Machine Learning via Transitions

Brendan van Rooyen

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

Except where otherwise indicated, this thesis is my own original work.

Brendan van Rooyen
25 September 2015

To my wife Sarah. Finally a return on your investment!

Acknowledgments

A great many people and institutions have helped me complete this thesis. Firstly, I would like to thank my parents Karen and Bruce, without their input I can safely say I would not be in this position! I was very kindly supported by both the Australian National University and National ICT Australia. I thank both for creating a fantastic environment for learning. I would like to thank all of my friends from both the ANU and NICTA, including but not limited to Lachlan Chislett, Tom Chen, Kiara Bruggeman, Beau Johnston, Alexandra Rodriguez, Avi Ruderman, Nikete Della Penna, Giorgio Patrini and Daniel McNamara. Thank you all for keeping me sane these last three years.

I was lucky enough to spend the entirety of my PhD within the machine learning research group in NICTA's Canberra lab. It was fantastic being around such a vibrant and engaged group of people. Thanks for all the afternoon teas, Friday lunches and of course Kioloa retreats! Special mention goes to Nishant Mehta, Cheng Soon Ong, Justin Domke, Christfried Webers, Felipe Trevizan and all the other members of the Friday lunch gang.

To my committee of supervisors, Mark Reid, Aditya Menon and Bob Williamson, it has been an absolute blast. I have learnt so much from all of you and have thoroughly enjoyed our interactions. Thanks for putting up with my eclectic interests, terse presentation and (particularly Bob) my developing writing skills. You have all shown me the joys of research. To Aditya, always remember to be unhinged!

Finally to my wife Sarah. Thanks for helping me through the frustration of edits, negative reviews and for putting up with the long pauses in our conversations while my mind was elsewhere. Without your help I would not be the person I am today. I can't wait for what comes next.

Abstract

This thesis presents a clear conceptual basis for theoretically studying machine learning problems. Machine learning methods afford means to automate the discovery of relationships in data sets. A relationship between quantities X and Y allows the prediction of one quantity given information of the other. It is these relationships that we make the central object of study. We call these relationships transitions.

A transition from a set X to a set Y is a function from X into the probability distributions on Y. Beginning with this simple notion, the thesis proceeds as follows:

- Utilizing tools from statistical decision theory, we develop an abstract language for quantifying the information present in a transition.
- We attack the problem of generalized supervision. Generalized supervision is the learning of classifiers from non-ideal data. An important example of this is the learning of classifiers from noisily labelled data. We demonstrate the virtues of our abstract treatment by producing generic methods for solving these problems, as well as producing generic upper bounds for our methods as well as lower bounds for any method that attempts to solve these problems.
- As a result of our study in generalized supervision, we produce means to define procedures that are robust to certain forms of corruption. We explore, in detail, procedures for learning classifiers that are robust to the effects of symmetric label noise. The result is a classification algorithm that is easier to understand, implement and parallelize than standard kernel based classification schemes, such as the support vector machine and logistic regression. Furthermore, we demonstrate the uniqueness of this method.
- Finally, we show how many feature learning schemes can be understood via our language. We present well motivated objectives for the task of learning features from unlabelled data, before showing how many standard feature learning methods (such as PCA, sparse coding, auto-encoders and so on) can be seen as minimizing surrogates to our objective functions.

Contents

Appendix **111**

List of Figures

1

Introduction

> ***The Problem:*** *The massive reduction in the cost of collecting, storing, transporting and processing data has meant an increasing need for tools to make sense of it. Unfortunately, the deployment of modern machine learning tools is more akin to a craft than an engineering discipline: the inference problems to be solved are often under-specified or ill-posed and the available tools are often adhoc - lacking generality, transparency, usability and interoperability. Our premise is that the root cause of these difficulties is a lack of a clear conceptual basis for machine learning as an information engineering discipline.*
>
> - Robert C. Williamson, *Reconceiving Machine Learning*

This thesis presents part of the required conceptual basis. Machine learning methods afford means to automate the discovery of relationships in data sets. A relationship between quantities X and Y allows the *prediction* of one quantity given information of the other. It is these relations that we make the central object of study. We call these relations *transitions*.

Definition 1.1. *A* transition *from a set X to a set Y is a function $T : X \to \mathbb{P}(Y)$, from X into the probability distributions on Y.*

Intuitively, a transition T summarizes the uncertainty in *predicting* the quantity Y given the observation of a quantity X. Many concepts in machine learning; such as forecasts, probabilistic models, experiments, algorithms, conditional probabilities, randomized decision rules, communication channels and so on, can *all* be understood abstractly as transitions. Transitions therefore will serve as a guiding light.

1.1 Outline of the Thesis

This thesis is the result of a three year journey that sought to understand transitions, with special focus paid to their application in machine learning theory. Some definitions are repeated, and there is slight variations in notation for each chapter. Ultimately, there is no single best notational system, effort has been placed into using the notation that most clearly explains the contents of the chapter.

Chapter one provides the abstract language underpinning the rest of the thesis. Utilizing tools from statistical decision theory, it shows how to *define* and *compare* the information content in transitions. We do this through the notion of *risk*. The risk has been shown previously to include a large number of information functions present in the literature, including the often used mutual information and KL-divergence [62; 105]. While most of the material is review, its presentation is greatly streamlined through the focus on transitions. Its novel contributions include; a generalization of the data processing theorem of information theory [46] (theorem 2.28) and means to calculate deficiency distances [85] via linear programming (lemma 2.35).

Chapter two considers the problem of learning from corrupted data. In the normal theoretical analysis of supervised learning algorithms, it is assumed the decision maker has access to *clean* data, that their observations are from the pattern they are expected to predict [25]. In the real world this is usually not the case, data is normally corrupted and real world data sets are amalgamations of data of variable type and quality. Understanding how to *learn from* and *compare* different corrupted data sets is therefore a problem of great practical importance. This chapter provides a simple correction to ERM style algorithms (theorem 3.2) that facilitates learning from a large class of corrupted data sets. Furthermore, upper and lower bounds for this problem (theorems 3.4 and 3.15) are presented. These bounds allow the *comparison* of different corrupted data sets.

Chapter three focuses on one *particular* corrupted learning problem, namely the learning of classifiers under symmetric label noise. A conceptually simple, easily parallelized and robust classification algorithm is motivated and analysed. This algorithm highlights the practical benefits that *focusing* on transitions can bring.

Finally, chapter four utilizes transitions in the study of feature learning methods. Study began in this direction to take up a challenge posed by Yann LeCun to the machine learning theory community at the Conference on Learning Theory 2013 [87]. It presents means to *quantify* the quality of learnt features *independently* of the supervised learning algorithm the features are used in. It is an attempt to provide a conceptual foundation of unsupervised, "deep" methods for the automated learning of features. Theorems 5.2 and 5.4 *characterize* when it is possible to learn generically good features from unlabelled data. Theorem 5.5 motivates several unsupervised learning algorithms as surrogate approaches to minimizing the quantities in theorem 5.2. Furthermore, we explore supervised feature learning algorithms and show their relationship to risks and deficiency presented in chapter one.

The following work was completed during the thesis [100]. While certainly in the same spirit as the other material, it has been excluded from the thesis as it does not fit as well with the theme of transitions as the work presented here.

2

Decision Theory, Transitions and Experiments

> **Science**
>
> The intellectual and practical activity encompassing the systematic study of the structure and behaviour of the physical and natural world through observation and experiment.

The scientific process is a means for turning the results of experiments into knowledge about the world in which we live. Much recent research effort has been directed toward *automating* the scientific process. To do this, one needs to *formulate* the scientific process in a *precise* mathematical language. This chapter specifies one such language. What is presented here is hardly new. The material leans much on great thinkers of times past [23; 53; 85; 125; 126] as well as more modern contributions [52; 67; 86; 105; 117]. It serves as the conceptual foundation for this thesis. The presentation is abstract; this is intentional. By laying bare the *basic* language, we remove the distraction that focusing on *specific* problems brings, and we expose the common elements all representable by transitions.

2.1 Basic Notation

We require the following notation. Let $\mathbb{R}_+$ be the set of non-negative real numbers. Let Y^X be the set of functions with domain X and range Y. For a set X define the functions $\mathrm{id}_X(x) = x$, and $\mathbf{1}_X(x) = 1$. For a function $f \in \mathbb{R}^{X \times Y}$ and $y \in Y$, we denote the *partial* function $f(-, y) \in \mathbb{R}^X$, with $f(-, y)(x) = f(x, y)$, with similar notation for fixing the first argument. We denote the *dual space* of $\mathbb{R}^X$, the set of linear maps $\mathbb{R}^X \to \mathbb{R}$, by $(\mathbb{R}^X)^*$. Finally, for a boolean predicate $p : X \to \{\mathsf{True}, \mathsf{False}\}$, let $[\![p(x)]\!] = 1$ if $p(x)$ is true and 0 otherwise. Other notation will be developed as necessary.

2.2 A Simple, Motivating Example

Consider the problem faced by a scientist in a laboratory. In front of them is a beaker, containing one of a number of possible substances. Available to them are a myriad of experiments that can be performed to identify the unknown substance. The scientist could attempt to ignite it, mix a bit of it with some other known substance and see what happens, x-ray a sample, throw some of it at high velocity toward an oncoming beam of electrons and so on. Due to time and budget constraints, only a limited number of experiments can be performed to ascertain the substances true identity. Therefore the scientist should focus their effort on the "most informative" experiments. Of course, what is informative is dependent on how the substance is to be *used*. For example, if the scientist wishes to sprinkle some of it on their food to enhance its flavour, misidentifying arsenic as table salt is a very bad idea. However, if they want to sprinkle it on the snails in their garden, this distinction is less important. The focus of this chapter is the abstract formulation of this problem. We present a language for making decisions under uncertainty, a definition of an experiment and finally means to quantify the information contained in an experiment.

2.3 The General Decision Problem

We consider the problem of how a decision maker, or scientist, uses observations from experiments to inform their decisions. Let Θ be a set of possible values of some unknown quantity, and A the set of actions available to the decision maker. The consequence of an action is measured by a loss function $L : \Theta \times A \to \mathbb{R}$. A negative loss represents a gain to the decision maker. In light of our previous example, Θ are the possible substances that could be in the beaker, A is what the decision maker can *do* with the substance (eat it, put it on snails and so on), and L measures the consequence of an action to the scientist ($L(\text{arsenic}, \text{eat})$ should be high). The norm of a loss function is given by its largest possible consequence (positive or negative), $\|L\|_\infty = \max_{\theta,a} |L(\theta, a)|$.

To avoid measure theoretic technicalities, we assume Θ to be finite and A to be closed, compact, set with L a continuous function. This ensures that infima of all the quantities defined can be replaced by minima. Ultimately the methods suggested in this thesis will (hopefully) run on a computer, meaning all real world objects we wish to simulate need to be approximated by elements of a finite set. For those that feel this restriction places severe limitations on the theory developed here, we point out that all the following can be proven in the more general setting, at the cost of a more technical presentation. For example the reader is directed to theorem 6.2.12 of Torgersen [117], for direction on how results for finite Θ can be extended to those for infinite Θ.

The decision maker seeks an action a that has low loss $L(\theta, a)$ on the true $\theta \in \Theta$. Due to their limited information concerning the phenomena, the decision maker does not know the exact value of θ. They may have a vague idea of which θ are more or less likely to occur. This uncertainty is represented by a probability distribution.

Definition 2.1. *A* probability distribution *on a set X is an element of $(\mathbb{R}^X)^*$, i.e. a linear function $\mathbb{E}_P : \mathbb{R}^X \to \mathbb{R}$, such that:*

1. $\mathbb{E}_P(\mathbf{1}_X) = 1$.
2. *If $f(x) \leq g(x)$, $\forall x \in X$ then $\mathbb{E}_P(f) \leq \mathbb{E}_P(g)$.*

The linear function $\mathbb{E}_P$ is called an expectation. For a large class of general topological spaces, this definition is equivalent to the usual one in terms of measures on sigma algebras [1] [79]. Focusing on the expectation operator rather than its representation via measures on sigma algebras and Lebesgue integrals provides means to "abstract away" the sample space. The function f can be thought of as a gamble taken by the decision maker, with $f(x)$ being the loss incurred if the outcome x is observed. The expectation $\mathbb{E}_P(f)$ is the total loss assigned to f by the decision maker. The gamble f is preferred to the gamble g if it has lower expected loss. Expectation is *one* way of ordering gambles. The first condition can be seen as a normalization, loss 1 is assigned to the constant gamble $\mathbf{1}_X$. The second condition can be seen as the sensible requirement that if f always offers lower loss than g, the decision maker always prefers f to g. If necessary, we use the notation $\mathbb{E}_{x\sim P} f(x)$ to make clear what quantity we are taking expectations over. We drop the subscript when this is clear from context. We also make use of the following infix notation to denote expectations,

$$\mathbb{E}_{x\sim P} f(x) = \langle P, f \rangle_X.$$

This notation for expectations is not standard, angled brackets are normally reserved to denote inner products. For finite spaces expectations are *exactly* inner products. We continue with both notations where appropriate. In particular the infix notation makes it easier to see connections to concepts in functional analysis, such as adjoint operators, that are so key to the ideas presented in chapter 3.

Denote the set of probability distributions on X by $\mathbb{P}(X)$, and the set of un-normalized distributions (those linear functions for which property (1) in definition 2.1 does not hold) by $\mathbb{P}^+(X)$. This is a convex subset of $(\mathbb{R}^X)^*$. For $x \in X$, denote the point mass distribution on x by δ_x. A probability distribution $P \in \mathbb{P}(\Theta)$ facilitates the decision making process. The decision maker acts by choosing the action with minimum expected loss,

$$\operatorname*{arg\,min}_{a\in A} \mathbb{E}_{\theta\sim P} L(\theta, a).$$

The key question is *which* distribution to use. To *discover* this distribution, the decision maker is guided by experiments. Before we discuss the language for defining

[1] The key is to identify sets with their indicator functions. $P \in (\mathbb{R}^X)^*$ defines a measure via $P(C) := P(\mathbf{1}_C)$.

and ordering experiments, we first focus on how to construct suitable loss functions. We show that the essential properties of a loss needed for the decision making process are encoded in its corresponding *entropy*,

$$\underline{L}(P) = \min_{a \in A} \mathbb{E}_{\theta \sim P} L(\theta, a).$$

We show that each entropy defines a *canonical* loss function. We also show that for the sake of developing theory, one only need work with *canonical* losses.

Aside: Loss versus Regret

It is quite natural in decision theory to work with the regret,

$$\Delta L(P, a) = \mathbb{E}_{\theta \sim P} L(\theta, a) - \min_{a' \in A} \mathbb{E}_{\theta \sim P} L(\theta, a'),$$

which measures the excess loss of the decision makers action versus the loss of the optimal action if they knew P. Here we focus on loss, elsewhere we focus on regret.

2.4 Representing Loss Functions

In this section we make heavy use of the infix notation for expectations, as well as partial functions. In its partial form, a loss provides a mapping $\text{partial}_L : A \to \mathbb{R}^\Theta$ with,

$$\text{partial}_L(a) = L(-, a) \in \mathbb{R}^\Theta.$$

In words, when the decision maker chooses an action a, they specify a function that takes the unknown and returns the loss incurred by the decision maker. Choosing an action is then equivalent to picking a partial loss function. In our notation,

$$\mathbb{E}_{\theta \sim P} L(\theta, a) = \langle P, L(-, a) \rangle_\Theta.$$

In many statistical problems, it is natural for the space of actions A to be the set of distributions over unknowns $\mathbb{P}(\Theta)$.

Definition 2.2. *A loss $L : \Theta \times \mathbb{P}(\Theta) \to \mathbb{R}$ is* proper *if for all distributions $P \in \mathbb{P}(\Theta)$,*

$$P \in \operatorname*{arg\,min}_{Q \in \mathbb{P}(\Theta)} \langle P, L(-, Q) \rangle_\Theta.$$

It is strictly proper *if P is the* unique *minimizer.*

Intuitively, a proper loss takes a prediction $Q \in \mathbb{P}(\Theta)$, and then penalizes the decision maker according to how much weight their prediction assigned to the unknown θ. Intuitively properness ensures that if the decision maker *knows* P, then they minimize their expected loss by *reporting* P. Proper losses constitute a well studied class of loss functions, that provide suitable surrogates for decision problems

[7; 31; 52; 52; 64; 67; 106; 133].

As will be shown, all ″sensible" losses are essentially re-parametrized proper losses. We show how to *construct* proper losses from their entropies. Furthermore, we show how to render any proper loss convex through a canonical re-parametrization. This allows the use of tools from convex analysis [26; 90] to aid in calculating optimal actions.

2.4.1 Entropy from Loss

Rather than working with probability distributions, we take the route of Williamson in [129] and work with un-normalized distributions. For any loss function L, define the *entropy* $\underline{L} : \mathbb{P}^+(\Theta) \to \mathbb{R}$,

$$\underline{L}(\mu) = \min_{a \in A} \langle \mu, L(-, a) \rangle_\Theta.$$

$\underline{L}(P)$ measures the uncertainty of the optimal action for the distribution P. The entropy is also called an *uncertainty function*, a *Bayes risk* or a *support function* [53; 105; 129]. It is concave and 1-homogeneous.

Definition 2.3. *A function $f : \mathbb{P}^+(\Theta) \to \mathbb{R}$ is 1-homogeneous if for all $x \in \mathbb{P}^+(\Theta)$ and for all $\lambda > 0$,*

$$f(\lambda x) = \lambda f(x).$$

2.4.2 Loss from Entropy

All loss functions give rise to an entropy. Conversely, the entropy encodes much information of its associated loss through its *super-gradients*, which include all the *Bayes* actions for the underlying loss.

Bayes Actions and Super-gradients

For any distribution P, define the *Bayes actions* for P as the set of minimizers,

$$A_P = \arg\min_{a \in A} \langle P, L(-, a) \rangle.$$

For any $a_P \in A_P$ we have $\underline{L}(P) = \langle P, L(-, a_P) \rangle$.

Definition 2.4 (Super-gradient of a concave function [90])**.** *Let $f : \mathbb{P}^+(\Theta) \to \mathbb{R}$ be a concave function. $v \in \mathbb{R}^\Theta$ is a* super-gradient *of f at the point x if for all $y \in \mathbb{P}^+(\Theta)$,*

$$\langle y - x, v \rangle + f(x) \geq f(y).$$

Denote the set of all super-gradients at a point x by $\partial f(x)$, and the set of all super-gradients by $\partial f = \cup_x \partial f(x)$. For differentiable concave functions, super-gradients are the same as regular gradients [90]. 1-homogeneous functions afford a very simple representation via their super-gradients.

Theorem 2.5 (Generalized Euler's Homogeneous Function Theorem [51])**.** *Let $f : \mathbb{P}^+(\Theta) \to \mathbb{R}$ be a concave 1-homogeneous function. Then for all x and for all $v \in \partial f(x)$,*

$$f(x) = \langle x, v \rangle.$$

Furthermore, $v \in \partial f(x) \implies v \in \partial f(\lambda x)$ for all $\lambda > 0$.

