Transitions, Losses, and Re-parameterizations: Elements of Prediction Games

Parameswaran Kamalaruban

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

September 2017

Declaration

The results in the thesis were produced under the supervision of Bob Williamson, and partly in collaboration with Xinhua Zhang. However, the majority of the work, approximately $70 - 80\%$, is my own. The main contributions of this thesis are three related parts. The first part of the thesis on the asymmetric learning problems is based on intensive discussions and technical advice from Bob Williamson. The main technical results in the second part on exp-concave loss functions appeared as a conference paper with Bob Williamson and Xinhua Zhang [1]. The results were discussed with my supervisors Bob Williamson and Xinhua Zhang, who gave me advice and direction. The results on the acceleration of online optimization methods are work in progress and contained in an unpublished manuscript [3]. This third part of the thesis is based on some technical advice from Tim van Erven.

Some of the material in the thesis has already been published elsewhere in collaboration with others, the details of which are (Conference Proceedings and Preprints):

1. **Parameswaran Kamalaruban**, Robert Williamson, and Xinhua Zhang. Exp-Concavity of Proper Composite Losses. In *Proceedings of The 28th Conference on Learning Theory,* pages 1035–1065, 2015.

2. Kush Bhatia, Prateek Jain, **Parameswaran Kamalaruban**, and Purushottam Kar. Efficient and Consistent Robust Time Series Analysis. *arXiv:1607.00146 [cs.LG],* 2016. URL http://arxiv.org/abs/1607.00146.

3. **Parameswaran Kamalaruban**. Improved Optimistic Mirror Descent for Sparsity and Curvature. *arXiv:1609.02383 [cs.LG],* 2016. URL http://arxiv.org/abs/1609.02383.

Parameswaran Kamalaruban
25 September 2017

To my parents.

Acknowledgments

I would like to express my gratitude to all the people whose help, advice, and support made significant contributions to this thesis.

First, I would like to warmly acknowledge the valuable guidance and the continuous support of my primary supervisor, Bob Williamson. His wide perspective and insight were instrumental in steering my research over the course of my studies. I have learned so much from him and have thoroughly enjoyed our interactions. A most special thanks go to Xinhua Zhang, who is my co-supervisor. I am very grateful for his constant support, availability, patience, for his thoughtful comments, sharp insights, constructive critics, and discussions.

I would like to thank the Australian Government for its great support for research. I was very kindly supported by both the Australian National University and Data 61 (then NICTA). I thank both for creating a fantastic environment for research. I was lucky to have helpful colleagues for technical and general discussions, among them Aditya Krishna Menon, Richard Nock, and Brendan van Rooyen.

My special thanks to Tim van Erven, and Prateek Jain for hosting me kindheartedly while visiting their research groups. Thank you to both of them for stimulating technical discussions.

My very heartfelt thanks go to my friends in Canberra who shared a lot of laughter, debates, and ideas. Special thanks to my long-time friend and house-mate Ajanthan, and I treasure much of our friendly conversations on various topics.

I would like to express my deep gratitude to my siblings (Nathan and Manju), and my long-time friends in Sri Lanka (Aravinthan, Pathmayogan, Prakash, and Manorathan). I still feel touched by the great trust and affection they showed me.

And finally, deepest felt thanks to my parents for their unconditional love, and without their input, I would not be the person I am today!

Abstract

This thesis presents some geometric insights into three different types of two player prediction games – namely general learning task, prediction with expert advice, and online convex optimization. These games differ in the nature of the opponent (stochastic, adversarial, or intermediate), the order of the players' move, and the utility function. The insights shed some light on the understanding of the intrinsic barriers of the prediction problems and the design of computationally efficient learning algorithms with strong theoretical guarantees (such as generalizability, statistical consistency, and constant regret etc.). The main contributions of the thesis are:

- Leveraging concepts from statistical decision theory, we develop a necessary toolkit for formalizing the prediction games mentioned above and quantifying the objective of them.

- We investigate the cost-sensitive classification problem which is an instantiation of the general learning task, and demonstrate the hardness of this problem by producing the lower bounds on the minimax risk of it.

  Then we analyse the impact of imposing constraints (such as corruption level, and privacy requirements etc.) on the general learning task. This naturally leads us to further investigation of strong data processing inequalities which is a fundamental concept in information theory.

  Furthermore, by extending the hypothesis testing interpretation of standard privacy definitions, we propose an asymmetric (prioritized) privacy definition.

- We study efficient merging schemes for prediction with expert advice problem and the geometric properties (*mixability* and *exp-concavity*) of the loss functions that guarantee constant regret bounds. As a result of our study, we construct two types of link functions (one using calculus approach and another using geometric approach) that can re-parameterize any binary mixable loss into an exp-concave loss.

- We focus on some recent algorithms for online convex optimization, which exploit the *easy nature of the data* (such as sparsity, predictable sequences, and curved losses) in order to achieve better regret bound while ensuring the protection against the worst case scenario. We unify some of these existing techniques to obtain new update rules for the cases when these easy instances occur together, and analyse the regret bounds of them.

Contents

List of Figures

Chapter 1

Introduction

A well-posed *learning problem* can be stated as follows: A *learning algorithm* is said to learn from *experience* E with respect to some *task* T and some *performance measure* P, if its performance on T, as measured by P, improves with experience E (Mitchell [1997]). Pattern recognition, regression estimation and density estimation are the three main learning problems described by Vapnik [1998].

Developing learning algorithms is very challenging in complicated problem settings with very high dimensional datasets. These challenges are both *theoretical* (tight error bounds relative to the best hypothesis in the benchmark class, generalizability, and statistical consistency) and *computational* (efficient formulation of the optimization problem, optimal memory usage and running time). Generally, in the machine learning literature, these two challenges are considered independently. Understanding the connection between these two aspects of the learning problem to better understand the problem itself and to develop efficient learning algorithms, is an important and challenging research topic.

Several important problems in machine learning and statistics can be viewed as a two player prediction game between a decision maker and nature. This thesis presents some geometric insights into three different types of two player prediction games - namely general learning task, prediction with expert advice, and online convex optimization. These games differ in the nature of the opponent (stochastic, adversarial, or intermediate), the order of the players' move, mode of the game (batch or sequential), and the utility function. These insights shed some light on the understanding of the intrinsic barriers of the prediction problems and the design of computationally efficient learning algorithms with strong theoretical guarantees (such as generalizability, statistical consistency, and constant regret etc.).

There are many different objects which help us understanding the learning problems better. These include loss function, regularizer, information, risk measure, regret, and divergence. Systematically studying various representations (weighted average of primitive elements, variational and dual) of these objects and connections between them proves very useful in developing modular based solutions to learning problems (Reid and Williamson [2011]). Certain properties of these objects are necessary for strong theoretical guarantees, whereas some other properties are useful in developing computationally efficient learning algorithms. Thus by studying the geometric characterization of the problem w.r.t. these notions, we may be able to design solutions which

are computationally efficient as well as having strong theoretical guarantees.

The rest of this chapter provides the background to, and a road map for, the rest of this thesis.

1.1 Thesis Outline

Chapter 2 introduces the general learning task which covers many practical problems in machine learning and statistics as special instantiations. The goal of the learner is to find the functions which reflect relationships in data and thus best explain unseen data. Using the decision theoretic concepts, we set up an abstract language of transitions to formalize this general learning task. Then we define several quantities associated with the performance of a learning algorithm for this task such as conditional risk and full risk, and some measures of the hardness of the task such as minimum Bayes risk and minimax risk.

Next we consider a specific instantiation of the general learning task - namely multi-class probability estimation problem. Finally we discuss the binary class probability estimation problem or classification (which is an instantiation of the multi-class probability estimation problem with $m = 2$) in detail.

The next three chapters contain the contributions of this thesis.

Chapter 3 mainly deals with the cost-sensitive classification problem, which is also an instantiation of the general learning task. This problem plays a crucial role in mission critical machine learning applications. We study the hardness of this problem and emphasize the impact of cost terms on the hardness.

Chapter 3 investigates the intrinsic barriers of the general learning task subject to constraints such as privacy, noisy transmission (with minimum corruption level), and resource limitation. This naturally leads us to the investigation of strong data processing inequalities. Despite extensive investigation tracing back to the 1950's, the geometric insights of strong data processing inequalities are still not fully understood. A comprehensive survey paper providing an overview of strong data processing inequalities was written by Raginsky [2014]. We continue existing investigations on strong data processing inequalities, and make a significant progress in the direction of filling this gap by focusing on the weighted integral representation of f-divergences. This guides us in the channel design for cost-sensitive constrained problems. Furthermore we propose a cost-sensitive privacy definition by extending the standard local privacy definitions, and provide a hypothesis testing based interpretation for it.

Chapter 4 considers the classical problem of *prediction with expert advice* (Cesa-Bianchi and Lugosi [2006]), in which the goal of the learner is to predict as well as the best expert in a given pool of experts, on any sequence of T outcomes. This framework encompasses several applications as special cases (Vovk [1995]) such as classifier aggregation, weather prediction etc. The regret bound of the learner depends on the merging scheme used to merge the experts' predictions and the nature of the loss function used to measure the performance. This problem has been widely studied and $O(\sqrt{T})$ and $O(\log T)$ regret bounds can be achieved for convex losses (Zinkevich [2003]) and strictly convex losses with bounded first and second derivatives (Hazan et al.

[2007a]) respectively. In special cases like the Aggregating Algorithm (Vovk [1995]) with mixable losses and the Weighted Average Algorithm (Kivinen and Warmuth [1999]) with exp-concave losses, it is possible to achieve $O(1)$ regret bounds.

Even though exp-concavity trivially implies mixability, the converse implication is not true in general. Thus by understanding the underlying relationship between these two notions we can gain the best of both algorithms (strong theoretical performance guarantees of the Aggregating Algorithm and the computational efficiency of the Weighted Average Algorithm). We study the general conditions on mixable losses under which they can be transformed into an exp-concave loss through a suitable *link function*. Under mild conditions, we construct two types of link functions (one using calculus approach and another using geometric approach) that can re-parameterize any binary mixable loss into an exp-concave loss.

Chapter 5 focuses on the *online convex optimization* problem which plays a key role in machine learning as it has interesting theoretical implications and important practical applications especially in the large scale setting where computational efficiency is the main concern (Shalev-Shwartz [2011]). Early approaches to this problem were conservative, in which the main focus was protection against the worst case scenario. But recently several algorithms have been developed for tightening the regret bounds in *easy data* instances such as sparsity (Duchi et al. [2011]), predictable sequences (Chiang et al. [2012]), and curved losses (strongly-convex, exp-concave, mixable etc.) (Hazan et al. [2007b]).

We unify some of these existing techniques to obtain new update rules for the cases when these easy instances occur together. First we analyse an adaptive and optimistic update rule which achieves tighter regret bound when the loss sequence is sparse and predictable. Then we explain an update rule that dynamically adapts to the curvature of the loss function and utilizes the predictable nature of the loss sequence as well. Finally we extend these results to composite losses.

Finally, Chapter 6 contains the conclusion of this thesis, and a discussion of possibilities for further research. Chapter 6 concludes and contains a summary of the key contributions of this thesis.

The following work was completed during the thesis: Bhatia et al. [2016]. In this work, we present and analyze a polynomial-time algorithm for consistent estimation of regression coefficients under adversarial corruptions. But it has been excluded from the thesis as it does not fit as well with our theme.

Some definitions are repeated, and there are slight variations in notation for each chapter. Ultimately, there is no single best notational system, the effort has been placed into using the notation that best suits the contents of the chapter.

Chapter 2

Elements of Decision and Information Theory

The focus of this chapter is the abstract formulation of the general learning task, where a decision maker uses observations from experiments to inform her decisions. We present a rigorous mathematical language for making decisions under uncertainty, and quantifying the hardness of the problem. The concepts or results that we review here are based upon both classical works in decision theory [Blackwell, 1951; DeGroot, 1962; Le Cam, 1964; Von Neumann and Morgenstern, 1944; Wald, 1949] as well as recent contributions [Dawid, 2007; Grünwald and Dawid, 2004; Le Cam, 2012; Reid and Williamson, 2011; Torgersen, 1991]. They serve as a necessary background for the rest of the thesis.

The chapter proceeds as follows. In section 2.2 we introduce the general learning task, formalize it using the language of transitions, and define some decision theoretic measures associated with the hardness of the problem. Then in section 2.3 we study the multi-class probability estimation problem, which is a specific instantiation of the general learning task. Finally in section 2.4 we review more specific problem of binary class probability estimation in detail, by introducing several decision theoretic notions associated with it.

2.1 Notation and General Definitions

We require the following notation and definitions for chapter 2 and chapter 3. Other notation will be developed as necessary.

Vectors and Matrices The real numbers are denoted $\mathbb{R}$, the non-negative reals $\mathbb{R}_+$ and the extended reals $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$; the rules of arithmetic with extended real numbers and the need for them in convex analysis are explained by Rockafellar [1970]. The integers and non-negative integers are denoted by $\mathbb{Z}$ and $\mathbb{Z}_+$ respectively. A superscript prime, A' denotes transpose of the matrix or vector A, except when applied to a real-valued function where it denotes derivative (f'). We denote the matrix multiplication of compatible matrices A and B by $A \cdot B$, so the inner product of two vectors $x, y \in \mathbb{R}^n$ is $x' \cdot y$. Let $[n] := \{1, \ldots, n\}$, and the n-simplex

$\Delta^n := \{(p_1, \ldots, p_n)' : 0 \leq p_i \leq 1, \forall i \in [n], \text{and} \sum_{i\in[n]} p_i = 1\}$. If x is an n-vector, $A = \text{diag}(x)$ is the $n \times n$ matrix with entries $A_{ii} = x_i$, $i \in [n]$ and $A_{ij} = 0$ for $i \neq j$. We use e_i^n to denote the ith n-dimensional unit vector, $e_i^n = (\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{n-i})'$ when $i \in [n]$, and define $e_i^n = 0_n$ when $i > n$. The n-vector $\mathbf{1}_n := (\underbrace{1, \ldots, 1}_{n})'$.

Convexity A set $\mathcal{S} \subseteq \mathbb{R}^d$ is said to be *convex* if for all $\lambda \in [0, 1]$ and for all points $s_1, s_2 \in \mathcal{S}$ the point $\lambda s_1 + (1 - \lambda)s_2 \in \mathcal{S}$. A function $\phi : \mathcal{S} \to \mathbb{R}$ defined on a convex set $\mathcal{S}$ is said to be a (proper) convex function if for all $\lambda \in [0, 1]$ and points $s_1, s_2 \in \mathcal{S}$ the function ϕ satisfies

$$\phi\left(\lambda s_1 + (1 - \lambda)s_2\right) \leq \lambda\phi\left(s_1\right) + (1 - \lambda)\phi\left(s_2\right).$$

A function is said to be concave if $-\phi$ is convex.

Given a finite set S and a weight vector w, the *convex combination* of the elements of the set w.r.t the weight vector is denoted by $\text{co}_w S$, and the *convex hull* of the set which is the set of all possible convex combinations of the elements of the set is denoted by co$S$ (Rockafellar [1970]). If $S, T \subset \mathbb{R}^n$, then the *Minkowski sum* $S \oplus T := \{s + t : s \in S, t \in T\}$.

The Perspective Transform and the Csiszár Dual When $\phi : \mathbb{R}_+ \to \overline{\mathbb{R}}$ is convex, the perspective transform of ϕ is defined for $\tau \in \mathbb{R}_+$ via

$$I_\phi\left(s, \tau\right) := \begin{cases} \tau\phi\left(s/\tau\right) & \tau > 0, s > 0 \\ 0 & \tau = 0, s = 0 \\ \tau\phi\left(0\right) & \tau > 0, s = 0 \\ s\phi'_\infty & \tau = 0, s > 0, \end{cases}$$

where $\phi\left(0\right) := \lim_{s\to 0} \phi\left(s\right) \in \overline{\mathbb{R}}$ and ϕ'_∞ is the *slope at infinity* defined as

$$\phi'_\infty := \lim_{s\to+\infty} \frac{\phi\left(s_0 + s\right) - \phi\left(s_0\right)}{s} = \lim_{s\to+\infty} \frac{\phi\left(s\right)}{s}$$

for every $s_0 \in \mathcal{S}$ where $\phi\left(s_0\right)$ is finite. This slope at infinity is only finite when $\phi\left(s\right) = O\left(s\right)$, that is, when ϕ grows at most linearly as s increases. When ϕ'_∞ is finite it measures the slope of the linear asymptote. The function $I_\phi : [0, \infty)^2 \to \overline{\mathbb{R}}$ is convex in both arguments Hiriart-Urruty and Lemaréchal [1993] and may take on the value $+\infty$ when s or τ is zero. It is introduced here because it will form the basis of the f-divergences.

The perspective transform can be used to define the *Csiszár dual* $\phi^\diamond : [0, \infty) \to \overline{\mathbb{R}}$ of a convex function $\phi : \mathbb{R}_+ \to \overline{\mathbb{R}}$ by letting

$$\phi^\diamond\left(\tau\right) := I_\phi\left(1, \tau\right) = \tau\phi\left(\frac{1}{\tau}\right)$$

for all $\tau \in (0, \infty)$ and $\phi^\diamond(0) := \phi'_\infty$. The original ϕ can be recovered from I_ϕ since $\phi(s) = I_\phi(s, 1)$.

The convexity of the perspective transform I_ϕ in both its arguments guarantees the convexity of the dual $\phi^\diamond$. Some simple algebraic manipulation shows that for all $s, \tau \in \mathbb{R}_+$

$$I_\phi(s, \tau) = I_{\phi^\diamond}(\tau, s).$$

This observation leads to a natural definition of symmetry for convex functions. We will call a convex function $\diamond$-symmetric (or simply symmetric when the context is clear) when its perspective transform is symmetric in its arguments. That is, ϕ is $\diamond$-symmetric when $I_\phi(s, \tau) = I_\phi(\tau, s)$ for all $s, \tau \in [0, \infty)$. Equivalently, ϕ is $\diamond$-symmetric if and only if $\phi^\diamond = \phi$.

Probabilities and Expectations Let Ω be a measurable space and let μ be a probability measure on Ω. Ω^n denotes the product space $\Omega \times \cdots \times \Omega$ endowed with the product measure μ^n. The notation $\mathsf{X} \sim \mu$ means X is randomly drawn according to the distribution μ. $\mathbb{P}_\mu[E]$ and $\mathop{\mathbb{E}}_{\mathsf{X}\sim\mu}[f(\mathsf{X})]$ will denote the probability of a statistical event E and the expectation of a random variable $f(\mathsf{X})$ with respect to μ respectively. We will use capital letters $\mathsf{X}, \mathsf{Y}, \mathsf{Z}, \ldots$ for random variables and lower-case letters $x, y, z, \ldots$ for their observed values in a particular instance. We will denote by $\mathcal{P}(\mathcal{X})$ the set of all probability distributions on an alphabet $\mathcal{X}$ and by $\mathcal{P}_*(\mathcal{X})$ the subset of $\mathcal{P}(\mathcal{X})$ consisting of all strictly positive distributions.

Metric Spaces The Hamming distance on $\mathbb{R}^n$ is defined as

$$\rho_{\mathrm{Ha}}(x, x') := \sum_{i=1}^{n} [\![x_i \neq x'_i]\!], \tag{2.1}$$

where $[\![P]\!] = 1$ if P is true and $[\![P]\!] = 0$ otherwise. Define the p-norm of $x \in \mathbb{R}^n$ as

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}. \tag{2.2}$$

Let $\ell_p^n = \left(\mathbb{R}^n, \|\cdot\|_p\right)$ and B_p^n denote the unit ball of ℓ_p^n. ℓ_∞^n is $\mathbb{R}^n$ endowed with the norm

$$\|x\|_\infty := \sup_{1 \leq i \leq n} |x_i|. \tag{2.3}$$

Let $L_\infty(\Omega)$ be the set of bounded functions on Ω with respect to the norm

$$\|f\|_\infty := \sup_{\omega \in \Omega} |f(\omega)| \tag{2.4}$$

and denote its unit ball by $B(L_\infty(\Omega))$. For a probability measure μ on a measurable space Ω and $1 \leq p \leq \infty$, let $L_p(\mu)$ be the space of measurable functions on Ω with a

finite norm

$$\|f\|_{L_p(\mu)} := \left(\int |f|^p d\mu \right)^{1/p}. \tag{2.5}$$

$\mathcal{Y}^{\mathcal{X}}$ represents the set of all measurable functions $f : \mathcal{X} \to \mathcal{Y}$. For a set $\mathcal{X}$ define the functions $id_{\mathcal{X}}(x) = x$, and $1_{\mathcal{X}}(x) = 1$. The set of all real-valued measurable functions on $\mathcal{X}$ is denoted by $\mathbb{R}^{\mathcal{X}}$; $\mathbb{R}^{\mathcal{X}}_{++}$ and $\mathbb{R}^{\mathcal{X}}_{+}$ are the subsets of $\mathbb{R}^{\mathcal{X}}$ consisting of all strictly positive and nonnegative measurable functions, respectively. Define $\bar{c} := 1 - c$, for $c \in [0, 1]$. We write $x \wedge y := \min(x, y)$. A mapping $t \mapsto \text{sign}(t)$ is defined by

$$\text{sign}(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ -1 & \text{otherwise} \end{cases}.$$

Throughout this thesis all absolute constants are denoted by c, C, or K.

2.2 General Learning Task

A *general learning task* in statistical decision theory can be viewed as a two player game between the *decision maker* (statistician or learner) and *nature* (environment or opponent) as follows: Given the parameter space Θ, observation space $\mathcal{O}$, and decision space $\mathcal{A}$, and the loss function $\ell : \Theta \times \mathcal{A} \to \mathbb{R}_+$,

- Nature chooses $\theta \in \Theta$, and generates the data $\mathsf{O} \sim P_\theta \in \mathcal{P}(\mathcal{O})$, where P_θ is the distribution determined by the parameter θ,
- the decision maker observes the data O, makes her own decision $a \in \mathcal{A}$ (deterministic or stochastic), and incurs loss with $\ell(\theta, a)$.

Throughout the thesis we assume Θ to be finite and $\mathcal{A}$ to be closed, compact, set in order to provide a clear presentation by avoiding the measure theoretic complexities. This ensures that infimum of all the quantities defined can be replaced by minimum. Note that all the results presented in the thesis are applicable to general cases as well, under suitable regularity assumptions. Torgersen [1991] (Theorem 6.2.12) shows how results for finite Θ can be extended to those for infinite Θ.

In order to formalize the above game, we develop an *abstract language* using the decision theoretic concepts. We start with the central object of this language called a *transition*.

2.2.1 Markov Kernel

We define a *Markov kernel* (also known as a *transition* or a *channel*) as follows:

Definition 2.1 ([Le Cam, 2012; Torgersen, 1991])**.** *A Markov kernel from a finite set $\mathcal{X}$ to a finite set $\mathcal{Y}$ (denoted by $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$) is a function from $\mathcal{X}$ to $\mathcal{P}(\mathcal{Y})$, the set of probability distributions on $\mathcal{Y}$.*

A *Markov kernel* $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ acts on probability distributions $\mu \in \mathcal{P}(\mathcal{X})$ by

$$T \circ \mu := \mathop{\mathbb{E}}_{\mathsf{X} \sim \mu} [T(\mathsf{X})] \in \mathcal{P}(\mathcal{Y})$$

or on functions $f \in \mathbb{R}^{\mathcal{Y}}$ by

$$(Tf)(x) := \mathop{\mathbb{E}}_{Y \sim T(x)} [f(Y)], \quad x \in \mathcal{X}.$$

The composition of two Markov kernels $T_1 : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $T_2 : \mathcal{Y} \rightsquigarrow \mathcal{Z}$, denoted by $T_2 T_1 : \mathcal{X} \rightsquigarrow \mathcal{Z}$, is defined by

$$T_2 T_1 f \;=\; T_1 (T_2 f), \quad f \in \mathbb{R}^{\mathcal{Z}}.$$

Denote the set of all Markov kernels from $\mathcal{X}$ to $\mathcal{Y}$ by $\mathcal{M}(\mathcal{X}, \mathcal{Y})$. If $\mathcal{X}$ and $\mathcal{Y}$ are finite, we can represent the distributions $P \in \mathcal{P}(\mathcal{X})$ by vectors in $\mathbb{R}^{|\mathcal{X}|}$, Markov kernels $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ by column stochastic matrices ($|\mathcal{Y}| \times |\mathcal{X}|$ positive matrices where the sum of all entries in each column is equal to 1), and composition by matrix multiplication. We can also verify that $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ is a closed convex subset of $\mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$, the set of all $|\mathcal{Y}| \times |\mathcal{X}|$ matrices. Note that the transition $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ induces a class of probability measures $\mathcal{P}_T(\mathcal{Y}) := \{P_x := T(x) \in \mathcal{P}(\mathcal{Y}) : x \in \mathcal{X}\}$. For a transition $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$, define $T(y \mid x) := \mathbb{P}_{T(x)}[\mathsf{Y} = y]$, where $T(x) \in \mathcal{P}_T(\mathcal{Y})$.

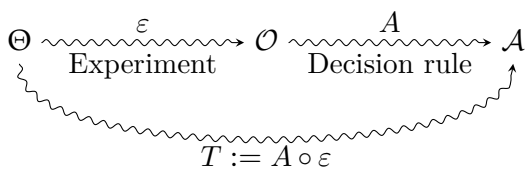
A function $f : \mathcal{X} \to \mathcal{Y}$ induces a Markov kernel $F : \mathcal{X} \rightsquigarrow \mathcal{Y}$ with $F(x) = \delta_{f(x)}$, a point mass distribution on $f(x)$. For every measure space $\mathcal{X}$, there are two special Markov kernels, the *completely informative* Markov kernel induced from the identity function $\mathrm{id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}$ (where $\mathrm{id}_{\mathcal{X}}(x) = x$), and the *completely uninformative* Markov kernel induced from the function $\bullet_{\mathcal{X}} : \mathcal{X} \to \bullet$ (where $\bullet_{\mathcal{X}}(x) = \bullet$, $\forall x \in \mathcal{X}$ and $\bullet \in \mathcal{Y}$).

Given $\mu \in \mathcal{P}(\mathcal{X})$, and $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$, let $D := \mu \otimes T \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denotes the joint probability measure of $(\mathsf{X}, \mathsf{Y}) \in \mathcal{X} \times \mathcal{Y}$ with $\mathbb{P}_D[\mathsf{X} = x] = \mathbb{P}_\mu[\mathsf{X} = x]$, and $\mathbb{P}_D[\mathsf{Y} = y \mid \mathsf{X} = x] = \mathbb{P}_{T(x)}[\mathsf{Y} = y]$.

We will now use this abstract language of transitions to formulate the general learning task introduced in the beginning of this section. This will enable us to analyse the intrinsic barriers or capacity of the task in a more generic way. Later, by using appropriate instantiations, we will derive important practical problems in machine learning and statistics.

2.2.2 Decision Theoretic Notions

The general learning task described above can be represented by the following transition diagram:



, (2.6)

where

- *Experiment* (denoted by $\varepsilon : \Theta \rightsquigarrow \mathcal{O}$) is a Markov kernel from the parameter space Θ to the observation space $\mathcal{O}$. If the true hypothesis is $\theta \in \Theta$, then the observed data is distributed according the probability measure $\varepsilon(\theta)$. The class of probability measures associated with this experiment is given by $\mathcal{P}_\varepsilon := \{P_\theta := \varepsilon(\theta) : \theta \in \Theta\}$.
- Stochastic *Decision rule* (denoted by $A : \mathcal{O} \rightsquigarrow \mathcal{A}$) is a Markov kernel from the observation space $\mathcal{O}$ to the action space $\mathcal{A}$. Upon observing data $o \in \mathcal{O}$, the learner will choose an action in $\mathcal{A}$ according to the distribution $A(o)$.

Remark 2.2. *We will depict the transitions (experiment and decision rule) associated with the learning task in a transition diagram, and we call it the 'transition diagram representation of the learning task' throughout the thesis.*

Loss and Regret: The quality of the composite relation $T := A \circ \varepsilon : \Theta \rightsquigarrow \mathcal{A}$ is measured by a loss function

$$\ell : \Theta \times \mathcal{A} \ni (\theta, a) \mapsto \ell(\theta, a) \in \mathbb{R}_+. \tag{2.7}$$

The general learning task can more compactly be represented as the pair (ℓ, ε) where $\mathcal{A}, \Theta, \mathcal{O}$ can be inferred from the type signatures of ℓ and ε. We usually encounter the loss relative to the best action defined formally as the *regret*

$$\Delta\ell : \Theta \times \mathcal{A} \ni (\theta, a) \;\mapsto\; \Delta\ell(\theta, a) := \ell(\theta, a) - \inf_{a' \in \mathcal{A}} \ell(\theta, a') \in \mathbb{R}_+. \tag{2.8}$$

Conditional Risk: The quality of the final action chosen by the decision maker when they use the composite relation $T : \Theta \rightsquigarrow \mathcal{A}$ (in fact the stochastic decision rule $A : \mathcal{O} \rightsquigarrow \mathcal{A}$ for a given experiment $\varepsilon : \Theta \rightsquigarrow \mathcal{O}$) can be evaluated using the notion of *conditional risk* (defined with an overloaded notation for the loss):

$$\ell : \Theta \times \mathcal{M}(\Theta, \mathcal{A}) \ni (\theta, T) \;\mapsto\; \ell(\theta, T) := \mathop{\mathbb{E}}_{\mathsf{A} \sim T(\theta)} [\ell(\theta, \mathsf{A})] \in \mathbb{R}_+, \tag{2.9}$$

where the term inside the expectation is the loss (2.7) of a random variable A when the true parameter is θ. We use the overloaded notation with a reason, which will become clear in section 2.3. Similarly we can define the conditional risk in terms of regret as follows (again with an overloaded notation for the regret):

$$\Delta\ell : \Theta \times \mathcal{M}(\Theta, \mathcal{A}) \ni (\theta, T) \;\mapsto\; \Delta\ell(\theta, T) := \mathop{\mathbb{E}}_{\mathsf{A} \sim T(\theta)} [\Delta\ell(\theta, \mathsf{A})] \in \mathbb{R}_+, \tag{2.10}$$

where the term inside the expectation is the regret (2.8) of a random variable A when the true parameter is θ.

For any fixed (unknown) parameter $\theta \in \Theta$, we can calculate the conditional risk of any composite relation T, and the goal is to find an optimal composite relation (in fact

an optimal stochastic decision rule for a given experiment). Two main approaches to find the best composite relation (or the best decision rule) are:

- *Bayesian approach* (average case analysis), which is more appropriate if the decision maker has some intuition about θ, given in the form of a prior probability distribution π, and
- *Minimax approach* (worst case analysis), which is more appropriate if the decision maker has no prior knowledge concerning θ.

The *conditional Bayesian risk* and *conditional max risk* are defined as,

$$L_\ell : \mathcal{P}(\Theta) \times \mathcal{M}(\Theta, \mathcal{A}) \ni (p, T) \;\mapsto\; L_\ell(p, T) := \mathop{\mathbb{E}}_{\mathsf{Y}\sim p} [\ell(\mathsf{Y}, T)] \in \mathbb{R}_+, \text{ and} \tag{2.11}$$

$$L^\star_\ell : \mathcal{M}(\Theta, \mathcal{A}) \ni T \;\mapsto\; L^\star_\ell(T) := \sup_{\theta\in\Theta} \ell(\theta, T) \in \mathbb{R}_+, \tag{2.12}$$

respectively. We measure the difficulty of the general learning task by the *conditional minimum Bayesian risk* and *conditional minimax risk* defined as,

$$\underline{L}_\ell : \mathcal{P}(\Theta) \ni p \;\mapsto\; \underline{L}_\ell(p) := \inf_{T\in\mathcal{M}(\Theta,\mathcal{A})} L_\ell(p, T) \in \mathbb{R}_+, \text{ and} \tag{2.13}$$

$$\underline{L}^\star_\ell : \cdot \;\mapsto\; \underline{L}^\star_\ell := \inf_{T\in\mathcal{M}(\Theta,\mathcal{A})} L^\star_\ell(T) \in \mathbb{R}_+, \tag{2.14}$$

respectively.

Remark 2.3. *By replacing ℓ by $\Delta\ell$ in* (2.11)*,* (2.12)*,* (2.13)*, and* (2.14) *, we obtain $L_{\Delta\ell}$, $L^\star_{\Delta\ell}$, $\underline{L}_{\Delta\ell}$ and $\underline{L}^\star_{\Delta\ell}$ respectively. One can do this transformation for all the concepts that we introduce below and obtain the 'regret' based notions.*

Full Risk: In the conditional quantities defined above, we have abstracted away the observation space $\mathcal{O}$ (i.e. no data setting). Now we consider the practical scenario with observations, and define the *full risk* of a stochastic decision rule $A : \mathcal{O} \rightsquigarrow \mathcal{A}$ as follows

$$\mathbb{L}_\ell : \Theta \times \mathcal{M}(\Theta, \mathcal{O}) \times \mathcal{M}(\mathcal{O}, \mathcal{A}) \ni (\theta, \varepsilon, A) \;\mapsto$$
$$\mathbb{L}_\ell(\theta, \varepsilon, A) := \ell(\theta, A \circ \varepsilon) = \mathop{\mathbb{E}}_{\mathsf{O}\sim\varepsilon(\theta)} \left[\mathop{\mathbb{E}}_{\mathsf{A}\sim A(\mathsf{O})} [\ell(\theta, \mathsf{A})] \right] \in \mathbb{R}_+, \tag{2.15}$$

where $\ell(\theta, A \circ \varepsilon)$ is the conditional risk (2.9) of the composite relation $A \circ \varepsilon$. Note that $A : \mathcal{O} \rightsquigarrow \mathcal{A}$ is a function of the observation in $\mathcal{O}$ which is distributed according the probability distribution associated with a parameter in Θ.

As in the conditional case, we define the *full Bayesian risk*, *full minimum Bayesian risk*, *full max risk*, and *full minimax risk* as follows:

$$\mathcal{R}_\ell : \mathcal{P}(\Theta) \times \mathcal{M}(\Theta, \mathcal{O}) \times \mathcal{M}(\mathcal{O}, \mathcal{A}) \ni (\pi, \varepsilon, A) \;\mapsto\; \mathcal{R}_\ell(\pi, \varepsilon, A) :=$$
$$\mathop{\mathbb{E}}_{\mathsf{Y}\sim\pi} [\mathbb{L}_\ell(\mathsf{Y}, \varepsilon, A)] \in \mathbb{R}_+,$$
$$\underline{\mathcal{R}}_\ell : \mathcal{P}(\Theta) \times \mathcal{M}(\Theta, \mathcal{O}) \ni (\pi, \varepsilon) \;\mapsto\; \underline{\mathcal{R}}_\ell(\pi, \varepsilon) :=$$

$$\begin{aligned}
& & & \inf_{A\in\mathcal{M}(\mathcal{O},\mathcal{A})} \mathcal{R}_\ell\left(\pi,\varepsilon,A\right)\in\mathbb{R}_+,\\
\mathcal{R}^\star_\ell : \mathcal{M}\left(\Theta,\mathcal{O}\right)\times\mathcal{M}\left(\mathcal{O},\mathcal{A}\right)\ni\left(\varepsilon,A\right) & \mapsto & & \mathcal{R}^\star_\ell\left(\varepsilon,A\right):=\\
& & & \sup_{\theta\in\Theta}\mathbb{L}_\ell\left(\theta,\varepsilon,A\right)\in\mathbb{R}_+,\text{ and}\\
\underline{\mathcal{R}}^\star_\ell : \mathcal{M}\left(\Theta,\mathcal{O}\right)\ni\varepsilon & \mapsto & & \underline{\mathcal{R}}^\star_\ell\left(\varepsilon\right):=\\
& & & \inf_{A\in\mathcal{M}(\mathcal{O},\mathcal{A})} \mathcal{R}^\star_\ell\left(\varepsilon,A\right)\in\mathbb{R}_+
\end{aligned}$$

respectively.

Let Y and O be random variables over Θ and $\mathcal{O}$ respectively. Also let $\theta\in\Theta$ and $o\in\mathcal{O}$. The experiment ε (in (2.6)) and a prior π on Θ induces a joint probability measure D on $\Theta\times\mathcal{O}$ and thus a transition $\eta_D:\mathcal{O}\rightsquigarrow\Theta$ (given by $\eta_D\left(\theta\mid o\right):=\mathbb{P}_D\left[\mathsf{Y}=\theta\mid\mathsf{O}=o\right]$) and a marginal distribution M_D on $\mathcal{O}$ (given by $M_D\left(o\right):=\mathbb{P}_D\left[\mathsf{O}=o\right]$). That is if $\Theta\times\mathcal{O}\ni\left(\mathsf{Y},\mathsf{O}\right)\sim D$, then we have

$$\begin{aligned}
\mathbb{P}_D\left[\mathsf{Y}=\theta,\mathsf{O}=o\right] &= \mathbb{P}_D\left[\mathsf{Y}=\theta\right]\cdot\mathbb{P}_D\left[\mathsf{O}=o\mid\mathsf{Y}=\theta\right] &= \pi\left(\theta\right)\cdot\varepsilon\left(o\mid\theta\right)\\
&= \mathbb{P}_D\left[\mathsf{O}=o\right]\cdot\mathbb{P}_D\left[\mathsf{Y}=\theta\mid\mathsf{O}=o\right] &= M_D\left(o\right)\cdot\eta_D\left(\theta\mid o\right).
\end{aligned}$$

Thus we can use the pairs (π,ε) and (M,η) interchangeably. We can define the full Bayesian risk and full minimum Bayesian risk in terms of (M,η) as follows:

$$\begin{aligned}
\widehat{\mathcal{R}}_\ell : \mathcal{P}\left(\mathcal{O}\right)\times\mathcal{M}\left(\mathcal{O},\Theta\right)\times\mathcal{M}\left(\mathcal{O},\mathcal{A}\right)\ni\left(M,\eta,A\right) & \mapsto & & \widehat{\mathcal{R}}_\ell\left(M,\eta,A\right):=\\
& & & \mathop{\mathbb{E}}_{\mathsf{O}\sim M}\left[\mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta(\mathsf{O})}\left[\ell\left(\mathsf{Y},A\left(\mathsf{O}\right)\right)\right]\right]\in\mathbb{R}_+\\
\underline{\widehat{\mathcal{R}}}_\ell : \mathcal{P}\left(\mathcal{O}\right)\times\mathcal{M}\left(\mathcal{O},\Theta\right)\ni\left(M,\eta\right) & \mapsto & & \underline{\widehat{\mathcal{R}}}_\ell\left(M,\eta\right):=\\
& & & \inf_{A\in\mathcal{M}(\mathcal{O},\mathcal{A})}\widehat{\mathcal{R}}_\ell\left(M,\eta,A\right)\in\mathbb{R}_+
\end{aligned}$$

At this point we note the following facts:

- Since

$$\mathop{\mathbb{E}}_{(\mathsf{Y},\mathsf{O})\sim D}\left[\ell\left(\mathsf{Y},A\left(\mathsf{O}\right)\right)\right]=\mathop{\mathbb{E}}_{\mathsf{O}\sim M}\left[\mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta(\mathsf{O})}\left[\ell\left(\mathsf{Y},A\left(\mathsf{O}\right)\right)\right]\right]=\mathop{\mathbb{E}}_{\mathsf{Y}\sim\pi}\left[\mathop{\mathbb{E}}_{\mathsf{O}\sim\varepsilon(\mathsf{Y})}\left[\ell\left(\mathsf{Y},A\left(\mathsf{O}\right)\right)\right]\right]$$

  we have that $\widehat{\mathcal{R}}_\ell\left(M,\eta,A\right)=\mathcal{R}_\ell\left(\pi,\varepsilon,A\right)$ and $\underline{\widehat{\mathcal{R}}}_\ell\left(M,\eta\right)=\underline{\mathcal{R}}_\ell\left(\pi,\varepsilon\right)$.
- $\widehat{\mathcal{R}}_\ell\left(M,\eta,A\right)=\mathop{\mathbb{E}}_{\mathsf{O}\sim M}\left[L_\ell\left(\eta\left(\mathsf{O}\right),A\left(\mathsf{O}\right)\right)\right]$ and $\underline{\widehat{\mathcal{R}}}_\ell\left(M,\eta\right)=\mathop{\mathbb{E}}_{\mathsf{O}\sim M}\left[\underline{L}_\ell\left(\eta\left(\mathsf{O}\right)\right)\right]$.

By using the minimax theorem (Komiya [1988]), we obtain the following result that relates the full minimum Bayesian risk and the full minimax risk.

Theorem 2.4. *Let Θ to be finite and $\mathcal{A}$ to be closed, compact, set with ℓ a continuous function. Then for all experiments ε,*

$$\underline{\mathcal{R}}^\star_\ell\left(\varepsilon\right)=\sup_{\pi\in\mathcal{P}(\Theta)}\underline{\mathcal{R}}_\ell\left(\pi,\varepsilon\right).$$

2.2.3 Repeated and Parallelized Transitions

Transitions can be *repeated*. For $P, Q \in \mathcal{P}(\mathcal{X})$, denote the product distribution by $P \otimes Q$. For any transition $T \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ we denote the *repeated transition* $T_n \in \mathcal{M}(\mathcal{X}, \mathcal{Y}^n), n \in \mathbb{Z}_+$, with,

$$T_n(x) = T(x) \otimes \cdots \otimes T(x) = T(x)^n, \tag{2.16}$$

the n-fold product of $T(x)$. Note that the transition T_n induces a probability space $\mathcal{P}_{T_n}(\mathcal{Y}^n) := \{P_x^n := T(x)^n \in \mathcal{P}(\mathcal{Y})^n : x \in \mathcal{X}\}$.

Transitions can also be combined in parallel. If $T_i \in \mathcal{M}(\mathcal{X}_i, \mathcal{Y}_i), i \in [k]$, are transitions then denote,

$$\bigotimes_{i=1}^{k} T_i \in \mathcal{M}\left(\times_{i=1}^{k} \mathcal{X}_i, \times_{i=1}^{k} \mathcal{Y}_i\right) \tag{2.17}$$

with $\bigotimes_{i=1}^{k} T_i(x) = T_1(x_1) \otimes \cdots \otimes T_n(x_n)$. For any transition $T \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ we denote the *parallelized transition* $T_{1:n} \in \mathcal{M}(\mathcal{X}^n, \mathcal{Y}^n), n \in \mathbb{Z}_+$, with,

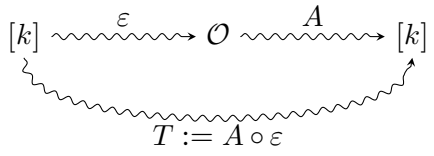
$$T_{1:n}(x) = \bigotimes_{i=1}^{n} T(x). \tag{2.18}$$

2.3 Multi-Class Probability Estimation Problem

We will now consider the special case when the prediction space is $\Theta = [k]$, and the action space is also $\mathcal{A} = [k]$. In this case, the loss function is written as

$$\ell : [k] \times [k] \ni (y, \widehat{y}) \;\mapsto\; \ell(y, \widehat{y}) \in \mathbb{R}_+. \tag{2.19}$$

The resulting problem is called the *k-class probability estimation (CPE)* problem (ℓ, ε) and can be represented by the following transition diagram:



. (2.20)

Define $T := A \circ \varepsilon : [k] \rightsquigarrow [k]$. As in the general learning problem, we define the conditional risk as follows (with overloaded notation):

$$\ell : [k] \times \mathcal{M}([k], [k]) \ni (y, T) \;\mapsto\; \ell(y, T) := \mathop{\mathbb{E}}_{\mathsf{Y} \sim T(y)} [\ell(y, \mathsf{Y})] \in \mathbb{R}_+, \tag{2.21}$$

where term inside the expectation is the loss (2.19) of a random variable Y given that the actual parameter is y. In this setting, it is common in the literature to refer the conditional risk as the loss function of the problem (it is also said to be *multi-CPE*

loss), and that's why we purposefully use overloaded notation for them. In fact, in Chapter 4 we call the conditional risk as the loss function of prediction with expert advice problem.

The conditional Bayesian risk and the conditional minimum Bayesian risk of this k-class probability estimation problem can be written as follows:

$$L_\ell : \Delta^k \times \mathcal{M}\left([k],[k]\right) \ni (p,T) \;\mapsto\; L_\ell\left(p,T\right) := \underset{\mathsf{Y}\sim p}{\mathbb{E}}\left[\ell\left(\mathsf{Y},T\right)\right] \in \mathbb{R}_+, \text{ and} \tag{2.22}$$

$$\underline{L}_\ell : \Delta^k \ni p \;\mapsto\; \underline{L}_\ell\left(p\right) := \inf_{T\in\mathcal{M}([k],[k])} L_\ell\left(p,T\right) \in \mathbb{R}_+ \tag{2.23}$$

respectively.

2.4 Binary Experiments

In this section we consider the k-class probability estimation problem with $k = 2$. Such a problem is known as a binary experiment. Here we review some important notions associated with the binary experiments such as loss, risk, ROC (Receiver Operating Characteristic) curves, information, and distance or divergence between probability distributions.

For consistency with much of the literature, we let $\Theta = \{1,2\}$, $P = \varepsilon\left(1\right)$, and $Q = \varepsilon\left(2\right)$. Thus a binary experiment can be simply represented (P,Q). The densities of P and Q with respect to some third reference distribution M over $\mathcal{O}$ will be defined by $dP = pdM$ and $dQ = qdM$ respectively. A central statistic in the study of binary experiments and statistical hypothesis testing is the likelihood ratio dP/dQ.

2.4.1 Hypothesis Testing

In the context of a binary experiment (P,Q), a *statistical test* is any function that assigns each instance $o \in \mathcal{O}$ to either P or Q. We will use the labels 1 and 2 for P and Q respectively and so a statistical test is any function $r : \mathcal{O} \to \{1,2\}$. The *classification rates* defined by a given test r are:

1. True positive rate $\mathrm{TP}_r := P\left(\mathcal{O}_r^1\right)$
2. True negative rate $\mathrm{TN}_r := Q\left(\mathcal{O}_r^2\right)$
3. False positive rate $\mathrm{FP}_r := Q\left(\mathcal{O}_r^1\right)$
4. False negative rate $\mathrm{FN}_r := P\left(\mathcal{O}_r^2\right)$

where $\mathcal{O}_r^1 := \{o \in \mathcal{O} : r\left(o\right) = 1\}$ and $\mathcal{O}_r^2 := \{o \in \mathcal{O} : r\left(o\right) = 2\}$. Since P and Q are distributions over $\mathcal{O} = \mathcal{O}_r^1 \cup \mathcal{O}_r^2$ and the positive and negative sets are disjoint we have that $\mathrm{TP} + \mathrm{FN} = 1$ and $\mathrm{FP} + \mathrm{TN} = 1$.

For a given binary experiment (P,Q), we define the following important quantities or notions associated with a statistical test r:

- The *power* $\beta_r := \mathrm{TP}_r$.

- The *size* $\alpha_r := \mathrm{FP}_r$.
- A test r is said to be the *most powerful* (MP) test of size $\alpha \in [0,1]$ if, $\alpha_r = \alpha$ and for all other tests r' such that $\alpha_{r'} \leq \alpha$ we have $1 - \beta_r \leq 1 - \beta_{r'}$.
- The *Neyman-Pearson function for the dichotomy* (P, Q) (Torgersen [1991])

$$\beta(\alpha) = \beta(\alpha, P, Q) := \sup_{r \in \{1,2\}^{\mathcal{O}}} \{\beta_r : \alpha_r \leq \alpha\}.$$

2.4.2 ROC curves

Often, statistical tests are obtained by applying a threshold τ_0 to a real-valued *test statistic* $\tau : \mathcal{O} \to \mathbb{R}$. In this case, the statistical test is $r(o) = 2 - [\![\tau(o) \geq \tau_0]\!]$. This leads to parameterized forms of prediction sets $\mathcal{O}^y_\tau(\tau_0) := \mathcal{O}^y_{[\![\tau \geq \tau_0]\!]}$ for $y \in \{1,2\}$, and the classification rates $\mathrm{TP}_\tau(\tau_0)$, $\mathrm{FP}_\tau(\tau_0)$, $\mathrm{FN}_\tau(\tau_0)$, and $\mathrm{TN}_\tau(\tau_0)$ which are defined analogously. By varying the threshold parameter a range of classification rates can be achieved. This observation leads to a well known graphical representation of test statistics known as the *receiver operating characteristic (ROC) curve*.

An ROC curve for the test statistic τ is simply a plot of the true positive rate of these classifiers as a function of their false positive rate as the threshold τ_0 varies over $\mathbb{R}$. Formally,

$$\mathrm{ROC}(\tau) := \{(\mathrm{FP}_\tau(\tau_0), \mathrm{TP}_\tau(\tau_0)) : \tau_0 \in \mathbb{R}\} \subset [0,1]^2.$$

A graphical example of an ROC curve is shown as the solid black line in Figure 2.1.

The Neyman-Pearson lemma (Neyman and Pearson [1933]) shows that for a fixed experiment (P, Q), the likelihood ratio $\tau^\star(o) = dP/dQ(o)$ is the most powerful test statistic for each choice of threshold τ_0. This guarantees that the ROC curve for the likelihood ratio $\tau^\star = dP/dQ$ will lie above, or *dominate*, that of any other test statistic τ as shown in Figure 2.1. This is an immediate consequence of the likelihood ratio being the most powerful test since for each false positive rate (or size) α it will have the largest true positive rate (or power) β of all tests (Eguchi and Copas [2001]). Thus $\mathrm{ROC}(dP/dQ)$ is the maximal ROC curve.

2.4.3 f-Divergences

The hardness of the binary classification problem depends on the distinguish-ability of the two probability distributions associated with it. The class of f-divergences ([Ali and Silvey, 1966; Csiszár, 1972]) provide a rich set of relations that can be used to measure the separation of the distributions in a binary experiment.

Definition 2.5. *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. For all distributions $P, Q \in \mathcal{P}(\mathcal{O})$ the f-divergence between P and Q is,*

$$\mathbb{I}_f(P, Q) = \mathop{\mathbb{E}}_{Q}\left[f\left(\frac{dP}{dQ}\right)\right] = \int_{\mathcal{O}} f\left(\frac{dP}{dQ}\right) dQ$$

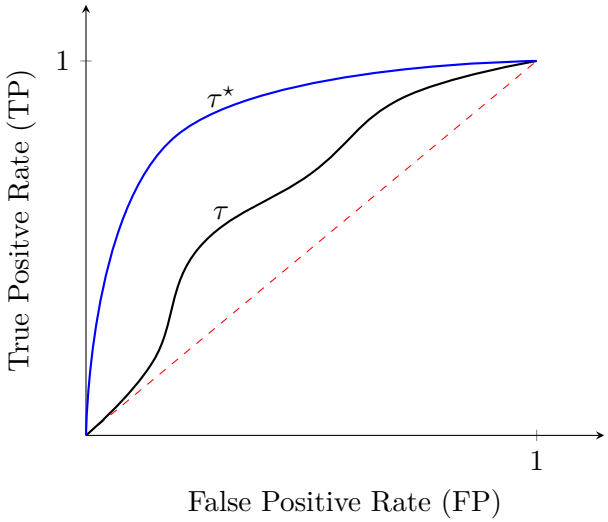


Figure 2.1: ROC curve for (a) an arbitrary statistical test τ (middle, black curve), (b) an optimal statistical test $\tau^\star$ (top, blue curve), and (c) an uninformative statistical test (dashed red line).

when P is absolutely continuous with respect to Q and equals ∞ otherwise.

The behavior of f is not specified at the endpoints of $(0, \infty)$ in the above definition. This is remedied via the perspective transform of f, which defines the limiting behavior of f. Given convex $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$ the *f-divergence of P from Q* is

$$\mathbb{I}_f(P, Q) := \mathop{\mathbb{E}}_{M}[I_f(p, q)] = \mathop{\mathbb{E}}_{O \sim M}[I_f(p(O), q(O))], \tag{2.24}$$

where I_f is the perspective transform of f.

Many commonly used divergences in probability, mathematical statistics and information theory are special cases of f-divergences. For example:

1. The Kullback-Leibler divergence (with $\mathrm{KL}(u) = u \log u$)

$$\mathbb{I}_{\mathrm{KL}}(P, Q) = D(P \,||\, Q) = \mathop{\mathbb{E}}_{Q}\left[\frac{dP}{dQ} \log \frac{dP}{dQ}\right]$$

2. The total variation distance (with $\mathrm{TV}(u) = |u - 1|$)

$$\mathbb{I}_{\mathrm{TV}}(P, Q) = d_{\mathrm{TV}}(P, Q) = \mathop{\mathbb{E}}_{Q}\left[\left|\frac{dP}{dQ} - 1\right|\right].$$

   Also for general measures μ and ν on $\mathcal{O}$, we define $d_{\mathrm{TV}}(\mu, \nu) = \int |d\mu - d\nu|$.

3. The χ^2-divergence (with $\chi^2(u) = (u-1)^2$)

$$\mathbb{I}_{\chi^2}(P, Q) = \chi^2(P \,||\, Q) = \mathop{\mathbb{E}}_{Q}\left[\left(\frac{dP}{dQ} - 1\right)^2\right]$$

4. The squared Hellinger distance (with $\mathrm{He}^2(u) = (\sqrt{u}-1)^2$)

$$\mathbb{I}_{\mathrm{He}^2}(P,Q) = \mathrm{He}^2(P,Q) = \mathop{\mathbb{E}}_{Q}\left[\left(\sqrt{\frac{dP}{dQ}}-1\right)^2\right]$$

We note the following properties of f-divergences:

- $\mathbb{I}_f(P,Q) \geq 0$ for all P and Q by Jensen's inequality
- $\mathbb{I}_f(Q,Q) = 0$ for all distributions Q since $f(1) = 0$
- $\mathbb{I}_f(P,Q) = \mathbb{I}_{f^\diamond}(Q,P)$ for all distributions P and Q (where $f^\diamond$ is the Csiszár dual of f) due to the symmetry of the perspective I_f. An f-divergence is symmetric if $\mathbb{I}_f(P,Q) = \mathbb{I}_f(Q,P)$ for all P, Q.
- Let $f : (0,\infty) \to \mathbb{R}$ be a convex function. Then for each $a, b \in \mathbb{R}$ the convex function $g(x) := f(x) + ax + b$ satisfies $\mathbb{I}_g(P,Q) = \mathbb{I}_f(P,Q)$ for all P and Q.
- The weak data processing theorem states that for all sets $\mathcal{O}, \widehat{\mathcal{O}}$, all transitions $T \in \mathcal{M}\left(\mathcal{O}, \widehat{\mathcal{O}}\right)$, all distributions $P, Q \in \mathcal{P}(\mathcal{O})$ and all f-divergences,

$$\mathbb{I}_f(T \circ P, T \circ Q) \leq \mathbb{I}_f(P,Q).$$

  Intuitively, adding noise never makes it easier to distinguish P and Q.

