definition

$$\xi_k := \frac{2}{k} \ln\left(\frac{|\mathcal{H}|^2\sqrt{k}}{\delta} \right), \quad \text{and} \quad n_0 := \max\{n : \xi_n \geq 1/2\}. \tag{8.52}$$

for $k \geq 1$, $\delta \in (0,1)$.

Lemma 8.19. *Let $\gamma > 0$, $\delta \in (0,1)$, and $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$, where $\mathcal{H}$ is a finite set. Further, let (ξ_k) and n_0 be as in* (8.52)*. Then for i.i.d. random variables $Z, Z_1, \ldots, Z_n$, we have, with probability at least $1-\delta$,*

$$L(h) - L(h') \leq \widehat{L}_n(h) - \widehat{L}_n(h') + 2\sqrt{\frac{\sum_{i=1}^n (\ell(h,Z_i) - \ell(h',Z_i))^2 \cdot \xi_n}{n}} + 4\xi_n, \tag{8.53}$$

for all $h, h' \in \mathcal{H}$ and all $n \geq n_0$.

The next lemma provides a way of bounding the square-root term in (8.53) under the Bernstein condition:

Lemma 8.20. *Let $\delta \in (0,1)$ and $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$, where $\mathcal{H}$ is a finite set. Further, let $B > 0$, $n \in \mathbb{N}$, and suppose that the (β, B)-Bernstein condition holds. Then, for (ξ_k) and n_0*

as in (8.52)*, we have, with probability at least* $1-\delta$*,*

$$\sqrt{\frac{2^5\sum_{i=1}^n(\ell(h,Z_i)-\ell(h',Z_i))^2\cdot\xi_n}{n}} \le \frac{L(h)-L(h_\star)}{2} + O\left(\max_{\beta'\in\{1,\beta\}}\left(\frac{\ln(n|\mathcal{H}|/\delta)}{n}\right)^{\frac{1}{2-\beta'}}\right), \quad (8.54)$$

for all $h\in\mathcal{H}$ *and* $n\ge n_0$*, where* $h_\star\in\arg\inf_{h\in\mathcal{H}}L(h)$*.*

Using the previous two lemmas, we derive the excess-risk rate of ERM under the Bernstein condition:

Lemma 8.21. *Let* $\mathcal{H}$ *be a finite set and* $\ell:\mathcal{H}\times\mathcal{Z}\to[0,1]$*. Further, let* $\beta\in[0,1]$*,* $B>0$*, and suppose that the* (β,B)*-Bernstein condition holds. Then, the* ERM $\tilde{h}_n\in$ *arg* $\inf_{h\in\mathcal{H}}\mathbb{E}_{\widehat{P}_n(Z)}[\ell(h,Z)]$ *satisfies,*

$$\mathbf{P}\left[\forall n\ge n_0,\ L(\tilde{h}_n)-L(h_\star)\le O\left(\left(\frac{\ln(n|\mathcal{H}|/\delta)}{n}\right)^{\frac{1}{2-\beta}}+\frac{\ln(n|\mathcal{H}|/\delta)}{n}\right)\right]\ge 1-\delta,$$

where n_0 *is as in* (8.52)*.*

8.10.2 Proofs of Theorems 8.10 and 8.11

For (δ_k) and n_0 as in (8.15) and (8.52), respectively, it will be useful to define the event

$$\mathcal{E}:=\left\{\forall n\ge n_0,\ \ L(\tilde{h}_n)-L(\hat{h}_{n-1})\le\widehat{L}_n(\tilde{h}_n)-\widehat{L}_n(\hat{h}_{n-1})+\delta_n\right\}, \quad (8.55)$$

where $(\tilde{h}_k)$ and $(\hat{h}_k)$ are as in Algorithm 2 with the choice of (δ_k) in (8.15). Observe that by Lemma 8.19, we have $\mathbf{P}[\mathcal{E}]\ge 1-\delta$. We begin by the proof of risk-monotonicity:

Proof of Theorem 8.10. Let $\Delta_n:=L(\hat{h}_n)-L(\hat{h}_{n-1})$. Using the definitions of $\mathcal{E}$ and δ_k as in (8.55) and (8.15) (with $k=n$), respectively, we have

$$\begin{aligned}\Delta_n &= (L(\tilde{h}_n)-L(\hat{h}_{n-1}))\cdot\mathbb{I}\{\hat{h}_n\not\equiv\hat{h}_{n-1}\}+(L(\hat{h}_n)-L(\hat{h}_{n-1}))\cdot\mathbb{I}\{\hat{h}_n\equiv\hat{h}_{n-1}\},\\ &= (L(\tilde{h}_n)-L(\hat{h}_{n-1}))\cdot\mathbb{I}\{\hat{h}_n\not\equiv\hat{h}_{n-1}\}. \end{aligned}\quad (8.56)$$

Now, when $\hat{h}_n\not\equiv\hat{h}_{n-1}$, then by Line 3 of Algorithm 2, we have

$$\widehat{L}_n(\tilde{h}_n)\le\widehat{L}_n(\hat{h}_{n-1})-\delta_n. \quad (8.57)$$

Using this and (8.56), we have that under the event $\mathcal{E}$,

$$\forall n\ge n_0,\quad L(\tilde{h}_n)-L(\hat{h}_{n-1})\le\widehat{L}_n(\tilde{h}_n)-\widehat{L}_n(\hat{h}_{n-1})+\delta_n\le 0.$$

This, combined with the fact that $\mathbf{P}[\mathcal{E}]\ge 1-\delta$ (Lemma 8.19) completes the proof. □

Proof of Theorem 8.11. Let n_0 be as in (8.52). Further, let δ_n be as in (8.15) (with $k = n$) and $\mathcal{E}$ be as in (8.55). It will be convenient to also consider the events:

$$\mathcal{E}' := \left\{ \forall n \geq n_0, \;\; L(\tilde{h}_n) - L(h_\star) \leq O\left(\left(\frac{\ln(n|\mathcal{H}|/\delta)}{n} \right)^{\frac{1}{2-\beta}} + \frac{\ln(n|\mathcal{H}|/\delta)}{n} \right) \right\},$$

$$\mathcal{E}'' := \left\{ \forall n \geq n_0, \quad \begin{array}{c} \sqrt{\dfrac{2^5 \sum_{i=1}^n (\ell(h, Z_i) - \ell(h', Z_i))^2 \xi_n}{n}} \leq \dfrac{L(h) - L(h_\star)}{2} \\ + O\left(\max_{\beta' \in \{1, \beta\}} \left(\dfrac{\ln \frac{n|\mathcal{H}|}{\delta}}{n} \right)^{\frac{1}{2-\beta'}} \right) \end{array} \right\}.$$

We note that by Lemmas 8.19, 8.20, and 8.21, we have

$$\min(\mathbf{P}[\mathcal{E}], \mathbf{P}[\mathcal{E}'], \mathbf{P}[\mathcal{E}'']) \geq 1 - \delta. \tag{8.58}$$

For the rest of this proof, we will assume the event $\mathcal{E} \cap \mathcal{E}' \cap \mathcal{E}''$ holds, and let $n \geq n_0$. We consider two cases pertaining to the condition in Line 3 of Algorithm 2:

Case 1. Suppose that the condition in Line 3 of Algorithm 2 is satisfied for $k = n$. In this case, we have by the assumption that $\mathcal{E}'$ is true:

$$L(\hat{h}_n) - L(h_\star) = L(\tilde{h}_n) - L(h_\star) \leq O\left(\left(\frac{\ln(n|\mathcal{H}|/\delta)}{n} \right)^{\frac{1}{2-\beta}} + \frac{\ln(n|\mathcal{H}|/\delta)}{n} \right), \tag{8.59}$$

Case 2. Now suppose the condition in Line 3 does not hold for $k = n$. This means that $\hat{h}_n \equiv \hat{h}_{n-1}$, and so

$$\widehat{L}_n(\hat{h}_n) - \widehat{L}_n(\tilde{h}_n) \leq \delta_n. \tag{8.60}$$

Thus, by the assumption that $\mathcal{E}'$ is true, we have,

$$\begin{aligned}
L(\hat{h}_n) &= L(\tilde{h}_n) + (L(\hat{h}_n) - L(\tilde{h}_n)), \\
&\leq L(\tilde{h}_n) + \widehat{L}_n(\hat{h}_n) - \widehat{L}_n(\tilde{h}_n) + \delta_n, && (\mathcal{E} \text{ is true}) \\
&\leq L(\tilde{h}_n) + 2\delta_n, && (\text{by } (8.60)) \\
&= L(\tilde{h}_n) + 4\sqrt{\frac{\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(\hat{h}_n, Z_i))^2 \cdot \xi_n}{n}} + 8\xi_n, && (\hat{h}_n \equiv \hat{h}_{n-1}) \\
&\leq L(\tilde{h}_n) + 4\sqrt{\frac{2\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(h_\star, Z_i))^2 \cdot \xi_n}{n}} + 8\xi_n \\
&\quad + \sqrt{\frac{2\sum_{i=1}^n (\ell(\hat{h}_n, Z_i) - \ell(h_\star, Z_i))^2 \cdot \xi_n}{n}},
\end{aligned} \tag{8.61}$$

where to obtain the last inequality, we used the fact that $(a-c)^2 \leq 2(a-b)^2 + 2(b-c)^2$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b, c \in \mathbb{R}_{\geq 0}$. Now, by (8.61) and the assumption that $\mathcal{E}''$ holds, we have

$$L(\hat{h}_n) - L(h_\star) \leq L(\tilde{h}_n) - L(h_\star) + \frac{L(\tilde{h}_n) - L(h_\star)}{2} + \frac{L(\hat{h}_n) - L(h_\star)}{2} + 4\xi_n + O\left(\max_{\beta' \in \{1,\beta\}} \left(\frac{\ln(n|\mathcal{H}|/\delta)}{n}\right)^{\frac{1}{2-\beta'}}\right),$$

which, after re-arranging, becomes

$$\frac{L(\hat{h}_n) - L(h_\star)}{2} \leq \frac{3(L(\tilde{h}_n) - L(h_\star))}{2} + O\left(\max_{\beta' \in \{1,\beta\}} \left(\frac{\ln(n|\mathcal{H}|/\delta)}{n}\right)^{\frac{1}{2-\beta'}}\right) + 4\xi_n. \tag{8.62}$$