We include a simple proof of this theorem for completeness.

Proof. Firstly, for all x and all $\lambda > 0$,

$$\langle \lambda x - x, v \rangle + f(x) \geq \lambda f(x),$$

which follows directly from the definition of a super-gradient at x and the 1-homogeneity of f. Re-arranging yields, $(1 - \lambda)(f(x) - \langle x, v \rangle) \geq 0$. Letting $\lambda \to 0^+$ yields $f(x) \geq \langle x, v \rangle$. Similarly, for all x and all $\lambda > 0$,

$$\langle x - \lambda x, v \rangle + \lambda f(x) \geq f(x),$$

which follows directly from the definition of a super-gradient at λx and the 1-homogeneity of f. Re-arranging yields, $(1 - \lambda)(f(x) - \langle x, v \rangle) \leq 0$. Letting $\lambda \to 0^+$ yields $f(x) \leq \langle x, v \rangle$, therefore $f(x) = \langle x, v \rangle$.

To prove the second claim, we have for all y and $\lambda > 0$,

$$\begin{aligned} &\langle y - x, v \rangle + f(x) \geq f(y) \\ &\langle \lambda y - \lambda x, v \rangle + f(\lambda x) \geq f(\lambda y), \end{aligned}$$

where the first line is by definition, and the second is by 1-homogeneity. As y is arbitrary, the claim is proved.

□

This theorem provides a corollary, that shows the super-gradients of a 1-homogeneous function have a property similar to properness.

Corollary 2.6. *Let $f : \mathbb{P}^+(\Theta) \to \mathbb{R}$ be a concave 1-homogeneous function. Then for all $x, y \in \mathbb{P}^+(\Theta)$ and for all $v_x \in \partial f(x)$, $v_y \in \partial f(y)$,*

$$\langle x, v_y \rangle \geq \langle x, v_x \rangle.$$

We now show that the partial loss of a Bayes action is a super-gradient of $\underline{L}$.

Theorem 2.7. *For all loss functions L and distributions P, $a_P \in A_P \Leftrightarrow L(-, a_P) \in \partial \underline{L}(P)$.*

Proof. For $a_P \in A_P$ we have for all $\mu \in \mathbb{P}^+(\Theta)$,

$$\langle \mu - P, L(-, a_P)\rangle + \underline{L}(P) = \langle \mu, L(-, a_P)\rangle \geq \min_{a\in A}\langle \mu, L(-,a)\rangle = \underline{L}(\mu).$$

Hence $L(-, a_P) \in \partial\underline{L}(P)$. For the converse, if $L(-, a_P) \in \partial\underline{L}(P)$ then,

$$\underline{L}(P) = \langle P, L(-, a_P)\rangle = \min_{a\in A}\langle P, L(-,a)\rangle,$$

meaning a is Bayes.

□

Therefore, once inadmissible actions are discarded, we can identify a loss with a subset of $\partial\underline{L}$. Rather than working with a subset $\partial\underline{L}$, it is advantageous to consider *all* of $\partial\underline{L}$.

Definition 2.8 (Canonical Loss)**.** *Let* $\underline{L} : \mathbb{P}^+(\Theta) \to \mathbb{R}$ *be a concave, 1-homogeneous function. Then its* canonical loss, $\mathcal{L} : \Theta \times \partial\underline{L} \to \mathbb{R}$ *is given by,* $\mathcal{L}(\theta, \ell) = \ell(\theta)$.

As will be shown, canonical losses can always be convexified. Furthermore, they maintain all of the properties of L needed for assessing the quality of decisions.

The Bayes Super Prediction Set

The process of *canonising* a loss, i.e. going from,

$$L \to \underline{L} \to \mathcal{L},$$

can create *extra* partial losses/actions that were not originally available to the decision maker under L. However, they gain no benefit from these extra actions. From any entropy define the *Bayes super prediction set*,

$$\mathcal{S}_{\underline{L}} := \left\{\ell \in \mathbb{R}^\Theta : \langle \mu, \ell\rangle \geq \underline{L}(\mu),\ \forall \mu \in \mathbb{R}^\Theta_+\right\}.$$

By the definition,

$$\min_{a\in A}\langle P, L(-,a)\rangle = \min_{\ell\in\mathcal{S}_{\underline{L}}}\langle P, \ell\rangle,\ \forall P \in \mathbb{P}(\Theta).$$

The Bayes super prediction set is precisely those partial losses that the decision maker need not use over the actions available to them, no matter the distribution P. The super prediction set is convex. Furthermore, the Bayes actions for $\mathcal{L}$ are the lower boundary of the super prediction set.

Lemma 2.9. *Let* $\underline{L} : \mathbb{P}^+(\Theta) \to \mathbb{R}$ *be a concave, 1-homogeneous function. Then* $\ell \in \partial\underline{L}$ *if and only if,*

$$\langle \mu, \ell\rangle \geq \underline{L}(\mu),\ \forall \mu \in \mathbb{P}^+(\Theta),$$

with equality holding for at least one μ.

The proof is a straightforward application of 1-homogeneity and super-gradients.

Admissible Actions

Bayes actions are one notion of optimal action. Admissibility affords another.

Definition 2.10. *Let L be a loss. An action a is* admissible *if there does not exist an action a^* such that,*

$$L(\theta, a^*) \leq L(\theta, a),\ \forall \theta \in \Theta,$$

with strict inequality for at least one θ.

Intuitively, an action is admissible if there is no other action that is obviously better. Bayesian actions are optimal *if* the decision maker has knowledge about the unknown, given in the form of a probability distribution. Interestingly, the class of admissible and Bayesian actions are the same for many loss functions.

Theorem 2.11 (Complete Class Theorem [24; 126])**.** *Let L be a loss such that* $\mathrm{im}(\mathrm{partial}_L)$ *is a convex subset of $\mathbb{R}^\Theta$. Then the set of Bayes actions for L is in 1-1 correspondence with the set of admissible actions for L.*

If the decision maker is allowed to used *randomized* actions, i.e. distributions over A with $L(\theta, Q) = \mathbb{E}_{a\sim Q} L(\theta, a)$, then all admissible actions are Bayesian actions.

Proper Losses

Using the canonical loss allows the construction of proper losses from entropies.

Lemma 2.12 (Loss from Entropy)**.** *Let $\underline{L} : \mathbb{P}^+(\Theta) \to \mathbb{R}$ be a concave, 1-homogeneous function and let $\nabla\underline{L} : \mathbb{P}^+(\Theta) \to \mathbb{R}^\Theta$ be a super-gradient function, $\nabla\underline{L}(\mu) \in \partial\underline{L}(\mu)$, $\forall\mu$. Then,*

$$L(\theta, Q) = \mathcal{L}(\theta, \nabla\underline{L}(Q)),$$

is a proper loss. Furthermore if $\underline{L}$ is strictly concave then L is strictly proper.

Regret for Proper Losses

Recall the notion of regret,

$$\Delta L(P, a) = \mathbb{E}_{\theta\sim P} L(\theta, a) - \min_{a'\in A} \mathbb{E}_{\theta\sim P} L(\theta, a').$$

For proper losses, the regret takes on the particularly elegant form,

$$\Delta L(P, Q) = \mathbb{E}_{\theta\sim P}\left[L(\theta, Q) - L(\theta, P)\right] = \langle P, \nabla\underline{L}(Q) - \nabla\underline{L}(P)\rangle.$$

The regret for a proper loss is also equal to the Bregman divergence between P and Q. We have,

$$\underbrace{\underline{L}(Q) + \langle P - Q, \nabla\underline{L}(Q)\rangle - \underline{L}(P)}_{\text{Bregman divergence induced by } \underline{L}} = \langle P, \nabla\underline{L}(Q) - \nabla\underline{L}(P)\rangle,$$

where we have used the fact that $\underline{L}(Q) = \langle Q, \nabla \underline{L}(Q) \rangle$.

2.4.3 Convexification of Losses in Canonical Form

The preceding shows how to *construct* losses, we begin with a convex 1-homogeneous function and then take super-gradients. Focus now turns to their convexification. Once convexified, the decision maker gains access to the large and ever growing literature on the minimization of convex functions to aid in the calculation of optimal actions. The development here closely follows that in [52], which focused on proper losses. Working with canonical versus proper losses streamlines the development. For example, for some proper losses lemma 2.14 fails to hold, while it does hold for all canonical losses. Furthermore, our result on convexification of canonical losses (theorem 2.16), is to the best of our knowledge novel.

Recall $\mathbf{1}_\Theta \in \mathbb{R}^\Theta$ is the function that always returns 1, and define $\mathbf{1}_\Theta^\perp$ to be its orthogonal complement in $\mathbb{R}^\Theta$, i.e. the functions $v \in \mathbb{R}^\Theta$ with,

$$\langle \mathbf{1}_\Theta, v \rangle = \sum_{z \in \Theta} v(z) = 0.$$

Define,

$$\Gamma_{\underline{L}} = \{ (\gamma, v) \in \mathbb{R} \times \mathbf{1}_\Theta^\perp : \gamma \mathbf{1}_\Theta + v \in \partial \underline{L} \ \}.$$

Lemma 2.13. *Let $(\gamma, v) \in \Gamma_{\underline{L}}$. Then γ is uniquely determined by v.*

Proof. Fix v and suppose there exists γ_1 and γ_2 with $\gamma_1 < \gamma_2$ and $\gamma_1 \mathbf{1}_\Theta + v, \gamma_2 \mathbf{1}_\Theta + v \in \partial \underline{L}$. By assumption, $\gamma_2 \mathbf{1}_\Theta + v$ is Bayes for some distribution P. But,

$$\langle P, \gamma_1 \mathbf{1}_\Theta + v \rangle = \gamma_1 + \langle P, v \rangle < \gamma_2 + \langle P, v \rangle = \langle P, \gamma_2 \mathbf{1}_\Theta + v \rangle,$$

a contradiction.

□

Thus we lose nothing by working with projections of losses onto $\mathbf{1}_\Theta^\perp$. Geometrically, we have the following sequence of maps,

$$\begin{tikzcd} \partial \underline{L} \arrow[r, "\mathrm{partial}_{\mathcal{L}}"] & \mathbb{R}^\Theta \arrow[r, "\mathrm{proj}_{\mathbf{1}_\Theta^\perp}"] & \mathbf{1}_\Theta^\perp, \end{tikzcd}$$

with $\mathrm{proj}_{\mathbf{1}_\Theta^\perp}$ the projection onto $\mathbf{1}_\Theta^\perp$. Lemma 2.13 shows that $\mathrm{proj}_{\mathbf{1}_\Theta^\perp} \circ \mathrm{partial}_{\mathcal{L}}$ is invertible. Define,

$$\hat{\Gamma}_{\underline{L}} = \mathrm{im}(\mathrm{proj}_{\mathbf{1}_\Theta^\perp} \circ \mathrm{partial}_{\mathcal{L}}) \subseteq \mathbf{1}_\Theta^\perp.$$

By lemma 2.13 $\hat{\Gamma}_{\underline{L}}$ is in 1-1 correspondence with $\partial \underline{L}$.

Lemma 2.14. *$\hat{\Gamma}_{\underline{L}}$ is a convex set.*

Proof. To show $\hat{\Gamma}_{\underline{L}}$ is convex, we are required to show that for all $\ell_1, \ell_2 \in \partial\underline{L}$ and all $\lambda \in [0,1]$ there is a constant γ such that,

$$\lambda\ell_1 + (1-\lambda)\ell_2 - \gamma\mathbf{1}_\Theta \in \partial\underline{L}.$$

By lemma 2.9, this is equivalent to,

$$\underbrace{\lambda\langle P,\ell_1\rangle + (1-\lambda)\langle P,\ell_2\rangle - \underline{L}(P)}_{\gamma(P)} - \gamma = \gamma(P) - \gamma \geq 0, \; \forall P \in \mathbb{P}(\Theta),$$

with equality holding for one P. Let $\gamma^* = \min_P \gamma(P)$, with P^* the distribution that achieves the minimum. Then,

$$\lambda\langle P,\ell_1\rangle + (1-\lambda)\langle P,\ell_2\rangle - \gamma^* \geq \underline{L}(P), \; \forall P \in \mathbb{P}(\Theta),$$

with equality for P^*. Therefore by lemma 2.9, $\lambda\ell_1 + (1-\lambda)\ell_2 + \gamma^*\mathbf{1}_\Theta \in \partial\underline{L}$.

□

Define the function $\Psi : \hat{\Gamma}_{\underline{L}} \to \mathbb{R}$ such that,

$$v + \Psi(v)\mathbf{1}_\Theta \in \partial\underline{L}.$$

By lemma 2.13, Ψ is well defined.

Lemma 2.15. Ψ *is a convex function.*

Proof. Let $v_1, v_2 \in \hat{\Gamma}_{\underline{L}}$ with $v_\lambda = \lambda v_1 + (1-\lambda)v_2$. Let their partial losses be,

$$\begin{aligned}\ell_1 &= v_1 + \Psi(v_1)\mathbf{1}_\Theta \\ \ell_2 &= v_2 + \Psi(v_2)\mathbf{1}_\Theta \\ \ell_\lambda &= \lambda v_1 + (1-\lambda)v_2 + \Psi(\lambda v_1 + (1-\lambda)v_2)\mathbf{1}_\Theta,\end{aligned}$$

respectively. By assumption, for all $\lambda \in [0,1]$ there exists a distribution P_λ such that,

$$\langle P_\lambda, \ell_\lambda\rangle \leq \langle P_\lambda, \ell\rangle, \; \forall \ell \in \partial\underline{L}.$$

Assume there is a λ^* such that,

$$\lambda^*\Psi(v_1) + (1-\lambda^*)\Psi(v_2) < \Psi(\lambda^* v_1 + (1-\lambda^*)v_2).$$

But then,

$$\langle P_{\lambda^*}, \lambda^*\ell_1 + (1-\lambda^*)\ell_2\rangle < \langle P_{\lambda^*}, \ell_{\lambda^*}\rangle,$$

a contradiction. □

This gives the following representation theorem for canonical losses.

Theorem 2.16 (Representation of Canonical Losses)**.** *Let $\underline{L} : \mathbb{P}^+(\Theta) \to \mathbb{R}$ be a concave, 1-homogeneous function. Then its canonical loss $\mathcal{L}$ can be represented as $\mathcal{L} : \Theta \times C \to \mathbb{R}$, with $C \subseteq \mathbf{1}_\Theta^\perp$ a convex set and,*

$$\mathcal{L}(\theta, v) = v(\theta) + \Psi(v),$$

for a convex function Ψ.

2.4.4 Example: Binary Decisions and Log Loss

For this example, take $\Theta = \{-1, 1\}$, with loss $L(-, p) = (-\log(1-p), -\log(p))$ for $p \in (0, 1)$, where p is the probability that $\theta = 1$. We plot this loss in 2.1. The partial losses are given by the red curve, the super prediction set in grey. The loss on negatives is plotted on the x-axis. In figure 2.2 we show geometrically how to produce canonical coordinates.

For canonical coordinates, We seek to decompose $L(-, p) = \gamma_1 v + \gamma_2 \mathbf{1}_\Theta$, where $v = \left(-\frac{1}{2}, \frac{1}{2}\right)$. Here we have projected $p = 0.8$ onto $\mathbf{1}_\Theta^\perp$. The length of the blue line is related to $\Psi(\gamma_1)$. Solving for γ_0 and γ_1 in terms of p gives,

$$\gamma_1 = \log\left(\frac{p}{1-p}\right) \text{ and } \gamma_2 = \frac{1}{2}\log\left(\frac{1}{p(1-p)}\right).$$

This equation can be easily solved for p, giving,

$$p = \frac{\exp(\gamma_1)}{1 + \exp(\gamma_1)} \text{ and } \gamma_2 = \Psi(\gamma_1) = \log\left(1 + e^{\gamma_1}\right) - \frac{\gamma_1}{2}.$$

The above relationship between p and γ_1 is exactly that given by the *canonical link* for log loss [104]. Finally for log loss,

$$\mathcal{L}(-, \gamma) = \gamma v + (\log\left(1 + e^{\gamma}\right) - \frac{\gamma}{2})\mathbf{1}_\Theta.$$

This yields,

$$\mathcal{L}(y, \gamma) = \log\left(1 + e^{-y\gamma}\right),$$

the usual form of logistic loss.

2.4.5 Misclassification and Linear Loss

For any observation space define the *misclassification* loss $L_{01} : \Theta \times \Theta \to \mathbb{R}$,

$$L_{01}(\theta, \theta') = [\![\theta \neq \theta']\!].$$

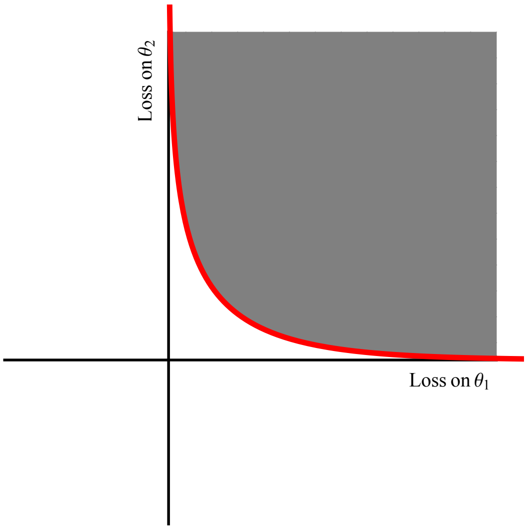


Fig. 2.1: Plot of super prediction set and its lower boundary for log loss, see text.

In words, the decision maker incurs loss 1 if their prediction is different to their observation and no loss otherwise. Allowing randomized actions gives *linear* loss,

$$L_{\text{linear}}(\theta, Q) = \mathbb{E}_{\theta' \sim Q} L_{01}(\theta, \theta').$$

In canonical form, linear loss can be written as $\mathcal{L}(-, v) = v + \frac{|\Theta|-1}{|\Theta|}\mathbf{1}_{\Theta}$. By theorem 2.16, linear loss is therefore the *primitive* loss, as it is the linear term in *all* other canonical losses. We will see in chapter 4, that linear loss provides means to learn classifiers.

2.5 Experiments

Recall that the decision maker chooses their actions by minimizing their expected loss,

$$\arg\min_{a \in A} \mathbb{E}_{\theta \sim P} L(\theta, a).$$

The key question is *which* distribution to use. To *discover* this distribution, the decision maker is guided by experiments. Let $\mathcal{Z}$ be a finite set of possible outcomes of an experiment. The outcome of the experiment, $z \in \mathcal{Z}$, is assumed related to the unknown, certain outcomes are more strongly linked to certain values of θ. The relationship between the unknown and the outcome of the experiment is modelled by a *transition*.