Remark 2.6. *Here we give a more general definition for f-divergence. Let $\phi : [0,\infty)^k \to \mathbb{R}$ be a convex function with $\phi(\mathbf{1}_k) = 0$, for some $k \in \mathbb{Z}_+$. For all experiments $\varepsilon : [k] \rightsquigarrow \mathcal{X}$ (with the parameter space $[k]$ and the observation space $\mathcal{X}$) the f-divergence of the experiment ε is,*

$$\mathbb{I}_\phi(\varepsilon) \;:=\; \mathop{\mathbb{E}}_{X \sim \varepsilon(k)}\left[\phi\left(t\left(X\right)\right)\right], \tag{2.25}$$

where $t : \mathcal{X} \to [0,\infty)^k$ is given by

$$t(x) \;:=\; \left(\frac{d\varepsilon(x \mid 1)}{d\varepsilon(x \mid k)}, \cdots, \frac{d\varepsilon(x \mid i)}{d\varepsilon(x \mid k)}, \cdots, 1\right), \text{ for } x \in \mathcal{X}.$$

By defining $f(t) := \phi(t,1)$ (with $k = 2$), we recover the binary f-divergence (Definition 2.5) for binary experiments $\varepsilon : [2] \rightsquigarrow \mathcal{X}$.

Integral Representations of f-divergences: Representation of f-divergences and loss functions as weighted average of *primitive* components (in the sense that they can be used to express other measures but themselves cannot be so expressed) is very useful in studying certain geometric properties of them using the weight function behavior. The following restatement of a theorem by Liese and Vajda [2006] provides such a representation for any f-divergence (confer Reid and Williamson [2011] for a proof):

Theorem 2.7. *Define $\overline{c} := 1 - c$, for $c \in [0,1]$, and let f be convex such that $f(1) = 0$. Then the f-divergence between P and Q can be written in a weighted integral form as follows:*

$$\mathbb{I}_f(P,Q) = \int_0^1 \mathbb{I}_{f_c}(P,Q)\,\gamma_f(c)\,dc, \tag{2.26}$$

where

$$f_c(t) = \overline{c} \wedge c - \overline{c} \wedge (ct) \tag{2.27}$$

and

$$\gamma_f(c) := \frac{1}{c^3} f''\left(\frac{\overline{c}}{c}\right). \tag{2.28}$$

For $c \in [0,1]$, the term $\mathbb{I}_{f_c}(P,Q)$ in (2.26) is called the *c-primitive f-divergence* and can be written as

$$\begin{aligned}
\mathbb{I}_{f_c}(P,Q) &= \int \left\{ \overline{c} \wedge c - \overline{c} \wedge \left(c\frac{dP}{dQ} \right) \right\} dQ && (2.29)\\
&= \overline{c} \wedge c - \int \overline{c}dQ \wedge cdP && (2.30)\\
&= \overline{c} \wedge c - \frac{1}{2} + \frac{1}{2}\int |cdP - \overline{c}dQ| && (2.31)\\
&= \frac{1}{2} d_{\mathrm{TV}}(cP, \overline{c}Q) - \frac{1}{2}|1 - 2c|, && (2.32)
\end{aligned}$$

where the first equality (2.29) is due to the definition of f-divergence and (2.27), and the third equality (2.31) is due to the following observation:

$$\begin{aligned}
\int |p - q| &= \int_{q \geq p} q - p + \int_{q<p} p - q\\
&= \int_{q \geq p} q + \int_{q<p} p - \int p \wedge q\\
&= 1 - \int_{q<p} q + 1 - \int_{q \geq p} p - \int p \wedge q\\
&= 2 - 2 - \int p \wedge q.
\end{aligned}$$

Comparison between f-Divergences: Consider the problem of maximizing or minimizing an f-divergence between two probability measures subject to a finite number of constraints on other f-divergences. Given divergences $\mathbb{I}_f$ and $\mathbb{I}_{f_i}, i \in [m]$ and non-negative real numbers $\alpha_1, \ldots, \alpha_m$, let

$$\begin{aligned}
U(\alpha_1, \ldots, \alpha_m) &:= \sup_{P,Q} \{ \mathbb{I}_f(P,Q) : \mathbb{I}_{f_i}(P,Q) \leq \alpha_i, \forall i \in [m] \}, \text{ and}\\
L(\alpha_1, \ldots, \alpha_m) &:= \inf_{P,Q} \{ \mathbb{I}_f(P,Q) : \mathbb{I}_{f_i}(P,Q) \geq \alpha_i, \forall i \in [m] \},
\end{aligned}$$

where the probability measures on the right hand sides above range over all possible measurable spaces. These large infinite-dimensional optimization problems can all be

reduced to optimization problems over small finite dimensional spaces as shown in the following theorem 2.8.

Define

$$\begin{aligned} U_n(\alpha_1,\ldots,\alpha_m) &:= \sup_{P,Q\in\mathcal{P}([n])} \{\mathbb{I}_f(P,Q) : \mathbb{I}_{f_i}(P,Q) \le \alpha_i, \forall i \in [m]\}, \text{ and} \\ L_n(\alpha_1,\ldots,\alpha_m) &:= \inf_{P,Q\in\mathcal{P}([n])} \{\mathbb{I}_f(P,Q) : \mathbb{I}_{f_i}(P,Q) \ge \alpha_i, \forall i \in [m]\}, \end{aligned}$$

where $\mathcal{P}([n])$ denotes the space of all probability measures defined on the finite set $[n]$.

Theorem 2.8 (Guntuboyina et al. [2014]). *For every $\alpha_1,\ldots,\alpha_m \ge 0$, we have*

$$U(\alpha_1,\ldots,\alpha_m) = U_{m+2}(\alpha_1,\ldots,\alpha_m)$$

Further if $\alpha_1,\ldots,\alpha_m$ are all finite, then

$$L(\alpha_1,\ldots,\alpha_m) = L_{m+2}(\alpha_1,\ldots,\alpha_m).$$

Suppose that $\mathbb{I}_f$ is an arbitrary f-divergence and that all divergences $\mathbb{I}_{f_i}, i \in [m]$ are c-primitive f-divergences (2.32)*. Then*

$$L(\alpha_1,\ldots,\alpha_m) = L_{m+1}(\alpha_1,\ldots,\alpha_m).$$

Now we introduce a closely related concept - namely the *joint range*.

Definition 2.9 (Joint Range). *Consider two f-divergences $\mathbb{I}_f(P,Q)$ and $\mathbb{I}_g(P,Q)$. Their joint range is a subset of $\mathbb{R}^2$ defined by*

$$\begin{aligned} J &:= \{(\mathbb{I}_f(P,Q), \mathbb{I}_g(P,Q)) : P,Q \in \mathcal{P}(\mathcal{X}) \text{ where } \mathcal{X} \text{ is some measurable space}\}, \\ J_k &:= \{(\mathbb{I}_f(P,Q), \mathbb{I}_g(P,Q)) : P,Q \in \mathcal{P}([k])\}. \end{aligned}$$

The region J seems difficult to characterize since we need to consider P, Q over all measurable spaces; on the other hand, the region J_k for small k is easy to obtain. The following theorem relates these two regions (J and J_k).

Theorem 2.10 (Harremoes and Vajda [2011]). $J = conv(J_2)$.

By Theorem 2.10, the region J is no more than the convex hull of J_2. In certain cases, it is easy to obtain a parametric formula of J_2. In those cases, we can systematically prove several important inequalities between two f-divergences via their joint range. For example using the joint range between the total variation and Hellinger divergence, it can be shown that ([Tsybakov, 2009; Polyanskiy and Wu, 2016; accessed March 30, 2017]):

$$\frac{1}{2}\mathrm{He}^2(P,Q) \le d_{\mathrm{TV}}(P,Q) \le \mathrm{He}(P,Q)\sqrt{1-\frac{\mathrm{He}^2(P,Q)}{4}}. \tag{2.33}$$

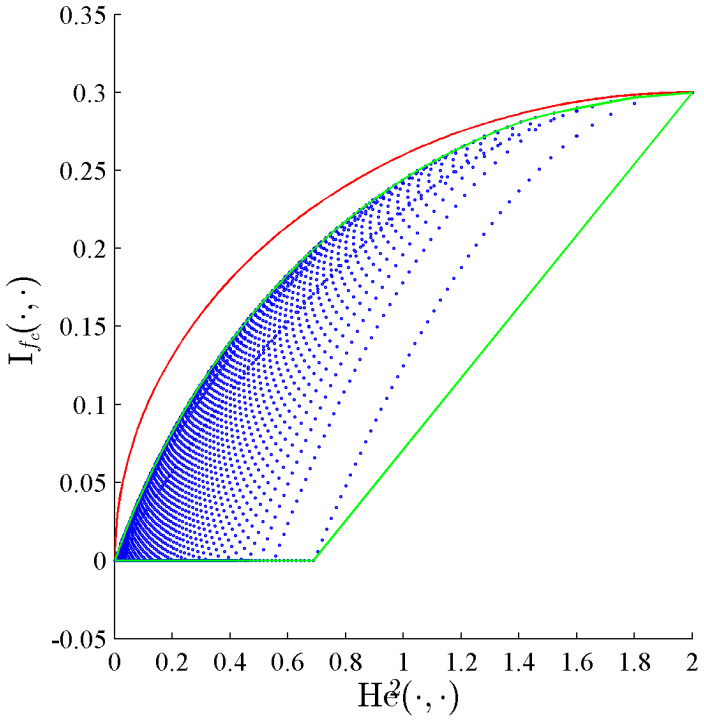


Figure 2.2: Joint range (J_2) of Hellinger distance and a c-primitive f-divergence (c=0.7) ($\cdots$), convex hull of J_2 (—), and a parametric curve $(c \wedge \bar{c}) \cdot \mathrm{He}(P,Q) \sqrt{1 - \frac{\mathrm{He}^2(P,Q)}{4}}$ (—).

We extend the above result to the c-primitive f-divergence as follows:

$$\mathbb{I}_{f_c}(P,Q) \;\leq\; (c \wedge \bar{c}) \cdot \mathrm{He}(P,Q) \sqrt{1 - \frac{\mathrm{He}^2(P,Q)}{4}}. \tag{2.34}$$

We use a mathematical software to plot (see Figure 2.2) the joint range between the c-primitive f-divergence and the Hellinger divergence which is given by the convex hull of

$$J_2 := \left\{ 2\left(1 - \sqrt{pq} - \sqrt{\bar{p}\bar{q}}\right), \frac{1}{2}\left(|cp - \bar{c}q| + |c\bar{p} - \bar{c}\bar{q}| - |2c-1|\right) : p,q \in [0,1] \right\}.$$

Then using this joint range, we verify that the bound given in (2.34) is indeed true. We also note that the bound in (2.34) is not tight but sufficient for our purposes (for analysing the hardness of the cost-sensitive classification problem in Chapter 3).

Sub-additive f-Divergences: Some f-divergences satisfy the sub-additivity property, which will be useful in analyzing the hardness of learning problems with repeated experiments (samples). The following lemma shows that both total variation and squared Hellinger divergences satisfy this property.

Lemma 2.11. *For all collections of distributions $P_i, Q_i \in \mathcal{P}(\mathcal{O}_i)$, $i \in [k]$*

$$d_{\mathrm{TV}}\left(\bigotimes_{i=1}^{k} P_i, \bigotimes_{i=1}^{k} Q_i\right) \leq \sum_{i=1}^{k} d_{\mathrm{TV}}(P_i, Q_i),$$

and

$$\mathrm{He}^2\left(\bigotimes_{i=1}^{k} P_i, \bigotimes_{i=1}^{k} Q_i\right) \leq \sum_{i=1}^{k} \mathrm{He}^2(P_i, Q_i).$$

Proof. Firstly $(P, Q) \mapsto d_{\mathrm{TV}}(P, Q)$ is a metric. Thus

$$\begin{aligned}
& d_{\mathrm{TV}}\left(\bigotimes_{i=1}^{k} P_i, \bigotimes_{i=1}^{k} Q_i\right) \\
= \; & d_{\mathrm{TV}}\left(P_1 \otimes \left(\bigotimes_{i=2}^{k} P_i\right), Q_1 \otimes \left(\bigotimes_{i=2}^{k} Q_i\right)\right) \\
\leq \; & d_{\mathrm{TV}}\left(P_1 \otimes \left(\bigotimes_{i=2}^{k} P_i\right), Q_1 \otimes \left(\bigotimes_{i=2}^{k} P_i\right)\right) + d_{\mathrm{TV}}\left(Q_1 \otimes \left(\bigotimes_{i=2}^{k} P_i\right), Q_1 \otimes \left(\bigotimes_{i=2}^{k} Q_i\right)\right) \\
= \; & d_{\mathrm{TV}}(P_1, Q_1) + d_{\mathrm{TV}}\left(\bigotimes_{i=2}^{k} P_i, \bigotimes_{i=2}^{k} Q_i\right),
\end{aligned}$$

where the second line follows by definition, the third follows from the triangle inequality and the forth is easily verified from the definition of $d_{\mathrm{TV}}(\cdot, \cdot)$. To complete the proof proceed inductively.

Let μ be a product measure on $\mathcal{O}_1 \times \mathcal{O}_2$, written as $\mu = \mu_1 \otimes \mu_2$, where $\mu_i := \mu \circ \pi_i$ denotes the image measure of the projection $\pi_i : \mathbb{R}^2 \ni (x_1, x_2) \mapsto \pi_i(x_1, x_2) = x_i$ w.r.t. μ. Also let $P = P_1 \otimes P_2$, and $Q = Q_1 \otimes Q_2$. Define $p := \frac{dP}{d\mu}$, $q := \frac{dQ}{d\mu}$, $p_1 := \frac{dP_1}{d\mu_1}$, $p_2 := \frac{dP_2}{d\mu_2}$, $q_1 := \frac{dQ_1}{d\mu_1}$, and $q_2 := \frac{dQ_2}{d\mu_2}$. Then, by Tonelli's theorem,

$$\begin{aligned}
1 - \frac{1}{2}\mathrm{He}^2(P, Q) &= \int \sqrt{pq}\, d\mu \\
&= \int \sqrt{p_1 q_1}\, d\mu_1 \cdot \int \sqrt{p_2 q_2}\, d\mu_2 \\
&= \left(1 - \frac{1}{2}\mathrm{He}^2(P_1, Q_1)\right) \cdot \left(1 - \frac{1}{2}\mathrm{He}^2(P_2, Q_2)\right).
\end{aligned}$$

Thus we have

$$\begin{aligned}
\mathrm{He}^2(P, Q) &= 2 - 2\left(1 - \frac{1}{2}\mathrm{He}^2(P_1, Q_1)\right) \cdot \left(1 - \frac{1}{2}\mathrm{He}^2(P_2, Q_2)\right) \\
&= \mathrm{He}^2(P_1, Q_1) + \mathrm{He}^2(P_2, Q_2) - \frac{1}{2}\mathrm{He}^2(P_1, Q_1)\,\mathrm{He}^2(P_2, Q_2) \\
&\leq \mathrm{He}^2(P_1, Q_1) + \mathrm{He}^2(P_2, Q_2).
\end{aligned}$$

To complete the proof proceed the above process iteratively. □

Chapter 3

Asymmetric Learning Problems

The central problem of this chapter is the cost-sensitive binary classification problem, where different costs are associated with different types of mistakes. Several important machine learning applications such as medical decision making, targeted marketing, and intrusion detection can be formalized as cost-sensitive classification setup (Abe et al. [2004]).

The chapter proceeds as follows. In section 3.1 we show that the abstract language of transitions introduced in chapter 2, is general enough to capture many of the existing practical problems in statistics and machine learning including the cost-sensitive classification problem. Then in section 3.2 we study the hardness of the cost-sensitive classification problem by extending the standard minimax lower bound of balanced binary classification problem (due to Massart and Nédélec [2006]) to cost-sensitive classification problem.

In section 3.3 we study the hardness of the constrained learning problem (specifically constrained cost-sensitive classification), which naturally leads us to a detailed investigation of strong data processing inequalities. After reviewing the known results in strong data processing inequalities, we make some novel progress in the direction of strong data processing inequalities for binary symmetric channels. We also extend the well-known contraction coefficient theorem (Cohen et al. [1993]) for total variational divergence to c-primitive f-divergences.

Finally in section 3.4 we study the local privacy requirement as a form of constraint on learning problem. We review the decision theoretic reduction of the local privacy requirement, and based on that we propose a prioritized (cost-sensitive) privacy definition.

3.1 Preliminaries and Background

General Learning Task: Consider the *General Learning Task* represented by the following transition diagram:

$$\Theta \overset{\varepsilon_n}{\rightsquigarrow} \mathcal{O}^n \overset{A}{\rightsquigarrow} \mathcal{A} \tag{3.1}$$

where Θ, $\mathcal{O}$, and $\mathcal{A}$ are *parameter*, *observation*, and *action* spaces respectively. The transitions ε_n and A denote *repeated experiment of* $\varepsilon : \Theta \rightsquigarrow \mathcal{O}$ and *algorithm* respectively. Note that the repeated experiment ε_n induces the class of probability measures given by $\mathcal{P}_{\varepsilon_n}(\mathcal{O}^n) := \{\varepsilon_n(\theta) := \varepsilon(\theta)^n : \theta \in \Theta\}$ (see Section 2.2.3). We recall the following objects introduced in Chapter 2 :

$$\begin{aligned}
\text{Loss} \quad & \ell : \Theta \times \mathcal{A} \to \mathbb{R} \\
\text{Regret} \quad & \Delta\ell(\theta, a) := \ell(\theta, a) - \inf_{a' \in \mathcal{A}} \ell(\theta, a') \\
\text{Full Risk} \quad & R_\ell(\varepsilon_n, \theta, A) := \mathop{\mathbb{E}}_{\mathsf{O}_1^n \sim \varepsilon_n(\theta)} \left[\mathop{\mathbb{E}}_{a \sim A(\mathsf{O}_1^n)} [\ell(\theta, a)] \right] \\
& R_{\Delta\ell}(\varepsilon_n, \theta, A) := \mathop{\mathbb{E}}_{\mathsf{O}_1^n \sim \varepsilon_n(\theta)} \left[\mathop{\mathbb{E}}_{a \sim A(\mathsf{O}_1^n)} [\Delta\ell(\theta, a)] \right] \\
\text{Full Minimax Risk} \quad & \underline{R}^\star_\ell(\varepsilon_n) := \inf_A \sup_{\theta \in \Theta} R_\ell(\varepsilon_n, \theta, A) \\
& \underline{R}^\star_{\Delta\ell}(\varepsilon_n) := \inf_A \sup_{\theta \in \Theta} R_{\Delta\ell}(\varepsilon_n, \theta, A).
\end{aligned}$$

One needs to carefully distinguish between the risk (and related notions) in terms of loss and regret based on the subscript (see Remark 2.3). The general learning task is compactly denoted by the tuple (ℓ, ε_n).

In order to demonstrate the generality of the language of transitions, below we discuss some specific instantiations (supervised learning, multi-class probability estimation, binary classification, and parameter estimation) of this general learning task.

Supervised Learning Problem: Let $\mathcal{X} \times \mathcal{Y}$ be a measurable space, and let D be an unknown joint probability measure on $\mathcal{X} \times \mathcal{Y}$. The set $\mathcal{X}$ is called the *instance space*, the set $\mathcal{Y}$ the *outcome space*. Let $S = \{(\mathsf{X}_i, \mathsf{Y}_i)\}_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$ be a finite training sample, where each pair $(\mathsf{X}_i, \mathsf{Y}_i)$ is generated independently according to the unknown probability measure D. Then the goal of a learning algorithm is to find a function $f : \mathcal{X} \to \mathcal{Y}$ which given a new instance $x \in \mathcal{X}$, predicts its label to be $\widehat{y} = f(x)$.

Here we rely on the fundamental assumption that both training and future (test) data are generated by the same fixed underlying probability measure D, which, although unknown, allows us to infer from training data to future data and therefore to generalize.

In order to measure the performance of a learning algorithm, we define an *error function* $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, where $d(y, \widehat{y})$ quantifies the discrepancy between the predicted value $\widehat{y}$ and the actual value y. The performance of any function $f : \mathcal{X} \to \mathcal{Y}$ is then measured in terms of its *generalization error*, which is defined as the expected error:

$$\mathrm{er}_d(f, D) := \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [d(\mathsf{Y}, f(\mathsf{X}))], \tag{3.2}$$

where the expectation is taken with respect to the probability measure D on the data (X, Y). The best estimate $f_D^\star \in \mathcal{Y}^{\mathcal{X}}$ is therefore the one for which the generalization

error is as small as possible, that is,

$$f_D^\star \;:=\; \operatorname*{arg\,min}_{f\in\mathcal{Y}^{\mathcal{X}}} \operatorname{er}_d\left(f, D\right). \tag{3.3}$$

The function $f_D^\star$ is called the target hypothesis.

In order to avoid functions which over-fit the training sample and do not generalize well on the test data, one usually imposes constraints on the function f. One way to impose constraints is by restricting the possible choices of functions to a fixed class of functions from which the learning algorithm chooses its hypothesis. This function class is called the *hypothesis class*. Given a fixed hypothesis class $\mathcal{F}\subseteq\mathcal{Y}^{\mathcal{X}}$, the goal of a learning algorithm is thus to choose the hypothesis function $f^\star$ in $\mathcal{F}$ which has the smallest generalization error on data drawn according to the underlying probability measure D,

$$f_{D,\mathcal{F}}^\star \;:=\; \operatorname*{arg\,min}_{f\in\mathcal{F}} \operatorname{er}_d\left(f, D\right). \tag{3.4}$$

We will assume in the following that such an $f_{D,\mathcal{F}}^\star$ exists.

The supervised learning problem can be derived from the general learning task (3.1) with the following instantiation:

- the observation space is $\mathcal{O}=\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}\subseteq\mathbb{R}^d$,
- the action space is $\mathcal{A}=\mathcal{F}\subseteq\mathcal{Y}^{\mathcal{X}}$,
- the learning algorithm is $A=\widehat{f}$, and
- the loss function is

$$\ell_d : \Theta\times\mathcal{F}\ni(\theta,f)\mapsto\ell_d\left(\theta,f\right):=\operatorname{er}_d\left(f,\varepsilon\left(\theta\right)\right)\in\mathbb{R}_+,$$

  where $\varepsilon\left(\theta\right)$ is the probability measure associated with the parameter $\theta\in\Theta$. One needs to carefully distinguish between the error function $d:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ which acts on the observation space, and the loss function $\ell_d:\Theta\times\mathcal{A}\to\mathbb{R}$ which acts on the parameter and decision spaces.

Then the transition diagram for this supervised learning problem (ℓ_d,ε_n) is

$$\Theta \overset{\varepsilon_n}{\rightsquigarrow} (\mathcal{X}\times\mathcal{Y})^n \overset{\widehat{f}}{\rightsquigarrow} \mathcal{F} \quad . \tag{3.5}$$

Binary Classification: When $\mathcal{Y}=\{-1,1\}$, the supervised learning task (3.5) is called binary classification, which is a central problem in machine learning (Devroye et al. [2013]). A common error function for binary classification is simply the zero-one error defined by $d_{0-1}\left(y,\widehat{y}\right)=[\![\widehat{y}\neq y]\!]$. In this case the generalization error of a classifier $f:\mathcal{X}\to\{-1,1\}$ w.r.t. a probability measure D is simply the probability that it predicts the wrong label on a randomly drawn example:

$$\operatorname{er}_{d_{0-1}}\left(f,D\right) \;:=\; \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D}\left[d_{0-1}\left(\mathsf{Y},f\left(\mathsf{X}\right)\right)\right] \;=\; \mathop{\mathbb{P}}_{(\mathsf{X},\mathsf{Y})\sim D}\left[f\left(\mathsf{X}\right)\neq\mathsf{Y}\right].$$

The optimal error over all possible classifiers $f : \mathcal{X} \to \{-1, 1\}$ for a given probability measure D is called the *Bayes error* (minimum generalization error) associated with D:

$$\underline{\mathrm{er}}_{d_{0-1}}(D) := \inf_{f \in \{-1,1\}^{\mathcal{X}}} \mathrm{er}_{d_{0-1}}(f, D). \tag{3.6}$$

It is easily verified that, if $\eta_D(x)$ is defined as the conditional probability (under D) of a positive label given x, $\eta_D(x) = \mathbb{P}_D[\mathsf{Y} = 1 \mid \mathsf{X} = x]$, then the classifier $f_D^\star : \mathcal{X} \to \{-1, 1\}$ given by

$$f_D^\star(x) = \begin{cases} 1 & \text{if } \eta_D(x) \geq 1/2 \\ -1 & \text{otherwise} \end{cases}$$

achieves the Bayes error. Such a classifier is termed a *Bayes classifier*. In general, η_D is unknown so the above classifier cannot be constructed directly.

By defining $\ell_{d_{0-1}} : \Theta \times \mathcal{F} \ni (\theta, f) \mapsto \ell_{d_{0-1}}(\theta, f) := \mathrm{er}_{d_{0-1}}(f, \varepsilon(\theta)) \in \mathbb{R}_+$, the binary classification problem $(\ell_{d_{0-1}}, \varepsilon_n)$ can be represented by the following transition diagram:

$$\Theta \overset{\varepsilon_n}{\rightsquigarrow} (\mathcal{X} \times \{-1, 1\})^n \overset{\widehat{f}}{\rightsquigarrow} \mathcal{F} \quad . \tag{3.7}$$

Note that the repeated experiment ε_n above induces the class of probability measures given by $\mathcal{P}_{\varepsilon_n}((\mathcal{X} \times \{-1, 1\})^n) := \{\varepsilon_n(\theta) := \varepsilon(\theta)^n \in \mathcal{P}(\mathcal{X} \times \{-1, 1\})^n : \theta \in \Theta\}$. Using the Bayes rule, the distribution $\mathbb{P}_{\varepsilon(\theta)}$ can be decomposed as follows:

$$\mathbb{P}_{\varepsilon(\theta)}[\mathsf{X} = x, \mathsf{Y} = 1] = \mathbb{P}_{\varepsilon(\theta)}[\mathsf{X} = x] \cdot \mathbb{P}_{\varepsilon(\theta)}[\mathsf{Y} = 1 \mid \mathsf{X} = x] = M_{\varepsilon(\theta)}(x) \cdot \eta_{\varepsilon(\theta)}(x),$$

where $M_{\varepsilon(\theta)}(x) := \mathbb{P}_{\varepsilon(\theta)}[\mathsf{X} = x]$ and $\eta_{\varepsilon(\theta)}(x) := \mathbb{P}_{\varepsilon(\theta)}[\mathsf{Y} = 1 \mid \mathsf{X} = x]$. For simplicity we will write $\mathbb{P}_{\varepsilon(\theta)}$, $M_{\varepsilon(\theta)}$, $\eta_{\varepsilon(\theta)}$, and $f^\star_{\varepsilon(\theta)}$ as $\mathbb{P}_\theta$, M_θ, η_θ, and $f^\star_\theta$ respectively.

Cost-sensitive Binary Classification: Suppose we are given gene expression profiles for some number of patients, together with labels for these patients indicating whether or not they had a certain form of a disease. We want to design a learning algorithm which automatically recognizes the diseased patient based on the gene expression profile of a patient. In this case, there are different costs associated with different types of mistakes (the health risk for a false label "no" is much higher than for a false "yes"), and the *cost-sensitive error function* (for $c \in (0, 1)$) can be used to capture this:

$$d_c : \mathcal{Y} \times \mathcal{Y} \ni (y, \widehat{y}) \mapsto d_c(y, \widehat{y}) := [\![\widehat{y} \neq y]\!] \cdot \{\overline{c} \cdot [\![y = 1]\!] + c \cdot [\![y = -1]\!]\},$$

where $\overline{c} := 1 - c$. Then the performance measure (loss function) associated with the above cost-sensitive error function is given by

$$\ell_{d_c} : \Theta \times \mathcal{F} \ni (\theta, f) \mapsto \ell_{d_c}(\theta, f) := \mathrm{er}_{d_c}(f, \varepsilon(\theta)) \in \mathbb{R}_+,$$

where

$$\begin{aligned}\mathrm{er}_{d_c}\left(f, \varepsilon\left(\theta\right)\right) &:= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim\varepsilon(\theta)}\left[d_c\left(\mathsf{Y}, f\left(\mathsf{X}\right)\right)\right] \\ &= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim\varepsilon(\theta)}\left[\left[\!\left[f\left(\mathsf{X}\right) \neq \mathsf{Y}\right]\!\right] \cdot \left\{\overline{c} \cdot \left[\!\left[\mathsf{Y} = 1\right]\!\right] + c \cdot \left[\!\left[\mathsf{Y} = -1\right]\!\right]\right\}\right].\end{aligned}$$

For any $\eta : \mathcal{X} \to [0,1]$, and $f : \mathcal{X} \to \{-1,1\}$, define the conditional generalization error (given $x \in \mathcal{X}$) as

$$\begin{aligned}\mathrm{er}_{d_c}\left(f, \eta; x\right) &:= \mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta(x)}\left[d_c\left(\mathsf{Y}, f\left(x\right)\right)\right] \\ &= \overline{c} \cdot \eta\left(x\right) \cdot \left[\!\left[f\left(x\right) \neq 1\right]\!\right] + c \cdot \overline{\eta\left(x\right)} \cdot \left[\!\left[f\left(x\right) \neq -1\right]\!\right],\end{aligned}$$

where $\overline{\eta\left(x\right)} := 1 - \eta\left(x\right)$. Then $\mathrm{er}_{d_c}\left(f, \eta; x\right)$ is minimized by

$$\begin{aligned}f^{\star}\left(x\right) &:= \mathop{\arg\min}_{f\in\{-1,1\}^{\mathcal{X}}} \mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta(x)}\left[d_c\left(\mathsf{Y}, f\left(x\right)\right)\right] \\ &= \mathrm{sign}\left(\overline{c} \cdot \eta\left(x\right) - c \cdot \overline{\eta\left(x\right)}\right) \\ &= \mathrm{sign}\left(\eta\left(x\right) - c\right),\end{aligned}$$

since $\mathrm{er}_{d_c}\left(f^{\star}, \eta; x\right) = \overline{c} \cdot \eta\left(x\right) \wedge c \cdot \overline{\eta\left(x\right)}$. In order to find the optimal classifier for each $\theta \in \Theta$ (associated joint probability measure $\varepsilon\left(\theta\right)$ on $\mathcal{X} \times \{-1,1\}$) w.r.t. the cost-sensitive loss function, we note that

$$\begin{aligned}\inf_{f\in\{-1,1\}^{\mathcal{X}}} \ell_{d_c}\left(\theta, f\right) &= \inf_{f\in\{-1,1\}^{\mathcal{X}}} \mathrm{er}_{d_c}\left(f, \varepsilon\left(\theta\right)\right) \\ &= \inf_{f\in\{-1,1\}^{\mathcal{X}}} \mathop{\mathbb{E}}_{\mathsf{X}\sim M_\theta}\left[\mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(\mathsf{X})}\left[d_c\left(\mathsf{Y}, f\left(\mathsf{X}\right)\right)\right]\right] \\ &= \mathop{\mathbb{E}}_{\mathsf{X}\sim M_\theta}\left[\inf_{f\in\{-1,1\}^{\mathcal{X}}} \mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(\mathsf{X})}\left[d_c\left(\mathsf{Y}, f\left(\mathsf{X}\right)\right)\right]\right] \\ &= \ell_{d_c}\left(\theta, f_\theta^{\star}\right),\end{aligned}$$

where $M_\theta\left(x\right) := \mathbb{P}_{\varepsilon(\theta)}\left[\mathsf{X} = x\right]$, $\eta_\theta\left(x\right) := \mathbb{P}_{\varepsilon(\theta)}\left[\mathsf{Y} = 1 \mid \mathsf{X} = x\right]$, and $f_\theta^{\star}$ is given by

$$f_\theta^{\star}\left(x\right) := \begin{cases} 1, & \text{if } \eta_\theta(x) \geq c \\ -1, & \text{otherwise} \end{cases}. \tag{3.8}$$

We instantiate the following objects related to the cost-sensitive classification problem

$$\begin{aligned}\text{Regret} \quad & \Delta\ell_{d_c}\left(\theta, f\right) := \ell_{d_c}\left(\theta, f\right) - \ell_{d_c}\left(\theta, f_\theta^{\star}\right) \\ \text{Full Risk} \quad & \mathcal{R}_{\Delta\ell_{d_c}}\left(\varepsilon_n, \theta, \widehat{f}\right) := \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\sim\varepsilon_n(\theta)}\left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)}\left[\Delta\ell_{d_c}\left(\theta, f\right)\right]\right]\end{aligned}$$

Full Minimax Risk $\underline{\mathcal{R}}^{\star}_{\Delta\ell_{d_c}}(\varepsilon_n) := \inf_{\widehat{f}} \sup_{\theta\in\Theta} \mathcal{R}_{\Delta\ell_{d_c}}\left(\varepsilon_n, \theta, \widehat{f}\right)$.

The following lemma from Scott et al. [2012] will be used later.

Lemma 3.1 (Scott et al. [2012])**.** *Consider the binary classification problem* (3.7)*. For any $f \in \mathcal{F}$ and $c \in (0,1)$,*

$$\Delta\ell_{d_c}(\theta, f) \;=\; \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X}\sim M_\theta} \left[|\eta_\theta(\mathsf{X}) - c| \cdot |f(\mathsf{X}) - f^{\star}_{\theta}(\mathsf{X})| \right],$$

where $f^{\star}_{\theta}$ is given by (3.8)*.*

Proof. Consider a fixed $x \in \mathcal{X}$. Recall that

$$f^{\star}_{\theta}(x) = \mathop{\arg\min}_{f\in\{-1,1\}^{\mathcal{X}}} \mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(x)} \left[d_c(\mathsf{Y}, f(x)) \right] = \operatorname{sign}\left(\eta_\theta(x) - c\right).$$

Therefore $\inf_{f\in\{-1,1\}^{\mathcal{X}}} \mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(x)} \left[d_c(\mathsf{Y}, f(x)) \right] = \mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(x)} \left[d_c(\mathsf{Y}, f^{\star}_{\theta}(x)) \right]$. This implies

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(x)} \left[d_c(\mathsf{Y}, f(x)) \right] - \mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(x)} \left[d_c(\mathsf{Y}, f^{\star}_{\theta}(x)) \right] \\
&= \bar{c}\,\eta_\theta(x)\,[\![f(x) \neq 1]\!] + c\,\overline{\eta_\theta(x)}\,[\![f(x) \neq -1]\!] \\
&\quad - \left\{ \bar{c}\,\eta_\theta(x)\,[\![f^{\star}_{\theta}(x) \neq 1]\!] + c\,\overline{\eta_\theta(x)}\,[\![f^{\star}_{\theta}(x) \neq -1]\!] \right\} \\
&= [\![f(x) \neq f^{\star}_{\theta}(x)]\!] \, |\eta_\theta(x) - c| \\
&= \frac{1}{2} \cdot |f(x) - f^{\star}_{\theta}(x)| \cdot |\eta_\theta(x) - c|.
\end{aligned}$$

Then the proof is completed by noting that

$$\begin{aligned}
\ell_{d_c}(\theta, f) - \ell_{d_c}(\theta, f^{\star}_{\theta}) \;&=\; \mathop{\mathbb{E}}_{\mathsf{X}\sim M_\theta} \left[\mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(\mathsf{X})} \left[d_c(\mathsf{Y}, f(\mathsf{X})) \right] - \mathop{\mathbb{E}}_{\mathsf{Y}\sim\eta_\theta(\mathsf{X})} \left[d_c(\mathsf{Y}, f^{\star}_{\theta}(\mathsf{X})) \right] \right] \\
&=\; \frac{1}{2} \mathop{\mathbb{E}}_{\mathsf{X}\sim M_\theta} \left[|f(\mathsf{X}) - f^{\star}_{\theta}(\mathsf{X})| \cdot |\eta_\theta(\mathsf{X}) - c| \right].
\end{aligned}$$

□

Parameter Estimation Problem: The main goal of a parameter problem is to accurately *reconstruct the parameters* of the original distribution from which the data is generated, using the loss function of the type $\rho : \Theta \times \Theta \to \mathbb{R}$. This problem is represented by the following transition diagram (with $\mathcal{A} = \Theta$, and $A = \widehat{\theta}$):

$$\Theta \overset{\varepsilon_n}{\rightsquigarrow} \mathcal{O}^n \overset{\widehat{\theta}}{\rightsquigarrow} \Theta \quad . \tag{3.9}$$

Let $\theta : \mathcal{P}(\mathcal{O}) \to \Theta$ denote a function defined on $\mathcal{P}(\mathcal{O})$, that is, a mapping $P \mapsto \theta(P)$. The goal of the algorithm $\widehat{\theta}$ is to estimate the parameter $\theta(P)$ based

on observations O_1^n drawn from the (unknown) distribution P. In certain cases, the parameter $\theta(P)$ uniquely determines the underlying distribution; for example, in the case of mean (θ) estimation problem from the normal distribution family $\mathcal{P}(\mathcal{O}) = \left\{\mathcal{N}(\theta, \Sigma) : \theta \in \mathbb{R}^d\right\}$ with known covariance matrix Σ, the parameter mapping $\theta(P) = \mathop{\mathbb{E}}_{\mathsf{O}\sim P}[\mathsf{O}]$ uniquely determines distributions in $\mathcal{P}(\mathcal{O})$. In other scenarios, however, θ does not uniquely determine the distribution (confer Duchi [2016; accessed March 30, 2017] for general treatment with this broader viewpoint of estimating functions of distributions). In this chapter we consider the one-to-one function $P \mapsto \theta(P)$.

Observe that the class of probability measures induced by the repeated experiment ε_n is written as $\mathcal{P}_{\varepsilon_n}(\mathcal{O}^n) := \{\varepsilon_n(\theta) := \varepsilon(\theta)^n \in \mathcal{P}(\mathcal{O})^n : \theta \in \Theta\}$. Let $\rho : \Theta \times \Theta \to \mathbb{R}$ be a pseudo metric (that is, it satisfies symmetry and the triangle inequality) on Θ. Then the minimax risk of this problem is defined as

$$\underline{\mathcal{R}}_\rho^\star(\varepsilon_n) \;:=\; \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathop{\mathbb{E}}_{\mathsf{O}_1^n\sim\varepsilon_n(\theta)} \left[\mathop{\mathbb{E}}_{\tilde{\theta}\sim\widehat{\theta}(\mathsf{O}_1^n)} \left[\rho\left(\theta, \tilde{\theta}\right) \right] \right]. \tag{3.10}$$

Hardness of a Problem via minimax lower bounds: Understanding the hardness or fundamental limits of a learning problem is important for practice for the following reasons:

- They give an estimate on the number of samples required for a good performance of a learning algorithm.
- They give an intuition about the quantities and structural properties which are essential for a learning process and therefore about which problems are inherently easier than others.
- They quantify the influence of parameters and indicate what prior knowledge is relevant in a learning setting and therefore they guide the analysis, design, and improvement of learning algorithms.

Note that the "hardness" here corresponds to lower bounds on sample complexity (and not computational complexity). We demonstrate the hardness of a learning problem (3.1) (and the instantiations of it) by obtaining lower bounds for the minimax risk $\underline{R}_\ell^\star(\varepsilon_n)$ of it.

In section 3.2 we review and extend techniques due to Le Cam [2012] and Assouad [1983] for obtaining minimax lower bounds for learning problems. Both techniques proceed by reducing the learning problem to an easier hypothesis testing problem [Tsybakov, 2009; Yang and Barron, 1999; Yu, 1997], then proving a lower bound on the probability of error in testing problems.

Le Cam's method, in its simplest form, provides lower bounds on the error in simple binary hypothesis testing problems, by using the connection between hypothesis testing and total variation distance.

Consider the parameter estimation problem (3.9) with $\Theta = \{-1, 1\}^m$ for some m, where the objective is to determine every bit of the underlying unknown parameter

$\theta \in \{-1,1\}^m$. In that setting, a key result known as *Assouad's lemma* says that the difficulty of estimating the entire bit string θ is related to the difficulty of estimating each bit of θ separately, assuming all other bits are already known.

3.2 Hardness of the Cost-sensitive Classification Problem

In this section we follow the presentation of Raginsky [2015; accessed March 30, 2017]. Before studying the hardness of the cost-sensitive classification, we study the hardness of the auxiliary problem of parameter estimation (3.9).

3.2.1 Minimax Lower Bounds for Parameter Estimation Problem

We derive the cost-dependent lower bound for $\underline{\mathcal{R}}^{\star}_{\rho}(\varepsilon_n)$ (defined in (3.10)) by extending the standard Le Cam and Assouad's techniques. We start with the two point method introduced by Lucien Le Cam for obtaining minimax lower bounds.

Proposition 3.2. *For any $c \in (0,1)$, the minimax risk $\underline{\mathcal{R}}^{\star}_{\rho}(\varepsilon_n)$ (given by (3.10)) of the parameter estimation problem (3.9) with (pseudo metric) loss function $\rho : \Theta \times \Theta \to \mathbb{R}$ is bounded from below as follows:*

$$\underline{\mathcal{R}}^{\star}_{\rho}(\varepsilon_n) \;\geq\; \sup_{\theta \neq \theta'} \left\{ \rho(\theta,\theta') \cdot \left(c \wedge \overline{c} - \mathbb{I}_{f_c}\left(\varepsilon_n(\theta), \varepsilon_n(\theta')\right)\right)\right\},$$

where f_c is given by (2.27).

Proof. Let $c \in (0,1)$ be arbitrary but fixed. Consider any two fixed parameters $\theta, \theta' \in \Theta$ s.t. $\theta \neq \theta'$ and an arbitrary estimator $\widehat{\theta} : \mathcal{O}^n \rightsquigarrow \Theta$. Let $P^n_{\theta} := \varepsilon_n(\theta)$, and $P^n_{\theta'} := \varepsilon_n(\theta')$ (associated probability densities can be written as dP^n_{θ} and $dP^n_{\theta'}$). For an arbitrary (but fixed) set of observations $o^n_1 \in \mathcal{O}^n$, when $\overline{c} \cdot dP^n_{\theta'}(o^n_1) \geq c \cdot dP^n_{\theta}(o^n_1)$, we have

$$\begin{aligned}
& c \cdot dP^n_{\theta}(o^n_1) \mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(o^n_1)} \left[\rho(\theta,\tilde{\theta})\right] + \overline{c} \cdot dP^n_{\theta'}(o^n_1) \mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(o^n_1)} \left[\rho(\theta',\tilde{\theta})\right] \\
= \;& c \cdot dP^n_{\theta}(o^n_1) \mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(o^n_1)} \left[\rho(\theta,\tilde{\theta}) + \rho(\theta',\tilde{\theta})\right] + \left(\overline{c} \cdot dP^n_{\theta'}(o^n_1) - c \cdot dP^n_{\theta}(o^n_1)\right) \mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(o^n_1)} \left[\rho(\theta',\tilde{\theta})\right] \\
\overset{(i)}{\geq} \;& c \cdot dP^n_{\theta}(o^n_1) \mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(o^n_1)} \left[\rho(\theta,\tilde{\theta}) + \rho(\theta',\tilde{\theta})\right] \\
\overset{(ii)}{\geq} \;& c \cdot dP^n_{\theta}(o^n_1)\, \rho(\theta,\theta'),
\end{aligned} \tag{3.11}$$

where (i) is due to $\overline{c} \cdot dP^n_{\theta'}(o^n_1) \geq c \cdot dP^n_{\theta}(o^n_1)$, and (ii) is due to the triangle inequality. Similarly, for the case where $\overline{c} \cdot dP^n_{\theta'}(o^n_1) \leq c \cdot dP^n_{\theta}(o^n_1)$, we get

$$c \cdot dP^n_{\theta}(o^n_1) \mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(o^n_1)} \left[\rho(\theta,\tilde{\theta})\right] + \overline{c} \cdot dP^n_{\theta'}(o^n_1) \mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(o^n_1)} \left[\rho(\theta',\tilde{\theta})\right] \;\geq\; \overline{c} \cdot dP^n_{\theta'}(o^n_1)\, \rho(\theta,\theta'). \tag{3.12}$$

By combining (3.11) and (3.12), and summing over all $o_1^n \in \mathcal{O}^n$, we get, for any two $\theta, \theta' \in \Theta$ and any estimator $\widehat{\theta}$,

$$\begin{aligned} & c \cdot \mathop{\mathbb{E}}_{\mathsf{O}_1^n \sim P_\theta^n} \left[\mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(\mathsf{O}_1^n)} \left[\rho(\theta, \tilde{\theta}) \right] \right] + \overline{c} \cdot \mathop{\mathbb{E}}_{\mathsf{O}_1^n \sim P_{\theta'}^n} \left[\mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(\mathsf{O}_1^n)} \left[\rho(\theta', \tilde{\theta}) \right] \right] && (3.13) \\ \geq \; & \rho(\theta, \theta') \cdot \int c dP_\theta^n \wedge \overline{c} dP_{\theta'}^n \\ = \; & \rho(\theta, \theta') \cdot \left(c \wedge \overline{c} - \mathbb{I}_{f_c}(P_\theta^n, P_{\theta'}^n) \right), && (3.14) \end{aligned}$$

where the last equality follows from the definition of c-primitive f-divergences (2.30). By taking the supremum of both sides over the choices of θ, θ' (since then the two terms in (3.13) collapse to one), we have

$$\sup_{\theta \in \Theta} \mathop{\mathbb{E}}_{\mathsf{O}_1^n \sim P_\theta^n} \left[\mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(\mathsf{O}_1^n)} \left[\rho(\theta, \tilde{\theta}) \right] \right] \; \geq \; \sup_{\theta \neq \theta'} \left\{ \rho(\theta, \theta') \cdot \left(c \wedge \overline{c} - \mathbb{I}_{f_c}(\varepsilon_n(\theta), \varepsilon_n(\theta')) \right) \right\}.$$

The proof is completed by taking the infimum of both sides over $\widehat{\theta}$. □

By setting $c = \frac{1}{2}$ in Proposition 3.2, we recover Le Cam's (Le Cam [2012]) minimax lower bound for parameter estimation problem (3.9):

$$\underline{\mathcal{R}}_\rho^\star(\varepsilon_n) \; \geq \; \frac{1}{2} \sup_{\theta \neq \theta'} \left\{ \rho(\theta, \theta') \cdot \left(1 - \frac{1}{2} d_{\mathrm{TV}}(\varepsilon_n(\theta), \varepsilon_n(\theta')) \right) \right\},$$

since $\mathbb{I}_{f_{\frac{1}{2}}}(P, Q) = \frac{1}{4} d_{\mathrm{TV}}(P, Q)$ (from (2.31) with $c = \frac{1}{2}$). Now we provide an auxiliary result which will be useful in deriving the cost-dependent minimax lower bounds via Assouad's lemma (Assouad [1983]).

Corollary 3.3. *Let π be any prior distribution on Θ, and let μ be any joint probability distribution of a random pair $(\theta, \theta') \in \Theta \times \Theta$, such that the marginal distributions of both θ and θ' are equal to π. Then for any $c \in (0, 1)$, the minimax risk $\underline{\mathcal{R}}_\rho^\star(\varepsilon_n)$ (given by (3.10)) of the parameter estimation problem (3.9) is bounded from below as follows:*

$$\underline{\mathcal{R}}_\rho^\star(\varepsilon_n) \; \geq \; \mathop{\mathbb{E}}_{(\theta, \theta') \sim \mu} \left[\rho(\theta, \theta') \cdot \left(c \wedge \overline{c} - \mathbb{I}_{f_c}(\varepsilon_n(\theta), \varepsilon_n(\theta')) \right) \right]$$

Proof. First observe that for any prior π

$$\underline{\mathcal{R}}_\rho^\star(\varepsilon_n) \; \geq \; \inf_{\widehat{\theta}} \mathop{\mathbb{E}}_{\theta \sim \pi} \left[\mathop{\mathbb{E}}_{\mathsf{O}_1^n \sim \varepsilon_n(\theta)} \left[\mathop{\mathbb{E}}_{\tilde{\theta} \sim \widehat{\theta}(\mathsf{O}_1^n)} \left[\rho(\theta, \tilde{\theta}) \right] \right] \right],$$

since the minimax risk can be lower bounded by the Bayesian risk (see Theorem 2.4). Then by taking expectation of both sides of (3.14) w.r.t μ and using the fact that, under μ, both θ and θ' have the same distribution π, the proof is completed. □

Using the above corollary and extending the standard Assouad's lemma, we derive the cost-dependent minimax lower bound for the parameter estimation problem (3.9).

Theorem 3.4. *Let $d \in \mathbb{N}$, $\Theta = \{-1,1\}^d$ and $\rho = \rho_{\mathrm{Ha}}$, where the Hamming distance ρ_{Ha} is given by (2.1). Then for any $c \in (0,1)$, the minimax risk of the parameter estimation problem (3.9) satisfies*

$$\underline{\mathcal{R}}^{\star}_{\rho_{\mathrm{Ha}}}(\varepsilon_n) \;\geq\; d\left(c \wedge \overline{c} - \max_{\theta,\theta':\rho_{\mathrm{Ha}}(\theta,\theta')=1} \mathbb{I}_{f_c}\left(\varepsilon_n(\theta), \varepsilon_n(\theta')\right)\right).$$

Proof. Recall that $\rho_{\mathrm{Ha}}(\theta,\theta') = \sum_{i=1}^d \rho_i(\theta,\theta')$, where $\rho_i(\theta,\theta') := [\![\theta_i \neq \theta'_i]\!]$, and each ρ_i is a pseudo metric. Let $\pi(\theta) = \frac{1}{2^d}, \forall \theta \in \{-1,1\}^d$. Also for each $i \in [d]$, let μ_i be the distribution in $\Theta \times \Theta$ such that any random pair $(\theta,\theta') \in \Theta \times \Theta$ drawn according to μ_i satisfies

1. $\theta \sim \pi$
2. $\rho_i(\theta,\theta') = 1$, and $\rho_{\mathrm{Ha}}(\theta,\theta') = 1$ (θ and θ' differ only in the i-th coordinate).