Combining (8.62) with the assumption that $\mathcal{E}'$ holds, we get

$$L(\hat{h}_n) - L(h_\star) \leq O\left(\left(\frac{\ln(n|\mathcal{H}|/\delta)}{n}\right)^{\frac{1}{2-\beta}} + \frac{\ln(n|\mathcal{H}|/\delta)}{n}\right).$$

This, together with (8.58) and a union bound yields the desired result. □

8.10.3 Proofs of Intermediate Results

Proof of Lemma 8.19. The proof follows by our new time-uniform concentration inequality in Theorem 8.9 with the function $f : (\mathcal{H} \times \mathcal{H}) \times \mathcal{Z} \to [0,1]$ defined by

$$f((h,h'),z) = \left(\ell(h,z) - \ell(h',z) + 1\right)/2.$$

The choice (ξ_k) and n_0 in (8.52) ensures that $\sqrt{\xi_n/n} \leq \xi_n$ and $1 - \xi_n \geq 1/2$ for all $n \geq n_0$. Thus, Theorem 8.9 implies that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathbb{E}_{P(Z)}[f((h,h'),Z)] \leq \mathbb{E}_{\widehat{P}_n(Z)}[f((h,h'),Z)] + 2\sqrt{\widehat{V}_n \cdot \xi_n} + 4\xi_n, \tag{8.63}$$

for all $h, h' \in \mathcal{H}$ and $n \geq n_0$, where $\widehat{V}_n$ is the (biased) sample variance:

$$\begin{aligned}
\widehat{V}_n &:= \frac{1}{n^2} \sum_{1 \leq i < j \leq n} (f((h,h'), Z_i) - f((h,h'), Z_j))^2, \\
&= \frac{1}{4n^2} \sum_{1 \leq i < j \leq n} (\ell(h, Z_i) - \ell(h', Z_i) - \ell(h, Z_j) + \ell(h', Z_j))^2, \\
&= \frac{1}{4} \mathbb{E}_{\widehat{P}_n(Z)} \left[\left(\ell(h, Z) - \ell(h', Z) - \mathbb{E}_{\widehat{P}_n(Z')}[\ell(h, Z') - \ell(h', Z')] \right)^2 \right], \\
&= \frac{1}{4} \left(\mathbb{E}_{\widehat{P}_n(Z)} \left[\left(\ell(h, Z) - \ell(h', Z) \right)^2 \right] - \mathbb{E}_{\widehat{P}_n(Z')}[\ell(h, Z') - \ell(h', Z')]^2 \right), \\
&\leq \frac{1}{4} \mathbb{E}_{\widehat{P}_n(Z)} \left[\left(\ell(h, Z) - \ell(h', Z) \right)^2 \right]. \qquad (8.64)
\end{aligned}$$

Plugging (8.64) into (8.63) and multiplying the resulting inequality by 2, leads to the desired inequality. □

Proof of Lemma 8.21. Let (ξ_k) and n_0 be as in (8.52), and $\mathcal{E}$ be as in (8.55). Further, consider the event $\mathcal{E}''$ defined in the proof of Theorem 8.11. To simplify notation, we define

$$\delta'_n := 2\sqrt{\frac{\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(h_\star, Z_i))^2 \cdot \xi_n}{n}} + 4\xi_n, \quad n \in \mathbb{N}.$$

We recall that by Lemmas 8.19 and 8.20, we have

$$\min(\mathbf{P}[\mathcal{E}], \mathbf{P}[\mathcal{E}'']) \geq 1 - \delta. \qquad (8.65)$$

For the rest of this proof, we will assume the event $\mathcal{E} \cap \mathcal{E}''$ holds, and let $n \geq n_0$. By the assumption that $\mathcal{E}$ holds, we have

$$\begin{aligned}
L(\tilde{h}_n) &= L(h_\star) + (L(\tilde{h}_n) - L(h_\star)), \\
&\leq L(h_\star) + \widehat{L}_n(\tilde{h}_n) - \widehat{L}_n(h_\star) + \delta'_n, && (\mathcal{E} \text{ is true}) \\
&\leq L(h_\star) + \delta'_n, && (\tilde{h}_n \text{ is the ERM}) \\
&= L(h_\star) + 2\sqrt{\frac{\sum_{i=1}^n (\ell(\tilde{h}_n, Z_i) - \ell(h_\star, Z_i))^2 \cdot \xi_n}{n}} + 4\xi_n. && (8.66)
\end{aligned}$$

Now by the assumption that $\mathcal{E}''$ holds, we can bound the middle term on the RHS of (8.66), leading to

$$\begin{aligned}
L(\tilde{h}_n) &= L(h_\star) + \frac{L(\tilde{h}_n) - L(h_\star)}{2} + O\left(\max_{\beta' \in \{1, \beta\}} \left(\frac{\ln(n|\mathcal{H}|/\delta)}{n} \right)^{\frac{1}{2-\beta'}} \right) + 4\xi_n, \\
&= L(h_\star) + \frac{L(\tilde{h}_n) - L(h_\star)}{2} + O\left(\max_{\beta' \in \{1, \beta\}} \left(\frac{\ln(n|\mathcal{H}|/\delta)}{n} \right)^{\frac{1}{2-\beta'}} \right). \qquad (8.67)
\end{aligned}$$

Combining (8.67) with (8.65), and applying a union bound, we obtain the desired result. □

Proof of Lemma 8.20. Let (ξ_k) and n_0 be as in (8.52), and $n \geq n_0$. We use the fact that $\sqrt{xy} \leq (\nu x + y/\nu)/2$, for all $\nu > 0$, and apply it to the LHS of (8.54) with

$$\nu = \frac{\eta}{8}, \quad x = \frac{1}{n}\sum_{i=1}^{n}(\ell(\tilde{h}_n, Z_i) - \ell(\hat{h}_{n-1}, Z_i))^2, \quad \text{and} \quad y = 2^5\xi_n,$$

which leads to, for all $\eta > 0$, and $c = 2^5$,

$$\begin{aligned} r_n(h) &:= \sqrt{\frac{c\sum_{i=1}^{n}(\ell(h, Z_i) - \ell(h_\star, Z_i))^2 \cdot \xi_n}{n}}, \\ &\leq \frac{\eta}{16n}\sum_{i=1}^{n}(\ell(h, Z_i) - \ell(h_\star, Z_i))^2 + \frac{4c\xi_n}{\eta}, \\ &\leq \frac{\eta}{16n}\sum_{i=1}^{n}(\ell(h, Z_i) - \ell(h_\star, Z_i))^2 + \frac{8c\ln(\sqrt{n}|\mathcal{H}|^2/\delta)}{n\eta}, \end{aligned} \tag{8.68}$$

where in the last inequality we have substituted the expression of ξ_n in (8.52). Now, let $C_\beta := \left((1-\beta)^{1-\beta}\beta^\beta\right)^{\frac{\beta}{1-\beta}} + 3/2(2B)^{\frac{1}{1-\beta}}$. By combining (8.68) and Lemma B.7, we get, for any $\delta \in (0,1)$ and $\eta \in [0, 1/2]$, with probability at least $1-\delta$,

$$\begin{aligned} \forall h \in \mathcal{H}, \forall n \geq n_0, \;\; r_n(h) &\leq (L(h) - L(h_\star))/2 + C_\beta \cdot \eta^{\frac{1}{1-\beta}}/4 + \frac{\ln(|\mathcal{H}|/\delta)}{2n\eta} \\ &\quad + \frac{8c\ln(\sqrt{n}|\mathcal{H}|^2/\delta)}{n\eta}, \\ &\leq (L(h) - L(h_\star))/2 + C_\beta \cdot \eta^{\frac{1}{1-\beta}}/4 \\ &\quad + \frac{(8c+1/2)\ln(\sqrt{n}|\mathcal{H}|^2/\delta)}{n \cdot \eta}, \end{aligned} \tag{8.69}$$

Now, minimizing the RHS of (8.69) over $\eta \in (0, 1/2)$ and invoking Lemma B.8, we get, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\begin{aligned} r_n(h) &\leq \frac{L(h) - L(h_\star)}{2} + \frac{C_\beta \cdot (3-2\beta)}{4(1-\beta)}\left(\frac{4(1-\beta)(8c+1/2)\ln(\sqrt{n}|\mathcal{H}|^2/\delta)}{C_\beta \cdot n}\right)^{\frac{1}{2-\beta}} \\ &\quad + \frac{2(16c+1/2)\ln(\sqrt{n}|\mathcal{H}|^2/\delta)}{n}, \\ &\leq \frac{L(h) - L(h_\star)}{2} + \frac{C_\beta^{\frac{1-\beta}{2-\beta}} \cdot (3-2\beta)}{4(1-\beta)}\left(\frac{4(1-\beta)(8c+1/2)\ln(\sqrt{n}|\mathcal{H}|^2/\delta)}{n}\right)^{\frac{1}{2-\beta}} \\ &\quad + \frac{2(8c+1/2)\ln(\sqrt{n}|\mathcal{H}|^2/\delta)}{n}, \end{aligned} \tag{8.70}$$

for all $h \in \mathcal{H}$ and all $n \geq n_0$. Combining (8.70) with the fact that $\beta \mapsto C_\beta^{\frac{1-\beta}{2-\beta}}$ is bounded in $[0, 1)$, we get the desired result. □

8.11 Summary and Future Work

In this chapter, we derived the first consistent and risk-monotonic algorithms for a general statistical learning setting with bounded losses. By definition, Risk-monotonicity avoids double descent, making GERM the first algorithm that provably mitigates the latter under the general setting we consider in this chapter. Surprisingly, GERM is able to achieve this without necessarily compromising on its excess-risk rate.

