Generalized Mixability via Entropic Duality

Mark D. Reid
Australian National University & NICTA

Rafael M. Frongillo
Microsoft Research

Robert C. Williamson
Australian National University & NICTA

Nishant Mehta
NICTA

June 25, 2014

Abstract

Mixability is a property of a loss which characterizes when fast convergence is possible in the game of prediction with expert advice. We show that a key property of mixability generalizes, and the exp and log operations present in the usual theory are not as special as one might have thought. In doing this we introduce a more general notion of Φ-mixability where Φ is a general entropy (*i.e.*, any convex function on probabilities). We show how a property shared by the convex dual of any such entropy yields a natural algorithm (the minimizer of a regret bound) which, analogous to the classical aggregating algorithm, is guaranteed a constant regret when used with Φ-mixable losses. We characterize precisely which Φ have Φ-mixable losses and put forward a number of conjectures about the optimality and relationships between different choices of entropy.

1 Introduction

The combination or aggregation of predictions is central to machine learning. Traditional Bayesian updating can be viewed as a particular way of aggregating information that takes account of prior information. Notions of "mixability" which play a key role in the setting of prediction with expert advice offer a more general way to aggregate by taking into account a loss function to evaluate predictions. As shown by Vovk [1], his more general "aggregating algorithm" reduces to Bayesian updating when log loss is used. However there is an implicit design variable in mixability that to date has not been fully exploited. The aggregating algorithm makes use of a distance between the current distribution and a prior which serves as a regularizer. In particular the aggregating algorithm uses the KL-divergence. We consider the general setting of an arbitrary loss and an arbitrary regularizer (in the form of a Bregman divergence) and show that we recover the core technical result of traditional mixability: if a loss is mixable in our generalized sense then there is a generalized aggregating algorithm which can be guaranteed to have constant regret. The generalized aggregating algorithm is developed by optimizing the bound that defines our new notion of mixability. Our approach relies heavily on dual representations of entropy functions defined on the probability simplex (hence the title). By doing so we gain new insight into why the original mixability argument works and a broader understanding of when constant regret guarantees are possible.

1.1 Mixability in Prediction With Expert Advice Games

A prediction with expert advice game is defined by its loss, a collection of experts that the player must compete against, and a fixed number of rounds. Each round the experts reveal their predictions to the player and then the player makes a prediction. An observation is then revealed to the experts and the player and all receive a penalty determined by the loss. The aim of the player is to keep its total loss close to that of the best expert once all the rounds have completed. The difference between the total loss of the player and the total loss of the best expert is called the regret and is the typically the focus of the analysis of this style of game. In particular, we are interested in when the regret is *constant*, that is, independent of the number of rounds played.

More formally, let X denote a set of possible *observations* and let $\mathcal{A}$ denote a set of *actions* or *predictions* the experts and player can perform. A *loss* $\ell : \mathcal{A} \to \mathbb{R}^X$ assigns the penalty $\ell_x(a)$ to predicting $a \in \mathcal{A}$ when $x \in X$ is observed. The set of experts is denoted Θ and the set of distributions over Θ is denoted Δ_Θ. In each round $t = 1, \ldots, T$, each expert $\theta \in \Theta$ makes a prediction $a_\theta^t \in \mathcal{A}$. These are revealed to the player who makes a prediction $\hat{a}^t \in \mathcal{A}$. Once observation $x^t \in X$ is revealed the experts receive loss $\ell_{x^t}(a_\theta^t)$ and the player receives loss $\ell_{x^t}(\hat{a}^t)$. The aim of the player is to minimize its *regret* $\mathrm{Regret}(T) := L^T - \min_\theta L_\theta^T$ where $L^T := \sum_{t=1}^T \ell_{x^t}(\hat{a}^t)$ and $L_\theta^T = \sum_{t=1}^T \ell_{x^t}(a_\theta^t)$. We will say the game has *constant regret* if there exists a player who can always make predictions that guarantee $\mathrm{Regret}(T) \leq R_{\ell,\Theta}$ for all T and all expert predictions $\{a_\theta^t\}_{t=1}^T$ where $R_{\ell,\Theta}$ is a constant that may depend on ℓ and Θ.

In [2, 3], Vovk showed that if the loss for a game satisfies a condition called mixability then a player making predictions using the aggregating algorithm (AA) will achieve constant regret.

Definition 1 (Mixability and the Aggregating Algorithm)**.** *Given $\eta > 0$, a loss $\ell : \mathcal{A} \to \mathbb{R}^X$ is* η-mixable *if, for all expert predictions $a_\theta \in \mathcal{A}$, $\theta \in \Theta$ and all mixture distributions $\mu \in \Delta_\Theta$ over experts there exists a prediction $\hat{a} \in \mathcal{A}$ such that for all outcomes $x \in X$ we have*

$$\ell_x(\hat{a}) \leq -\eta^{-1} \log \sum_{\theta \in \Theta} \exp\left(-\eta \ell_x(a_\theta)\right) \mu_\theta. \tag{1}$$

The aggregating algorithm *starts with a mixture $\mu^0 \in \Delta_\Theta$ over experts. In round t, experts predict a_θ^t and the player predicts the $\hat{a}^t \in \mathcal{A}$ guaranteed by the η-mixability of ℓ so that* (1) *holds for $\mu = \mu^{t-1}$ and $a_\theta = a_\theta^t$. Upon observing x^t, the mixture $\mu^t \in \Delta_\Theta$ is set so that $\mu_\theta^t \propto \mu_\theta^{t-1} e^{-\eta \ell_{x^t}(a_\theta^t)}$.*

Mixability can be seen as a weakening of exp-concavity (see [4, §3.3]) that requires just enough of the loss to ensure constant regret.

Theorem 1 (Mixability implies constant regret [3])**.** *If a loss ℓ is η-mixable then the aggregating algorithm will achieve* $\mathrm{Regret}(T) = \eta^{-1} \log |\Theta|$.

1.2 Contributions

The key contributions of this paper are as follows. We provide a new general definition (Definition 2) of mixability and an induced generalized aggregating algorithm (Definition 3) and show (Theorem 2) that prediction with expert advice using a Φ-mixable loss and the associated generalized aggregating algorithm is guaranteed to have constant regret. The proof illustrates that the log and exp functions that arise in the classical aggregating algorithm are themselves not special, but rather it is a translation invariant property of the convex conjugate of and entropy Φ defined on a probability simplex that is the crucial property that leads to constant regret.

We characterize (Theorem 4) for which entropies Φ there exists Φ-mixable losses via the Legendre property. We show that Φ-mixability of a loss can be expressed directly in terms of the Bayes risk associated with the loss (Definition 4 and Theorem 3), reflecting the situation that holds for classical mixability [5]. As part of this analysis we show that proper losses are quasi-convex (Lemma 6) which, to the best of our knowledge appears to be a new result.

1.3 Related Work

The starting point for mixability and the aggregating algorithm is the work of [3, 2]. The general setting of prediction with expert advice is summarized in [4, Chapters 2 and 3]. There one can find a range of results that study different aggregation schemes and different assumptions on the losses (exp-concave, mixable). Variants of the aggregating algorithm have been studied for classically mixable losses, with a trade-off between tightness of the bound (in a constant factor) and the computational complexity [6]. Weakly mixable losses are a generalization of mixable losses. They have been studied in [7] where it is shown there exists a variant of the aggregating algorithm that achieves regret $C\sqrt{T}$ for some constant C. Vovk [1, in §2.2] makes the observation that his Aggregating Algorithm reduces to Bayesian mixtures in the case of the log loss game. See also the discussion in [4, page 330] relating certain aggregation schemes to Bayesian updating.

The general form of updating we propose is similar to that considered by Kivinen and Warmuth [8] who consider finding a vector w minimizing $d(w, s) + \eta L(y_t, w \cdot x_t)$ where s is some starting vector, (x_t, y_t) is the instance/label observation at round t and L is a loss. The key difference between their formulation and ours is that our loss term is (in their notation) $w \cdot L(y_t, x_t)$ – *i.e.*, the linear combination of the losses of the x_t at y_t and not the loss of their inner product. Online methods of density estimation for exponential families are discussed in [9, §3] where the authors compare the online and offline updates of the same sequence and make heavy use of the relationship between the KL divergence between members of an exponential family and an associated Bregman divergence between the parameters of those members. The analysis of mirror descent [10] shows that it achieves constant regret when the entropic regularizer is used. However, there is no consideration regarding whether similar results extend to other entropies defined on the simplex.