2.5.1 Transitions and their Algebra

Definition 2.17. *A* transition *from a set X to a set Y is a function $T : X \to \mathbb{P}(Y)$.*

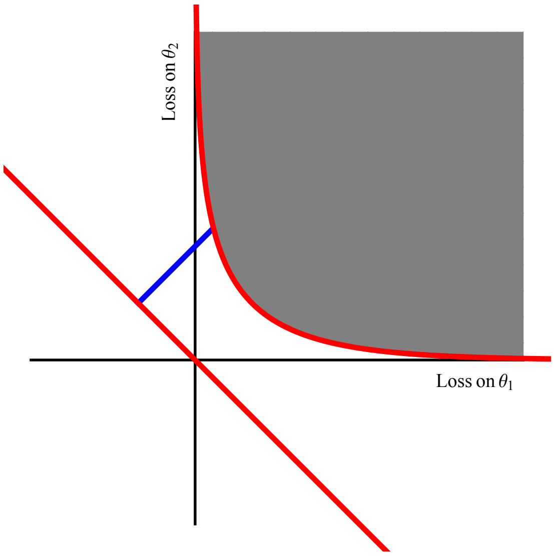


Fig. 2.2: Construction of canonical coordinates for log loss, see text.

Denote the set of all transitions from X to Y by $\mathbb{T}(X, Y)$. Transitions (or Markov kernels), constitute a modern approach to conditional probability [36; 39; 85; 117]. The distribution $T(x)$ is how the decision maker summarizes their uncertainty about Y if the true value of X is x. Every function $\phi \in Y^X$ defines a transition with,

$$\langle \phi(x), f \rangle_Y = f(\phi(x)), \ \forall f \in \mathbb{R}^Y.$$

Such a transition is called *deterministic*. Transitions can also be thought of as dual mappings, $T : (\mathbb{R}^X)^* \to (\mathbb{R}^Y)^*$. We define,

$$\langle T(\alpha), f \rangle_Y = \langle \alpha, \langle T(x), f \rangle_Y \rangle_X, \ \forall f \in \mathbb{R}^Y, \ \forall \alpha \in (\mathbb{R}^X)^*.$$

The function $T^*(f)(x) = \langle T(x), f \rangle_Y$ is the *pullback* of f by T. Formally, the operator T^* is the *adjoint* or *transpose* of T. Transitions can be *composed*. For transitions $T \in \mathbb{T}(X, Y)$ and $S \in \mathbb{T}(Y, Z)$ we can define $S \circ T \in \mathbb{T}(X, Z)$ with,

$$\langle S \circ T(x), f \rangle_Z = \langle T(x), \langle S(y), f \rangle_Z \rangle_Y, \ \forall f \in \mathbb{R}^Z.$$

In usual notation, this is just iterated expectation,

$$\langle S \circ T(x), f \rangle_Z = \mathbb{E}_{y \sim T(x)} \mathbb{E}_{z \sim S(y)} f(z).$$

Intuitively, this can be seen as "marginalizing" over Y in the Markov chain,

$$X \to Y \to Z.$$

If X and Y are finite sets, a transition $T \in \mathbb{T}(X, Y)$ can be represented by a column stochastic matrix, with composition given by matrix multiplication.

Transitions can also be combined in *parallel*. For $P, Q \in \mathbb{P}(X)$, denote the product distribution by $P \otimes Q$. If $T_i \in \mathbb{T}(X_i, Y_i)$, $i \in [1; k]$, are transitions then denote,

$$\otimes_{i=1}^{k} T_i \in \mathbb{T}(\times_{i=i}^{k} X_i, \times_{i=1}^{k} Y_i)$$

with $\otimes_{i=1}^{k} T_i(x) = T_1(x_1) \otimes \cdots \otimes T_k(x_k)$. Transitions can also be *replicated*. For any transition $T \in \mathbb{T}(X, Y)$ we denote the *replicated transition* $T_n \in \mathbb{T}(X, Y^n)$, $n \in \{1, 2, \dots\}$, with,

$$T_n(x) = \underbrace{T(x) \otimes \dots T(x)}_{n \text{ times}} = T(x)^n,$$

the n-fold product of $T(x)$. A distribution $P \in \mathbb{P}(X)$ and a transition $T \in \mathbb{T}(X, Y)$ can be combined into a *joint* distribution $P \ltimes T \in \mathbb{P}(X \times Y)$, with,

$$\langle P \ltimes T, f \rangle_{X \times Y} = \langle P, \langle T(x), f_x \rangle_Y \rangle_X = \mathbb{E}_{x \sim P} \mathbb{E}_{y \sim T(x)} f(x, y), \ \forall f \in \mathbb{R}^{X \times Y}.$$

Bayes theorem provides means to *disintegrate* [36; 111] a joint distribution $P_X \ltimes T_{X \to Y}$ into a joint distribution $P_Y \ltimes T_{Y \to X}$, where $P_X \in \mathbb{P}(X)$ and $T_{X \to Y} \in \mathbb{T}(X, Y)$, and $P_Y \in \mathbb{P}(Y)$ and $T_{Y \to X} \in \mathbb{T}(X, Y)$. Disintegration theorems hold in very general spaces, not just the cases considered here.

2.5.2 Comparing Experiments

An *experiment* is a transition $e \in \mathbb{T}(\Theta, \mathcal{Z})$. We call $\mathcal{Z}$ the observation space of the experiment. The distribution $e(\theta)$ summarizes the decision maker's uncertainty in the observation when θ is the value of the unknown. After observing the results of an experiment, the decision maker is tasked with choosing a suitable action. They do this via a *learning algorithm*.

A *learning algorithm* is a transition $\mathcal{A} \in \mathbb{T}(\mathcal{Z}, A)$. $\mathcal{A}(z)$ summarizes the decision makers uncertainty in which action to choose, given an observation $z \in \mathcal{Z}$. We define the *risk*,

$$\mathcal{R}_L(\theta, e, \mathcal{A}) = \mathbb{E}_{z \sim e(\theta)} \mathbb{E}_{a \sim \mathcal{A}(z)} L(\theta, a).$$

The risk measures the quality of the final action chosen by the decision maker when they use the learning algorithm $\mathcal{A}$, after performing experiment e, assuming θ is the true value of the unknown. The risk does not provide a single number for the comparison of experiments, rather it provides an entire *risk profile*. To compare risks directly, the decision maker can use the *Bayesian* or *max* risks defined as,

$$\mathcal{R}_L^{\pi}(e, \mathcal{A}) := \mathbb{E}_{\theta \sim \pi} \mathcal{R}_L(\theta, e, \mathcal{A}) \text{ and } \mathcal{R}_L(e, \mathcal{A}) := \sup_{\theta} \mathcal{R}_L(\theta, e, \mathcal{A}),$$

respectively. The Bayesian risk is more appropriate if the decision maker has some intuition about θ, given in the form of a prior probability distribution π. The max risk is more appropriate if the decision maker has no prior knowledge concerning θ.

These quantities allow the decision maker to *compare* the usefulness of experiment, learning algorithm pairs. To compare experiments directly, we assume the decision maker uses the *best* learning algorithm. Define the minimum Bayesian risk and minimax risk as,

$$\underline{\mathcal{R}}_L^\pi(e) := \min_{\mathcal{A}} \mathcal{R}_L^\pi(e, \mathcal{A}) \text{ and } \underline{\mathcal{R}}_L(e) := \min_{\mathcal{A}} \mathcal{R}_L(e, \mathcal{A}),$$

respectively. The minimum Bayes risk and the minimax risk are deeply related.

Theorem 2.18. *For all experiments e and loss functions L,*

$$\underline{\mathcal{R}}_L(e) = \sup_{\pi \in \mathbb{P}(\Theta)} \underline{\mathcal{R}}_L^\pi(e).$$

The proof is a simple application of the minimax theorem [82]. In light of this theorem, we focus on Bayesian risks for the remainder. All results have a minimax equivalent.

We also point out to the reader that all notions here have relative versions. For example the *relative risk* is defined as,

$$\Delta\mathcal{R}_L(\theta, e, \mathcal{A}) = \mathcal{R}_L(\theta, e, \mathcal{A}) - \inf_{a \in A} L(\theta, a),$$

which measures the risk relative to knowing θ.

Abstracting Away the Observation: Risk as Loss

Ultimately, what matters to the decision maker is not the exact details of the experiment and their learning algorithm. What matters is that the distribution $\mathcal{A} \circ e(\theta)$ places high weight on actions that are suitable for θ. We can think of risk as a loss,

$$\mathcal{R}_L : \Theta \times \mathbb{T}(\Theta, A) \to \mathbb{R},$$

with $\mathcal{R}_L(\theta, \mathcal{A}) = \mathbb{E}_{a \sim \mathcal{A}(\theta)} L(\theta, a)$. Different experiments allow the decision maker access to different subsets of $\mathbb{T}(\Theta, A)$.

Admissible and Bayesian Learning Algorithms

The optimal learning algorithm will in general depend on their prior knowledge about the unknown. Even without this knowledge, the decision maker can remove rules that are obviously not optimal.

Definition 2.19. *Let e be an experiment. A learning algorithm $\mathcal{A}$ is* admissible for *e if there does not exist a learning algorithm $\mathcal{A}'$ with,*

$$\mathcal{R}_L(\theta, e, \mathcal{A}') \leq \mathcal{R}_L(\theta, e, \mathcal{A}), \ \forall \theta \in \Theta$$

with strict inequality for at least one θ.

Intuitively, a learning algorithm is admissible if it is not obviously worse than some other learning algorithm. If the decision maker has prior π, they can minimize the Bayesian risk by using a Bayesian learning algorithm.

Definition 2.20. *Let e be an experiment and π a prior. A learning algorithm $\mathcal{A}^*$ is* Bayes for (π, e) *if,*

$$\mathcal{A}^* \in \operatorname*{arg\,min}_{\mathcal{A}} \mathcal{R}_L^{\pi}(e, \mathcal{A}).$$

Much like the case for Bayesian actions, the decision maker need only consider Bayesian learning algorithms.

Theorem 2.21 (Complete Class Theorem [126]). *A learning algorithm $\mathcal{A}$ is admissible for e if and only if there exists a prior π such that $\mathcal{A}$ is Bayes for (π, e).*

The above theorem says that Bayesian algorithms provide all rules that a sensible decision maker should use. Picking a particular admissible algorithm is *equivalent* to picking a prior π and minimizing the Bayesian risk against that prior. While statistically, admissible algorithms afford no obvious improvements, they may be hard to implement. Our language as is does not take this into account. The study of inadmissible algorithms and their risks is therefore a worthwhile endeavour.

Bayes optimal algorithms admit a simple representation. A prior π and an experiment e define a joint distribution on pairs $\Theta \times \mathcal{Z}$ in the obvious way. Let $\pi_{\mathcal{Z}}$ be the marginal distribution over the observation space, and $\eta \in \mathbb{T}(\mathcal{Z}, \Theta)$ be the induced conditional distribution of unknowns given observations. Then,

$$\underline{\mathcal{R}}_L^{\pi}(e) = \mathbb{E}_{z \sim \pi_{\mathcal{Z}}} \underline{L}(\eta(z)),$$

with Bayes optimal algorithm,

$$\mathcal{A}^*(z) := \operatorname*{arg\,min}_{a \in A} \mathbb{E}_{\theta \sim \eta(z)} L(\theta, a).$$

We stress that this algorithm is prior dependent, η depends on the prior π and the experiment T.

2.5.3 When is One Experiment Always Better than Another?

Let e and e' be experiments. Suppose that due to constraints, the decision maker can only perform one of these two experiments. The decision maker can compare the Bayes or minimax risks of the two experiments, however this involves a (perhaps difficult) calculation. Furthermore, if the loss function of interest changes, then the ordering of the experiments might change. We seek qualitative results concerning when e is *always* better than e' no matter what the loss or prior distribution.

Definition 2.22. *Let $e \in \mathbb{T}(\Theta, \mathcal{Z})$ and $e' \in \mathbb{T}(\Theta, \mathcal{Z}')$ be experiments. e* divides *e' (written $e \mid e'$) if there exists a transition $T \in \mathbb{T}(\mathcal{Z}, \mathcal{Z}')$ such that $e' = T \circ e$.*

Intuitively, $e \mid e'$ if e' is e with extra noise T. We make this intuition precise with theorem 2.24. For an experiment e, let $\mathbb{T}_e(\Theta, A)$ be the set of transitions from Θ to A that e divides.

Theorem 2.23. *e divides e' if and only if for all action sets A,*

$$\mathbb{T}_{e'}(\Theta, A) \subseteq \mathbb{T}_e(\Theta, A).$$

Proof. The forward implication follows simply from the definition. For the converse, take $A = \mathcal{Z}'$ and note $e' \in \mathbb{T}_{e'}(\Theta, \mathcal{Z}')$. By assumption,

$$\mathbb{T}_{e'}(\Theta, \mathcal{Z}') \subseteq \mathbb{T}_e(\Theta, \mathcal{Z}').$$

As $e' \in \mathbb{T}_{e'}(\Theta, \mathcal{Z}')$, this implies there exists a T with $T \circ e = e'$.

□

The Blackwell-Sherman-Stein Theorem and Sufficiency

Theorem 2.24 (Blackwell-Sherman-Stein Theorem [24])**.** *Let e and e' be experiments. $e \mid e'$ if and only if for all action sets, loss functions and priors,*

$$\underline{\mathcal{R}}_L^\pi(e) \leq \underline{\mathcal{R}}_L^\pi(e').$$

We prove the forward implication, called the data processing theorem. The proof of the converse will come later, as a simple corollary of the randomization theorem.

Proof. For any learning algorithm $\mathcal{A}' \in \mathbb{T}(\mathcal{Z}', A)$ consider the learning algorithm $\mathcal{A} = \mathcal{A}' \circ T \in \mathbb{T}(\mathcal{Z}, A)$. As $e' = T \circ e$, it is easy to verify that,

$$\mathcal{R}_L^\pi(e', \mathcal{A}') = \mathcal{R}_L^\pi(T \circ e, \mathcal{A}') = \mathcal{R}_L^\pi(e, \mathcal{A}' \circ T) = \mathcal{R}_L^\pi(e, \mathcal{A}).$$

To complete the proof take minima over $\mathcal{A}$ and $\mathcal{A}'$.

□

We say e and e' are *equivalent* experiments (written $e \cong e'$) if both $e \mid e'$ and $e' \mid e$. Equivalent experiments have equivalent risks. A key notion in statistics is that of *sufficiency*. A *sufficient statistic* is a function of the observation that loses none of the information contained in e. The Blackwell-Sherman-Stein theorem provides means to define and understand sufficiency. Identifying and exploiting sufficient statistics allows the decision maker to compress the information contained in the observation, without losing information.

Definition 2.25. *Let $e \in \mathbb{T}(\Theta, \mathcal{Z})$ be an experiment. A transition $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is* sufficient *for e if $T \circ e \cong e$.*

By the Blackwell-Sherman-Stein theorem, sufficient statistics maintain all information in the observation, under the assumption the decision maker uses the best learning algorithm for each experiment.

For any set Θ of unknowns there is a *most* informative and a *least* informative experiment. Recall the identity function id_Θ, $\mathrm{id}_\Theta(\theta) = \theta$. For any experiment e, we have $e \circ \mathrm{id}_\Theta = e$. Therefore, id_Θ divides any experiment. Intuitively, id_Θ provides the decision maker the *exact* value of θ. This experiment has risk,

$$\underline{\mathcal{R}}_L^\pi(\mathrm{id}_\Theta) = \mathbb{E}_{\theta\sim\pi} \min_{a\in A} L(\theta, a).$$

For any set X, define the *terminal* transition $\bullet_X \in \{1\}^X$ with $\bullet_X(x) = 1$ for all X. Intuitively this transition throws away all information about X. Much like the identity transition divides every experiment, the terminal transition is divided by every experiment. For all experiments e, $\bullet_\Theta = \bullet_{\mathcal{Z}} \circ e$. This experiment has risk,

$$\underline{\mathcal{R}}_L^\pi(\bullet_\Theta) = \underline{L}(\pi).$$

By the data processing theorem,

$$\underline{\mathcal{R}}_L^\pi(\mathrm{id}_\Theta) \leq \underline{\mathcal{R}}_L^\pi(e) \leq \underline{\mathcal{R}}_L^\pi(\bullet_\Theta).$$

Relationship to the Standard Data Processing Theorem

Definition 2.26. *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $f(1) = 0$. For all distributions $P, Q \in \mathbb{P}(\mathcal{Z})$ the* f-divergence *between P and Q is,*

$$D_f(P, Q) = \langle P, f\left(\tfrac{dQ}{dP}\right)\rangle,$$

if Q is absolutely continuous with respect to P and is undefined otherwise.

f-divergences provide one means of measuring the *dissimilarity* of two probability distributions. They include many standard measures of dissimilarity, including the KL-divergence, the Hellinger divergence and variational distance [2; 48]. The standard data processing theorem states that for all sets $\mathcal{Z}, \tilde{\mathcal{Z}}$, all transitions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$, all distributions $P, Q \in \mathbb{P}(\mathcal{Z})$ and all f-divergences,

$$D_f(T(P), T(Q)) \leq D_f(P, Q).$$

Intuitively, adding noise always makes it harder to distinguish P and Q. This theorem is actually theorem 2.24 in disguise. The pair $P, Q \in \mathbb{P}(\mathcal{Z})$ *define* a transition $e \in \mathbb{T}(\{-1, 1\}, \mathcal{Z})$, with $e(1) = P$ and $e(-1) = Q$.

Theorem 2.27 (Minimum Bayes Risk is an f-divergence [105]). *For all f-divergences, there exists a loss function $L : \{-1, 1\} \times A \to \mathbb{R}$ and a prior π such that for all experiments*

$e \in \mathbb{T}(\{-1,1\},\mathcal{Z})$,

$$\underline{\mathcal{R}}_L^\pi(\bullet_\Theta) - \underline{\mathcal{R}}_L^\pi(e) = \underline{L}(\pi) - \underline{\mathcal{R}}_L^\pi(e) = D_f(e(-1),e(1)).$$

The data processing theorem for f-divergences follows directly from our data processing theorem. In [62] this correspondence is developed further to create multi-f-divergences. These results highlight the foundational role played by the Bayesian risk.

Digression: Generalised Data Processing Theorems

Key to the proof of the data processing theorem is the insight that if $e \mid e'$ then,

$$\mathbb{T}_{e'}(\Theta,A) \subseteq \mathbb{T}_e(\Theta,A).$$

Intuitively, this means more learning algorithms are available to the decision maker after performing the experiment e than e'. This suggests another means to construct quantities that satisfy data processing theorems.

Theorem 2.28. *Let $\psi : \mathbb{T}(\Theta,A) \to \mathbb{R}$ and define the information measure,*

$$I_\psi(e) = \min_{\mathcal{A}\in\mathbb{T}_e(\Theta,A)} \psi(\mathcal{A}).$$

If $e \mid e'$ then $I_\psi(e) \leq I_\psi(e')$.