Computational considerations. From a computational perspective, the main setback of Algorithm 2 is that returning the final hypothesis $\hat{h}_n$ requires performing n-ERM computations to evaluate the intermediate hypotheses $(\tilde{h}_k)_{k \in [n]}$. Nevertheless, in practice, ERM solutions for sample sizes k and $k+1$ may be close to each other, and this fact may be leveraged to efficiently generate the ERM sequence $(\tilde{h}_k)$. When the loss ℓ is convex in the first argument, it is possible to efficiently generate the final predictor $\hat{h}_n$ using tools from online convex optimization (see Section 8.5). However, in general it is unclear whether risk-monotonicity can be achieved without the (greedy) for-loop procedure of Algorithm 2.

We note also that if one only wants a decreasing risk after some sample size $s \in \mathbb{N}$ then computing the hypotheses $(\hat{h}_k)_{k<s}$ is unnecessary. In this case, the for-loop in Algorithm 2 need only start at $k = s$; the resulting hypotheses would satisfy the monotonicity condition in (8.1) for all $n \geq s$ (the proofs of Theorems 8.2 and 8.5 can easily be modified to show this).

Extensions. Some important questions remain open along the axes of assumptions. In particular, can we remove the boundedness condition on the loss while retaining risk-monotonicity? Lifting the boundedness assumption may be key in resolving another COLT open problem [Grünwald and Kotłowski, 2011] regarding achievable risk rates of log-loss Bayesian predictors[7]. Our results build foundations for these avenues, which are promising subjects for future work.

[7] We remark that their setting is slightly different from ours, in that their predictors may be improper, i.e. not within the hypothesis class $\mathcal{H}$.

Chapter 9

Conclusion

In this thesis, we have explored different aspects of adaptivity in machine learning. We started by looking at adaptivity in online learning—a prominent learning setting in machine learning. Specifically, we considered the online convex optimization setting with either bounded or unbounded decision sets. For the former, we developed an algorithm based on MetaGrad that fully adapts to virtually all parameters of interest in the bounded OCO setting, such as curvature, Lipschitz constant, and time horizon.

In the unbounded setting, we have derived two scale-free and parameter-free algorithms (FreeGrad and Matrix-FreeGrad) with state-of-the-art regret guarantees. We also showed that it is possible to be scale-free and parameter-free simultaneously at the cost of an additional term in the regret that scales with $\|\boldsymbol{w}\|^3$. Furthermore, we proved a lower bound that shows that this term cannot be avoided in general.

In the statistical learning setting, we have presented new data-dependent generalization bounds that have the advantage of becoming small when the learning algorithm used is stable or whenever the problem at hand is "easy" (as characterized by the Bernstein condition). We have also derived sharp concentration inequalities and generalization bounds for the CVaR—an alternative risk measure used in risk-sensitive applications—by reducing CVaR estimation to estimating an expectation from empirical means. Finally, by leveraging some ideas from Chapter 5 (data-dependent bounds), we devised the first algorithm in the general statistical learning setting with bounded losses that not only has a monotonic risk curve, but whose excess risk converges at a fast rate under the Bernstein "easiness" condition.

In the next section, we outline some interesting future research directions.

9.1 Future Work

Online learning in full-information. In Chapter 4, we derived a regret lower-bound for any scale-free algorithm which insists on a root-T-type regret in the unbounded OCO setting (T denotes the length of the horizon). This lower bound includes a penalty involving the cubed norm of the comparator. In recent works [Jun and Orabona, 2019b; Van der Hoeven, 2019b], it was observed that this penalty is not needed in a practical stochastic setting where the norm of the losses may

only be unbounded through stochastic noise. However, strong assumptions on the underlying noise distributions were made in these works to arrive at this result. Whether the cubed norm penalty can be avoided in general remains an interesting open question [Orabona and Cutkosky, 2020]. A positive answer would have several important implications. For example, an algorithm, say A, which avoids the penalty mentioned above, can be used together with modern online-to-batch conversion techniques [Cutkosky, 2019a] to produce the *first* algorithm for stochastic optimization that can adapt to the norm of the function minimizer. Algorithm A can also be useful in differential privacy, in particular, Local Differential Privacy (LDP) [Jun and Orabona, 2019b; Van der Hoeven, 2019b] (see also Chapter 1).

While studying the online convex optimization setting is a good starting point for understanding and characterizing the fundamental limits of adaptivity in machine learning, there are many practical applications where the objective function is non-convex (e.g. when training deep NNs). Developing parameter-free algorithms for the online non-convex optimization setting is an exciting future work direction [Orabona and Cutkosky, 2020].

Online learning in partial information. Despite existing reductions to online learning [Lykouris et al., 2018], various questions remain open around the fundamental limits of adaptivity in specific bandit settings. An important open problem here is that of contextual bandit model selection [Foster et al., 2020], where the goal is to select the best model out of a set of candidates using a number of samples that scales only with the complexity of the best model. Another problem involving adaptivity (with a more practical motivation) is that of achieving a small-loss bound in contextual bandits using only an ERM oracle [Agarwal et al., 2017].

Stepping up from the bandit setting to reinforcement learning, an interesting future research direction would be to extend the result due to Mhammedi et al. [2020a] to the case where the dynamics of the latent state are locally (instead of globally) linear. Achieving this would constitute a significant breakthrough in continuous control.

Closing the gap between theory and practice in statistical learning. The gap between the theoretical and observed generalization performance of NNs remains elusive [Nagarajan and Kolter, 2019]. Closing this gap is an important goal towards understanding and improving current algorithms. Although PAC-Bayesian bounds may be tighter than alternatives in some cases, they have a major drawback. They do not allow for deterministic algorithms; those who output hypotheses that are deterministic functions of the observed samples. Instead, the output is required to be a non-degenerate distribution over hypotheses; a degenerate distribution will blow up the generalization bound due to the presence of a KL term between this (posterior) distribution and some prior. However, this restriction seems to be merely an artifact of the analysis rather than a real phenomenon. In practice, good generalization is witnessed for algorithms that output point predictors; for example, those obtained from the last iterate of stochastic gradient descent. Understanding this gap between

theory and practice and ultimately deriving non-vacuous generalization bounds for deterministic algorithms is an important goal.

Generalization bounds for coherent risk measures. For the goal of obtaining generalization bounds for alternative risk measures, deriving a PAC-Bayesian bound for CVaR_α (as we did in Chapter 6) is only scratching the surface. In fact, there exists a larger family of risk measures for which generalization bounds do not exist. *Coherent Risk Measures* (CRM) is one such family of measures that possess properties that make it desirable in many machine learning applications. In fact, Williamson and Menon [2019] showed that these risk measures are good candidate objectives to optimize to ensure a certain degree of fairness when learning from data. I conjecture that the choice of a risk measure can become part of the learning process itself, in the same way as the choice of the loss can be made part of this process [Walder and Nock, 2020]. In this case, having generalization bounds for CRMs will be crucial. The results presented in Chapters 6 and 7 may be a good starting point towards this goal.

Appendices

A Technical Results for Chapter 7

Lemma A.1. *Let $R, \widehat{R}, A, B > 0$. If $R \leq \widehat{R} + \sqrt{RA} + B$, then*

$$\begin{aligned} R &\leq \widehat{R} + \frac{A}{2} + B + \sqrt{A\widehat{R} + A^2/4 + AB}, \\ &\leq \widehat{R} + \sqrt{\widehat{R}A} + 2B + A. \end{aligned}$$

Proof. If $R \leq \widehat{R} + \sqrt{RA} + B$, then for all $\eta > 0$,

$$R \leq \widehat{R} + \frac{\eta}{2}R + \frac{A}{2\eta} + B,$$

which after rearranging, becomes,

$$R \leq \frac{\widehat{R}}{1-\eta/2} + \frac{A}{2\eta \cdot (1-\eta/2)} + \frac{B}{1-\eta/2}, \quad \text{for } \eta \notin \{0, 2\}. \tag{1}$$

The minimizer of the RHS of (1) is given by

$$\eta = \eta_\star := \frac{-A + \sqrt{A^2 + 4AB + 4A\widehat{R}}}{2(B + \widehat{R})}.$$

Plugging this η into (1), yields,

$$\begin{aligned} R &\leq \widehat{R} + \frac{A}{2} + B + \frac{1}{2}\sqrt{4A\widehat{R} + A^2 + 4AB}, \\ &\leq \widehat{R} + A + 2B + \sqrt{A\widehat{R}}, \end{aligned} \tag{2}$$

where (2) follows by the facts that

$$A^2 + 4AB \leq (A + 2B)^2 \quad \textit{and} \quad \sqrt{4\widehat{R}A + (A+2B)^2} \leq \sqrt{4\widehat{R}A} + A + 2B.$$

□

Lemma A.2. *For a random variable Z, $c > 0$, and $p \in \mathbb{N}$, we have*

$$\mathbb{E}[Z^p \cdot \mathbb{I}\{Z \geq c\}] = v^p \cdot \mathbf{P}[Z \geq c] + \int_{v^p}^{+\infty} \mathbf{P}[Z^p \geq u]\mathrm{d}u.$$

Proof. Let $Y := Z^p \cdot \mathbb{Z} \geq$. Since $Y \geq 0$, we have

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^{+\infty} \mathbf{P}[Y \geq u]\mathrm{d}u, \\
&= \int_0^{v^p} \mathbf{P}[Y \geq u]\mathrm{d}u + \int_{v^p}^{+\infty} \mathbf{P}[Y \geq u]\mathrm{d}u, \\
&= \int_0^{v^p} \mathbf{P}[Z \geq u]\mathrm{d}u + \int_{v^p}^{+\infty} \mathbf{P}[Z^p \cdot \mathbb{I}\{Z \geq c\} \geq u]\mathrm{d}u, \\
&= v^p \cdot \mathbf{P}[Z \geq c] + \int_{v^p}^{+\infty} \mathbf{P}[Z^p \geq u]\mathrm{d}u.
\end{aligned}$$