We stress that the idea of the more general regularization and updates is hardly new. See for example the discussion of potential based methods in [4] and other references later in the paper. The key novelty is the generalized notion of mixability, the name of which is justified by the key new technical result — a constant regret bound assuming the general mixability condition achieved via a generalized algorithm which can be seen as intimately related to mirror descent. Crucially, our result depends on some properties of the conjugates of potentials defined over probabilities that do not hold for potential functions defined over more general spaces.

2 Generalized Mixability and Aggregation via Convex Duality

In this section we introduce our generalizations of mixability and the aggregating algorithm. One feature of our approach is the way the generalized aggregating algorithm falls out of the definition of generalized mixability as the minimizer of the mixability bound. Our approach relies on concepts and results from convex analysis. Terms not defined below can be found in a reference such as [11].

2.1 Definitions and Notation

A convex function $\Phi : \Delta_\Theta \to \mathbb{R}$ is called an *entropy* (on Δ_Θ) if it is proper (*i.e.*, $-\infty < \Phi \neq +\infty$), convex[1], and lower semi-continuous. In the following example and elsewhere we use $\mathbb{1}$ to denote the vector $\mathbb{1}_\theta = 1$ for all $\theta \in \Theta$ so that $|\Theta|^{-1}\mathbb{1} \in \Delta_\Theta$ is the uniform distribution over Θ.

Example 1 (Entropies)**.** *The* (negative) Shannon entropy $H(\mu) := \sum_\theta \mu_\theta \log \mu_\theta$*; the* quadratic entropy $Q(\mu) := \sum_\theta (\mu - |\Theta|^{-1}\mathbb{1})^2$*; the* Tsallis entropies $S_\alpha(\mu) := \alpha^{-1}\left(\sum_\theta \mu_\theta^{\alpha+1} - 1\right)$ *for* $\alpha \in (-1, 0) \cup (0, \infty)$*; and the* Rényi entropies $R_\alpha(\mu) = \alpha^{-1}\left(\log \sum_\theta \mu_\theta^{\alpha+1}\right)$*, for* $\alpha \in (-1, 0)$*. We note that both Tsallis and Rényi entropies limit to Shannon entropy* $\alpha \to 0$ *(cf. [12, 13]).*

Let $\langle \mu, v\rangle$ denote the inner product between $\mu \in \Delta_\Theta$ and $v \in \Delta_\Theta^*$, the dual space of Δ_Θ. The *Bregman divergence* associated with a suitably differentiable entropy Φ on Δ_Θ is given by

$$D_\Phi(\mu, \mu') = \Phi(\mu) - \Phi(\mu') - \langle \mu - \mu', \nabla\Phi(\mu')\rangle \tag{2}$$

for all $\mu \in \Delta_\Theta$ and $\mu' \in \mathrm{ri}(\Delta_\Theta)$, the relative interior of Δ_Θ. Given an entropy $\Phi : \Delta_\Theta \to \mathbb{R}$, we define its *entropic dual* to be $\Phi^*(v) := \sup_{\mu \in \Delta_\Theta} \langle \mu, v\rangle - \Phi(\mu)$ where $v \in \Delta_\Theta^*$, *i.e.*, the dual space to Δ_Θ. Note that one could also write the supremum over $\mathbb{R}^\Theta$ by setting $\Phi(\mu) = +\infty$ for $\mu \notin \Delta_\Theta$ so that Φ^* is just the usual convex dual

[1] While the information theoretic notion of Shannon entropy as a measure of uncertainty is concave, it is convenient for us to work with convex functions on the simplex which can be thought of as certainty measures.

(cf. [11]). Thus, all of the standard results about convex duality also hold for entropic duals provided some care is taken with the domain of definition. We note that although the regular convex dual of H defined over all of $\mathbb{R}^\Theta$ is $v \mapsto \sum_\theta \exp(v_\theta - 1)$ its entropic dual is $H^*(v) = \log \sum_\theta \exp(v_\theta)$.

For differentiable Φ, it is known [11] that the supremum defining Φ^* is attained at $\mu = \nabla\Phi^*(v)$. That is,

$$\Phi^*(v) = \langle \nabla\Phi^*(v), v\rangle - \Phi(\nabla\Phi^*(v)). \tag{3}$$

A similar result holds for Φ by applying this result to Φ^* and using $\Phi = (\Phi^*)^*$. We will make repeated use of two easy established properties of entropic duals (see Appendix A.1 for proof).

Lemma 1. *If Φ is an entropy over Δ_Θ and $\Phi_\eta := \eta^{-1}\Phi$ denotes a scaled version of Φ then 1) for all $\eta > 0$ we have $\Phi^*_\eta(v) = \eta^{-1}\Phi^*(\eta v)$; and 2) the entropic dual Φ^* is* translation invariant *– i.e., for all $v \in \Delta^*_\Theta$ and $\alpha \in \mathbb{R}$ we have $\Phi^*(v+\alpha\mathbb{1}) = \Phi^*(v)+\alpha$ and hence for differentiable Φ^* we have $\nabla\Phi^*(v + \alpha\mathbb{1}) = \nabla\Phi^*(v)$.*

The translation invariance if Φ^* is central to our analysis. It is what ensures our Φ-mixability inequality (4) "telescopes" when it is summed. The proof of the original mixability result (Theorem 1) uses a similar telescoping argument that works due to the interaction of $\log$ and $\exp$ terms in Definition 1. Our results show that this telescoping property is not due to any special properties of $\log$ and $\exp$, but rather because of the translation invariance of the entropic dual of Shannon entropy, H. The following analysis generalizes that of the original work on mixability precisely because this property holds for the dual of any entropy.

2.2 Φ-Mixability and the Generalized Aggregating Algorithm

For convenience, we will use $A \in \mathcal{A}^\Theta$ to denote a collection of expert predictions and $A_\theta \in \mathcal{A}$ to denote the prediction of expert θ. Abusing notation slightly, we will write $\ell(A) \in \mathbb{R}^{X\times\Theta}$ for the matrix of loss values $[\ell_x(A_\theta)]_{x,\theta}$, and $\ell_x(A) = [\ell_x(A_\theta)]_\theta \in \mathbb{R}^\Theta$ for the vector of losses for each expert θ on outcome x.

Definition 2 (Φ-mixability)**.** *Let Φ be an entropy on Δ_Θ. A loss $\ell : \mathcal{A} \to \mathbb{R}^X$ is Φ-mixable if for all $A \in \mathcal{A}^\Theta$, all $\mu \in \Delta_\Theta$, there exists an $\hat{a} \in \mathcal{A}$ such that for all $x \in X$*

$$\ell_x(\hat{a}) \leq \mathrm{Mix}^\Phi_{\ell,x}(A, \mu) := \inf_{\mu' \in \Delta_\Theta} \langle \mu', \ell_x(A)\rangle + D_\Phi(\mu', \mu). \tag{4}$$

The term on the right-hand side of (4) has some intuitive appeal. Since $\langle \mu', A\rangle = \mathbb{E}_{\theta\sim\mu'}[\ell_x(A_\theta)]$ (*i.e.*, the expected loss of an expert drawn at random according to μ') we can view the optimization as a trade off between finding a mixture μ' that tracks the expert with the smallest loss upon observing outcome x and keeping μ' close to μ, as measured by D_Φ. In the special case when Φ is Shannon entropy, ℓ is log loss, and expert predictions $A_\theta \in \Delta_X$ are distributions over X such an optimization is equivalent to Bayesian updating [14].

To see that Φ-mixability is indeed a generalization of Definition 1, we make use of an alternative form for the right-hand side of the bound in the Φ-mixability definition that "hides" the infimum inside Φ^*. As shown in Appendix A.1 this is a straightforward consequence of (3).

Lemma 2. *The mixability bound*

$$\mathrm{Mix}^{\Phi}_{\ell,x}(A,\mu) = \Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \ell_x(A)). \tag{5}$$

Hence, for $\Phi = \eta^{-1}H$ we have $\mathrm{Mix}^{\Phi}_{\ell,x}(A,\mu) = -\eta^{-1}\log\sum_\theta \exp(-\eta\ell_x(A_\theta))\mu_\theta$ which is the bound in Definition 1.