Proof. If $e \mid e'$ then $\mathbb{T}_{e'}(\Theta,A) \subseteq \mathbb{T}_e(\Theta,A)$. Therefore,

$$\min_{\mathcal{A}\in\mathbb{T}_e(\Theta,A)} \psi(\mathcal{A}) \leq \min_{\mathcal{A}'\in\mathbb{T}_{e'}(\Theta,A)} \psi(\mathcal{A}').$$

□

We recover the usual data processing theorem by taking,

$$\psi(\mathcal{A}) = \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{a\sim\mathcal{A}(\theta)}L(\theta,\mathcal{A}).$$

Remarkably, the proof of theorem 2.28 makes no reference to expected risks, transitions, or even probability distributions! Much recent work has been directed toward characterizing functions that satisfy a data processing theorem [10; 49; 71]. Invariably, these approaches "cook the books", adding extra constraints until KL-divergence and mutual information are discovered to be the only such functions, and the standard data processing theorem of information theory is recovered [46].

Many other uncertainty calculus exist, based on more exotic means of making decisions than relatively simple probability theory (see [70] for some). For example in robust statistics, the linear functions defining probabilities are replaced with convex functions, probabilities replaced with upper and lower probabilities (see chapter 10 of [75]). In the emerging fields of non-commutative probability [42; 124], a field with close ties to quantum theory, the commutative algebra of functions is replaced with general non commuting algebras. All these different systems could potentially be used in place of probability. The generality of theorem 2.28 would seem to indicate that such a theorem will excess in these other systems.

Deficiency and Quantitative Data Processing Theorems

The converse of the Blackwell-Sherman-Stein theorem states that if e does *not* divide e' then there is a loss function and prior that renders e' more useful. Furthermore, if e' does not divide e then there is a loss function and prior that renders e more useful. The gap in risks is quantified by the *deficiency*.

Definition 2.29. *Let $P, Q \in \mathbb{P}(\mathcal{Z})$ be distributions. The* variational distance *between P and Q is,*

$$V(P,Q) := \sup_{f \in [0,1]^{\mathcal{Z}}} |\mathbb{E}_P(f) - \mathbb{E}_Q(f)| .$$

Intuitively, variational distance is the maximum difference in assigned loss when making decisions via P or Q. The variational distance is a metric on probability distributions. It is an f-divergence with $f(x) = |x-1|$ [2; 48]. This means the variational divergence satisfies a data processing inequality, for all transitions $T \in \mathbb{T}(\mathcal{Z}, \mathcal{Z}')$,

$$V(T(P), T(Q)) \leq V(P,Q).$$

Definition 2.30. *Let $e \in \mathbb{T}(\Theta, \mathcal{Z})$ and $e' \in \mathbb{T}(\Theta, \mathcal{Z}')$ be experiments. The* directed deficiency *from e to e' is,*

$$\zeta^\pi(e,e') := \min_{T \in \mathbb{T}(\mathcal{Z},\mathcal{Z}')} \mathbb{E}_{\theta \sim \pi} V(T \circ e(\theta), e'(\theta)).$$

The directed deficiency provides means to quantify how close e is to dividing e'. $\zeta^\pi(e,e') = 0$ for all priors if and only if $e \mid e'$. The *deficiency* is defined as,

$$\Xi^\pi(e,e') := \max\{\zeta^\pi(e,e'), \zeta^\pi(e',e)\}.$$

Deficiency measures how close to equivalence e and e' are. $\Xi^\pi(e,e') = 0$ for all priors if and only if $e \cong e'$. The directed deficiency provides a *quantitative* version of the Blackwell-Sherman-Stein theorem.

Theorem 2.31 (Randomization Theorem [85])**.** *Fix $\epsilon > 0$ and a prior π. Let e and e' be experiments. Then,*

$$\underline{\mathcal{R}}_L^\pi(e) \leq \underline{\mathcal{R}}_L^\pi(e') + \epsilon \, \|L\|_\infty$$

for all action sets and loss functions, if and only if $\zeta^\pi(e,e') \leq \epsilon$.

We present the proof appearing in [117], with some streamlining.

Proof. We begin with the reverse implication. As $\xi^{\pi}(e, e') \leq \epsilon$, there exists a transition $T \in \mathbb{T}(\mathcal{Z}, \mathcal{Z}')$ such that,

$$\mathbb{E}_{\theta \sim \pi} V(T \circ e(\theta), e'(\theta)) \leq \epsilon.$$

Now fix a learning algorithm $\mathcal{A}' \in \mathbb{T}(\mathcal{Z}', A)$, and consider $\mathcal{A} = \mathcal{A}' \circ T$ as in the diagram below.

$$\begin{tikzcd}
& \Theta \arrow[dl, rightsquigarrow, "e"'] \arrow[dr, rightsquigarrow, "e'"] & & \\
\mathcal{Z} \arrow[rr, rightsquigarrow, "T"] & & \mathcal{Z}' \arrow[r, rightsquigarrow, "\mathcal{A}'"] & A
\end{tikzcd}$$

We have,

$$\begin{aligned}
\mathcal{R}^{\pi}_{L}(e, \mathcal{A}) - \mathcal{R}^{\pi}_{L}(e', \mathcal{A}') &= \mathbb{E}_{\theta \sim \pi} \left[\mathbb{E}_{a \sim \mathcal{A} \circ e(\theta)} L(\theta, a) - \mathbb{E}_{a \sim \mathcal{A}' \circ e'(\theta)} L(\theta, a) \right] \\
&\leq \mathbb{E}_{\theta \sim \pi} V(\mathcal{A} \circ e(\theta), \mathcal{A}' \circ e'(\theta)) \left\| L \right\|_{\infty} \\
&= \mathbb{E}_{\theta \sim \pi} V(\mathcal{A}' \circ T \circ e(\theta), \mathcal{A}' \circ e'(\theta)) \left\| L \right\|_{\infty} \\
&\leq \mathbb{E}_{\theta \sim \pi} V(T \circ e(\theta), e'(\theta)) \left\| L \right\|_{\infty} \\
&\leq \epsilon \left\| L \right\|_{\infty}
\end{aligned}$$

where the first line follows from the definition of the Bayesian risk, the second follows from the definition of the variational distance, the third from the definition of $\mathcal{A}$, the fourth as variational distance is an f-divergence and therefore satisfies a data processing inequality and finally from our assumptions on T. The proof is completed by taking a minimum over $\mathcal{A}'$ and $\mathcal{A}$.

For the forward implication, first fix a set of actions A and a learning algorithm $\mathcal{A}' \in \mathbb{T}(\mathcal{Z}', A)$ and define the function,

$$\phi(L, \mathcal{A}) = \mathcal{R}^{\pi}_{L}(e, \mathcal{A}) - \mathcal{R}^{\pi}_{L}(e', \mathcal{A}') - \epsilon \left\| L \right\|_{\infty}.$$

Note that ϕ is affine in $\mathcal{A}$ and concave in L. By the conditions in the theorem,

$$\sup_{L} \min_{\mathcal{A}} \phi(L, \mathcal{A}) \leq 0.$$

By the minimax theorem [82] or strong convex duality [90], there exists a saddle point $(L^*, \mathcal{A}^*)$ with,

$$\phi(L^*, \mathcal{A}^*) = \min_{\mathcal{A}} \sup_{L} \phi(L, \mathcal{A}) = \sup_{L} \min_{\mathcal{A}} \phi(L, \mathcal{A}) \leq 0.$$

This implies,

$$\mathcal{R}^{\pi}_{L}(e, \mathcal{A}^*) \leq \mathcal{R}^{\pi}_{L}(e', \mathcal{A}') + \epsilon \left\| L \right\|_{\infty}, \ \forall L.$$

This means $\mathbb{E}_{\theta \sim \pi} V(\mathcal{A}^* \circ e(\theta), \mathcal{A}' \circ e'(\theta)) \leq \epsilon$, from the definition of variational distance. Note that $\mathcal{A}'$ and the action set A are arbitrary. To complete the proof, take

$A = \mathcal{Z}'$ and $\mathcal{A}' = \mathrm{id}_{\mathcal{Z}'}$. The transition T is then given by $\mathcal{A}^*$.

□

The proof of the reverse implication of the Blackwell-Sherman-Stein theorem can be recovered by setting $\epsilon = 0$. The randomization theorem shows there is a deep connection between differences of risks and deficiency. The following theorem makes this connection precise.

Theorem 2.32. *Let e and e' be experiments. For all priors π,*

$$\Xi^\pi(e, e') = \sup_{L : \|L\|_\infty \neq 0} \frac{|\underline{\mathcal{R}}_L^\pi(e) - \underline{\mathcal{R}}_L^\pi(e')|}{\|L\|_\infty}.$$

For the proof we require the following simple lemma.

Lemma 2.33. *For $x, y \in \mathbb{R}$ if $\forall \epsilon \in \mathbb{R}$, $x \leq \epsilon \Leftrightarrow y \leq \epsilon$ then $x = y$.*

Proof. Suppose that $x \neq y$ and without loss of generality assume that $x < y$. Set $\epsilon = \frac{x+y}{2}$. Then $x \leq \epsilon$ and $y > \epsilon$, which implies the contrapositive. □

We now prove the theorem.

Proof. If $\Xi^\pi(e, e') \leq \epsilon$ then $\xi^\pi(e, e') \leq \epsilon$ and $\xi^\pi(e', e) \leq \epsilon$. By the randomization theorem,

$$\frac{|\underline{\mathcal{R}}_L^\pi(e) - \underline{\mathcal{R}}_L^\pi(e')|}{\|L\|_\infty} \leq \epsilon, \ \forall L : \|L\|_\infty \neq 0.$$

Conversely, if,

$$\sup_{L : \|L\|_\infty \neq 0} \frac{|\underline{\mathcal{R}}_L^\pi(e) - \underline{\mathcal{R}}_L^\pi(e')|}{\|L\|_\infty} \leq \epsilon,$$

then $\underline{\mathcal{R}}_L^\pi(e) \leq \underline{\mathcal{R}}_L^\pi(e') + \epsilon \|L\|_\infty$ and $\underline{\mathcal{R}}_L^\pi(e') \leq \underline{\mathcal{R}}_L^\pi(e) + \epsilon \|L\|_\infty$. By the randomization theorem, this means $\Xi^\pi(e, e') \leq \epsilon$. This combined with the above lemma completes the proof.

□

The randomization theorem can be used to define *quantitative* versions of concepts such as sufficiency. This was Le Cam's original motivation for defining the quantity in [85]. T is approximately sufficient for e if $\Xi^\pi(e, T \circ e)$ is small. Deficiency provides a metric on experiments.

Theorem 2.34. *Let π be a prior that assigns non zero probability to each unknown. Then Ξ^π is a metric on experiments modulo equivalence.*

Proof. Ξ^π is obviously non-negative and symmetric. We are required to show that it satisfies the triangle inequality. Let e, e', e'' be experiments, with T and T' transitions as in the diagram below.

$$\begin{tikzcd}
 & \Theta \arrow[dl, squiggly, "e"'] \arrow[d, squiggly, "e'"] \arrow[dr, squiggly, "e''"] & \\
\mathcal{Z} \arrow[r, squiggly, "T"] & \mathcal{Z}' \arrow[r, squiggly, "T'"] & \mathcal{Z}''
\end{tikzcd}$$

We have for all θ,

$$\begin{aligned} V(e''(\theta), T' \circ T \circ e(\theta)) &\leq V(e''(\theta), T' \circ e'(\theta)) + V(T' \circ e'(\theta), T' \circ T \circ e(\theta)) \\ &\leq V(e''(\theta), T' \circ e'(\theta)) + V(e'(\theta), T \circ e(\theta)), \end{aligned}$$

where we have used the fact that the variational distance is a metric, followed by the data processing inequality. Averaging over π and taking a minimum over T and T' yields,

$$\xi^\pi(e, e'') \leq \xi^\pi(e, e') + \xi^\pi(e', e'').$$

Reversing the direction and taking maximums yields the desired result. □

Calculating Deficiency

The variational distance can be calculated as the l_1 distance,

$$V(P, Q) = \frac{1}{2} \sum_{z \in \mathcal{Z}} |P(z) - Q(z)| \,.$$

Experiments $e \in \mathbb{T}(\Theta, \mathcal{Z})$ can be represented by a $|\mathcal{Z}| \times |\Theta|$ column stochastic matrix, $T_{ij} \geq 0$ and $\sum_i T_{ij} = 1$ for all j. Furthermore, the prior distribution π can be represented by a vector in $\mathbb{R}^{|\Theta|}$. Using these representations, the directed deficiency can be calculated via linear programming.

Lemma 2.35. *Let e and e' experiments with their stochastic matrix representation given by E and E' respectively. Then $\xi^\pi(e, e')$ can be calculated via the following linear program,*

$$\begin{aligned} &\min_{M_{ij}, T_{ij}} \sum_{i=1}^{|\mathcal{Z}'|} \sum_{j=1}^{|\Theta|} M_{ij} \\ &\textit{subject to} \\ &M_{ij}, T_{ij} \geq 0 \textit{ and } \left| \pi_j E'_{ij} - \pi_j \left[TE\right]_{ij} \right| \leq M_{ij} \ \forall i, j \\ &\sum_{i=1}^{|\mathcal{Z}'|} T_{ij} = 1 \ \forall j, \end{aligned}$$

where $[TE]_{ij}$ is the ij entry of TE.

Proof. The constraints $T_{ij} \geq 0$ and $\sum_{i=1}^{|\mathcal{Z}'|} T_{ij} = 1 \ \forall j$, ensure that T is a stochastic matrix.

Taking the final constraint and summing over i and j yields,

$$\begin{aligned}\sum_{i=1}^{|\mathcal{Z}'|}\sum_{j=1}^{|\Theta|} M_{ij} &\geq \sum_{i=1}^{|\mathcal{Z}'|}\sum_{j=1}^{|\Theta|} \left|\pi_j E'_{ij} - \pi_j \left[TE\right]_{ij}\right| \\ &= \sum_{j=1}^{|\Theta|} \pi_j \sum_{i=1}^{|\mathcal{Z}'|} \left|E'_{ij} - \left[TE\right]_{ij}\right| \\ &= \mathbb{E}_{\theta\sim\pi} V(e'(\theta), T \circ e(\theta)).\end{aligned}$$

Equality is attained in the above if $M_{ij} = \left|\pi_j E'_{ij} - \pi_j \left[TE\right]_{ij}\right|$. Minimizing over M and T yields an optimal solution of $\xi^\pi(e, e')$.

□

2.6 Preview of the Remainder of the Thesis

While the ideas of the previous subsections originated in theoretical statistics [24; 60; 86; 117] they can be readily applied to machine learning problems. The main distinction is that statistics focuses on *parametric families* and loss functions of type $L : \Theta \times \Theta \to \mathbb{R}$. The goal is to accurately *reconstruct parameters*. In machine learning one is interested in *predicting the observations* of the experiment well. The remainder of this thesis turns to the *quantification* of the usefulness of different experiments for different prediction problems.

2.6.1 Prediction Problems

A central problem in machine learning is that of *prediction*. Given side information $x \in X$, the goal of the decision maker is to correctly *predict* a label $y \in Y$. To do so, the decision maker specifies a function $f \in \hat{Y}^X$, that should have low expected loss,

$$\mathbb{E}_{(x,y)\sim P} \ell(y, f(x)),$$

where $\ell : Y \times \hat{Y} \to \mathbb{R}$ is a loss function that measures the suitability of a prediction $\hat{y}$. Note Y and $\hat{Y}$ need not be the same set, for example in conditional probability estimation $\hat{Y} = \mathbb{P}(Y)$. If P is known to the decision maker, then this is a problem of optimization. In general P is unknown, but the decision maker has access to a sample S of n iid draws from P, $\{(x_i, y_i)\}_{i=1}^n$. The sample is used as a proxy for expectation under Y, and the decision maker returns the function in a restricted class $\mathcal{F} \subseteq \hat{Y}^X$,

$$f_S = \arg\min_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)).$$

This is known as the empirical risk minimization (ERM) rule [122]. Of central interest are bounds on the *expected loss* of the ERM rule,

$$\mathbb{E}_{S\sim P^n}\mathbb{E}_{(x,y)\sim P}\ell(y, f_S(x))$$

that hold *regardless* of the true value of P. What is *unknown* to the decision maker is the distribution P. The decision maker *acts* by specifying a function $f \in \hat{Y}^X$. The loss incurred to the decision maker is the expected predictive performance of f. The assumption that the sample is comprised of iid draws from P can be seen as an *experiment* $e \in \mathbb{T}(\mathbb{P}(X \times Y), (X \times Y)^n)$, that maps each distribution to its n-fold product. The ERM rule can be understood as a *particular* learning algorithm. The *risk* for any learning algorithm $\mathcal{A}$ is,

$$\mathcal{R}_L(P, e, \mathcal{A}) = \mathbb{E}_{S\sim P^n}\mathbb{E}_{f\sim\mathcal{A}(S)}\mathbb{E}_{(x,y)\sim P}\ell(y, f(x)).$$

Requiring that the risk is small for all P is then *exactly* a requirement that the minimax risk of $\mathcal{A}$ is small.

For the remainder of the thesis we focus on prediction problems for *general* experiments, where the decision maker is given access to data that is not of the form of iid draws from P.

3

Learning in the Presence of Corruption

> *In the spirit of science, there really is no such thing as a "failed experiment". Any test that yields valid data is a valid test.*
>
> \- Adam Savage, *Mythbusters*

The goal of supervised learning is to find a function in some hypothesis class that accurately predicts the relationship between instances and labels. Such a function should have low expected loss according to the true distribution of instances and labels, P. The decision maker is not given direct access to P, but rather a training set comprising n iid samples from P. There are many algorithms for solving this problem (for example empirical risk minimization) and this problem is well understood.

There are many other types of data one could learn from. For example in semi-supervised learning [37] the decision maker is given n instance label pairs and m instances devoid of labels. In learning with noisy labels [4; 81; 101], the decision maker observes instance label pairs where the observed labels have been corrupted by some noise process. There are many other variants including, but not limited to, learning with label proportions [103], learning with partial labels [45], multiple instance learning [95] as well as combinations of the above.

What is currently lacking is a general theory of learning from corrupted data, as well as means to *compare* the relative usefulness of different data types. Such a theory is required if one wishes to make informed economic decisions on which data sets to acquire. For example, are n clean data better or worse than n_1 noisy labels and n_2 partial labels?

To answer this question we first place the problem of corrupted learning into the abstract language of statistical decision theory. We then develop general lower and upper bounds on the risk relative to the amount of corruption of the clean data. Finally we show examples of problems that fit into this abstract framework.