□

Lemma A.3. *For $\alpha > 0$ and a non-negative random variable X, we have $\mathrm{C}[X] \leq \mathbb{E}[X]/\alpha$.*

Proof. For $\mu_\alpha := \inf\{\mu \in \mathbb{R} : \mathbf{P}[X \geq \mu] \leq \alpha\}$, we have

$$\mathrm{C}[X] = \mathbb{E}[X \mid X \geq \mu] = \mathbb{E}[X \cdot \mathbb{I}\{X \geq \mu\}]/\alpha \leq \mathbb{E}[X]/\alpha,$$

where the last inequality follows by positivity of X. □

Lemma A.4. *For a random variable X, we have*

$$|\mathrm{C}[X]| \leq \mathrm{C}[|X|].$$

Proof. By (7.4a), there exists a set $\mathcal{Q} \subset \{\Omega \to \mathbb{R}\}$ such that $\mathrm{C}[X] = \sup_{Q \in \mathcal{Q}} \mathbb{E}[XQ]$. Therefore, by Jensen's inequality, we have

$$|\mathrm{C}[X]| = \sup_{Q \in \mathcal{Q}} |\mathbb{E}[XQ]| \leq \sup_{Q \in \mathcal{Q}} \mathbb{E}[|X|Q] = \mathrm{C}[|X|].$$

□

Lemma A.5. *Let $v, y > 0$, $\alpha \in (0,1)$, and X be a real random variable. Further, let $\mu_{\alpha,i} := \inf\{\mu \in \mathbb{R} \mid \mathbf{P}[X^i \geq \mu^i] \leq \alpha\}$, for $i \in [2]$. If X satisfies (7.7) with $p, \lambda > 0$, then for $i \in [2]$, $y > 0$, and f as in (7.23)*

$$\begin{aligned}
&\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq (\lambda \ln(v/y))^{1/p}\}] \leq y\lambda^{1/p} \cdot \left(\ln(v/y)^{1/p} + v\Gamma(1/p, \ln(v/y))/(yp)\right), \\
&\quad \textit{and} \quad \mathbb{E}[X^2 \cdot \mathbb{I}\{X^i \geq \mu^i_{\alpha,i}\}] \leq \alpha\lambda^{2/p} f_{p/2}(\alpha).
\end{aligned}$$

Furthermore, if X satisfies (7.8) with $p > 2$, then for $i \in [2]$ and $y > 0$,

$$\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq (v/y)^{1/p}\}] \leq \frac{py(v/y)^{1/p}}{p-1}. \tag{3}$$

Proof. We start with the case of finite exponential moment. Let $\mu_{\alpha,i}, i \in [2]$, be as in the lemma statement. By Lemma A.2, we have

$$\begin{aligned}\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq (\lambda \ln(v/y))^{1/p}\}] &\leq y(\lambda \ln(v/y))^{1/p} + \int_{(\lambda \ln(v/y))^{1/p}}^{+\infty} \mathbf{P}[|X| \geq u]\mathrm{d}u, \\ &\leq y(\lambda \ln(v/y))^{1/p} + v \int_{(\lambda \ln(v/y))^{1/p}}^{+\infty} e^{-\frac{u^p}{\lambda}} \mathrm{d}u, \\ &\leq y(\lambda \ln(v/y))^{1/p} + v\lambda^{1/p}\Gamma(1/p, \ln(v/y))/p.\end{aligned}$$

Similarity, by Lemma A.2,

$$\begin{aligned}\mathbb{E}[X^2 \cdot \mathbb{I}\{X^i \geq \mu^i_{\alpha,i}\}] &= \mathbb{E}[X^2 \cdot \mathbb{I}\{\mu^i_{\alpha,i} \leq X^i \leq (\lambda \ln(v/\alpha))^{i/p}\}] \\ &\quad + \mathbb{E}[X^2 \cdot \mathbb{I}\{X^i \geq (\lambda \ln(v/\alpha))^{i/p}\}], \\ &\leq \mathbb{E}[X^2 \cdot \mathbb{I}\{\mu^i_{\alpha,i} \leq X^i \leq (\lambda \ln(v/\alpha))^{i/p}\}] \\ &\quad + \mathbb{E}[X^2 \cdot \mathbb{I}\{X^2 \geq (\lambda \ln(v/\alpha))^{2/p}\}], \\ &\leq 2\alpha(\lambda \ln(v/\alpha))^{2/p} + \int_{(\lambda \ln(v/\alpha))^{2/p}}^{+\infty} \mathbf{P}[X^2 \geq u]\mathrm{d}u, \\ &\leq 2\alpha(\lambda \ln(v/\alpha))^{2/p} + v \int_{(\lambda \ln(v/\alpha))^{2/p}}^{+\infty} e^{-\frac{u^{p/2}}{\lambda}} \mathrm{d}u, \\ &\leq 2\alpha(\lambda \ln(v/\alpha))^{2/p} + 2v\lambda^{2/p}\Gamma(2/p, \ln(v/\alpha))/p.\end{aligned}$$

We now move to case where X has a finite pth order moment for $p > 2$. As above, by Lemma A.2,

$$\begin{aligned}\mathbb{E}[|X| \cdot \mathbb{I}\{|X| \geq (v/y)^{1/p}\}] &\leq y(v/y)^{1/p} + \int_{(v/y)^{1/p}}^{+\infty} \mathbf{P}[|X| \geq u]\mathrm{d}u, \\ &\leq y(v/y)^{1/p} + v \int_{(v/y)^{1/p}}^{+\infty} u^{-p}\mathrm{d}u, \\ &\leq y(v/y)^{1/p} + \frac{y(v/y)^{1/p}}{p-1}, \\ &= \frac{p}{p-1} y(v/y)^{1/p}.\end{aligned}$$

□

Lemma A.6. *Let $\alpha \in (0,1)$ and Z be a real random variable. Then, for $\mu_\alpha := \inf\{\mu \in \mathbb{R} \mid \mathbf{P}[Z \geq \mu]\}$ and $\mu_{\alpha,2} := \inf\{\mu \in \mathbb{R} \mid \mathbf{P}[|Z| \geq \mu]\}$, we have*

$$\mathbb{E}[(\mu_\alpha + \alpha^{-1}\mathbb{E}[[Z - \mu_\alpha]_+])^2] \leq \frac{1}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{Z \geq \mu_\alpha\}] + \frac{3}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{|Z| \geq \mu_{\alpha,2}\}].$$

Proof of Lemma A.6. Let $\mu_{\alpha,2} := \inf\{\mu \in \mathbb{R} \mid \mathbf{P}[|Z| \geq \mu] \leq \alpha\}$. Using the fact that

$|\mu_\alpha| \leq \mu_{\alpha,2}$ and expanding the square term, we get

$$\begin{aligned}
\mathbb{E}[(\mu_\alpha + [Z-\mu_\alpha]_+/\alpha)^2] &= \mu_\alpha^2 + \frac{2\mu_\alpha \mathbb{E}[[Z-\mu_\alpha]_+]}{\alpha} + \frac{1}{\alpha^2}\mathbb{E}[[Z-\mu_\alpha]_+^2], \\
&= \mu_\alpha^2 + 2\mu_\alpha(\mathrm{C}[Z]-\mu_\alpha) + \frac{1}{\alpha^2}\mathbb{E}[(Z-\mu_\alpha)^2 \cdot \mathbb{I}\{Z \geq \mu_\alpha\}], \\
&= \mu_\alpha^2 + 2\mu_\alpha(\mathrm{C}[Z]-\mu_\alpha) + \frac{1}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{Z \geq \mu_\alpha\}] \\
&\quad - \frac{2\mu_\alpha}{\alpha^2}\mathbb{E}[Z \cdot \mathbb{I}\{Z \geq \mu_\alpha\}] + \frac{\mu_\alpha^2 \cdot \mathbf{P}[Z \geq \mu_\alpha]}{\alpha^2}, \\
&= -\mu_\alpha^2 + 2\mu_\alpha \mathrm{C}[Z] + \frac{1}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{Z \geq \mu_\alpha\}] - \frac{2\mu_\alpha \mathrm{C}[Z]}{\alpha} + \frac{\mu_\alpha^2}{\alpha}, \\
&\leq \frac{1}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{Z \geq \mu_\alpha\}] + \frac{4\mu_{\alpha,2}|\mathrm{C}[Z]|}{\alpha} + \frac{(\mu_{\alpha,2})^2}{\alpha}, \\
&\leq \frac{1}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{Z \geq \mu_\alpha\}] + \frac{2\mu_{\alpha,2}\mathrm{C}[|Z|]}{\alpha} + \frac{(\mu_{\alpha,2})^2}{\alpha}, \qquad (4) \\
&\leq \frac{1}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{Z \geq \mu_\alpha\}] + \frac{3\mathrm{C}[|Z|]^2}{\alpha}, \qquad (5)
\end{aligned}$$

where (4) follows by Lemma A.4, and (5) follows by the fact that $\mu_{\alpha,2} \leq \mathrm{C}[|Z|]$. Now, by the dual formulation of CVaR in (7.4a), we have

$$\begin{aligned}
\mathrm{C}[|Z|]^2 &= \left(\sup_{Q \in \mathcal{Q}} \mathbb{E}[|Z|Q]\right)^2, \\
&\leq \sup_{Q \in \mathcal{Q}} \mathbb{E}[|Z|^2 Q^2], && \text{(Jensen's inequality)} \\
&\leq \sup_{Q \in \mathcal{Q}} \mathbb{E}[|Z|^2 Q]/\alpha, && \text{(By def. of } Q \text{ in (7.4a))} \\
&= \mathrm{C}[Z^2]/\alpha, \\
&= \frac{1}{\alpha^2}\mathbb{E}[Z^2 \cdot \mathbb{I}\{|Z| \geq \mu_{\alpha,2}\}].
\end{aligned}$$