We now define a generalization of the Aggregating Algorithm of Definition 1 that very naturally relates to our definition of Φ-mixability: starting with some initial distribution over experts, the algorithm repeatedly incorporates the information about the experts' performances by finding the minimizer μ' in (4).

Definition 3 (Generalized Aggregating Algorithm)**.** *The algorithm begins with a mixture distribution $\mu^0 \in \Delta_\Theta$ over experts. On round t, after receiving expert predictions $A^t \in \mathcal{A}^\Theta$, the* generalized aggregating algorithm *(GAA) predicts any $\hat{a} \in \mathcal{A}$ such that $\ell_x(\hat{a}) \le \mathrm{Mix}^{\Phi}_{\ell,x}(A^t,\mu^{t-1})$ for all x which is guaranteed to exist by the Φ-mixability of ℓ. After observing $x^t \in X$, the GAA updates the mixture $\mu^{t-1} \in \Delta_\Theta$ by setting*

$$\mu^t := \operatorname*{arg\,min}_{\mu' \in \Delta_\Theta} \left\langle \mu', \ell_{x^t}(A^t)\right\rangle + D_\Phi(\mu', \mu^{t-1}). \tag{6}$$

We now show that this updating process simply aggregates the per-expert losses $\ell_x(A)$ in the dual space Δ^*_Θ with $\nabla\Phi(\mu^0)$ as the starting point. The GAA is therefore closely related to mirror descent techniques [10].

Lemma 3. *The GAA updates μ^t in* (6) *satisfy $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t)$ for all t and so*

$$\nabla\Phi(\mu^T) = \nabla\Phi(\mu^0) - \sum_{t=1}^{T} \ell_{x^t}(A^t). \tag{7}$$

The proof is given in Appendix A.1. Finally, to see that the above is indeed a generalization of the Aggregating Algorithm from Definition 1 we need only apply Lemma 3 and observe that for $\Phi = \eta^{-1}H$ we have $\nabla\Phi(\mu) = \eta^{-1}(\log(\mu) + \mathbb{1})$ and so $\log\mu^t = \log\mu^{t-1} - \eta\ell_{x^t}(A^t)$. Exponentiating this vector equality element-wise gives $\mu^t_\theta \propto \mu^{t-1}_\theta \exp(-\eta\ell_{x^t}(A^t_\theta))$.

3 Properties of Φ-mixability

In this section we establish a number of key properties for Φ-mixability, the most important of these being that Φ-mixability implies constant regret. We also show that Φ-mixability is not a vacuous concept for Φ other than Shannon entropy by showing that any Legendre Φ has Φ-mixable losses and that this is a necessary condition for such losses to exist.

3.1 Φ-mixability Implies Constant Regret

Theorem 2. *If $\ell : \mathcal{A} \to \mathbb{R}^X$ is Φ-mixable then there is a family of strategies parameterized by $\mu \in \Delta_\Theta$ which, for any sequence of observations $x^1, \ldots, x^T \in X$ and sequence of expert predictions $A^1, \ldots, A^T \in \mathcal{A}^\Theta$, plays a sequence $\hat{a}^1, \ldots, \hat{a}^T \in \mathcal{A}$ such that for all $\theta \in \Theta$*

$$\sum_{t=1}^{T} \ell_{x^t}(\hat{a}^t) \leq \sum_{t=1}^{T} \ell_{x^t}(A^t_\theta) + D_\Phi(\delta_\theta, \mu). \tag{8}$$

The proof is in Appendix A.2 and is a straight-forward consequence of Lemma 2 and the translation invariance of Φ^*. The standard notion of mixability is recovered when $\Phi = \frac{1}{\eta} H$ for $\eta > 0$ and H the Shannon entropy on Δ_Θ. In this case, Theorem 1 is obtained as a corollary for $\mu = |\Theta|^{-1}\mathbb{1}$, the uniform distribution over Θ. A compelling feature of our result is that it gives a natural interpretation of the constant $D_\Phi(\delta_\theta, \pi)$ in the regret bound: if π is the initial guess as to which expert is best before the game starts, the "price" that is paid by the player is exactly how far (as measured by D_Φ) the initial guess was from the distribution that places all its mass on the best expert.

The following example computes mixability bounds for the alternative entropies introduced in §2.1. They will be discussed again in §4.2 below.

Example 2. *Consider games with $K = |\Theta|$ experts and $\mu = K^{-1}\mathbb{1}$. For the (negative) Shannon entropy, the regret bound from Theorem 2 is $D_H(\delta_\theta, \mu) = \log K$. For quadratic entropy the regret bound is $D_Q(\delta_\theta, \mu) = 1 - \frac{2(K-1)}{K^2}$. For the family of Tsallis entropies the regret bound given by $D_{S_\alpha}(\delta_\theta, K^{-1}\mathbb{1}) = \alpha^{-1}(1 - K^{-\alpha})$. For the family of Rényi entropies the regret bound becomes $D_{R_\alpha}(\delta_\theta, K^{-1}\mathbb{1}) = \log K$.*

A second, easily established result concerns the mixability of scaled entropies. The proof is by observing that in (4) the only term in the definition of $\mathrm{Mix}^{\Phi_\eta}_{\ell,x}$ involving η is $D_{\Phi_\eta} = \frac{1}{\eta} D_\Phi$. The quantification over $A, \mu, \hat{a}, \mu'$ and x in the original definition have been translated into infima and suprema.

Lemma 4. *The function $M(\eta) := \inf_{A,\mu} \sup_{\hat{a}} \inf_{\mu',x} \ \mathrm{Mix}^{\Phi_\eta}_{\ell,x}(A, \mu) - \ell_x(\hat{a})$ is non-increasing.*

This implies that there is a well-defined maximal $\eta > 0$ for which a given loss ℓ is Φ_η-mixable since Φ_η-mixability is equivalent to $M(\eta) \geq 0$. We will call this maximal η the Φ-*mixability constant* for ℓ and denote it $\eta(\ell, \Phi) := \sup\{\eta > 0 : M(\eta) \geq 0\}$. This constant is central to the discussion in Section 4.3 below.

3.2 Φ-Mixability of Proper Losses and Their Bayes Risks

Entropies are known to be closely related to the Bayes risk of what are called proper losses or proper scoring rules [15, 16]. Here, the predictions are distributions over outcomes, *i.e.*, points in Δ_X. To highlight this we will use p, $\hat{p}$ and P instead of a, $\hat{a}$ and A to denote actions. If a loss $\ell : \Delta_X \to \mathbb{R}^X$ is used to assign a penalty $\ell_x(\hat{p})$ to a prediction $\hat{p}$ upon outcome x it is said to be *proper* if its expected value under $x \sim p$ is minimized by predicting $\hat{p} = p$. That is, for all $p, \hat{p} \in \Delta_X$ we have

$$\mathbb{E}_{x\sim p}\left[\ell_x(\hat{p})\right] = \langle p, \ell(\hat{p})\rangle \geq \langle p, \ell(p)\rangle =: -F^{\ell}(p)$$

where $-F^{\ell}$ is the *Bayes risk* of ℓ and is necessarily concave [5], thus making $F^{\ell} : \Delta_X \to \mathbb{R}$ convex and thus an entropy. The correspondence also goes the other way: given any convex function $F : \Delta_X \to \mathbb{R}$ we can construct a unique proper loss [17]. The following representation can be traced back to [18] but is expressed here using convex duality.

Lemma 5. *If $F : \Delta_X \to \mathbb{R}$ is a differentiable entropy then the loss $\ell^F : \Delta_X \to \mathbb{R}$ defined by*

$$\ell^F(p) := F^*(\nabla F(p))\mathbb{1} - \nabla F(p) \tag{9}$$

is proper.

It is straight-forward to show that the proper loss associated with the negative Shannon entropy $\Phi = H$ is the log loss, that is, $\ell^H(\mu) := -\left(\log \mu(\theta)\right)_{\theta\in\Theta}$.

This connection between losses and entropies lets us define the Φ-mixability of a proper loss strictly in terms of its associated entropy. This is similar in spirit to the result in [5] which shows that the original mixability (for $\Phi = H$) can be expressed in terms of the relative curvature of Shannon entropy and the loss's Bayes risk. We use the following definition to explore the optimality of Shannon mixability in Section 4.3 below.