The main contributions of this chapter are:

- Novel, general means to construct methods for learning from corrupted data based on a generalization of the method of unbiased estimators presented by Natarajan et al. in [101] and implicit in the earlier work of Kearns [81] (theorems 3.2 and 3.3).
- Novel lower bounds on the risk of corrupted learning (theorem 3.14).
- Means to understand *compositions* of corruptions (lemmas 3.12 and 3.18).
- Upper and lower bounds on the risk of learning from combinations of corrupted data (theorems 3.4 and 3.15).
- Analyses of the tightness of the above bounds.

In doing so we provide means to rank different types of corrupted data, through the utilization of our upper and lower bounds. These results greatly extend the state of the art in Crammer et al. [47], both in scope and in completeness. Their results only apply to the learning of binary classifiers with label noise, and they only provide upper bounds.

While not the complete story for *all* problems, the contributions outlined above make progress toward the final goal of informed economic decisions regarding the acquisition of data sets of differing quality.

Proofs are mostly relegated to the appendix of this chapter.

3.1 The Standard Prediction Problem

We consider a general prediction problem. Let $\ell : \mathcal{Z} \times A \to \mathbb{R}$ be a loss. We assume that $\mathcal{Z}$ is finite. Ultimately we are interested in supervised learning problems with finite label space Y and corruptions only on the labels. All of the techniques developed for finite $\mathcal{Z}$ can be transferred to this setting. For the simplicity of presentation, we assume A is finite. Our bounds for finite A can be extended to infinite A via PAC-Bayesian bounds or covering number arguments. We state and prove the more general results in the appendix to this chapter.

For a distribution $P \in \mathbb{P}(\mathcal{Z})$, the goal of the decision maker is to minimize the *regret*,

$$\Delta\ell(P, Q) = \mathbb{E}_{z\sim P}\mathbb{E}_{a\sim Q}\ell(z, a) - \inf_{a\in A} \mathbb{E}_{z\sim P}\ell(z, a).$$

To do so they are given access to n iid draws from P. This can be realised as the experiment $e_n \in \mathbb{T}(\mathbb{P}(\mathcal{Z}), \mathcal{Z}^n)$, with $e_n(P) = P^n$, the n-fold product of P. This is an example of a replicated experiment. They seek a learning algorithm $\mathcal{A} \in \mathbb{T}(\mathcal{Z}^n, A)$ with low *relative* risk,

$$\Delta\mathcal{R}_\ell(P, e_n, \mathcal{A}) = \mathbb{E}_{S\sim P^n}\Delta\ell(P, \mathcal{A}(S)).$$

Supervised learning can be seen as a specific instance of this more general problem.

3.1.1 Corrupted Prediction Problems

Due to limitations in the measurement apparatus available to the decision maker, rather than observing $z \in \mathcal{Z}$, it is often the case that the decision maker observes a corrupted $\tilde{z}$ in a potentially different observation space $\tilde{\mathcal{Z}}$. We model the corruption process via a transition $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$. For example, we may wish to learn a relationship between measured symptoms and a medical diagnosis, as provided to us by an expert. To do so, rather than being given the experts data, we are given access to data from one of their apprentices. Here T models the hypothesized link between the experts and apprentices data. The goal of predicting as well as the expert remains.

For convenience we define the corrupted experiment $\tilde{e} = T \circ e$. We order the utility of different corruptions via the *relative minimax risk*,

$$\Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n) = \min_{\mathcal{A}}\max_{P}\Delta\mathcal{R}_\ell(P, \tilde{e}_n, \mathcal{A}).$$

All of the results that follow still hold for the minimum Bayesian risk. Ideally we wish to compare $\Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n)$ with $\Delta\underline{\mathcal{R}}_\ell(e_n)$, the minimum risk of the corrupted and the clean experiments. By the general data processing (theorem 2.28), $\Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n) \geq \Delta\underline{\mathcal{R}}_\ell(e_n)$, however this does not allow one to *rank* the utility of *different* T.

Even after many years of directed research, in general we can not compute $\Delta\underline{\mathcal{R}}_\ell(e_n)$ exactly, let alone $\Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n)$ for general corruptions. Consequently our effort for the remaining turns to upper and lower bounds of $\Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n)$.

3.2 Corruption Corrected Losses

When convenient we use the shorthand $T(P) = \tilde{P}$. Natarajan et al. [101] introduced a method of learning classifiers from data subjected to label noise, called the "method of unbiased estimators". Here we show that this method can be generalized to other corruptions. Recall from section 2.5.1, a transition $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ provides a linear map $T : (\mathbb{R}^{\mathcal{Z}})^* \to (\mathbb{R}^{\tilde{\mathcal{Z}}})^*$ with,

$$\langle T(\alpha), \tilde{f}\rangle_{\tilde{\mathcal{Z}}} = \langle \alpha, T^*\tilde{f}\rangle_{\mathcal{Z}},\ \forall\tilde{f} \in \mathbb{R}^{\tilde{\mathcal{Z}}},\ \forall\alpha \in (\mathbb{R}^{\mathcal{Z}})^*,$$

where $T^*(\tilde{f}) = \langle T(z), \tilde{f}\rangle_{\tilde{\mathcal{Z}}}$. In words, T allows functions of the corrupted sample to be "pulled back" to functions of the clean sample. We wish to go in the other direction; to *transfer* functions of clean samples to those of corrupted samples.

Definition 3.1. *A transition $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is* reconstructible *if T has a left inverse; that is there exists a linear map $R : (\mathbb{R}^{\tilde{\mathcal{Z}}})^* \to (\mathbb{R}^{\mathcal{Z}})^*$ such that $R \circ T = \mathrm{id}_{(\mathbb{R}^{\mathcal{Z}})^*}$.*

Intuitively, T is reconstructible if there is some transformation that "undoes" the effects of T. In general R is not a transition, the inverse of a stochastic matrix need not be stochastic (see section 3.2.1). Many forms of corrupted learning are reconstructible, including semi-supervised learning, learning with label noise and learning with partial labels for all but a few pathological cases. Section 3.9 contains several worked examples.

We call a left inverse of T a *reconstruction*. For concreteness, we can always take,

$$R = (T^*T)^{-1}T^*,$$

the Moore-Penrose pseudo inverse of T. In general it will be useful to consider other reconstructions. Reconstructible transitions are exactly those where we can *transfer* a function of the clean z to one of the corrupted $\tilde{z}$ while preserving expectations. By properties of adjoints,

$$\langle P, f \rangle = \langle R \circ T(P), f \rangle = \langle T(P), R^*(f) \rangle.$$

In words, to take expectations of f with samples from $\tilde{P}$ we use the corruption corrected $\tilde{f} = R^*(f)$. Recall the partial loss function $\ell(-, a) \in \mathbb{R}^{\mathcal{Z}}$. Using R we can reconstruct the partial loss from corrupted examples.

Theorem 3.2 (Corruption Corrected Loss). *For all reconstructible $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and loss functions $\ell : \mathcal{Z} \times A \to \mathbb{R}$ define the* corruption corrected *loss $\ell_T : \tilde{\mathcal{Z}} \times A \to \mathbb{R}$, with*

$$\ell_T(-, a) = R^*(\ell(-, a)),\ \forall a \in A.$$

Then for all distributions $P \in \mathbb{P}(\mathcal{Z})$, $\ell(P, a) = \ell_T(\tilde{P}, a)$.

3.2.1 A Worked Example: Learning with Symmetric Label Noise

When learning under symmetric label noise, the decision maker is required to predict a binary label $y \in \{-1, 1\}$, where $y \sim P$. Rather than observing the true y, the decision maker observes $\tilde{y}$, where $\tilde{y} = y$ with probability $1 - \sigma$ and $\tilde{y} = -y$ with probability σ. This process can be modelled by the following transition and reconstruction respectively;

$$T = \begin{pmatrix} 1-\sigma & \sigma \\ \sigma & 1-\sigma \end{pmatrix},\ R^* = \frac{1}{1-2\sigma} \begin{pmatrix} 1-\sigma & -\sigma \\ -\sigma & 1-\sigma \end{pmatrix}.$$

Note that R is *not* a transition as some of the entries of R are negative. For a loss $\ell : \{-1, 1\} \times A \to \mathbb{R}$, we have,

$$\begin{pmatrix} \ell_T(-1, a) \\ \ell_T(1, a) \end{pmatrix} = \frac{1}{1-2\sigma} \begin{pmatrix} 1-\sigma & -\sigma \\ -\sigma & 1-\sigma \end{pmatrix} \begin{pmatrix} \ell(-1, a) \\ \ell(1, a) \end{pmatrix},$$

or more compactly,

$$\ell_T(y, a) = \frac{(1-\sigma)\ell(y, a) - \sigma\ell(-y, a)}{1-2\sigma}.$$

This is equivalent to the original "method of unbiased estimators presented" in [101]. In section 3.9 several examples of corruption corrected losses are given.

3.2.2 Uses of Corruption Corrected Losses in Supervised Learning

In supervised learning $\mathcal{Z} = X \times Y$ and the goal is to find a function that predicts Y from X with low expected loss. Given a suitable function class $\mathcal{F} \subseteq \hat{Y}^X$ and a loss $\ell : Y \times \hat{Y} \to \mathbb{R}$, one attempts to find,

$$f^* = \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim P} \ell(y, f(x)).$$

If we assume the labels have been corrupted by a corruption $T \in \mathbb{T}(Y, \tilde{Y})$, we can correct for the corruption and solve for,

$$\operatorname*{arg\,min}_{f \in \mathcal{F}} \mathbb{E}_{(x,\tilde{y})\sim \tilde{P}} \ell_T(\tilde{y}, f(x)).$$

3.3 Upper Bounds for Corrupted Learning

Minimizing ℓ_T on a sample $\tilde{S}^n \sim \tilde{P}$ provides a means to learn from corrupted data. Let $\ell(S, a) = \frac{1}{|S|} \sum_{z \in S} \ell(z, a)$, the average loss on the sample.

By an application of the PAC-Bayes bound (see theorem A.11 of the appendix to the thesis) one has for all algorithms $\mathcal{A} \in \mathbb{T}(\tilde{\mathcal{Z}}^n, A)$ and distributions $P \in \mathbb{P}(\mathcal{Z})$,

$$\mathbb{E}_{\tilde{S}\sim\tilde{P}^n} \ell_T(\tilde{P}, \mathcal{A}(\tilde{S})) \leq \mathbb{E}_{\tilde{S}\sim\tilde{P}^n} \ell_T(\tilde{S}, \mathcal{A}(\tilde{S})) + \|\ell_T\|_\infty \sqrt{\frac{2\log(|A|)}{n}}.$$

Since, by the construction of ℓ_T, $\ell_T(\tilde{P}, \mathcal{A}(\tilde{S})) = \ell(P, \mathcal{A}(\tilde{S}))$, the above bound yields the following theorem.

Theorem 3.3. *For all reconstructible $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$, algorithms $\mathcal{A} \in \mathbb{T}(\tilde{\mathcal{Z}}^n, A)$, distributions $P \in \mathbb{P}(\mathcal{Z})$ and bounded loss functions ℓ,*

$$\mathbb{E}_{\tilde{S}\sim\tilde{P}^n} \ell(P, \mathcal{A}(\tilde{S})) \leq \mathbb{E}_{\tilde{S}\sim\tilde{P}^n} \ell_T(\tilde{S}, \mathcal{A}(\tilde{S})) + \|\ell_T\|_\infty \sqrt{\frac{2\log(|A|)}{n}}.$$

A similar result also holds with high probability on draws from $\tilde{P}^n$. This bound shows that ERM converges to the optimum $a \in A$ as $\frac{\|\ell_T\|_\infty}{\sqrt{n}}$ for learning with corrupted data versus $\frac{\|\ell\|_\infty}{\sqrt{n}}$ for learning with clean data. Therefore, the ratio $\frac{\|\ell_T\|_\infty}{\|\ell\|_\infty}$ measures the relative difficulty of corrupted versus clean learning, as judged solely by the upper bound.

3.3.1 Upper Bounds for Combinations of Corrupted Data

Recall that our final goal is to quantify the utility of a data set comprising different corrupted data. For example in learning with noisy labels out of n data, there could be n_1 clean, n_2 slightly noisy and n_3 very noisy samples and so on. More generally we assume access to a corrupted sample $\tilde{S}$, made up of k different types of corrupted data, with $\tilde{S}_i{\sim}\tilde{P}^{n_i}$, $i \in [1;k]$.

Theorem 3.4. *Let $T_i \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_i)$ be a collection of k reconstructible transitions. Let $\tilde{P} = \otimes_{i=i}^k \tilde{P}_i^{n_i}$, $\tilde{\mathcal{Z}} = \times_{i=1}^k \tilde{\mathcal{Z}}_i^{n_i}$, $n = \sum_{i=1}^k n_i$ and $r_i = \frac{n_i}{n}$. Then for all algorithms $\mathcal{A} \in \mathbb{T}(\tilde{\mathcal{Z}}, A)$, distributions $P \in \mathbb{P}(\mathcal{Z})$ and bounded loss functions ℓ,*

$$\mathbb{E}_{\tilde{S}\sim\tilde{P}}\ell(P, \mathcal{A}(\tilde{S})) \leq \mathbb{E}_{\tilde{S}\sim\tilde{P}} \sum_{i=1}^k r_i \ell_{T_i}(\tilde{S}_i, \mathcal{A}(\tilde{S})) + K\sqrt{\frac{2\log(|A|)}{n}},$$

where $K = \sqrt{\sum_{i=1}^k r_i \|\ell_{T_i}\|_\infty^2}$.

A similar result also holds with high probability on draws from $\tilde{P}$. Theorem 3.4 is a generalization of the final bound appearing in Crammer et al. [47] that only pertains to symmetric label noise and binary classification. Theorem 3.4 suggests the following means of choosing data sets. Let c_i be the cost of acquiring data corrupted by T_i. First, choose data from the T_i with lowest $c_i\|\ell_{T_i}\|_\infty^2$ until picking more violates the budget constraint. Then choose data from the second lowest and so on.

One must be careful when comparing upper bounds, as there may exist alternate methods for learning from the corrupted sample with better properties. In the next section we present arguments indicating this is not the case.

3.4 Lower Bounds for Corrupted Learning

Thus far we have developed upper bounds for ERM algorithms. In particular we have found that reconstructible corruption does not affect the *rate* at which learning occurs, it only affects constants in the upper bound. Can we do better? Are these constants *tight*? To answer this question we develop lower bounds for corrupted learning.

Here we review Le Cam's method [86], a powerful technique for generating lower bounds for decision problems that very often gives the correct rate and dependence on constants (including being able to reproduce the standard VC dimension lower bounds for classification presented in [97]). In recent times it has been used to establish lower bounds for: differentially private learning [59], learning in a distributed set up [135], function evaluations required in convex optimization [1] as well as generic lower bounds in statistical estimation problems [130]. We show how this method can be extended using the strong data processing theorem [27; 41] to provide a general tool for lower bounding the possible performance attainable in corrupted prediction problems.

We stress here that these techniques apply to general experiments $e \in \mathbb{T}(\Theta, \mathcal{Z})$, and general loss functions $\ell : \Theta \times A \to \mathbb{R}$, and not only the predictive problems of interest here.

3.4.1 Le Cam's Method and Minimax Lower Bounds

Le Cam's method proceeds by reducing a general decision problem to an easier binary classification problem, before relating the best possible performance on this classification problem to the minimax risk. Let Θ be a set of unknowns, $e \in \mathbb{T}(\Theta, \mathcal{Z})$ an experiment and $\ell : \Theta \times A \to \mathbb{R}$ a loss. Recall the regret $\Delta\ell(\theta, a) = \ell(\theta, a) - \inf_{a'} \ell(\theta, a')$, and define the *separation* $\rho : \Theta \times \Theta \to \mathbb{R}$,

$$\rho(\theta_1, \theta_2) = \inf_a \Delta\ell(\theta_1, a) + \Delta\ell(\theta_2, a).$$

The separation measures how hard it is to act well against both θ_1 and θ_2 simultaneously.

Lemma 3.5. *For all experiments e, loss functions ℓ and $\theta_1, \theta_2 \in \Theta$,*

$$\Delta\underline{\mathcal{R}}_\ell(e) \geq \rho(\theta_1, \theta_2)\left(\frac{1}{4} - \frac{1}{4}V(e(\theta_1), e(\theta_2))\right).$$

where V is the variational divergence.

The reader is reminded that all proofs for this and following chapters are relegated to the appendix. This lower bound is a trade off between distances measured by ρ and statistical distances measured by the variational divergence. A decision problem is easy if proximity in variational divergence of $e(\theta_1)$ and $e(\theta_2)$ (hard to distinguish θ_1 and θ_2 statistically) implies proximity of θ_1 and θ_2 in ρ (hard to distinguish θ_1 and θ_2 with actions).

Replication and Rates

We wish to lower bound how the risk decreases as n grows. The following lemma provides a simple way to do this.

Lemma 3.6. *For all collections of distributions $P_i, Q_i \in \mathbb{P}(\mathcal{Z}_i)$, $i \in [1;k]$,*

$$V(\otimes_{i=1}^k P_i, \otimes_{i=1}^k Q_i) \leq \sum_{i=1}^k V(P_i, Q_i).$$

We make use of the specific case where $P_i = P$ and $Q_i = Q$ for all i. Lemma 3.6 and Lemma 3.5 yield the following.

Lemma 3.7. *For all experiments e, loss functions ℓ, $\theta_1, \theta_2 \in \Theta$ and n,*

$$\Delta\underline{\mathcal{R}}_\ell(e_n) \geq \rho(\theta_1, \theta_2)\left(\frac{1}{4} - \frac{n}{4}V(e(\theta_1), e(\theta_2))\right).$$

To use lemma 3.7, one defines $\theta_1 = \phi_1(n)$ and $\theta_2 = \phi_2(n)$ for $n \in [0, \infty)$, with the property,

$$\frac{1}{4} - \frac{n}{4}V(e(\theta_1), e(\theta_2)) \geq \frac{1}{8},$$

or equivalently $V(e(\theta_1), e(\theta_2)) \leq \frac{1}{2n}$. This can always be done, for example by taking $\phi_1 = \phi_2$. This yields a lower bound of,

$$\Delta\underline{\mathcal{R}}_\ell(e_n) \geq \frac{1}{8}\rho(\phi_1(n), \phi_2(n)).$$

To obtain *tight* lower bounds, ϕ needs to be designed in a problem dependent fashion. However, as our goal here is to reason *relatively* we assume that ϕ is given.

Other Methods for Obtaining Minimax Lower Bounds

There are many other techniques for lower bounds in terms of functions of pairwise KL divergences [131] (for example Assouad's method) as well as functions of pairwise f-divergences [68]. While such methods are often required to get tighter lower bounds, all of what follows can be applied to these more intricate lower bounding techniques. For the sake of conceptual clarity, we proceed with Le Cam's method.