Combining this with (5) yields the desired result. □

B Technical Results for Chapter 8

In this appendix, it will be convenient to adopt the ESI notation [Koolen et al., 2016]:

Definition B.1 (Exponential Stochastic Inequality (ESI) notation)**.** *Let $\eta > 0$, and X, Y be any two random variables. We define*

$$X \trianglelefteq_\eta Y \iff X - Y \trianglelefteq_\eta 0 \iff \mathbb{E}\left[e^{\eta(X-Y)}\right] \leq 1.$$

In what follows, we denote by $X_h(Z) := \ell(h, Z) - \ell(h_*, Z)$, $h \in \mathcal{H}$, the excess-loss

random variable, where h_* is the risk minimizer within $\mathcal{H}$. Let

$$\Phi_\eta := \frac{1}{\eta} \ln \mathbb{E}_{P(Z)} \left[e^{-\eta X_h(Z)} \right] \tag{6}$$

be the *normalized cumulant generating function* of $X_h(Z)$. We note that since the loss ℓ takes values in the interval $[0,1]$, we have

$$X_h(z) \in [-1,1], \quad \text{for all } (h,z) \in \mathcal{H} \times \mathcal{Z}.$$

We now present some existing results pertaining to the excess-loss random variable $X_h(Z)$ and its normalized cumulant generating function, which will be useful in the proof of Theorem 8.5:

Lemma B.1 (Koolen et al. [2016]). *Let $h \in \mathcal{H}$, X_h, and Φ_η be as above. Then, for all $\eta \geq 0$,*

$$\alpha_\eta \cdot X_h(Z)^2 - X_h(Z) \trianglelefteq_\eta \Phi_{2\eta} + \alpha_\eta \cdot \Phi_{2\eta}^2, \quad \textit{where } \alpha_\eta := \frac{\eta}{1+\sqrt{1+4\eta^2}}.$$

Lemma B.2 (Koolen et al. [2016]). *If the (β, B)-Bernstein condition holds for $(\beta, B) \in [0,1] \times \mathbb{R}_{>0}$, then for Φ_η as in* (6)*, it holds that*

$$\Phi_\eta \leq (B\eta)^{\frac{1}{1-\beta}}, \quad \textit{for all } \eta \in (0,1].$$

Lemma B.3 (Cesa-Bianchi and Lugosi [2006]). *For Φ_η as in* (6)*, it holds that*

$$\Phi_\eta \leq \frac{\eta}{2}, \quad \textit{for all } \eta \in \mathbb{R}.$$

Lemma B.4 (Cesa-Bianchi and Lugosi [2006]). *The excess-loss random variable $X_h(Z)$ satisfies*

$$X_h(Z) - \mathbb{E}_{P(Z)}[X_h(Z)] \trianglelefteq_\eta \eta \cdot \mathbb{E}_{P(Z)}[X_h(Z)^2], \quad \textit{for all } \eta \in [0,1].$$

The following useful proposition is imported from [Mhammedi et al., 2019a]:

Proposition B.5. [ESI Transitivity and Chain Rule] *(a) Let $Z_1, \ldots, Z_n$ be any random variables on $\mathcal{Z}$ (not necessarily independent). If for some $(\gamma_i)_{i\in[n]} \in (0,+\infty)^n$, $Z_i \trianglelefteq_{\gamma_i} 0$, for all $i \in [n]$, then*

$$\sum_{i=1}^n Z_i \trianglelefteq_{\nu_n} 0, \quad \textit{where } \nu_n := \left(\sum_{i=1}^n \frac{1}{\gamma_i} \right)^{-1} \quad (\textit{so if } \forall i \in [n], \gamma_i = \gamma > 0 \textit{ then } \nu_n = \gamma/n).$$

(b) Suppose now that $Z_1, \ldots, Z_n$ are i.i.d. and let $f : \mathcal{Z} \times \bigcup_{i=1}^n \mathcal{Z}^i \to \mathbb{R}$ be any real-valued function. If for some $\eta > 0$, $f(Z_i; z_{<i}) \trianglelefteq_\eta 0$, for all $i \in [n]$ and all $z_{<i} \in \mathcal{Z}^{i-1}$, then $\sum_{i=1}^n f(Z_i; Z_{<i}) \trianglelefteq_\eta 0$.

To proof of our time-uniform concentration in Theorem 8.9, we will require the following generalization of Markov's inequality (we state the version found in Howard

et al. [2020]):

Lemma B.6 (Ville's inequality). *If $(M_n)_{n\geq 0}$ is a non-negative supermartingale, then for any $a > 0$,*

$$\mathbf{P}[\exists n \geq 1 : M_n \geq a] \leq \frac{M_0}{a}.$$

The upcoming lemmas will help us bound the sequences of gaps (δ_k) in (8.7) and (8.15) under the Bernstein condition.

Lemma B.7. *Let $\beta \in [0,1]$, $B > 0$, $n \in \mathbb{N}$. Further, let $Z, Z_1, \ldots, Z_n \in \mathcal{Z}$ be i.i.d. random variables and suppose that the (β, B)-Bernstein condition holds for the loss $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1]$, where $\mathcal{H}$ is a finite set. Then, for any $\eta \in [0, 1/2]$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\forall h \in \mathcal{H}, \forall n \geq 1, \quad \frac{\eta}{n}\sum_{i=1}^{n}(\ell(h, Z_i) - \ell(h_\star, Z_i))^2 \leq 8(L(h) - L(h_\star)) + 4C_\beta \cdot \eta^{\frac{1}{1-\beta}} + \frac{8\ln(|\mathcal{H}|/\delta)}{n\eta}, \tag{7}$$

where $h_\star \in \arg\inf_{h\in\mathcal{H}} L(h)$ and $C_\beta := \left((1-\beta)^{1-\beta}\beta^\beta\right)^{\frac{\beta}{1-\beta}} + 3/2(2B)^{\frac{1}{1-\beta}}$.

Proof. Let $\delta \in (0,1)$ and define $X_h(z) := \ell(h,z) - \ell(h_*, z)$, for $z \in \mathcal{Z}$. For any $\eta \in [0, 1/2]$ and $h \in \mathcal{H}$ our strategy is to show that, under the (β, B)-Bernstein condition,

$$M_n := \exp\left(\eta^2 \sum_{i=1}^{n} X_h(Z_i)^2/8 - n\eta\mathbb{E}_{P(Z)}[X_h(Z)] + nC_\beta \cdot \eta^{\frac{2-\beta}{1-\beta}}/2\right), \tag{8}$$

is a non-negative supermartingale. After that, invoking Ville's inequality (Lemma B.6) and applying a union bound over $h \in \mathcal{H}$ implies the desired result.

Under the (β, B)-Bernstein condition, Lemmas B.1-B.3 imply, for all $\eta \in [0, 1/2]$ and $i \geq 1$,

$$\eta \cdot X_h(Z_i)^2/4 \trianglelefteq_\eta X_h(Z_i) + 3/2\,(2B\eta)^{\frac{1}{1-\beta}}, \tag{9}$$

where we used the fact that $\alpha_\eta = \frac{\eta}{1+\sqrt{1+4\eta^2}} \geq \eta/4$, for all $0 \leq \eta \leq 1/2$ (α_η is involved in Lemma B.1). Now, due to the Bernstein inequality (Lemma B.4), we have for all $\eta \in [0, 1/2]$ and $i \geq 1$,

$$\begin{aligned} X_h(Z_i) &\trianglelefteq_\eta \mathbb{E}_{P(Z)}[X_h(Z)] + \eta \cdot \mathbb{E}_{P(Z)}[X_h(Z)^2], \\ &\trianglelefteq_\eta \mathbb{E}_{P(Z)}[X_h(Z)] + \eta \cdot \mathbb{E}_{P(Z)}[X_h(Z)]^\beta, \quad \text{(by the Bernstein condition)} \\ &\trianglelefteq_\eta 2\mathbb{E}_{P(Z)}[X_h(Z)] + c_\beta^{\frac{\beta}{1-\beta}} \cdot \eta^{\frac{1}{1-\beta}}, \quad \text{where } c_\beta := (1-\beta)^{1-\beta}\beta^\beta. \end{aligned} \tag{10}$$

The last inequality follows by the fact that $z^\beta = c_\beta \cdot \inf_{\nu>0}\{z/\nu + \nu^{\frac{\beta}{1-\beta}}\}$, for $z \geq 0$ (in our case, we set $\nu = c_\beta \eta$ to get to (10)). By chaining (9) with (10) using Proposition B.5-(a), we get:

$$\begin{aligned}\eta \cdot X_h(Z_i)^2/4 &\trianglelefteq_{\frac{\eta}{2}} 2\mathbb{E}_{P(Z)}[X_h(Z)] + c_\beta^{\frac{\beta}{1-\beta}} \cdot \eta^{\frac{1}{1-\beta}} + 3/2(2B\eta)^{\frac{1}{1-\beta}}. \\ &\trianglelefteq_{\frac{\eta}{2}} 2\mathbb{E}_{P(Z)}[X_h(Z)] + C_\beta \cdot \eta^{\frac{1}{1-\beta}}. \end{aligned} \tag{11}$$

Since the random variables $Z_1, \ldots, Z_n$ are i.i.d., (11) implies that M_n in (8) is a non-negative supermartingale. Thus, by Ville's inequality in Lemma B.6, we have, for any $\delta \in (0,1)$ and $h \in \mathcal{H}$,

$$\begin{aligned}\delta &\geq \mathbf{P}[\exists n \geq 1, M_n \geq \delta^{-1}], \\ &= \mathbf{P}\left[\exists n \geq 1,\ \eta \sum_{i=1}^n X_h(Z_i)^2 \geq 8n\mathbb{E}_{P(Z)}[X_h(Z)] + 4nC_\beta \cdot \eta^{\frac{1}{1-\beta}} + \frac{2\ln\delta^{-1}}{\eta} \geq \frac{8\ln\delta^{-1}}{\eta}\right].\end{aligned}$$