Definition 4. *An entropy $F : \Delta_X \to \mathbb{R}$ is* Φ-mixable *if*

$$\sup_{P,\mu} F^*\left(\left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x - \Phi^*(\nabla\Phi(\mu))\mathbb{1}\right) \leq 0 \tag{10}$$

where ℓ^F is as in Lemma 5 and the supremum is over expert predictions $P \in \Delta_X^\Theta$ and mixtures over experts $\mu \in \Delta_\Theta$.

Although this definition appears complicated due to the handling of vectors in $\mathbb{R}^X$ and $\mathbb{R}^\Theta$, it has a natural interpretation in terms of *risk measures* from mathematical finance [19]. Given some convex function $\alpha : \Delta_X \to \mathbb{R}$, its associated risk measure is its dual $\rho(v) := \sup_{p\in\Delta_X} \langle p, -v\rangle - \alpha(p) = \alpha^*(-v)$ where v is a *position* meaning v_x is some monetary value associated with outcome x occurring. Due to its translation invariance, the quantity $\rho(v)$ is often interpreted as the amount of "cash" (*i.e.*, outcome independent value) an agent would ask for to take on the uncertain position v. Observe that the risk ρ^F for when $\alpha = F$ satisfies $\rho^F \circ \ell^F = 0$ so that $\ell^F(p)$ is always a ρ^F-risk free position. If we now interpret $\mu^* = \nabla\Phi(\mu)$ as a position over outcomes in Θ and Φ^* as a risk for $\alpha = \Phi$ the term $\left\{\Phi^*(\mu^* - \ell^F_x(P))\right\}_x - \Phi^*(\mu^*)\mathbb{1}$ can be seen as the change in ρ^Φ risk when shifting position μ^* to $\mu^* - \ell^F_x(P)$ for each possible outcome x. Thus, the mixability condition in (10) can be viewed as a requirement that a ρ^F-risk free change in positions over Θ always be ρ^F-risk free.

The following theorem shows that the entropic version of Φ-mixability Definition 4 is equivalent to the loss version in Definition 2 in the case of proper losses. Its proof can be found in Appendix A.3 and relies on Sion's theorem and the facts that proper losses are quasi-convex. This latter fact appears to be new so we state it here as a separate lemma and prove it in Appendix A.1.

Lemma 6. *If $\ell : \Delta_X \to \mathbb{R}$ is proper then $p' \mapsto \langle p, \ell(p') \rangle$ is quasi-convex for all $p \in \Delta_X$.*

Theorem 3. *If $\ell : \Delta_X \to \mathbb{R}^X$ is proper and has Bayes risk $-F$ then F is an entropy and ℓ is Φ-mixable if and only if F is Φ-mixable.*

The entropic form of mixability in (10) shares some similarities with expressions for the classical mixability constants given in [20] for binary outcome games and in [5] for general games. Our expression for the mixability is more general than the previous two being both for binary and non-binary outcomes and for general entropies. It is also arguably more efficient since the optimization in [5] for non-binary outcomes requires inverting a Hessian matrix at each point in the optimization.

3.3 Characterizing and Comparing Φ-mixability

Although Theorem 2 recovers the already known constant regret bound for Shannon-mixable losses, it is natural to ask whether the result is vacuous or not for other entropies. That is, do there exist Φ-mixable losses for Φ other than Shannon entropy? If so, do such Φ-mixable losses exist for any entropy Φ? The next theorem answers both of these questions, showing that the existence of "non-trivial" Φ-mixable losses is intimately related to the behaviour of an entropy's gradient at the simplex's boundary. Specifically, an entropy Φ is said to be *Legendre* [21] if: a) Φ is strictly convex in $\operatorname{int}(\Delta_\Theta)$; and b) $\|\nabla\Phi(\mu)\| \to \infty$ as $\mu \to \mu_b$ for any μ_b on the boundary of Δ_Θ.

We will say a loss is *non-trivial* if there exist distinct actions which are optimal for distinct outcomes (see A.4 for formal definition). This, for example, rules out constant losses – *i.e.*, $\ell(a) = k \in \mathbb{R}^X$ for all $a \in \mathcal{A}$ – are easily[2] seen to be Φ-mixable for any Φ. For technical reasons we will further restrict our attention to *curved* losses by which we mean those losses with strictly concave Bayes risks. We conjecture that the following theorem also holds for non-curved losses.

Theorem 4. *There exist non-trivial, curved Φ-mixable losses if and only if the entropy Φ is Legendre.*

The proof is in Appendix A.4. From this result we can deduce that there are no Q-mixable losses. Also, since it is easy to show the derivatives ∇S_α and ∇R_α are unbounded for $\alpha \in (0, 1)$, the entropies S_α and R_α are Legendre. Thus there exist S_α- and R_α-mixable losses when $\alpha \in (-1, 0)$.

4 Conclusions and Open Questions

The main purpose of this work was to shed new light on mixability by casting it within the broader notion of Φ-mixability. We showed that the constant regret bounds enjoyed by mixability losses are due to the translation invariance of entropic duals, and

[2] The inequality in (4) reduces to $0 \leq \inf_{\mu'} D_\Phi(\mu', \mu)$ which is true for all Bregman divergences.

so are also enjoyed by any Φ-mixable loss. The definitions and technical machinery presented here allow us to ask precise questions about entropies and the optimality of their associated aggregating algorithms.

4.1 Are All Legendre Entropies "Equivalent"?

Since Theorem 4 shows the existence of Φ-mixable losses, a natural question concerns the relationship between the sets of losses that are mixable for different choices of Φ. For example, are there losses that are H-mixable but not S_α-mixable, or vice-versa? We conjecture that essentially all Legendre entropies Φ have the same Φ-mixable losses up to a scaling factor.

Conjecture 1. *Let Φ be a entropy on Δ_Θ and ℓ be a Φ-mixable loss. If Ψ is a Legendre entropy on Δ_Θ then there exists an $\eta > 0$ such that ℓ is $\eta^{-1}\Psi$-mixable.*

Some intuition for this conjecture is derived from observing that $\operatorname{Mix}_{\ell,x}^{\Psi_\eta} = \eta^{-1} \operatorname{Mix}_{\eta\ell,x}^{\Psi}$ and that as $\eta \to 0$ the function $\eta\ell$ behaves like a constant loss and will therefore be mixable. This means that scaling up $\operatorname{Mix}_{\eta\ell,x}^{\Psi}$ by η^{-1} should make it larger than $\operatorname{Mix}_{\ell,x}^{\Phi}$. However, some subtlety arises in ensuring that this dominance occurs uniformly.

4.2 Asymptotic Behaviour

There is a lower bound due to Vovk [3] for general losses ℓ which shows that if one is allowed to vary the number of rounds T and the number of experts $K = |\Theta|$, then no regret bound can be better than the optimal regret bound obtained by Shannon mixability. Specifically, for a fixed loss ℓ with optimal Shannon mixability constant η_ℓ, suppose that for some $\eta' > \eta_\ell$ we have a regret bound of the form $(\log K)/\eta'$ as well as some strategy L for the learner that supposedly satisfies this regret bound. Vovk's lower bound shows, for this η' and L, that there exists an instantiation of the prediction with expert advice game with T large enough and K roughly exponential in T (and both are still finite) for which the alleged regret bound will fail to hold at the end of the game with non-zero probability. The regime in which Vovk's lower bound holds suggests that the best achievable regret with respect to the number of experts grows as $\log K$. Indeed, there is a lower bound for general losses ℓ that shows the regret of the best possible algorithm on games using ℓ must grow like $\Omega(\log_2 K)$ [20].

The above lower bound arguments apply when the number of experts is large (*i.e.*, exponential in the number of rounds) or if we consider the dynamics of the regret bound as K grows. This leaves open the question of the best possible regret bound for moderate and possibly fixed K which we formally state in the next section. This question that serves as a strong motivation for the study of generalized mixability considered here. Note also that the above lower bounds are consistent with the fact that there cannot be non-trivial, Φ-mixable losses for non-Legendre Φ (*e.g.*, the quadratic entropy Q) since the growth of the regret bound as a function of K (cf. Example 2) is less than $\log K$ and hence violates the above lower bounds.