Then the marginal distribution of θ' under μ_i is

$$\sum_{\theta \in \{-1,1\}^d} \mu_i(\theta,\theta') = \frac{1}{2^d} \sum_{\theta \in \{-1,1\}^d} [\![\theta_i \neq \theta'_i \text{ and } \theta_j = \theta'_j, j \neq i]\!] = \frac{1}{2^d} = \pi(\theta'),$$

since by construction of μ, $\rho_{\mathrm{Ha}}(\theta,\theta') = 1$ and for each θ' there is only one θ that differs from it in a single coordinate. Now consider

$$\begin{aligned}
\underline{\mathcal{R}}^{\star}_{\rho_{\mathrm{Ha}}}(\varepsilon_n) &\overset{(i)}{\geq} \inf_{\widehat{\theta}} \mathop{\mathbb{E}}_{\theta\sim\pi}\left[\mathop{\mathbb{E}}_{\mathsf{O}_1^n\sim\varepsilon_n(\theta)}\left[\mathop{\mathbb{E}}_{\tilde{\theta}\sim\widehat{\theta}(\mathsf{O}_1^n)}\left[\rho_{\mathrm{Ha}}\left(\theta,\tilde{\theta}\right)\right]\right]\right] \\
&\overset{(ii)}{=} \inf_{\widehat{\theta}} \sum_{i=1}^d \mathop{\mathbb{E}}_{\theta\sim\pi}\left[\mathop{\mathbb{E}}_{\mathsf{O}_1^n\sim\varepsilon_n(\theta)}\left[\mathop{\mathbb{E}}_{\tilde{\theta}\sim\widehat{\theta}(\mathsf{O}_1^n)}\left[\rho_i\left(\theta,\tilde{\theta}\right)\right]\right]\right] \\
&\geq \sum_{i=1}^d \inf_{\widehat{\theta}} \mathop{\mathbb{E}}_{\theta\sim\pi}\left[\mathop{\mathbb{E}}_{\mathsf{O}_1^n\sim\varepsilon_n(\theta)}\left[\mathop{\mathbb{E}}_{\tilde{\theta}\sim\widehat{\theta}(\mathsf{O}_1^n)}\left[\rho_i\left(\theta,\tilde{\theta}\right)\right]\right]\right] \\
&\overset{(iii)}{\geq} \sum_{i=1}^d \mathop{\mathbb{E}}_{(\theta,\theta')\sim\mu_i}\left[\rho_i(\theta,\theta') \cdot \left(c \wedge \overline{c} - \mathbb{I}_{f_c}\left(\varepsilon_n(\theta), \varepsilon_n(\theta')\right)\right)\right] \\
&\overset{(iv)}{=} \sum_{i=1}^d \mathop{\mathbb{E}}_{(\theta,\theta')\sim\mu_i}\left[\left(c \wedge \overline{c} - \mathbb{I}_{f_c}\left(\varepsilon_n(\theta), \varepsilon_n(\theta')\right)\right)\right] \\
&\geq \sum_{i=1}^d \min_{\theta,\theta':\rho_{\mathrm{Ha}}(\theta,\theta')=1} \left(c \wedge \overline{c} - \mathbb{I}_{f_c}\left(\varepsilon_n(\theta), \varepsilon_n(\theta')\right)\right) \\
&= d\left(c \wedge \overline{c} - \max_{\theta,\theta':\rho_{\mathrm{Ha}}(\theta,\theta')=1} \mathbb{I}_{f_c}\left(\varepsilon_n(\theta), \varepsilon_n(\theta')\right)\right),
\end{aligned}$$

where (i) is due to the fact that the minimax risk is lower bounded by the Bayesian risk (see Theorem 2.4), (ii) is due to $\rho_{\mathrm{Ha}}(\theta,\theta') = \sum_{i=1}^d \rho_i(\theta,\theta')$, (iii) is by Corollary 3.3, and (iv) is by the fact that $\rho_i(\theta,\theta') = 1$ under μ_i for every i. □

If we re-normalize $\mathbb{I}_{f_c}(\cdot,\cdot)$ by $c \wedge \bar{c}$, and define $\mathbb{I}^*_{f_c}(\cdot,\cdot) := \frac{1}{c\wedge\bar{c}}\mathbb{I}_{f_c}(\cdot,\cdot)$, then the minimax lower bound in Theorem 3.4 can be written as follows:

$$\underline{\mathcal{R}}^{\star}_{\rho_{\mathrm{Ha}}}(\varepsilon_n) \;\geq\; d\cdot(c\wedge\bar{c})\left[1-\max_{\theta,\theta':\rho_{\mathrm{Ha}}(\theta,\theta')=1}\mathbb{I}^*_{f_c}(\varepsilon_n(\theta),\varepsilon_n(\theta'))\right].$$

Also note that, by setting $c=\frac{1}{2}$ in Theorem 3.4, we recover the standard Assouad's lemma (Assouad [1983]):

$$\underline{\mathcal{R}}^{\star}_{\rho_{\mathrm{Ha}}}(\varepsilon_n) \;\geq\; \frac{d}{2}\left(1-\frac{1}{2}\max_{\theta,\theta':\rho_{\mathrm{Ha}}(\theta,\theta')=1} d_{\mathrm{TV}}(\varepsilon_n(\theta),\varepsilon_n(\theta'))\right).$$

We use the following two properties of the Hellinger distance $\mathrm{He}^2(P,Q)$ (shown in Chapter 2) to derive a more practically useful version of Assouad's lemma:

- $\mathbb{I}_{f_c}(P,Q) \;\leq\; (c\wedge\bar{c})\cdot\mathrm{He}(P,Q)$, for all distributions $P,Q\in\mathcal{P}(\mathcal{O})$ (refer (2.34))
- $\mathrm{He}^2\left(\bigotimes_{i=1}^k P_i, \bigotimes_{i=1}^k Q_i\right) \;\leq\; \sum_{i=1}^k \mathrm{He}^2(P_i,Q_i)$, for all distributions $P_i,Q_i\in\mathcal{P}(\mathcal{O}_i)$, $i\in[k]$

Armed with these facts, we prove the following version of Assouad's lemma:

Corollary 3.5. *Let $\mathcal{O}$ be some set and $c\in[0,1]$. Define*

$$\mathcal{P}_{\varepsilon}(\mathcal{O}) := \left\{\varepsilon(\theta)\in\mathcal{P}(\mathcal{O}) : \theta\in\{-1,1\}^d\right\}$$

be a class of probability measures induced by the transition $\varepsilon:\{-1,1\}^d \rightsquigarrow \mathcal{O}$. Suppose that there exists some function $\alpha:[0,1]\to\mathbb{R}_+$, such that

$$\mathrm{He}^2(\varepsilon(\theta),\varepsilon(\theta')) \leq \alpha(c), \qquad \text{if } \rho_{\mathrm{Ha}}(\theta,\theta')=1,$$

i.e. the two probability distributions $\varepsilon(\theta)$ and $\varepsilon(\theta')$ (associated with the two parameters θ and θ' which differ only in one coordinate) are sufficiently close w.r.t. Hellinger distance. Then the minimax risk of the parameter estimation problem (3.9) *with parameter space $\Theta=\{-1,1\}^d$ and the loss function $\rho=\rho_{\mathrm{Ha}}$ is bounded below by*

$$\underline{\mathcal{R}}^{\star}_{\rho_{\mathrm{Ha}}}(\varepsilon_n) \;\geq\; d\cdot(c\wedge\bar{c})\cdot\left(1-\sqrt{\alpha(c)\,n}\right). \tag{3.15}$$

Proof. For any two $\theta,\theta'\in\Theta$ with $\rho_{\mathrm{Ha}}(\theta,\theta')=1$, we have

$$\begin{aligned}
\mathbb{I}_{f_c}(\varepsilon_n(\theta),\varepsilon_n(\theta')) &\leq (c\wedge\bar{c})\cdot\mathrm{He}(\varepsilon_n(\theta),\varepsilon_n(\theta')) \\
&\leq (c\wedge\bar{c})\cdot\sqrt{\sum_{i=1}^n \mathrm{He}^2(\varepsilon(\theta),\varepsilon(\theta'))} \\
&\leq (c\wedge\bar{c})\cdot\sqrt{\alpha(c)\,n}
\end{aligned}$$

Substituting this bound into Theorem 3.4 completes the proof. □

The number of training samples n appear in the minimax lower bound (3.15). Thus the hardness of the problem can be expressed as a function of the sample size along with other problem specific parameters.

3.2.2 Minimax Lower Bounds for Cost-sensitive Classification Problem

A natural question to ask regarding cost-sensitive classification problem is how does the hardness of the problem depend upon the cost parameter $c \in [0,1]$. Let $\mathcal{F} \subseteq \{-1,1\}^{\mathcal{X}}$ be the action space and $h \in [0, c \wedge \bar{c}]$ be the margin parameter whose interpretation is explained below. Then we choose a parameter space $\Theta_{h,\mathcal{F}}$ (thus the experiment $\varepsilon_{h,\mathcal{F}} : \Theta_{h,\mathcal{F}} \rightsquigarrow (\mathcal{X} \times \{-1,1\})$) such that:

1. $\forall \theta \in \Theta_{h,\mathcal{F}}$, $f_\theta^\star \in \mathcal{F}$, where $f_\theta^\star$ is given by (3.8). That is we restrict the parameter space s.t. the Bayes classifier associated with each choice of parameter lies within the predetermined function class $\mathcal{F}$.

2. $$|\eta_\theta(\mathsf{X}) - c| \geq h \text{ a.s. } \forall \theta \in \Theta_{h,\mathcal{F}}. \tag{3.16}$$

   This condition is a generalized notion of *Massart noise condition* with margin $h \in [0, c \wedge \bar{c}]$ (Massart and Nédélec [2006]). The motivation for this condition is well established by Massart and Nédélec [2006]. They have argued that under certain "margin" type conditions ([Vapnik and Chervonenkis, 1974; Tsybakov, 2004]) like this, it is possible to design learning algorithms for the binary classification problem, with better rates compared to the case where no such condition is satisfied.

Thus we consider the problem represented by following transition diagram

$$\Theta_{h,\mathcal{F}} \overset{\varepsilon_n}{\rightsquigarrow} (\mathcal{X} \times \{-1,1\})^n \overset{\widehat{f}}{\rightsquigarrow} \mathcal{F} \quad , \tag{3.17}$$

and the minimax risk (in terms of regret) of it given by

$$\underline{\mathcal{R}}^\star_{\Delta\ell_{d_c}}(\varepsilon_n) := \inf_{\widehat{f}} \sup_{\theta \in \Theta_{h,\mathcal{F}}} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(\theta)} \left[\mathop{\mathbb{E}}_{f \sim \widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\Delta\ell_{d_c}(\theta, f)\right] \right]. \tag{3.18}$$

The following is a generalization of the result proved in [Massart and Nédélec, 2006, Theorem 4] for $c = \frac{1}{2}$.

Theorem 3.6. *Let $\mathcal{F}$ be a VC class of binary-valued functions on $\mathcal{X}$ with VC dimension (refer section 3.6.1) $V \geq 2$. Then for any $n \geq V$ and any $h \in [0, c \wedge \bar{c}]$, the minimax risk* (3.18) *of the cost-sensitive binary classification problem* (3.17) *is lower bounded as*

follows:

$$\underline{\mathcal{R}}^{\star}_{\Delta\ell_{d_c}}(\varepsilon_n) \;\geq\; K\cdot(c\wedge\overline{c})\cdot\min\left(\sqrt{\frac{(c\wedge\overline{c})V}{n}},(c\wedge\overline{c})\cdot\frac{V}{nh}\right)$$

where $K>0$ is some absolute constant.

Proof. Instantiate $\Theta=\mathcal{A}=B:=\{-1,1\}^{V-1}$, $\mathcal{O}=\mathcal{X}\times\{-1,1\}$, and $A=\widehat{b}$ in the general learning task (3.1). Then the resulting parameter estimation problem can be represented by the following transition diagram:

$$B \rightsquigarrow^{\varepsilon_n} \mathcal{O}^n \rightsquigarrow^{\widehat{b}} B \quad .$$

Let $\mathcal{P}_{\varepsilon_n}(\mathcal{O}^n):=\{\varepsilon_n(b):=\varepsilon(b)^n\in\mathcal{P}(\mathcal{O})^n : b\in B\}$ be the class of probability measures induced by the experiment ε_n. Then the minimax risk of this problem w.r.t. Hamming distance ρ_{Ha} is given by

$$\underline{\mathcal{R}}^{\star}_{\rho_{\mathrm{Ha}}}(\varepsilon_n) \;=\; \inf_{\widehat{b}}\max_{b\in B}\mathop{\mathbb{E}}_{\mathsf{O}_1^n\sim\varepsilon_n(b)}\left[\mathop{\mathbb{E}}_{b'\sim\widehat{b}(\mathsf{O}_1^n)}\left[\rho_{\mathrm{Ha}}(b,b')\right]\right].$$

Observe that $\mathbb{P}_{\varepsilon(b)}[\mathsf{X}=x,\mathsf{Y}=y]=\mathbb{P}_{\varepsilon(b)}[\mathsf{X}=x]\cdot\mathbb{P}_{\varepsilon(b)}[\mathsf{Y}=y|\mathsf{X}=x]$ for $b\in B$ (by Bayes rule). For simplicity, we will write $\mathbb{P}_{\varepsilon(b)}[\cdot]$ as $\mathbb{P}_b[\cdot]$. Now we will construct these distributions.

Construction of marginal distribution $\mathbb{P}_b[\mathsf{X}=x], x\in\mathcal{X}$: Since $\mathcal{F}$ is a *VC class* with *VC dimension* V, $\exists\{x_1,...,x_V\}\subset\mathcal{X}$ that is shattered, i.e. for any $\beta\in\{-1,1\}^V, \exists f\in\mathcal{F}$ s.t. $f(x_i)=\beta_i,\forall i\in[V]$. Given $p\in[0,1/(V-1)]$, for each $b\in B$, let

$$\mathbb{P}_b[\mathsf{X}=x]=\begin{cases}p, & \text{if } x=x_i \text{ for some } i\in[V-1]\\ 1-(V-1)p, & \text{if } x=x_V\\ 0, & \text{otherwise}\end{cases} \tag{3.19}$$

A particular value for p will be chosen later.

Construction of conditional distribution $\mathbb{P}_b[\mathsf{Y}=y|\mathsf{X}=x], y\in\{-1,1\}, x\in\mathcal{X}$: For each $b\in B$, let

$$\eta_b(x):=\mathbb{P}_b[\mathsf{Y}=1|\mathsf{X}=x]=\begin{cases}c-h, & \text{if } x=x_i \text{ for some } i\in[V-1], \text{ and } b_i=-1\\ c+h, & \text{if } x=x_i \text{ for some } i\in[V-1], \text{ and } b_i=1\\ 0, & \text{otherwise.}\end{cases} \tag{3.20}$$

Then the corresponding Bayes classifier can be given as follows:

$$f_b^{\star}(x)=\begin{cases}-1, & \text{if } x=x_i \text{ for some } i\in[V-1], \text{ and } b_i=-1\\ 1, & \text{if } x=x_i \text{ for some } i\in[V-1], \text{ and } b_i=1\\ -1, & \text{otherwise}\end{cases} \tag{3.21}$$

Now we show that $\{\varepsilon(b) : b \in B\} \subseteq \{\varepsilon(\theta) : \theta \in \Theta_{h,\mathcal{F}}\}$. First of all, from (3.20) we see that $|\eta_b(x) - c| \geq h$ for all x (indeed, $|\eta_b(x) - c| = h$ when $x \in \{x_1, ..., x_{V-1}\}$, and $|\eta_b(x) - c| = c$ otherwise). Second, because $\{x_1, ..., x_V\}$ is shattered by $\mathcal{F}$, there exists at least one $f \in \mathcal{F}$, such that $f_b^\star(x) = f(x)$ for all $x \in \{x_1, ..., x_V\}$. Thus, we get $B \subset \Theta_{h,\mathcal{F}}$.

Reduction to Parameter Estimation Problem: We start with the following observation

$$\underline{\mathcal{R}}^\star_{\Delta\ell_{d_c}}(\varepsilon_n) \;\geq\; \inf_{\widehat{f}} \max_{b\in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\ell_{d_c}(b, f) - \ell_{d_c}(b, f_b^\star)\right] \right],$$

since $B \subset \Theta_{h,\mathcal{F}}$. Define $M_\theta(x) := \mathbb{P}_\theta[\mathsf{X} = x]$, and $\eta_\theta(x) := \mathbb{P}_\theta[\mathsf{Y} = 1 \mid \mathsf{X} = x]$, for $x \in \mathcal{X}$. By Lemma 3.1, for any classifier $f : \mathcal{X} \to \{-1, 1\}$ and any $\theta \in \Theta_{h,\mathcal{F}}$, we have

$$\ell_{d_c}(\theta, f) - \ell_{d_c}(\theta, f_\theta^\star) \;=\; \frac{1}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X}\sim M_\theta} \left[|\eta_\theta(\mathsf{X}) - c| \cdot |f(\mathsf{X}) - f_\theta^\star(\mathsf{X})|\right].$$

If $\theta \in \Theta_{h,\mathcal{F}}$, then using the above equation and the margin condition (3.16) we get

$$\ell_{d_c}(\theta, f) - \ell_{d_c}(\theta, f_\theta^\star) \;\geq\; \frac{h}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X}\sim M_\theta} \left[|f(\mathsf{X}) - f_\theta^\star(\mathsf{X})|\right] = \frac{h}{2} \cdot \|f - f_\theta^\star\|_{L_1(M_\theta)},$$

where $\|f\|_{L_1(M_\theta)}$ is given by (2.5) with $p = 1$ and $\mu = M_\theta$. Since there is no confusion, we can simply drop M_θ and write the L_1 norm as $\|\cdot\|_{L_1}$. Hence we have

$$\begin{aligned}
&\inf_{\widehat{f}} \max_{b\in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\ell_{d_c}(b, f) - \ell_{d_c}(b, f_b^\star)\right] \right] \\
\geq\; &\frac{h}{2} \cdot \inf_{\widehat{f}} \max_{b\in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\|f - f_b^\star\|_{L_1}\right] \right].
\end{aligned}$$

Define

$$b_f \;:=\; \arg\min_{b\in B} \|f - f_b^\star\|_{L_1}.$$

Then for any $b \in B$,

$$\left\|f_{b_f}^\star - f_b^\star\right\|_{L_1} \leq \left\|f_{b_f}^\star - f\right\|_{L_1} + \|f - f_b^\star\|_{L_1} \leq 2\|f - f_b^\star\|_{L_1},$$

where the first inequality is due to the triangle inequality and the second follows from the definitions of b_f and $f_\theta^\star$. Thus we have

$$\begin{aligned}
&\inf_{\widehat{f}} \max_{b\in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\ell_{d_c}(b, f) - \ell_{d_c}(b, f_b^\star)\right] \right] \\
\geq\; &\frac{h}{4} \cdot \inf_{\widehat{f}} \max_{b\in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\left\|f_{b_f}^\star - f_b^\star\right\|_{L_1}\right] \right]
\end{aligned}$$

$$= \frac{h}{4} \cdot \inf_{\widehat{b}} \max_{b \in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{b' \sim \widehat{b}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\left\| f_{b'}^\star - f_b^\star \right\|_{L_1} \right] \right].$$

For any two $b, b' \in B$, we have

$$\begin{aligned} \left\| f_{b'}^\star - f_b^\star \right\|_{L_1} &= \int_{\mathcal{X}} |f_{b'}^\star(x) - f_b^\star(x)| \mathbb{P}_b \left[\mathsf{X} = x \right] dx \\ &= p \sum_{i=1}^{V-1} |f_{b'}^\star(x_i) - f_b^\star(x_i)| \\ &= p \sum_{i=1}^{V-1} |b_i' - b_i| \\ &= 2p \cdot \rho_{\mathrm{Ha}}\left(b, b'\right), \end{aligned}$$

where the second and third equalities are from (3.19) and (3.21). Finally we get

$$\begin{aligned} & \inf_{\widehat{f}} \max_{b \in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{f \sim \widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\ell_{d_c}\left(b, f\right) - \ell_{d_c}\left(b, f_b^\star\right) \right] \right] \\ \geq\ & \frac{ph}{2} \cdot \inf_{\widehat{b}} \max_{b \in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{b' \sim \widehat{b}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\rho_{\mathrm{Ha}}\left(b, b'\right) \right] \right] \\ =\ & \frac{ph}{2} \cdot \underline{\mathcal{R}}_{\rho_{\mathrm{Ha}}}^\star\left(\varepsilon_n\right). \end{aligned} \tag{3.22}$$

Applying Assouad's Lemma: For any two $b, b' \in B$ we have

$$\begin{aligned} \mathrm{He}^2\left(\varepsilon\left(b\right), \varepsilon\left(b'\right)\right) &= \sum_{i=1}^{V} \sum_{y \in \{-1,1\}} \left(\sqrt{\mathbb{P}_b\left[\mathsf{X} = x_i, \mathsf{Y} = y\right]} - \sqrt{\mathbb{P}_{b'}\left[\mathsf{X} = x_i, \mathsf{Y} = y\right]} \right)^2 \\ &= p \sum_{i=1}^{V-1} \sum_{y \in \{-1,1\}} \left(\sqrt{\mathbb{P}_b\left[\mathsf{Y} = y | \mathsf{X} = x_i\right]} - \sqrt{\mathbb{P}_{b'}\left[\mathsf{Y} = y | \mathsf{X} = x_i\right]} \right)^2 \\ &= p \sum_{i=1}^{V-1} [\![b_i \neq b_i']\!] \left\{ \left(\sqrt{c-h} - \sqrt{c+h} \right)^2 + \left(\sqrt{\overline{c}-h} - \sqrt{\overline{c}+h} \right)^2 \right\} \\ &= 2p \left(1 - \sqrt{c^2 - h^2} - \sqrt{\overline{c}^2 - h^2} \right) \rho_{\mathrm{Ha}}\left(b, b'\right), \end{aligned}$$

where the second and third equalities are from (3.19) and (3.20). Thus the condition of the Corollary 3.5 is satisfied with

$$2p \left(1 - \sqrt{c^2 - h^2} - \sqrt{\overline{c}^2 - h^2} \right) \ \leq\ 4p \frac{h^2}{c \wedge \overline{c}} \ =:\ \alpha\left(c\right),$$

where the inequality is from Lemma 3.7 (see below). Therefore we get

$$\begin{aligned}
&\inf_{\widehat{f}} \max_{b\in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\sim\varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\ell_{d_c}(b,f) - \ell_{d_c}(b,f_b^\star)\right]\right] \\
&\geq \frac{ph(V-1)}{2}\left(c\wedge\overline{c} - c\wedge\overline{c}\cdot\sqrt{4p\frac{h^2}{c\wedge\overline{c}}n}\right) \\
&= \frac{ph(V-1)}{2}\left(c\wedge\overline{c} - 2h\sqrt{c\wedge\overline{c}\cdot pn}\right),
\end{aligned}$$

where the first inequality is due to (3.15) and (3.22). If we let $p = \frac{c\wedge\overline{c}}{9nh^2}$, then the term in the parentheses will be equal to $\frac{c\wedge\overline{c}}{3}$, and

$$\inf_{\widehat{f}} \max_{b\in B} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\sim\varepsilon_n(b)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\ell_{d_c}(b,f) - \ell_{d_c}(b,f_b^\star)\right]\right] \;\geq\; \frac{(c\wedge\overline{c})^2(V-1)}{54nh},$$

assuming that the condition $p \leq 1/(V-1)$ holds. This will be the case if $h \geq \sqrt{\frac{c\wedge\overline{c}(V-1)}{9n}}$. Therefore

$$\underline{\mathcal{R}}^\star_{\Delta\ell_{d_c}}(\varepsilon_n) \;\geq\; \frac{(c\wedge\overline{c})^2(V-1)}{54nh}, \quad \text{if } h \geq \sqrt{\frac{c\wedge\overline{c}(V-1)}{9n}}. \tag{3.23}$$

If $h \leq \sqrt{\frac{c\wedge\overline{c}(V-1)}{9n}}$, we can use the above construction with $\tilde{h} = \sqrt{\frac{c\wedge\overline{c}(V-1)}{9n}}$. Then, because $\Theta_{\tilde{h},\mathcal{F}} \subseteq \Theta_{h,\mathcal{F}}$ whenever $\tilde{h} \geq h$, we see that

$$\begin{aligned}
&\inf_{\widehat{f}} \sup_{\theta\in\Theta_{h,\mathcal{F}}} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\sim\varepsilon_n(\theta)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\ell_{d_c}(\theta,f) - \ell_{d_c}(\theta,f_\theta^\star)\right]\right] \\
&\geq \inf_{\widehat{f}} \sup_{\theta\in\Theta_{\tilde{h},\mathcal{F}}} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\sim\varepsilon_n(\theta)} \left[\mathop{\mathbb{E}}_{f\sim\widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\ell_{d_c}(\theta,f) - \ell_{d_c}(\theta,f_\theta^\star)\right]\right] \\
&\geq \frac{(c\wedge\overline{c})^2(V-1)}{54n\tilde{h}} \\
&= \frac{(c\wedge\overline{c})^{\frac{3}{2}}}{18}\sqrt{\frac{V-1}{n}}, \quad \text{if } h \leq \sqrt{\frac{c\wedge\overline{c}(V-1)}{9n}}.
\end{aligned} \tag{3.24}$$

Observe that $\frac{(c\wedge\overline{c})^2(V-1)}{54nh} \leq \frac{(c\wedge\overline{c})^{\frac{3}{2}}}{18}\sqrt{\frac{V-1}{n}}$ if $h \geq \sqrt{\frac{c\wedge\overline{c}(V-1)}{9n}}$, and $\frac{(c\wedge\overline{c})^2(V-1)}{54nh} > \frac{(c\wedge\overline{c})^{\frac{3}{2}}}{18}\sqrt{\frac{V-1}{n}}$ otherwise. Then combining (3.23) and (3.24) completes the proof. □

Lemma 3.7. *For $h \in [0, c\wedge\overline{c}]$, we have $1 - \sqrt{c^2-h^2} - \sqrt{\overline{c}^2-h^2} \;\leq\; 2\frac{h^2}{c\wedge\overline{c}}$.*

Proof. Let $A = 1 - \sqrt{c^2-h^2} - \sqrt{\overline{c}^2-h^2}$. Take series expansion of A w.r.t. h to get

$$A \;=\; \frac{1}{2}\left(\frac{1}{c}+\frac{1}{\overline{c}}\right)h^2 + \frac{1}{8}\left(\frac{1}{c^3}+\frac{1}{\overline{c}^3}\right)h^4 + \frac{1}{16}\left(\frac{1}{c^5}+\frac{1}{\overline{c}^5}\right)h^6 + \frac{5}{128}\left(\frac{1}{c^7}+\frac{1}{\overline{c}^7}\right)h^8$$

$$+\frac{7}{256}\left(\frac{1}{c^9}+\frac{1}{\overline{c}^9}\right)h^{10}+\frac{21}{1024}\left(\frac{1}{c^{11}}+\frac{1}{\overline{c}^{11}}\right)h^{12}+\frac{33}{2048}\left(\frac{1}{c^{13}}+\frac{1}{\overline{c}^{13}}\right)h^{14}+\cdots.$$

Now $\frac{1}{2}\left(\frac{1}{c}+\frac{1}{\overline{c}}\right)\leq\frac{1}{c}\vee\frac{1}{\overline{c}}=\frac{1}{c\wedge\overline{c}}$ (since average is less than maximum). Thus

$$A\leq h\left[\frac{h}{c\wedge\overline{c}}+\frac{1}{4}\frac{h^3}{(c\wedge\overline{c})^3}+\frac{1}{8}\frac{h^5}{(c\wedge\overline{c})^5}+\frac{5}{64}\frac{h^7}{(c\wedge\overline{c})^7}+\frac{7}{128}\frac{h^9}{(c\wedge\overline{c})^9}+\frac{21}{512}\frac{h^{11}}{(c\wedge\overline{c})^{11}}+\cdots\right].$$

Now we have

$$\begin{aligned}h\leq c\wedge\overline{c}&\implies\frac{h}{c\wedge\overline{c}}\leq 1\\&\implies\left(\frac{h}{c\wedge\overline{c}}\right)^{\alpha}\leq\frac{h}{c\wedge\overline{c}},\forall\alpha\geq 1.\end{aligned}$$

Thus

$$\begin{aligned}A\leq&h\left[\frac{h}{c\wedge\overline{c}}+\frac{h}{c\wedge\overline{c}}\left\{\frac{1}{4}+\frac{1}{8}+\frac{5}{64}+\frac{7}{128}+\frac{21}{512}+\cdots\right\}\right]\\\leq&h\left[\frac{2h}{c\wedge\overline{c}}\right]=\frac{2h^2}{c\wedge\overline{c}},\end{aligned}$$

where the second inequality follows from the fact that (can be shown with the aid of computer or using the properties of gamma function)

$$\frac{1}{4}+\frac{1}{8}+\frac{5}{64}+\frac{7}{128}+\frac{21}{512}+\cdots\leq 1.$$

□

When $h=0$ (or being too small), we get a minimax lower bound of order $(c\wedge\overline{c})^{\frac{3}{2}}\cdot\sqrt{\frac{V}{n}}$, and when $h=c\wedge\overline{c}$, we obtain a bound of the order $(c\wedge\overline{c})\cdot\frac{V}{n}$.

When $c=1/2$ in Theorem 3.6, we recover the standard minimax lower bounds for balanced binary classification (with zero-one loss function) presented in [Massart and Nédélec, 2006, Theorem 4]:

$$\underline{\mathcal{R}}^{\star}_{\Delta\ell^{0-1}}(\varepsilon_n)\ \geq\ K'\cdot\min\left(\sqrt{\frac{V}{n}},\frac{V}{nh}\right),$$

for some constant $K'>0$.

It would be interesting to study the hardness of the following classification problem settings which are closely related to the binary cost-sensitive classification problem that we considered in this thesis:

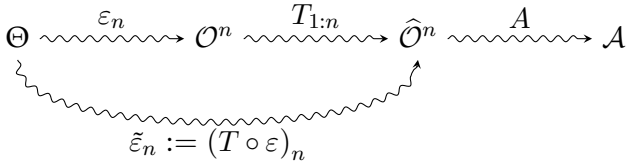
1. Cost-sensitive classification with example dependent costs ([Zadrozny and Elkan, 2001; Zadrozny et al., 2003]).

2. Binary classification problem w.r.t. generalized performance measures (Koyejo et al. [2014]) such as arithmetic, geometric and harmonic means of the true posi-

tive and true negative rates. These measures are more appropriate for imbalanced classification problem ([Cardie and Nowe, 1997; Elkan, 2001]) than the usual classification accuracy.

3.3 Constrained Learning Problem

In the normal theoretical analysis of the learning problem (3.1), it is assumed that the decision maker has access to clean data (in $\mathcal{O}^n$), that their observations are from the pattern they are expected to predict. In the real world, this is usually not the case, data is normally corrupted or needs to be corrupted to meet privacy requirements. We can formalize these constraints (noisy data and privacy requirements) in the language of transitions by introducing the channel $T : \mathcal{O} \rightsquigarrow \widehat{\mathcal{O}}$, where $\widehat{\mathcal{O}}$ is the new observation space for the decision maker. The *Constrained learning task* (denoted by $(\ell, \varepsilon_n, T_{1:n})$, where ε_n is the repeated experiment (2.16), and $T_{1:n}$ is the parallelized transition (2.18)), can be represented by the following transition diagram:



. (3.25)

For convenience we define the corrupted experiment $\tilde{\varepsilon} := T \circ \varepsilon$, and denote the repeated corrupted experiment by $\tilde{\varepsilon}_n$. We study the hardness of this constrained learning problem (3.25) by producing the minimax lower bounds of it. From the *Weak Data Processing Inequality* ($\mathbb{I}_f(\tilde{\varepsilon}_n) \leq \mathbb{I}_f(\varepsilon_n)$), we have

$$\underline{\mathcal{R}}^\star_\ell(\tilde{\varepsilon}_n) \;\geq\; \underline{\mathcal{R}}^\star_\ell(\varepsilon_n),$$

i.e. the constrained problem is harder than the original unconstrained problem. Then the minimax lower bounds for the unconstrained problem (3.1) are applicable to this constrained task as well. However, this provides us with no means to compare various choices of channel T for a given problem.

For some T, the weak data processing theorem can be strengthened, in the sense that one can find a constant called the *contraction coefficient* $\eta_f(T) < 1$ (formally defined in Definition 3.8) such that $\mathbb{I}_f(\tilde{\varepsilon}_n) \leq \eta_f(T)\,\mathbb{I}_f(\varepsilon_n)$, for all experiments $\varepsilon : \Theta \rightsquigarrow \mathcal{O}$. The strengthened inequality is called the *Strong Data Processing Inequality*. The contraction coefficient $\eta_f(T)$ provides a means to measure the amount of corruption present in T. For example if T is constant and maps all input distributions to the same output distribution, then $\eta_f(T) = 0$. If T is an invertible function, then $\eta_f(T) = 1$. Together with the minimax lower bounds of unconstrained problem, this strong data processing inequality leads to meaningful lower bounds that allow the comparison of different corrupted experiments. In what follows we will present some new results on strong data processing inequalities. Specifically, we will derive an explicit closed form

for the contraction coefficient of any channel w.r.t. c-primitive f-divergence, and obtain efficiently computable lower bounds for the contraction coefficients of binary symmetric channels w.r.t. symmetric f-divergences.

3.3.1 Strong Data Processing Inequalities

Consider a channel $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$, with finite input ($\mathcal{X}$) and output ($\mathcal{Y}$) spaces (also let $|\mathcal{X}| \geq 2$ and $|\mathcal{Y}| \geq 2$). Observe that T can also be viewed as an experiment with a different parameter space $\mathcal{X}$. The classes of probability measures generated by the experiment ε and the channel T are given by $\mathcal{P}_{\varepsilon}(\mathcal{O}) := \{\varepsilon(\theta) \in \mathcal{P}(\mathcal{O}) : \theta \in \Theta\}$ and $\mathcal{P}_T(\mathcal{Y}) := \{T(x) \in \mathcal{P}(\mathcal{Y}) : x \in \mathcal{X}\}$ respectively.

Strong data processing inequalities has become an intensive research area in the information theory community recently (see Raginsky [2014] and references therein). Early work includes Ahlswede and Gács [1976] and Dobrushin [1956a] (see Cohen et al. [1993] for further history).

Below we formally define the contraction coefficient $\eta_f(T)$ of the channel T w.r.t. f-divergence. Recall that $\mathcal{P}_*(\mathcal{X})$ means the set of all strictly positive distributions over $\mathcal{X}$.

Definition 3.8. *Given a transition $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $\mu \in \mathcal{P}_*(\mathcal{X})$, we define*

$$\begin{aligned} \eta_f(\mu, T) &:= \sup_{\nu \neq \mu, \nu \in \mathcal{P}(\mathcal{X})} \frac{\mathbb{I}_f(T \circ \nu, T \circ \mu)}{\mathbb{I}_f(\nu, \mu)} \\ \eta_f(T) &:= \sup_{\mu \in \mathcal{P}_*(\mathcal{X})} \eta_f(\mu, T). \end{aligned}$$

If $\eta_f(\mu, T) < 1$, we say that T satisfies the strong data processing inequality (SDPI) at μ w.r.t. f-divergence i.e.

$$\mathbb{I}_f(T \circ \nu, T \circ \mu) \leq \eta_f(\mu, T) \mathbb{I}_f(\nu, \mu)$$

for all $\nu \in \mathcal{P}(\mathcal{X})$. Moreover if $\eta_f(T) < 1$ we say that T satisfies the SDPI w.r.t. f-divergence i.e.

$$\mathbb{I}_f(T \circ \nu, T \circ \mu) \leq \eta_f(T) \mathbb{I}_f(\nu, \mu)$$

for all $\nu \in \mathcal{P}(\mathcal{X})$ and $\mu \in \mathcal{P}_*(\mathcal{X})$.

Cohen et al. [1993] showed that the contraction coefficient $\eta_f(T)$ of any channel T with respect to any f-divergence is universally upper-bounded by the so-called *Dobrushin's coefficient* of T [Dobrushin, 1956a,b]. Dobrushin's coefficient is extensively studied in the context of Markov chains (see Paz [1971] and Isaacson and Madsen [1976] for detailed discussions).

Theorem 3.9. *Define the Dobrushin's coefficient [Dobrushin, 1956a,b] of a channel $T \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$ as*

$$\vartheta(T) := \frac{1}{2} \max_{x, x' \in \mathcal{X}} d_{\mathrm{TV}}(T(x), T(x')). \tag{3.26}$$

Then the contraction coefficient of the channel T w.r.t. any f-divergence is bounded above as follows

$$\eta_f(T) \;\leq\; \eta_{\mathrm{TV}}(T) \;=\; \vartheta(T), \tag{3.27}$$

where $\eta_{\mathrm{TV}}(T)$ is the contraction coefficient of the channel T w.r.t. total variation divergence (i.e. $\mathrm{TV}(x) = |x-1|$).

We are interested in the question of how loose the bound in Theorem 3.9 can be. For this we study the contraction coefficients w.r.t. c-primitive f-divergences (2.32). Recall that any f-divergence can be written as a weighted integral of c-primitive f-divergences as given in Theorem 2.7.

As a starting point, we define the *generalized Dobrushin's coefficient* of a channel $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ as follows

$$\vartheta_c(T) \;:=\; \frac{1}{c \wedge \bar{c}} \cdot \max_{x,x' \in \mathcal{X}} \mathbb{I}_{f_c}(T(x), T(x')), \tag{3.28}$$

where $c \in (0,1)$ and $\vartheta_0(T) = \vartheta_1(T) := 0$. Note that when $c = \frac{1}{2}$, this gives the standard Dobrushin's coefficient. From the definition (3.28) we can note the following properties of $\vartheta_c(T)$:

1. For a given T, $\vartheta_c(T)$ is symmetric in c about $\frac{1}{2}$ i.e. $\vartheta_c = \vartheta_{\bar{c}}$.
2. For the fully-informative channel we have $\vartheta_c(T_{id}) = 1$ (can be easily shown from the definition), and for the non-informative channel $\vartheta_c(T_{\bullet\mathcal{X}}) = 0$. Thus for any channel T we have $0 \leq \vartheta_c(T) \leq 1$.
3. For two channels T_1 and T_2, it is possible that $\vartheta_c(T_1) > \vartheta_c(T_2)$ and $\vartheta_{c'}(T_1) < \vartheta_{c'}(T_2)$ i.e. optimal channel choice based on contraction coefficient will depend on c (see Figure 3.1).

We have plotted the generalized Dobrushin's coefficient for binary channels $T : [2] \rightsquigarrow [2]$, and ternary channels $T : [3] \rightsquigarrow [3]$. Based on those observations (see Figure 3.1), we conjecture the following properties of $\vartheta_c(T)$:

1. For a given T, and for any $0 < c < c' < 0.5$, $\vartheta_c(T) \leq \vartheta_{c'}(T)$.
2. The maximum point of the curve $\vartheta_c(T)$ always occurs at $c = \frac{1}{2}$ (this matches the fact that $\eta_{\mathrm{TV}}(T)$ upper bounds the contraction coefficient $\eta_f(T)$ of any f-divergence).

We now relate the contraction coefficient $\eta_{f_c}(T)$ w.r.t. c-primitive f-divergence to the generalized Dobrushin's coefficient $\vartheta_c(T)$ defined above.

Theorem 3.10. *For any $T \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, and $c \in [0,1]$ we have*

$$\eta_{f_c}(T) \;=\; \vartheta_c(T)$$

where $\eta_{f_c}(T)$ is the contraction coefficient of the channel T w.r.t. c-primitive f-divergence.

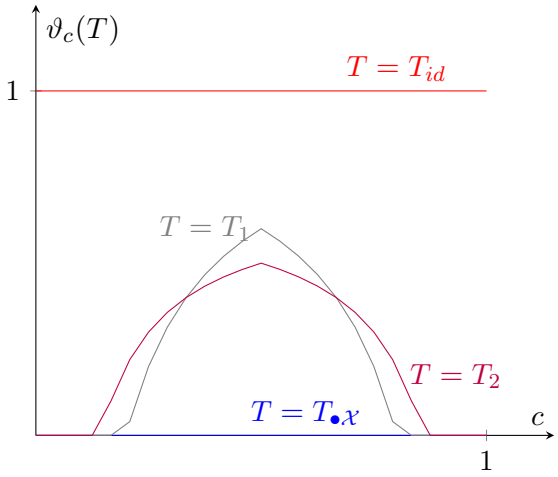


Figure 3.1: Generalized Dobrushin's coefficient of two arbitrary channels $T_1 = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$ (—) and $T_2 = \begin{bmatrix} 0.1 & 0.6 \\ 0.9 & 0.4 \end{bmatrix}$ (—), fully-informative channel T_{id} (—), and non-informative channel $T_{\bullet\mathcal{X}}$ (—).

Proof. We follow the proof of Theorem 3.1 in [Raginsky, 2014].

We start with the following generalization of *strong Markov contraction lemma* from Cohen et al. [1993]: for any signed measure $\tilde{\nu}$ on $\mathcal{X}$, any $c \in [0,1]$ and any Markov kernel $T \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, we have

$$\|T \circ \tilde{\nu}\|_{\mathrm{TV}} \leq \vartheta_c(T) \|\tilde{\nu}\|_{\mathrm{TV}} + (1 - \vartheta_c(T)) |\tilde{\nu}(\mathcal{X})|, \tag{3.29}$$

where the total variation norm $\|\tilde{\nu}\|_{\mathrm{TV}}$ of signed measure $\tilde{\nu}$ is given by $\|\tilde{\nu}\|_{\mathrm{TV}} := \sum_{x \in \mathcal{X}} |\tilde{\nu}(x)|$. This can be shown by simply following through the steps of the proof of Lemma 3.2 in Cohen et al. [1993]. Let $\tilde{\nu} = c\nu - \bar{c}\mu$, where ν and μ are probability measures on $\mathcal{X}$. Then $T \circ \tilde{\nu} = cT \circ \nu - \bar{c}T \circ \mu$ and $|\tilde{\nu}(\mathcal{X})| = |2c - 1|$. By using (3.29) and the definition of c-primitive f-divergence (2.32), we get

$$\begin{aligned} & d_{\mathrm{TV}}(cT \circ \nu, \bar{c}T \circ \mu) \;\leq\; \vartheta_c(T)\, d_{\mathrm{TV}}(c\nu, \bar{c}\mu) + (1 - \vartheta_c(T)) |2c - 1| \\ \Longrightarrow\; & d_{\mathrm{TV}}(cT \circ \nu, \bar{c}T \circ \mu) - |2c - 1| \;\leq\; \vartheta_c(T) \{d_{\mathrm{TV}}(c\nu, \bar{c}\mu) - |2c - 1|\} \\ \Longrightarrow\; & \mathbb{I}_{f_c}(T \circ \nu, T \circ \mu) \;\leq\; \vartheta_c(T) \cdot \mathbb{I}_{f_c}(\nu, \mu). \end{aligned}$$

Now it remains to show that this bound is achieved for some probability measures μ and ν.

To that end, let us first assume that $|\mathcal{X}| > 2$. Let $x_0, x_1 \in \mathcal{X}$ achieve the maximum in (3.28), pick some $\epsilon_1, \epsilon_2, \epsilon \in (0,1)$ such that $\epsilon_1 \neq \epsilon_2, \epsilon_1 + \epsilon < 1, \epsilon_2 + \epsilon < 1$, and consider the following probability distributions:

- ν that puts the mass $1 - \epsilon_1 - \epsilon$ on x_0, ϵ_1 on x_1, and distributes the remaining mass of ϵ evenly among the set $\mathcal{X} \setminus \{x_0, x_1\}$;
- μ that puts the mass $1 - \epsilon_2 - \epsilon$ on x_0, ϵ_2 on x_1, and distributes the remaining

mass of ϵ evenly among the set $\mathcal{X} \setminus \{x_0, x_1\}$.

Then a simple calculation (using a mathematical software) gives

$$\mathbb{I}_{f_c}(\nu,\mu) = \frac{1}{2}\left\{|c\epsilon_1 - \overline{c}\epsilon_2| + |c\overline{(\epsilon_1+\epsilon)} - \overline{c}\overline{(\epsilon_2+\epsilon)}| + (|\mathcal{X}|-2)\cdot|c\epsilon - \overline{c}\epsilon| - |2c-1|\right\}$$
$$\mathbb{I}_{f_c}(T\circ\nu, T\circ\mu) = \vartheta_c(T)\cdot\mathbb{I}_{f_c}(\nu,\mu).$$

For $|\mathcal{X}| = 2$, the idea is the same, except that there is no need for the extra slack ϵ. □

From the above theorem, for any $T \in \mathcal{M}(\mathcal{X},\mathcal{Y})$ and for all $P, Q \in \mathcal{P}(\mathcal{X})$, we have

$$\mathbb{I}_{f_c}(T\circ P, T\circ Q) \;\le\; \vartheta_c(T)\,\mathbb{I}_{f_c}(P,Q).$$

Thus for any general f-divergence (using the weighted integral representation) we get

$$\begin{aligned}\mathbb{I}_f(T\circ P, T\circ Q) &= \int_0^1 \gamma_f(c)\,\mathbb{I}_{f_c}(T\circ P, T\circ Q)\,dc \\ &\le \int_0^1 \vartheta_c(T)\,\gamma_f(c)\,\mathbb{I}_{f_c}(P,Q)\,dc \\ &\le \vartheta_{\frac{1}{2}}(T)\cdot\mathbb{I}_f(P,Q),\end{aligned} \tag{3.30}$$

where the last inequality is due to the fact that $\vartheta_c(T) \le \vartheta_{\frac{1}{2}}(T)$. Even though this doesn't fully answer the question of how loose the universal bound (3.27) can be, (3.30) along with the Figure 3.1, sheds some light in that direction.

Remark 3.11. *Let f be a convex function with $f(1) = 0$, which can be written in the following form*

$$f(u) = \alpha u + \beta u^2 + \int_0^\infty \left(\frac{tu}{1+t^2} - \frac{u}{u+t}\right) v(dt),$$

where $\alpha \in \mathbb{R}, \beta \ge 0$, and v is a non-negative measure on $\mathbb{R}_+$ such that $\int_0^\infty \frac{1}{1+t^2} v(dt) < \infty$. Note that for such function f, we have

$$f''(u) \;=\; 2\beta + 2\int_0^\infty \frac{t}{(u+t)^3} v(dt)$$

and thus

$$\gamma_f(c) \;=\; \frac{2}{c^3} f''\left(\frac{\overline{c}}{c}\right) = \frac{2}{c^3}\left(\beta + \int_0^\infty \frac{t}{\left(\frac{\overline{c}}{c}+t\right)^3} v(dt)\right). \tag{3.31}$$

Raginsky [2014] has shown that for this class of functions, $\eta_f(T) = S(T)^2$, where

$$S(T) \;:=\; \sup_{\mu}\sup_{f,g} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim\mu\otimes T}[f(\mathsf{X})g(\mathsf{Y})],$$

for $\mathbb{E}[f(\mathsf{X})] = \mathbb{E}[g(\mathsf{Y})] = 0$ and $\mathbb{E}[f(\mathsf{X})^2] = \mathbb{E}[g(\mathsf{Y})^2] = 1$. Still it is hard to compute $S(T)$ for any non-trivial channels.

Following divergences are generated from the functions that satisfy the above mentioned conditions:

- KL-*divergence satisfies* (3.31) *with* $\gamma_{\mathrm{KL}}(c) = \frac{1}{c^2 \bar{c}}$, $\beta = 0$, *and* $v(dt) = dt$.
- χ^2-*divergence satisfies* (3.31) *with* $\gamma_{\chi^2}(c) = \frac{1}{c^3}$, $\beta = 1$, *and* $v(dt) = 0$.
- *squared Hellinger divergence satisfies* (3.31) *with* $\gamma_{\mathrm{He}^2}(c) = \frac{1}{2(c\bar{c})^{\frac{3}{2}}}$, $\beta = 0$, *and* $v(dt) = \frac{2}{\pi\sqrt{t}} dt$.

3.3.2 Binary Symmetric Channels

Despite the extensive research in the strong data processing inequalities, an efficiently computable closed form for the contraction coefficient of a channel w.r.t. most of the f-divergences are not known. Indeed it is not understood at least for the simplest case of symmetric channels. Here we point to a technical report Makur and Polyanskiy [2016], which attempts to find simpler criteria for a given channel T being dominated by a symmetric channel W (in the sense that $\mathrm{I}_{\mathrm{KL}}(W \circ \mu, W \circ \nu) \geq \mathrm{I}_{\mathrm{KL}}(T \circ \mu, T \circ \nu)$, $\forall \mu, \nu \in \mathcal{P}(\mathcal{X})$). This suggests that obtaining an efficiently computable closed form for the contraction coefficient of symmetric channels might guide us in upper bounding the contraction coefficient of general channels.

In any case, since $\eta_f(T)$ is not known in a computable form for any non-trivial channels, we will now consider the *binary symmetric channel* (BSC) $T : [2] \rightsquigarrow [2]$. Let X and Y be the input and output random variables of the channel respectively. This BSC can be written in a matrix form as follows

$$T = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}, \quad p \in [0,1]$$

where the rows represent the outputs of the channel, and the columns represent the inputs. The (i,j)-th entry of the matrix represents $\mathbb{P}[\mathsf{Y} = i \mid \mathsf{X} = j]$.

To better understand the insights of the contraction coefficients of a BSC w.r.t. f-divergences, we consider the restrictive setting of symmetric f-divergences with symmetric experiments. First we define the following classes:

$$\begin{aligned} \mathcal{F}_{\mathrm{symm}} &:= \{f : (0,\infty) \to \mathbb{R} : f \text{ is convex}, f(1) = 0, \text{ and } f(x) = f^{\diamond}(x), \forall x\}, \\ \mathcal{P}_{\mathrm{symm}} &:= \left\{ P : P = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix} \text{ for some } p \in (0,1) \right\}. \end{aligned}$$

We note the following properties of the function class $\mathcal{F}_{\mathrm{symm}}$:

- $\mathrm{I}_f(P,Q) = \mathrm{I}_f(Q,P)$ for all distributions P and Q and $f \in \mathcal{F}_{\mathrm{symm}}$.
- For any $f, g \in \mathcal{F}_{\mathrm{symm}}$ and $\alpha, \beta \in \mathbb{R}$, we have $\alpha f + \beta g \in \mathcal{F}_{\mathrm{symm}}$.
- Defining a function $f \in \mathcal{F}_{\mathrm{symm}}$ only in the domain $(0,1]$ is sufficient, as we can extend it for $x \in [1,\infty)$ by using the symmetric property $f(x) = f^{\diamond}(x) = x \cdot f\left(\frac{1}{x}\right)$.

Define $f_{\text{tv}}(x) := |x-1|$, $f_{\text{tri}}(x) := \frac{(x-1)^2}{x+1}$, $f_{\text{He}}(x) := \left(\sqrt{x}-1\right)^2$, $f_{\text{tvtri}}(x) := \frac{f_{\text{tv}}(x)+f_{\text{tri}}(x)}{2}$, $f_{\text{tvHe}}(x) := \frac{f_{\text{tv}}(x)+f_{\text{He}}(x)}{2}$, and $f_{\text{triHe}}(x) := \frac{f_{\text{tri}}(x)+f_{\text{He}}(x)}{2}$. Here $f_{\text{tv}}, f_{\text{tri}}$, and f_{He} are associated with total variation, triangular discrimination, and squared Hellinger divergences respectively (Reid and Williamson [2011]). One can easily verify that f_{tv}, f_{tri}, f_{He}, f_{tvtri}, f_{tvHe}, $f_{\text{triHe}} \in \mathcal{F}_{\text{symm}}$.

Note that the composition of the channel $T = \begin{bmatrix} t & 1-t \\ 1-t & t \end{bmatrix} \in \mathcal{P}_{\text{symm}}$ and the symmetric experiment $E = \begin{bmatrix} e & 1-e \\ 1-e & e \end{bmatrix} \in \mathcal{P}_{\text{symm}}$ can be written as follows

$$T \circ E = \begin{bmatrix} te+(1-t)(1-e) & t(1-e)+(1-t)e \\ t(1-e)+(1-t)e & te+(1-t)(1-e) \end{bmatrix} = \begin{bmatrix} s(t,e) & 1-s(t,e) \\ 1-s(t,e) & s(t,e) \end{bmatrix},$$

where $s(t,e) := te+(1-t)(1-e)$. The auxiliary notion that we are mainly interested here is:

$$\eta_f^{\text{symm}}(T) := \sup_{E \in \mathcal{P}_{\text{symm}}} \frac{\mathbb{I}_f(T \circ E)}{\mathbb{I}_f(E)} \quad \text{where } f \in \mathcal{F}_{\text{symm}}, \text{ and } T \in \mathcal{P}_{\text{symm}}. \tag{3.32}$$

Observe that $\eta_f^{\text{symm}}(T) \leq \eta_f(T)$ for any BSC T (where $\eta_f(T)$ is the contraction coefficient of T w.r.t. f-divergence). The above identity is simplified in the following Lemma, using the symmetric nature of the objects involved:

Lemma 3.12. *For any channel* $T = \begin{bmatrix} t & 1-t \\ 1-t & t \end{bmatrix} \in \mathcal{P}_{\text{symm}}$ *and* $f \in \mathcal{F}_{\text{symm}}$*, we have*

$$\eta_f^{\text{symm}}(T) = \sup_{e \in (0,1)} \frac{s(t,e) \cdot f\left(\frac{1-s(t,e)}{s(t,e)}\right)}{e \cdot f\left(\frac{1-e}{e}\right)}, \tag{3.33}$$

where $s(t,e) := te+(1-t)(1-e)$.

Proof. For any binary symmetric channel $A = \begin{bmatrix} a & 1-a \\ 1-a & a \end{bmatrix}$ (with $a \in (0,1)$) and $f \in \mathcal{F}_{\text{symm}}$, we have

$$\begin{aligned}
\mathbb{I}_f(A) &= \mathbb{I}_f\left(\begin{bmatrix} a \\ 1-a \end{bmatrix}, \begin{bmatrix} 1-a \\ a \end{bmatrix}\right) \\
&= \int f\left(\frac{dP}{dQ}\right) dQ \\
&= f\left(\frac{a}{1-a}\right) \cdot (1-a) + f\left(\frac{1-a}{a}\right) \cdot a \\
&= (1-a) \cdot \frac{a}{1-a} \cdot f\left(\frac{1-a}{a}\right) + a \cdot f\left(\frac{1-a}{a}\right) \\
&= 2a \cdot f\left(\frac{1-a}{a}\right),
\end{aligned}$$

where the fourth equality is due to $f(t) = f^{\diamond}(t)$. Thus for any symmetric channel $T \in \mathcal{P}_{\text{symm}}$ and symmetric experiment $E \in \mathcal{P}_{\text{symm}}$, we get

$$\begin{aligned} \mathbb{I}_f(E) &= 2e \cdot f\left(\frac{1-e}{e}\right) \\ \mathbb{I}_f(T \circ E) &= 2s \cdot f\left(\frac{1-s}{s}\right), \end{aligned}$$

where $s = te + (1-t)(1-e)$. Thus the proof is completed by taking the ratio between the above two divergences. □

We can better understand $\eta_f^{\text{symm}}(T)$ by defining the following functions (for $e, t \in (0,1)$)

$$F_f(e) := e \cdot f\left(\frac{1-e}{e}\right) \tag{3.34}$$

$$g_f(t,e) := \frac{F_f(s(t,e))}{F_f(e)} = \frac{s(t,e) \cdot f\left(\frac{1-s(t,e)}{s(t,e)}\right)}{e \cdot f\left(\frac{1-e}{e}\right)}. \tag{3.35}$$

Therefore $\eta_f^{\text{symm}}(T)$ can be compactly written as follows (from Lemma 3.12)

$$\eta_f^{\text{symm}}(T) = \sup_{e \in (0,1)} g_f(t,e). \tag{3.36}$$

We attempt to characterize $\eta_f^{\text{symm}}(T)$, by studying the behavior of $g_f(t,e)$. First we note the symmetric nature of $F_f(e)$ in the following lemma:

Lemma 3.13. *Let $f \in \mathcal{F}_{\text{symm}}$. Then F_f defined in* (3.34) *is convex, non-negative, and symmetric about $\frac{1}{2}$ with $F_f\left(\frac{1}{2}\right) = 0$.*

Proof. First we show that $f(x) \geq 0, \forall x \in (0,\infty)$ by using the facts that $f(1) = 0$ and $f(x) = f^{\diamond}(x) = xf\left(\frac{1}{x}\right)$. Observe that showing $f(x) \geq 0, \forall x \in (0,1)$ is sufficient. Suppose that $\exists x \in (0,1)$ s.t. $f(x) < 0$. Then for $x' = \frac{1}{x} \in (1,\infty)$, we have $f(x') = f^{\diamond}(x') = x' f\left(\frac{1}{x'}\right) < 0$. But $f(1) = 0$ and f is convex. This is a contradiction. Thus $f(x) \geq 0, \forall x \in (0,1)$.