From this, a union bound over $h \in \mathcal{H}$ implies the desired result. □

Lemma B.8. *For $A, B > 0$, we have*

$$\inf_{\eta \in (0,1/2)} \left\{A\eta^{\frac{1}{1-\beta}} + B/\eta\right\} \leq \frac{A(3-2\beta)}{1-\beta}\left(\frac{(1-\beta)B}{A}\right)^{\frac{1}{2-\beta}} + 2B. \tag{12}$$

Proof. The unconstrained minimizer of the LHS of (12) is given by

$$\eta_\star := \left(\frac{(1-\beta)B}{A}\right)^{\frac{1-\beta}{2-\beta}}.$$

If $\eta_\star \leq 1/2$, then

$$\inf_{\eta \in (0,1/2]} \left\{A\eta^{\frac{1}{1-\beta}} + B/\eta\right\} \leq A\eta_\star^{\frac{1}{1-\beta}} + B/\eta_\star = \frac{A(2-\beta)}{1-\beta}\left(\frac{(1-\beta)B}{A}\right)^{\frac{1}{2-\beta}}. \tag{13}$$

Now if $\eta_\star > 1/2$, we have $(1/2)^{\frac{1}{1-\beta}} < \left(\frac{(1-\beta)B}{A}\right)^{\frac{1}{2-\beta}}$, and so, we have

$$\begin{aligned}\inf_{\eta \in (0,1/2]} \left\{A\eta^{\frac{1}{1-\beta}} + B/\eta\right\} &\leq A(1/2)^{\frac{1}{1-\beta}} + 2B, \\ &\leq A\left(\frac{(1-\beta)B}{A}\right)^{\frac{1}{2-\beta}} + 2B. \end{aligned} \tag{14}$$

By combining (13) and (14) we get the desired result. □

Bibliography

Agarwal, A.; Krishnamurthy, A.; Langford, J.; Luo, H.; and Schapire, R. E., 2017. Open problem: First-order regret bounds for contextual bandits. *Proceedings of Machine Learning Research vol*, 65 (2017), 1–4. (cited on page 234)

Agarwal, N.; Gonen, A.; and Hazan, E., 2019a. Learning in non-convex games with an optimization oracle. In *Conference on Learning Theory*, 18–29. PMLR. (cited on page 2)

Agarwal, N.; Hazan, E.; and Singh, K., 2019b. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 10175–10184. (cited on page 2)

Agarwal, N.; Hazan, E.; and Singh, K., 2019c. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/78719f11fa2df9917de3110133506521-Paper.pdf. (cited on page 2)

Agrawal, S.; Koolen, W. M.; and Juneja, S., 2020a. Optimal best-arm identification methods for tail-risk measures. *CoRR*, abs/2008.07606 (2020). (cited on page 8)

Agrawal, S.; Koolen, W. M.; and Juneja, S., 2020b. Optimal best-arm identification methods for tail-risk measures. *arXiv*, (Aug. 2020). (cited on page 165)

Ahmadi, M.; Xiong, X.; and Ames, A. D., 2021. Risk-averse planning via cvar barrier functions: Application to bipedal robot locomotion. https://arxiv.org/abs/2011.01578. (cited on page 8)

Artzner, P.; Delbaen, F.; Eber, J.-M.; and Heath, D., 1999. Coherent measures of risk. *Mathematical finance*, 9, 3 (1999), 203–228. (cited on page 163)

Audibert, J.-Y., 2004a. Classification under polynomial entropy and margin assumptions and randomized estimators. (2004). (cited on page 208)

Audibert, J.-Y., 2004b. PAC-Bayesian statistical learning theory. *These de doctorat de l'Université Paris*, 6 (2004), 29. (cited on page 209)

Bartlett, P. L.; Jordan, M. I.; and McAuliffe, J. D., 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101, 473 (2006), 138–156. (cited on page 209)

Bartlett, P. L. and Mendelson, S., 2006a. Empirical minimization. *Probability Theory and Related Fields*, 135, 3 (2006), 311–334. (cited on pages 8 and 202)

Bartlett, P. L. and Mendelson, S., 2006b. Empirical minimization. *Probability Theory and Related Fields*, 135, 3 (2006), 311–334. (cited on pages 202 and 209)

Baudry, D.; Gautron, R.; Kaufmann, E.; and Maillard, O.-A., 2020. Thompson sampling for cvar bandits. *arXiv preprint arXiv:2012.05754*, (2020). (cited on page 165)

Belkin, M.; Hsu, D.; Ma, S.; and Mandal, S., 2018. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, (2018). (cited on page 201)

Belkin, M.; Hsu, D.; and Xu, J., 2019. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, (2019). (cited on pages 9 and 201)

Ben-David, S.; Loker, D.; Srebro, N.; and Sridharan, K., 2012. Minimizing the misclassification error rate using a surrogate convex loss. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 83–90. (cited on page 202)

Ben-David, S.; Srebro, N.; and Urner, R., 2011. Universal learning vs. no free lunch results. In *Philosophy and Machine Learning Workshop NIPS*. (cited on page 203)

Bhat, S. P. and Prashanth, L., 2019. Concentration of risk measures: A Wasserstein distance approach. In *Advances in Neural Information Processing Systems*, 11739–11748. (cited on pages 143, 165, and 168)

Boucheron, S.; Bousquet, O.; and Lugosi, G., 2005. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9 (2005), 323–375. (cited on pages 202, 207, and 208)

Bousquet, O.; Boucheron, S.; and Lugosi, G., 2004. *Introduction to Statistical Learning Theory*. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-28650-9. doi:10.1007/978-3-540-28650-9_8. https://doi.org/10.1007/978-3-540-28650-9_8. (cited on pages 7, 202, 207, and 210)

Bousquet, O. and Elisseeff, A., 2002. Stability and generalization. *The Journal of Machine Learning Research*, 2 (2002), 499–526. (cited on pages 1 and 7)

Brown, D. B., 2007. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35, 6 (2007), 722–730. (cited on pages 8, 164, 165, 173, 178, and 183)

Bubeck, S.; Cesa-Bianchi, N.; and Lugosi, G., 2013. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59, 11 (2013), 7711–7717. (cited on pages 168 and 200)

Bubeck, S.; Devanur, N. R.; Huang, Z.; and Niazadeh, R., 2017. Online auctions and multi-scale online learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17 (Cambridge, Massachusetts, USA, 2017), 497–514. Association for Computing Machinery, New York, NY, USA. (cited on page 5)

Cassel, A. and Koren, T., 2020. Bandit linear control. *arXiv preprint arXiv:2007.00759*, (2020). (cited on page 2)

Cassel, A.; Mannor, S.; and Zeevi, A., 2018. A general approach to multi-armed bandits under risk criteria. In *Proceedings of the 31st Conference On Learning Theory*, vol. 75 of *Proceedings of Machine Learning Research*, 1295–1306. PMLR. (cited on page 166)

Cesa-Bianchi, N.; Conconi, A.; and Gentile, C., 2004. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50, 9 (2004), 2050–2057. (cited on page 2)

Cesa-Bianchi, N. and Lugosi, G., 2006. *Prediction, learning, and games*. Cambridge university press. (cited on pages 2 and 241)

Cesa-Bianchi, N.; Mansour, Y.; and Stoltz, G., 2007. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66, 2-3 (2007), 321–352. (cited on page 61)

Cheema, P. and Sugiyama, M., 2020. A geometric look at double descent risk: Volumes, singularities, and distinguishabilities. *arXiv preprint arXiv:2006.04366*, (2020). (cited on page 201)

Chen, L.; Min, Y.; Belkin, M.; and Karbasi, A., 2020. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, (2020). (cited on pages 9 and 201)

Chen, Y. F.; Xu, M.; and Zhang, Z. G., 2009. A risk-averse newsvendor model under the CVaR criterion. *Operations Research*, 57, 4 (2009), 1040–1044. (cited on page 143)

Chow, Y. and Ghavamzadeh, M., 2014. Algorithms for CVaR Optimization in MDPs. In *Advances in Neural Information Processing Systems 27*, 3509–3517. (cited on page 143)

Chow, Y.; Tamar, A.; Mannor, S.; and Pavone, M., 2015. Risk-sensitive and robust decision-making: a cvar optimization approach. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 1522–1530. (cited on page 8)

Curi, S.; Levy, K. Y.; Jegelka, S.; and Krause, A., 2020. Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (cited on page 165)

Cutkosky, A., 2019a. Anytime online-to-batch, optimism and acceleration. In *International Conference on Machine Learning*, 1446–1454. (cited on pages 2, 6, 85, 214, and 234)

Cutkosky, A., 2019b. Combining online learning guarantees. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, 895–913. (cited on pages 5 and 7)

Cutkosky, A., 2020a. Better full-matrix regret via parameter-free online learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (cited on page 6)

Cutkosky, A., 2020b. Parameter-free, dynamic, and strongly-adaptive online learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 2250–2259. (cited on page 5)

Cutkosky, A. and Boahen, K., 2017. Online learning without prior information. In *Conference on Learning Theory*, 643–677. (cited on page 6)

Cutkosky, A. and Orabona, F., 2018. Black-box reductions for parameter-free online learning in Banach spaces. In *Conference On Learning Theory, COLT 2018*, 1493–1529. (cited on pages 4, 5, and 6)

Cutkosky, A. and Sarlós, T., 2019. Matrix-free preconditioning in online learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 1455–1464. (cited on page 6)

d'Ascoli, S.; Sagun, L.; and Biroli, G., 2020. Triple descent and the two kinds of overfitting: Where & why do they appear? *arXiv preprint arXiv:2006.03509*, (2020). (cited on page 201)

Deng, Z.; Kammoun, A.; and Thrampoulidis, C., 2019. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, (2019). (cited on page 201)