4.3 Is There An "Optimal" Entropy?

Since we believe that Φ-mixability for Legendre Φ yield the same set of losses, we can ask whether, for a fixed loss ℓ, some Φ give better regret bounds than others. These bounds depend jointly on the largest η such that ℓ is Φ_η-mixable and the value of $D_\Phi(\delta_\theta, \mu)$. We can define the optimal regret bound one can achieve for a particular loss ℓ using the generalized aggregating algorithm with $\Phi_\eta := \frac{1}{\eta}\Phi$ for some $\eta > 0$. This allows us to compare entropies on particular losses, and we can say that an entropy *dominates* another if its optimal regret bound is better for all losses ℓ. Recalling the definition of the maximal Φ-mixability constant from Lemma 4, we can determine a quantity of more direct interest: the best regret bound one can obtain using a scaled copy of Φ. Recall that if ℓ is Φ-mixable, then the best regret bound one can achieve from the generalized aggregating algorithm is $\inf_\mu \sup_\theta D_\Phi(\delta_\theta, \mu)$. We can therefore define the best regret bound for ℓ on a scaled version of Φ to be $R_{\ell,\Phi} := \eta(\ell, \Phi)^{-1} \inf_\mu \sup_\theta D_\Phi(\delta_\theta, \mu)$ which simply corresponds to the regret bound for the entropy $\Phi_{\eta(\ell,\Phi)}$. Note a crucial property of $R_{\ell,\Phi}$, which will be very useful in comparing entropies: $R_{\ell,\Phi} = R_{\ell,\alpha\Phi}$ for all $\alpha > 0$. (This follows from the observation that $\eta(\ell, \alpha\Phi) = \eta(\ell, \Phi)/\alpha$.) That is, $R_{\ell,\Phi}$ is independent of the particular scaling we choose for Φ.

We can now use $R_{\ell,\Phi}$ to define a scale-invariant relation over entropies. Define $\Phi \geq_\ell \Psi$ if $R_{\ell,\Phi} \leq R_{\ell,\Psi}$, and $\Phi \geq_* \Psi$ if $\Phi \geq_\ell \Psi$ for all losses ℓ. In the latter case we say Φ *dominates* Ψ. By construction, if one entropy dominates another its regret bound is guaranteed to be tighter and therefore its aggregating algorithm will achieve better worst-case regret. As discussed above, one natural candidate for a universally dominant entropy is the Shannon entropy.

Conjecture 2. *For all choices of Θ, the negative Shannon entropy dominates all other entropies. That is, $H \geq_* \Phi$ for all Θ and all convex Φ on Δ_Θ.*

Although we have not been able to prove this conjecture we were able to collect some positive evidence in the form of Table 1. Here, we took the entropic form of Φ-mixability from Definition 4 and implemented[3] it as an optimization problem in the language R and computed $\eta(\ell^F, \Phi)$ for F and Φ equal to the entropies introduced in Example 1 for two expert games with two outcomes. The maximal η (and hence the optimal regret bounds) for each pair was found doing a binary search for the zero-crossing of $M(\eta)$ from Lemma 4 and then applying the bounds from Example 2. Although we were expecting the dominant entropy for each loss ℓ^F to be its "matching" entropy (*i.e.*, $\Phi = F$), as can be seen from the table the optimal regret bound for every loss was obtained in the column for H. However, one interesting feature for these matching cases is that the optimal η (shown in parentheses) is always equal to 1.

Conjecture 3. *Suppose $|X| = |\Theta|$ so that $\Delta_\Theta = \Delta_X$. Given a Legendre $\Phi : \Delta_\Theta \to \mathbb{R}$ and its associated proper loss $\ell^\Phi : \Delta_X \to \mathbb{R}^X$, the maximal η such that ℓ^Φ is $\eta^{-1}\Phi$-mixable is $\eta = 1$.*

We conjecture that this pattern will hold for matching entropies and losses for larger numbers of experts and outcomes and hope to test or prove this in future work.

[3] In order to preserve anonymity the code will not be made available until after publication.

Table 1: Mixability and optimal regrets for pairs of losses and entropies in 2 outcome/2 experts games. Entries show the regret bound $\eta^{-1}D_\Phi(\delta_\theta, \frac{1}{2}\mathbb{1})$ for the maximum η (in parentheses).

	Entropy						
Loss	H	$S_{-.1}$	$S_{-.5}$	$S_{-.9}$	$R_{-.1}$	$R_{-.5}$	$R_{-.9}$
log	0.69 (1*)	0.74 (.97)	1.17 (.71)	5.15 (.19)	0.77 (0.9)	1.38 (0.5)	6.92 (0.1)
ℓ^Q	0.34 (2)	0.37 (1.9)	0.58 (1.4)	2.57 (0.4)	0.38 (1.8)	0.69 (1)	3.45 (0.2)
$\ell^{S_{-.5}}$	0.49 (1.4)	0.53 (1.4)	0.82 (1*)	3.64 (.26)	0.54 (1.3)	0.98 (.71)	4.90 (.14)
$\ell^{R_{.5}}$	0.34 (2)	0.37 (1.9)	0.58 (1.4)	2.57 (.37)	0.38 (1.8)	0.69 (1*)	3.46 (0.2)

Acknowledgments

We would like to thank Matus Telgarsky for help with restricted duals, Brendan van Rooyen for noting that there are no quadratic mixable losses, and Harish Guruprasad for identifying a flaw in an earlier "proof" of the quasi-convexity of proper losses. Mark Reid is supported by an ARC Discovery Early Career Research Award (DE130101605) and part of this work was developed while he was visiting Microsoft Research. NICTA is funded by the Australian Government and as an ARC ICT Centre of Excellence.

References

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

[2] Volodya Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT)*, pages 371–383, 1990.

[3] Volodya Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 51–60. ACM, 1995.

[4] Nicolo Cesa-Bianchi. *Prediction, learning, and games*. Cambridge University Press, 2006.

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

[6] Jyrki Kivinen and Manfred K Warmuth. Averaging expert predictions. In *Computational Learning Theory*, pages 153–167. Springer, 1999.

[7] Yuri Kalnishkan and Michael V. Vyugin. The weak aggregating algorithm and weak mixability. *Journal of Computer and System Sciences*, 74:1228–1244, 2008.

[8] Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

[9] Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

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

[11] J.B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer Verlag, 2001.

[12] Tomasz Maszczyk and Włodzisław Duch. Comparison of shannon, renyi and tsallis entropy used in decision trees. In *Artificial Intelligence and Soft Computing–ICAISC 2008*, pages 643–651. Springer, 2008.

[13] Tim Van Erven and Peter Harremoës. R\'enyi divergence and kullback-leibler divergence. *arXiv preprint arXiv:1206.2459*, 2012.

[14] Peter M Williams. Bayesian conditionalisation and the principle of minimum information. *British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

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

[16] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[17] Elodie Vernet, Robert C Williamson, and Mark D Reid. Composite multiclass losses. In *NIPS*, volume 24, pages 1224–1232, 2011.

[18] Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[19] Hans Föllmer and Alexander Schied. Stochastic finance, volume 27 of de gruyter studies in mathematics, 2004.

[20] David Haussler, Jyrki Kivinen, and Manfred K Warmuth. Sequential prediction of individual sequences under general loss functions. *Information Theory, IEEE Transactions on*, 44(5):1906–1925, 1998.

[21] R.T. Rockafellar. *Convex analysis*. Princeton University Press, 1997.

[22] Frederick A. Valentine. *Convex Sets*. McGraw-Hill, New York, 1964.

[23] Jacob D Abernethy and Rafael M Frongillo. A characterization of scoring rules for linear properties. *Journal of Machine Learning Research-Proceedings Track*, 23:27–1, 2012.