3.4.2 Measuring the Amount of Corruption

Rather than the experiment e, in corrupted learning we work with the corrupted experiment $\tilde{e}$. The data processing theorem for f-divergences [105] states that,

$$V(T(P), T(Q)) \leq V(P, Q), \ \forall P, Q \in \mathbb{P}(\mathcal{Z}).$$

Thus any lower bound achieved by Le Cam's method for e can be directly transferred to one for $\tilde{e}$. However, this provides us with no means to rank different T. For some T, the data processing theorem can be *strengthened*, in the sense that one can find $\alpha(T) < 1$ such that,

$$V(T(P), T(Q)) \leq \alpha(T)V(P, Q), \ \forall P, Q \in \mathbb{P}(\mathcal{Z}).$$

The coefficient $\alpha(T)$ provides a means to measure the amount of corruption present in T. For example if T is constant and maps all P to the same distribution, then $\alpha(T) = 0$. If T is an invertible function, then $\alpha(T) = 1$. Together with lemma 3.7 this strong data processing theorem leads to meaningful lower bounds that allow the comparison of different corrupted experiments.

3.4.3 A Generic Strong Data Processing Theorem.

In light of the comments in section 3.4.1, following Cohen and Kempermann [41] we present a strong data processing theorem that works for all f-divergences. Recall the definition of an f-divergence from chapter two.

Definition 3.8. *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $f(1) = 0$. For all distributions $P, Q \in \mathbb{P}(\mathcal{Z})$ the f-*divergence *between P and Q is,*

$$D_f(P, Q) = \langle P, f\left(\tfrac{dQ}{dP}\right)\rangle.$$

Both the variational and KL divergence are examples of f divergences. For fixed T we seek an $\alpha(T)$ such that,

$$D_f(T(P), T(Q)) \leq \alpha(T) D_f(P, Q) \ \forall P, Q, f.$$

To do so we first relate the amount T *contracts* P and Q to a certain deconstruction for transitions before proving when such a deconstruction can occur.

Lemma 3.9. *For all transitions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and distributions $P, Q \in \mathbb{P}(\mathcal{Z})$, if there exists $F, G \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and $\lambda \in [0, 1]$ such that $T = \lambda F + (1 - \lambda) G$ with $F(P) = F(Q)$ then $D_f(T(P), T(Q)) \leq (1 - \lambda) D_f(P, Q)$ for all f.*

Hence the amount T contracts P and Q is related to the amount of T that fixes P and Q. We seek the largest λ such that a decomposition $T = \lambda F + (1 - \lambda) G$ is always possible, no matter what pair of distributions F is required to fix.

Lemma 3.10. *For all transitions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ define $\lambda(T) = \min_{i,j} \sum_k \min(T_{k,i}, T_{k,j})$. Then $\lambda \leq \lambda(T)$ if and only if for all pairs of distributions P, Q there exists a decomposition,*

$$T = \lambda F + (1 - \lambda) G,$$

with $F, G \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and $F(P) = F(Q)$.

Theorem 3.11 (Strong Data Processing)**.** *For all transitions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ define $\alpha(T) = 1 - \lambda(T)$. Then for all P, Q, f,*

$$D_f(T(P), T(Q)) \leq \alpha(T) D_f(P, Q).$$

The proof is a simple application of lemma 3.9 and lemma 3.10. It is easy to see that $0 \leq \alpha(T) \leq 1$. Furthermore $\alpha(T) = 0$ if and only if all of the columns of T are the same. While this α may not be the tightest for a given f, it will be shown in the next section that it is the tightest possible for variational divergence. Furthermore, it is *generic* and as such can be applied in all lower bounding methods mentioned in section 3.4.1.

3.4.4 Relating α to Variational Divergence

It can be shown [41] that $\alpha(T) = \max_{x_1,x_2} V(T(x_1), T(x_2)) = \frac{1}{2}\max_{i,j}\sum_k |T_{ki} - T_{kj}|$, the maximum $L1$ distance between the columns of A [105]. Furthermore,

$$\alpha(T) = \sup_{P,Q\in\mathbb{P}(X)} \frac{V(T(P),T(Q))}{V(P,Q)} = \sup_{v\in\Omega} \frac{\|T(v)\|_1}{\|v\|_1},$$

where $\Omega = \{v : \sum v_i = 0, v \neq 0\}$. Hence $\alpha(T)$ is the operator 1-norm of T when restricted to Ω. The above also shows that $\alpha(T)$ provides the tightest strong data processing theorem possible when using variational divergence.

We have the following compositional property of α.

Lemma 3.12. *For all transitions $T_1 \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_1)$ and $T_2 \in \mathbb{T}(\tilde{\mathcal{Z}}_1, \tilde{\mathcal{Z}}_2)$,*

$$\alpha(T_2 \circ T_1) \leq \alpha(T_2)\alpha(T_1) \leq \min(\alpha(T_2), \alpha(T_1)).$$

Hence $T_2 \circ T_1$ is at least as corrupt as either of the T_i.

The first use of $\alpha(T)$ occurred in the work of Dobrushin [58] where it is called the coefficient of ergodicity and is used (much like in [27]) to prove rates of convergence of Markov chains to their stationary distribution.

3.4.5 Lower bounds Relative to the Amount of Corruption

The strong data processing theorem and Le Cam's method provide lower bounds for corrupted decisions problems.

Lemma 3.13. *For all experiments e, loss functions ℓ, $\theta_1, \theta_2 \in \Theta$, n and corruptions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$,*

$$\Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n) \geq \rho(\theta_1, \theta_2)\left(\frac{1}{4} - \frac{\alpha(T)n}{4} V(e(\theta_1), e(\theta_2))\right).$$

The proof is a simple application of lemma 3.7 and the strong data processing theorem. Suppose we have proceeded as in section 3.4.1, defining $\theta_1 = \phi_1(n)$ and $\theta_2 = \phi_2(n)$ with $V(e(\theta_1), e(\theta_2)) \leq \frac{1}{2n}$. Letting $\tilde{\theta}_1 = \phi_1(\alpha(T)n)$ and $\tilde{\theta}_2 = \phi_2(\alpha(T)n)$ gives $V(e(\tilde{\theta}_1), e(\tilde{\theta}_2)) \leq \frac{1}{2\alpha(T)n}$. Furthermore,

$$\Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n) \geq \frac{1}{8}\rho(\phi_1(\alpha(T)n), \phi_2(\alpha(T)n)).$$

In words, if ever Le Cam's method gives a lower bound of $f(n)$ for repetitions of the clean experiment, we obtain a lower bound of $f(\alpha(T)n)$ for repetitions of the corrupted experiment.

Theorem 3.14. *For all experiments $e \in \mathbb{T}(\Theta, \mathcal{Z})$ and corruptions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$, if Le Cam's method yields a lower bound $\Delta\underline{\mathcal{R}}_\ell(e_n) \geq f(n)$ then $\Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n) \geq f(\alpha(T)n)$.*

In particular if Le Cam's method yields a lower bound of $\frac{C}{\sqrt{n}}$ for the clean problem, as is usual for many machine learning problems, theorem 3.14 yields a lower bound of $\frac{C}{\sqrt{\alpha(T)n}}$ for the corrupted problem. The *rate* at which one learns is unaffected, only the constants. A penalty factor $\alpha(T)$ is unavoidable no matter what learning algorithm is used.

3.4.6 Lower Bounds for Combinations of Corrupted Data

As in section 3.3.1 we present lower bounds for combinations of corrupted data. For example in learning with noisy labels out of n data, there could be n_1 clean, n_2 slightly noisy and n_3 very noisy samples and so on.

Theorem 3.15. *Let $T_i \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_i)$, $i \in [1;k]$, be k reconstructible transitions. Let $T = \otimes_{i=i}^k T_i^{n_i}$ with $n = \sum_{i=i}^k n_k$. If Le Cam's method yields a lower bound $\Delta\underline{\mathcal{R}}_\ell(e_n) \geq f(n)$ then $\Delta\underline{\mathcal{R}}_\ell(T \circ e_n) \geq f(K)$ where $K = \left(\sum\limits_{i=1}^k \alpha(T_i)n_i\right)$.*

As in section 3.3.1, this bound suggest means of choosing data sets, via the following integer program,

$$\operatorname*{arg\,max}_{n_1,n_2\ldots n_k} \sum_{i=1}^k \alpha(T_i)n_i \ \text{ subject to} \sum_{i=1}^k c_i n_i \leq C,$$

where c_i is the cost of acquiring data corrupted by T_i and C is the maximum total cost. This is exactly the unbounded knapsack problem [50] which admits the following near optimal greedy algorithm. First, choose data from the T_i with highest $\frac{\alpha(T_i)}{c_i}$ until picking more violates the constraints. Then pick from the second highest and so on.

3.4.7 Applying the Bounds to Supervised Learning

Our lower bounds as stated apply to *general* decision problems. Of particular interest are supervised learning problems where $\mathcal{Z} = X \times Y$, and the corruption is entirely on the labels, $T \in \mathbb{T}(Y, \tilde{Y})$, with the instances unaffected. We show here how our lower bounds can be applied to such problems.

Any distribution $P \in \mathbb{P}(X \times Y)$ can be decomposed as a marginal over instances π_X, and a transition $\eta \in \mathbb{T}(X, Y)$. With slight abuse of notation we write $T(P)$ for the label corrupted P. Normally when constructing lower bounds for supervised learning problems, the marginal distribution over instances, π_X is fixed and the conditional distribution η is varied. For example, in Massart et al [97], lower bounds for learning binary classifiers are constructed. There the marginal distribution over instances, π_X, is assumed to be concentrated on a set that the particular class of binary classifiers under consideration shatters. For distributions P_1 and P_2 with the same marginals over instances,

$$V(P_1, P_2) = \mathbb{E}_{x \sim \pi_X} V(\eta_1(x), \eta_2(x)).$$

Furthermore, when corrupted by noise only on the labels,

$$V(T(P_1), T(P_2)) = \mathbb{E}_{x \sim \pi_X} V(T \circ \eta_1(x), T \circ \eta_2(x)) \leq \alpha(T) V(P_1, P_2).$$

Therefore only properties of the label corruption are required to apply our lower bounding techniques.

3.5 Measuring the Tightness of the Upper Bounds and Lower Bounds

In the previous sections we have shown upper bounds that depend on $\|\ell_T\|_\infty$ as well as lower bounds that depend on $\alpha(T)$. Here we compare these bounds.

Recall from theorem 3.2 $\ell_T(-, a) = R^*(\ell(-, a))$. The worst case ratio $\frac{\|\ell_T\|_\infty}{\|\ell\|_\infty}$ is determined by the *operator norm* of R^*. For a linear map $R : \mathbb{R}^X \to \mathbb{R}^Y$ define,

$$\|R\|_1 := \sup_{v \in \mathbb{R}^X} \frac{\|Rv\|_1}{\|v\|_1} \text{ and } \|R\|_\infty := \sup_{v \in \mathbb{R}^X} \frac{\|Rv\|_\infty}{\|v\|_\infty}$$

which are two operator norms of R. They are equal to the maximum absolute column and row sum of R respectively [22]. Hence $\|R\|_1 = \|R^*\|_\infty$.

Lemma 3.16. *For all losses ℓ, $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and reconstructions R, $\frac{\|\ell_T\|_\infty}{\|\ell\|_\infty} \leq \|R^*\|_\infty$.*

Lemma 3.17. *If $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is reconstructible, with reconstruction R, then,*

$$\frac{1}{\alpha(T)} \leq 1 / \left(\inf_{u \in \mathbb{R}^X} \frac{\|Tu\|_1}{\|u\|_1} \right) \leq \|R^*\|_\infty.$$

Note that for lower bounds we look at the *best* case separation of columns of T, for upper bounds we essentially use the *worst*. We also get the following compositional theorem.

Lemma 3.18. *If $T_1 \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_1)$ and $T_2 \in \mathbb{T}(\tilde{\mathcal{Z}}_1, \tilde{\mathcal{Z}}_2)$ are reconstructible, with reconstructions R_1 and R_2, then $T_2 \circ T_1$ is reconstructible with reconstruction $R_1 \circ R_2$. Furthermore,*

$$\frac{1}{\alpha(T_1)\alpha(T_2)} \leq \|R_1 \circ R_2\|_1 \leq \|R_1\|_1 \|R_2\|_1.$$

Proof. The first statement is obvious. For the first inequality simply use lemma 3.17 followed by lemma 3.12. The second inequality is an easy to prove property of operator norms. □

3.5.1 Comparing Theorems 3.3 and 3.14

We have shown the following implication, for all reconstructible T,

$$\frac{C_1}{\sqrt{n}} \leq \Delta\underline{\mathcal{R}}_\ell(e_n) \leq \frac{C_2\|\ell\|_\infty}{\sqrt{n}} \Rightarrow \frac{C_1}{\sqrt{\alpha(T)n}} \leq \Delta\underline{\mathcal{R}}_\ell(\tilde{e}_n) \leq \frac{C_2\|\ell_T\|_\infty}{\sqrt{n}}.$$

By lemma 3.17, in the worst case $\|\ell_T\|_\infty \geq \frac{\|\ell\|_\infty}{\alpha(T)}$. Thus in the worst case over all losses, we arrive at upper and lower bounds for the corrupted problem that are at least factor of $\frac{1}{\sqrt{\alpha(T)}}$ apart. We do not know if this is the fault of our upper or lower bounding techniques. However, for *specific* ℓ and T this gap can be smaller.

For example, in the problem of learning with symmetric label noise discussed in section 3.2.1, with misclassification loss ℓ_{01},

$$\alpha(T) = 1 - 2\sigma \text{ and } \|\ell_{01,T}\| = \frac{1-\sigma}{1-2\sigma},$$

respectively. The worst case ratio of upper and lower bounds over all losses is of order $\frac{1}{\sqrt{1-2\sigma}}$. For misclassification loss the actual ratio is $\frac{1-\sigma}{\sqrt{1-2\sigma}}$. For all $\sigma \in [0, \frac{2}{10}]$, i.e. up to 0.2 flip probability, this ratio is never larger than $\frac{4}{\sqrt{15}} \approx 1.03$.

3.5.2 Comparing Theorems 3.4 and 3.15

Assuming c_T is the cost of acquiring data corrupted by T, theorem 3.15 ranks the utility of different corruptions by $\frac{1}{\|\ell_T\|_\infty^2 c_T}$ where as theorem 3.15 ranks by $\frac{\alpha(T)}{c_T}$. By lemma 3.17, $\frac{1}{\alpha(T)}$ is a proxy for $\frac{\|\ell\|_\infty}{\|\ell_T\|_\infty}$ meaning both theorems are "doing the same thing". In theorems 3.15 and 3.4 we have, respectively, best case and a worst case loss specific method for choosing data sets. Theorem 3.4 combined with 1emma 3.16 provides a worst case loss insensitive method for choosing data sets.

3.6 Canonical Losses and Convexity

Recall the notion of a canonical loss from theorem 2.16 of chapter 2. All canonical losses remain convex when corrected for corruption.

Theorem 3.19 (Preservation of Convexity)**.** *Let $\mathcal{L} : \mathcal{Z} \times C \to \mathbb{R}$ be a canonical loss, i.e. $C \subseteq \mathbf{1}_{\mathcal{Z}}^{\perp}$ is a convex set and,*

$$\mathcal{L}(z,v) = v(z) + \Psi(v),$$

for a convex function Υ. For all reconstructible $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ there exists a reconstruction R with,

$$\mathcal{L}_T(\tilde{z}, v) = \langle R(\delta_{\tilde{z}}), v\rangle + \Upsilon(v).$$

Furthermore this loss is convex in v.

Therefore we need not abandon the framework of convex surrogates when the corruption is known.

3.7 Learning when the Corruption Process is Partially Known

Thus far we have considered the problem of learning when T is known. Here we consider the problem of when $T \in \mathcal{C}$, a subset of possible reconstructible corruptions $\mathcal{C} \subset \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$. For example when learning classifiers under symmetric label noise [4], the corruption is of the form,

$$T_\sigma = \begin{pmatrix} 1-\sigma & \sigma \\ \sigma & 1-\sigma \end{pmatrix},$$

where $\sigma \in (0, \frac{1}{2})$. There are three ways in which one can proceed.

If we assume access to a "gold standard" sample $S{\sim}P$ as well as a corrupted sample $\tilde{S}$, we can use methods akin to those in Kearns [81]. One covers the set C to some tolerance ϵ with a finite cover $\{T_i\}_{i=1}^k$. For each T_i in the cover, estimate an action a_i using ℓ_{T_i} and the corrupted sample. Finally, choose the a_i that best predicts the gold standard sample. Using theorem 3.3, we know that for a large enough corrupted sample, one of the a_i has performance close to that of the optimal a.

One can attempt to *estimate* T from the corrupted sample. Under certain distributional assumptions (such as separability), Menon et al. [100] surveys methods for estimating T for the problem of learning under asymmetric label noise. While there is currently no firm theory on the performance of these estimators, they are shown in [100] to work empirically. These methods can be easily extended to general corruptions.

In both of the above methods, operator norms can provide suitable losses/metrics that can guide their use.

Lemma 3.20. *Let $T, T' \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ be reconstructible. Then,*

$$\|\ell_T - \ell_{T'}\|_\infty = \|R - R'\|_1 \, \|\ell\|_\infty .$$

The quantity $\|R - R'\|_1$ is a statistically motivated distance that can be used when covering $\mathcal{C}$. Furthermore, it can be used when designing loss functions for estimating T.

Finally, one can look for loss functions that are robust to $\mathcal{C}$. We explore this approach further in section 3.22 and in chapter 4.

3.8 Conclusion

Real world data sets are amalgamations of data of variable type and quality. Understanding how to *learn from* and *compare* different corrupted data sets is therefore a problem of great practical importance. Theorems 3.4 and 3.15 provide means to do this. Future work will attempt to further refine these methods as well as extend the framework to non reconstructible problems such as multiple instance learning and learning with label proportions.

Appendix to Chapter 3

3.9 Examples

Here we show examples of common corrupted machine learning problems.

3.9.1 Noisy Labels

We consider the problem of learning from noisy binary labels [4; 101]. Here σ_i is the probability that class i is has its label flipped. We have,

$$T = \begin{pmatrix} 1-\sigma_{-1} & \sigma_1 \\ \sigma_{-1} & 1-\sigma_1 \end{pmatrix} \; R^* = \frac{1}{1-\sigma_{-1}-\sigma_1} \begin{pmatrix} 1-\sigma_1 & -\sigma_{-1} \\ -\sigma_1 & 1-\sigma_{-1} \end{pmatrix}.$$

This yields,

$$\ell_T(y,a) = \frac{(1-\sigma_{-y})\ell(y,a) - \sigma_y \ell(-y,a)}{1-\sigma_{-1}-\sigma_1}.$$

The above equation is lemma 1 in Natarajan et al. [101] and is the original method of unbiased estimators. Interestingly, even if ℓ is positive, ℓ_T can be negative. If the noise is symmetric with $\sigma_{-1} = \sigma_1 = \sigma$ and ℓ is 01 loss then,

$$\ell_T(y,a) = \frac{\ell_{01}(y,a) - \sigma}{1-2\sigma},$$

which is just a rescaled and shifted version of 01 loss. If we work in the realizable setting, i.e. there is some $f \in \mathcal{F}$ with,

$$\mathbb{E}_{(x,y)\sim P} \ell_{01}(y, f(x)) = 0,$$

then the above provides an interesting correspondence between learning with symmetric label noise and learning under distributions with large Tsybakov margin [6]. Taking $\sigma = \frac{1}{2} - h$ with P *separable* in turn implies $\tilde{P}$ has Tsybakov margin h. This means bounds developed for this setting, such as in Massart et al [97], can be transferred to the setting of learning with symmetric label noise. Our lower bound reproduces the results of Massart et al [97].