Derezinski, M.; Liang, F. T.; and Mahoney, M. W., 2020. Exact expressions for double descent and implicit regularization via surrogate random design. *Advances in Neural Information Processing Systems*, 33 (2020). (cited on pages 9 and 201)

Dereziński, M.; Liang, F.; and Mahoney, M. W., 2019. Exact expressions for double descent and implicit regularization via surrogate random design. (cited on page 201)

Dudley, R. M., 1984. A course on empirical processes. In *Ecole d'été de Probabilités de Saint-Flour XII-1982*, 1–142. Springer. (cited on pages 202 and 208)

Duin, R. P., 1995. Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, vol. 2, 957–964. (cited on page 201)

Duin, R. P., 2000. Classifiers in almost empty spaces. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 2, 1–7. IEEE. (cited on page 201)

Dziugaite, G. K.; Drouin, A.; Neal, B.; Rajkumar, N.; Caballero, E.; Wang, L.; Mitliagkas, I.; and Roy, D. M., 2020. In search of robust measures of generalization. *arXiv preprint arXiv:2010.11924*, (2020). (cited on page 1)

Erven, T. V.; Mehta, N. A.; Reid, M. D.; and Williamson, R. C., 2015. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16 (2015), 1793–1861. (cited on page 209)

Feurer, M. and Hutter, F., 2019. Hyperparameter optimization. In *Automated Machine Learning*, 3–33. Springer, Cham. (cited on page 1)

Foster, D. J.; Kale, S.; Mohri, M.; and Sridharan, K., 2017. Parameter-free online learning via model selection. In *Advances in Neural Information Processing Systems*, 6020–6030. (cited on pages 4 and 5)

Foster, D. J.; Krishnamurthy, A.; and Luo, H., 2020. Open problem: Model selection for contextual bandits. vol. 125 of *Proceedings of Machine Learning Research*, 3842–3846. PMLR. (cited on page 234)

Foster, D. J. and Simchowitz, M., 2020. Logarithmic regret for adversarial online control. *arXiv preprint arXiv:2003.00189*, (2020). (cited on page 2)

Freedman, D. A., 1975. On tail probabilities for martingales. *the Annals of Probability*, (1975), 100–118. (cited on pages 6, 10, 203, and 212)

Gaillard, P. and and, G. S., 2014. A second-order bound with excess losses. In *Proc. of the 27th Annual Conference on Learning Theory (COLT)*, 176–196. (cited on page 3)

Galichet, N.; Sebag, M.; and Teytaud, O., 2013. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, 245–260. PMLR. (cited on page 165)

Grünwald, P. D. and Kotłowski, W., 2011. Bounds on individual risk for log-loss predictors. In *Proceedings of the 24th Annual Conference on Learning Theory*, 813–816. (cited on page 231)

Guedj, B., 2019. A primer on pac-bayesian learning. *CoRR*, abs/1901.05353 (2019). (cited on page 8)

Hardt, M.; Recht, B.; and Singer, Y., 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 1225–1234. (cited on page 7)

Hazan, E., 2016a. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2, 3–4 (2016), 157–325. (cited on page 2)

Hazan, E., 2016b. Introduction to online convex optimization. *Found. Trends Optim.*, 2, 3–4 (Aug. 2016), 157–325. doi:10.1561/2400000013. https://doi.org/10.1561/2400000013. (cited on pages 2, 3, and 214)

Hazan, E.; Agarwal, A.; and Kale, S., 2007. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69, 2-3 (2007), 169–192. (cited on pages 3 and 61)

Héliou, A.; Martin, M.; Mertikopoulos, P.; and Rahier, T., 2020. Online non-convex optimization with imperfect feedback. *arXiv preprint arXiv:2010.08496*, (2020). (cited on page 2)

Howard, S. R.; Ramdas, A.; McAuliffe, J.; and Sekhon, J., 2018. Time-uniform, non-parametric, nonasymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, (2018). (cited on page 212)

Howard, S. R.; Ramdas, A.; McAuliffe, J.; Sekhon, J.; et al., 2020. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17 (2020), 257–317. (cited on pages 185 and 241)

Huo, X. and Fu, F., 2017. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society Open Science*, 4, 11 (2017). (cited on page 143)

Jamieson, K. G.; Malloy, M.; Nowak, R. D.; and Bubeck, S., 2014. lil' UCB : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, vol. 35 of *JMLR Workshop and Conference Proceedings*, 423–439. JMLR.org. (cited on pages 164, 166, 178, 179, 191, and 194)

Jun, K. and Orabona, F., 2019a. Parameter-free locally differentially private stochastic subgradient descent. *CoRR*, abs/1911.09564 (2019). http://arxiv.org/abs/1911.09564. (cited on page 4)

Jun, K.-S. and Orabona, F., 2019b. Parameter-free online convex optimization with sub-exponential noise. In *Conference on Learning Theory*, 1802–1823. (cited on pages 2, 212, 213, 216, 217, 233, and 234)

Kagrecha, A.; Nair, J.; and Jagannathan, K., 2019a. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc. (cited on page 165)

Kagrecha, A.; Nair, J.; and Jagannathan, K. P., 2019b. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 11269–11278. (cited on page 165)

Kalai, A. and Vempala, S. S., 2003. Efficient algorithms for online decision problems. In *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, 26–40. (cited on page 2)

Karnin, Z.; Koren, T.; and Somekh, O., 2013. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning (ICML)*. (cited on page 180)

Kempka, M.; Kotlowski, W.; and Warmuth, M. K., 2019a. Adaptive scale-invariant online algorithms for learning linear models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, 3321–3330. (cited on page 4)

Kempka, M.; Kotłowski, W.; and Warmuth, M. K., 2019b. Adaptive scale-invariant online algorithms for learning linear models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, 3321–3330. (cited on pages 61 and 85)

Khani, F.; Raghunathan, A.; and Liang, P., 2019. Maximum weighted loss discrepancy. *arXiv preprint arXiv:1906.03518*, (2019). (cited on page 163)

Kolla, R. K.; Prashanth, L. A.; Bhat, S. P.; and Jagannathan, K., 2019a. Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47, 1 (2019), 16–20. (cited on page 8)

Kolla, R. K.; Prashanth, L. A.; Bhat, S. P.; and Jagannathan, K., 2019b. Concentration bounds for empirical conditional value-at-risk: The unbounded case. *Operations Research Letters*, 47, 1 (2019), 16–20. (cited on page 165)

Koolen, W. M.; Grünwald, P. D.; and Van Erven, T., 2016. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems*, 4457–4465. (cited on pages 209, 240, and 241)

Kotlowski, W., 2017. Scale-invariant unconstrained online learning. In *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, 412–433. (cited on page 4)

Kotłowski, W., 2017. Scale-invariant unconstrained online learning. In *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, 412–433. (cited on pages 61 and 85)

Krämer, N., 2009. On the peaking phenomenon of the lasso in model selection. *arXiv preprint arXiv:0904.4416*, (2009). (cited on page 201)

Kusuoka, S., 2001. On law invariant coherent risk measures. In *Advances in mathematical economics*, 83–95. Springer. (cited on pages 8 and 163)

Laforgue, P.; Clémenccon, S.; and Bertail, P., 2019. On medians of (randomized) pairwise means. In *International Conference on Machine Learning*, 1272–1281. PMLR. (cited on page 163)

Langford, J. and Shawe-Taylor, J., 2003. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, 439–446. (cited on page 143)

Lattimore, T. and Szepesvári, C., 2019. *Bandit Algorithms*. Cambridge University Press. (cited on page 181)

Lee, J.; Park, S.; and Shin, J., 2020. Learning bounds for risk-sensitive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (cited on page 165)

Lerasle, M.; Szabó, Z.; Mathieu, T.; and Lecué, G., 2019. Monk outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning*, 3782–3793. PMLR. (cited on pages 8 and 163)

Loog, M., 2015. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE transactions on pattern analysis and machine intelligence*, 38, 3 (2015), 462–475. (cited on pages 9 and 202)

Loog, M. and Duin, R. P., 2012. The dipping phenomenon. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 310–317. Springer. (cited on pages 9 and 202)

Loog, M.; Viering, T.; and Mey, A., 2019. Minimizers of the empirical risk and risk monotonicity. In *Advances in Neural Information Processing Systems*, 7476–7485. (cited on pages 9, 201, 202, 203, and 205)

Luo, H. and Schapire, R. E., 2015. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, 1286–1304. PMLR. (cited on page 3)

Lykouris, T.; Sridharan, K.; and Tardos, É., 2018. Small-loss bounds for online learning with partial information. In *Conference on Learning Theory*, 979–986. (cited on pages 2 and 234)

Massart, P. and Nédélec, É., 2006. Risk bounds for statistical learning. *The Annals of Statistics*, 34, 5 (2006), 2326–2366. (cited on page 209)

Maurer, A., 2004. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, (2004). (cited on page 143)

Maurer, A. and Pontil, M., 2009. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, (2009). (cited on pages 10, 117, 170, 185, 187, 188, 203, 206, 219, 222, and 223)

McAllester, D. A., 2003. Pac-bayesian stochastic model selection. *Machine Learning*, 51, 1 (2003), 5–21. (cited on pages 9 and 199)

McDiarmid, C., 1989. *On the method of bounded differences*, 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press. doi:10.1017/CBO9781107359949.008. (cited on page 215)

McMahan, B., 2011. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 525–533. JMLR Workshop and Conference Proceedings. (cited on page 3)

McMahan, B. and Abernethy, J., 2013. Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems*, 2724–2732. (cited on page 4)

Mcmahan, B. and Streeter, M., 2012. No-regret algorithms for unconstrained online convex optimization. In *Advances in neural information processing systems*, 2402–2410. (cited on page 4)

McMahan, H. B. and Orabona, F., 2014. Unconstrained online linear learning in Hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, 1020–1039. (cited on page 4)

McMahan, H. B. and Streeter, M. J., 2010. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 244–256. (cited on page 4)