Consider

$$F_f\left(\frac{1}{2}+\epsilon\right) = \left(\frac{1}{2}+\epsilon\right) \cdot f\left(\frac{1-\left(\frac{1}{2}+\epsilon\right)}{\frac{1}{2}+\epsilon}\right) = \left(\frac{1}{2}+\epsilon\right) \cdot f\left(\frac{\frac{1}{2}-\epsilon}{\frac{1}{2}+\epsilon}\right)$$

and

$$F_f\left(\frac{1}{2}-\epsilon\right) = \left(\frac{1}{2}-\epsilon\right) \cdot f\left(\frac{1-\left(\frac{1}{2}-\epsilon\right)}{\frac{1}{2}-\epsilon}\right) = \left(\frac{1}{2}-\epsilon\right) \cdot f\left(\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}\right).$$

Then using the property $f(x) = xf\left(\frac{1}{x}\right)$, one can easily see that $F_f\left(\frac{1}{2}+\epsilon\right) = F_f\left(\frac{1}{2}-\epsilon\right)$. Thus $F_f(x)$ is even symmetric about $\frac{1}{2}$.

- $F_f\left(\frac{1}{2}\right) = 0$ since $f(1) = 0$.
- $F_f(x) \geq 0, \forall x \in (0,1)$ since $f(x) \geq 0, \forall x \in (0,\infty)$.
- $F_f(x)$ is convex because it is a perspective transform of f which is convex.

□

Note that for any $f, g \in \mathcal{F}_{\text{symm}}$ and $\alpha, \beta \in \mathbb{R}$, we have $F_{\alpha f+\beta g}(x) = \alpha F_f(x) + \beta F_g(x)$. By using the symmetric nature of $F_f(\cdot)$, we can further simplify the identity $\eta_f^{\text{symm}}(T)$. Since $\eta_f^{\text{symm}}\left(\begin{bmatrix} t & 1-t \\ 1-t & t \end{bmatrix}\right) = \eta_f^{\text{symm}}\left(\begin{bmatrix} 1-t & t \\ t & 1-t \end{bmatrix}\right)$, hereafter we assume $t \geq \frac{1}{2}$ without loss of generality.

Lemma 3.14. *For a given fixed channel* $T = \begin{bmatrix} t & 1-t \\ 1-t & t \end{bmatrix} \in \mathcal{P}_{\text{symm}}$ *with* $t \geq \frac{1}{2}$ *(wlog) and* $f \in \mathcal{F}_{\text{symm}}$*, define*

$$\phi_f^t(\epsilon) \;:=\; \frac{F_f\left(\frac{1}{2}+c_t\epsilon\right)}{F_f\left(\frac{1}{2}+\epsilon\right)}, \tag{3.37}$$

where $c_t = 2t-1 \in [0,1)$ *and* $\epsilon \in \left(-\frac{1}{2}, \frac{1}{2}\right)$*. Then we have*

$$\eta_f^{\text{symm}}(T) \;=\; \sup_{\epsilon\in[0,1/2)} \phi_f^t(\epsilon). \tag{3.38}$$

Proof. Let $e = \frac{1}{2} + \epsilon$ where $\epsilon \in \left(-\frac{1}{2}, \frac{1}{2}\right)$. Then we have

$$\begin{aligned} s(t,e) \;&=\; \frac{1}{2} + \epsilon(2t-1) = \frac{1}{2} + c_t\epsilon, \text{ where } c_t = 2t-1 \in [0,1) \\ g_f(t,e) \;&=\; \frac{F_f(s(t,e))}{F_f(e)} = \frac{F_f\left(\frac{1}{2}+c_t\epsilon\right)}{F_f\left(\frac{1}{2}+\epsilon\right)} = \phi_f^t(\epsilon). \end{aligned}$$

Observe that $\phi_f^t(\epsilon)$ is symmetric about 0 since $F_f(\cdot)$ is symmetric about $\frac{1}{2}$. Then by using (3.36) we get

$$\eta_f^{\text{symm}}(T) = \sup_{e\in(0,1)} g_f(t,e) = \sup_{\epsilon\in(-1/2,1/2)} \phi_f^t(\epsilon) = \sup_{\epsilon\in[0,1/2)} \phi_f^t(\epsilon).$$

□

Let $L_f(\epsilon) := F_f(1/2+\epsilon)$. Then for fixed $c_t \in [0,1)$ we have

$$\phi_f^t(\epsilon) = \frac{L_f(c_t\epsilon)}{L_f(\epsilon)}, \text{ where } \epsilon \in [0,1/2).$$

Note that $L_f(0) = 0, L_f(\cdot) \geq 0$ and L_f is convex (for $f \in \mathcal{F}_{\text{symm}}$). Since we want to study the behavior of $\phi_f^t(\epsilon)$, we consider the derivative of it

$$\left(\phi_f^t\right)'(\epsilon) = \frac{\partial}{\partial \epsilon}\phi_f^t(\epsilon) = \frac{c_t L_f'(c_t\epsilon) L_f(\epsilon) - L_f(c_t\epsilon) L_f'(\epsilon)}{L_f(\epsilon)^2}.$$

Based on this we can observe two important behavior patterns of $\phi_f^t(\epsilon)$:

1. If $\left(\phi_f^t\right)'(\epsilon) \leq 0$, $\forall \epsilon \in (0, 1/2)$, then $\phi_f^t(\epsilon)$ is maximized at $\epsilon \to 0$, minimized at $\epsilon \to 1/2$. That is
$$\lim_{\epsilon \to 1/2} \phi_f^t(\epsilon) \leq \phi_f^t(\epsilon) \leq \lim_{\epsilon \to 0} \phi_f^t(\epsilon)$$
which is equivalent to
$$\lim_{e \to 1} g_f(t,e) \leq g_f(t,e) \leq \lim_{e \to 1/2} g_f(t,e). \tag{3.39}$$

2. If $\left(\phi_f^t\right)'(\epsilon) = 0$, $\forall \epsilon \geq 0$, then $\phi_f^t(\epsilon)$ is equal for all $\epsilon \in (0, 1/2)$. That is
$$\lim_{\epsilon \to 1/2} \phi_f^t(\epsilon) = \phi_f^t(\epsilon) = \lim_{\epsilon \to 0} \phi_f^t(\epsilon)$$
which is equivalent to
$$\lim_{e \to 1} g_f(t,e) = g_f(t,e) = \lim_{e \to 1/2} g_f(t,e). \tag{3.40}$$

For the above two cases we have

$$\eta_f^{\text{symm}}(T) = \sup_{e \in (0,1)} g_f(t,e) = \lim_{e \to 1/2} g_f(t,e).$$

Note that $g_f(t, 1/2)$ is not well defined. But for the second case above, where $\left(\phi_f^t\right)'(\epsilon) = 0, \forall \epsilon \in (0, 1/2)$, we can obtain an efficiently computable closed form for $\eta_f^{\text{symm}}(T)$. The following proposition characterizes the subclass of $\mathcal{F}_{\text{symm}}$ which satisfies this condition.

Proposition 3.15. *Define* $h_\alpha(x) := \frac{|1-x|^\alpha}{(1+x)^{\alpha-1}}$, *for* $x \in (0,1]$ *and* $\alpha \in \mathbb{R}$. *Then*

$$\mathcal{F}_{\text{symm}}^* := \Big\{ f : (0,\infty) \to \mathbb{R} : \forall x \in (0,1], f(x) = K \cdot h_{\alpha_f}(x) \text{ for some } K > 0, \alpha_f \geq 1, \\ \text{and } \forall x \in [1,\infty), f(x) = f^\diamond(x) \Big\} \subseteq \mathcal{F}_{\text{symm}}.$$

For any $T = \begin{bmatrix} t & 1-t \\ 1-t & t \end{bmatrix} \in \mathcal{P}_{\text{symm}}$ *(with* $t \geq \frac{1}{2}$*) and* $f \in \mathcal{F}_{\text{symm}}^*$, *we get*

$$\eta_f^{\text{symm}}(T) \;=\; \lim_{e \to 1} g_f(t,e) \;=\; (2t-1)^{\alpha_f}.$$

Proof. For any $f \in \mathcal{F}_{\text{symm}}^*$, we have $f(1) = 0$, $f(x) = f^\diamond(x)$, and f is convex (since

h_α is convex for $\alpha \geq 1$). Thus $\mathcal{F}^*_{\text{symm}} \subseteq \mathcal{F}_{\text{symm}}$.

If

$$\frac{c_t L'_f(c_t\epsilon)}{L_f(c_t\epsilon)} = \frac{L'_f(\epsilon)}{L_f(\epsilon)},$$

then $\left(\phi_f^t\right)'(\epsilon) = 0$, $\forall \epsilon \in (0, 1/2)$ (thus $\eta_f^{\text{symm}}(T) = \lim_{e\to 1} g_f(t,e)$). By letting $\psi = \log L_f$, the above condition can be written as follows,

$$c_t \psi'(c_t\epsilon) = \psi'(\epsilon)$$

that is we require ψ' to be (-1)-homogeneous. For a function $\psi'(x) = \alpha x^{-1}$ which is (-1)-homogeneous, we have (for some constant $K > 0$)

$$\begin{aligned}
\psi'(x) &= \alpha\frac{1}{x},\ x \geq 0 \text{ (to enforce symmetry)} \\
\Longleftarrow \psi(x) &= \alpha \log x + \log K = \log L_f(x) \\
\Longleftarrow L_f(x) &= Kx^\alpha = F_f(1/2 + x) \\
\Longleftarrow F_f(y) &= K(y - 1/2)^\alpha = yf\left(\frac{1-y}{y}\right), \text{ where } y = 1/2 + x \geq 1/2 \\
\Longleftarrow f(z) &= K \cdot \frac{(1-z)^\alpha}{(1+z)^{\alpha-1}}, \text{ where } z = \frac{1-y}{y} \leq 1.
\end{aligned}$$

That is for any $f \in \mathcal{F}^*_{\text{symm}}$, we have $\left(\phi_f^t\right)'(\epsilon) = 0$, $\forall \epsilon \in (0, 1/2)$. Thus for any $f \in \mathcal{F}^*_{\text{symm}}$, we get

$$\eta_f^{\text{symm}}(T) = \lim_{e\to 1} g_f(t,e) = \frac{F_f(t)}{\lim_{e\to 1} F_f(e)} = \frac{t \cdot h_{\alpha_f}\left(\frac{1-t}{t}\right)}{\lim_{x\to 0} h_{\alpha_f}(x)} = \frac{(2t-1)^{\alpha_f}}{1}.$$

□

Note that $f_{\text{tv}}, f_{\text{tri}} \in \mathcal{F}^*_{\text{symm}}$ with $\alpha_{f_{\text{tv}}} = 1$ and $\alpha_{f_{\text{tri}}} = 2$ (recall that $f_{\text{tv}}(t) = |t-1|$, and $f_{\text{tri}}(t) = \frac{(t-1)^2}{t+1}$). Thus from the above proposition and (3.40), for any $T = \begin{bmatrix} t & 1-t \\ 1-t & t \end{bmatrix} \in \mathcal{P}_{\text{symm}}$ (with $t \geq 1/2$), we have

$$\eta_{f_{\text{tv}}}^{\text{symm}}(T) = \lim_{e\to 1/2} g_{f_{\text{tv}}}(t,e) = 2t - 1$$

and

$$\eta_{f_{\text{tri}}}^{\text{symm}}(T) = \lim_{e\to 1/2} g_{f_{\text{tri}}}(t,e) = (2t-1)^2.$$

$g_{f_{\text{tv}}}(t,e)$ and $g_{f_{\text{tri}}}(t,e)$ are shown in Figures 3.2 and 3.3 respectively.

For $f_{\mathrm{He}}(t) = \left(\sqrt{t}-1\right)^2$, we have

$$\begin{aligned}
F_{f_{\mathrm{He}}}(e) &= 1 - 2\sqrt{e\cdot(1-e)}, \quad e \in (0,1) \\
L_{f_{\mathrm{He}}}(\epsilon) &= 1 - \sqrt{1-4\epsilon^2}, \quad \epsilon \in [0,1/2) \\
\phi^t_{f_{\mathrm{He}}}(\epsilon) &= \frac{1-\sqrt{1-4c_t^2\epsilon^2}}{1-\sqrt{1-4\epsilon^2}}, \quad c_t = 2t-1 \in (0,1) \\
\lim_{\epsilon\to 0}\phi^t_{f_{\mathrm{He}}}(\epsilon) &= c_t^2 \\
\lim_{\epsilon\to 1/2}\phi^t_{f_{\mathrm{He}}}(\epsilon) &= 1-\sqrt{1-c_t^2}.
\end{aligned}$$

By using simple calculations, one can easily verify that (see Figure 3.4)

$$\lim_{\epsilon\to 1/2}\phi^t_{f_{\mathrm{He}}}(\epsilon) \le \phi^t_{f_{\mathrm{He}}}(\epsilon) \le \lim_{\epsilon\to 0}\phi^t_{f_{\mathrm{He}}}(\epsilon) = (2t-1)^2 = \eta^{\mathrm{symm}}_{f_{\mathrm{He}}}(T).$$

We observed that f_{tvtri}, f_{tvHe}, and f_{triHe} also satisfy (3.39) (see Figures 3.5, 3.6, and 3.7):

$$\begin{aligned}
\lim_{e\to 1} g_{f_{\mathrm{tvtri}}}(t,e) \le g_{f_{\mathrm{tvtri}}}(t,e) \le \lim_{e\to 1/2} g_{f_{\mathrm{tvtri}}}(t,e) &= \eta^{\mathrm{symm}}_{f_{\mathrm{tvtri}}}(T) = 2t-1 \\
\lim_{e\to 1} g_{f_{\mathrm{tvHe}}}(t,e) \le g_{f_{\mathrm{tvHe}}}(t,e) \le \lim_{e\to 1/2} g_{f_{\mathrm{tvHe}}}(t,e) &= \eta^{\mathrm{symm}}_{f_{\mathrm{tvHe}}}(T) = 2t-1 \\
\lim_{e\to 1} g_{f_{\mathrm{triHe}}}(t,e) \le g_{f_{\mathrm{triHe}}}(t,e) \le \lim_{e\to 1/2} g_{f_{\mathrm{triHe}}}(t,e) &= \eta^{\mathrm{symm}}_{f_{\mathrm{triHe}}}(T) = (2t-1)^2.
\end{aligned}$$

Thus for all binary symmetric channels, and certain subset of symmetric f-divergences, we are able to obtain a lower bound (of the form $\eta^{\mathrm{symm}}_f(T) \le \eta_f(T)$) on the contraction coefficients. At this stage, we point out the following possible extensions for the above exercise (some of them will follow through the above approach to certain level):

- relax the symmetric experiments restriction in $\eta^{\mathrm{symm}}_f(T)$, to obtain $\eta_f(T)$ for $f \in \mathcal{F}_{\mathrm{symm}}$ and $T \in \mathcal{P}_{\mathrm{symm}}$.
- extend the study of $\eta^{\mathrm{symm}}_f(T)$ to k-ary symmetric channels and experiments with $k > 2$.
- extend the study of $\eta^{\mathrm{symm}}_f(T)$ to non-symmetric f.

3.3.3 Hardness of Constrained Learning Problem

We now use the strong data processing inequalities and minimax lower bound techniques to analyse the hardness of the constrained learning problem that we introduced in the beginning of this section. First we generalize Le Cam's (Proposition 3.2) and Assouad's (Theorem 3.4 and Corollary 3.5) results for the constrained parameter estimation problem ((3.25) with $\Theta = \mathcal{A}$, $A = \widehat{\theta}$, and $\ell = \rho$). Most of these generalizations follows directly from the original versions, thus don't require any proof.

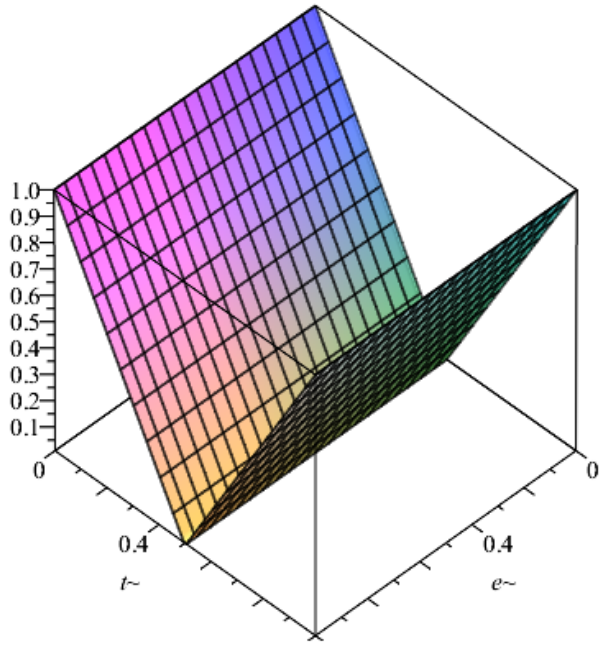


Figure 3.2: $g_{f_{\text{tv}}}(t, e)$ of a binary symmetric channel w.r.t. total variation divergence.

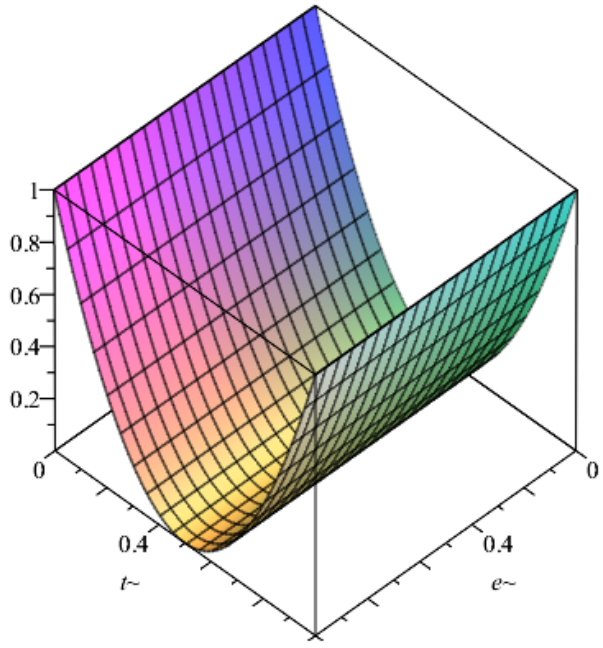


Figure 3.3: $g_{f_{\text{tri}}}(t, e)$ of a binary symmetric channel w.r.t. triangular discrimination divergence.

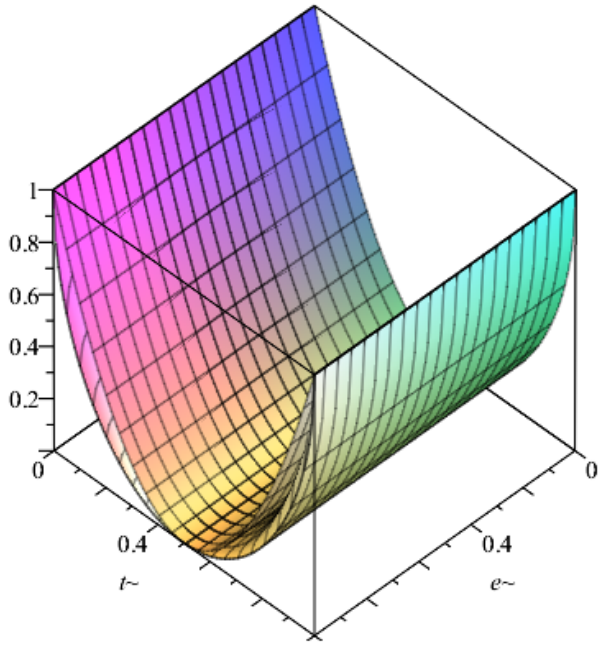


Figure 3.4: $g_{f_{\mathrm{He}}}(t, e)$ of a binary symmetric channel w.r.t. symmetric squared Hellinger divergence (sandwiched according to (3.39)).

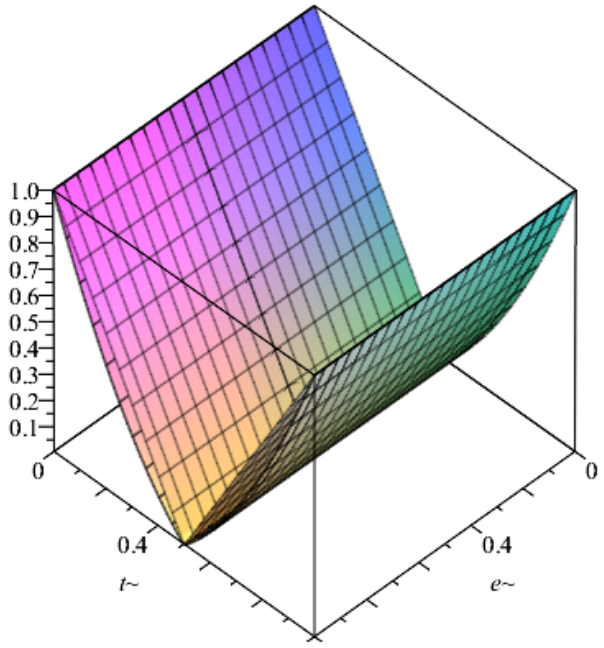


Figure 3.5: $g_{f_{\mathrm{tvtri}}}(t, e)$ of a binary symmetric channel w.r.t. $\mathbb{I}_{f_{\mathrm{tvtri}}}$ (sandwiched according to (3.39)).

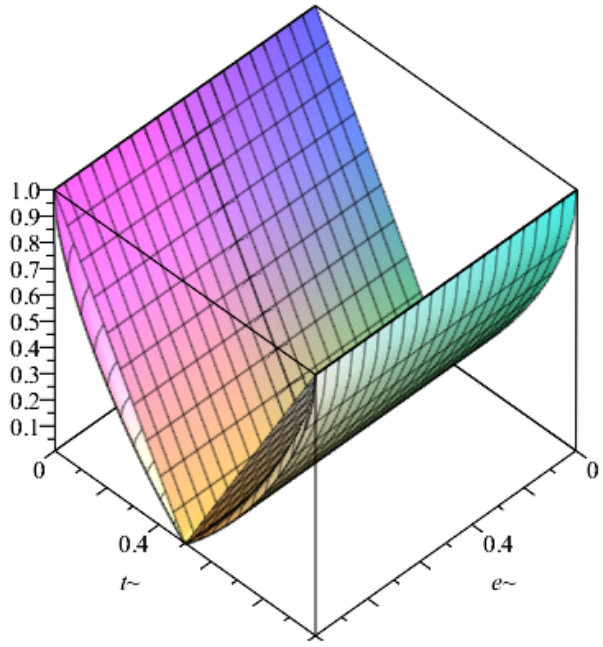


Figure 3.6: $g_{f_{\text{tvHe}}}(t, e)$ of a binary symmetric channel w.r.t. $\mathbb{I}_{f_{\text{tvHe}}}$ (sandwiched according to (3.39)).

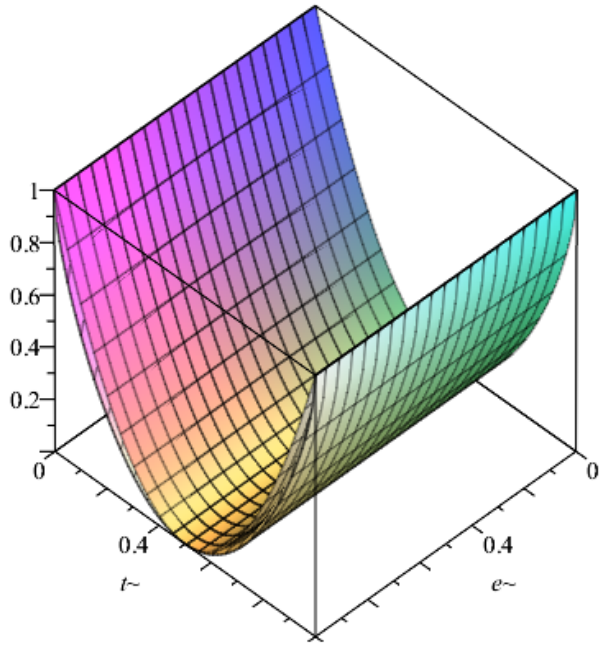


Figure 3.7: $g_{f_{\text{triHe}}}(t, e)$ of a binary symmetric channel w.r.t. $\mathbb{I}_{f_{\text{triHe}}}$ (sandwiched according to (3.39)).

Proposition 3.16. *For any $c \in (0,1)$, the minimax risk of the constrained parameter estimation problem (*(3.25) *with $\Theta = \mathcal{A}$, $A = \widehat{\theta}$, and $\ell = \rho$) is lower bounded as*

$$\underline{R}^{\star}_{\rho}\left(\tilde{\varepsilon}_n\right) \;\geq\; \sup_{\theta \neq \theta'} \left\{ \rho\left(\theta, \theta'\right) \cdot \left[\frac{1}{2} - n\left(\frac{1}{2} - c \wedge \overline{c} \right) - \vartheta_c\left(T\right) n \cdot \mathbb{I}_{f_c}\left(\varepsilon\left(\theta\right), \varepsilon\left(\theta'\right)\right) \right] \right\}.$$

Proof. From Proposition 3.2, we have that

$$\underline{R}^{\star}_{\rho}\left(\tilde{\varepsilon}_n\right) \;\geq\; \sup_{\theta \neq \theta'} \left\{ \rho\left(\theta, \theta'\right) \cdot \left(c \wedge \overline{c} - \cdot \mathbb{I}_{f_c}\left(\tilde{\varepsilon}_n\left(\theta\right), \tilde{\varepsilon}_n\left(\theta'\right)\right)\right)\right\}.$$

Further from Lemma 2.11, we have that

$$d_{\mathrm{TV}}\left(c\tilde{\varepsilon}_n\left(\theta\right), \overline{c}\tilde{\varepsilon}_n\left(\theta'\right)\right) \;\leq\; n d_{\mathrm{TV}}\left(c\tilde{\varepsilon}\left(\theta\right), \overline{c}\tilde{\varepsilon}\left(\theta'\right)\right).$$

Thus we have

$$\begin{aligned} \mathbb{I}_{f_c}\left(\tilde{\varepsilon}_n\left(\theta\right), \tilde{\varepsilon}_n\left(\theta'\right)\right) &\leq n \mathbb{I}_{f_c}\left(\tilde{\varepsilon}\left(\theta\right), \tilde{\varepsilon}\left(\theta'\right)\right) + \left(c \wedge \overline{c} - \frac{1}{2} \right) \cdot \left(1 - n\right) \\ &\leq \vartheta_c\left(T\right) n \mathbb{I}_{f_c}\left(\tilde{\varepsilon}\left(\theta\right), \tilde{\varepsilon}\left(\theta'\right)\right) + \left(c \wedge \overline{c} - \frac{1}{2} \right) \cdot \left(1 - n\right). \end{aligned}$$

□

Theorem 3.17. *Let $\Theta = \{-1,1\}^d$ and $\rho = \rho_{\mathrm{Ha}}$ (defined in* (2.1)*). Then for any $c \in (0,1)$, the minimax risk of the constrained parameter estimation problem (*(3.25) *with $\Theta = \mathcal{A}$, $A = \widehat{\theta}$, and $\ell = \rho$) is lower bounded as*

$$\underline{R}^{\star}_{\rho_{\mathrm{Ha}}}\left(\tilde{\varepsilon}_n\right) \;\geq\; d \left(c \wedge \overline{c} - \max_{\theta, \theta' : \rho_{\mathrm{Ha}}(\theta, \theta') = 1} \mathbb{I}_{f_c}\left(\tilde{\varepsilon}_n\left(\theta\right), \tilde{\varepsilon}_n\left(\theta'\right)\right) \right)$$

Corollary 3.18. *Let $\mathcal{O}$ be some set and $c \in (0,1)$. Define*

$$\mathcal{P}_{\varepsilon}\left(\mathcal{O}\right) := \left\{ \varepsilon\left(\theta\right) \in \mathcal{P}\left(\mathcal{O}\right) : \theta \in \{-1,1\}^d \right\}$$

be a class of probability measures induced by the transition $\varepsilon : \{-1,1\}^d \rightsquigarrow \mathcal{O}$. Suppose that there exists some cost-dependent constant $\alpha\left(c\right) > 0$, such that

$$\mathrm{He}^2\left(\varepsilon\left(\theta\right), \varepsilon\left(\theta'\right)\right) \leq \alpha\left(c\right), \qquad \text{if } \rho_{\mathrm{Ha}}\left(\theta, \theta'\right) = 1.$$

The minimax risk of the constrained parameter estimation problem ((3.25) *with $\Theta = \mathcal{A}$, $A = \widehat{\theta}$, and $\ell = \rho_{\mathrm{Ha}}$) is lower bounded as*

$$\underline{\mathcal{R}}^{\star}_{\rho_{\mathrm{Ha}}}\left(\tilde{\varepsilon}_n\right) \;\geq\; d \cdot \left(c \wedge \overline{c}\right) \cdot \left(1 - \sqrt{\alpha\left(c\right) \eta_{\mathrm{He}^2}\left(T\right) n} \right), \tag{3.41}$$

where T is as per (3.25) *and $\eta_{\mathrm{He}^2}\left(T\right)$ is the contraction coefficient of T w.r.t. squared Hellinger distance.*

Proof. For any two $\theta, \theta' \in \Theta$ with $\rho_{\mathrm{Ha}}(\theta, \theta') = 1$, we have

$$\begin{aligned}
\mathbb{I}_{f_c}\left(\tilde{\varepsilon}_n(\theta), \tilde{\varepsilon}_n(\theta')\right) &\leq (c \wedge \bar{c}) \cdot \mathrm{He}\left(\tilde{\varepsilon}_n(\theta), \tilde{\varepsilon}_n(\theta')\right) \\
&\leq (c \wedge \bar{c}) \cdot \sqrt{\sum_{i=1}^{n} \mathrm{He}^2\left(\tilde{\varepsilon}(\theta), \tilde{\varepsilon}(\theta')\right)} \\
&\leq (c \wedge \bar{c}) \cdot \sqrt{\sum_{i=1}^{n} \eta_{\mathrm{He}^2}(T)\, \mathrm{He}^2\left(\varepsilon(\theta), \varepsilon(\theta')\right)} \\
&\leq (c \wedge \bar{c}) \cdot \sqrt{\alpha(c)\, \eta_{\mathrm{He}^2}(T)\, n}
\end{aligned}$$

□

Consider the corrupted cost-sensitive binary classification problem represented by the following transition diagram:

$$\Theta_{h,\mathcal{F}} \overset{\varepsilon_n}{\rightsquigarrow} (\mathcal{X} \times \{-1,1\})^n \overset{T_{1:n}}{\rightsquigarrow} (\mathcal{X} \times \{-1,1\})^n \overset{\widehat{f}}{\rightsquigarrow} \mathcal{F}, \tag{3.42}$$

and the minimax risk of it given by

$$\underline{\mathcal{R}}^{\star}_{\Delta\ell_{d_c}}(\tilde{\varepsilon}_n) := \inf_{\widehat{f}} \sup_{\theta \in \Theta_{h,\mathcal{F}}} \mathop{\mathbb{E}}_{\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n \sim \tilde{\varepsilon}_n(\theta)} \left[\mathop{\mathbb{E}}_{f \sim \widehat{f}\left(\{(\mathsf{X}_i,\mathsf{Y}_i)\}_{i=1}^n\right)} \left[\Delta\ell_{d_c}(\theta, f)\right] \right]. \tag{3.43}$$

Theorem 3.19. *Let $\mathcal{F}$ be a VC class of binary-valued functions on $\mathcal{X}$ with VC dimension $V \geq 2$. Then for any $n \geq V$ and any $h \in [0, c \wedge \bar{c}]$, the minimax risk* (3.43) *of the corrupted cost-sensitive binary classification* (3.42) *is lower bounded as follows:*

$$\underline{\mathcal{R}}^{\star}_{\Delta\ell_{d_c}}(\tilde{\varepsilon}_n) \;\geq\; K \cdot (c \wedge \bar{c}) \cdot \min\left(\sqrt{\frac{(c \wedge \bar{c})V}{\eta_{\mathrm{He}^2}(T)\, n}}, (c \wedge \bar{c}) \cdot \frac{V}{\eta_{\mathrm{He}^2}(T)\, nh} \right)$$

where $K > 0$ is some absolute constant.

The number of samples that appear in the minimax lower bound of the original learning problem is scaled by the contraction coefficient $\eta_f(T)$ in the case of corrupted learning problem. Hence the rate is unaffected, only the constants. However, a penalty of factor $\eta_f(T)$ is unavoidable no matter what learning algorithm is used, suggesting that $\eta_f(T)$ is a valid way of measuring the amount of corruption.

3.4 Cost-sensitive Privacy Notions

Suppose a trustworthy data curator gathers sensitive data from a large number of data providers, with the goal of learning statistical facts about the underlying population.

A data analyst makes a statistical query on the sensitive dataset from the data curator. Thus the main challenge for the data curator is to send back a randomized response such that the utility of the task of the data analyst is increased while maintaining the privacy of the data providers. This requires a formal definition of privacy, and differential privacy has been put forth as such formalization (Dwork et al. [2006]). Differential privacy requires that the data analyst knows no more about any individual in the sensitive dataset after the analysis is completed, than she knew before the analysis was begun. That is the impact on the data provider is the same independent of whether or not he was in the analysis. It is possible to reduce the problem of enforcing differential privacy to a statistical decision problem (Wasserman and Zhou [2010]). We exploit this observation and extend it further (see section 3.4.2).

In a more restrictive requirement than the differential privacy, called "local privacy" ([Duchi et al., 2013; Warner, 1965]), the data providers don't even trust the data curator collecting the data. When the sensitive data to be protected is other than the value of a single individual, it is common to consider different definitions for privacy requirements.

A *privacy mechanism* is an algorithm that takes as input a database, a universe $\mathcal{V}$ of data types (of the database), and optionally a set of queries, and produce a randomized response. The privacy mechanism can be represented by a transition $T : \mathcal{O} \rightsquigarrow \widehat{\mathcal{O}}$. Below we represent some of the privacy enforced settings via transition diagrams:

- We need to protect an *abstract set of secrets* $\mathcal{X}$ (for example geographical locations of army base points) from the data analyst, who wants to learn some summary statistic about the probability distribution which generated the secrets i.e. something about the actual parameter θ from the parameter space Θ. This setting is represented by the following transition diagram

$$\Theta \overset{\varepsilon}{\rightsquigarrow} \mathcal{X} \overset{T}{\rightsquigarrow} \mathcal{Z} \quad , \tag{3.44}$$

  where T is the privacy mechanism and $\mathcal{Z}$ is the new outcome space observed by the data analyst. When the outcome space $\mathcal{Z} = \mathcal{X}$, the resulting transition diagram is

$$\Theta \overset{\varepsilon}{\rightsquigarrow} \mathcal{X} \overset{T}{\rightsquigarrow} \mathcal{X} \quad . \tag{3.45}$$

- We need to protect the *entries of the database* by releasing a *sanitized database* (this approach is also referred to as non-interactive method). Let the database universe be $\mathcal{V}$. Then by repeatedly applying the privacy mechanism T in the transition diagram (3.45) with $\mathcal{X} = \mathcal{V}$, over all the entries of the database, we get

$$\Theta \overset{\varepsilon_n}{\rightsquigarrow} \mathcal{V}^n \overset{T_{1:n}}{\rightsquigarrow} \mathcal{V}^n \quad . \tag{3.46}$$

- In comparison to the above non-interactive approach, it is possible to protect the entries of the database by corrupting the response for the database query

appropriately. This interactive (query dependent) method can be represented by the following transition digram

$$\Theta \overset{\varepsilon}{\rightsquigarrow} \mathcal{X} \xrightarrow{f} \mathcal{Y} \overset{H}{\rightsquigarrow} \mathcal{Z} \quad , \tag{3.47}$$

where $\mathcal{X} = \mathcal{V}^n$ is the database (with universe $\mathcal{V}$) to be protected, f is a query on the database, and H is the privacy mechanism over the query outcome space $\mathcal{Y}$. We need to enforce restrictions on the composite mechanism $T = H \circ f$ in order to protect the elements of $\mathcal{X}$. By appropriate tailoring, these restrictions can be reduced to the restrictions on the mechanism H depending on f.

Based on the discussions above, without loss of generality, we only consider the privacy definitions for the setting represented by the transition diagram (3.44), with finite $\mathcal{X}$ and $\mathcal{Z}$.

3.4.1 Symmetric Local Privacy

First we briefly review the *(symmetric) local privacy* notion which is well studied in the literature ([Dwork, 2008; Duchi et al., 2013]). Consider the setting represented by the transition diagram (3.44) with finite $\mathcal{X}$ and $\mathcal{Z}$. The (symmetric) local privacy imposes *indistinguishability* between pairs of secrets and protects all of them equally:

Definition 3.20 ([Duchi et al., 2013])**.** *Given $\epsilon > 0$, let $\mathcal{M}(\mathcal{X}, \mathcal{Z}; \epsilon) \subseteq \mathcal{M}(\mathcal{X}, \mathcal{Z})$ denote the set of all ϵ-locally private mechanisms where*

$$T \in \mathcal{M}(\mathcal{X}, \mathcal{Z}; \epsilon) \iff \frac{T(z \mid x_i)}{T(z \mid x_j)} \leq \mathrm{e}^{\epsilon}, \forall x_i, x_j \in \mathcal{X}, z \in \mathcal{Z}. \tag{3.48}$$

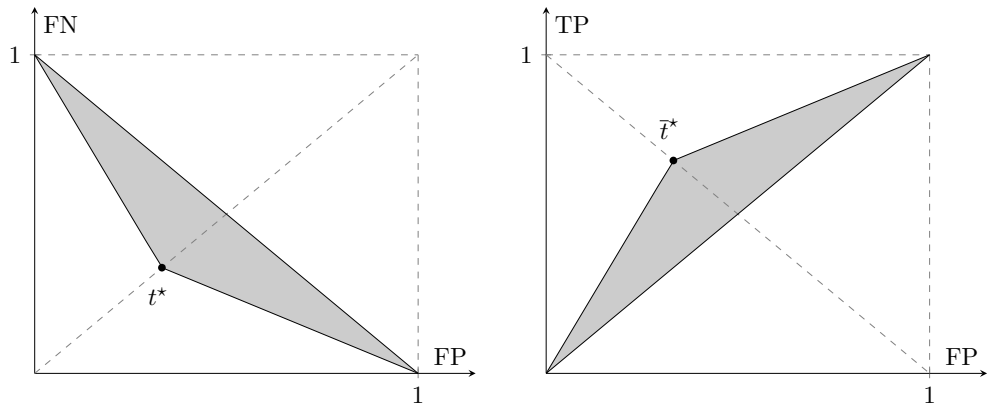
Below we provide a hypothesis testing based interpretation of the above definition, essentially noted by Wasserman and Zhou [2010].

Hypothesis Testing Interpretation: Based on the random outcome in $\mathcal{Z}$ from the privacy mechanism T, we want determine whether it is generated by the secret x_i or x_j. Let the labels 1 and 0 correspond to the probability measures $T(x_i)$ and $T(x_j)$ respectively. Consider a statistical test (recall from section 2.4.1) $r_{ij} : \mathcal{Z} \to \{0, 1\}$. Then the false negative and false positive rates of this test are given by

$$\mathrm{FN}_{r_{ij}} := \sum_{z \in \mathcal{Z}} T(z \mid x_i) [\![r_{ij}(z) = 0]\!], \text{ and} \tag{3.49}$$

$$\mathrm{FP}_{r_{ij}} := \sum_{z \in \mathcal{Z}} T(z \mid x_j) [\![r_{ij}(z) = 1]\!], \tag{3.50}$$

respectively. The ϵ-local privacy condition on a mechanism T is equivalent to the following set of constraints on the false negative and false positive rates:



(a) $t^\star = \left(\frac{1}{e^\epsilon+1}, \frac{1}{e^\epsilon+1}\right)$ (b) $\overline{t}^\star = \left(\frac{1}{e^\epsilon+1}, \frac{e^\epsilon}{e^\epsilon+1}\right)$

Figure 3.8: Operational characteristic representation of ϵ-local privacy mechanisms (with $\epsilon = 0.7$).

Theorem 3.21 ([Kairouz et al., 2014])**.** *For any $\epsilon > 0$, a mechanism $T \in \mathcal{M}(\mathcal{X}, \mathcal{Z})$ is ϵ-locally private if and only if the following conditions are satisfied for all $x_i, x_j \in \mathcal{X}$, and all statistical tests $r_{ij} : \mathcal{Z} \to \{0, 1\}$:*

$$\mathrm{FN}_{r_{ij}} + e^\epsilon \cdot \mathrm{FP}_{r_{ij}} \geq 1, \tag{3.51}$$

$$e^\epsilon \cdot \mathrm{FN}_{r_{ij}} + \mathrm{FP}_{r_{ij}} \geq 1. \tag{3.52}$$

The above operational interpretation says that it is impossible to get both small false negative and false positive rates from data obtained via a ϵ-locally private mechanism. The above characterization is graphically represented in Figure 3.8, where the shaded region of the left side diagram (Figure 3.8(a)) can be mathematically written as follows

$$\mathcal{S}(\epsilon) := \{(\mathrm{FP}, \mathrm{FN}) : \mathrm{FN} + e^\epsilon \cdot \mathrm{FP} \geq 1, \text{ and } e^\epsilon \cdot \mathrm{FN} + \mathrm{FP} \geq 1\}. \tag{3.53}$$

We define the *privacy region* of a mechanism T with respect to x_i and x_j as

$$\mathcal{S}(T, x_i, x_j) := \mathrm{conv}\left(\left\{\left(\mathrm{FP}_{r_{ij}}, \mathrm{FN}_{r_{ij}}\right) : \text{ for all } r_{ij} : \mathcal{Z} \to \{0, 1\}\right\}\right), \tag{3.54}$$

where $\mathrm{conv}(\cdot)$ is the convex hull of a set. The following corollary, which follows immediately from Theorem 3.21, gives a necessary and sufficient condition for a mechanism to be ϵ-locally private.

Corollary 3.22. *A mechanism T is ϵ-locally private if and only if $\mathcal{S}(T, x_i, x_j) \subseteq \mathcal{S}(\epsilon)$ for all $x_i, x_j \in \mathcal{X}$.*

3.4.2 Non-homogeneous Local Privacy

Now if we want to protect some secrets more than others we need to break the inherent symmetry in the privacy definition of the previous section. Here we introduce an

asymmetric privacy notion which is a simple extension of Chatzikokolakis et al. [2013]. We replace the undirected pairwise cost terms in the definition of Chatzikokolakis et al. [2013] with directed cost terms in order to enforce asymmetry.

Definition 3.23. *Define $n := |\mathcal{X}|$. Given $\mathbb{R}_+^{n\times n}$ matrix C (with $(i,j)^{th}$ entry given by the 'directed' cost $C_{ij} \in [0,1]$), let $\mathcal{M}(\mathcal{X},\mathcal{Z};C) \subseteq \mathcal{M}(\mathcal{X},\mathcal{Z})$ denote the set of all C-locally private mechanisms where*

$$T \in \mathcal{M}(\mathcal{X},\mathcal{Z};C) \iff \frac{T(z \mid x_i)}{T(z \mid x_j)} \leq \mathrm{e}^{C_{ij}},\ \forall x_i, x_j \in \mathcal{X}, z \in \mathcal{Z}.$$

When C is a symmetric matrix with 0's as the diagonal entries and ϵ's as the off-diagonal entries, we recover the usual ϵ-local privacy requirement.

Suppose we want to prioritize only $x_{i^\star}$'s privacy and treat others equally. In this case we can choose C be a symmetric matrix with 0's as the diagonal entries, $(c\cdot\epsilon)$'s (where $c \in [0,1]$) in the $i^\star$-th row (except the diagonal term), $(\overline{c}\cdot\epsilon)$'s in the $i^\star$-th column (except the diagonal term), and $(0.5\cdot\epsilon)$'s in other places:

$$\begin{bmatrix} 0 & 0.5 & \dots & \overline{c} & \dots & 0.5 \\ 0.5 & 0 & \dots & \overline{c} & \dots & 0.5 \\ \vdots & \vdots & & \vdots & & \vdots \\ c & c & \dots & 0 & \dots & c \\ \vdots & \vdots & & \vdots & & \vdots \\ 0.5 & 0.5 & \dots & \overline{c} & \dots & 0 \end{bmatrix} \cdot \epsilon.$$

Hypothesis Testing Interpretation: We extend the hypothesis testing based interpretation of the ϵ-local privacy definition, to this general case. Then the C-local privacy condition on a mechanism T is equivalent to the following set of constraints on the false negative and false positive rates:

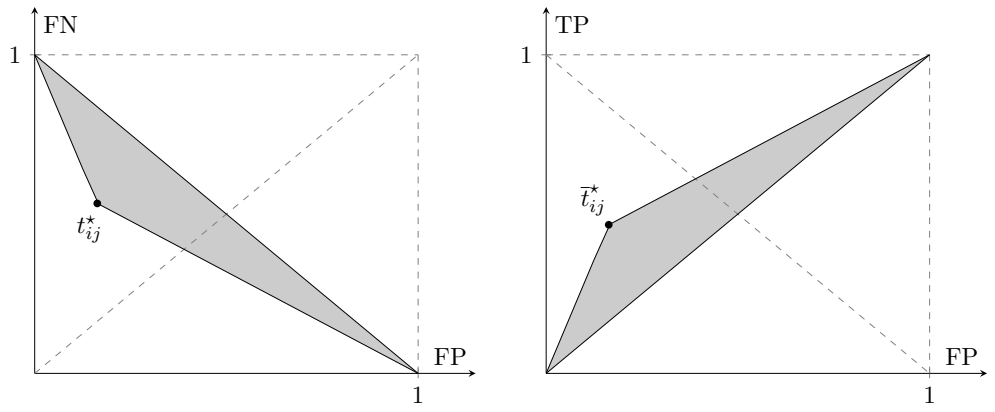
Theorem 3.24. *For any $C \in \mathbb{R}_+^{n\times n}$, a mechanism $T \in \mathcal{M}(\mathcal{X},\mathcal{Z})$ is C-locally private if and only if the following conditions are satisfied for all $x_i, x_j \in \mathcal{X}$, and all statistical tests $r_{ij} : \mathcal{Z} \to \{0,1\}$:*

$$\mathrm{FN}_{r_{ij}} + \mathrm{e}^{C_{ij}} \cdot \mathrm{FP}_{r_{ij}} \geq 1, \tag{3.55}$$

$$\mathrm{e}^{C_{ji}} \cdot \mathrm{FN}_{r_{ij}} + \mathrm{FP}_{r_{ij}} \geq 1. \tag{3.56}$$

Proof. From the definition of C-local privacy, for any statistical test $r_{ij} : \mathcal{Z} \to \{0,1\}$, we have

$$\begin{aligned} T(z \mid x_i) &\leq \mathrm{e}^{C_{ij}} \cdot T(z \mid x_j) \\ \implies \sum_{z\in\mathcal{Z}} T(z \mid x_i)\,[\![r_{ij}=1]\!] &\leq \mathrm{e}^{C_{ij}} \cdot \sum_{z\in\mathcal{Z}} T(z \mid x_j)\,[\![r_{ij}=1]\!] \\ \implies 1 - \mathrm{FN}_{r_{ij}} &\leq \mathrm{e}^{C_{ij}} \cdot \mathrm{FP}_{r_{ij}}, \end{aligned}$$



(a) $t^\star_{ij} = (\alpha, \beta)$ (b) $\overline{t}^\star_{ij} = (\alpha, 1-\beta)$

Figure 3.9: Operational characteristic representation of C-local privacy mechanisms (with $C_{ij} = 1.05$ and $C_{ji} = 0.45$). Note that $C_{ij} > C_{ji}$. Observe that $\alpha = \frac{\mathrm{e}^{C_{ji}}-1}{\mathrm{e}^{(C_{ij}+C_{ji})}-1}$ and $\beta = \frac{\mathrm{e}^{C_{ij}}-1}{\mathrm{e}^{(C_{ij}+C_{ji})}-1}$.

and

$$\begin{aligned} T(z \mid x_j) &\leq \mathrm{e}^{C_{ji}} \cdot T(z \mid x_i) \\ \implies \sum_{z \in \mathcal{Z}} T(z \mid x_j) [\![r_{ij} = 0]\!] &\leq \mathrm{e}^{C_{ij}} \cdot \sum_{z \in \mathcal{Z}} T(z \mid x_i) [\![r_{ij} = 0]\!] \\ \implies 1 - \mathrm{FP}_{r_{ij}} &\leq \mathrm{e}^{C_{ji}} \cdot \mathrm{FN}_{r_{ij}}. \end{aligned}$$

□

The above characterization is graphically represented in Figure 3.9, where the shaded region of the left side diagram (Figure 3.9(a)) can be mathematically written as follows

$$\mathcal{S}(C, x_i, x_j) := \left\{ (\mathrm{FP}, \mathrm{FN}) : \mathrm{FN} + \mathrm{e}^{C_{ij}} \cdot \mathrm{FP} \geq 1, \text{ and } \mathrm{e}^{C_{ji}} \cdot \mathrm{FN} + \mathrm{FP} \geq 1 \right\}. \quad (3.57)$$

Note that unlike Figure 3.8 which holds $\forall i \neq j$, here in general we get a different picture for each choice of i and j. The following corollary, which follows immediately from Theorem 3.24, gives a necessary and sufficient condition on the privacy region for C-local privacy.

Corollary 3.25. *A mechanism T is C-locally private if and only if $\mathcal{S}(T, x_i, x_j) \subseteq \mathcal{S}(C, x_i, x_j)$ for all $x_i, x_j \in \mathcal{X}$.*

To facilitate the mechanism design, we define the following

$$\mathcal{S}(C, x_i) := \bigcap_{x_j \in \mathcal{X} \setminus x_i} \mathcal{S}(C, x_i, x_j), \quad (3.58)$$

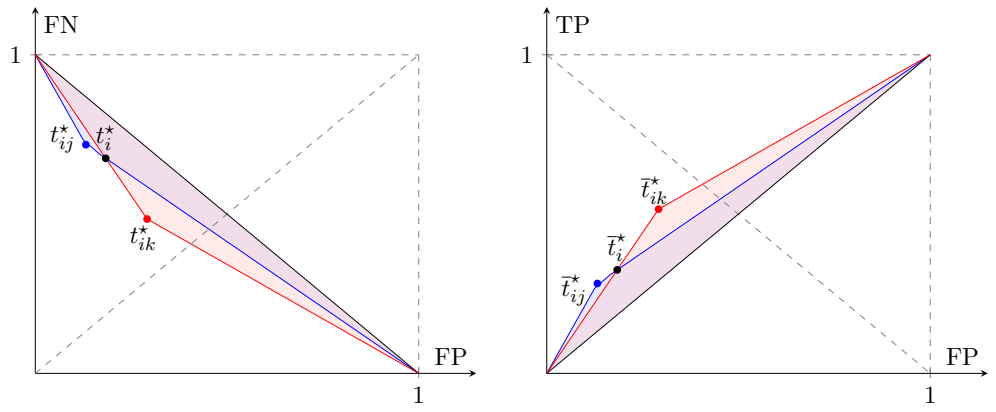


Figure 3.10: Feasible region for $T(\cdot \mid x_i)$ under C-local privacy. The point $t_i^\star$ will impact the optimal privacy mechanism's $T(\cdot \mid x_i)$.

using which we can design $T(\cdot \mid x_i)$. This is illustrated in Figure 3.10.

Figure 3.11 is plotted for two different C: one with $C_{ij} = 1.05$ and $C_{ji} = 0.45$, and the other with $C_{ij} = C_{ji} = 0.75$. This diagram shows how much we lose by prioritizing someone's privacy than others. It can be observed that the permissible ROC region for the privacy mechanism gets shrunk (compared to the equal privacy case) when we enforce prioritized privacy for someone.

3.5 Conclusion

The cost-sensitive classification problem plays a crucial role in mission critical machine learning applications. We have studied the hardness of this problem and emphasized the impact of cost terms on the hardness.

Strong data processing inequalities (SDPI) are very useful in analysing the hardness of constrained learning problems. Despite extensive investigation, the geometric insights of the SDPI are not fully understood. This chapter provides some direction. To this end, we have derived an explicit form for the contraction coefficient of any channel w.r.t. c-primitive f-divergence, and we have obtained efficiently computable lower bound for the contraction coefficient of any binary symmetric channel w.r.t. any symmetric f-divergence.

We pose the following open problems as future directions:

- There are some divergences other than f-divergences which satisfy the weak data processing inequality, such as Neyman-Pearson α-divergences ([Polyanskiy and Verdú, 2010; Raginsky, 2011]). Thus it would be interesting to study strong data processing inequalities w.r.t. those divergences as well.

- Recently people have attempted to relate several types of channel ordering to the strong data processing inequalities ([Makur and Polyanskiy, 2016; Polyanskiy

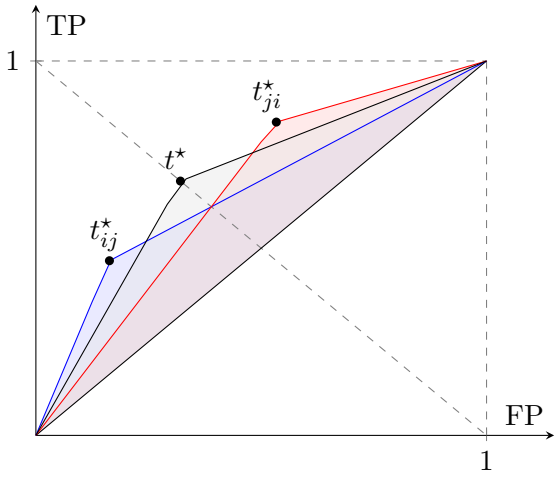


Figure 3.11: Comparison between ϵ-local privacy (with $\epsilon = 0.75$) and C-local privacy (with $C_{ij} = 1.05$, and $C_{ji} = 0.45$).

and Wu, 2015]). It would be interesting to study the relationship between the statistical deficiency based channel ordering (Raginsky [2011]) and the strong data processing inequalities.

- Wider exploration of asymmetric privacy notions.

3.6 Appendix

3.6.1 VC Dimension

A measure of complexity in learning theory should reflect which learning problems are inherently easier than others. The standard approach in statistical theory is to define the complexity of the learning problem through some notion of "richness", "size", "capacity" of the hypothesis class.

The complexity measure proposed in Vapnik and Chervonenkis [1971], the *Vapnik-Chervonenkis (VC) dimension* is a *combinatorial* measure of the richness of classes of binary-valued functions when evaluated on samples. VC-dimension is independent of the underlying probability measure and of the particular sample, and hence is worst-case estimate with regard to these quantities.