Below is a table of the relevant parameters for learning with noisy binary labels. These results directly extend those presented in [81] that considered only the case of symmetric label noise.

Learning with Label Noisy (Binary)

T	$\begin{pmatrix} 1-\sigma_{-1} & \sigma_1 \\ \sigma_{-1} & 1-\sigma_1 \end{pmatrix}$
R^*	$\frac{1}{1-\sigma_{-1}-\sigma_1}\begin{pmatrix} 1-\sigma_1 & -\sigma_{-1} \\ -\sigma_1 & 1-\sigma_{-1} \end{pmatrix}$
$\alpha(T)$	$\lvert 1-\sigma_{-1}-\sigma_1\rvert$
$\lVert R^*\rVert_\infty$	$\frac{1}{\lvert 1-\sigma_{-1}-\sigma_1\rvert}\max(1-\sigma_{-1}+\sigma_1, 1-\sigma_1+\sigma_{-1})$
$\lVert \ell_{01,T}\rVert_\infty$	$\frac{1}{\lvert 1-\sigma_{-1}-\sigma_1\rvert}\max(1-\sigma_{-1}, 1-\sigma_1, \sigma_{-1}, \sigma_1)$

We see that as long as $\sigma_{-1}+\sigma_1 \neq 1$, T is reconstructible. The pattern we see in this table is quite common. $\lVert R^*\rVert_\infty$ tends to be marginally greater than $\frac{1}{\alpha(T)}$, with $\lVert \ell_{01,T}\rVert_\infty$ less than both. In the symmetric case our lower bound reproduces that of Aslam and Decatur [5].

Finally, when working with symmetric label noise ($\sigma_{-1} = \sigma_1 = \sigma$),

$$\lVert R_\sigma - R_{\sigma'}\rVert_1 = \frac{2\lvert\sigma-\sigma'\rvert}{\lvert 1-2\sigma\rvert\lvert 1-2\sigma'\rvert}.$$

For fixed true noise rate σ, the presence of a factor $\lvert 1-2\sigma'\rvert$ in the denominator means that underestimating σ is preferred to overestimating. Hence when designing estimates for σ, those with negative bias might perform better than those that are unbiased or are positively biased. Furthermore, when covering noise rates, as per the discussion in 3.7, more focus should be given to higher noise rates than to lower.

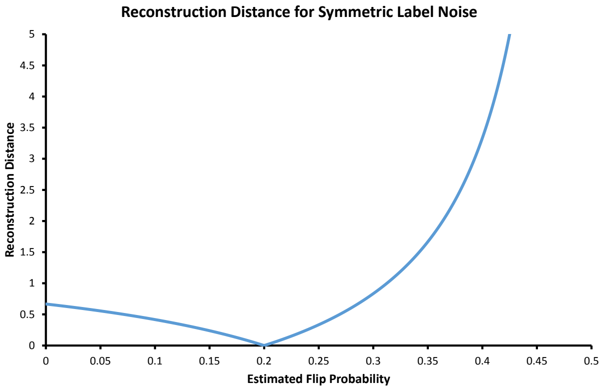


Fig. 3.1: Plot of $\lVert R_\sigma - R_{\sigma'}\rVert_1$ for $\sigma = 0.2$. $\lVert R_\sigma - R_{\sigma'}\rVert_1$ is a measure of how far apart two corruptions are. This distance measure can be used when constructing estimators for the corruption process T. See text.

3.9.2 Semi-Supervised Learning

We consider the problem of semi-supervised learning [37]. Here $1 - \sigma_i$ is the probability class i has a missing label. We first consider the easier symmetric case where $\sigma_{-1} = \sigma_1 = \sigma$.

Symmetric Semi-Supervised Learning

T	$\begin{pmatrix} \sigma & 0 \\ 0 & \sigma \\ 1-\sigma & 1-\sigma \end{pmatrix}$
R^*	$\begin{pmatrix} \frac{1-2\sigma+2\sigma^2}{1-3\sigma+5\sigma^2-3\sigma^3} & \frac{-\sigma^2}{1-3\sigma+5\sigma^2-3\sigma^3} \\ \frac{-\sigma^2}{1-3\sigma+5\sigma^2-3\sigma^3} & \frac{1-2\sigma+2\sigma^2}{1-3\sigma+5\sigma^2-3\sigma^3} \\ \frac{\sigma}{1-2\sigma+3\sigma^2} & \frac{\sigma}{1-2\sigma+3\sigma^2} \end{pmatrix}$
$\alpha(T)$	σ
$\lVert R^* \rVert_\infty$	$\frac{1}{\sigma}$
$\lVert \ell_{01,T} \rVert_\infty$	$\frac{1-2\sigma+2\sigma^2}{2\sigma+3\sigma-5\sigma^2}$

Once again $\lVert \ell_{01,T} \rVert_\infty \leq \frac{1}{\alpha(T)}$. Our lower bound confirms that in general unlabelled data does not help [11]. Rather than using the method of unbiased estimators, one could simply throw away the unlabelled data leaving behind σn labelled instances on average. To make further progress in this problem, as noted elsewhere, normally one assumes some form of compatibility between the marginal distribution of instances and the optimal classifier. In principle, restricted versions of Le Cams method and the strong data processing inequality could be used to give lower bounds under these different assumptions. As our interest here are minimax bounds, we do not pursue these methods.

Semi-Supervised Learning

T	$\begin{pmatrix} \sigma_{-1} & 0 \\ 0 & \sigma_1 \\ 1-\sigma_{-1} & 1-\sigma_1 \end{pmatrix}$
$\alpha(T)$	$\max_i \sigma_i$

Other parameters for the more general case are omitted due to complexity (they involve the maximum of three 4th order rational equations). They are available in closed form.

3.9.3 Three Class Symmetric Label Noise

Here we present parameters for the three class variant of symmetric label noise. We have $\tilde{Y} = Y = \{1, 2, 3\}$ with $P(\tilde{Y} = \tilde{y} | Y = y) = 1 - \sigma$, if $y = \tilde{y}$ and $\frac{\sigma}{2}$ otherwise.

Learning with Symmetric Label Noisy (Multiclass)

T	$\begin{pmatrix} 1-\sigma & \frac{\sigma}{2} & \frac{\sigma}{2} \\ \frac{\sigma}{2} & 1-\sigma & \frac{\sigma}{2} \\ \frac{\sigma}{2} & \frac{\sigma}{2} & 1-\sigma \end{pmatrix}$
R^*	$\begin{pmatrix} \frac{2-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} \\ \frac{-\sigma}{2-3\sigma} & \frac{2-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} \\ \frac{-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} & \frac{2-\sigma}{2-3\sigma} \end{pmatrix}$
$\alpha(T)$	$\lvert 1-\frac{3}{2}\sigma \rvert$
$\lVert R^* \rVert_\infty$	$\frac{2+\sigma}{\lvert 2-3\sigma \rvert}$
$\lVert \ell_{01,T} \rVert_\infty$	$\frac{2}{\lvert 2-3\sigma \rvert}\max(\sigma, 1-\sigma)$

We see that as long as $\sigma \neq \frac{2}{3}$, T is reconstructible. Once again $\lVert \ell_{01,T} \rVert_\infty \leq \frac{1}{\alpha(T)}$.

3.9.4 Partial Labels

Here we follow [45] with $Y = \{1, 2, 3\}$ and $\tilde{Y} = \{0, 1\}^Y$ the set of partial labels. A partial label of $(0, 1, 1)$ indicates that the true label is either 2 or 3 but not 1. We assume that a partial label always includes the true label as one of the possibilities and furthermore that spurious labels are added with probability σ.

Learning with Partial Labels

T	$\begin{pmatrix} 0 & 0 & (1-\sigma)^2 \\ 0 & (1-\sigma)^2 & 0 \\ 0 & (1-\sigma)\sigma & (1-\sigma)\sigma \\ (1-\sigma)^2 & 0 & 0 \\ (1-\sigma)\sigma & 0 & (1-\sigma)\sigma \\ (1-\sigma)\sigma & (1-\sigma)\sigma & 0 \\ \sigma^2 & \sigma^2 & \sigma^2 \end{pmatrix}$
$\alpha(T)$	$1-\sigma$

We see that as long as $\sigma \neq 1$, T is reconstructible. In this case $\lVert \ell_{01,T} \rVert_\infty$ and $\lVert R^* \rVert_\infty$ are given by more complicated expressions (however they are both available in closed form). We display their interrelation in a graph in below. To the best of our knowledge, no upper and lower bounds are present in the literature for this problem.

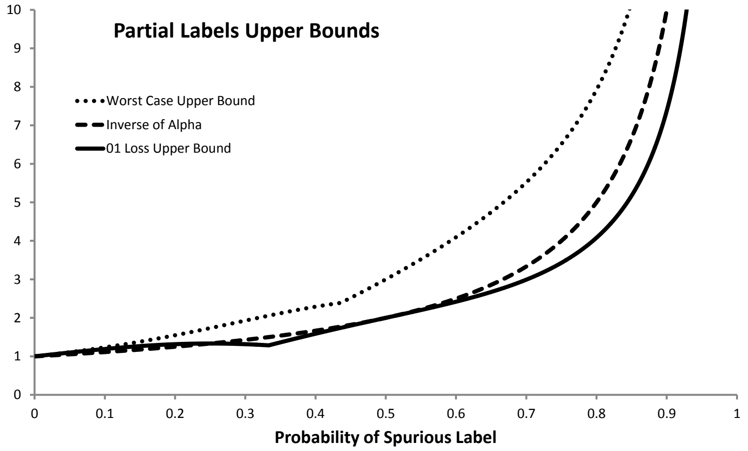


Fig. 3.2: Upper and lower bounds for the problem of learning from partial labels, see text.

3.10 Proof of Theorem 3.4

We actually prove a more general theorem, that works for infinite action sets.

Theorem. *Let $T_i \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_i)$ be a collection of k reconstructible transitions. Let $\tilde{P} = \otimes_{i=i}^{k} \tilde{P}_i^{n_i}$, $\tilde{\mathcal{Z}} = \times_{i=1}^{k} \tilde{\mathcal{Z}}_i^{n_i}$, $n = \sum_{i=1}^{k} n_i$ and $r_i = \frac{n_i}{n}$. Then for all algorithms $\mathcal{A} \in \mathbb{T}(\tilde{\mathcal{Z}}, A)$, priors $\pi \in \mathbb{P}(A)$, distributions $P \in \mathbb{P}(\mathcal{Z})$ and bounded loss functions ℓ,*

$$\mathbb{E}_{\tilde{S}\sim\tilde{P}}\ell(P,\mathcal{A}(\tilde{S})) \le \mathbb{E}_{\tilde{S}\sim\tilde{P}}\sum_{i=1}^{k} r_i \ell_{T_i}(\tilde{S}_i,\mathcal{A}(\tilde{S})) + K\sqrt{\frac{2\mathbb{E}_{S\sim P^n} D_{KL}(\mathcal{A}(S),\pi)}{n}}.$$

where $K = \sqrt{\sum_{i=1}^{k} r_i \|\ell_{T_i}\|_{\infty}^{2}}$.

Proof. Define $L(\tilde{S}, a) = \sum_{i=1}^{k} \sum_{\tilde{z}\in\tilde{S}_i} \ell_{T_i}(z_i, a)$, the sum of the corrupted losses on the sample. We have by theorem A.3 of the appendix,

$$\mathbb{E}_{\tilde{S}\sim Q}\mathbb{E}_{a\sim\mathcal{A}(\tilde{S})} - \frac{1}{\beta}\log(\mathbb{E}_{\tilde{S}'\sim Q} e^{-\beta L(\tilde{S}',a)}) \le \mathbb{E}_{\tilde{S}\sim Q}\left[L(\tilde{S},\mathcal{A}(\tilde{S})) + \frac{D_{KL}(\mathcal{A}(S),\pi)}{\beta}\right]$$
$$\sum_{i=1}^{k} n_i \mathbb{E}_{\tilde{S}\sim Q}\mathbb{E}_{a\sim\mathcal{A}(\tilde{S})} - \frac{1}{\beta}\log(\mathbb{E}_{\tilde{z}\sim\tilde{P}_i} e^{-\beta\ell_{T_i}(\tilde{z},a)}) \le \mathbb{E}_{\tilde{S}\sim Q}\left[L(\tilde{S},\mathcal{A}(\tilde{S})) + \frac{D_{KL}(\mathcal{A}(S),\pi)}{\beta}\right]$$

where the first line follows from the theorem and the second from properties of the cumulant generating function. Invoking lemma A.8 of the appendix yields,

$$\sum_{i=1}^{k} n_i \left(\mathbb{E}_{\tilde{S}\sim Q} \ell_{T_i}(\tilde{P}_i, \mathcal{A}(\tilde{S})) - \frac{\|\ell_{T_i}\|_\infty^2 \beta}{2} \right) \leq \mathbb{E}_{\tilde{S}\sim Q} \left[L(\tilde{S}, \mathcal{A}(\tilde{S})) + \frac{D_{KL}(\mathcal{A}(S), \pi)}{\beta} \right].$$

As the T_i are reconstructible,

$$\mathbb{E}_{\tilde{S}\sim Q} \ell(P, \mathcal{A}(\tilde{S})) \leq \frac{1}{n} \mathbb{E}_{\tilde{S}\sim Q} \left[L(\tilde{S}, \mathcal{A}(\tilde{S})) + \frac{D_{KL}(\mathcal{A}(S), \pi)}{\beta} \right] + \frac{\left(\sum_{i=1}^{k} r_i \|\ell_{T_i}\|_\infty^2 \right) \beta}{2}.$$

Optimizing over β yields the desired result.

□

Theorem 3.4 is recovered by taking A finite, π uniform on A and upper bounding $D_{KL}(\mathcal{A}(S), \pi) \leq \log(|A|)$.

3.11 Le Cam's Method and Minimax Lower Bounds

The development here closely follows [59] with some streamlining. We consider a general decision problem with unknowns Θ, observation space $\mathcal{Z}$ and loss $\ell : \Theta \times A \to \mathbb{R}$. Recall the regret,

$$\Delta\ell(\theta, a) = \ell(\theta, a) - \inf_{a' \in A} \ell(\theta, a').$$

For any learning algorithm $\mathcal{A} \in \mathbb{T}(\mathcal{Z}, A)$, we wish to lower bound,

$$\sup_{\theta} \mathbb{E}_{z\sim e(\theta)} \Delta\ell(\theta, \mathcal{A}(z)).$$

The method proceeds by reducing a general decision problem to an easier binary classification problem. We consider a supremum over a restricted set $\{\theta_1, \theta_2\}$. Using Markov's inequality we then relate this to the minimum 01 loss in a particular binary classification problem. Finally one finds a lower bound for this quantity. With $\theta{\sim}\{\theta_1, \theta_2\}$ meaning θ is drawn uniformly at random from the set $\{\theta_1, \theta_2\}$, we have,

$$\begin{aligned} \sup_{\theta} \mathbb{E}_{z\sim e(\theta)} \mathbb{E}_{a\sim \mathcal{A}(z)} \Delta\ell(\theta, a) &\geq \sup_{\{\theta_1, \theta_2\}} \mathbb{E}_{z\sim e(\theta)} \mathbb{E}_{a\sim \mathcal{A}(z)} \Delta\ell(\theta, a) \\ &\geq \mathbb{E}_{\theta\sim\{\theta_1,\theta_2\}} \mathbb{E}_{z\sim e(\theta)} \mathbb{E}_{a\sim \mathcal{A}(z)} \Delta\ell(\theta, a) \\ &\geq \delta \mathbb{E}_{\theta\sim\{\theta_1,\theta_2\}} \mathbb{E}_{z\sim e(\theta)} \mathbb{E}_{a\sim \mathcal{A}(z)} [\![\Delta\ell(\theta, a) \geq \delta]\!]. \end{aligned}$$

Recall the *separation* $\rho : \Theta \times \Theta \to \mathbb{R}$, $\rho(\theta_1, \theta_2) = \inf_a \Delta\ell(\theta_1, a) + \Delta\ell(\theta_2, a)$. The separation measures how hard it is to act well against both θ_1 and θ_2 simultaneously. We now assume $\rho(\theta_1, \theta_2) > 2\delta$. Define $f : A \to \{\theta_1, \theta_2, \text{error}\}$ where $f(a) = \theta_i$ if $\Delta\ell(\theta_i, a) < \delta$ and error otherwise. This function is well defined as if there exists an action a with $\Delta\ell(\theta_1, a) < \delta$ and $\Delta\ell(\theta_2, a) < \delta$ then $\rho(\theta_1, \theta_2) < 2\delta$, a contradiction. Let $\hat{\mathcal{A}}$ be the classifier that first draws $a \sim \mathcal{A}(z)$ and then outputs $f(a)$. We have,

$$\begin{aligned}\sup_\theta \mathbb{E}_{z\sim e(\theta)} \mathbb{E}_{a\sim\mathcal{A}(z)} \Delta\ell(\theta, a) &\geq \delta \mathbb{E}_{\theta\sim\{\theta_1,\theta_2\}} \mathbb{E}_{z\sim e(\theta)} \mathbb{E}_{\theta'\sim\hat{\mathcal{A}}(z)} [\![\theta \neq \theta']\!] \\ &\geq \delta \inf_{\hat{\mathcal{A}}\in\mathbb{T}(\mathcal{Z},\Theta)} \mathbb{E}_{\theta\sim\{\theta_1,\theta_2\}} \mathbb{E}_{z\sim e(\theta)} \mathbb{E}_{\theta'\sim\hat{\mathcal{A}}(z)} [\![\theta \neq \theta']\!] \\ &= \delta \left(\frac{1}{2} - \frac{1}{2} V(e(\theta_1), e(\theta_2)) \right),\end{aligned}$$

where the first line is a rewriting of of the previous in terms of the classifier $\hat{\mathcal{A}}$, the second takes an infimum over all classifiers and the final line is a standard result in theoretical statistics [105]. Taking $\delta = \frac{\rho(\theta_1,\theta_2)}{2}$ yields lemma 3.5.