Mehta, N., 2017. Fast rates with high probability in exp-concave statistical learning. In *Artificial Intelligence and Statistics*, 1085–1093. PMLR. (cited on pages 202 and 209)

Mei, S. and Montanari, A., 2019. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, (2019). (cited on pages 9 and 201)

Mhammedi, Z.; Foster, D. J.; Simchowitz, M.; Misra, D.; Sun, W.; Krishnamurthy, A.; Rakhlin, A.; and Langford, J., 2020a. Learning the linear quadratic regulator from nonlinear observations. *arXiv preprint arXiv:2010.03799*, (October 2020). (cited on page 234)

Mhammedi, Z.; Grünwald, P.; and Guedj, B., 2019a. Pac-bayes un-expected bernstein inequality. In *Advances in Neural Information Processing Systems*, 12202–12213. (cited on pages 208 and 241)

Mhammedi, Z.; Guedj, B.; and Williamson, R. C., 2020b. PAC-Bayesian bound for the conditional value at risk. *arXiv:2006.14763 [cs, stat]*, (June 2020). (cited on page 9)

Mhammedi, Z.; Guedj, B.; and Williamson, R. C., 2020c. Pac-bayesian bound for the conditional value at risk. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (cited on pages 165, 166, 171, 173, and 199)

Mhammedi, Z. and Koolen, W. M., 2020. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory, COLT 2020*, vol. 125, 2858–2887. (cited on pages 4, 203, 212, 216, and 217)

Mhammedi, Z.; Koolen, W. M.; and Van Erven, T., 2019b. Lipschitz adaptivity with multiple learning rates in online learning. In *Conference on Learning Theory, COLT 2019*, 2490–2511. (cited on pages 61 and 85)

Morimura, T.; Sugiyama, M.; Kashima, H.; Hachiya, H.; and Tanaka, T., 2010. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 799–806. (cited on page 143)

Nagarajan, V. and Kolter, J. Z., 2019. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems 32*, 11615–11626. (cited on pages 117 and 234)

Nakkiran, P., 2019. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, (2019). (cited on pages 9 and 201)

Nakkiran, P.; Kaplun, G.; Bansal, Y.; Yang, T.; Barak, B.; and Sutskever, I., 2020a. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1g5sA4twr. (cited on pages 9 and 201)

Nakkiran, P.; Venkat, P.; Kakade, S.; and Ma, T., 2020b. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, (2020). (cited on pages 9, 201, and 203)

Neu, G.; Jonsson, A.; and Gómez, V., 2017. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, (2017). (cited on page 2)

Neyshabur, B.; Tomioka, R.; and Srebro, N., 2015. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, 1376–1401. PMLR. (cited on page 207)

Opper, M., 2001. Learning to generalize. *Frontiers of Life, 3(part 2), pp.763-775.*, (2001). (cited on page 201)

Opper, M. and Kinzel, W., 1996. Statistical mechanics of generalization. In *Models of neural networks III*, 151–209. Springer. (cited on page 201)

Orabona, F., 2013. Dimension-free exponentiated gradient. In *Advances in Neural Information Processing Systems*, 1806–1814. (cited on page 4)

Orabona, F., 2014a. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, 1116–1124. (cited on page 4)

Orabona, F., 2014b. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 1116–1124. (cited on page 4)

Orabona, F. and Cutkosky, A., 2020. Parameter-free learning and optimization algorithms. In *ICML Tutorial*. ICML. https://parameterfree.com/icml-tutorial/. (cited on page 234)

Orabona, F. and Pál, D., 2016a. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems*, 577–585. (cited on page 4)

Orabona, F. and Pál, D., 2016b. Open problem: Parameter-free and scale-free online algorithms. In *Conference on Learning Theory*, 1659–1664. (cited on pages 6, 61, and 85)

Orabona, F. and Pál, D., 2021. Parameter-free stochastic optimization of variationally coherent functions. *CoRR*, abs/2102.00236 (2021). (cited on pages 5 and 6)

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A., 2017. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2817–2826. (cited on page 143)

Prashanth, L. and Ghavamzadeh, M., 2013. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in neural information processing systems*, 252–260. (cited on pages 8, 143, and 164)

Prashanth, L. A.; Jagannathan, K. P.; and Kolla, R. K., 2020. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, vol. 119 of *Proceedings of Machine Learning Research*, 5577–5586. PMLR. (cited on pages 165, 168, and 175)

Reid, M., 2017. Generalization bounds. In *Encyclopedia of Machine Learning and Data Mining*, 556. (cited on page 7)

Reid, M. D.; Frongillo, R. M.; Williamson, R. C.; and Mehta, N., 2015. Generalized mixability via entropic duality. In *Conference on Learning Theory*, 1501–1522. (cited on pages 4 and 11)

Rockafellar, R. T., 1997. Convex analysis. Princeton University Press, Princeton, NJ. (cited on page 8)

Rockafellar, R. T. and Uryasev, S., 2013. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18, 1-2 (2013), 33–53. (cited on pages 167, 176, 177, and 184)

Ross, S.; Mineiro, P.; and Langford, J., 2013a. Normalized online learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 537–545. (cited on page 61)

Ross, S.; Mineiro, P.; and Langford, J., 2013b. Normalized online learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. (cited on page 85)

Shalev-Shwartz, S., 2012. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4, 2 (Feb. 2012), 107–194. (cited on page 2)

Shalev-Shwartz, S. and Ben-David, S., 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press. (cited on page 201)

Simchowitz, M., 2020. Making non-stochastic control (almost) as easy as stochastic. *arXiv preprint arXiv:2006.05910*, (2020). (cited on page 2)

Smola, A. J.; Bartlett, P. J.; Schuurmans, D.; and Schölkopf, B., 2000. *Advances in large margin classifiers*. MIT press. (cited on page 201)

Soma, T. and Yoshida, Y., 2020. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, (2020). (cited on page 165)

Spigler, S.; Geiger, M.; d'Ascoli, S.; Sagun, L.; Biroli, G.; and Wyart, M., 2018. A jamming transition from under-to over-parametrization affects loss landscape and generalization. *arXiv preprint arXiv:1810.09665*, (2018). (cited on page 201)

Srebro, N.; Sridharan, K.; and Tewari, A., 2010. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, 2199–2207. (cited on page 6)

Suggala, A. S. and Netrapalli, P., 2020. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, 845–861. PMLR. (cited on page 2)

Takeda, A. and Kanamori, T., 2009. A robust approach based on conditional value-at-risk measure to statistical learning problems. *European Journal of Operational Research*, 198, 1 (2009), 287 – 296. (cited on page 143)

Talagrand, M., 1994. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, (1994), 28–76. (cited on pages 202 and 208)

Tamar, A.; Glassner, Y.; and Mannor, S., 2015. Optimizing the CVaR via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. (cited on page 143)

Tamkin, A.; Keramati, R.; Dann, C.; and Brunskill, E., 2019. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop*. (cited on page 165)

Thomas, P. and Learned-Miller, E., 2019a. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, 6225–6233. PMLR. (cited on page 8)

Thomas, P. and Learned-Miller, E., 2019b. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, 6225–6233. PMLR. (cited on page 165)

Torossian, L.; Garivier, A.; and Picheny, V., 2019. $\mathcal{X}$-armed bandits: Optimizing quantiles, cvar and other risks. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, vol. 101 of *Proceedings of Machine Learning Research*, 252–267. PMLR, Nagoya, Japan. (cited on page 165)

Van der Hoeven, D., 2019a. User-specified local differential privacy in unconstrained adaptive online learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 14080–14089. (cited on page 2)

Van der Hoeven, D., 2019b. User-specified local differential privacy in unconstrained adaptive online learning. In *Advances in Neural Information Processing Systems*, 14103–14112. (cited on pages 4, 233, and 234)

Van der Hoeven, D.; Van Erven, T.; and Kotłowski, W., 2018. The many faces of exponential weights in online learning. In *Conference On Learning Theory*, 2067–2092. (cited on page 3)

Van Erven, T.; Grünwald, P. D.; Mehta, N. A.; Reid, M. D.; and Williamson, R. C., 2015. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16 (2015), 1793–1861. (cited on page 202)

Van Erven, T. and Koolen, W. M., 2016. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 3666–3674. (cited on pages 3 and 61)

Van Erven, T.; Koolen, W. M.; and van der Hoeven, D., 2021. Metagrad: Adaptation using multiple learning rates in online learning. *CoRR*, abs/2102.06622 (2021). https://arxiv.org/abs/2102.06622. (cited on pages 3 and 61)

Viering, T. and Loog, M., 2021. The shape of learning curves: a review. *arXiv preprint arXiv:2103.10948*, (2021). (cited on page 9)

Viering, T.; Mey, A.; and Loog, M., 2019a. Open problem: Monotonicity of learning. In *Conference on Learning Theory*, 3198–3201. (cited on pages 9, 202, 203, 205, and 211)

Viering, T. J.; Mey, A.; and Loog, M., 2019b. Making learners (more) monotone. (cited on pages 9, 203, and 211)

Vovk, V., 1998. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56, 2 (1998), 153–173. (cited on pages 4 and 11)

Walder, C. and Nock, R., 2020. All your loss are belong to bayes. (cited on page 235)

Wang, Y. and Gao, F., 2010. Deviation inequalities for an estimator of the conditional value-at-risk. *Operations research letters*, 38, 3 (2010), 236–239. (cited on pages 8, 164, 165, and 173)

Williamson, R. and Menon, A., 2019. Fairness risk measures. In *International Conference on Machine Learning*, 6786–6797. (cited on pages 8, 143, 163, and 235)

Wintenberger, O., 2017. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106, 1 (2017), 119–141. (cited on pages 61 and 85)

Zhivotovskiy, N. and Hanneke, S., 2018. Localization of vc classes: Beyond local rademacher complexities. *Theoretical Computer Science*, 742 (2018), 27–49. (cited on pages 202 and 208)

Zinkevich, M., 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, 928–936. (cited on page 3)