A Appendix

A.1 Proof of Lemmas

Proof of Lemma 1. To show 1) we observe that $(\eta^{-1}\Phi)^*(v) = \sup_p \langle v, p\rangle - \eta^{-1}\Phi(p) = \eta^{-1}\sup_p \langle \eta v, p\rangle - \Phi(p) = \eta^{-1}\Phi^*(\eta v)$. For 2), we note that the definition of the dual implies $\Phi^*(v + \alpha\mathbb{1}) = \sup_{\mu\in\Delta_\Theta} \langle \mu, v + \alpha\mathbb{1}\rangle - \Phi(\mu) = \sup_{\mu\in\Delta_\Theta} \langle \mu, v\rangle - \Phi(\mu) + \alpha = \Phi^*(v) + \alpha$ since $\langle \mu, \mathbb{1}\rangle = 1$. Taking derivatives of both sides gives the final part of the lemma. □

Proof of Lemma 2. By definition $\Phi^*(\nabla\Phi(\mu) - v) = \sup_{\mu'\in\Delta_\Theta} \langle \mu', \nabla\Phi(\mu) - v\rangle - \Phi(\mu')$ and using (3) gives $\Phi^*(\nabla\Phi(\mu)) = \langle \mu, \nabla\Phi(\mu)\rangle - \Phi(\mu)$. Subtracting the former from the latter gives $\langle \mu, \nabla\Phi(\mu)\rangle - \Phi(\mu) - \left[\sup_{\mu'\in\Delta_\Theta} \langle \mu', \nabla\Phi(\mu) - v\rangle - \Phi(\mu')\right]$ which, when rearranged gives $\inf_{\mu'\in\Delta_\Theta} \Phi(\mu') - \Phi(\mu) - \langle \nabla\Phi(\mu), \mu' - \mu\rangle + \langle \mu', v\rangle$ establishing the result.

When $\Phi = H$ – *i.e.*, Φ is the (negative) Shannon entropy – we have that $\nabla\Phi(\mu) = \log\mu + \mathbb{1}$, that $\Phi^*(v) = \log\sum_\theta \exp(v_\theta)$, and so $\nabla\Phi^*(v) = \exp(v)/\sum_\theta \exp(v_\theta)$, where $\log$ and $\exp$ are interpreted as acting point-wise on the vector μ. By Lemma 1, $\Phi^*(\nabla\Phi(\mu)) = \Phi^*(\log\mu + \mathbb{1}) = \Phi^*(\log(\mu)) + 1 = 1$ since $\Phi^*(\log(\mu_\theta)) = \log\sum_\theta \mu_\theta = 0$. Similarly, $\Phi^*(\nabla\Phi(\mu) - \ell_x(A)) = \Phi^*(\log(\mu) - \ell_x(A)) + 1 = \log\sum_\theta \mu_\theta \exp(-\ell_x(A)) + 1$. Substituting this into Lemma 2 and applying the second part of Lemma 1 shows that $\mathrm{Mix}_{\ell,x}^{\eta^{-1}H}(A, \mu) = -\eta^{-1}\log\sum_\theta \exp(-\eta\ell_x(A_\theta))$, recovering the right-hand side of the inequality in Definition 1. □

Proof of Lemma 5. By eq. (3) we have $F^*(\nabla F(p)) = \langle p, \nabla F(p)\rangle - F(p)$, giving us

$$\begin{aligned}\left\langle p, \ell^F(p')\right\rangle - \left\langle p, \ell^F(p)\right\rangle &= \Big(\langle p', \nabla F(p')\rangle - F(p') - \langle p, \nabla F(p')\rangle\Big) \\ &\quad - \Big(\langle p, \nabla F(p)\rangle - F(p) - \langle p, \nabla F(p)\rangle\Big) \\ &= D_F(p, p'),\end{aligned}$$

from which propriety follows. □

Proof of Lemma 3. By considering the Lagrangian $\mathcal{L}(\mu, a) = \langle \mu, \ell_{x^t}(A)\rangle + D_\Phi(\mu, \mu^{t-1}) + \alpha(\langle \mu, \mathbb{1}\rangle - 1)$ and setting its derivative to zero we see that the minimizing μ^t must satisfy $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t) - \alpha^t\mathbb{1}$ where $\alpha^t \in R$ is the dual variable at step t. For convex Φ, the functions $\nabla\Phi^*$ and $\nabla\Phi$ are inverses [11] so $\mu^t = \nabla\Phi^*(\nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t) - a^t\mathbb{1}) = \nabla\Phi^*(\nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t))$ by the translation invariance of Φ^* (Lemma 1). This means the constants α^t are arbitrary and can be ignored. Thus, the mixture updates satisfy the relation in the lemma and summing over $t = 1, \ldots, T$ gives (7). □

Proof of Lemma 6. Let $n = |X|$ and fix an arbitrary $p \in \Delta_X$. The function $f_p(q) = \langle p, \ell(q)\rangle$ is quasi-convex if its α sublevel sets $F_p^\alpha := \{q \in \Delta_X : \langle p, \ell(q)\rangle \leq \alpha\}$ are convex for all $\alpha \in \mathbb{R}$. Let $g(p) := \inf_q f_p(q)$ and fix an arbitrary $\alpha > g(p)$ so that $F_p^\alpha \neq \emptyset$. Let $Q_p^\alpha := \{v \in \mathbb{R}^n : \langle p, v\rangle \leq \alpha\}$ so $F_p^\alpha = \{q \in \Delta_X : \ell(q) \in Q_p^\alpha\}$.

Denote by $h_q^\beta := \{v \colon \langle v, q\rangle = \beta\}$ the hyperplane in direction $q \in \Delta_X$ with offset $\beta \in \mathbb{R}$ and by $H_q^\beta := \{v \colon \langle v, q\rangle \geq \beta\}$ the corresponding half-space. Since ℓ is proper, its *superprediction set* $\mathcal{S}_\ell = \{\lambda \in \mathbb{R}^n : \exists q \in \Delta_X \forall x \in X \lambda_x \geq \ell_x(q)\}$ (see [17, Prop. 17]) is supported at $x = \ell(q)$ by the hyperplane $h_q^{g(q)}$ and furthermore since $\mathcal{S}_\ell$ is convex, $\mathcal{S}_\ell = \bigcap_{q\in\Delta_X} H_q^{g(q)}$.

Let

$$V_p^\alpha := \bigcap_{v\in\ell(\Delta_X)\cap Q_p^\alpha} H_{\ell^{-1}(v)}^{g(\ell^{-1}(v))} = \bigcap_{q\in F_p^\alpha} H_q^{g(q)}$$

(see figure 1). Since V_p^α is the intersection of halfspaces it is convex. Note that a given half-space $H_q^{g(q)}$ is supported by exactly one hyperplane, namely $h_q^{g(q)}$. Thus the set of hyperplanes that support V_p^α is $\{h_q^{g(q)} \colon q \in F_p^\alpha\}$ If $u \in F_p^\alpha$ then there is a hyperplane in direction u that supports V_p^α and its offset is given by

$$\sigma_{V_p^\alpha}(u) := \inf_{v\in V_p^\alpha} \langle u, v\rangle = g(p) > -\infty$$

whereas if $u \notin F_p^\alpha$ then for all $\beta \in \mathbb{R}$, h_u^β does not support V_p^α and hence $\sigma_{V_p^\alpha}(u) = -\infty$. Thus we have shown

$$\left(u \notin W_p^\alpha\right) \Leftrightarrow \Big(\sigma_{V_p^\alpha}(u) = -\infty\Big).$$

Observe that $\sigma_{V_p^\alpha}(u) = -s_{V_p^\alpha}(-u)$ where $s_C(u) = \sup_{v\in C}\langle u, v\rangle$ is the support function of a set C. It is known [22, Theorem 5.1] that the "domain of definition" of a support function $\{u \in \mathbb{R}^n \colon s_C(u) < +\infty\}$ for a convex set C is always convex. Thus $G_p^\alpha := \{u \in \Delta_X \colon \sigma_{V_p^\alpha}(u) > -\infty\} = \{u \in \mathbb{R}^n \colon \sigma_{V_p^\alpha}(u) > -\infty\} \cap \Delta_X$ is always convex because it is the intersection of convex sets. Finally by observing that

$$G_p^\alpha = \{p \in \Delta_X \colon \ell(p) \in \ell(\Delta_X) \cap Q_p^\alpha\} = F_p^\alpha$$

we have shown that F_p^α is convex. Since $p \in \Delta_X$ and $\alpha \in \mathbb{R}$ were arbitrary we have thus shown that f_p is quasi-convex for all $p \in \Delta_X$.