We use the notation x_1^m for a sequence $(x_1, \ldots, x_m) \in \mathcal{X}^m$, and for a class of binary-valued functions $\mathcal{F} \subseteq \{-1, 1\}^{\mathcal{X}}$, we denote by $\mathcal{F}_{|x_1^m}$ the *restriction* of $\mathcal{F}$ to x_1^m:

$$\mathcal{F}_{|x_1^m} = \{(f(x_1), \ldots, f(x_m)) \mid f \in \mathcal{F}\}.$$

Define the *m-th shatter coefficient* of $\mathcal{F}$ as follows:

$$\mathbb{S}_m(\mathcal{F}) := \max_{x_1^m \in \mathcal{X}^m} |\mathcal{F}_{|x_1^m}|.$$

Definition 3.26. *Let* $\mathcal{F} \subseteq \{-1,1\}^{\mathcal{X}}$ *and let* $x_1^m = (x_1, \ldots, x_m) \in \mathcal{X}^m$. *We say* x_1^m *is shattered by* $\mathcal{F}$ *if* $|\mathcal{F}_{|x_1^m}| = 2^m$*; i.e. if* $\forall \mathbf{b} \in \{-1,1\}^m, \exists f_{\mathbf{b}} \in \mathcal{F}$ *s.t.* $(f_{\mathbf{b}}(x_1), \ldots, f_{\mathbf{b}}(x_m)) = \mathbf{b}$. *The Vapnik-Chervonenkis (VC) dimension of* $\mathcal{F}$*, denoted by* $VCdim(\mathcal{F})$*, is the cardinality of the largest set of points in* $\mathcal{X}$ *that can be shattered by* $\mathcal{F}$*:*

$$VCdim(\mathcal{F}) = \max\{m \in \mathbb{N} \mid \mathbb{S}_m(\mathcal{F}) = 2^m\}.$$

If $\mathcal{F}$ *shatters arbitrarily large sets of points in* $\mathcal{X}$*, then* $VCdim(\mathcal{F}) = \infty$. *If* $VCdim(\mathcal{F}) < \infty$*, we say that* $\mathcal{F}$ *is a VC class.*

Chapter 4

Exp-concavity of Proper Composite Losses

Loss functions are the means by which the quality of a prediction in learning problem is evaluated. A composite loss (the composition of a class probability estimation (CPE) loss with an invertible link function which is essentially just a re-parameterization) is proper if its risk is minimized when predicting the true underlying class probability (a formal definition is given later). In Williamson et al. [2016], there is an argument that shows that there is no point in using losses that are neither proper nor proper composite as they are inadmissible. Flexibility in the choice of loss function is important to tailor the solution to a learning problem (Buja et al. [2005], Hand [1994], Hand and Vinciotti [2003]), and it could be attained by characterizing the set of loss functions using natural parameterizations.

The goal of the learner in a *game of prediction with expert advice* (which is formally described in section 4.1.5) is to predict as well as the best expert in the given pool of experts. The regret bound of the learner depends on the merging scheme used to merge the experts' predictions and the type of loss function used to measure the performance. It has already been shown that constant regret bounds are achievable for mixable losses when the Aggregating Algorithm is the merging scheme (Vovk [1995]), and for exp-concave losses when the Weighted Average Algorithm is the merging scheme (Kivinen and Warmuth [1999]). We can see that the exp-concavity trivially implies mixability. Even though the converse implication is not true in general, under some re-parameterization we can make it possible. This chapter discusses general conditions on proper losses under which they can be transformed to an exp-concave loss through a suitable link function. In the binary case, these conditions give two concrete formulas (Proposition 4.1 and Corollary 4.8) for link functions that can transform β-mixable proper losses into β-exp-concave, proper, composite losses. The explicit form of the link function given in Proposition 4.1 is derived using the same geometric construction used in van Erven [2012].

Further we extend the work by Williamson et al. [2016], to provide a complete characterization of the exp-concavity of the proper composite multi-class losses in terms of the Bayes risk associated with the underlying proper loss, and the link function. The mixability of proper losses (mixability of a proper composite loss is equivalent

to the mixability of its generating proper loss) is studied in Van Erven et al. [2012]. Using these characterizations (for the binary case), in Corollary 4.8 we derive an *exp-concavifying link* function that can also transform any β-mixable proper loss into a β-exp-concave composite loss. Since for the multi-class losses these conditions do not hold in general, we propose a geometric approximation approach (Proposition 4.2) which takes a parameter ϵ and transforms the mixable loss function appropriately on a subset S_ϵ of the prediction space. When the prediction space is Δ^n, any prediction belongs to the subset S_ϵ for sufficiently small ϵ. In the conclusion we provide a way to use the Weighted Average Algorithm with learning rate β for proper β-mixable but non-exp-concave loss functions to achieve $O(1)$ regret bound.

The exp-concave losses achieve $O(\log T)$ regret bound in online convex optimization algorithms, which is a more general setting of online learning problems. Thus the exp-concavity characterization of composite losses could be helpful in constructing exp-concave losses for online learning problems.

The chapter is organized as follows. In Section 4.1 we formally introduce the loss function, several loss types, conditional risk, proper composite losses and a game of prediction with expert advice. In Section 4.2 we consider our main problem — whether one can always find a link function to transform β-mixable losses into β-exp-concave losses. Section 4.3 concludes with a brief discussion. The impact of the choice of substitution function on the regret of the learner is explored via experiments in Appendix 4.4.1. In Appendix 4.4.2, we discuss the mixability conditions of *probability games* with continuous outcome space. Detailed proofs are in Appendix 4.4.3.

4.1 Preliminaries and Background

This section provides the necessary background on loss functions, conditional risks, and the sequential prediction problem.

4.1.1 Notation

We use the following notation throughout this chapter. A superscript prime, A' denotes transpose of the matrix or vector A, except when applied to a real-valued function where it denotes derivative (f'). We denote the matrix multiplication of compatible matrices A and B by $A \cdot B$, so the inner product of two vectors $x, y \in \mathbb{R}^n$ is $x' \cdot y$. Let $[n] := \{1, ..., n\}$, $\mathbb{R}_+ := [0, \infty)$ and the n-simplex $\Delta^n := \{(p_1, ..., p_n)' : 0 \leq p_i \leq 1, \forall i \in [n], \text{and} \sum_{i\in[n]} p_i = 1\}$. If x is a n-vector, $A = \text{diag}(x)$ is the $n \times n$ matrix with entries $A_{i,i} = x_i$, $i \in [n]$ and $A_{i,j} = 0$ for $i \neq j$. If $A - B$ is positive definite (resp. semi-definite), then we write $A \succ B$ (resp. $A \succcurlyeq B$). We use e_i^n to denote the ith n-dimensional unit vector, $e_i^n = (0, ..., 0, 1, 0, ...0)'$ when $i \in [n]$, and define $e_i^n = 0_n$ when $i > n$. The n-vector $\mathbf{1}_n := (1, ..., 1)'$. We write $[\![P]\!] = 1$ if P is true and $[\![P]\!] = 0$ otherwise. Given a set S and a weight vector w, the *convex combination* of the elements of the set w.r.t the weight vector is denoted by $\text{co}_w S$, and the *convex hull* of the set which is the set of all possible convex combinations of the elements of the set is denoted by co$S$ (Rockafellar [1970]). If $S, T \subset \mathbb{R}^n$, then the *Minkowski sum*

$S \oplus T := \{s + t : s \in S, t \in T\}$. $\mathcal{Y}^{\mathcal{X}}$ represents the set of all functions $f : \mathcal{X} \to \mathcal{Y}$. We say $f : C \subset \mathbb{R}^n \to \mathbb{R}^n$ is *monotone* (resp. *strictly monotone*) on C when for all x and y in C,

$$(f(x) - f(y))' \cdot (x - y) \geq 0 \quad \text{resp.} \quad (f(x) - f(y))' \cdot (x - y) > 0;$$

confer Hiriart-Urruty and Lemaréchal [1993]. Other notation (the Kronecker product $\otimes$, the Jacobian D, and the Hessian H) is defined in Appendix A of Van Erven et al. [2012]. $\mathsf{D}f(v)$ and $\mathsf{H}f(v)$ denote the Jacobian and Hessian of $f(v)$ w.r.t. v respectively. When it is not clear from the context, we will explicitly mention the variable; for example $\mathsf{D}_{\tilde{v}} f(v)$ where $v = h(\tilde{v})$.

4.1.2 Loss Functions

For a prediction problem with an instance space $\mathcal{X}$, outcome space $\mathcal{Y}$ and prediction space $\mathcal{V}$, a loss function $\ell : \mathcal{Y} \times \mathcal{V} \to \mathbb{R}_+$ (bivariate function representation) can be defined to assign a penalty $\ell(y, v)$ for predicting $v \in \mathcal{V}$ when the actual outcome is $y \in \mathcal{Y}$. When the outcome space $\mathcal{Y} = [n]$, $n \geq 2$, the loss function ℓ is called a *multi-class loss* and it can be expressed in terms of its partial losses $\ell_i := \ell(i, \cdot)$ for any outcome $i \in [n]$, as

$$\ell(y, v) = \sum_{i \in [n]} [\![y = i]\!] \ell_i(v).$$

The vector representation of the multi-class loss is given by $\ell : \mathcal{V} \to \mathbb{R}_+^n$, which assigns a vector $\ell(v) = (\ell_1(v), ..., \ell_n(v))'$ to each prediction $v \in \mathcal{V}$. A loss is differentiable if all of its partial losses are differentiable. In this thesis, we will use the bivariate function representation $(\ell(y, v))$ to denote a general loss function and the vector representation for multi-class loss functions.

The *super-prediction set* of a binary loss ℓ is defined as

$$S_\ell := \{x \in \mathbb{R}^n : \exists v \in \mathcal{V}, x \geq \ell(v)\},$$

where inequality is component-wise. For any dimension n and $\beta \geq 0$, the β-exponential operator $E_\beta : [0, \infty]^n \to [0, 1]^n$ is defined by

$$E_\beta(x) := (e^{-\beta x_1}, ..., e^{-\beta x_n})'.$$

For $\beta > 0$ it is clearly invertible with inverse

$$E_\beta^{-1}(z) = -\beta^{-1}(\ln z_1, ..., \ln z_n)'.$$

The β-exponential transformation of the super-prediction set is given by

$$E_\beta(S_\ell) := \{(e^{-\beta x_1}, ..., e^{-\beta x_n})' \in \mathbb{R}^n : (x_1, ..., x_n)' \in S_\ell\}, \quad \beta > 0.$$

A multi-class loss ℓ is

- *convex* if $f(v) = \ell_y(v)$ is convex in v for all $y \in [n]$,

- α-*exp-concave* (for $\alpha > 0$) if $f(v) = e^{-\alpha \ell_y(v)}$ is concave in v for all $y \in [n]$ (Cesa-Bianchi and Lugosi [2006]),
- *weakly mixable* if the super-prediction set S_ℓ is convex (Kalnishkan and Vyugin [2005]), and
- β-*mixable* (for $\beta > 0$) if the set $E_\beta(S_\ell)$ is convex (Vovk and Zhdanov [2009]; Vovk [1995]).

The *mixability constant* β_ℓ of a loss ℓ is the largest β such that ℓ is β-mixable; i.e.

$$\beta_\ell := \sup \{\beta > 0 : \ell \text{ is } \beta\text{-mixable}\}.$$

If the loss function ℓ is α-exp-concave (resp. β-mixable) then it is α'-exp-concave for any $0 < \alpha' \leq \alpha$ (resp. β'-mixable for any $0 < \beta' \leq \beta$), and its λ-scaled version ($\lambda\ell$) for some $\lambda > 0$ is $\frac{\alpha}{\lambda}$-exp-concave (resp. $\frac{\beta}{\lambda}$-mixable). If the loss ℓ is α-exp-concave, then it is convex and α-mixable (Cesa-Bianchi and Lugosi [2006]).

For a multi-class loss ℓ, if the prediction space $\mathcal{V} = \Delta^n$ then it is said to be *multi-class probability estimation (CPE) loss*, where the predicted values are directly interpreted as probability estimates: $\ell : \Delta^n \to \mathbb{R}^n_+$. We will say a multi-CPE loss is *fair* whenever $\ell_i(e^n_i) = 0$, for all $i \in [n]$. That is, there is no loss incurred for perfect prediction. Examples of multi-CPE losses include

1. the *square loss* $\ell^{\text{sq}}_i(q) := \sum_{j\in[n]} (\llbracket i = j \rrbracket - q_j)^2$,
2. the *log loss* $\ell^{\log}_i(q) := -\log q_i$,
3. the *absolute loss* $\ell^{\text{abs}}_i(q) := \sum_{j\in[n]} |\llbracket i = j \rrbracket - q_j|$, and
4. the *0-1 loss* $\ell^{01}_i(q) := \llbracket i \notin \arg\max_{j\in[n]} q_j \rrbracket$.

4.1.3 Conditional and Full Risks

Let X and Y be random variables taking values in the instance space $\mathcal{X}$ and the outcome space $\mathcal{Y} = [n]$ respectively. Let D be the joint distribution of (X, Y) and for $x \in \mathcal{X}$, denote the conditional distribution by $p(x) = (p_1(x), ..., p_n(x))'$ where $p_i(x) := P(\mathsf{Y} = i | \mathsf{X} = x)$, $\forall i \in [n]$, and the marginal distribution by $M(x) := P(\mathsf{X} = x)$. For any multi-CPE loss ℓ, the *conditional Bayes risk* is defined as

$$L_\ell : \Delta^n \times \Delta^n \ni (p, q) \mapsto L_\ell(p, q) = \mathbb{E}_{\mathsf{Y}\sim p}[\ell_{\mathsf{Y}}(q)] = p' \cdot \ell(q) = \sum_{i\in[n]} p_i \ell_i(q) \in \mathbb{R}_+ , \quad (4.1)$$

where $\mathsf{Y} \sim p$ represents a Multinomial distribution with parameter $p \in \Delta^n$. The *full Bayes risk* of the estimator function $q : \mathcal{X} \to \Delta^n$ is defined as

$$\widehat{\mathcal{R}}_\ell(M, p, q) := \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D}[\ell_{\mathsf{Y}}(q(\mathsf{X}))] = \mathbb{E}_{\mathsf{X}\sim M}[L_\ell(p(\mathsf{X}), q(\mathsf{X}))].$$

Furthermore the *minimum full Bayes risk* is defined as

$$\widehat{\underline{\mathcal{R}}}_\ell(M,p) := \inf_{q\in(\Delta^n)^{\mathcal{X}}} \widehat{\mathcal{R}}_\ell(M,p,q) = \mathbb{E}_{\mathsf{X}\sim M}[\underline{L}_\ell(p(\mathsf{X}))],$$

where $\underline{L}_\ell(p) = \inf_{q\in\Delta^n} L_\ell(p,q)$ is the *minimum conditional Bayes risk* and is always concave (Gneiting and Raftery [2007]). If ℓ is fair, $\underline{L}_\ell(e_i^n) = \ell_i(e_i^n) = 0$. One can understand the effect of choice of loss in terms of the conditional perspective (Reid and Williamson [2011]), which allows one to ignore the marginal distribution M of X which is typically unknown.

4.1.4 Proper and Composite Losses

A multi-CPE loss $\ell : \Delta^n \to \mathbb{R}_+^n$ is said to be *proper* if for all $p \in \Delta^n$, $\underline{L}_\ell(p) = L_\ell(p,p) = p' \cdot \ell(p)$ (Buja et al. [2005], Gneiting and Raftery [2007]), and *strictly proper* if $\underline{L}_\ell(p) < L_\ell(p,q)$ for all $p, q \in \Delta^n$ and $p \neq q$. It is easy to see that the log loss, square loss, and 0-1 loss are proper while absolute loss is not. Furthermore, both log loss and square loss are strictly proper while 0-1 loss is proper but not strictly proper.

Given a proper loss $\ell : \Delta^n \to \mathbb{R}_+^n$ with differentiable Bayes conditional risk $\underline{L}_\ell : \Delta^n \mapsto \mathbb{R}_+$, in order to be able to calculate derivatives easily, following Van Erven et al. [2012] we define

$$\tilde{\Delta}^n := \left\{(p_1,\ldots,p_{n-1})' : p_i \geq 0,\ \sum_{i=1}^{n-1} p_i \leq 1\right\} \tag{4.2}$$

$$\Pi_\Delta : \mathbb{R}_+^n \ni p = (p_1,\ldots,p_n)' \mapsto \tilde{p} = (p_1,\ldots,p_{n-1})' \in \mathbb{R}_+^{n-1} \tag{4.3}$$

$$\Pi_\Delta^{-1} : \tilde{\Delta}^n \ni \tilde{p} = (\tilde{p}_1,\ldots,\tilde{p}_{n-1})' \mapsto p = (\tilde{p}_1,\ldots,\tilde{p}_{n-1}, 1-\textstyle\sum_{i=1}^{n-1}\tilde{p}_i)' \in \Delta^n \tag{4.4}$$

$$\tilde{\underline{L}}_\ell : \tilde{\Delta}^n \ni \tilde{p} \mapsto \underline{L}_\ell(\Pi_\Delta^{-1}(\tilde{p})) \in \mathbb{R}_+ \tag{4.5}$$

$$\tilde{\ell} : \tilde{\Delta}^n \ni \tilde{p} \mapsto \Pi_\Delta(\ell(\Pi_\Delta^{-1}(\tilde{p}))) \in \mathbb{R}_+^{n-1}. \tag{4.6}$$

Let $\tilde{\psi} : \tilde{\Delta}^n \to \mathcal{V} \subseteq \mathbb{R}_+^{n-1}$ be continuous and strictly monotone (hence invertible) for some convex set $\mathcal{V}$. It induces $\psi : \Delta^n \to \mathcal{V}$ via

$$\psi := \tilde{\psi} \circ \Pi_\Delta. \tag{4.7}$$

Clearly ψ is continuous and invertible with $\psi^{-1} = \Pi_\Delta^{-1} \circ \tilde{\psi}^{-1}$. We can now extend the notion of properness to the prediction space $\mathcal{V}$ from Δ^n using this link function. Given a proper loss $\ell : \Delta^n \to \mathbb{R}_+^n$, a *proper composite loss* $\ell^\psi : \mathcal{V} \to \mathbb{R}_+^n$ for multi-class probability estimation is defined as $\ell^\psi := \ell \circ \psi^{-1} = \ell \circ \Pi_\Delta^{-1} \circ \tilde{\psi}^{-1}$. We can easily see that the conditional Bayes risks of the composite loss ℓ^ψ and the underlying proper loss ℓ are equal ($\underline{L}_\ell = \underline{L}_{\ell^\psi}$). Every continuous proper loss has a convex super-prediction set (Williamson et al. [2016]). Thus they are weakly mixable. Since by applying a link function the super-prediction set won't change (as it is just a re-parameterization), all proper composite losses are also weakly mixable.

4.1.5 Game of Prediction with Expert Advice

Let $\mathcal{Y}$ be the outcome space, $\mathcal{V}$ be the prediction space, and $\ell : \mathcal{Y} \times \mathcal{V} \to \mathbb{R}_+$ be the loss function, then a *game of prediction with expert advice* represented by the tuple $(\mathcal{Y}, \mathcal{V}, \ell)$ can be described as follows: for each trial $t = 1, ..., T$,

- N experts make their prediction $v_t^1, ..., v_t^N \in \mathcal{V}$
- the learner makes his own decision $v_t \in \mathcal{V}$
- the environment reveals the actual outcome $y_t \in \mathcal{Y}$

Let $S = (y_1, ..., y_T)$ be the outcome sequence in T trials. Then the *cumulative loss* of the learner over S is given by $L_{S,\ell} := \sum_{t=1}^T \ell(y_t, v_t)$, of the i-th expert is given by $L_{S,\ell}^i := \sum_{t=1}^T \ell(y_t, v_t^i)$, and the *regret* of the learner is given by $R_{S,\ell} := L_{S,\ell} - \min_i L_{S,\ell}^i$. The goal of the learner is to predict as well as the best expert; to which end the learner tries to minimize the regret.

When using the exponential weights algorithm (which is an important family of algorithms in game of prediction with expert advice), at the end of each trial, the weight of each expert is updated as $w_{t+1}^i \propto w_t^i \cdot e^{-\eta \ell(y_t, v_t^i)}$ for all $i \in [N]$, where η is the learning rate and w_t^i is the weight of the i^{th} expert at time t (the weight vector of experts at time t is denoted by $w_t = (w_t^1, ..., w_t^N)'$). Then based on the weights of experts, their predictions are merged using different merging schemes to make the learner's prediction. The Aggregating Algorithm and the Weighted Average Algorithm are two important algorithms in the family of exponential weights algorithm.

Consider multi-class games with outcome space $\mathcal{Y} = [n]$. In the Aggregating Algorithm with learning rate β, first the loss vectors of the experts and their weights are used to make a *generalized prediction* $g_t = (g_t(1), ..., g_t(n))'$ which is given by

$$\begin{aligned} g_t &:= E_\beta^{-1}\left(\mathrm{co}_{w_t}\{E_\beta\left((\ell_1(v_t^i), ..., \ell_n(v_t^i))'\right)\}_{i\in[N]}\right) \\ &= E_\beta^{-1}\left(\sum_{i\in[N]} w_t^i (e^{-\beta\ell_1(v_t^i)}, ..., e^{-\beta\ell_n(v_t^i)})'\right). \end{aligned}$$

Then this generalized prediction is mapped into a permitted prediction v_t via a *substitution function* such that $(\ell_1(v_t), ..., \ell_n(v_t))' \le c(\beta)(g_t(1), ..., g_t(n))'$, where the inequality is element-wise and the constant $c(\beta)$ depends on the learning rate. If ℓ is β-mixable, then $E_\beta(S_\ell)$ is convex, so $\mathrm{co}\{E_\beta(\ell(v)) : v \in \mathcal{V}\} \subseteq E_\beta(S_\ell)$, and we can always choose a substitution function with $c(\beta) = 1$. Consequently for β-mixable losses, the learner of the Aggregating Algorithm is guaranteed to have regret bounded by $\frac{\log N}{\beta}$ (Vovk [1995]).

In the Weighted Average Algorithm with learning rate α, the experts' predictions are simply merged according to their weights to make the learner's prediction $v_t = \mathrm{co}_{w_t}\{v_t^i\}_{i\in[N]}$, and this algorithm is guaranteed to have a $\frac{\log N}{\alpha}$ regret bound for α-exp-concave losses (Kivinen and Warmuth [1999]). In either case it is preferred to have bigger values for the constants β and α to have better regret bounds. Since an α-exp-concave loss is β-mixable for some $\beta \ge \alpha$, the regret bound of the Weighted

Average Algorithm is worse than that of the Aggregating Algorithm by a small constant factor. In Vovk [2001], it is noted that $(\ell_1(\text{co}_{w_t}\{v_t^i\}_i), ..., \ell_n(\text{co}_{w_t}\{v_t^i\}_i))' \leq g_t = E_\beta^{-1}\left(\text{co}_{w_t}\{E_\beta\left((\ell_1(v_t^i), ..., \ell_n(v_t^i))'\right)\}_i\right)$ is always guaranteed only for β-exp-concave losses. Thus for α-exp-concave losses, the Weighted Average Algorithm is equivalent to the Aggregating Algorithm with the weighted average of the experts' predictions as its substitution function and α as the learning rate for both algorithms.

Even though the choice of substitution function will not have any impact on the regret bound and the weight update mechanism of the Aggregating Algorithm, it will certainly have impact on the actual regret of the learner over a given sequence of outcomes. According to the results given in Appendix 4.4.1 (where we have empirically compared some substitution functions), this impact on the actual regret varies depending on the outcome sequence, and in general the regret values for practical substitution functions don't differ much — thus we can stick with a computationally efficient substitution function.

4.2 Exp-Concavity of Proper Composite Losses

Exp-concavity of a loss is desirable for better (logarithmic) regret bounds in online convex optimization algorithms, and for efficient implementation of exponential weights algorithms. In this section we will consider whether one can always find a link function that can transform a β-mixable proper loss into β-exp-concave composite loss — first by using the geometry of the set $E_\beta(S_\ell)$ (Section 4.2.1), and then by using the characterization of the composite loss in terms of the associated Bayes risk (Sections 4.2.2, and 4.2.3).

4.2.1 Geometric approach

In this section we will use the same construction used by van Erven [2012] to derive an explicit closed form of a link function that could re-parameterize any β-mixable loss into a β-exp-concave loss, under certain conditions which are explained below. Given a multi-class loss $\ell : \mathcal{V} \to \mathbb{R}_+^n$, define

$$\ell(\mathcal{V}) := \{\ell(v) : v \in \mathcal{V}\}, \tag{4.8}$$

$$\mathcal{B}_\beta := \text{co}E_\beta(\ell(\mathcal{V})). \tag{4.9}$$

For any $g \in \mathcal{B}_\beta$ let $c(g) := \sup\{c \geq 0 : (g + c\mathbf{1}_n) \in \mathcal{B}_\beta\}$. Then the "north-east" boundary of the set $\mathcal{B}_\beta$ is given by $\partial_{\mathbf{1}_n}\mathcal{B}_\beta := \{g + c(g) : g \in \mathcal{B}_\beta\}$. The following proposition is the main result of this section.

Proposition 4.1. *Assume ℓ is strictly proper and it satisfies the condition : $\partial_{\mathbf{1}_n}\mathcal{B}_\beta \subseteq E_\beta(\ell(\mathcal{V}))$ for some $\beta > 0$. Define $\psi(p) := JE_\beta(\ell(p))$ for all $p \in \Delta^n$, where $J = [I_{n-1}, -\mathbf{1}_{n-1}]$. Then ψ is invertible, and $\ell \circ \psi^{-1}$ is β-exp-concave over $\psi(\Delta^n)$, which is a convex set.*

The condition stated in the above proposition is satisfied by any β-mixable proper loss in the binary case ($n = 2$), but it is not guaranteed in the multi-class case where

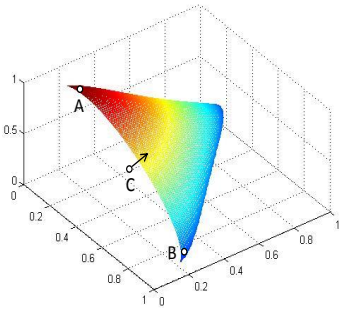


Figure 4.1: Ray "escaping" in $\mathbf{1}_n$ direction. More evidence in Figure 4.12 in Appendix 4.4.4.

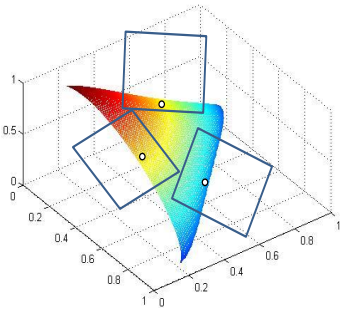


Figure 4.2: Adding "faces" to block rays in (almost) all positive directions.

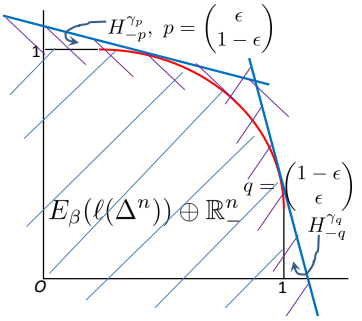


Figure 4.3: Sub-exp-prediction set extended by removing near axis-parallel supporting hyperplanes.

$n > 2$. In the binary case the link function can be given as $\psi(p) = e^{-\beta\ell_1(p)} - e^{-\beta\ell_2(p)}$ for all $p \in \Delta^2$.

Unfortunately, the condition that $\partial_{\mathbf{1}_n}\mathcal{B}_\beta \subseteq E_\beta(\ell(\mathcal{V}))$ is generally not satisfied; an example based on squared loss ($\beta = 1$ and $n = 3$ classes) is shown in Figure 4.1, where for A and B in $E_\beta(\ell(\mathcal{V}))$, the mid-point C can travel along the ray of direction $\mathbf{1}_3$ without hitting any point in the exp-prediction set $E_\beta(\ell(\mathcal{V}))$. Therefore we resort to *approximating* a given β-mixable loss by a sequence of β-exp-concave losses parameterised by positive constant ϵ, while the approximation approaches the original loss in some appropriate sense as ϵ tends to 0. Without loss of generality, we assume $\mathcal{V} = \Delta^n$.

Inspired by Proposition 4.1, a natural idea to construct the approximation is by adding "faces" to the exp-prediction set such that all rays in the $\mathbf{1}_n$ direction will be blocked. Technically, it turns out more convenient to add faces that block rays in (almost) all directions of positive orthant. See Figure 4.2 for an illustration. In particular, we extend the "rim" of the exp-prediction set by hyperplanes that are ϵ close to axis-parallel. The key challenge underlying this idea is to design an appropriate *parameterisation* of the surrogate loss $\tilde{\ell}_\epsilon$, which not only produces such an extended exp-prediction set, but also ensures that $\tilde{\ell}_\epsilon(p) = \ell(p)$ for almost all $p \in \Delta^n$ as $\epsilon \downarrow 0$.

Given a β-mixable loss ℓ, its sub-exp-prediction set defined as follows must be convex:

$$T_\ell := E_\beta(\ell(\Delta^n)) \oplus \mathbb{R}_-^n = \{E_\beta(\ell(p)) - x : p \in \Delta^n, x \in \mathbb{R}_+^n\}. \tag{4.10}$$

Note T_ℓ extends infinitely in any direction $p \in \mathbb{R}_-^n$. Therefore it can be written in terms of supporting hyperplanes as

$$T_\ell = \bigcap_{p \in \Delta^n} H_{-p}^{\gamma_p}, \quad \text{where} \quad \gamma_p = \min_{x \in T_\ell} x' \cdot (-p), \text{ and } H_{-p}^{\gamma_p} := \{x : x' \cdot (-p) \geq \gamma_p\}. \tag{4.11}$$

To extend the sub-exp-prediction set with "faces", we remove some hyperplanes involved in (4.11) that correspond to the ϵ "rim" of the simplex (see Figure 4.3 for an

illustration in 2-D)

$$T_\ell^\epsilon := \bigcap_{p\in\Delta_\epsilon^n} H_{-p}^{\gamma_p}, \quad \text{where} \quad \Delta_\epsilon^n := \{p \in \Delta^n : \min_i p_i > \epsilon\}. \tag{4.12}$$

Since $\epsilon > 0$, for any $p \in \Delta_\epsilon^n$, $E_\beta^{-1}(H_{-p}^{\gamma_p} \cap \mathbb{R}_+^n)$ is exactly the super-prediction set of a log-loss with appropriate scaling and shifting (see proof in Appendix 4.4.3). So it must be convex. Therefore $E_\beta^{-1}(T_\ell^\epsilon \cap \mathbb{R}_+^n) = \bigcap_{p\in\Delta_\epsilon^n} E_\beta^{-1}(H_{-p}^{\gamma_p} \cap \mathbb{R}_+^n)$ must be convex, and its recession cone is clearly $\mathbb{R}_+^n$. This guarantees that the following loss is proper over $p \in \Delta^n$ [Williamson, 2014, Proposition 2]:

$$\tilde{\ell}_\epsilon(p) = \operatorname*{arg\,min}_{z\in E_\beta^{-1}(T_\ell^\epsilon\cap\mathbb{R}_+^n)} p' \cdot z, \tag{4.13}$$

where the argmin must be attained uniquely (Appendix 4.4.3). Our next proposition states that $\tilde{\ell}_\epsilon$ meets all the requirements of approximation suggested above.

Proposition 4.2. *For any $\epsilon > 0$, $\tilde{\ell}_\epsilon$ satisfies the condition $\partial_{\mathbf{1}_n}\mathcal{B}_\beta \subseteq E_\beta(\ell(\mathcal{V}))$. In addition, $\tilde{\ell}_\epsilon = \ell$ over a subset $S_\epsilon \subseteq \Delta^n$, where for any p in the relative interior of Δ^n, $p \in S_\epsilon$ for sufficiently small ϵ i.e. $\lim_{\epsilon\downarrow 0} \operatorname{vol}(\Delta^n \setminus S_\epsilon) = 0$.*

Note $\|\tilde{\ell}_\epsilon(p) - \ell(p)\|$ is not bounded for $p \notin S_\epsilon$. While the result does not show that *all* β-mixable losses can be made β-exp-concave, it is suggestive that such a result may be obtainable by a different argument.

4.2.2 Calculus approach

Proper composite losses are defined by the proper loss ℓ and the link ψ. In this section we will characterize the exp-concave proper composite losses in terms of $(\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{p}), \mathsf{D}\tilde{\psi}(\tilde{p}))$. The following proposition provides the identities of the first and second derivatives of the proper composite losses (Williamson et al. [2016]).

Proposition 4.3. *For all $i \in [n]$, $\tilde{p} \in \mathring{\tilde{\Delta}}^n$ (the interior of $\tilde{\Delta}^n$), and $v = \tilde{\psi}(\tilde{p}) \in \mathcal{V} \subseteq \mathbb{R}_+^{n-1}$ (so $\tilde{p} = \tilde{\psi}^{-1}(v)$),*

$$\mathsf{D}\ell_i^\psi(v) = -(e_i^{n-1} - \tilde{p})' \cdot k(\tilde{p}), \tag{4.14}$$

$$\mathsf{H}\ell_i^\psi(v) = -\left((e_i^{n-1} - \tilde{p})' \otimes I_{n-1}\right) \cdot \mathsf{D}_v[k(\tilde{p})] + k(\tilde{p})' \cdot [\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1}, \tag{4.15}$$

where

$$k(\tilde{p}) := -\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{p}) \cdot [\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1}. \tag{4.16}$$

The term $k(\tilde{p})$ can be interpreted as the curvature of the Bayes risk function $\underline{\tilde{L}}_\ell$ relative to the rate of change of the link function $\tilde{\psi}$. In the binary case where $n = 2$, above proposition reduces to

$$(\ell_1^\psi)'(v) = -(1-\tilde{p})k(\tilde{p}) \quad ; \quad (\ell_2^\psi)'(v) = \tilde{p}k(\tilde{p}), \tag{4.17}$$

$$(\ell_1^\psi)''(v) = \frac{-(1-\tilde{p})k'(\tilde{p}) + k(\tilde{p})}{\tilde{\psi}'(\tilde{p})}, \tag{4.18}$$

$$(\ell_2^\psi)''(v) \quad = \quad \frac{\tilde{p}k'(\tilde{p}) + k(\tilde{p})}{\tilde{\psi}'(\tilde{p})}, \tag{4.19}$$

where $k(\tilde{p}) = \frac{-\tilde{\underline{L}}_\ell''(\tilde{p})}{\tilde{\psi}'(\tilde{p})} \geq 0$ and so $\frac{d}{dv}k(\tilde{p}) = \frac{d}{d\tilde{p}}k(\tilde{p}) \cdot \frac{d}{dv}\tilde{p} = \frac{k'(\tilde{p})}{\tilde{\psi}'(\tilde{p})}$.

A loss $\ell : \Delta^n \to \mathbb{R}_+^n$ is α-exp-concave (i.e. $\Delta^n \ni q \mapsto \ell_y(q)$ is α-exp-concave for all $y \in [n]$) if and only if the map $\Delta^n \ni q \mapsto L_\ell(p,q) = p' \cdot \ell(q)$ is α-exp-concave for all $p \in \Delta^n$. It can be easily shown that the maps $v \mapsto \ell_y^\psi(v)$ are α-exp-concave if and only if $\mathsf{H}\ell_y^\psi(v) \succcurlyeq \alpha \mathsf{D}\ell_y^\psi(v)' \cdot \mathsf{D}\ell_y^\psi(v)$. By applying Proposition 4.3 we obtain the following characterization of the α-exp-concavity of the composite loss ℓ^ψ.

Proposition 4.4. *A proper composite loss $\ell^\psi = \ell \circ \psi^{-1}$ is α-exp-concave (with $\alpha > 0$ and $v = \tilde{\psi}(\tilde{p})$) if and only if for all $\tilde{p} \in \mathring{\tilde{\Delta}}^n$ and for all $i \in [n]$*

$$\left((e_i^{n-1} - \tilde{p})' \otimes I_{n-1}\right) \cdot \mathsf{D}_v[k(\tilde{p})] \preccurlyeq k(\tilde{p})' \cdot [\mathsf{D}\tilde{\psi}(\tilde{p})]^{-1} - \alpha k(\tilde{p})' \cdot (e_i^{n-1} - \tilde{p}) \cdot (e_i^{n-1} - \tilde{p})' \cdot k(\tilde{p}). \tag{4.20}$$

Based on this characterization, we can determine which loss functions can be exp-concavified by a chosen link function and how much a link function can exp-concavify a given loss function. In the binary case ($n = 2$), the above proposition reduces to the following.

Proposition 4.5. *Let $\tilde{\psi} : [0,1] \to \mathcal{V} \subseteq \mathbb{R}$ be an invertible link and $\ell : \Delta^2 \to \mathbb{R}_+^2$ be a strictly proper binary loss with weight function $w(\tilde{p}) := -\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{p}) = -\tilde{\underline{L}}_\ell''(\tilde{p})$. Then the binary composite loss $\ell^\psi := \ell \circ \Pi_\Delta^{-1} \circ \tilde{\psi}^{-1}$ is α-exp-concave (with $\alpha > 0$) if and only if*

$$-\frac{1}{\tilde{p}} + \alpha w(\tilde{p})\tilde{p} \quad \leq \quad \frac{w'(\tilde{p})}{w(\tilde{p})} - \frac{\tilde{\psi}''(\tilde{p})}{\tilde{\psi}'(\tilde{p})} \quad \leq \quad \frac{1}{1-\tilde{p}} - \alpha w(\tilde{p})(1-\tilde{p}), \quad \forall \tilde{p} \in (0,1). \tag{4.21}$$

The following proposition gives an easier to check necessary condition for the binary proper losses that generate an α-exp-concave (with $\alpha > 0$) binary composite loss given a particular link function. Since scaling a loss function will not affect what a sensible learning algorithm will do, it is possible to normalize the loss functions by normalizing their weight functions by setting $w(\frac{1}{2}) = 1$. By this normalization we are scaling the original loss function by $\frac{1}{w(\frac{1}{2})}$ and the super-prediction set is scaled by the same factor. If the original loss function is β-mixable (resp. α-exp-concave), then the normalized loss function is $\beta w(\frac{1}{2})$-mixable (resp. $\alpha w(\frac{1}{2})$-exp-concave).

Proposition 4.6. *Let $\tilde{\psi} : [0,1] \to \mathcal{V} \subseteq \mathbb{R}$ be an invertible link and $\ell : \Delta^2 \to \mathbb{R}_+^2$ be a strictly proper binary loss with weight function $w(\tilde{p}) := -\mathsf{H}\tilde{\underline{L}}_\ell(\tilde{p}) = -\tilde{\underline{L}}_\ell''(\tilde{p})$ normalised such that $w(\frac{1}{2}) = 1$. Then the binary composite loss $\ell^\psi := \ell \circ \Pi_\Delta^{-1} \circ \tilde{\psi}^{-1}$ is α-exp-concave (with $\alpha > 0$) only if*

$$\frac{\tilde{\psi}'(\tilde{p})}{\tilde{p}(2\tilde{\psi}'(\frac{1}{2}) - \alpha(\tilde{\psi}(\tilde{p}) - \tilde{\psi}(\frac{1}{2})))} \lesseqgtr w(\tilde{p}) \lesseqgtr \frac{\tilde{\psi}'(\tilde{p})}{(1-\tilde{p})(2\tilde{\psi}'(\frac{1}{2}) + \alpha(\tilde{\psi}(\tilde{p}) - \tilde{\psi}(\frac{1}{2})))}, \quad \forall \tilde{p} \in (0,1), \tag{4.22}$$

where $\lesseqgtr$ denotes $\leq$ for $\tilde{p} \geq \frac{1}{2}$ and denotes $\geq$ for $\tilde{p} \leq \frac{1}{2}$.

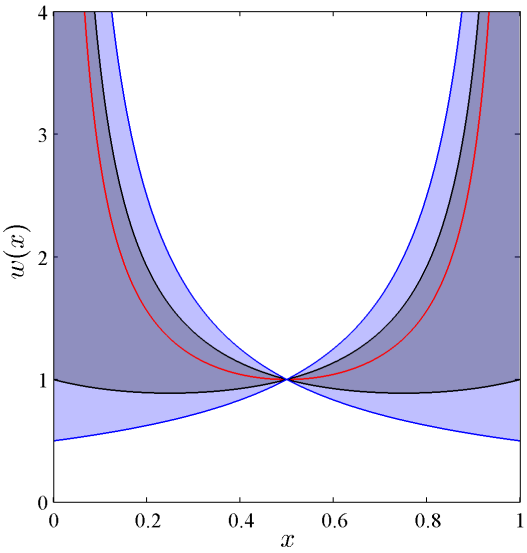


Figure 4.4: Necessary but not sufficient region of normalised weight functions to ensure α-exp-concavity and convexity of proper losses (— $\alpha = 4$; — $\alpha = 2$; — convexity).

Proposition 4.5 provides necessary *and* sufficient conditions for the exp-concavity of binary composite losses, whereas Proposition 4.6 provides simple necessary *but not* sufficient conditions. By setting $\alpha = 0$ in all the above results we have obtained for exp-concavity, we recover the convexity conditions for proper and composite losses which are already derived by Reid and Williamson [2010] for the binary case and Williamson et al. [2016] for multi-class.

4.2.3 Link functions

A proper loss can be exp-concavified ($\alpha > 0$) by some link function only if the loss is mixable ($\beta_\ell > 0$) and the maximum possible value for exp-concavity constant is the mixability constant of the loss (since the link function won't change the super-prediction set and an α-exp-concave loss is always β-mixable for some $\beta \geq \alpha$).

By applying the *identity link* $\tilde{\psi}(\tilde{p}) = \tilde{p}$ in (4.21) we obtain the necessary and sufficient conditions for a binary proper loss to be α-exp-concave (with $\alpha > 0$) as given by,

$$-\frac{1}{\tilde{p}} + \alpha w(\tilde{p})\tilde{p} \;\leq\; \frac{w'(\tilde{p})}{w(\tilde{p})} \;\leq\; \frac{1}{1-\tilde{p}} - \alpha w(\tilde{p})(1-\tilde{p}), \quad \forall \tilde{p} \in (0,1). \tag{4.23}$$

By substituting $\tilde{\psi}(\tilde{p}) = \tilde{p}$ in (4.22) we obtain the following necessary but not sufficient (simpler) constraints for a normalized binary proper loss to be α-exp-concave

$$\frac{1}{\tilde{p}(2-\alpha(\tilde{p}-\frac{1}{2}))} \;\lesseqgtr\; w(\tilde{p}) \;\lesseqgtr\; \frac{1}{(1-\tilde{p})(2+\alpha(\tilde{p}-\frac{1}{2}))}, \quad \forall \tilde{p} \in (0,1), \tag{4.24}$$

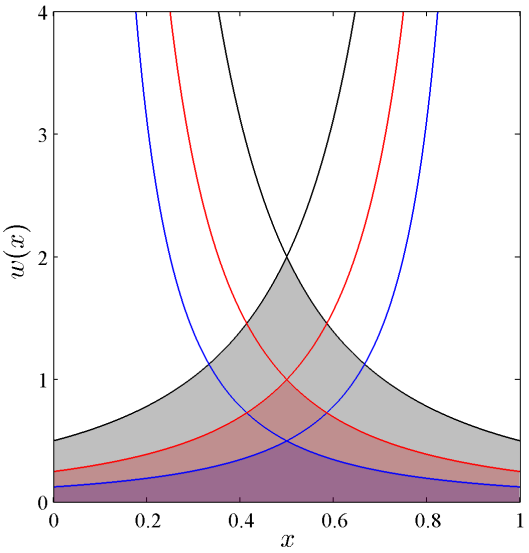


Figure 4.5: Necessary and sufficient region of unnormalised weight functions to ensure α-exp-concavity of composite losses with canonical link (— $\alpha = 2$; — $\alpha = 4$; — $\alpha = 8$).

which are illustrated as the shaded region in Figure 4.4 for different values of α. Observe that normalized proper losses can be α-exp-concave only for $0 < \alpha \leq 4$. When $\alpha = 4$, only the normalized weight function of log loss ($w_{\ell^{\log}}(\tilde{p}) = \frac{1}{4\tilde{p}(1-\tilde{p})}$) will satisfy (4.24), and when $\alpha > 4$, the allowable (necessary) $w(\tilde{p})$ region to ensure α-exp-concavity vanishes. Thus normalized log loss is the most exp-concave normalized proper loss. Observe (from Figure 4.4) that normalized square loss ($w_{\ell^{\text{sq}}}(\tilde{p}) = 1$) is at most 2-exp-concave. Further from (4.24), if $\alpha' > \alpha$, then the allowable $w(\tilde{p})$ region to ensure α'-exp-concavity will be within the region for α-exp-concavity, and also any allowable $w(\tilde{p})$ region to ensure α-exp-concavity will be within the region for convexity, which is obtained by setting $\alpha = 0$ in (4.24). Here we recall the fact that, if the normalized loss function is α-exp-concave, then the original loss function is $\frac{\alpha}{w(\frac{1}{2})}$-exp-concave. The following theorem provides *sufficient* conditions for the exp-concavity of binary proper losses.

Theorem 4.7. *A binary proper loss $\ell : \Delta^2 \to \mathbb{R}^2_+$ with the weight function $w(\tilde{p}) = -\underline{\tilde{L}}''_\ell(\tilde{p})$ normalized such that $w(\frac{1}{2}) = 1$ is α-exp-concave (with $\alpha > 0$) if*

$$w(\tilde{p}) \;=\; \frac{1}{\tilde{p}\left(2 - \int_{\tilde{p}}^{1/2} a(t)dt\right)} \quad \textit{for } a(\tilde{p}) \textit{ s.t.}$$

$$\left[\frac{\alpha(1-\tilde{p})}{\tilde{p}} - \frac{2}{\tilde{p}(1-\tilde{p})}\right] + \frac{1}{\tilde{p}(1-\tilde{p})}\int_{\tilde{p}}^{1/2} a(t)dt \;\leq\; a(\tilde{p}) \;\leq\; -\alpha,\ \forall \tilde{p} \in (0, 1/2],$$

and

$$w(\tilde{p}) \;=\; \frac{1}{(1-\tilde{p})\left(2-\int_{\frac{1}{2}}^{\tilde{p}} b(t)dt\right)} \quad \textit{for } b(\tilde{p}) \textit{ s.t.}$$
$$\left[\frac{\alpha\tilde{p}}{(1-\tilde{p})} - \frac{2}{\tilde{p}(1-\tilde{p})}\right] + \frac{1}{\tilde{p}(1-\tilde{p})}\int_{1/2}^{\tilde{p}} b(t)dt \;\leq\; b(\tilde{p}) \;\leq\; -\alpha,\, \forall \tilde{p} \in [\tfrac{1}{2}, 1).$$

For square loss we can find that $a(\tilde{p}) = \frac{-1}{\tilde{p}^2}$ and $b(\tilde{p}) = \frac{-1}{(1-\tilde{p})^2}$ will satisfy the above sufficient condition with $\alpha = 4$ and for log loss $a(\tilde{p}) = b(\tilde{p}) = -4$ will satisfy the sufficient condition with $\alpha = 4$. It is also easy to see that for symmetric losses $a(\tilde{p})$ and $b(\tilde{p})$ will be symmetric.

When the *canonical link* function $\tilde{\psi}_\ell(\tilde{p}) := -\mathsf{D}\underline{\tilde{L}}_\ell(\tilde{p})'$ is combined with a strictly proper loss to form ℓ^{ψ_ℓ}, since $\mathsf{D}\tilde{\psi}_\ell(\tilde{p}) = -\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{p})$, the first and second derivatives of the composite loss become considerably simpler as follows

$$\mathsf{D}\ell_i^{\psi_\ell}(v) \;=\; -(e_i^{n-1} - \tilde{p})', \tag{4.25}$$
$$\mathsf{H}\ell_i^{\psi_\ell}(v) \;=\; -[\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{p})]^{-1}. \tag{4.26}$$

Since a proper loss ℓ is β-mixable if and only if $\beta\mathsf{H}\underline{\tilde{L}}_\ell(\tilde{p}) \succcurlyeq \mathsf{H}\underline{\tilde{L}}_{\ell^{\log}}(\tilde{p})$ for all $\tilde{p} \in \mathring{\tilde{\Delta}}^n$ (Van Erven et al. [2012]), by applying the canonical link any β-mixable proper loss will be transformed to α-exp-concave proper composite loss (with $\beta \geq \alpha > 0$) but $\alpha = \beta$ is not guaranteed in general. In the binary case, since $\tilde{\psi}'_\ell(\tilde{p}) = -\underline{\tilde{L}}''_\ell(\tilde{p}) = w(\tilde{p})$, we get

$$w(\tilde{p}) \leq \frac{1}{\alpha\tilde{p}^2} \quad \text{and} \quad w(\tilde{p}) \leq \frac{1}{\alpha(1-\tilde{p})^2}, \quad \forall \tilde{p} \in (0,1), \tag{4.27}$$

as the necessary and sufficient conditions for ℓ^{ψ_ℓ} to be α-exp-concave. In this case when $\alpha \to \infty$ the allowed region vanishes (since for proper losses $w(\tilde{p}) \geq 0$). From Figure 4.5 it can be seen that, if the normalized loss function satisfies

$$w(\tilde{p}) \leq \frac{1}{4\tilde{p}^2} \quad \text{and} \quad w(\tilde{p}) \leq \frac{1}{4(1-\tilde{p})^2}, \quad \forall \tilde{p} \in (0,1),$$

then the composite loss obtained by applying the canonical link function on the unnormalized loss with weight function $w_{\text{org}}(\tilde{p})$ is $\frac{4}{w_{\text{org}}(\frac{1}{2})}$-exp-concave.

We now consider whether one can always find a link function that can transform a β-mixable proper loss into β-exp-concave composite loss. In the binary case, such a link function exists and is given in the following corollary.

Corollary 4.8. *Let $w_\ell(\tilde{p}) = -\underline{\tilde{L}}''_\ell(\tilde{p})$. The exp-concavifying link function $\tilde{\psi}^*_\ell$ defined via*

$$\tilde{\psi}^*_\ell(\tilde{p}) = \frac{w_{\ell^{\log}}(\frac{1}{2})}{w_\ell(\frac{1}{2})}\int_0^{\tilde{p}} \frac{w_\ell(v)}{w_{\ell^{\log}}(v)}dv, \quad \forall \tilde{p} \in [0,1] \tag{4.28}$$

(which is a valid strictly increasing link function) will always transform a β-mixable proper loss ℓ into β-exp-concave composite loss $\ell^{\psi^*_\ell}$, where $\ell_y^{\psi^*_\ell}(v) = \ell_y \circ \Pi_\Delta^{-1} \circ (\tilde{\psi}^*_\ell)^{-1}(v)$.

For log loss, the exp-concavifying link is equal to the identity link and the canonical link could be written as $\int_0^{\tilde{p}} w_\ell(v)dv$. If ℓ is a binary proper loss with weight function $w_\ell(\tilde{p})$, then we can define a new proper loss ℓ^{mix} with weight function $w_{\ell^{\text{mix}}}(\tilde{p}) = \frac{w_\ell(\tilde{p})}{w_{\ell^{\log}}(\tilde{p})}$. Then applying the exp-concavifying link $\tilde{\psi}^*_\ell$ on the original loss ℓ is equivalent to applying the canonical link $\tilde{\psi}_\ell$ on the new loss ℓ^{mix}.

The links constructed by the geometric and calculus approaches can be completely different (see Appendix 4.4.4, 4.4.5, and 4.4.6). The former can be further varied by replacing $\mathbf{1}_n$ with any direction in the positive orthant, and the latter can be arbitrarily rescaled. Furthermore, as both links satisfy (4.21) with $\alpha = \beta$, any appropriate interpolation also works.

4.3 Conclusions

If a loss is β-mixable, one can run the Aggregating Algorithm with learning rate β and obtain a $\frac{\log N}{\beta}$ regret bound. Similarly a $\frac{\log N}{\alpha}$ regret bound can be attained by the Weighted Average Algorithm with learning rate α, when the loss is α-exp-concave. Vovk [2001] observed that the weighted average of the expert predictions (Kivinen and Warmuth [1999]) will be a *perfect* (in the technical sense defined in Vovk [2001]) substitution function for the Aggregating Algorithm if and only if the loss function is exp-concave. Thus if we have to use a proper, mixable but non-exp-concave loss function ℓ for a sequential prediction (online learning) problem, an $O(1)$ regret bound could be achieved by the following two approaches:

- Use the Aggregating Algorithm (Vovk [1995]) with the *inverse loss* ℓ^{-1} (Williamson [2014]) as the universal substitution function.

- Apply the exp-concavifying link ($\tilde{\psi}^*_\ell$) on ℓ, derive the β_ℓ-exp-concave composite loss $\ell^{\psi^*_\ell}$. Then use the Weighted Average Algorithm (Kivinen and Warmuth [1999]) with $\ell^{\psi^*_\ell}$ to obtain the learner's prediction in the transformed domain ($v_{\text{avg}} \in \psi^*_\ell(\tilde{\Delta}^n)$). Finally output the inverse link value of this prediction ($(\psi^*_\ell)^{-1}(v_{\text{avg}})$).

In either approach we are faced with a computational problem of evaluating an inverse function. But in the binary class case the inverse of a strictly monotone function can be efficiently evaluated using one sided bisection method (or lookup table). So in conclusion, the latter approach can be more convenient and efficient in computation than the former.

When $n = 2$, we have shown that one can always transform a beta-mixable proper loss into beta-exp-concave proper composite loss using either geometric link function (Proposition 4.1) or calculus-based link function (Corollary 4.8). When $n > 2$, we observed that the square loss (which is a mixable proper loss) cannot be exp-concavified via the geometric link. And by the calculus approach, it is hard to obtain an explicit form for exp-concavifying link function when $n > 2$. Thus when $n > 2$, characterization of proper mixable losses that are exp-concavifiable still remains an open problem. For

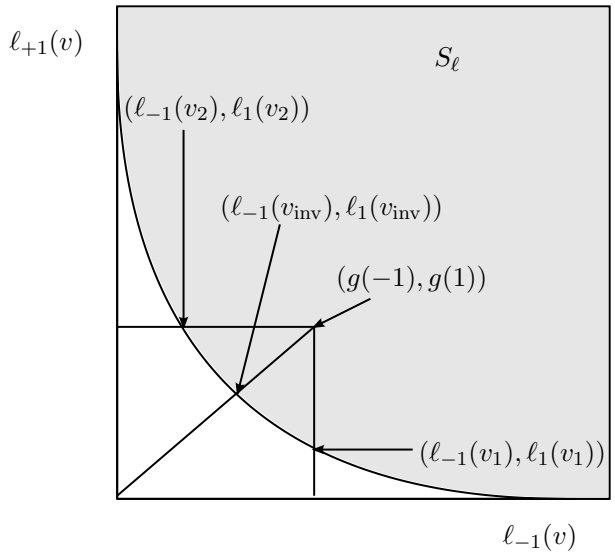


Figure 4.6: Super-prediction set (S_ℓ) of a binary game $(\{-1, 1\}, [0, 1], \ell)$, where a generalized prediction is represented by $(g(-1), g(1))$.

this, we may need to consider other possible ways to construct link functions Williamson [2014].

4.4 Appendix

4.4.1 Substitution Functions

We consider the following choices of substitution functions — *Best look ahead*, *Worst look ahead*, *Inverse loss* (Williamson [2014]) and *Weighted average* (Kivinen and Warmuth [1999]). The first two choices are just hypothetical ones as they look ahead the actual outcome first and then choose their predictions to incur low and high loss respectively, still they give us an idea about how well or worse the Aggregating Algorithm performs over a given outcome sequence. The weighted average is a computationally efficient and easily implementable substitution function, but it is applicable only for exp-concave losses.

For a binary game represented by $(\{-1, 1\}, [0, 1], \ell)$ and shown in the Figure 4.6,

- If the outcome $y = -1$, the Best look ahead and the Worst look ahead will choose the predictions v_2 and v_1 and incur losses $\ell_{-1}(v_2)$ and $\ell_{-1}(v_1) = g(-1)$ respectively; and if $y = 1$, they will choose v_1 and v_2 and suffer losses $\ell_1(v_1)$ and $\ell_1(v_2) = g(1)$ respectively.

- The Inverse loss will choose the prediction v_{inv} such that $\frac{\ell_1(v_{\text{inv}})}{\ell_{-1}(v_{\text{inv}})} = \frac{g(1)}{g(-1)}$ and will incur a loss $\ell(y, v_{\text{inv}})$, and the Weighted average will choose $v_{\text{avg}} = \sum_i w^i v^i$ (where w^i and v^i are the weight and the prediction of the i-the expert respectively) and will incur a loss $\ell(y, v_{\text{avg}})$.

Further if the loss function ℓ is chosen to be the square loss (which is both 2-mixable and $\frac{1}{2}$-exp-concave), then we have $v_1 = \sqrt{g(-1)}$ (since $\ell_{-1}(v_1) = v_1^2 = g(-1)$), $v_2 = 1 - \sqrt{g(1)}$ (since $\ell_1(v_2) = (1-v_2)^2 = g(1)$), and $v_{\text{inv}} = \frac{\sqrt{g(-1)}}{\sqrt{g(-1)}+\sqrt{g(1)}}$ (since $\frac{(1-v_{\text{inv}})^2}{v_{\text{inv}}^2} = \frac{g(1)}{g(-1)}$). Thus for a binary square loss game over an outcome sequence $y_1, ..., y_T$, the cumulative losses of the Aggregating Algorithm for different choices of substitution function are given as follows:

- Best look ahead: $\sum_1^T \left(1 - \sqrt{g_t(-y_t)}\right)^2$
- Worst look ahead: $\sum_1^T g_t(y_t)$
- Inverse loss: $\sum_1^T \left(y_t - \frac{\sqrt{g_t(0)}}{\sqrt{g_t(0)}+\sqrt{g_t(1)}}\right)^2$
- Weighted average: $\sum_1^T \left(y_t - \sum_i w_i^t v_i^t\right)^2$

Some experiments are conducted on a binary square loss game to compare these substitution functions. For this, binary outcome sequences of 100 elements are generated using the Bernoulli distribution with success probabilities 0.5, 0.7, 0.9, and 1.0 (these sequences are represented by $\{y_t\}_{p=0.5}$, $\{y_t\}_{p=0.7}$, $\{y_t\}_{p=0.9}$ and $\{y_t\}_{p=1.0}$ respectively). Furthermore the following expert settings are used:

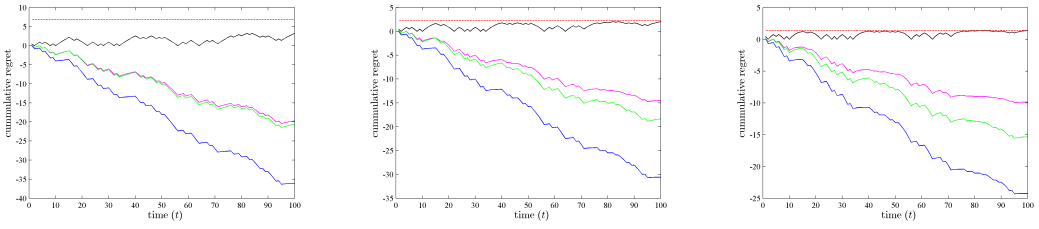
- 2 experts where one expert always make the prediction $v = 0$, and the other one always makes the prediction $v = 1$. This setting is represented by $\{E_t\}_{\text{set.1}}$.
- 3 experts where two experts are as in the previous setting, and the other one is always accurate expert. This setting is represented by $\{E_t\}_{\text{set.2}}$.
- 101 constant experts where the prediction values of the experts are from 0 to 1 with equal interval. This setting is represented by $\{E_t\}_{\text{set.3}}$.

The results of these experiments are presented in the figures 4.7,4.8,4.9, and 4.10. From these figures, it can be seen that for the expert setting $\{E_t\}_{\text{set.1}}$, the difference between the regret values of the worst look ahead and the best look ahead substitution functions relative to the theoretical regret bound is very high, whereas that relative difference is very low for the expert setting $\{E_t\}_{\text{set.3}}$. Further the performance of the Aggregating Algorithm over a real dataset is shown in the Figure 4.11. From these results for both simulated dataset (for all three expert settings) and real dataset, observe that the difference between the regret values of the inverse loss and the weighted average substitution functions relative to the theoretical regret bound is very low.

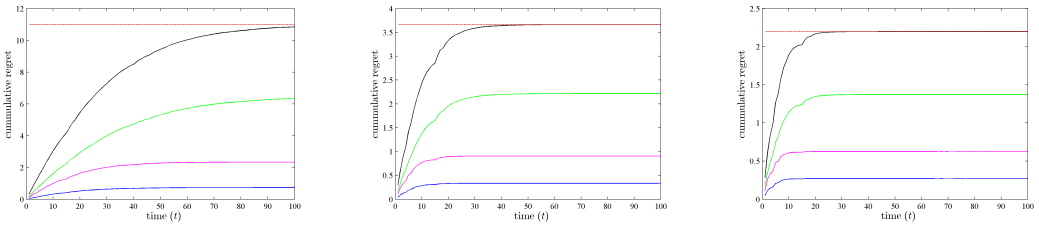
4.4.2 Probability Games with Continuous outcome space

We consider an important class of prediction problem called *probability games* (as explained by Vovk [2001]), in which the prediction v and the outcome y are probability distributions in some set (for example a finite set of the form $[n]$). A special class of

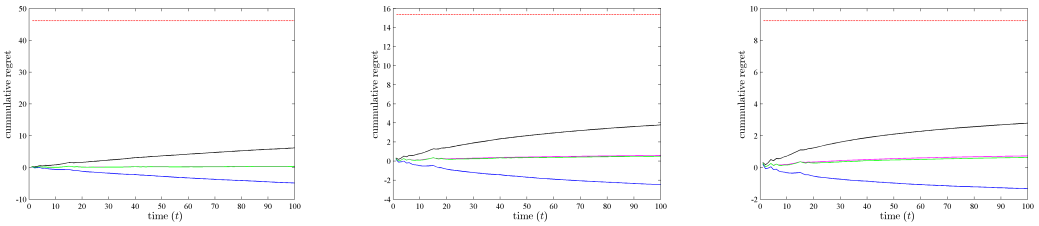
Figure 4.7: Cumulative regret of the Aggregating Algorithm over the outcome sequence $\{y_t\}_{p=0.5}$ for different choices of substitution functions (Best look ahead(—), Worst look ahead(—), Inverse loss(—), and Weighted average(—)) with learning rate η and expert setting $\{E_t\}_i$ (theoretical regret bound is shown by - - -).



(a) $\eta = 0.1, \{E_t\}_{\text{set.1}}$ (b) $\eta = 0.3, \{E_t\}_{\text{set.1}}$ (c) $\eta = 0.5, \{E_t\}_{\text{set.1}}$

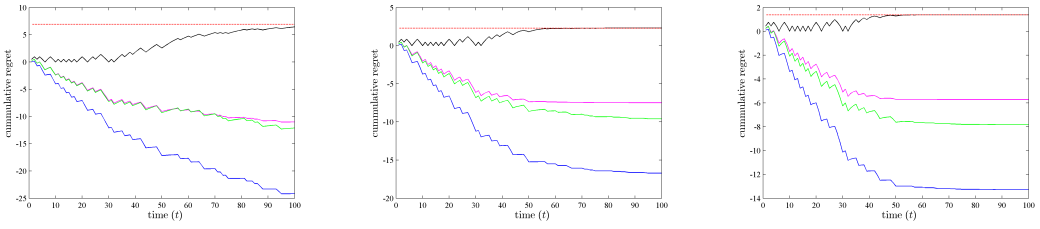


(d) $\eta = 0.1, \{E_t\}_{\text{set.2}}$ (e) $\eta = 0.3, \{E_t\}_{\text{set.2}}$ (f) $\eta = 0.5, \{E_t\}_{\text{set.2}}$

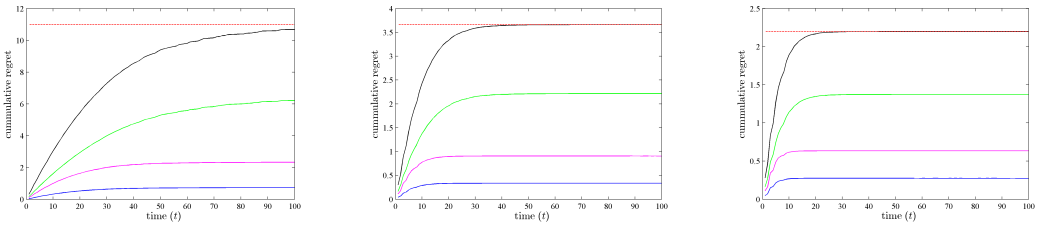


(g) $\eta = 0.1, \{E_t\}_{\text{set.3}}$ (h) $\eta = 0.3, \{E_t\}_{\text{set.3}}$ (i) $\eta = 0.5, \{E_t\}_{\text{set.3}}$

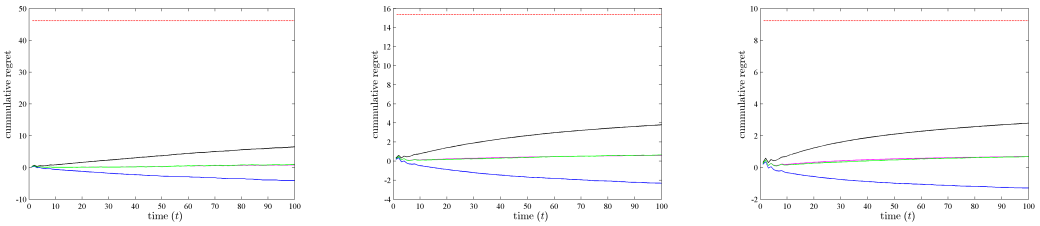
Figure 4.8: Cumulative regret of the Aggregating Algorithm over the outcome sequence $\{y_t\}_{p=0.7}$ for different choices of substitution functions (Best look ahead(—), Worst look ahead(—), Inverse loss(—), and Weighted average(—)) with learning rate η and expert setting $\{E_t\}_i$ (theoretical regret bound is shown by - - -).



(a) $\eta = 0.1, \{E_t\}_{\text{set.1}}$ (b) $\eta = 0.3, \{E_t\}_{\text{set.1}}$ (c) $\eta = 0.5, \{E_t\}_{\text{set.1}}$

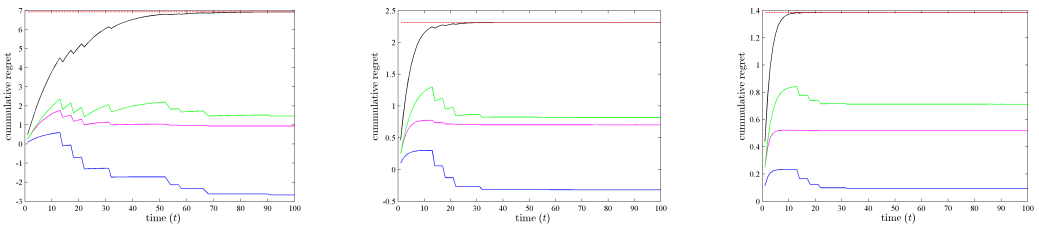


(d) $\eta = 0.1, \{E_t\}_{\text{set.2}}$ (e) $\eta = 0.3, \{E_t\}_{\text{set.2}}$ (f) $\eta = 0.5, \{E_t\}_{\text{set.2}}$

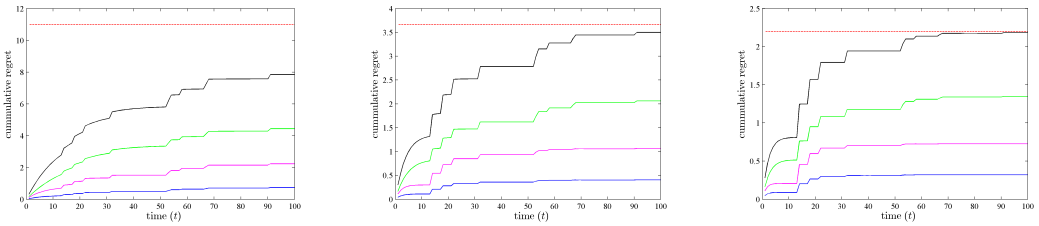


(g) $\eta = 0.1, \{E_t\}_{\text{set.3}}$ (h) $\eta = 0.3, \{E_t\}_{\text{set.3}}$ (i) $\eta = 0.5, \{E_t\}_{\text{set.3}}$

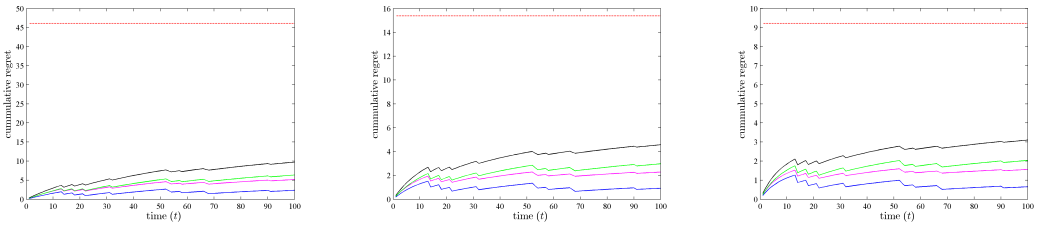
Figure 4.9: Cumulative regret of the Aggregating Algorithm over the outcome sequence $\{y_t\}_{p=0.9}$ for different choices of substitution functions (Best look ahead(—), Worst look ahead(—), Inverse loss(—), and Weighted average(—)) with learning rate η and expert setting $\{E_t\}_i$ (theoretical regret bound is shown by - - -).



(a) $\eta = 0.1, \{E_t\}_{\text{set.1}}$ (b) $\eta = 0.3, \{E_t\}_{\text{set.1}}$ (c) $\eta = 0.5, \{E_t\}_{\text{set.1}}$

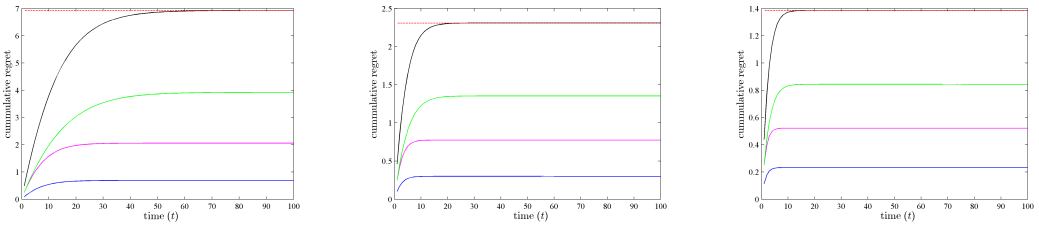


(d) $\eta = 0.1, \{E_t\}_{\text{set.2}}$ (e) $\eta = 0.3, \{E_t\}_{\text{set.2}}$ (f) $\eta = 0.5, \{E_t\}_{\text{set.2}}$

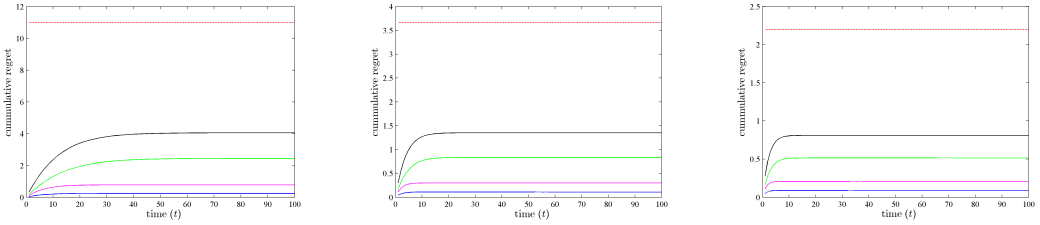


(g) $\eta = 0.1, \{E_t\}_{\text{set.3}}$ (h) $\eta = 0.3, \{E_t\}_{\text{set.3}}$ (i) $\eta = 0.5, \{E_t\}_{\text{set.3}}$

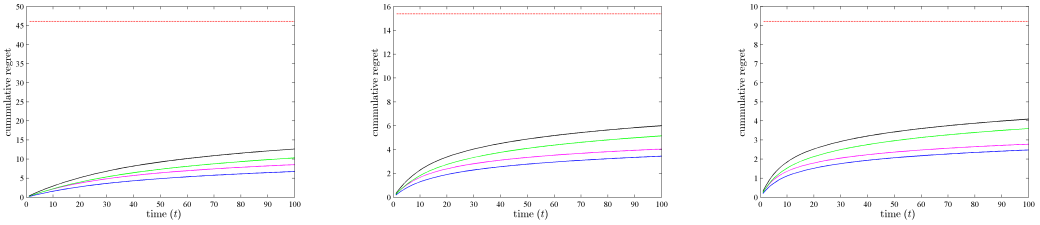
Figure 4.10: Cumulative regret of the Aggregating Algorithm over the outcome sequence $\{y_t\}_{p=1.0}$ for different choices of substitution functions (Best look ahead(—), Worst look ahead(—), Inverse loss(—), and Weighted average(—)) with learning rate η and expert setting $\{E_t\}_i$ (theoretical regret bound is shown by - - -).



(a) $\eta = 0.1, \{E_t\}_{\text{set.1}}$ (b) $\eta = 0.3, \{E_t\}_{\text{set.1}}$ (c) $\eta = 0.5, \{E_t\}_{\text{set.1}}$

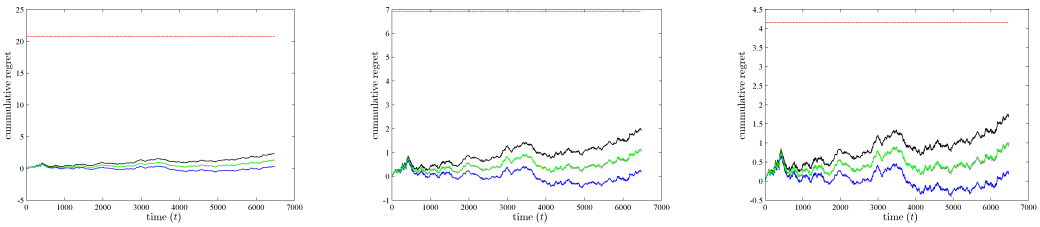


(d) $\eta = 0.1, \{E_t\}_{\text{set.2}}$ (e) $\eta = 0.3, \{E_t\}_{\text{set.2}}$ (f) $\eta = 0.5, \{E_t\}_{\text{set.2}}$



(g) $\eta = 0.1, \{E_t\}_{\text{set.3}}$ (h) $\eta = 0.3, \{E_t\}_{\text{set.3}}$ (i) $\eta = 0.5, \{E_t\}_{\text{set.3}}$

Figure 4.11: Cumulative regret of the Aggregating Algorithm over the football dataset as used by Vovk and Zhdanov [2009], for different choices of substitution functions (Best look ahead(—), Worst look ahead(—), Inverse loss(—), and Weighted average(—)) with learning rate η (theoretical regret bound is shown by - - -).



(a) $\eta = 0.1$ (b) $\eta = 0.3$ (c) $\eta = 0.5$

loss functions called *Bregman Loss Functions* (defined below) would be an appropriate choice for the probability games.

Given a differentiable convex function $\phi : \mathcal{S} \to \mathbb{R}$ defined on a convex set $\mathcal{S} \subset \mathbb{R}^d$ and two points $s_0, s \in \mathcal{S}$ the *Bregman divergence* of s from s_0 is defined as

$$B_\phi(s, s_0) := \phi(s) - \phi(s_0) - (s - s_0)' \cdot \mathsf{D}\phi(s_0),$$

where $\mathsf{D}\phi(s_0)$ is the gradient of ϕ at s_0. For any strictly convex function $\phi : \tilde{\Delta}^n \to \mathbb{R}$, differentiable over the interior of $\tilde{\Delta}^n$, the *Bregman Loss Function* (BLF, Banerjee et al. [2005]) $\ell_\phi : \Delta^n \times \Delta^n \to \mathbb{R}_+$ with generator ϕ is given by

$$\ell_\phi(y, v) := B_\phi(\tilde{y}, \tilde{v}) = \phi(\tilde{y}) - \phi(\tilde{v}) - (\tilde{y} - \tilde{v})' \cdot \mathsf{D}\phi(\tilde{v}); \quad y, v \in \Delta^n, \tag{4.29}$$

where $\tilde{y} = \Pi_\Delta(y)$, and $\tilde{v} = \Pi_\Delta(v)$. Since the conditional Bayes risk of a strictly proper loss is strictly concave, any differentiable strictly proper loss $\ell : \Delta^n \to \mathbb{R}^n_+$ will generate a BLF ℓ_ϕ with generator $\phi = -\underline{\tilde{L}}_\ell$. Further if ℓ is fair, $\ell_i(v) = \ell_\phi(e^n_i, v)$; i.e. reconstruction is possible. For example the *Kullback-Leibler loss* given by $\ell_{KL}(y, v) := \sum_{i=1}^n y(i) \log \frac{y(i)}{v(i)}$, is a BLF generated by the log loss which is strictly proper.

The following lemma (multi-class extension of a result given by Haussler et al. [1998]) provides the mixability condition for probability games.

Lemma 4.9. *For given $\ell : \Delta^n \times \Delta^n \to \mathbb{R}_+$, assume that for all $\tilde{y}, \tilde{v_1}, \tilde{v_2} \in \tilde{\Delta}^n$ (let $y = \Pi_\Delta^{-1}(\tilde{y}), v_1 = \Pi_\Delta^{-1}(\tilde{v_1})$, and $v_2 = \Pi_\Delta^{-1}(\tilde{v_2})$), the function g defined by*

$$g(\tilde{y}, \tilde{v_1}, \tilde{v_2}) = \frac{\beta}{c(\beta)} \ell(y, v_1) - \beta \ell(y, v_2) \tag{4.30}$$

satisfies

$$\mathsf{H}_{\tilde{y}} g(\tilde{y}, \tilde{v_1}, \tilde{v_2}) + \mathsf{D}_{\tilde{y}} g(\tilde{y}, \tilde{v_1}, \tilde{v_2}) \cdot (\mathsf{D}_{\tilde{y}} g(\tilde{y}, \tilde{v_1}, \tilde{v_2}))' \succcurlyeq 0. \tag{4.31}$$

If

$$\exists \tilde{v^*} \in \tilde{\Delta}^n \text{ s.t. } \ell(y, v^*) \leq -\frac{c(\beta)}{\beta} \log \int e^{-\beta \ell(y,v)} P(d\tilde{v}) \tag{4.32}$$

holds for the vertices $\tilde{y} = e^{n-1}_i, i \in [n]$, then it holds for all values $\tilde{y} \in \tilde{\Delta}^n$ (where $y = \Pi_\Delta^{-1}(\tilde{y})$, $v^* = \Pi_\Delta^{-1}(\tilde{v^*})$ and $v = \Pi_\Delta^{-1}(\tilde{v})$).

Proof. From (4.32)

$$\frac{\beta}{c(\beta)} \ell(y, v^*) + \log \int e^{-\beta \ell(y,v)} P(d\tilde{v}) \leq 0.$$

By exponentiating both sides we get

$$e^{\frac{\beta}{c(\beta)} \ell(y, v^*)} \cdot \int e^{-\beta \ell(y,v)} P(d\tilde{v}) \leq 1.$$

Denoting the left hand side of the above inequality by $f(\tilde{y})$ we have

$$f(\tilde{y}) = \int e^{g(\tilde{y},\tilde{v^*},\tilde{v})} P(d\tilde{v}).$$

Since the Hessian of f w.r.t. $\tilde{y}$ given by

$$\mathsf{H}_{\tilde{y}} f(\tilde{y}) = \int e^{g(\tilde{y},\tilde{v^*},\tilde{v})} \left(\mathsf{H}_{\tilde{y}} g(\tilde{y},\tilde{v^*},\tilde{v}) + \mathsf{D}_{\tilde{y}} g(\tilde{y},\tilde{v^*},\tilde{v}) \cdot (\mathsf{D}_{\tilde{y}} g(\tilde{y},\tilde{v^*},\tilde{v}))'\right) P(d\tilde{v})$$

is positive semi-definite (by (4.31)), $f(\tilde{y})$ is convex in $\tilde{y}$. So the maximum values of f for $\tilde{y} \in \tilde{\Delta}^n$ occurs for some $\tilde{y} = e_i^{n-1}, i \in [n]$. And by noting that, (4.32) is equivalent to $f(\tilde{y}) \leq 1$ for $\tilde{y} = e_i^{n-1}, i \in [n]$, the proof is completed. □

The next proposition shows that the mixability and exp-concavity of a strictly proper loss is carried over to the BLF generated by it.

Proposition 4.10. *For a strictly proper fair loss $\ell : \Delta^n \to \mathbb{R}_+^n$, and the BLF $\ell_\phi : \Delta^n \times \Delta^n \to \mathbb{R}_+$ generated by ℓ with $\phi = -\underline{\tilde{L}}_\ell$, if ℓ is β-mixable (resp. α-exp-concave), then ℓ_ϕ is also β-mixable (resp. α-exp-concave).*

Proof. From (4.30) and (4.29), for the BLF ℓ_ϕ we have

$$\begin{aligned}
g(\tilde{y},\tilde{v_1},\tilde{v_2}) &= \frac{\beta}{c(\beta)}\{\phi(\tilde{y}) - \phi(\tilde{v_1}) - (\tilde{y}-\tilde{v_1})' \cdot \mathsf{D}\phi(\tilde{v_1})\} \\
&\qquad - \beta\{\phi(\tilde{y}) - \phi(\tilde{v_2}) - (\tilde{y}-\tilde{v_2})' \cdot \mathsf{D}\phi(\tilde{v_2})\}, \\
\mathsf{D}_{\tilde{y}} g(\tilde{y},\tilde{v_1},\tilde{v_2}) &= \frac{\beta}{c(\beta)}\{\mathsf{D}\phi(\tilde{y}) - \mathsf{D}\phi(\tilde{v_1})\} - \beta\{\mathsf{D}\phi(\tilde{y}) - \mathsf{D}\phi(\tilde{v_2})\}, \\
\mathsf{H}_{\tilde{y}} g(\tilde{y},\tilde{v_1},\tilde{v_2}) &= \frac{\beta}{c(\beta)}\mathsf{H}\phi(\tilde{y}) - \beta\mathsf{H}\phi(\tilde{y}).
\end{aligned}$$

And since $x \cdot x' \succcurlyeq 0, \forall x \in \mathbb{R}^n$, (4.31) is satisfied for all $\tilde{y}, \tilde{v_1}, \tilde{v_2} \in \tilde{\Delta}^n$ when $c(\beta) = 1$, which is the mixability condition (in addition requiring $\tilde{v^*} = \int \tilde{v} P(d\tilde{v})$ in (4.32) is the exp-concavity condition). Then by applying Lemma 4.9 proof is completed. □

As an application of Proposition 4.10, we can see that both Kullback-Leibler loss and log loss are 1-mixable and 1-exp-concave.

4.4.3 Proofs

Proof. **(Proposition 4.1)** We first prove that the set $\psi(\Delta^n)$ is convex. For any $p, q \in \Delta^n$, by assumption there exists $c \geq 0$ and $r \in \Delta^n$ such that $\frac{1}{2}(E_\beta(\ell(p)) + E_\beta(\ell(q))) + c\mathbf{1}_n = E_\beta(\ell(r))$. Therefore $\frac{1}{2}(\psi(p) + \psi(q)) = \psi(r)$, which implies the convexity of the set $\psi(\Delta^n)$.

Let $T : \mathbb{R}^{n-1} \ni (e^{-\beta z_1} - e^{-\beta z_n}, ..., e^{-\beta z_{n-1}} - e^{-\beta z_n})' \to (e^{-\beta z_1}, ..., e^{-\beta z_n})' \in [0,1]^n$. Note this mapping from low dimension to high dimension is well defined because if there are two different z and $\overline{z}$ in $\ell(\mathcal{V})$ such that $JE_\beta(z) = JE_\beta(\overline{z})$, then there must

be $c \neq 0$ such that $E_\beta(z) + c\mathbf{1} = E_\beta(\overline{z})$. This means $z > \overline{z}$ or $z < \overline{z}$, which violates the strict properness of ℓ.

Since for any $v = \psi(p) = JE_\beta(\ell(p))$ we have $p = \ell^{-1}\left(E_\beta^{-1}(Tv)\right)$, the link ψ is invertible (ℓ is invertible if it is strictly proper (Williamson et al. [2016]), and E_β is invertible for $\beta > 0$).

Now $\ell \circ \psi^{-1}$ is β-exp-concave if for all $p, q \in \Delta^n$

$$E_\beta\left(\ell \circ \psi^{-1}\left(\frac{1}{2}(\psi(p) + \psi(q))\right)\right) \geq \frac{1}{2}E_\beta\left(\ell \circ \psi^{-1}(\psi(p))\right) + \frac{1}{2}E_\beta\left(\ell \circ \psi^{-1}(\psi(q))\right). \tag{4.33}$$

The right-hand side is obviously $\frac{1}{2}\left(E_\beta(\ell(p)) + E_\beta(\ell(q))\right)$. Let $r = \psi^{-1}\left(\frac{1}{2}(\psi(p) + \psi(q))\right) \in \Delta^n$. Then

$$JE_\beta(\ell(r)) = \psi(r) = \frac{1}{2}\left(JE_\beta(\ell(p)) + JE_\beta(\ell(q))\right).$$

Therefore $\frac{1}{2}\left(E_\beta(\ell(p)) + E_\beta(\ell(q))\right) = E_\beta(\ell(r)) + c\mathbf{1}_n$ for some $c \in \mathbb{R}$. To establish (4.33), it suffices to show $c \leq 0$. But this is guaranteed by the condition assumed. □

Proof. **(Proposition 4.2)** We first show that for a half space $H_{-p}^{\gamma_p}$ defined in (4.11) with $p \in \Delta_\epsilon^n$, $E_\beta^{-1}(H_{-p}^{\gamma_p} \cap \mathbb{R}_+^n)$ must be the super-prediction set of a scaled and shifted log loss. In fact, as $p_i > 0$, clearly $\gamma_p = \min_{x \in T_\ell} x' \cdot (-p) < 0$. Define a new loss $\tilde{\ell}_i^{\log}(q) = -\frac{1}{\beta}\log(-\frac{\gamma_p}{p_i}q_i)$ over $q \in \Delta^n$. Then $S_{\tilde{\ell}^{\log}} \subseteq E_\beta^{-1}(H_{-p}^{\gamma_p} \cap \mathbb{R}_+^n)$ can be seen from

$$\sum_i (-p_i)\exp(-\beta\tilde{\ell}_i^{\log}(q)) = \sum_i (-p_i)(-\frac{\gamma_p}{p_i}q_i) = \gamma_p. \tag{4.34}$$

Conversely, for any u such that $u_i > 0$ and $(-p)' \cdot u = \gamma_p$, simply choose $q_i = -\frac{u_i p_i}{\gamma_p}$. Then $q \in \Delta^n$ and $E_\beta(\tilde{\ell}^{\log}(q)) = u$. In summary, $E_\beta^{-1}(H_{-p}^{\gamma_p} \cap \mathbb{R}_+^n)$ is the super-prediction set of $\tilde{\ell}^{\log}$.

To prove Proposition 4.2, we first show that for any point $a \in T_\ell^\epsilon$ and any direction d from the relative interior of the positive orthant (which includes the $\mathbf{1}_n$ direction), the ray $\{a + rd : r \geq 0\}$ will be blocked by a boundary point of T_ℓ^ϵ. This is because by the definition of T_ℓ^ϵ in (4.12), the largest value of r to guarantee $a + rd \in T_\ell^\epsilon$ can be computed by

$$r^* := \sup\{r \geq 0 : a + rd \in T_\ell^\epsilon\} = \sup\{r \geq 0 : (a + rd)' \cdot (-p) \geq \gamma_p, \forall p \in \Delta_\epsilon^n\} \tag{4.35}$$

must be finite and attained. Denote $x = a + r^*d$, which must be on the boundary of T_ℓ^ϵ because

$$-x' \cdot p \geq \gamma_p, \text{ for all } p \in \Delta_\epsilon^n, \tag{4.36}$$

$$\text{and } -x' \cdot p^* = \gamma_{p^*} \text{ for some } p^* \in \Delta_\epsilon^n \text{ (not necessarily unique).} \tag{4.37}$$

In order to prove the first statement of Proposition 4.2, it suffices to show that for any point x on the north-east boundary of T_ℓ^ϵ, there exists a $q \in \Delta^n$ such that $E_\beta(\tilde{\ell}_\epsilon(q)) = x$. Suppose x satisfies (4.36) and (4.37). Then consider the (shifted/scaled) log loss $\tilde{\ell}^{\log}$ that corresponds to $H_{p^*}^{\gamma_{p^*}}$. Because log loss is strictly proper, there must be a unique

$q \in \Delta^n$ such that the hyperplane $H_0 := \{z : q' \cdot z = q' \cdot E_\beta^{-1}(x)\}$ supports the super-prediction set of $\tilde{\ell}^{\log}$ (i.e., $E_\beta^{-1}(H_{-p^*}^{\gamma_{p^*}} \cap \mathbb{R}_+^n)$) at $E_\beta^{-1}(x)$. Since $E_\beta^{-1}(T_\ell^\epsilon \cap \mathbb{R}_+^n)$ is a convex subset of $E_\beta^{-1}(H_{-p^*}^{\gamma_{p^*}} \cap \mathbb{R}_+^n)$, this hyperplane also supports $E_\beta^{-1}(T_\ell^\epsilon \cap \mathbb{R}_+^n)$ at $E_\beta^{-1}(x)$. Therefore $E_\beta^{-1}(x)$ is an optimal solution to the problem in the definition of $\tilde{\ell}_\epsilon(q)$ in (4.13). Finally observe that it must be the unique optimal solution, because if there were another solution which also lies on H_0, then by the convexity of the super-prediction set of $\tilde{\ell}_\epsilon$, the line segment between them must also lie on the prediction set of $\tilde{\ell}_\epsilon$. This violates the mixability condition of $\tilde{\ell}_\epsilon$, because by construction its sub-exp-prediction set is convex.

In order to check where $\ell(p) = \tilde{\ell}_\epsilon(p)$, a sufficient condition is that the normal direction d on the exp-prediction set evaluated at $E_\beta(\ell(p))$ satisfies $d_i / \sum_j d_j > \epsilon$. Simple calculus shows that $d_i \propto p_i \exp(\beta \ell_i(p))$. Therefore as long as p is in the relative interior of Δ^n, $d_i / \sum_j d_j > \epsilon$ can always be satisfied by choosing a sufficiently small ϵ. And for each fixed ϵ, the set S_ϵ mentioned in the theorem consists exactly of all such p that satisfies this condition. □

Proof. **(Proposition 4.5)** When $n = 2$, (4.17), (4.18) and (4.19) and the positivity of $\tilde{\psi}'$ simplify (4.20) to the two conditions:

$$\begin{aligned} (1-\tilde{p})\, k'(\tilde{p}) &\leq k(\tilde{p}) - \alpha\, \tilde{\psi}'(\tilde{p})\, k(\tilde{p})^2\, (1-\tilde{p})^2, \\ -\tilde{p}\, k'(\tilde{p}) &\leq k(\tilde{p}) - \alpha\, \tilde{\psi}'(\tilde{p})\, k(\tilde{p})^2\, \tilde{p}^2, \end{aligned}$$

for all $\tilde{p} \in (0,1)$. These two conditions can be merged as follows

$$-\frac{1}{\tilde{p}} + \alpha\, \tilde{\psi}'(\tilde{p})\, k(\tilde{p})\, \tilde{p} \;\leq\; \frac{k'(\tilde{p})}{k(\tilde{p})} \;\leq\; \frac{1}{1-\tilde{p}} - \alpha\, \tilde{\psi}'(\tilde{p})\, k(\tilde{p})\, (1-\tilde{p}), \quad \forall \tilde{p} \in (0,1).$$

By noting that $k(\tilde{p}) = \frac{w(\tilde{p})}{\tilde{\psi}'(\tilde{p})}$ and $k'(\tilde{p}) = \frac{w'(\tilde{p})\tilde{\psi}'(\tilde{p}) - w(\tilde{p})\tilde{\psi}''(\tilde{p})}{\tilde{\psi}'(\tilde{p})^2}$ completes the proof. □

Proof. **(Proposition 4.6)** Let $g(\tilde{p}) = \frac{1}{w(\tilde{p})}$ and so $g'(\tilde{p}) = -\frac{1}{w(\tilde{p})^2} w'(\tilde{p})$, $g(v) = \int_{\frac{1}{2}}^{v} g'(\tilde{p}) d\tilde{p} + g(\frac{1}{2})$ and $g(\frac{1}{2}) = \frac{1}{w(\frac{1}{2})} = 1$. By dividing all sides of (4.21) by $-w(\tilde{p})$ and applying the substitution we get,

$$\frac{1}{\tilde{p}} g(\tilde{p}) - \alpha\tilde{p} \;\geq\; g'(\tilde{p}) - \Phi_{\tilde{\psi}}(\tilde{p}) g(\tilde{p}) \;\geq\; -\frac{1}{1-\tilde{p}} g(\tilde{p}) + \alpha(1-\tilde{p}), \quad \forall \tilde{p} \in (0,1), \quad (4.38)$$

where $\Phi_{\tilde{\psi}}(\tilde{p}) := -\frac{\tilde{\psi}''(\tilde{p})}{\tilde{\psi}'(\tilde{p})}$. If we take the first inequality of (4.38) and rearrange it we obtain,

$$-\alpha \geq \left(g'(\tilde{p}) \frac{1}{\tilde{p}} - g(\tilde{p}) \frac{1}{\tilde{p}^2} \right) - \Phi_{\tilde{\psi}}(\tilde{p}) \frac{g(\tilde{p})}{\tilde{p}} = \left(\frac{g(\tilde{p})}{\tilde{p}} \right)' - \Phi_{\tilde{\psi}}(\tilde{p}) \left(\frac{g(\tilde{p})}{\tilde{p}} \right), \quad \forall \tilde{p} \in (0,1). \quad (4.39)$$

Multiplying (4.39) by $e^{-\int_0^{\tilde{p}} \Phi_{\tilde{\psi}}(t)dt}$ will result in,

$$\begin{aligned} -\alpha e^{-\int_0^{\tilde{p}} \Phi_{\tilde{\psi}}(t)dt} &\geq \left(\frac{g(\tilde{p})}{\tilde{p}}\right)' e^{-\int_0^{\tilde{p}} \Phi_{\tilde{\psi}}(t)dt} + \left(\frac{g(\tilde{p})}{\tilde{p}}\right) e^{-\int_0^{\tilde{p}} \Phi_{\tilde{\psi}}(t)dt}(-\Phi_{\tilde{\psi}}(\tilde{p})) \\ &= \left(\frac{g(\tilde{p})}{\tilde{p}} e^{-\int_0^{\tilde{p}} \Phi_{\tilde{\psi}}(t)dt}\right)', \quad \forall \tilde{p} \in (0,1). \end{aligned} \tag{4.40}$$

Since

$$-\int_0^{\tilde{p}} \Phi_{\tilde{\psi}}(t)dt = -\int_0^{\tilde{p}} -\frac{\tilde{\psi}''(t)}{\tilde{\psi}'(t)}dt = \int_0^{\tilde{p}} (\log \tilde{\psi}'(t))'dt = \log \frac{\tilde{\psi}'(\tilde{p})}{\tilde{\psi}'(0)},$$

(4.40) is reduced to

$$\begin{aligned} -\alpha \frac{\tilde{\psi}'(\tilde{p})}{\tilde{\psi}'(0)} &\geq \left(\frac{g(\tilde{p})}{\tilde{p}} \frac{\tilde{\psi}'(\tilde{p})}{\tilde{\psi}'(0)}\right)', \quad \forall \tilde{p} \in (0,1) \\ \Rightarrow -\alpha \tilde{\psi}'(\tilde{p}) &\geq \left(\frac{g(\tilde{p})}{\tilde{p}} \tilde{\psi}'(\tilde{p})\right)', \quad \forall \tilde{p} \in (0,1). \end{aligned}$$

For $v \geq \frac{1}{2}$ we thus have

$$\begin{aligned} -\alpha \int_{\frac{1}{2}}^{v} \tilde{\psi}'(\tilde{p})d\tilde{p} &\geq \int_{\frac{1}{2}}^{v} \left(\frac{g(\tilde{p})}{\tilde{p}} \tilde{\psi}'(\tilde{p})\right)' d\tilde{p}, \quad \forall v \in [1/2, 1) \\ \Rightarrow -\alpha(\tilde{\psi}(v) - \tilde{\psi}(\frac{1}{2})) &\geq \left(\frac{g(v)}{v}\tilde{\psi}'(v) - \frac{g(\frac{1}{2})}{\frac{1}{2}}\tilde{\psi}'(\frac{1}{2})\right) \\ &= \left(\frac{1}{w(v)v}\tilde{\psi}'(v) - 2\tilde{\psi}'(\frac{1}{2})\right), \quad \forall v \in [1/2, 1) \\ \Rightarrow w(v) &\geq \frac{\tilde{\psi}'(v)}{v(2\tilde{\psi}'(\frac{1}{2}) - \alpha(\tilde{\psi}(v) - \tilde{\psi}(\frac{1}{2})))}, \quad \forall v \in [1/2, 1). \end{aligned}$$

Also by considering $v \leq \frac{1}{2}$ case as above, we get

$$\frac{\tilde{\psi}'(v)}{v(2\tilde{\psi}'(\frac{1}{2}) - \alpha(\tilde{\psi}(v) - \tilde{\psi}(\frac{1}{2})))} \lesseqgtr w(v), \quad \forall v \in (0,1).$$

Finally by following the similar steps for the second inequality of (4.38), the proof will be completed. □

Here we provide an integral inequalities related result (without proof) due to Beesack and presented in Dragomir [2000].

Theorem 4.11. *Let y and k be continuous and f and g Riemann integrable functions on $J = [\alpha, \beta]$ with g and k nonnegative on J. If*

$$y(x) \geq f(x) + g(x) \int_{\alpha}^{x} y(t)k(t)dt, \quad x \in J,$$

then

$$y(x) \;\geq\; f(x) + g(x)\int_{\alpha}^{x} f(t)k(t)\exp\left(\int_{t}^{x} g(r)k(r)dr\right)dt, \quad x \in J.$$

The result remains valid if $\int_{\alpha}^{x}$ *is replaced by* $\int_{x}^{\beta}$ *and* $\int_{t}^{x}$ *by* $\int_{x}^{t}$ *throughout.*

Using the above theorem, we get the following simplified test for the conditions in Theorem 4.7:

$$-\alpha + \frac{\alpha}{2\tilde{p}^2} - \frac{2}{\tilde{p}^2} \;\leq\; a(\tilde{p}) \;\implies\; \left[\frac{\alpha(1-\tilde{p})}{\tilde{p}} - \frac{2}{\tilde{p}(1-\tilde{p})}\right] + \frac{1}{\tilde{p}(1-\tilde{p})}\int_{\tilde{p}}^{1/2} a(t)dt \;\leq\; a(\tilde{p})$$

and

$$\frac{\alpha\tilde{p}}{(1-\tilde{p})} + \frac{2\alpha\tilde{p}-\alpha-4}{2(1-\tilde{p})^2} \;\leq\; b(\tilde{p}) \;\implies\; \left[\frac{\alpha\tilde{p}}{(1-\tilde{p})} - \frac{2}{\tilde{p}(1-\tilde{p})}\right] + \frac{1}{\tilde{p}(1-\tilde{p})}\int_{1/2}^{\tilde{p}} b(t)dt \;\leq\; b(\tilde{p}).$$

Above two identities are proved in the following proof of Theorem 4.7.

Proof. **(Theorem 4.7)** The necessary *and* sufficient condition for the exp-concavity of proper losses is given by (4.23). But from (4.39), we can see that (4.23) is equivalent to

$$\left(\frac{g(\tilde{p})}{\tilde{p}}\right)' \leq -\alpha, \quad \forall\tilde{p} \in (0,1), \tag{4.41}$$

and

$$-\left(\frac{g(\tilde{p})}{1-\tilde{p}}\right)' \leq -\alpha, \quad \forall\tilde{p} \in (0,1), \tag{4.42}$$

where $g(\tilde{p}) = \frac{1}{w(\tilde{p})}$ with $w(\frac{1}{2}) = 1$; i.e. (4.23) if and only if (4.41) & (4.42).

Now if we choose the weight function $w(\tilde{p})$ as follows

$$w(\tilde{p}) = \frac{1}{\tilde{p}\left(2 + \int_{1/2}^{\tilde{p}} a(t)dt\right)}, \tag{4.43}$$

such that $a(t) \leq -\alpha$, then (4.41) will be satisfied (since (4.43) $\implies 2 + \int_{1/2}^{\tilde{p}} a(t)dt = \frac{1}{w(\tilde{p})\tilde{p}} = \frac{g(\tilde{p})}{\tilde{p}} \implies a(\tilde{p}) = \left(\frac{g(\tilde{p})}{\tilde{p}}\right)'$). Similarly the weight function $w(\tilde{p})$ given by

$$w(\tilde{p}) = \frac{1}{(1-\tilde{p})\left(2 - \int_{1/2}^{\tilde{p}} b(t)dt\right)}, \tag{4.44}$$

with $b(t) \leq -\alpha$ will satisfy (4.42) (since (4.44) $\implies 2 - \int_{1/2}^{\tilde{p}} b(t)dt = \frac{1}{w(\tilde{p})(1-\tilde{p})} = \frac{g(\tilde{p})}{1-\tilde{p}} \implies -b(\tilde{p}) = \left(\frac{g(\tilde{p})}{1-\tilde{p}}\right)'$). To satisfy both (4.41) and (4.42) at the same time (then obviously (4.23) will be satisfied), we can make the two forms of the weight function ((4.43) and (4.44)) equivalent with the appropriate choice of $a(t)$ and $b(t)$. This can be done in two cases.

In the first case, for $\tilde{p} \in (0, 1/2]$ we can fix the weight function $w(\tilde{p})$ as given by (4.43) and choose $a(t)$ such that,

- $a(t) \leq -\alpha$ (then (4.41) is satisfied) and
- (4.43) = (4.44) $\implies$ $b(t) \leq -\alpha$ (then (4.42) is satisfied).

But (4.43) = (4.44) for all $\tilde{p} \in (0, 1/2]$ if and only if

$$\begin{aligned}
& \tilde{p}\left(2 - \int_{\tilde{p}}^{1/2} a(t)dt\right) = (1-\tilde{p})\left(2 + \int_{\tilde{p}}^{1/2} b(t)dt\right), \quad \tilde{p} \in (0, 1/2] \\
\iff \quad & \frac{\tilde{p}}{1-\tilde{p}}\left(2 - \int_{\tilde{p}}^{1/2} a(t)dt\right) - 2 = \int_{\tilde{p}}^{1/2} b(t)dt, \quad \tilde{p} \in (0, 1/2] \\
\iff \quad & \frac{\tilde{p}}{1-\tilde{p}}a(\tilde{p}) + \frac{1}{(1-\tilde{p})^2}\left(2 - \int_{\tilde{p}}^{1/2} a(t)dt\right) = -b(\tilde{p}), \quad \tilde{p} \in (0, 1/2],
\end{aligned}$$

where the last step is obtained by differentiating both sides w.r.t $\tilde{p}$. Thus the constraint (4.43) = (4.44) $\implies$ $b(t) \leq -\alpha$ can be given as

$$\begin{aligned}
& \frac{\tilde{p}}{1-\tilde{p}}a(\tilde{p}) + \frac{1}{(1-\tilde{p})^2}\left(2 - \int_{\tilde{p}}^{1/2} a(t)dt\right) \;\geq\; \alpha, \quad \tilde{p} \in (0, 1/2] \\
\iff \quad & a(\tilde{p}) \;\geq\; \left[\frac{\alpha(1-\tilde{p})}{\tilde{p}} - \frac{2}{\tilde{p}(1-\tilde{p})}\right] + \frac{1}{\tilde{p}(1-\tilde{p})}\int_{\tilde{p}}^{1/2} a(t)dt, \quad \tilde{p} \in (0, 1/2] \\
\iff \quad & a(\tilde{p}) \;\geq\; f(\tilde{p}) + g(\tilde{p})\int_{\tilde{p}}^{1/2} a(t)k(t)dt, \quad \tilde{p} \in (0, 1/2],
\end{aligned}$$

where $f(\tilde{p}) = \left[\frac{\alpha(1-\tilde{p})}{\tilde{p}} - \frac{2}{\tilde{p}(1-\tilde{p})}\right]$, $g(\tilde{p}) = \frac{1}{\tilde{p}(1-\tilde{p})} = \frac{1}{\tilde{p}} + \frac{1}{(1-\tilde{p})}$ and $k(t) = 1$. Now by applying Theorem 4.11 we have

$$a(\tilde{p}) \;\geq\; f(\tilde{p}) + g(\tilde{p})\int_{\tilde{p}}^{1/2} f(t)k(t)\exp\left(\int_{\tilde{p}}^{t} g(r)k(r)dr\right)dt, \quad \tilde{p} \in (0, 1/2].$$

Since

$$\int_{\tilde{p}}^{t} g(r)k(r)dr = \int_{\tilde{p}}^{t} \frac{1}{r} + \frac{1}{(1-r)}dr = \left[\ln r - \ln(1-r)\right]_{\tilde{p}}^{t} = \ln\left(\frac{t}{(1-t)}\frac{(1-\tilde{p})}{\tilde{p}}\right),$$

$$\begin{aligned}
\int_{\tilde{p}}^{1/2} f(t)k(t)\exp\left(\int_{\tilde{p}}^{t} g(r)k(r)dr\right)dt &= \int_{\tilde{p}}^{1/2} f(t)\frac{t}{(1-t)}\frac{(1-\tilde{p})}{\tilde{p}}dt \\
&= \frac{(1-\tilde{p})}{\tilde{p}}\int_{\tilde{p}}^{1/2}\left[\frac{\alpha(1-t)}{t} - \frac{2}{t(1-t)}\right]\frac{t}{(1-t)}dt \\
&= \frac{(1-\tilde{p})}{\tilde{p}}\int_{\tilde{p}}^{1/2} \alpha - \frac{2}{(1-t)^2}dt \\
&= \frac{(1-\tilde{p})}{\tilde{p}}\left[\alpha t - \frac{2}{1-t}\right]_{\tilde{p}}^{1/2} \\
&= \frac{(1-\tilde{p})}{\tilde{p}}\left[\frac{\alpha}{2} - 4 - \alpha\tilde{p} + \frac{2}{1-\tilde{p}}\right],
\end{aligned}$$

we get

$$\begin{aligned} a(\tilde{p}) &\geq \left[\frac{\alpha(1-\tilde{p})}{\tilde{p}} - \frac{2}{\tilde{p}(1-\tilde{p})}\right] + \frac{1}{\tilde{p}(1-\tilde{p})}\frac{(1-\tilde{p})}{\tilde{p}}\left[\frac{\alpha}{2} - 4 - \alpha\tilde{p} + \frac{2}{1-\tilde{p}}\right], \quad \tilde{p} \in (0, 1/2] \\ &= -\alpha + \frac{\alpha}{2\tilde{p}^2} - \frac{2}{\tilde{p}^2}, \quad \tilde{p} \in (0, 1/2]. \end{aligned}$$

Similarly in the second case, for $\tilde{p} \in [1/2, 1)$ we can fix the weight function $w(\tilde{p})$ as given by (4.44) and choose $b(t)$ such that,

- $b(t) \leq -\alpha$ (then (4.42) is satisfied) and
- (4.43) = (4.44) $\implies a(t) \leq -\alpha$ (then (4.41) is satisfied).