□

A.2 Proof of Theorem 2

Proof of Theorem 2. Applying Lemma 2 to the assumption that ℓ is Φ-mixable means that for μ equal to the updates μ^t from Definition 3 and A^t equal to the expert predictions at round t, there must exist an $\hat{a}^t \in \Delta_X$ such that

$$\ell_{x^t}(\hat{a}^t) \leq \Phi^*(\nabla\Phi(\mu^{t-1})) - \Phi^*(\nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t))$$

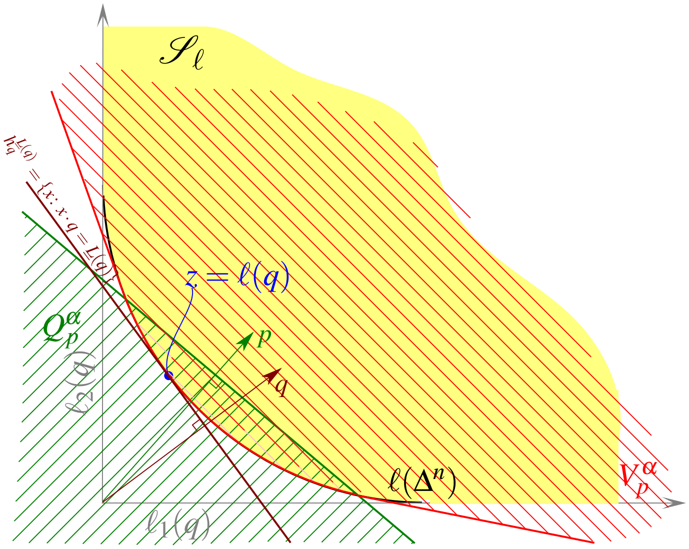

Figure 1: Visualization of construction in proof of Lemma 6.

for all $x^t \in X$. Summing these bounds over $t = 1, \dots, T$ gives

$$\sum_{t=1}^{T} \ell_{x^t}(p^t) \leq \sum_{t=1}^{T} \Phi^*(\nabla\Phi(\mu^{t-1})) - \Phi^*(\nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t))$$

$$= \Phi^*(\nabla\Phi(\mu^0)) - \Phi^*(\nabla\Phi(\mu^T)) \tag{11}$$

$$= \inf_{\mu' \in \Delta_\Theta} \left\langle \mu', \sum_{t=1}^{T} \ell_{x^T}(A^t) \right\rangle + D_\Phi(\mu', \mu^0) \tag{12}$$

$$\leq \left\langle \mu', \sum_{t=1}^{T} \ell_{x^t}(A^t) \right\rangle + D_\Phi(\mu', \mu^0) \qquad \text{for all } \mu' \in \Delta_\Theta \tag{13}$$

Line (11) above is because $\nabla\Phi(\mu^t) = \nabla\Phi(\mu^{t-1}) - \ell_{x^t}(A^t)$ by Lemma 3 and the series telescopes. Line (12) is obtained by applying (6) from Lemma 3 and matching equations (5) and (4). Setting $\mu' = \delta_\theta$ and noting $\langle \delta_\theta, \ell(A^t) \rangle = \ell_{x^t}(A^t_\theta)$ gives the required result. □

A.3 Proof of Theorem 3

We first establish a general reformulation of Φ-mixability that holds for arbitrary ℓ by converting the quantifiers in the definition of Φ-mixability from Lemma 2 for ℓ into an expression involving infima and suprema. We then further refine this by assuming

$\ell = \ell^F$ is proper (and thus quasi-convex) and has Bayes risk F.

$$\begin{aligned}
&\inf_{A,\mu} \sup_{\hat{a}} \inf_{x} \; \Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \ell^F_x(A)) - \ell^F_x(\hat{a}) \geq 0 \\
\iff &\inf_{A,\mu} \sup_{\hat{a}} \inf_{p} \; \left\langle p, \left\{\Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x \right\rangle - \left\langle p, \ell^F_x(\hat{p}) \right\rangle \geq 0
\end{aligned} \tag{14}$$

where the term in braces is a vector in $\mathbb{R}^X$. The infimum over x is switched to an infimum over distributions over $p \in \Delta_X$ because the optimization over p will be achieved on the vertices of the simplex as it is just an average over random variables over X.

From here on we assume that $\ell = \ell^F$ is proper and adjust our notation to emphasis that actions $\hat{a} = \hat{p}$ and $A = P$ are distributions. Note that the new expression is linear – and therefore convex in p – and, by Lemma 6, we know ℓ^F is quasi-convex and so the function being optimized in (14) is quasi-concave in $\hat{p}$. We can therefore apply Sion's theorem to swap $\inf_p$ and $\sup_{\hat{p}}$ which means ℓ^F is Φ-mixable if and only if

$$\begin{aligned}
&\inf_{P,\mu} \inf_{p} \sup_{\hat{p}} \; \left\langle p, \left\{\Phi^*(\nabla\Phi(\mu)) - \Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x \right\rangle - \left\langle p, \ell^F_x(\hat{p}) \right\rangle \geq 0 \\
\iff &\inf_{P,\mu} \inf_{p} \; \Phi^*(\nabla\Phi(\mu)) - \left\langle p, \left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x) \right\rangle + F(p) \geq 0 \\
\iff &\inf_{P,\mu} \; \Phi^*(\nabla\Phi(\mu)) - F^*(\left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x) \geq 0
\end{aligned}$$

The second line above is obtained by recalling that, by the definition of ℓ^F, its Bayes risk is F. We now note that the inner infimum over p passes through $\Phi^*(\nabla\Phi(\mu))$ so that the final two terms are just the convex dual for F evaluated at $\left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x$. Finally, by translation invariance of F^* we can pull the $\Phi^*(\pi^*)$ term inside F^* to simplify further so that the loss ℓ^F with Bayes risk F is Φ-mixable if and only if

$$\inf_{P,\mu} -F^*\left(\left\{\Phi^*(\nabla\Phi(\mu) - \ell^F_x(P))\right\}_x - \Phi^*(\nabla\Phi(\mu))\mathbb{1}\right) \geq 0.$$

Applying Lemma 5 to write ℓ^F in terms of F and passing the sign through the infimum and converting it to a supremum gives the required result.

A.4 Proof of Theorem 4

We will make use the following formulation of mixability,

$$M(\eta) := \inf_{A\in\mathcal{A},\,\pi\in\Delta_\Theta} \; \sup_{\hat{a}\in\mathcal{A}} \; \inf_{\mu\in\Delta_\Theta,\,x\in X} \; \langle \mu, \ell_x(A)\rangle + \frac{1}{\eta} D_\Phi(\mu,\pi) - \ell_x(\hat{a}), \tag{15}$$

so that ℓ is Φ_η-mixable if and only if $M(\eta) \geq 0$.

We call a loss ℓ *nontrivial* if there exist x^*, x' and a^*, a' such that

$$a' \in \arg\min\{\ell_{x^*}(a) : \ell_{x'}(a) = \inf_{a\in\mathcal{A}} \ell_{x'}(a)\} \text{ and } \inf_{a\in\mathcal{A}} \ell_{x^*}(a) = \ell_{x^*}(a^*) < \ell_{x^*}(a')\;. \tag{16}$$

Intuitively, this means that there exist distinct actions which are optimal for different outcomes x^*, x'. Note that in particular, among all optimum actions for x', a' has the lowest loss on x^*.

Lemma 7. *Suppose ℓ has a strictly concave Bayes risk L. Then given any distinct $\mu^*, \mu' \in \Delta_\Theta$, there is some $A \in \mathcal{A}$ and $x^*, x' \in X$ such that for all $\hat{a} \in \mathcal{A}$ we have at least one of the following:*

$$\langle \mu^*, \ell_{x^*}(A) \rangle < \ell_{x^*}(\hat{a}) \ , \quad \langle \mu', \ell_{x'}(A) \rangle < \ell_{x'}(\hat{a}) \ . \tag{17}$$

Proof. Let θ^* be an expert such that $\alpha := \mu^*_{\theta^*} > \mu'_{\theta^*} =: \beta$, which exists as $\mu^* \neq \mu'$. Pick arbitrary $x^*, x' \in X$ and let $p^*, p' \in \Delta_X$ with support only on $\{x^*, x'\}$ and $p^*_{x^*} = \alpha/(\alpha+\beta), p'_{x^*} = (1-\alpha)/(2-\alpha-\beta)$. Now let $a^* = \arg\min_{a\in\mathcal{A}} \mathbb{E}_{x\sim p^*}\left[\ell_x(a)\right]$, $a' = \arg\min_{a\in\mathcal{A}} \mathbb{E}_{x\sim p'}\left[\ell_x(a)\right]$, and set A such that $A_{\theta^*} = a^*$ and $A_\theta = a'$ for all other $\theta \in \Theta$.

Now suppose there is some $\hat{a} \in \mathcal{A}$ violating eq. (17). Then in particular,

$$\begin{aligned}
\tfrac{1}{2}\left(\ell_{x^*}(\hat{a}) + \ell_{x'}(\hat{a})\right) &\leq \tfrac{1}{2}\left(\langle \mu^*, \ell_{x^*}(A)\rangle + \langle \mu', \ell_{x'}(A)\rangle\right) \\
&= \tfrac{1}{2}\left(\alpha\ell_{x^*}(a^*) + (1-\alpha)\ell_{x^*}(a') + \beta\ell_{x'}(a^*) + (1-\beta)\ell_{x'}(a')\right) \\
&= \tfrac{\alpha+\beta}{2}\left(\tfrac{\alpha}{\alpha+\beta}\ell_{x^*}(a^*) + \tfrac{\beta}{\alpha+\beta}\ell_{x'}(a^*)\right) + \tfrac{2-\alpha-\beta}{2}\left(\tfrac{1-\alpha}{2-\alpha-\beta}\ell_{x^*}(a') + \tfrac{1-\beta}{2-\alpha-\beta}\ell_{x'}(a')\right) \\
&= \tfrac{\alpha+\beta}{2}L(p^*) + \left(1 - \tfrac{\alpha+\beta}{2}\right)L(p') \ .
\end{aligned}$$

Letting $\bar{p} \in \Delta_X$ with $\bar{p}_{x^*} = \bar{p}_{x'} = 1/2$, observe that $\bar{p} = \frac{\alpha+\beta}{2}p^* + (1 - \frac{\alpha+\beta}{2})p'$. But by the above calculation, we have $L(\bar{p}) \leq \frac{\alpha+\beta}{2}L(p^*) + (1 - \frac{\alpha+\beta}{2})L(p')$, thus violating strict concavity of L. □

Non-Legendre $\Longrightarrow$ no nontrivial mixable ℓ with strictly convex Bayes risk:

To show that no non-constant Φ-mixable losses exist, we must exhibit a $\pi \in \Delta_\Theta$ and an $A \in \mathcal{A}$ such that for all $\hat{a} \in \mathcal{A}$ we can find a $\mu \in \Delta_\Theta$ and $x \in X$ satisfying $\langle \mu, \ell_x(A)\rangle + \frac{1}{\eta}D_\Phi(\mu, \pi) - \ell_x(\hat{a}) < 0$. Since Φ is non-Legendre it must either (1) fail strict convexity, or (2) have a point on the boundary with bounded derivative; we will consider each case separately.

(1) Assume that Φ is not strictly convex; then we have some $\mu^* \neq \mu'$ such that $D_\Phi(\mu^*, \mu') = 0$. By Lemma 7 with these two distributions, we have some A and x^*, x' such that for all $\hat{a}$, either (i) $\langle \mu^*, \ell_{x^*}(A)\rangle < \ell_{x^*}(\hat{a})$ or (ii) $\langle \mu', \ell_{x'}(A)\rangle < \ell_{x'}(\hat{a})$. We set $\pi = \mu'$; in case (i) we take $\mu = \mu^*$ and $x = x^*$, and in (ii) we take $\mu = \mu'$ and $x = x'$, but as $\frac{1}{\eta}D_\Phi(\mu, \pi) = 0$ in both cases, we have $M(\eta) < 0$ for all η.

(2) Now assume instead that we have some μ' on the boundary of Δ_Θ with bounded $\|\nabla\Phi(\mu')\| = C < \infty$. Because μ' is on the boundary of Δ_Θ there is at least one expert $\theta^* \in \Theta$ for which $\mu'_{\theta^*} = 0$. Pick x^*, x', a^*, a' from the definition of nontrivial, eq. (16). In particular, note that $\ell_{x^*}(a^*) < \ell_{x^*}(a')$. Let $\pi = \mu'$ and $A \in \mathcal{A}$ such that $A_{\theta^*} = a^*$ and $A_\theta = a'$ for all other θ.

Now suppose $\hat{a} \in \mathcal{A}$ has $\ell_{x'}(\hat{a}) > \ell_{x'}(a')$. Then taking $\mu = \pi$ puts all weights on experts predicting a' while keeping $D_\Phi(\mu, \pi) = 0$, so choosing $x = x'$ gives $M(\eta) < 0$ for all η. Otherwise, $\ell_{x'}(\hat{a}) = \ell_{x'}(a')$, which by eq. (16) implies $\ell_{x^*}(\hat{a}) \geq \ell_{x^*}(a')$.

Let $\mu^\alpha = \pi + \alpha(\delta_{\theta^*} - \pi)$, where δ_{θ^*} denotes the point distribution on θ^*. Calculating, we have

$$\begin{aligned} M(\eta) &= \langle \mu^\alpha, \ell_{x^*}(A) \rangle + \tfrac{1}{\eta} D_\Phi(\mu^\alpha, \pi) - \ell_{x^*}(\hat{a}) \\ &= (1-\alpha)\ell_{x^*}(a') + \alpha \ell_{x^*}(a^*) + \tfrac{1}{\eta} D_\Phi(\mu^\alpha, \pi) - \ell_{x^*}(\hat{a}) \\ &\leq (1-\alpha)\ell_{x^*}(\hat{a}) + \alpha \ell_{x^*}(a^*) + \tfrac{1}{\eta} D_\Phi(\mu^\alpha, \pi) - \ell_{x^*}(\hat{a}) \\ &= \alpha(\ell_{x^*}(a^*) - \ell_{x^*}(\hat{a})) + \tfrac{1}{\eta} D_f(\alpha, 0), \end{aligned}$$

where $f(\alpha) = \Phi(\mu^\alpha) = \Phi(\pi + \alpha(\delta_{\theta^*} - \pi))$. As $\nabla_\pi \Phi$ is bounded, so is $f'(0)$. Now as $\lim_{\epsilon \to 0} D_f(x+\epsilon, x)/\epsilon = 0$ for any scalar convex f with bounded $f'(x)$ (see e.g. [21, Theorem 24.1] and [23]), we see that for any $c > 0$ we have some $\alpha > 0$ such that $D_f(\alpha, 0) < c\alpha$. Taking $c = \eta(\ell_{x^*}(\hat{a}) - \ell_{x^*}(a^*)) > 0$ then gives $M(\eta) < 0$.

Legendre $\implies$ $\exists$ mixable ℓ:

Assuming Φ is Legendre, we need only show that some non-constant ℓ is Φ-mixable. As $\nabla_\pi \Phi$ is infinite on the boundary, π must be in the relative interior of Δ_Θ; otherwise $D_\Phi(\mu, \pi) = \infty$ for $\mu \neq \pi$.

Take $\mathcal{A} = \Delta_X$ and $\ell(p, x) = \|p - \delta_x\|^2$ to be the 2-norm squared loss. Now for all μ in the interior of Δ_Θ and $P \in \Delta_X^\Theta$, we have $\langle \mu, \ell_x(P) \rangle = \sum_\theta \mu_\theta \|P_\theta - \delta_x\|^2 \geq \|\bar{p} - \delta_x\|^2$ by convexity, where $\bar{p} = \sum_\theta \mu_\theta P_\theta$. In fact, as μ is in the interior, this inequality is strict, and remains so if replace μ by μ' with $\|\mu' - \mu\| < \epsilon$ for some ϵ sufficiently small. Now for all μ, P the algorithm can take $\hat{p} = \bar{p}$, and we can always choose $\eta = \inf_{x, \mu' : \|\mu' - \mu\| = \epsilon} D_\Phi(\mu', \mu)/(\epsilon \ell_{\max}) > 0$, so either $\|\mu - \pi\| < \epsilon$ in which case we are fine by the above, or μ is far enough away that the D_Φ term dominates the algorithm's loss. (Here $\ell_{\max}$ is just $\max_{p,x} \ell_x(p)$, which is bounded, and $D_\Phi(\mu', \mu) > 0$ as Φ is strictly convex.) So if Φ is Legendre, squared loss is Φ-mixable.