But (4.43) = (4.44) for all $\tilde{p} \in [1/2, 1)$ if and only if

$$\begin{aligned} & \tilde{p}\left(2 + \int_{1/2}^{\tilde{p}} a(t)dt\right) = (1-\tilde{p})\left(2 - \int_{1/2}^{\tilde{p}} b(t)dt\right), \quad \tilde{p} \in [1/2, 1) \\ \iff & \int_{\tilde{p}}^{1/2} a(t)dt = \frac{1-\tilde{p}}{\tilde{p}}\left(2 - \int_{1/2}^{\tilde{p}} b(t)dt\right) - 2, \quad \tilde{p} \in [1/2, 1) \\ \iff & a(\tilde{p}) = -\frac{1-\tilde{p}}{\tilde{p}}b(\tilde{p}) - \frac{1}{\tilde{p}^2}\left(2 - \int_{1/2}^{\tilde{p}} b(t)dt\right), \quad \tilde{p} \in [1/2, 1), \end{aligned}$$

where the last step is obtained by differentiating both sides w.r.t $\tilde{p}$. Thus the constraint (4.43) = (4.44) $\implies a(t) \leq -\alpha$ can be given as

$$\begin{aligned} & \frac{1-\tilde{p}}{\tilde{p}}b(\tilde{p}) + \frac{1}{\tilde{p}^2}\left(2 - \int_{1/2}^{\tilde{p}} b(t)dt\right) \geq \alpha, \quad \tilde{p} \in [1/2, 1) \\ \iff & b(\tilde{p}) \geq \left[\frac{\alpha\tilde{p}}{(1-\tilde{p})} - \frac{2}{\tilde{p}(1-\tilde{p})}\right] + \frac{1}{\tilde{p}(1-\tilde{p})}\int_{1/2}^{\tilde{p}} b(t)dt, \quad \tilde{p} \in [1/2, 1) \\ \iff & b(\tilde{p}) \geq f(\tilde{p}) + g(\tilde{p})\int_{1/2}^{\tilde{p}} b(t)k(t)dt, \quad \tilde{p} \in [1/2, 1), \end{aligned}$$

where $f(\tilde{p}) = \left[\frac{\alpha\tilde{p}}{(1-\tilde{p})} - \frac{2}{\tilde{p}(1-\tilde{p})}\right]$, $g(\tilde{p}) = \frac{1}{\tilde{p}(1-\tilde{p})} = \frac{1}{\tilde{p}} + \frac{1}{(1-\tilde{p})}$ and $k(t) = 1$. Again by applying Theorem 4.11 we have

$$b(\tilde{p}) \geq f(\tilde{p}) + g(\tilde{p})\int_{1/2}^{\tilde{p}} f(t)k(t)\exp\left(\int_t^{\tilde{p}} g(r)k(r)dr\right)dt, \quad \tilde{p} \in [1/2, 1).$$

Since

$$\int_t^{\tilde{p}} g(r)k(r)dr = \int_t^{\tilde{p}} \frac{1}{r} + \frac{1}{(1-r)}dr = [\ln r - \ln(1-r)]_t^{\tilde{p}} = \ln\left(\frac{\tilde{p}}{(1-\tilde{p})}\frac{(1-t)}{t}\right),$$

$$\begin{aligned} \int_{1/2}^{\tilde{p}} f(t)k(t)\exp\left(\int_t^{\tilde{p}} g(r)k(r)dr\right)dt &= \int_{1/2}^{\tilde{p}} f(t)\frac{\tilde{p}}{(1-\tilde{p})}\frac{(1-t)}{t}dt \\ &= \frac{\tilde{p}}{(1-\tilde{p})}\int_{1/2}^{\tilde{p}}\left[\frac{\alpha t}{(1-t)} - \frac{2}{t(1-t)}\right]\frac{(1-t)}{t}dt \end{aligned}$$

$$
\begin{aligned}
&= \frac{\tilde{p}}{(1-\tilde{p})}\int_{1/2}^{\tilde{p}} \alpha - \frac{2}{t^2}dt \\
&= \frac{\tilde{p}}{(1-\tilde{p})}\left[\alpha t + \frac{2}{t}\right]_{1/2}^{\tilde{p}} \\
&= \frac{\tilde{p}}{(1-\tilde{p})}\left[\alpha\tilde{p} + \frac{2}{\tilde{p}} - \frac{\alpha}{2} - 4\right],
\end{aligned}
$$

we get

$$
\begin{aligned}
b(\tilde{p}) &\geq \left[\frac{\alpha\tilde{p}}{(1-\tilde{p})} - \frac{2}{\tilde{p}(1-\tilde{p})}\right] + \frac{1}{\tilde{p}(1-\tilde{p})}\frac{\tilde{p}}{(1-\tilde{p})}\left[\alpha\tilde{p} + \frac{2}{\tilde{p}} - \frac{\alpha}{2} - 4\right], \quad \tilde{p} \in [1/2, 1) \\
&= \frac{\alpha\tilde{p}}{(1-\tilde{p})} + \frac{\alpha\tilde{p}}{(1-\tilde{p})^2} - \frac{\alpha}{2(1-\tilde{p})^2} - \frac{2}{(1-\tilde{p})^2}, \quad \tilde{p} \in [1/2, 1).
\end{aligned}
$$

□

Proof. **(Corollary 4.8)** We have to show that $\tilde{\psi}_\ell^*$ will satisfy (4.21) with $\alpha = \beta$, for all β-mixable proper loss functions. Since

$$
\frac{(\tilde{\psi}_\ell^*)''(\tilde{p})}{(\tilde{\psi}_\ell^*)'(\tilde{p})} = \frac{\frac{w_\ell'(\tilde{p})w_{\ell^{\log}}(\tilde{p}) - w_\ell(\tilde{p})w_{\ell^{\log}}'(\tilde{p})}{w_{\ell^{\log}}(\tilde{p})^2}}{\frac{w_\ell(\tilde{p})}{w_{\ell^{\log}}(\tilde{p})}} = \frac{w_\ell'(\tilde{p})}{w_\ell(\tilde{p})} - \frac{w_{\ell^{\log}}'(\tilde{p})}{w_{\ell^{\log}}(\tilde{p})} = \frac{w_\ell'(\tilde{p})}{w_\ell(\tilde{p})} - (\log w_{\ell^{\log}}(\tilde{p}))', \tag{4.45}
$$

by substituting $\tilde{\psi} = \tilde{\psi}_\ell^*$ and $\alpha = \beta$ in (4.21) we have,

$$
\begin{aligned}
& -\frac{1}{\tilde{p}} + \beta w_\ell(\tilde{p})\tilde{p} \leq \frac{w_\ell'(\tilde{p})}{w_\ell(\tilde{p})} - \frac{(\tilde{\psi}_\ell^*)''(\tilde{p})}{(\tilde{\psi}_\ell^*)'(\tilde{p})} \leq \frac{1}{1-\tilde{p}} - \beta w_\ell(\tilde{p})(1-\tilde{p}), \quad \forall\tilde{p} \in (0,1) \\
\iff & -\frac{1}{\tilde{p}} + \beta w_\ell(\tilde{p})\tilde{p} \leq (\log w_{\ell^{\log}}(\tilde{p}))' \leq \frac{1}{1-\tilde{p}} - \beta w_\ell(\tilde{p})(1-\tilde{p}), \quad \forall\tilde{p} \in (0,1) \\
\iff & -\frac{1}{\tilde{p}} + \beta w_\ell(\tilde{p})\tilde{p} \leq -\frac{1}{\tilde{p}} + \frac{1}{1-\tilde{p}} \leq \frac{1}{1-\tilde{p}} - \beta w_\ell(\tilde{p})(1-\tilde{p}), \quad \forall\tilde{p} \in (0,1) \\
\iff & \beta \leq \frac{1}{\tilde{p}(1-\tilde{p})w_\ell(\tilde{p})} = \frac{w_{\ell^{\log}}(\tilde{p})}{w_\ell(\tilde{p})}, \quad \forall\tilde{p} \in (0,1),
\end{aligned}
$$

which is true for all β-mixable binary proper loss functions. From (4.45)

$$
\begin{aligned}
& \left(\log(\tilde{\psi}_\ell^*)'(\tilde{p})\right)' = (\log w_\ell(\tilde{p}))' - (\log w_{\ell^{\log}}(\tilde{p}))' = \left(\log\frac{w_\ell(\tilde{p})}{w_{\ell^{\log}}(\tilde{p})}\right)', \\
\Rightarrow & \left[\log(\tilde{\psi}_\ell^*)'(\tilde{p})\right]_{1/2}^{\tilde{p}} = \left[\log\frac{w_\ell(\tilde{p})}{w_{\ell^{\log}}(\tilde{p})}\right]_{1/2}^{\tilde{p}}, \\
\Rightarrow & \log(\tilde{\psi}_\ell^*)'(\tilde{p}) - \log(\tilde{\psi}_\ell^*)'\left(\frac{1}{2}\right) = \log\frac{w_\ell(\tilde{p})}{w_{\ell^{\log}}(\tilde{p})} \cdot \frac{w_{\ell^{\log}}(\frac{1}{2})}{w_\ell(\frac{1}{2})},
\end{aligned}
$$

it can be seen that a design choice of $(\tilde{\psi}_\ell^*)'\left(\frac{1}{2}\right) = 1$ is made in the construction of this link function. □

4.4.4 Squared Loss

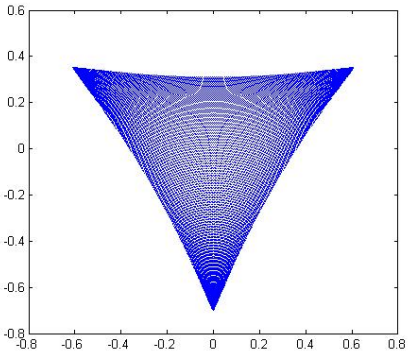


Figure 4.12: Projection of the exp-prediction set of square loss ($\beta = 1$) along the $\mathbf{1}_3$ direction. By the apparent lack of convexity of the projection, the condition $\partial_{\mathbf{1}_n}\mathcal{B}_\beta \subseteq E_\beta(\ell(\mathcal{V}))$ in Proposition 4.1 does not hold in this case.

In this section we will consider the multi-class squared loss with partial losses given by $\ell_i^{\text{sq}}(p) := \sum_{j\in[n]}(\llbracket i = j \rrbracket - p_j)^2$. The Bayes risk of this loss is $\underline{\tilde{L}}_{\ell^{\text{sq}}}(\tilde{p}) = 1 - \sum_{i=1}^{n-1} p_i^2 - (1 - \sum_{i=1}^{n-1} p_i)^2$. Thus the Hessian of the Bayes risk is given by

$$\mathsf{H}\underline{\tilde{L}}_{\ell^{\text{sq}}}(\tilde{p}) = 2\begin{pmatrix} -2 & -1 & \cdots & -1 \\ -1 & -2 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & -2 \end{pmatrix}.$$

For the identity link, from (4.16) we get $k_{\text{id}}(\tilde{p}) = -\mathsf{H}\underline{\tilde{L}}_{\ell^{\text{sq}}}(\tilde{p})$ and $\mathsf{D}_v[k_{\text{id}}(\tilde{p})] = 0$ since $\mathsf{D}\tilde{\psi}(\tilde{p}) = I_{n-1}$. Thus from (4.20), the multi-class squared loss is α-exp-concave (with $\alpha > 0$) if and only if for all $\tilde{p} \in \mathring{\Delta}^n$ and for all $i \in [n]$

$$\begin{aligned} & 0 \preccurlyeq k_{\text{id}}(\tilde{p}) - \alpha k_{\text{id}}(\tilde{p}) \cdot (e_i^{n-1} - \tilde{p}) \cdot (e_i^{n-1} - \tilde{p})' \cdot k_{\text{id}}(\tilde{p}) \\ \iff & k_{\text{id}}(\tilde{p})^{-1} \succcurlyeq \alpha(e_i^{n-1} - \tilde{p}) \cdot (e_i^{n-1} - \tilde{p})'. \end{aligned} \tag{4.46}$$

Similarly for the canonical link, from (4.25) and (4.26), the composite loss is α-exp-concave (with $\alpha > 0$) if and only if for all $\tilde{p} \in \mathring{\Delta}^n$ and for all $i \in [n]$

$$k_{\text{id}}(\tilde{p})^{-1} = -[\mathsf{H}\underline{\tilde{L}}_{\ell^{\text{sq}}}(\tilde{p})]^{-1} \succcurlyeq \alpha(e_i^{n-1} - \tilde{p}) \cdot (e_i^{n-1} - \tilde{p})'. \tag{4.47}$$

From (4.46) and (4.47), it can be seen that for the multi-class squared loss the level of exp-concavification by identity link and canonical link are same. When $n = 2$, since $k_{\text{id}}(\tilde{p}) = 4$, the condition (4.46) is equivalent to

$$\frac{1}{4} \geq \alpha(e_i^{n-1} - \tilde{p}) \cdot (e_i^{n-1} - \tilde{p})', \quad i \in [2], \forall \tilde{p} \in (0, 1)$$

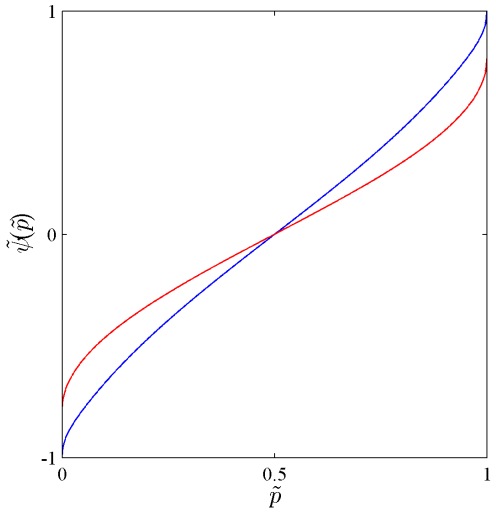


Figure 4.13: Exp-concavifying link functions for binary boosting loss constructed by Proposition 4.1 (—) and Corollary 4.8 (—).

$$\iff \quad \alpha \leq \frac{1}{4\tilde{p}^2} \quad \text{and} \quad \alpha \leq \frac{1}{4(1-\tilde{p})^2}, \quad \forall \tilde{p} \in (0,1)$$
$$\iff \quad \alpha \leq \frac{1}{4}.$$

When $n = 3$, using the fact that a 2×2 matrix is positive semi-definite if its trace and determinant are both non-negative, it can be easily verified that the condition (4.46) is equivalent to $\alpha \leq \frac{1}{12}$.

For binary squared loss, the link functions constructed by geometric (Proposition 4.1) and calculus (Corollary 4.8) approach are:

$$\tilde{\psi}(\tilde{p}) = e^{-2(1-\tilde{p})^2} - e^{-2\tilde{p}^2} \quad \text{and} \quad \tilde{\psi}^*_\ell(\tilde{p}) = \frac{4}{4}\int_0^{\tilde{p}} \frac{w_{\ell^{\text{sq}}}(v)}{w_{\ell^{\log}}(v)} dv = 4\left(\frac{\tilde{p}^2}{2} - \frac{\tilde{p}^3}{3}\right),$$

respectively. By applying these link functions we can get 1-exp-concave composite squared loss.

4.4.5 Boosting Loss

Consider the binary "boosting loss" (Buja et al. [2005]) with partial losses given by

$$\ell_1^{\text{boost}}(\tilde{p}) = \frac{1}{2}\sqrt{\frac{1-\tilde{p}}{\tilde{p}}} \quad \text{and} \quad \ell_2^{\text{boost}}(\tilde{p}) = \frac{1}{2}\sqrt{\frac{\tilde{p}}{1-\tilde{p}}}, \quad \forall \tilde{p} \in (0,1).$$

This loss has weight function

$$w_{\ell^{\text{boost}}}(\tilde{p}) = \frac{1}{4(\tilde{p}(1-\tilde{p}))^{3/2}}, \quad \forall \tilde{p} \in (0,1).$$

By applying the results of Van Erven et al. [2012], we can show that this loss is mixable with mixability constant 2 (since $\beta_\ell = \inf_{\tilde{p}\in(0,1)} \frac{w_{\ell^{\log}}(\tilde{p})}{w_\ell(\tilde{p})}$).

Now we can check the level of exp-concavification of this loss for different choices of link functions. By considering the identity link $\tilde{\psi}(\tilde{p}) = \tilde{p}$, from (4.23)

$$\begin{aligned}
-\frac{1}{\tilde{p}} + \alpha w_{\ell^{\text{boost}}}(\tilde{p})\tilde{p} &\leq \frac{w'_{\ell^{\text{boost}}}(\tilde{p})}{w_{\ell^{\text{boost}}}(\tilde{p})}, \quad \forall \tilde{p} \in (0,1) \\
\Rightarrow -\frac{1}{\tilde{p}} + \alpha w_{\ell^{\text{boost}}}(\tilde{p})\tilde{p} &\leq 6 w_{\ell^{\text{boost}}}(\tilde{p})\sqrt{\tilde{p}(1-\tilde{p})}(2\tilde{p}-1), \quad \forall \tilde{p} \in (0,1) \\
\Rightarrow \alpha\tilde{p} - 6\sqrt{\tilde{p}(1-\tilde{p})}(2\tilde{p}-1) &\leq \frac{1}{w_{\ell^{\text{boost}}}(\tilde{p})\tilde{p}}, \quad \forall \tilde{p} \in (0,1) \\
\Rightarrow \alpha &\leq 8\sqrt{\frac{1-\tilde{p}}{\tilde{p}}}(\tilde{p} - 1/4), \quad \forall \tilde{p} \in (0,1) \\
\Rightarrow \alpha &\leq 0,
\end{aligned}$$

we see that the boosting loss is non-exp-concave. Similarly from (4.27)

$$\alpha \leq \frac{1}{w_{\ell^{\text{boost}}}(\tilde{p})\tilde{p}^2} = 4\sqrt{\frac{1-\tilde{p}}{\tilde{p}}}(1-\tilde{p}), \quad \forall \tilde{p} \in (0,1) \tag{4.48}$$

it can be seen that the RHS of (4.48) approaches 0 as $p \to 1$, thus it is not possible to exp-concavify (for some $\alpha > 0$) this loss using the canonical link. For binary boosting loss, the link functions constructed by geometric (Proposition 4.1) and calculus (Corollary 4.8) approach are:

$$\tilde{\psi}(\tilde{p}) = e^{-\sqrt{\frac{1-\tilde{p}}{\tilde{p}}}} - e^{-\sqrt{\frac{\tilde{p}}{1-\tilde{p}}}} \quad \text{and} \quad \tilde{\psi}^*_\ell(\tilde{p}) = \frac{4}{2}\int_0^{\tilde{p}} \frac{w_{\ell^{\text{boost}}}(v)}{w_{\ell^{\log}}(v)} dv = \frac{1}{2}\arcsin(-1+2\tilde{p}),$$

respectively (as shown in Figure 4.13). By applying these link functions we can get 2-exp-concave composite boosting loss.

4.4.6 Log Loss

By using the results from this paper and Van Erven et al. [2012] one can easily verify that the multi-class log loss is both 1-mixable and 1-exp-concave. For binary log loss, the link functions constructed by geometric (Proposition 4.1) and calculus (Corollary 4.8) approach are:

$$\tilde{\psi}(\tilde{p}) = e^{\log\tilde{p}} - e^{\log 1-\tilde{p}} = 2\tilde{p} - 1 \quad \text{and} \quad \tilde{\psi}^*_\ell(\tilde{p}) = \frac{4}{4}\int_0^{\tilde{p}} \frac{w_{\ell^{\log}}(v)}{w_{\ell^{\log}}(v)} dv = \tilde{p},$$

respectively.

Chapter 5

Accelerating Optimization for Easy Data

The Online Convex Optimization (OCO) problem plays a key role in machine learning as it has interesting theoretical implications and important practical applications especially in the large scale setting where computational efficiency is the main concern. [Shalev-Shwartz, 2011] provides a detailed analysis of the OCO problem setting and discusses several applications of this paradigm - online regression, prediction with expert advice, and online ranking.

Given a convex set $\Omega \subseteq \mathbb{R}^n$ and a set $\mathcal{F}$ of convex functions, the OCO problem can be formulated as a repeated game between a learner and an adversary. At each time step $t \in [T]$, the learner chooses a point $x_t \in \Omega$, then the adversary reveals the loss function $f_t \in \mathcal{F}$, and the learner suffers a loss of $f_t(x_t)$. The learner's goal is to minimize the regret (w.r.t. any $x^* \in \Omega$) which is given by

$$R(\{f_t\}_{t=1}^T, x^*) := \sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(x^*).$$

For example, consider the online linear regression problem ([Shalev-Shwartz, 2011] (Example 2.1)). At each time step t, the learner receives a feature vector $x_t \in \mathbb{R}^d$, and predicts $p_t \in \mathbb{R}$. Then the adversary reveals the true value $y_t \in \mathbb{R}$, and the learner pays the loss $|y_t - p_t|$. When the learner's prediction is of the form $p_t = \langle w_t, x_t \rangle$, and she needs to compete with the set of linear predictors, this problem can be cast in the OCO framework by setting $f_t(w_t) = |\langle w_t, x_t \rangle - y_t|$.

Abernethy et al. [2008] analyzed the OCO problem from a minimax perspective (where each player plays optimally for their benefit), and showed that $R(\{f_t\}_{t=1}^T, x^*) \approx \Omega\left(\sqrt{T}\right)$ for arbitrary sequence of convex losses $\{f_t\}_{t=1}^T$, and for any strategy of the learner. But the adversary choosing f_t need not to be malicious always, for example the f_t might be drawn from a distribution.

There are two main classes of update rules which attain the above minimax regret bound $O\left(\sqrt{T}\right)$ (thus called minimax optimal updates), namely Follow The Regularized Leader (FTRL) and Mirror Descent. In this work we consider the latter class. Given a strongly convex function (formally defined later) ψ and a learning rate $\eta > 0$, standard

mirror descent update is given by

$$x_{t+1} = \arg\min_{x \in \Omega} \eta \langle g_t, x_t \rangle + \mathcal{B}_\psi(x, x_t), \tag{5.1}$$

where $g_t \in \partial f_t(x_t)$ and $\mathcal{B}_\psi(\cdot, \cdot)$ is Bregman divergence (formally defined later). Shalev-Shwartz [2011] provides a comprehensive survey of analysis techniques for this non-adaptive algorithm family, where the learning rate is fixed for all rounds and chosen with knowledge of T.

Easy Data Instances: It is well understood that the minimax optimal algorithms achieve a regret bound of $O(\sqrt{T})$, which cannot be improved for arbitrary sequences of convex losses [Zinkevich, 2003]. But in practice there are several *easy data* instances such as sparsity, predictable sequences and curved losses, in which much tighter regret bounds are achievable. These tighter bounds translate to much better performance in practice, especially for high dimensional but sparse problems (McMahan [2014]). Even though minimax analysis gives robust algorithms, they are overly conservative on *easy data*. Now we consider some of the existing algorithms that automatically adapt to the *easy data* to learn faster while being robust to worst case as well.

[Duchi et al., 2011] replaced the single static regularizer ψ in the standard mirror descent update 5.1 by a data dependent sequence of regularizers. This is a fully adaptive approach as it doesn't require any prior knowledge about the bound on the term given by $\sum_{t=1}^{T} \|g_t\|^2$ to construct the regularizers. Further for a particular choice of regularizer sequence they achieved a regret bound of the form

$$R(\{f_t\}_{t=1}^{T}, x^*) = O\left(\max_t \|x_t - x^*\|_\infty \sum_{i=1}^{n} \sqrt{\sum_{t=1}^{T} g_{t,i}^2} \right),$$

which is better than the minimax optimal bound ($O\left(G\sqrt{T}\right)$, where G is the worst case magnitude of gradients) when the gradients of the losses are sparse and the prediction space is box-shaped.

[Chiang et al., 2012; Rakhlin and Sridharan, 2012] have shown that an optimistic prediction $\tilde{g}_{t+1}$ of the next gradient g_{t+1} at time t can be used to achieve tighter regret bounds in the case where the loss functions are generated by some predictable process e.g. i.i.d losses with small variance and slowly changing gradients. For the general convex losses, the regret bound of this optimistic approach is $O\left(\sqrt{\sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_*^2}\right)$. But this is a non-adaptive approach since one requires knowledge of the upper bound on $\sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_*^2$ to set the optimal value for the learning rate. Instead we can employ the standard doubling trick to obtain similar bound with slightly worst constants.

Online optimization with curved losses (strong-convex, exp-concave, mixable etc.) is easier than linear losses. When the loss functions are uniformly exp-concave or strongly convex, $O(\log T)$ regret bounds are achieved with appropriate choice of regularizers [Hazan et al., 2007a,b]. But this bound will become worse when the uniform lower bound on the convexity parameters is much smaller. In that case [Hazan et al.,

2007b] proposed an algorithm that can adapt to the convexity of the loss functions, and achieves $O(\sqrt{T})$ regret bounds for arbitrary convex losses and $O(\log T)$ for uniformly strong-convex losses.

Chapter Outline: Even though [McMahan, 2014] has shown equivalence between mirror descent and a variant of FTRL (namely FTRL-Prox) algorithms with adaptive regularizers, no such mapping is available between optimistic mirror descent and optimistic FTRL updates. Recently [Mohri and Yang, 2015] have combined adaptive FTRL and optimistic FTRL updates to achieve tighter regret bounds for sparse and predictable sequences. In section 5.2 we extend this unification to obtain adaptive and optimistic mirror descent updates. We obtained a factor of $\sqrt{2}$ improvement in the regret bound compared to that of [Mohri and Yang, 2015], because in their regret analysis they could not apply the strong FTRL lemma from [McMahan, 2014].

In section 5.3 we consider the adaptive and optimistic mirror descent update with strongly convex loss functions. In this case we achieve tighter logarithmic regret bound without a priori knowledge about the lower bound on the strong-convexity parameters, in similar spirit of [Hazan et al., 2007b]. We also present a curvature adaptive optimistic algorithm that interpolates the results for general convex losses and strongly-convex losses.

In practice the original convex optimization problem itself can have a regularization term associated with the constraints of the problem and generally it is not preferable to linearize those (possibly non-smooth) regularization terms. In section 5.4 we extend all our results to such composite objectives as well.

The main contributions of this chapter are:

- An adaptive and optimistic mirror descent update that achieves tighter regret bounds for sparse and predictable sequences (Section 5.2).
- Improved optimistic mirror descent algorithm that adapts to the curvature of the loss functions (Section 5.3).
- Extension of the unified update rules to the composite objectives (Section 5.4).

Omitted proofs are given in section 5.6.1.

5.1 Notation and Background

We use the following notation throughout. For $n \in \mathbb{Z}^+$, let $[n] := \{1, ..., n\}$. The ith element of a vector $x \in \mathbb{R}^n$ is denoted by $x_i \in \mathbb{R}$, and for a time dependent vector $x_t \in \mathbb{R}^n$, the ith element is $x_{t,i} \in \mathbb{R}$. The inner product between two vectors $x, y \in \mathbb{R}^n$ is written as $\langle x, y \rangle$. The gradient of a differentiable function f at $x \in \mathbb{R}^n$ is denoted by $\nabla f(x)$ or $f'(x)$. A superscript T, A^T denotes transpose of the matrix or vector A. Given $x \in \mathbb{R}^n$, $A = \text{diag}(x)$ is the $n \times n$ matrix with entries $A_{ii} = x_i$, $i \in [n]$ and $A_{ij} = 0$ for $i \neq j$. Similarly given $B \in \mathbb{R}^{n \times n}$, $A = \text{diag}(B)$ is the $n \times n$ matrix with entries $A_{ii} = B_{ii}$, $i \in [n]$ and $A_{ij} = 0$ for $i \neq j$. For a symmetric positive

definite matrix $A \in S^n_{++}$, we have that $\forall x \neq 0, x^T Ax > 0$. If $A - B \in S^n_{++}$, then we write $A \succ B$. The square root of $A \in S^n_{++}$ is the unique matrix $X \in S^n_{++}$ such that $XX = A$ and it is denoted as $A^{\frac{1}{2}}$. We use the compressed summation notation $H_{a:b}$ as shorthand for $\sum_{s=a}^{b} H_s$, where H_s can be a scalar, vector, matrix, or function. Given a norm $\|\cdot\|$, its dual norm is defined as follows $\|y\|_* := \sup_{x:\|x\|\leq 1} \langle x, y\rangle$. For a time varying norm $\|\cdot\|_{(t)}$, its dual norm is written as $\|\cdot\|_{(t),*}$. The dual norm of the Mahalanobis norm $\|x\|_A := \sqrt{x^T Ax}$ is given by $\|y\|_{A^{-1}} = \sqrt{y^T A^{-1} y}$.

Given a convex set $\Omega \subseteq \mathbb{R}^n$ and a convex function $f : \Omega \to \mathbb{R}$, $\partial f(x)$ denotes the sub-differential of f at x which is defined as $\partial f(x) := \{g : f(y) \geq f(x) + \langle g, y - x\rangle, \forall y \in \Omega\}$. A function $f : \Omega \to \mathbb{R}$ is α-strongly convex with respect to a general norm $\|\cdot\|$ if for all $x, y \in \Omega$

$$f(x) \;\geq\; f(y) + \langle g, x - y\rangle + \frac{\alpha}{2}\|x - y\|^2, \, g \in \partial f(y).$$

The Bregman divergence with respect to a differentiable function g is defined as follows

$$\mathcal{B}_g(x, y) \;:=\; g(x) - g(y) - \langle \nabla g(y), x - y\rangle.$$

Observe that the function g is α-strongly convex with respect to $\|\cdot\|$ if and only if for all $x, y \in \Omega : \mathcal{B}_g(x, y) \geq \frac{\alpha}{2}\|x - y\|^2$. In this chapter we use the following properties of Bregman divergences

- Linearity: $\mathcal{B}_{\alpha\psi+\beta\phi}(x, y) = \alpha\mathcal{B}_\psi(x, y) + \beta\mathcal{B}_\phi(x, y)$.
- Generalized triangle inequality: $\mathcal{B}_\psi(x, y) + \mathcal{B}_\psi(y, z) = \mathcal{B}_\psi(x, z) + \langle x - y, \nabla\psi(z) - \nabla\psi(y)\rangle$.

The following proposition [Srebro et al., 2011; Beck and Teboulle, 2003] is handy in deriving explicit update rules for mirror descent algorithms.

Proposition 5.1. *Suppose ψ is strictly convex and differentiable, and y satisfies the condition $\nabla\psi(y) = \nabla\psi(u) - g$. Then*

$$\arg\min_{x\in\Omega} \{\langle g, x\rangle + \mathcal{B}_\psi(x, u)\} = \arg\min_{x\in\Omega} \mathcal{B}_\psi(x, y).$$

5.2 Adaptive and Optimistic Mirror Descent

When the sequence of losses f_t's (in fact their sub-gradients g_t's) are predictable, many authors have recently considered variance (regret) bounds (Hazan and Kale [2010]) that depend only on the deviation of g_t from its average, or path length (regret) bounds (Chiang et al. [2012]) in terms of $g_t - g_{t-1}$. Rakhlin and Sridharan [2012] present an optimistic learning framework that yields such bounds for any mirror descent algorithm. In this framework, the learner is given a sequence of 'hints' $\tilde{g}_{t+1}(g_1, ..., g_t)$ of what g_{t+1} might be. Then the learner chooses x_{t+1} based on the optimistically predicted sub-gradient $\tilde{g}_{t+1}$ along with already observed sub-gradients $g_1, ..., g_t$. For the

Algorithm 1 Adaptive and Optimistic Mirror Descent

Input: regularizers $r_0, r_1 \geq 0$, scheme for selecting r_t for $t \geq 2$.
Initialize: $x_1, \widehat{x}_1 = 0 \in \Omega$.
for $t = 1$ **to** T **do**
Predict $\widehat{x}_t$, observe f_t, and incur loss $f_t(\widehat{x}_t)$.
Compute $g_t \in \partial f_t(\widehat{x}_t)$ and $\tilde{g}_{t+1}(g_1, ..., g_t)$.
Choose r_{t+1} s.t. $r_{0:t+1}$ is 1-strongly convex w.r.t. $\|\cdot\|_{(t+1)}$.
Update

$$x_{t+1} = \arg\min_{x \in \Omega} \langle g_t, x \rangle + \mathcal{B}_{r_{0:t}}(x, x_t), \tag{5.2}$$

$$\widehat{x}_{t+1} = \arg\min_{x \in \Omega} \langle \tilde{g}_{t+1}, x \rangle + \mathcal{B}_{r_{0:t+1}}(x, x_{t+1}). \tag{5.3}$$

end for

optimistic sub-gradient prediction choices of $\tilde{g}_{t+1} = \frac{1}{t}\sum_{s=1}^{t} g_s$ (reasonable prediction when the adversary is iid) and $\tilde{g}_{t+1} = g_t$ (reasonable prediction for slow varying data), we obtain the variance bound and the path length bound respectively.

Given a 1-strongly convex function ψ, and a learning rate $\eta > 0$, the optimistic mirror descent update is equivalent to the following two stage updates

$$\begin{aligned} x_{t+1} &= \arg\min_{x \in \Omega} \eta\langle g_t, x \rangle + \mathcal{B}_{\psi}(x, x_t) \\ \widehat{x}_{t+1} &= \arg\min_{x \in \Omega} \eta\langle \tilde{g}_{t+1}, x \rangle + \mathcal{B}_{\psi}(x, x_{t+1}). \end{aligned}$$

Adaptive and Optimistic mirror descent update is obtained by replacing the static regularizer ψ by a sequence of data dependent regularizers r_t's, which are chosen such that $r_{0:t}$ is 1-strongly convex with respect to $\|\cdot\|_{(t)}$ (here we use the compressed summation notation $r_{0:t}(x) = \sum_{s=0}^{t} r_s(x)$). The unified update is given in Algorithm 1. Note that the regularizer r_{t+1} is constructed at time t (based on the data observed only up to time t) and is used in the second stage update (5.3). Also observe that by setting $\tilde{g}_t = 0$ for all t in Algorithm 1 we recover a slightly modified adaptive mirror descent update given by $x_{t+1} = \arg\min_{x \in \Omega} \langle g_t, x \rangle + \mathcal{B}_{r_{0:t}}(x, x_t)$, where r_t can depend only on $g_1, ..., g_{t-1}$.

In order to obtain a regret bound for Algorithm 1, we first consider the *instantaneous linear regret* (w.r.t. any $x^* \in \Omega$) of it given by $\langle \widehat{x}_t - x^*, g_t \rangle$. The following lemma is a generalization of Lemma 5 from [Chiang et al., 2012] for time varying norms, which gives a bound on the instantaneous linear regret of Algorithm 1.

Lemma 5.2. *The instantaneous linear regret of Algorithm 1 w.r.t. any $x^* \in \Omega$ is bounded from above as follows*

$$\langle \widehat{x}_t - x^*, g_t \rangle \leq \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) + \tfrac{1}{2}\|g_t - \tilde{g}_t\|^2_{(t),*}.$$

Proof. Consider

$$\langle g_t, \widehat{x}_t - x^* \rangle = \langle g_t - \tilde{g}_t, \widehat{x}_t - x_{t+1} \rangle + \langle \tilde{g}_t, \widehat{x}_t - x_{t+1} \rangle + \langle g_t, x_{t+1} - x^* \rangle. \tag{5.4}$$

By the fact that $\langle a, b \rangle \leq \|a\| \|b\|_* \leq \frac{1}{2} \|a\|^2 + \frac{1}{2} \|b\|_*^2$, we have

$$\langle g_t - \tilde{g}_t, \widehat{x}_t - x_{t+1} \rangle \ \leq \ \tfrac{1}{2} \|\widehat{x}_t - x_{t+1}\|_{(t)}^2 + \tfrac{1}{2} \|g_t - \tilde{g}_t\|_{(t),*}^2 .$$

The first-order optimality condition [Boyd and Vandenberghe, 2004] for

$$x^* = \arg\min_{x \in \Omega} \langle g, x \rangle + \mathcal{B}_\psi(x, y)$$

is given by

$$\langle x^* - z, g \rangle \ \leq \ \mathcal{B}_\psi(z, y) - \mathcal{B}_\psi(z, x^*) - \mathcal{B}_\psi(x^*, y), \forall z \in \Omega.$$

By applying the above condition for (5.3) and (5.2) we have respectively

$$\begin{aligned} \langle \widehat{x}_t - x_{t+1}, \tilde{g}_t \rangle \ &\leq \ \mathcal{B}_{r_{0:t}}(x_{t+1}, x_t) - \mathcal{B}_{r_{0:t}}(x_{t+1}, \widehat{x}_t) - \mathcal{B}_{r_{0:t}}(\widehat{x}_t, x_t), \\ \langle x_{t+1} - x^*, g_t \rangle \ &\leq \ \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) - \mathcal{B}_{r_{0:t}}(x_{t+1}, x_t). \end{aligned}$$

Thus by (5.4) we have

$$\begin{aligned} & \langle g_t, \widehat{x}_t - x^* \rangle \\ \leq \ & \tfrac{1}{2} \|\widehat{x}_t - x_{t+1}\|_{(t)}^2 + \tfrac{1}{2} \|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) - \mathcal{B}_{r_{0:t}}(x_{t+1}, \widehat{x}_t) \\ \leq \ & \tfrac{1}{2} \|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) \end{aligned}$$

where the second inequality is due to 1-strong convexity of $r_{0:t}$ w.r.t. $\|\cdot\|_{(t)}$. □

The following lemma is already proven by [Chiang et al., 2012] and used in the proof of our Theorem 5.8.

Lemma 5.3. *For Algorithm 1 we have,* $\|\widehat{x}_t - x_{t+1}\|_{(t)} \ \leq \ \|g_t - \tilde{g}_t\|_{(t),*}$.

The following regret bound holds for Algorithm 1 with a sequence of general convex functions f_t's:

Theorem 5.4. *The regret of Algorithm 1 w.r.t. any* $x^* \in \Omega$ *is bounded by*

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \ \leq \ \tfrac{1}{2} \sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_{(t),*}^2 + \sum_{t=1}^{T} \mathcal{B}_{r_t}(x^*, x_t) + \mathcal{B}_{r_0}(x^*, x_1) - \mathcal{B}_{r_{0:T}}(x^*, x_{T+1}).$$

Proof. Consider

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*)$$

$$\begin{aligned} &\leq \sum_{t=1}^{T} \langle g_t, \widehat{x}_t - x^* \rangle \\ &\leq \sum_{t=1}^{T} \tfrac{1}{2} \left\| g_t - \tilde{g}_t \right\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}), \end{aligned}$$

where the first inequality is due to the convexity of f_t and the second one is due to Lemma 5.2. Then the following simplification of the sum of Bregman divergence terms completes the proof.

$$\begin{aligned} &\sum_{t=1}^{T} \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) \\ =\ & \mathcal{B}_{r_0}(x^*, x_1) - \mathcal{B}_{r_{0:T}}(x^*, x_{T+1}) + \sum_{t=1}^{T} \mathcal{B}_{r_t}(x^*, x_t) \end{aligned}$$

□

Now we analyse the performance of Algorithm 1 with specific choices of regularizer sequences. First we recover the non-adaptive optimistic mirror descent [Chiang et al., 2012] and its regret bound as a corollary of Theorem 5.4.

Corollary 5.5. *Given* 1*-strongly convex (w.r.t.* $\|\cdot\|$*) function* ψ*, define* $\mathcal{R}_{max}(x^*) := \max_{x \in \Omega} \mathcal{B}_\psi(x^*, x) - \min_{x \in \Omega} \mathcal{B}_\psi(x^*, x) = \max_{x \in \Omega} \mathcal{B}_\psi(x^*, x)$*. If* r_t*'s are given by* $r_0(x) = \frac{1}{\eta}\psi(x)$ *(for* $\eta > 0$*) and* $r_t(x) = 0,\ \forall t \geq 1$*, then the regret of Algorithm 1 w.r.t. any* $x^* \in \Omega$ *is bounded as follows*

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \ \leq\ \frac{\eta}{2} \sum_{t=1}^{T} \left\| g_t - \tilde{g}_t \right\|_*^2 + \frac{1}{\eta} \mathcal{R}_{max}(x^*).$$

Further if $\sum_{t=1}^{T} \left\| g_t - \tilde{g}_t \right\|_*^2 \leq Q$*, then by choosing* $\eta = \sqrt{\frac{2\mathcal{R}_{max}(x^*)}{Q}}$*, we have*

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \ \leq\ \sqrt{2\mathcal{R}_{max}(x^*) Q}.$$

Proof. For the given choice of regularizers, we have $r_{0:t}(x) = \frac{1}{\eta}\psi(x)$ and $\mathcal{B}_{r_{0:t}}(x, y) = \frac{1}{\eta}\mathcal{B}_\psi(x, y)$. Since $r_{0:t}$ is 1-strongly convex w.r.t. $\frac{1}{\sqrt{\eta}} \|\cdot\|$, we have $\|\cdot\|_{(t)} = \frac{1}{\sqrt{\eta}} \|\cdot\|$ and $\|\cdot\|_{(t),*} = \sqrt{\eta} \|\cdot\|_*$. Then the corollary directly follows from Theorem 5.4. □

In this non-adaptive case we need to know an upper bound of $\sum_{t=1}^{T} \left\| g_t - \tilde{g}_t \right\|_*^2$ in advance to choose the optimal value for η. Instead we can employ the standard doubling trick to obtain similar bounds with slightly worst constants.

By leveraging the techniques from [Duchi et al., 2011] we can adaptively construct regularizers based on the observed data. The following corollary describes a regularizer construction scheme for Algorithm 1 which is fully adaptive and achieves a regret guarantee that holds at anytime.

Corollary 5.6. *Given* $\Omega \subseteq \times_{i=1}^{n}[-R_i, R_i]$, *let*

$$\begin{aligned} G_0 &= 0 && (5.5)\\ G_1 &= \gamma^2 I \ \ s.t. \ \ \gamma^2 I \succcurlyeq (g_t - \tilde{g}_t)(g_t - \tilde{g}_t)^T, \forall t && (5.6)\\ G_t &= (g_{t-1} - \tilde{g}_{t-1})(g_{t-1} - \tilde{g}_{t-1})^T, \forall t \geq 2 && (5.7)\\ Q_{1:t} &= diag\left(\frac{1}{R_1}, ..., \frac{1}{R_n}\right) diag\left(G_{1:t}\right)^{\frac{1}{2}}. \end{aligned}$$

If r_t*'s are given by* $r_0(x) = 0$ *and* $r_t(x) = \frac{1}{2\sqrt{2}}\|x\|_{Q_t}^2$, *then the regret of Algorithm 1 w.r.t. any* $x^* \in \Omega$ *is bounded by*

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \leq 2\sqrt{2}\sum_{i=1}^{n} R_i \sqrt{\gamma^2 + \sum_{t=1}^{T-1}\left(g_{t,i} - \tilde{g}_{t,i}\right)^2}.$$

Proof. By letting $\eta = \sqrt{2}$ for the given sequence of regularizers, we get $r_{0:t}(x) = \frac{1}{2\eta}\|x\|_{Q_{1:t}}^2$. Since $r_{0:t}$ is 1-strongly convex w.r.t. $\frac{1}{\sqrt{\eta}}\|\cdot\|_{Q_{1:t}}$, we have $\|\cdot\|_{(t)} = \frac{1}{\sqrt{\eta}}\|\cdot\|_{Q_{1:t}}$ and $\|\cdot\|_{(t),*} = \sqrt{\eta}\|\cdot\|_{Q_{1:t}^{-1}}$. By using the facts that $\text{diag}\left(\alpha_1, ..., \alpha_n\right)^{\frac{1}{2}} = \text{diag}\left(\sqrt{\alpha_1}, ..., \sqrt{\alpha_n}\right)$ and $\text{diag}\left(\beta_1, ..., \beta_n\right) \cdot \text{diag}\left(\gamma_1, ..., \gamma_n\right) = \text{diag}\left(\beta_1\gamma_1, ..., \beta_n\gamma_n\right)$, the (i,i)-th entry of the diagonal matrix $Q_{1:t}$ can be given as

$$\begin{aligned} \left(Q_{1:t}\right)_{ii} &= \frac{1}{R_i}\sqrt{\text{diag}\left(\gamma^2 I + \sum_{s=1}^{t-1}\left(g_s - \tilde{g}_s\right)\left(g_s - \tilde{g}_s\right)^T\right)_{ii}}\\ &= \frac{1}{R_i}\sqrt{\gamma^2 + \sum_{s=1}^{t-1}\left(g_{s,i} - \tilde{g}_{s,i}\right)^2}. \end{aligned}$$

Now by Theorem 5.4 the regret bound of Algorithm 1 with this choice of regularizer sequence can be given as follows

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \ \leq \ \frac{1}{2}\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_{(t),*}^2 + \sum_{t=1}^{T}\mathcal{B}_{r_t}(x^*, x_t).$$

Consider

$$\begin{aligned} &\frac{1}{2}\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_{(t),*}^2\\ =\ &\frac{1}{2}\sum_{t=1}^{T}\eta\|g_t - \tilde{g}_t\|_{Q_{1:t}^{-1}}^2\\ =\ &\frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(g_{t,i} - \tilde{g}_{t,i}\right)^2\left(Q_{1:t}\right)_{ii}^{-1} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(g_{t,i}-\tilde{g}_{t,i}\right)^2\frac{R_i}{\sqrt{\gamma^2+\sum_{s=1}^{t-1}\left(g_{s,i}-\tilde{g}_{s,i}\right)^2}}\\
&\leq \frac{\eta}{2}\sum_{i=1}^{n}R_i\sum_{t=1}^{T}\frac{\left(g_{t,i}-\tilde{g}_{t,i}\right)^2}{\sqrt{\sum_{s=1}^{t}\left(g_{s,i}-\tilde{g}_{s,i}\right)^2}}\\
&\leq \eta\sum_{i=1}^{n}R_i\sqrt{\sum_{t=1}^{T}\left(g_{t,i}-\tilde{g}_{t,i}\right)^2}\\
&\leq \eta\sum_{i=1}^{n}R_i\sqrt{\gamma^2+\sum_{t=1}^{T-1}\left(g_{t,i}-\tilde{g}_{t,i}\right)^2},
\end{aligned}
$$

where the first and third inequalities are due to the fact that $\gamma^2 \geq \left(g_{t,i}-\tilde{g}_{t,i}\right)^2$ for all $t \in [T]$, and the second inequality is due to the fact that for any non-negative real numbers $a_1, a_2, ..., a_n$: $\sum_{i=1}^{n}\frac{a_i}{\sqrt{\sum_{j=1}^{i}a_j}} \leq 2\sqrt{\sum_{i=1}^{n}a_i}$. Also observing that

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathcal{B}_{r_t}(x^*,x_t)\\
&= \sum_{t=1}^{T}\frac{1}{2\eta}\left\|x^*-x_t\right\|_{Q_t}^2\\
&= \frac{1}{2\eta}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(x_i^*-x_{t,i}\right)^2\left(Q_t\right)_{ii}\\
&\leq \frac{1}{2\eta}\sum_{i=1}^{n}\left(2R_i\right)^2\sum_{t=1}^{T}\left(Q_t\right)_{ii}\\
&= \frac{2}{\eta}\sum_{i=1}^{n}R_i^2\left(Q_{1:T}\right)_{ii}\\
&= \frac{2}{\eta}\sum_{i=1}^{n}R_i\sqrt{\gamma^2+\sum_{t=1}^{T-1}\left(g_{t,i}-\tilde{g}_{t,i}\right)^2}
\end{aligned}
$$

completes the proof. □

The regret bound obtained in the above corollary is much tighter than that of [Duchi et al., 2011] and [Chiang et al., 2012] when the sequence of loss functions are sparse and predictable. Consider an adversary that is benign and sparse (having non-zero components in fixed locations). In this case, the predictor can learn the non-zero locations of the actual gradient after few iterations. Then $\tilde{g}_t$ will also be mostly zero in the locations where g_t is zero.

Since we are using per-coordinate learning rates implicitly we get better bounds for the case where only certain coordinates of the gradients are accurately predictable as well. Even when the loss sequence is completely unpredictable, the above bound is not much worse than a constant factor of the bound in [Duchi et al., 2011]. For

constructive examples confer [Mohri and Yang, 2016, Section 2.2].

By using Proposition 5.1 we can derive explicit forms of the update rules given by (5.2) and (5.3) with regularizers constructed in Corollary 5.6. For $y_{t+1} = x_t - \sqrt{2}Q_{1:t}^{-1} g_t$ and $\widehat{y}_{t+1} = x_{t+1} - \sqrt{2}Q_{1:t+1}^{-1}\tilde{g}_{t+1}$, the updates (5.2) and (5.3) can be given as $x_{t+1} = \arg\min_{x\in\Omega} \frac{1}{2}\|x - y_{t+1}\|_{Q_{1:t}}^2$ and $\widehat{x}_{t+1} = \arg\min_{x\in\Omega} \frac{1}{2}\|x - \widehat{y}_{t+1}\|_{Q_{1:t+1}}^2$ respectively.

The next corollary explains a regularizer construction method with full matrix learning rates, which is an extension of Corollary 5.6. But this approach is computationally not preferable, especially in high dimensions, as it costs $O(n^2)$ per round of operations.

Corollary 5.7. *Define* $D := \sup_{x,y\in\Omega} \|x - y\|_2$. *Let* $Q_{1:t} = (G_{1:t})^{\frac{1}{2}}$, *where* G_t*'s are given by* (5.5),(5.6) *and* (5.7). *If* r_t*'s are given by* $r_0(x) = 0$ *and* $r_t(x) = \frac{1}{\sqrt{2}D}\|x\|_{Q_t}^2$, *then the regret of Algorithm 1 w.r.t. any* $x^* \in \Omega$ *is bounded by*

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \leq \sqrt{2}D\, tr\left(Q_{1:T}\right).$$

The improvement is a bit more subtle in this case, and it is problem dependent as well. Since this method is not computationally efficient we haven't discussed it in detail. Please confer [Duchi et al., 2011, Section 1.3] for an example.

5.3 Optimistic Mirror Descent with Curved Losses

The following theorem provides a regret bound of Algorithm 1 for the case where f_t is H_t-strongly convex with respect to some general norm $\|\cdot\|$. Since this theorem is an extension of Theorem 2.1 from [Hazan et al., 2007b] for the Optimistic Mirror Descent, this inherits the properties mentioned there such as : r_t's can be chosen without the knowledge of uniform lower bound on H_t's, and $O(\log T)$ bound can be achieved even when some $H_t \leq 0$ as long as $\frac{H_{1:t}}{t} > 0$.

Theorem 5.8. *Let* f_t *is* H_t*-strongly convex w.r.t.* $\|\cdot\|$ *and* $H_t \leq \gamma$ *for all* $t \in [T]$. *If* r_t*'s are given by* $r_0(x) = 0$, $r_1(x) = \frac{\gamma}{4}\|x\|^2$, *and* $r_t(x) = \frac{H_{t-1}}{4}\|x\|^2$ *for all* $t \geq 2$, *then the regret of Algorithm 1 w.r.t. any* $x^* \in \Omega$ *is bounded by*

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \;\leq\; 3\sum_{t=1}^{T} \frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t}} + \frac{\gamma}{4}\|x^* - x_1\|^2.$$

Proof. For the given choice of regularizers, we have $r_{0:t}(x) = \frac{H_{1:t-1}+\gamma}{4}\|x\|^2$ and

$$\mathcal{B}_{r_{0:t}}(x, y) = \frac{H_{1:t-1} + \gamma}{4}\|x - y\|^2.$$

Since $r_{0:t}$ is 1-strongly convex w.r.t. $\sqrt{\frac{H_{1:t-1}+\gamma}{2}}\|\cdot\|$, we have $\|\cdot\|_{(t)} = \sqrt{\frac{H_{1:t-1}+\gamma}{2}}\|\cdot\|$ and

$\|\cdot\|_{(t),*} = \sqrt{\frac{2}{H_{1:t-1}+\gamma}}\,\|\cdot\|_*$. Thus for any $x^* \in \Omega$ we have

$$\begin{aligned}
& f_t(\widehat{x}_t) - f_t(x^*) \\
\leq\ & \langle g_t, \widehat{x}_t - x^* \rangle - \frac{H_t}{2}\,\|\widehat{x}_t - x^*\|^2 \\
\leq\ & \tfrac{1}{2}\,\|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) - \frac{H_t}{2}\,\|\widehat{x}_t - x^*\|^2 \\
=\ & \frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t-1}+\gamma} + \frac{H_{1:t-1}+\gamma}{4}\,\|x^* - x_t\|^2 - \frac{H_{1:t-1}+\gamma}{4}\,\|x^* - x_{t+1}\|^2 - \frac{H_t}{2}\,\|\widehat{x}_t - x^*\|^2,
\end{aligned}$$

where the first inequality is due to the strong convexity of f_t, and the second inequality is due to Lemma 5.2. Observe that

$$\begin{aligned}
& \sum_{t=1}^{T} \frac{H_{1:t-1}+\gamma}{4} \left\{ \|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2 \right\} \\
=\ & \sum_{t=1}^{T} \|x^* - x_{t+1}\|^2 \left\{ \frac{H_{1:t}+\gamma}{4} - \frac{H_{1:t-1}+\gamma}{4} \right\} + \frac{\gamma}{4}\,\|x^* - x_1\|^2 - \frac{H_{1:T}+\gamma}{4}\,\|x^* - x_{T+1}\|^2 \\
\leq\ & \sum_{t=1}^{T} \frac{H_t}{4}\,\|x^* - x_{t+1}\|^2 + \frac{\gamma}{4}\,\|x^* - x_1\|^2,
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{t=1}^{T} \frac{H_t}{4}\,\|x^* - x_{t+1}\|^2 - \frac{H_t}{2}\,\|\widehat{x}_t - x^*\|^2 \\
=\ & \sum_{t=1}^{T} \frac{H_t}{4} \left\{ \|x^* - \widehat{x}_t + \widehat{x}_t - x_{t+1}\|^2 - 2\,\|x^* - \widehat{x}_t\|^2 \right\} \\
\leq\ & \sum_{t=1}^{T} \frac{H_t}{2}\,\|\widehat{x}_t - x_{t+1}\|^2 \\
\leq\ & \sum_{t=1}^{T} \frac{H_{1:t-1}+\gamma}{2}\,\|\widehat{x}_t - x_{t+1}\|^2 \\
\leq\ & 2\sum_{t=1}^{T} \frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t-1}+\gamma},
\end{aligned}$$

where the first inequality is obtained by applying the triangular inequality of norms the fact that $(a+b)^2 \leq 2a^2 + 2b^2$, the second inequality is due to the facts that $H_t \leq \gamma$ and $H_{1:t-1} \geq 0$, and the third inequality is due to Lemma 5.3.

Now by summing up the instantaneous regrets and using the above observation we get

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \ \leq\ 3\sum_{t=1}^{T} \frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t-1}+\gamma} + \frac{\gamma}{4}\,\|x^* - x_1\|^2$$

$$\leq \; 3\sum_{t=1}^{T} \frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t}} + \frac{\gamma}{4}\|x^* - x_1\|^2 ,$$

where the last inequality is due to the fact that $H_t \leq \gamma$. □

In the above theorem if $H_t \geq H > 0$ and $\|g_t - \tilde{g}_t\|_* \leq 1$ (w.l.o.g) for all t, then it obtain a regret bound of the form $O\left(\log \sum_{t=1}^{T} \|g_t - \tilde{g}_t\|_*^2\right)$, using the fact that if $a_t \leq 1$ for all $t \in [T]$, then $\sum_{t=1}^{T} \frac{a_t}{t} = O\left(\log \sum_{t=1}^{T} a_t\right)$. When H is small, however, this guaranteed regret can still be large.

Now instead of running Algorithm 1 on the observed sequence of f_t's, we use the modified sequence of loss functions of the form

$$\tilde{f}_t(x) := f_t(x) + \frac{\lambda_t}{2}\|x - \widehat{x}_t\|^2 , \; \lambda_t \geq 0, \tag{5.8}$$

which is already considered in [Do et al., 2009] for the non-optimistic mirror descent case. Given f_t is H_t-strongly convex with respect to $\|\cdot\|$, $\tilde{f}_t$ is $(H_t + \lambda_t)$-strongly convex. Also note that $\partial \tilde{f}_t(\widehat{x}_t) = \partial f_t(\widehat{x}_t)$ because the gradient of $\|x - \widehat{x}_t\|^2$ is 0 when evaluated at $\widehat{x}_t$ [Do et al., 2009]. Thus in the updates (5.2) and (5.3) the terms g_t and $\tilde{g}_{t+1}$ remain unchanged, only the regularizers r_t's will change appropriately. By applying Theorem 5.8 for the modified sequence of losses given by (5.8) we obtain the following corollary.

Corollary 5.9. *Let* $2R = \sup_{x,y\in\Omega} \|x - y\|$. *Also let* f_t *be* H_t*-strongly convex w.r.t.* $\|\cdot\|$, $H_t \leq \gamma$, *and* $\lambda_t \leq \delta$, *for all* $t \in [T]$. *If Algorithm 1 is performed on the modified functions* $\tilde{f}_t$*'s with the regularizers* r_t*'s given by* $r_0(x) = 0$, $r_1(x) = \frac{\gamma+\delta}{4}\|x\|^2$, *and* $r_t(x) = \frac{H_{t-1}+\lambda_{t-1}}{4}\|x\|^2$ *for all* $t \geq 2$, *then for any sequence* $\lambda_1, ..., \lambda_T \geq 0$, *we get*

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \; \leq \; 2R^2\lambda_{1:T} + 3\sum_{t=1}^{T} \frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t} + \lambda_{1:t}} + \frac{\gamma + \delta}{4}\|x^* - x_1\|^2 .$$

In the above corollary if we consider the two terms that depend on λ_t's, the first term increases and the second term deceases with the increase of λ_t's. Based on the online balancing heuristic approach [Hazan et al., 2007b], the positive solution of $2R^2\lambda_t = 3\frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t}+\lambda_{1:t}}$ is given by

$$\lambda_t = \frac{\sqrt{(H_{1:t} + \lambda_{1:t-1})^2 + \frac{6\|g_t - \tilde{g}_t\|_*^2}{R^2}} - (H_{1:t} + \lambda_{1:t-1})}{2}.$$

The resulting algorithm with the above choice of λ_t is given in Algorithm 2. By using the Lemma 3.1 from [Hazan et al., 2007b] we obtain the following regret bound for Algorithm 2.

Theorem 5.10. *The regret of Algorithm 2 on the sequence of* f_t*'s with curvature*

Algorithm 2 Curvature Adaptive and Optimistic Mirror Descent

Input: $r_0(x) = 0$ and $r_1(x) = \frac{\gamma+\delta}{4}\|x\|^2$.
Initialize: $x_1, \widehat{x}_1 = 0 \in \Omega$.
for $t = 1$ **to** T **do**
Predict $\widehat{x}_t$, observe f_t, and incur loss $f_t(\widehat{x}_t)$.
Compute $g_t \in \partial f_t(\widehat{x}_t)$ and $\tilde{g}_{t+1}(g_1, ..., g_t)$.
Compute $\lambda_t = \frac{\sqrt{(H_{1:t}+\lambda_{1:t-1})^2 + \frac{6\|g_t - \tilde{g}_t\|_*^2}{R^2}} - (H_{1:t}+\lambda_{1:t-1})}{2}$
Define $r_{t+1}(x) = \frac{H_t+\lambda_t}{4}\|x\|^2$.
Update

$$\begin{aligned} x_{t+1} &= \arg\min_{x\in\Omega} \langle g_t, x\rangle + \mathcal{B}_{r_{0:t}}(x, x_t), \\ \widehat{x}_{t+1} &= \arg\min_{x\in\Omega} \langle \tilde{g}_{t+1}, x\rangle + \mathcal{B}_{r_{0:t+1}}(x, x_{t+1}). \end{aligned}$$

end for

$H_t \geq 0$ is bounded by

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \;\leq\; \frac{\gamma+\delta}{4}\|x^* - x_1\|^2 + 2\inf_{\lambda_1^*,\ldots,\lambda_T^*}\left\{2R^2\lambda_{1:T}^* + 3\sum_{t=1}^{T}\frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t}+\lambda_{1:t}^*}\right\}.$$

Thus the Algorithm 2 achieves a regret bound which is competitive with the bound achievable by the best offline choice of parameters λ_t's. From the above theorem we obtain the following two corollaries which show that Algorithm 2 achieves intermediate rates between $O\left(\sqrt{\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_*^2}\right)$ and $O\left(\log\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_*^2\right)$ depending on the curvature of the losses.

Corollary 5.11. *For any sequence of convex loss functions f_t's, the bound on the regret of Algorithm 2 is $O\left(\sqrt{\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_*^2}\right)$.*

Proof. Let $\lambda_1^* = \sqrt{\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_*^2}$, and $\lambda_t^* = 0$ for all $t > 1$.

$$\begin{aligned} & 2R^2\lambda_{1:T}^* + 3\sum_{t=1}^{T}\frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t}+\lambda_{1:t}^*} \\ =\; & 2R^2\sqrt{\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_*^2} + 3\sum_{t=1}^{T}\frac{\|g_t - \tilde{g}_t\|_*^2}{0 + \sqrt{\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_*^2}} \\ =\; & \left(2R^2+3\right)\sqrt{\sum_{t=1}^{T}\|g_t - \tilde{g}_t\|_2^2}. \end{aligned}$$

□

Algorithm 3 Adaptive and Optimistic Mirror Descent with Composite Losses

Input: regularizers $r_0, r_1 \geq 0$, composite losses $\{\psi_t\}_t$ where $\psi_t \geq 0$.
Initialize: $x_1, \widehat{x}_1 = 0 \in \Omega$.
for $t = 1$ **to** T **do**
Predict $\widehat{x}_t$, observe f_t, and incur loss $f_t(\widehat{x}_t) + \psi_t(\widehat{x}_t)$.
Compute $g_t \in \partial f_t(\widehat{x}_t)$ and $\tilde{g}_{t+1}(g_1, ..., g_t)$.
Construct r_{t+1} s.t. $r_{0:t+1}$ is 1-strongly convex w.r.t. $\|\cdot\|_{(t+1)}$.
Update

$$x_{t+1} = \arg\min_{x \in \Omega} \langle g_t, x \rangle + \psi_t(x) + \mathcal{B}_{r_{0:t}}(x, x_t), \tag{5.9}$$

$$\widehat{x}_{t+1} = \arg\min_{x \in \Omega} \langle \tilde{g}_{t+1}, x \rangle + \psi_{t+1}(x) + \mathcal{B}_{r_{0:t+1}}(x, x_{t+1}). \tag{5.10}$$

end for

Corollary 5.12. *Suppose* $\|g_t - \tilde{g}_t\|_* \leq 1$ *(w.l.o.g) and* $H_t \geq H > 0$ *for all* $t \in [T]$. *Then the bound on the regret of Algorithm 2 is* $O\left(\log \sum_{t=1}^T \|g_t - \tilde{g}_t\|_*^2\right)$.

Proof. Set $\lambda_t^* = 0$ for all t.

$$\begin{aligned} 2R^2\lambda_{1:T}^* + 3\sum_{t=1}^T \frac{\|g_t - \tilde{g}_t\|_*^2}{H_{1:t} + \lambda_{1:t}^*} &= 0 + 3\sum_{t=1}^T \frac{\|g_t - \tilde{g}_t\|_*^2}{Ht + 0} \\ &= O\left(\log \sum_{t=1}^T \|g_t - \tilde{g}_t\|_*^2\right), \end{aligned}$$

where the last inequality is due to the fact that if $a_t \leq 1$ for all $t \in [T]$, then $\sum_{t=1}^T \frac{a_t}{t} = O\left(\log \sum_{t=1}^T a_t\right)$. □

The results obtained here can be extended to the applications discussed in [Do et al., 2009; Orabona et al., 2010] to obtain much tighter results.

5.4 Composite Losses

Here we consider the case when observed loss function f_t is composed with some non-negative (possibly non-smooth) convex regularizer term ψ_t to impose certain constraints on the original problem. In this case we generally do not want to linearize the additional regularizer term, thus in the update rules given by (5.2) and (5.3) we include ψ_t and ψ_{t+1} respectively without linearizing them. This extension is presented in Algorithm 3.

The following lemma provides a bound on the instantaneous regret of Algorithm 3.

Lemma 5.13. *The instantaneous regret of Algorithm 3 w.r.t. any* $x^* \in \Omega$ *can be bounded as follows*

$$\{f_t(\widehat{x}_t) + \psi_t(\widehat{x}_t)\} - \{f_t(x^*) + \psi_t(x^*)\} \leq \tfrac{1}{2}\|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}).$$

Proof. The instantaneous regret of the algorithm can be bounded as below using the convexity of f_t

$$\{f_t(\widehat{x}_t) + \psi_t(\widehat{x}_t)\} - \{f_t(x^*) + \psi_t(x^*)\} \;\leq\; \langle g_t, \widehat{x}_t - x^*\rangle + \{\psi_t(\widehat{x}_t) - \psi_t(x^*)\}\,.$$

Now consider

$$\langle g_t, \widehat{x}_t - x^*\rangle \;=\; \langle g_t - \tilde{g}_t, \widehat{x}_t - x_{t+1}\rangle + \langle \tilde{g}_t, \widehat{x}_t - x_{t+1}\rangle + \langle g_t, x_{t+1} - x^*\rangle. \tag{5.11}$$

By the fact that $\langle a, b\rangle \leq \|a\| \|b\|_* \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|_*^2$, we have

$$\langle g_t - \tilde{g}_t, \widehat{x}_t - x_{t+1}\rangle \;\leq\; \tfrac{1}{2}\|\widehat{x}_t - x_{t+1}\|_{(t)}^2 + \tfrac{1}{2}\|g_t - \tilde{g}_t\|_{(t),*}^2$$

The first-order optimality condition for $x^* = \operatorname*{arg\,min}_{x\in\Omega} \langle g, x\rangle + f(x) + \mathcal{B}_\psi(x, y)$ and for $z \in \Omega$,

$$\langle x^* - z, g\rangle \;\leq\; \langle z - x^*, f'(x^*)\rangle + \mathcal{B}_\psi(z, y) - \mathcal{B}_\psi(z, x^*) - \mathcal{B}_\psi(x^*, y).$$

By applying the above condition for (5.10) and (5.9) we have respectively

$$\langle \widehat{x}_t - x_{t+1}, \tilde{g}_t\rangle \;\leq\; \langle \psi_t'(\widehat{x}_t), x_{t+1} - \widehat{x}_t\rangle + \mathcal{B}_{r_{0:t}}(x_{t+1}, x_t) - \mathcal{B}_{r_{0:t}}(x_{t+1}, \widehat{x}_t) - \mathcal{B}_{r_{0:t}}(\widehat{x}_t, x_t)$$

$$\langle x_{t+1} - x^*, g_t\rangle \;\leq\; \langle \psi_t'(x_{t+1}), x^* - x_{t+1}\rangle + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) - \mathcal{B}_{r_{0:t}}(x_{t+1}, x_t).$$

Thus by (5.11) we have

$$\begin{aligned}
&\langle g_t, \widehat{x}_t - x^*\rangle + \{\psi_t(\widehat{x}_t) - \psi_t(x^*)\} \\
\leq\;& \tfrac{1}{2}\|\widehat{x}_t - x_{t+1}\|_{(t)}^2 + \tfrac{1}{2}\|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) - \mathcal{B}_{r_{0:t}}(x_{t+1}, \widehat{x}_t) \\
&+ \psi_t(\widehat{x}_t) - \psi_t(x^*) + \langle \psi_t'(\widehat{x}_t), x_{t+1} - \widehat{x}_t\rangle + \langle \psi_t'(x_{t+1}), x^* - x_{t+1}\rangle \\
\leq\;& \tfrac{1}{2}\|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) \\
&+ \psi_t(\widehat{x}_t) + \langle \psi_t'(\widehat{x}_t), x_{t+1} - \widehat{x}_t\rangle + \langle \psi_t'(x_{t+1}), x^* - x_{t+1}\rangle - \psi_t(x^*) \\
\leq\;& \tfrac{1}{2}\|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) \\
&+ \psi_t(x_{t+1}) + \langle \psi_t'(x_{t+1}), x^* - x_{t+1}\rangle - \psi_t(x^*) \\
\leq\;& \tfrac{1}{2}\|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) + \psi_t(x^*) - \psi_t(x^*) \\
=\;& \tfrac{1}{2}\|g_t - \tilde{g}_t\|_{(t),*}^2 + \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1})
\end{aligned}$$

where the second inequality is due to 1-strong convexity of $r_{0:t}$ w.r.t. $\|\cdot\|_{(t)}$, and the third and fourth inequalities are due to the convexity of ψ_t at $\widehat{x}_t$ and x_{t+1} respectively. □

From the above lemma we can observe that the instantaneous regret of Algorithm 3 is exactly equal to that of the non-composite version (Algorithm 1). Thus all the improvements that we discussed in the previous sections for the non-composite case are also applicable to composite losses as well.

5.5 Discussion

Early approaches to the OCO problem were conservative, in which the main focus was protection against the worst case scenario. But recently several algorithms have been developed for tightening the regret bounds in easy data instances such as sparsity, predictable sequences, and curved losses. We have unified some of these existing techniques to obtain new update rules for the cases when these easy instances occur together. First we have analysed an adaptive and optimistic update rule which achieves tighter regret bound when the loss sequence is sparse and predictable (Algorithm 1). Then we have analysed an update rule that dynamically adapts to the curvature of the loss function and utilizes the predictable nature of the loss sequence as well (Algorithm 2). Finally we have extended these results to composite losses (Algorithm 3).

We also note that the regret bounds given in this chapter can be converted into convergence bounds for batch stochastic problems using online-to-batch conversion techniques [Cesa-Bianchi et al., 2004; Kakade and Tewari, 2009].

5.6 Appendix

5.6.1 Proofs

Proof. **(Proposition 5.1)** Observe that

$$\begin{aligned}
& \arg\min_{x\in\Omega} \mathcal{B}_\psi(x,y) \\
= \; & \arg\min_{x\in\Omega} \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle \\
= \; & \arg\min_{x\in\Omega} \psi(x) - \langle \nabla\psi(y), x\rangle \\
= \; & \arg\min_{x\in\Omega} \psi(x) - \langle \nabla\psi(u) - g, x\rangle \\
= \; & \arg\min_{x\in\Omega} \langle g, x\rangle + \psi(x) - \psi(u) - \langle \nabla\psi(u), x-u\rangle \\
= \; & \arg\min_{x\in\Omega} \langle g, x\rangle + \mathcal{B}_\psi(x,u).
\end{aligned}$$

□

Proof. **(Lemma 5.3)** Since $r_{0:t}$ is 1-strongly convex w.r.t. $\|\cdot\|_{(t)}$ we have

$$\begin{aligned}
& \mathcal{B}_{r_{0:t}}(\widehat{x}_t, x_{t+1}) \\
= \; & r_{0:t}(\widehat{x}_t) - r_{0:t}(x_{t+1}) - \langle \nabla r_{0:t}(x_{t+1}), \widehat{x}_t - x_{t+1}\rangle \\
\geq \; & \tfrac{1}{2}\left\|\widehat{x}_t - x_{t+1}\right\|_{(t)}^2,
\end{aligned}$$

and

$$\begin{aligned}
& \mathcal{B}_{r_{0:t}}(x_{t+1}, \widehat{x}_t) \\
= \; & r_{0:t}(x_{t+1}) - r_{0:t}(\widehat{x}_t) - \langle \nabla r_{0:t}(\widehat{x}_t), x_{t+1} - \widehat{x}_t\rangle
\end{aligned}$$

$$\geq \tfrac{1}{2} \|x_{t+1} - \widehat{x}_t\|_{(t)}^2 .$$

Adding these two bounds, we obtain

$$\|\widehat{x}_t - x_{t+1}\|_{(t)}^2 \leq \langle \nabla r_{0:t}(\widehat{x}_t) - \nabla r_{0:t}(x_{t+1}), \widehat{x}_t - x_{t+1} \rangle. \tag{5.12}$$

Suppose y_{t+1} and $\widehat{y}_t$ satisfy the conditions $\nabla r_{0:t}(y_{t+1}) = \nabla r_{0:t}(x_t) - g_t$ and $\nabla r_{0:t}(\widehat{y}_t) = \nabla r_{0:t}(x_t) - \tilde{g}_t$ respectively. Then by applying Proposition 5.1 to the updates in (5.3) and (5.2) of Algorithm 1, we obtain

$$\begin{aligned} x_{t+1} &= \arg\min_{x \in \Omega} \mathcal{B}_{r_{0:t}}(x, y_{t+1}) \\ \widehat{x}_t &= \arg\min_{x \in \Omega} \mathcal{B}_{r_{0:t}}(x, \widehat{y}_t). \end{aligned}$$

By applying the first order optimality condition for the above two optimization statements, we have

$$\begin{aligned} \langle \nabla r_{0:t}(x_{t+1}) - \nabla r_{0:t}(y_{t+1}), \widehat{x}_t - x_{t+1} \rangle &\geq 0 \\ \langle \nabla r_{0:t}(\widehat{x}_t) - \nabla r_{0:t}(\widehat{y}_t), x_{t+1} - \widehat{x}_t \rangle &\geq 0, \end{aligned}$$

respectively. Combining these two bounds, we obtain

$$\langle \nabla r_{0:t}(\widehat{y}_t) - \nabla r_{0:t}(y_{t+1}), \widehat{x}_t - x_{t+1} \rangle \geq \langle \nabla r_{0:t}(\widehat{x}_t) - \nabla r_{0:t}(x_{t+1}), \widehat{x}_t - x_{t+1} \rangle.$$

By combining the above result with (5.12), we obtain

$$\begin{aligned} & \|\widehat{x}_t - x_{t+1}\|_{(t)}^2 \\ \leq\ & \langle \nabla r_{0:t}(\widehat{y}_t) - \nabla r_{0:t}(y_{t+1}), \widehat{x}_t - x_{t+1} \rangle \\ \leq\ & \|\nabla r_{0:t}(\widehat{y}_t) - \nabla r_{0:t}(y_{t+1})\|_{(t),*} \|\widehat{x}_t - x_{t+1}\|_{(t)} , \end{aligned}$$

by a generalized Cauchy-Schwartz inequality. Dividing both sides by $\|\widehat{x}_t - x_{t+1}\|_{(t)}$, we have

$$\begin{aligned} & \|\widehat{x}_t - x_{t+1}\|_{(t)} \\ \leq\ & \|\nabla r_{0:t}(\widehat{y}_t) - \nabla r_{0:t}(y_{t+1})\|_{(t),*} \\ =\ & \|(\nabla r_{0:t}(x_t) - \tilde{g}_t) - (\nabla r_{0:t}(x_t) - g_t)\|_{(t),*} \\ =\ & \|g_t - \tilde{g}_t\|_{(t),*} . \end{aligned}$$

□

Proof. **(Corollary 5.7)** By letting $\eta = \frac{D}{\sqrt{2}}$ for the given sequence of regularizers, we get $r_{0:t}(x) = \frac{1}{2\eta} \|x\|_{Q_{1:t}}^2$. Since $r_{0:t}$ is 1-strongly convex w.r.t. $\frac{1}{\sqrt{\eta}} \|\cdot\|_{Q_{1:t}}$, we have $\|\cdot\|_{(t)} = \frac{1}{\sqrt{\eta}} \|\cdot\|_{Q_{1:t}}$ and $\|\cdot\|_{(t),*} = \sqrt{\eta} \|\cdot\|_{Q_{1:t}^{-1}}$. By Theorem 5.4 the regret bound of

Algorithm 1 with this choice of regularizer sequence can be given as follows

$$\sum_{t=1}^{T} f_t(\widehat{x}_t) - f_t(x^*) \;\leq\; \tfrac{1}{2}\sum_{t=1}^{T} \|g_t - \tilde{g}_t\|^2_{(t),*} + \sum_{t=1}^{T} \mathcal{B}_{r_t}(x^*, x_t).$$

Consider

$$\begin{aligned}
& \tfrac{1}{2}\sum_{t=1}^{T} \|g_t - \tilde{g}_t\|^2_{(t),*} \\
=\;& \tfrac{1}{2}\sum_{t=1}^{T} \eta \, \|g_t - \tilde{g}_t\|^2_{Q_{1:t}^{-1}} \\
=\;& \frac{\eta}{2}\sum_{t=1}^{T} (g_t - \tilde{g}_t)\, Q_{1:t}^{-1}\, (g_t - \tilde{g}_t)^T \\
=\;& \frac{\eta}{2}\sum_{t=1}^{T} (g_t - \tilde{g}_t) \left(\gamma^2 I + G_{2:t}\right)^{-\frac{1}{2}} (g_t - \tilde{g}_t)^T \\
\leq\;& \frac{\eta}{2}\sum_{t=1}^{T} (g_t - \tilde{g}_t) \left(G_{2:t+1}\right)^{-\frac{1}{2}} (g_t - \tilde{g}_t)^T \\
\leq\;& \eta \operatorname{tr}\left(G_{2:T+1}^{\frac{1}{2}}\right) \\
\leq\;& \eta \operatorname{tr}\left(\left(\gamma^2 I + G_{2:T}\right)^{\frac{1}{2}}\right) \\
=\;& \eta \operatorname{tr}\left(Q_{1:T}\right),
\end{aligned}$$

where the first inequality is due to the facts that $\gamma^2 I \succcurlyeq G_{t+1}$ and $A \succcurlyeq B \succcurlyeq 0 \Rightarrow A^{\frac{1}{2}} \succcurlyeq B^{\frac{1}{2}}$ and $B^{-1} \succcurlyeq A^{-1}$, the second inequality is due to the fact that $\sum_{t=1}^{T} a_t^T \left(\sum_{s=1}^{t} a_s a_s^T\right)^{-\frac{1}{2}} a_t \leq 2 \cdot \operatorname{tr}\left(\left(\sum_{t=1}^{T} a_t a_t^T\right)^{\frac{1}{2}}\right)$ (see Lemma 10 from [Duchi et al., 2011]), and the third inequality is due to the fact that $\gamma^2 I \succcurlyeq G_{T+1}$. Also observing that

$$\begin{aligned}
& \sum_{t=1}^{T} \mathcal{B}_{r_t}(x^*, x_t) \\
=\;& \sum_{t=1}^{T} \frac{1}{2\eta} \|x^* - x_t\|^2_{Q_t} \\
\leq\;& \frac{1}{2\eta}\sum_{t=1}^{T} \|x^* - x_t\|_2^2 \, \lambda_{\max}\left(Q_t\right) \\
\leq\;& \frac{1}{2\eta}\sum_{t=1}^{T} \|x^* - x_t\|_2^2 \operatorname{tr}\left(Q_t\right)
\end{aligned}$$

Algorithm 4 Adaptive Mirror Descent

Input: regularizers $r_0 \geq 0$.
Initialize: $x_1 = 0 \in \Omega$.
for $t = 1$ **to** T **do**
Predict x_t, observe f_t, and incur loss $f_t(x_t)$.
Compute $g_t \in \partial f_t(x_t)$.
Construct r_t s.t. $r_{0:t}$ is 1-strongly convex w.r.t. $\|\cdot\|_{(t)}$.
Update

$$x_{t+1} = \arg\min_{x \in \Omega} \langle g_t, x \rangle + \mathcal{B}_{r_{0:t}}(x, x_t). \tag{5.13}$$

end for

$$\begin{aligned} &\leq \frac{1}{2\eta} \sum_{t=1}^{T} D^2 \mathrm{tr}\left(Q_t\right) \\ &= \frac{D^2}{2\eta} \mathrm{tr}\left(Q_{1:T}\right). \end{aligned}$$

completes the proof. □

5.6.2 Mirror Descent with β-convex losses

Given a convex set $\Omega \subseteq \mathbb{R}^n$ and $\beta > 0$, a function $f : \Omega \to \mathbb{R}$ is β-convex, if for all $x, y \in \Omega$

$$f(x) \geq f(y) + \langle g, x - y \rangle + \beta \|x - y\|^2_{gg^T}, \; g \in \partial f(y).$$

As in Theorem 5.8, we can obtain regret bound for the case when the loss function f_t is β_t-convex (which is broader class than exp-concave losses) as well. But for the resulting bound we cannot apply Lemma 3.1 from [Hazan et al., 2007b] to obtain a near optimal closed form solution of λ_t.

Lemma 5.14. *The instantaneous linear regret of Algorithm 4 w.r.t. any $x^* \in \Omega$ can be bounded as follows*

$$\langle x_t - x^*, g_t \rangle \leq \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) + \tfrac{1}{2} \|g_t\|^2_{(t),*}.$$

Proof. By the first-order optimality condition for (5.13) we have,

$$\langle x - x_{t+1}, g_t + \nabla r_{0:t}(x_{t+1}) - \nabla r_{0:t}(x_t) \rangle \geq 0 \tag{5.14}$$

Consider

$$\begin{aligned} &\langle x_t - x^*, g_t \rangle \\ = \; &\langle x_{t+1} - x^*, g_t \rangle + \langle x_t - x_{t+1}, g_t \rangle \\ \leq \; &\langle x^* - x_{t+1}, \nabla r_{0:t}(x_{t+1}) - \nabla r_{0:t}(x_t) \rangle + \langle x_t - x_{t+1}, g_t \rangle \end{aligned}$$

$$
\begin{aligned}
&= \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) - \mathcal{B}_{r_{0:t}}(x_{t+1}, x_t) + \langle x_t - x_{t+1}, g_t \rangle \\
&\leq \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) - \mathcal{B}_{r_{0:t}}(x_{t+1}, x_t) + \tfrac{1}{2} \|x_t - x_{t+1}\|_{(t)}^2 + \tfrac{1}{2} \|g_t\|_{(t),*}^2 \\
&\leq \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) + \tfrac{1}{2} \|g_t\|_{(t),*}^2 ,
\end{aligned}
$$

where the first inequality is due to (5.14), the second equality is due to the fact that $\langle \nabla\psi(a) - \nabla\psi(b), c - a \rangle = \mathcal{B}_\psi(c, b) - \mathcal{B}_\psi(c, a) - \mathcal{B}_\psi(a, b)$, the second inequality is due to the fact that $\langle a, b \rangle \leq \|a\| \|b\|_* \leq \frac{1}{2} \|a\|^2 + \frac{1}{2} \|b\|_*^2$, and the third inequality is due to the 1-strong convexity of $r_{0:t}$ w.r.t. $\|\cdot\|_{(t)}$. □

Theorem 5.15. *Let f_t is β_t-convex, $\forall t \in [T]$. If r_t's are given by*

$$
r_t(x) = \|x\|_{h_t}^2 , \text{ where } h_0 = I_{n\times n} \text{ and } h_t = \beta_t g_t g_t^T \text{ for } t \geq 1, \tag{5.15}
$$

then the regret of Algorithm 4 w.r.t. any $x^* \in \Omega$ is bounded by

$$
\sum_{t=1}^T f_t(x_t) - f_t(x^*) \;\leq\; \frac{1}{4} \sum_{t=1}^T \|g_t\|_{h_{0:t}^{-1}}^2 + \|x^* - x_1\|_2^2 .
$$

Proof. For the choice of regularizer sequence $\{r_t\}$ given by (5.15), we have $r_{0:t}(x) = \|x\|_{h_{0:t}}^2$ and $\mathcal{B}_{r_{0:t}}(x, y) = \frac{1}{2} \left(\sqrt{2} \|x - y\|_{h_{0:t}} \right)^2$. Since $r_{0:t}$ is 1-strongly convex w.r.t. $\sqrt{2} \|\cdot\|_{h_{0:t}}$, we have $\|\cdot\|_{(t)} = \sqrt{2} \|\cdot\|_{h_{0:t}}$ and $\|\cdot\|_{(t),*} = \frac{1}{\sqrt{2}} \|\cdot\|_{h_{0:t}^{-1}}$.

For any $x^* \in \Omega$

$$
\begin{aligned}
& f_t(x_t) - f_t(x^*) \\
\leq\; & \langle g_t, x_t - x^* \rangle - \beta_t \|x^* - x_t\|_{g_t g_t^T}^2 \\
\leq\; & \mathcal{B}_{r_{0:t}}(x^*, x_t) - \mathcal{B}_{r_{0:t}}(x^*, x_{t+1}) + \tfrac{1}{2} \|g_t\|_{(t),*}^2 - \|x^* - x_t\|_{\beta_t g_t g_t^T}^2 \\
=\; & \|x^* - x_t\|_{h_{0:t}}^2 - \|x^* - x_{t+1}\|_{h_{0:t}}^2 - \|x^* - x_t\|_{h_t}^2 + \tfrac{1}{2} \|g_t\|_{(t),*}^2 ,
\end{aligned}
$$

where the first inequality is due to the β_t-convexity of $f_t(\cdot)$, and the second inequality is due to Lemma 5.14. By summing all the instantaneous regrets we get

$$
\begin{aligned}
& \sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(x^*) \\
\leq\; & \sum_{t=1}^T \left\{ \|x^* - x_t\|_{h_{0:t}}^2 - \|x^* - x_t\|_{h_{0:t-1}}^2 - \|x^* - x_t\|_{h_t}^2 \right\} \\
& + \|x^* - x_1\|_{h_0}^2 - \|x^* - x_{T+1}\|_{h_{0:T}}^2 + \tfrac{1}{2} \sum_{t=1}^T \|g_t\|_{(t),*}^2 \\
\leq\; & \|x^* - x_1\|_2^2 + \frac{1}{4} \sum_{t=1}^T \|g_t\|_{h_{0:t}^{-1}}^2 .
\end{aligned}
$$

□

Now instead of running Algorithm 4 on the observed sequence of f_t's, we use the modified sequence of loss functions of the form

$$\tilde{f}_t(x) := f_t(x) + \lambda_t g(x),\ \lambda_t \geq 0, \tag{5.16}$$

where $g(x)$ is 1-convex. By following the proof of Theorem 5.15 for the modified sequence of losses given by (5.16) we obtain the following corollary.

Theorem 5.16. *Let $g(x)$ be a 1-convex function, $A^2 = \sup_{x\in\Omega} g(x)$ and*

$$B = \sup_{x\in\Omega} \left\|g'(x)\right\|_{\left(g'(x)g'(x)^T\right)^{-1}}.$$

Also let f_t be β_t-convex ($\beta_t \geq 0$), $\forall t \in [T]$. If Algorithm 4 is performed on the modified functions $\tilde{f}_t$'s with the regularizers r_t's given by

$$r_t(x) = \|x\|^2_{h_t},\ \textit{where } h_0 = I_{n\times n},\ \textit{and } h_t = \beta_t g_t g_t^T + \lambda_t g'(x_t)g'(x_t)^T,\ \textit{for } t \geq 1, \tag{5.17}$$

then for any sequence $\lambda_1, ..., \lambda_T \geq 0$, we get

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \ \leq\ \left(A^2 + \frac{B^2}{2}\right)\lambda_{1:T} + \tfrac{1}{2}\sum_{t=1}^{T} \|g_t\|^2_{h_{0:t}^{-1}} + \|x^* - x_1\|_2^2.$$

Proof. Since f_t is β_t-convex and g is 1-convex, for any $x^* \in \Omega$ we have

$$\begin{aligned}
&\{f_t(x_t) + \lambda_t g(x_t)\} - \{f_t(x^*) + \lambda_t g(x^*)\}\\
=\ & f_t(x_t) - f_t(x^*) + \lambda_t\{g(x_t) - g(x^*)\}\\
\leq\ & \langle g_t, x_t - x^*\rangle - \beta_t \|x^* - x_t\|^2_{g_t g_t^T} + \lambda_t\left\{\langle g'(x_t), x_t - x^*\rangle - \|x^* - x_t\|^2_{g'(x_t)g'(x_t)^T}\right\}\\
=\ & \langle g_t + \lambda_t g'(x_t), x_t - x^*\rangle - \|x^* - x_t\|^2_{\beta_t g_t g_t^T + \lambda_t g'(x_t)g'(x_t)^T}.
\end{aligned}$$

By following the similar steps from the proof of Theorem 5.15 we get

$$\sum_{t=1}^{T} f_t(x_t) + \lambda_t g(x_t) - \left\{\sum_{t=1}^{T} f_t(x^*) + \lambda_t g(x^*)\right\} \ \leq\ \frac{1}{4}\sum_{t=1}^{T} \left\|g_t + \lambda_t g'(x_t)\right\|^2_{h_{0:t}^{-1}} + \|x^* - x_1\|_2^2.$$

By using the facts that $\|x+y\|_A^2 \leq 2\|x\|_A^2 + 2\|y\|_A^2$, $h_{0:t} \succcurlyeq h_t \succcurlyeq \lambda_t g'(x_t)g'(x_t)^T$, and $\|g'(x_t)\|_{\left(g'(x_t)g'(x_t)^T\right)^{-1}} \leq B$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(x_t) + \lambda_t g(x_t) - \left\{\sum_{t=1}^{T} f_t(x^*) + \lambda_t g(x^*)\right\}\\
\leq\ & \tfrac{1}{2}\sum_{t=1}^{T}\left\{\|g_t\|^2_{h_{0:t}^{-1}} + \lambda_t^2\left\|g'(x_t)\right\|^2_{h_{0:t}^{-1}}\right\} + \|x^* - x_1\|_2^2\\
\leq\ & \tfrac{1}{2}\sum_{t=1}^{T}\left\{\|g_t\|^2_{h_{0:t}^{-1}} + \lambda_t^2\left\|g'(x_t)\right\|^2_{\left(\lambda_t g'(x_t)g'(x_t)^T\right)^{-1}}\right\} + \|x^* - x_1\|_2^2
\end{aligned}$$

$$\leq \frac{1}{2}\sum_{t=1}^{T}\|g_t\|^2_{h^{-1}_{0:t}} + \frac{B^2}{2}\lambda_{1:T} + \|x^* - x_1\|_2^2 .$$

By neglecting the $g(x_t)$ terms in the L.H.S. and using the fact that $g(x^*) \leq A^2$ we get

$$\sum_{t=1}^{T} f_t(x_t) \leq \sum_{t=1}^{T} f_t(x^*) + A^2\lambda_{1:T} + \frac{1}{2}\sum_{t=1}^{T}\|g_t\|^2_{h^{-1}_{0:t}} + \|x^* - x_1\|_2^2 + \frac{B^2}{2}\lambda_{1:T}.$$

□

But we cannot apply Lemma 3.1 from [Hazan et al., 2007b] for the above regret bound to obtain a near optimal closed form solution to λ_t. One could employ an optimization algorithm to find the optimal λ_t.

Chapter 6

Conclusion

This thesis studied the problem of bounding the performance of machine learning algorithms in both statistical and adversarial setting, and designing computationally efficient algorithms with strong theoretical guarantees.

The major contributions of this thesis are:

- An investigation of the influence of cost terms on the hardness of the cost-sensitive classification problem by extending the minimax lower bound analysis for balanced binary classification (Theorem 3.6).
- A relationship between the contraction coefficient of a channel w.r.t. c-primitive f-divergence, and a generalized form of Dobrushin's coefficient (Theorem 3.10).
- An increased understanding of contraction coefficients of binary symmetric channels w.r.t. any symmetric f-divergence (Section 3.3.2).
- A complete characterization of the exp-concavity of any proper composite loss (Proposition 4.4). Using this characterization and the mixability condition of proper losses (Van Erven et al. [2012]), we showed that it is possible to re-parameterize any β-mixable binary proper loss into a β-exp-concave composite loss with the same β (Corollary 4.8).
- Analysis of unified update rules of the accelerated online convex optimization algorithms (Sections 5.2 and 5.3). Improved regret bounds were achieved by exploiting the easy nature of the sequence of outcomes.

By studying the geometry of prediction problems we have obtained insights into the factors which conspire to make the problem hard. These insights have enabled us to place bounds on the learning algorithm's ability to accurately predict the unseen data, and guided us in the design of better solutions.

Throughout the thesis, we have pointed out a number of open questions which we feel are important. These include the following:

1. Extend the study of the contraction coefficients of binary symmetric channels (Section 3.3.2) w.r.t. symmetric f-divergences to k-ary symmetric channels (with $k > 2$) and general f-divergences.

2. There are some divergences other than f-divergences which satisfy the weak data processing inequality, such as Neyman-Pearson α-divergences ([Polyanskiy and Verdú, 2010; Raginsky, 2011]). Thus it would be interesting to study strong data processing inequalities w.r.t. those divergences as well.

3. Recently people have attempted to relate several types of channel ordering to the strong data processing inequalities ([Makur and Polyanskiy, 2016; Polyanskiy and Wu, 2015]). It is worth to explore the relationship between the statistical deficiency based channel ordering (Raginsky [2011]) and the strong data processing inequalities.

4. It would be interesting to study the hardness of the cost-sensitive classification with example dependent costs ([Zadrozny and Elkan, 2001; Zadrozny et al., 2003]), and the binary classification problem w.r.t. generalized performance measures (Koyejo et al. [2014]) such as arithmetic, geometric and harmonic means of the true positive and true negative rates.

5. Further study could be undertaken in applying the cost-sensitive privacy notion to some real-world problems.

6. We illustrated the impact of the choice of substitution function in Aggregating Algorithm with experiments conducted on a synthetic dataset and a number of different real-world data sets (Section 4.4.1). A theoretical understanding of the choice of substitution functions would be very useful.

7. An efficiently computable β-exp-concavifying link function for β-mixable multi-class proper losses, is still not known. At least showing a negative result would be worth, for example, showing that it is not possible to exp-concavify a multi-class (with $n > 2$) square loss with same mixability constant.

8. Develop adaptive and optimistic variants of second order online learning algorithms such as online Newton step (Hazan et al. [2007a]), efficiently using sketching methods (Woodruff [2014]).

Exploring and exploiting the geometric structure of the learning problem will serve as a guiding light in this future work.

Bibliography

Abe, N.; Zadrozny, B.; and Langford, J., 2004. An iterative method for multi-class cost-sensitive learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 3–11. ACM. (cited on page 23)

Abernethy, J.; Bartlett, P. L.; Rakhlin, A.; and Tewari, A., 2008. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the Annual Conference on Learning Theory*. (cited on page 97)

Ahlswede, R. and Gács, P., 1976. Spreading of sets in product spaces and hypercontraction of the Markov operator. *The annals of probability*, 4 (1976), 925–939. (cited on page 41)

Ali, S. M. and Silvey, S. D., 1966. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, (1966), 131–142. (cited on page 15)

Assouad, P., 1983. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296, 23 (1983), 1021–1024. (cited on pages 29, 31, and 33)

Banerjee, A.; Guo, X.; and Wang, H., 2005. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51, 7 (2005), 2664–2669. (cited on page 85)

Beck, A. and Teboulle, M., 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31, 3 (2003), 167–175. (cited on page 100)

Bhatia, K.; Jain, P.; Kamalaruban, P.; and Kar, P., 2016. Efficient and consistent robust time series analysis. *arXiv preprint arXiv:1607.00146*, (2016). (cited on page 3)

Blackwell, D., 1951. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, vol. 1, 93–102. (cited on page 5)

Boyd, S. and Vandenberghe, L., 2004. *Convex optimization*. Cambridge university press. (cited on page 102)

Buja, A.; Stuetzle, W.; and Shen, Y., 2005. Loss functions for binary class probability estimation and classification: Structure and applications. (2005). (cited on pages 65, 69, and 95)

Cardie, C. and Nowe, N., 1997. Improving minority class prediction using case-specific feature weights. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 57–65. (cited on page 40)

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 112)

Cesa-Bianchi, N. and Lugosi, G., 2006. *Prediction, learning, and games*. Cambridge university press. (cited on pages 2 and 68)

Chatzikokolakis, K.; Andrés, M. E.; Bordenabe, N. E.; and Palamidessi, C., 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, 82–102. Springer. (cited on page 60)

Chiang, C.-K.; Yang, T.; Lee, C.-J.; Mahdavi, M.; Lu, C.-J.; Jin, R.; and Zhu, S., 2012. Online optimization with gradual variations. In *Proceedings of the Annual Conference on Learning Theory*, 6.1–6.20. (cited on pages 3, 98, 100, 101, 102, 103, and 105)

Cohen, J. E.; Iwasa, Y.; Rautu, G.; Ruskai, M. B.; Seneta, E.; and Zbaganu, G., 1993. Relative entropy under mappings by stochastic matrices. *Linear algebra and its applications*, 179 (1993), 211–235. (cited on pages 23, 41, and 43)

Csiszár, I., 1972. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2, 1-4 (1972), 191–213. (cited on page 15)

Dawid, A. P., 2007. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59, 1 (2007), 77–93. (cited on page 5)

DeGroot, M. H., 1962. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33 (1962), 404–419. (cited on page 5)

Devroye, L.; Györfi, L.; and Lugosi, G., 2013. *A probabilistic theory of pattern recognition*, vol. 31. Springer Science & Business Media. (cited on page 25)

Do, C. B.; Le, Q. V.; and Foo, C.-S., 2009. Proximal regularization for online and batch learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 257–264. ACM. (cited on pages 108 and 110)

Dobrushin, R. L., 1956a. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1, 1 (1956), 65–80. (cited on page 41)

Dobrushin, R. L., 1956b. Central limit theorem for nonstationary Markov chains. II. *Theory of Probability & Its Applications*, 1, 4 (1956), 329–383. (cited on page 41)

Dragomir, S. S., 2000. Some Gronwall type inequalities and applications. *RGMIA Monographs, Victoria University, Australia*, 19 (2000). (cited on page 89)

Duchi, J., 2016; accessed March 30, 2017. Statistics and information theory. https://stanford.edu/class/stats311/. (cited on page 29)

Duchi, J.; Hazan, E.; and Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12 (2011), 2121–2159. (cited on pages 3, 98, 103, 105, 106, and 114)

Duchi, J.; Jordan, M.; and Wainwright, M., 2013. Local privacy and statistical minimax rates. In *54th Annual Symposium on Foundations of Computer Science (FOCS)*, 429–438. IEEE. (cited on pages 57 and 58)

Dwork, C., 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, 1–19. Springer. (cited on page 58)

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A., 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 265–284. Springer. (cited on page 57)

Eguchi, S. and Copas, J., 2001. Recent developments in discriminant analysis from an information geometric point of view. *Journal of the Korean Statistical Society*, 30 (2001), 247–264. (cited on page 15)

Elkan, C., 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, vol. 17, 973–978. Lawrence Erlbaum Associates Ltd. (cited on page 40)

Gneiting, T. and Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 477 (2007), 359–378. (cited on page 69)

Grünwald, P. D. and Dawid, A. P., 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics*, 32 (2004), 1367–1433. (cited on page 5)

Guntuboyina, A.; Saha, S.; and Schiebinger, G., 2014. Sharp inequalities for f-divergences. *IEEE Transactions on Information Theory*, 60, 1 (2014), 104–121. (cited on page 19)

Hand, D. J., 1994. Deconstructing statistical questions. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 157, 3 (1994), 317–356. (cited on page 65)

Hand, D. J. and Vinciotti, V., 2003. Local versus global models for classification problems: Fitting models where it matters. *The American Statistician*, 57, 2 (2003), 124–131. (cited on page 65)

Harremoes, P. and Vajda, I., 2011. On pairs of f-divergences and their joint range. *IEEE Transactions on Information Theory*, 57, 6 (2011), 3230–3235. (cited on page 19)

Haussler, D.; Kivinen, J.; and Warmuth, M. K., 1998. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44, 5 (1998), 1906–1925. (cited on page 85)

Hazan, E.; Agarwal, A.; and Kale, S., 2007a. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69, 2-3 (2007), 169–192. (cited on pages 2, 98, and 120)

Hazan, E. and Kale, S., 2010. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine Learning*, 80, 2-3 (2010), 165–188. (cited on page 100)

Hazan, E.; Rakhlin, A.; and Bartlett, P. L., 2007b. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems*, vol. 20, 65–72. (cited on pages 3, 98, 99, 106, 108, 115, and 118)

Hiriart-Urruty, J.-B. and Lemaréchal, C., 1993. *Convex analysis and minimization algorithms I: Fundamentals*, vol. 305. Springer. (cited on pages 6 and 67)

Isaacson, D. L. and Madsen, R. W., 1976. *Markov chains, theory and applications*, vol. 4. Wiley New York. (cited on page 41)

Kairouz, P.; Oh, S.; and Viswanath, P., 2014. Extremal mechanisms for local differential privacy. In *Advances in neural information processing systems*, vol. 27, 2879–2887. (cited on page 59)

Kakade, S. M. and Tewari, A., 2009. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, vol. 22, 801–808. (cited on page 112)

Kalnishkan, Y. and Vyugin, M. V., 2005. The weak aggregating algorithm and weak mixability. In *Proceedings of the Annual Conference on Learning Theory*, 188–203. Springer. (cited on page 68)

Kivinen, J. and Warmuth, M. K., 1999. Averaging expert predictions. In *Proceedings of the Annual Conference on Learning Theory*, 153–167. Springer. (cited on pages 3, 65, 70, 78, and 79)

Komiya, H., 1988. Elementary proof for Sion's minimax theorem. *Kodai Mathematical Journal*, 11, 1 (1988), 5–7. (cited on page 12)

Koyejo, O. O.; Natarajan, N.; Ravikumar, P. K.; and Dhillon, I. S., 2014. Consistent binary classification with generalized performance metrics. In *Advances in Neural Information Processing Systems*, vol. 27, 2744–2752. (cited on pages 39 and 120)

Le Cam, L., 1964. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, (1964), 1419–1455. (cited on page 5)

Le Cam, L., 2012. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media. (cited on pages 5, 8, 29, and 31)

Liese, F. and Vajda, I., 2006. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52, 10 (2006), 4394–4412. (cited on page 17)

Makur, A. and Polyanskiy, Y., 2016. Comparison of channels: criteria for domination by a symmetric channel. ArXiv:1609.06877 (cs.IT). (cited on pages 45, 62, and 120)

Massart, P. and Nédélec, É., 2006. Risk bounds for statistical learning. *The Annals of Statistics*, (2006), 2326–2366. (cited on pages 23, 34, and 39)

McMahan, H. B., 2014. A survey of algorithms and analysis for adaptive online learning. ArXiv:1403.3465 (cs.LG). (cited on pages 98 and 99)

Mitchell, T., 1997. *Machine Learning*. McGraw-Hill, Inc. (cited on page 1)

Mohri, M. and Yang, S., 2015. Accelerating optimization via adaptive prediction. *arXiv preprint*, (2015). ArXiv:1509.05760 (stat.ML). (cited on page 99)

Mohri, M. and Yang, S., 2016. Accelerating online convex optimization via adaptive prediction. In *Artificial Intelligence and Statistics*, 848–856. (cited on page 106)

Neyman, J. and Pearson, E., 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London Series A*, 231 (1933), 289–337. (cited on page 15)

Orabona, F.; Jie, L.; and Caputo, B., 2010. Online-batch strongly convex multi kernel learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 787–794. IEEE. (cited on page 110)

Paz, A., 1971. *Introduction to probabilistic automata (Computer science and applied mathematics)*. Academic Press, Inc. (cited on page 41)

Polyanskiy, Y. and Verdú, S., 2010. Arimoto channel coding converse and Rényi divergence. In *48th Annual Conference on Communication, Control, and Computing (Allerton)*, 1327–1333. IEEE. (cited on pages 62 and 120)

Polyanskiy, Y. and Wu, Y., 2015. Strong data-processing inequalities for channels and Bayesian networks. ArXiv:1508.06025 (cs.IT). (cited on pages 62 and 120)

Polyanskiy, Y. and Wu, Y., 2016; accessed March 30, 2017. Lecture notes on information theory. http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf. (cited on page 19)

Raginsky, M., 2011. Shannon meets Nlackwell and Le Cam: Channels, codes, and statistical experiments. In *International Symposium on Information Theory (ISIT)*, 1220–1224. IEEE. (cited on pages 62, 63, and 120)

Raginsky, M., 2014. Strong data processing inequalities and ϕ-Sobolev inequalities for discrete channels. ArXiv:1411.3575 (cs.IT). (cited on pages 2, 41, 43, and 44)

Raginsky, M., 2015; accessed March 30, 2017. Minimax lower bounds (statistical learning theory). http://maxim.ece.illinois.edu/teaching/fall15b/notes/minimax.pdf. (cited on page 30)

Rakhlin, A. and Sridharan, K., 2012. Online learning with predictable sequences. *arXiv preprint*, (2012). ArXiv:1208.3728 (stat.ML). (cited on pages 98 and 100)

Reid, M. D. and Williamson, R. C., 2010. Composite binary losses. *Journal of Machine Learning Research*, 11 (2010), 2387–2422. (cited on page 75)

Reid, M. D. and Williamson, R. C., 2011. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12 (2011), 731–817. (cited on pages 1, 5, 17, 46, and 69)

Rockafellar, R. T., 1970. *Convex analysis*. Princeton University Press. (cited on pages 5, 6, and 66)

Scott, C. et al., 2012. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6 (2012), 958–992. (cited on page 28)

Shalev-Shwartz, S., 2011. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4, 2 (2011), 107–194. (cited on pages 3, 97, and 98)

Srebro, N.; Sridharan, K.; and Tewari, A., 2011. On the universality of online mirror descent. In *Advances in neural information processing systems*, vol. 24, 2645–2653. (cited on page 100)

Torgersen, E., 1991. *Comparison of statistical experiments*. Cambridge University Press. (cited on pages 5, 8, and 15)

Tsybakov, A. B., 2004. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32 (2004), 135–166. (cited on page 34)

Tsybakov, A. B., 2009. *Introduction to Nonparametric Estimation*. Springer. (cited on pages 19 and 29)

van Erven, T., 2012. From exp-concavity to mixability. http://www.timvanerven.nl/blog/2012/12/from-exp-concavity-to-mixability/. (cited on pages 65 and 71)

Van Erven, T.; Reid, M. D.; and Williamson, R. C., 2012. Mixability is Bayes risk curvature relative to log loss. *Journal of Machine Learning Research*, 13, 1 (2012), 1639–1663. (cited on pages 66, 67, 69, 77, 96, and 119)

Vapnik, V., 1998. *Statistical learning theory*. Wiley New York. (cited on page 1)

Vapnik, V. and Chervonenkis, A. Y., 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16, 2 (1971), 264. (cited on page 63)

Vapnik, V. N. and Chervonenkis, A. Y., 1974. *Theory of Pattern Recognition [in Russian]*. Nauka. (cited on page 34)

Von Neumann, J. and Morgenstern, O., 1944. *Theory of games and economic behavior*. Princeton University Press. (cited on page 5)

Vovk, V., 1995. A game of prediction with expert advice. In *Proceedings of the Annual Conference on Learning Theory*, 51–60. ACM. (cited on pages 2, 3, 65, 68, 70, and 78)

Vovk, V., 2001. Competitive on-line statistics. *International Statistical Review*, 69, 2 (2001), 213–248. (cited on pages 71, 78, and 80)

Vovk, V. and Zhdanov, F., 2009. Prediction with expert advice for the Brier game. *Journal of Machine Learning Research*, 10 (2009), 2445–2471. (cited on pages 68 and 84)

Wald, A., 1949. Statistical decision functions. *The Annals of Mathematical Statistics*, 11 (1949), 165–205. (cited on page 5)

Warner, S. L., 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 309 (1965), 63–69. (cited on page 57)

Wasserman, L. and Zhou, S., 2010. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105, 489 (2010), 375–389. (cited on pages 57 and 58)

Williamson, R. C., 2014. The geometry of losses. In *Proceedings of the Annual Conference on Learning Theory*, 1078–1108. (cited on pages 73, 78, and 79)

Williamson, R. C.; Vernet, E.; and Reid, M. D., 2016. Composite multiclass losses. *Journal of Machine Learning Research*, 17, 223 (2016), 1–52. (cited on pages 65, 69, 73, 75, and 87)

Woodruff, D. P., 2014. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10, 1–2 (2014), 1–157. (cited on page 120)

Yang, Y. and Barron, A., 1999. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27 (1999), 1564–1599. (cited on page 29)

Yu, B., 1997. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, 423–435. Springer. (cited on page 29)

Zadrozny, B. and Elkan, C., 2001. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the 7th International Conference on Knowledge discovery and data mining*, 204–213. ACM. (cited on pages 39 and 120)

Zadrozny, B.; Langford, J.; and Abe, N., 2003. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the Third IEEE International Conference on Data Mining*, 435–442. IEEE. (cited on pages 39 and 120)

Zinkevich, M., 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the International Conference on Machine Learning*, 928–936. (cited on pages 2 and 98)