3.12 Extension of Le Cam's Method to Bayesian Risk

Rather than lower bounding $\sup_\theta \mathbb{E}_{z\sim e(\theta)} \Delta\ell(\theta, \mathcal{A}(z))$, a Bayesian with some knowledge about the unknown, given in the form of a prior $\pi \in \mathbb{P}(\Theta)$, wishes to lower bound the Bayesian risk,

$$\mathbb{E}_{\theta\sim\pi} \mathbb{E}_{z\sim e(\theta)} \Delta\ell(\theta, \mathcal{A}(z)).$$

Following from the second line of the derivation of Le Cam's method, we have a lower bound,

$$\begin{aligned}\mathbb{E}_{\theta\sim\{\theta_1,\theta_2\}} \mathbb{E}_{z\sim e(\theta)} \mathbb{E}_{a\sim\mathcal{A}(z)} \Delta\ell(\theta, a) &= \frac{1}{2}\mathcal{R}_\ell(\theta_1, e, \mathcal{A}) + \frac{1}{2}\mathcal{R}_\ell(\theta_2, e, \mathcal{A}) \\ &\geq \rho(\theta_1, \theta_2) \left(\frac{1}{4} - \frac{1}{4} V(e(\theta_1), e(\theta_2)) \right).\end{aligned}$$

Let $\mu \in \mathbb{P}(\Theta \times \Theta)$ be any distribution with *both* marginals over Θ equal to π. Averaging over this distribution we have,

$$\mathbb{E}_{\theta\sim\pi} \mathcal{R}_\ell(\theta, e, \mathcal{A}) \geq \mathbb{E}_{(\theta_1,\theta_2)\sim\mu} \rho(\theta_1, \theta_2) \left(\frac{1}{4} - \frac{1}{4} V(e(\theta_1), e(\theta_2)) \right).$$

This insight leads to a Bayesian version of lemma 3.5.

Lemma 3.21. *Let $\mu \in \mathbb{P}(\Theta \times \Theta)$ be any distribution with* both *marginals over Θ equal to π. Then for all experiments e and loss functions ℓ,*

$$\Delta\underline{\mathcal{R}}_\ell^\pi(e) \geq \mathbb{E}_{(\theta_1,\theta_2)\sim\mu} \rho(\theta_1, \theta_2) \left(\frac{1}{4} - \frac{1}{4} V(e(\theta_1), e(\theta_2)) \right).$$

Using this in place of lemma 3.5 leads to Bayesian versions of theorems 3.14 and 3.15.

3.13 Proof of Lemma 3.6

Proof. Firstly V is a *metric* on $\mathbb{P}(\times_{n=1}^{k}\mathcal{Z}_i)$ [105]. Thus,

$$\begin{aligned}V(\otimes_{i=1}^{k}P_i, \otimes_{i=1}^{k}Q_i) &= V(P_1 \otimes (\otimes_{i=2}^{k}P_i), Q_1 \otimes (\otimes_{i=2}^{k}Q_i))\\ &\leq V(P_1 \otimes (\otimes_{i=2}^{k}P_i), Q_1 \otimes (\otimes_{i=2}^{k}P_i)) + V(Q_1 \otimes (\otimes_{i=2}^{k}P_i), Q_1 \otimes (\otimes_{i=2}^{k}Q_i))\\ &= V(P_1, Q_1) + V(\otimes_{i=2}^{k}P_i, \otimes_{i=2}^{k}Q_i),\end{aligned}$$

where the first line is by definition, the second as V is a metric and the third is easily verified from the definition of V. To complete the proof proceed inductively. □

3.14 Proof of Lemma 3.9

Proof.

$$\begin{aligned}D_f(T(P), T(Q)) &= D_f(\lambda F(P) + (1-\lambda)G(P), \lambda F(Q) + (1-\lambda)G(Q))\\ &\leq \lambda D_f(F(P), F(Q)) + (1-\lambda)D_f(G(P), G(Q))\\ &= (1-\lambda)D_f(G(P), G(Q))\\ &\leq (1-\lambda)D_f(P, Q),\end{aligned}$$

where the first line follows from the definition, the second from the joint convexity of f-divergences [105], the third because $F(P) = F(Q)$ and $D_f(P, P) = 0$ and finally the fourth is from the standard data processing inequality.

□

3.15 Proof of Lemma 3.10

The proof of the forward implication is lemma 2 of [27]. We prove the reverse implication.

Proof. As this decomposition works for all pairs of distributions we can take $P = \delta_{x_i} = e_i$ and $Q = \delta_{x_j} = e_j$. As $F(P) = F(Q)$ we must have $F_{ki} = F_{kj} = v_k$ for all k. As all of the entries of $(1-\lambda)G$ are positive, we have $\lambda v_k \leq T_{ki}$ and $\lambda v_k \leq T_{kj}$. Hence $\lambda v_k \leq \min(T_{ki}, T_{kj})$. Summing over k and remembering that F is column stochastic gives $\lambda \leq \sum_k \min(T_{k,i}, T_{k,j})$. As i and j are arbitrary we have the desired result. □

3.16 Proof of Theorem 3.15

Proof. Let

$$T = \otimes_{i=i}^{k}T_i^{n_i} = \underbrace{T_1 \otimes \cdots \otimes T_1}_{n_1 \text{ times}} \otimes \underbrace{T_2 \otimes \cdots \otimes T_2}_{n_2 \text{ times}} \cdots \otimes \underbrace{T_k \otimes \cdots \otimes T_k}_{n_k \text{ times}}.$$

One has $T(e_n(\theta)) = T_1(e(\theta))^{n_1} \otimes T_2(e(\theta))^{n_2} \otimes \cdots \otimes T_k(e(\theta))^{n_k}$. By lemma 3.6,

$$\begin{aligned} V(T(e_n(\theta_1)), T(e_n(\theta_2)) &\leq \sum_{i=1}^{k} n_i V(T_i(e(\theta_1)), T_i(e(\theta_2))) \\ &\leq \left(\sum_{i=1}^{k} \alpha(T_i) n_i \right) V(e(\theta_1), e(\theta_2)). \end{aligned}$$

Now proceed as in the proof of theorem 3.14. □

3.17 Proof of Lemma 3.12

Proof.

$$\begin{aligned} \alpha(T_2 T_1) &= \sup_{P,Q \in \mathbb{P}(\mathcal{Z})} \frac{\|T_2 \circ T_1(P) - T_2 \circ T_1(Q)\|_1}{\|P - Q\|_1} \\ &= \sup_{P,Q \in \mathbb{P}(\mathcal{Z})} \frac{\|T_2 \circ T_1(P) - T_2 \circ T_1(Q)\|_1}{\|T_1(P) - T_2(Q)\|_1} \frac{\|T_1(P) - T_2(Q)\|_1}{\|P - Q\|_1} \\ &\leq \sup_{P,Q \in \mathbb{P}(\mathcal{Z})} \frac{\|T_2 \circ T_1(P) - T_2 \circ T_1(Q)\|_1}{\|T_1(P) - T_2(Q)\|_1} \sup_{P,Q \in \mathbb{P}(\mathcal{Z})} \frac{\|T_1(P) - T_2(Q)\|_1}{\|P - Q\|_1} \\ &\leq \sup_{P,Q \in \mathbb{P}(\tilde{\mathcal{Z}}_1)} \frac{\|T_2(P) - T_2(Q)\|_1}{\|P - Q\|_1} \sup_{P,Q \in \mathbb{P}(\mathcal{Z})} \frac{\|T_1(P) - T_2(Q)\|_1}{\|P - Q\|_1} \\ &= \alpha(T_2)\alpha(T_1) \end{aligned}$$

Where the first line follows from the definitions, the second follows if $T_1(P) \neq T_2(Q)$ and the rest are simple rearrangements. For the final inequality, remember that $\alpha(T) \leq 1$. □

3.18 Proof of Lemma 3.16

Proof. By definition $\|\tilde{\ell}\|_\infty = \sup_{z,a} |\tilde{\ell}(z,a)| = \sup_a \|\tilde{\ell}_a\|_\infty$. Hence,

$$\begin{aligned} \|\tilde{\ell}\|_\infty &= \sup_a \|\tilde{\ell}_a\|_\infty \\ &\leq \sup_a \|R^*\|_\infty \|\ell_a\|_\infty \\ &= \|R^*\|_\infty \|\ell\|_\infty, \end{aligned}$$

where the second line follows from the definition of the operator norm $\|R^*\|_\infty$. □

3.19 Proof of Lemma 3.17

Proof. Firstly $\|R\|_1 = \|R^*\|_\infty$ [22]. From the definition of $\|R\|_1$ we have,

$$
\begin{aligned}
\|R\|_1 &= \sup_{v \in \mathbb{R}^Y} \frac{\|Rv\|_1}{\|v\|_1} \\
&\geq \sup_{u \in \mathbb{R}^X} \frac{\|RTu\|_1}{\|Tu\|_1} \\
&= \sup_{u \in \mathbb{R}^X} \frac{\|u\|_1}{\|Tu\|_1} \\
&= 1 / \left(\inf_{u \in \mathbb{R}^X} \frac{\|Tu\|_1}{\|u\|_1} \right).
\end{aligned}
$$

This proves the first inequality. Recall one of the equivalent definitions of $\alpha(T)$ from section 3.4.4,

$$
\alpha(T) = \sup_{v \in \Omega} \frac{\|T(v)\|_1}{\|v\|_1},
$$

where $\Omega = \{v \in \mathbb{R}^X : \sum v_i = 0, v \neq 0\}$. Hence $\inf_{u \in \mathbb{R}^X} \frac{\|Tu\|_1}{\|u\|_1} \leq \alpha(T)$.

□

3.20 Proof of Theorem 3.19

Proof. As $\mathcal{L}$ is canonical, its partial loss function is given by $\mathcal{L}(-, v) = v + \Psi(v)\mathbf{1}_{\mathcal{Z}}$. By definition, the partial loss,

$$
\mathcal{L}_T(-, v) = R^*(\mathcal{L}(-, v)) = R^*(v) + \Psi(v)R^*(\mathbf{1}_{\mathcal{Z}}).
$$

If $|\mathcal{Z}| = |\tilde{\mathcal{Z}}|$, then T is reconstructible if and only if T is invertible. As T is column stochastic,

$$
\mathbf{1}_{\mathcal{Z}} = T^*(\mathbf{1}_{\tilde{\mathcal{Z}}}).
$$

This yields,

$$
R^*(\mathbf{1}_{\mathcal{Z}}) = R^*T^*(\mathbf{1}_{\tilde{\mathcal{Z}}}) = \mathbf{1}_{\tilde{\mathcal{Z}}}.
$$

For the more general case where $|\mathcal{Z}| < |\tilde{\mathcal{Z}}|$, we have for all T and all $v \in \mathbf{1}_{\mathcal{Z}}^{\perp}$,

$$
\begin{aligned}
\langle T(v), \mathbf{1}_{\tilde{\mathcal{Z}}} \rangle &= \langle v, T^*(\mathbf{1}_{\tilde{\mathcal{Z}}}) \rangle \\
&= \langle v, \mathbf{1}_{\mathcal{Z}} \rangle \\
&= 0.
\end{aligned}
$$

Therefore $T(\mathbf{1}_{\mathcal{Z}}^{\perp}) \subseteq \mathbf{1}_{\tilde{\mathcal{Z}}}^{\perp}$. As left inverses as not unique, we can further restrict R to those with $R(\mathbf{1}_{\tilde{\mathcal{Z}}}^{\perp}) \subseteq \mathbf{1}_{\mathcal{Z}}^{\perp}$, or dually those with $R^*(\mathbf{1}_{\mathcal{Z}}) = \mathbf{1}_{\tilde{\mathcal{Z}}}$. There is always such an R, as the restriction of T to $\mathbf{1}_{\mathcal{Z}}^{\perp}$ is also left invertible. Furthermore, $T(\mathbf{1}_{\mathcal{Z}}) \notin \mathbf{1}_{\tilde{\mathcal{Z}}}^{\perp}$, as $T(\mathbf{1}_{\mathcal{Z}})$ nonnegative entries. Therefore, we can take R restricted to $\mathbf{1}_{\tilde{\mathcal{Z}}}^{\perp}$ to be the left inverse of T restricted to $\mathbf{1}_{\mathcal{Z}}^{\perp}$, with $RT(\mathbf{1}_{\mathcal{Z}}) = \mathbf{1}_{\mathcal{Z}}$. Such an R can then be extended to all of $\mathbb{R}^{\tilde{\mathcal{Z}}}$. Finally, by definition, $\mathcal{L}_T(\tilde{z}, v) = \langle \delta_{\tilde{z}}, \mathcal{L}_T(-, v) \rangle$, yielding,

$$\begin{aligned}\mathcal{L}_T(\tilde{z}, v) &= \langle \delta_{\tilde{z}}, R^*(v) \rangle + \langle \delta_{\tilde{z}}, R^*(\mathbf{1}_{\mathcal{Z}}) \rangle \Psi(v) \\ &= \langle R(\delta_{\tilde{z}}), v \rangle + \Psi(v),\end{aligned}$$

where the last line is by properties of adjoints. This function is the sum of two functions, one linear in v the other convex and is therefore convex in v.

□

3.21 Corrupted Learning when Clean Learning is Fast

The contents of this chapter largely solve the problem of learning from corrupted data, when learning on the original problem occurs at the standard $\frac{1}{\sqrt{n}}$ rate. There are many conditions under which clean learning is fast, here we focus on the Bernstein condition presented in [17; 120].

Definition 3.22. *Let $P \in \mathbb{P}(\mathcal{Z})$, ℓ a loss and $a_P = \arg\min_a \mathbb{E}_{z\sim P}\ell(z,a)$. A pair (ℓ, P) satisfies the* Bernstein condition *with constant K if for all $a \in A$,*

$$\mathbb{E}_{z\sim P}(\ell(z,a) - \ell(z,a_P))^2 \leq K\, \mathbb{E}_{z\sim P}\ell(z,a) - \ell(z,a_P)$$

When A is finite, such a condition leads to $\frac{1}{n}$ rates of convergence. From theorem A.12 we have the following theorem.

Theorem 3.23. *(Fast Rates for ERM) Let $\mathcal{A}$ be ERM with A finite. If (ℓ, P) satisfies the Bernstein condition then for some constant $C > 0$,*

$$\mathbb{E}_{S\sim P^n}\ell(P, \mathcal{A}(S)) - \ell(P, a_P) \leq \frac{C\log(|A|)}{n}.$$

Furthermore with probability at least $1-\delta$ on a draw from P^n one has,

$$\ell(P, \mathcal{A}(S)) - \ell(P, a_P) \leq \frac{C\left(\log(|A|) + \log\left(\frac{1}{\delta}\right)\right)}{n}.$$

Proof. First, define $\ell_P(z,a) = \ell(z,a) - \ell(z,a_P)$. ℓ_P measures the loss relative to the best action for the distribution P. It is easy to verify that for bounded ℓ, $\|\ell_P\|_\infty \leq 2\|\ell\|_\infty$. We now utilize theorem A.12 with ℓ_P and π uniform on A. This yields,

$$\mathbb{E}_{S\sim P^n}\left[\ell_P(P, \mathcal{A}(S)) - \gamma\mathbb{E}_{z\sim P}\ell_P^2(z, \mathcal{A}(S))\right] \leq \frac{1}{n}\mathbb{E}_{S\sim P^n}\left[\ell_P(S, \mathcal{A}(S)) + \|\ell_P,\|_\infty\left(\frac{\log(|A|)}{\beta}\right)\right]$$

with $\gamma = \frac{(e^\beta - 1 - \beta)}{\beta\|\ell_P\|_\infty}$. Firstly ERM minimizes the right hand side of the bound meaning,

$$\frac{1}{n}\mathbb{E}_{S\sim P^n}\left[\ell_P(S, \mathcal{A}(S)) + \|\ell_P\|_\infty\left(\frac{\log(|A|)}{\beta}\right)\right] \leq \frac{1}{n}\left[\|\ell_P\|_\infty\left(\frac{\log(|A|)}{\beta}\right)\right].$$

To see this consider the algorithm that always outputs a_P, this algorithm generalizes very well however it may be suboptimal on the sample. Secondly (ℓ, P) satisfies the Bernstein condition with constant K. Therefore,

$$(1-\gamma K)\mathbb{E}_{S\sim P^n}\ell_P(P,\mathcal{A}(S)) \leq \frac{1}{n}\left[\|\ell_P\|_\infty\left(\frac{\log(|A|)}{\beta}\right)\right].$$

Finally chose β small enough so that $\gamma K \leq 1$. This can always be done as $\gamma \to 0$ as $\beta \to 0_+$. The high probability version proceeds in a similar way.

□

A natural question to ask is when does $(\ell_T, \tilde{P})$ satisfy the Bernstein condition?

Theorem 3.24. *If $(\ell_T, \tilde{P})$ satisfies the Bernstein condition with constant K then (ℓ, P) also satisfies the Bernstein condition with the same constant.*

Proof.

$$\begin{aligned}
K\mathbb{E}_{z\sim P}\ell(z,a) - \ell(z,a_P) &= K\mathbb{E}_{\tilde{z}\sim\tilde{P}}\ell_T(z,a) - \ell_T(z,a_P)\\
&\geq \mathbb{E}_{\tilde{z}\sim\tilde{P}}(\ell_T(\tilde{z},a) - \ell_T(\tilde{z},a_P))^2\\
&= \mathbb{E}_{z\sim P}\mathbb{E}_{\tilde{z}\sim T(z)}(\ell_T(\tilde{z},a) - \ell_T(\tilde{z},a_P))^2\\
&\geq \mathbb{E}_{z\sim P}(\mathbb{E}_{\tilde{z}\sim T(z)}\ell_T(\tilde{z},a) - \mathbb{E}_{\tilde{z}\sim T(z)}\ell_T(\tilde{z},a_P))^2\\
&= \mathbb{E}_{z\sim P}(\ell(z,a) - \ell(z,a_P))^2,
\end{aligned}$$

where the first line follows from the definition of ℓ and because $a_P = a_{\tilde{P}}$, the second as $(\ell_T, \tilde{P})$ satisfies the Bernstein condition and finally we have used the convexity of $f(x) = x^2$.

□

This theorem (almost) rules out pathological behaviour where ERM learns quickly from corrupted data and yet slowly for clean data. The converse of theorem 3.24 is not true, for example consider the case of PAC learning versus PAC learning with arbitrary instance dependent noise. In some cases the Bernstein condition can be transfered from the clean problem to the corrupted problem, as we now explore.

Definition 3.25. *Let $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and ℓ a loss. A pair (ℓ, T) are η-*compatible *if for all $z \in \mathcal{Z}$ and $a_1, a_2 \in A$,*

$$\mathbb{E}_{\tilde{z}\sim T(z)}(\ell_T(\tilde{z},a_1) - \ell_T(\tilde{z},a_2))^2 \leq \eta(\ell(z,a_1) - \ell(z,a_2))^2.$$

Theorem 3.26. *If the pair (ℓ, P) satisfies the Bernstein condition with constant K and the pair (ℓ, T) are η-compatible then $(\ell_T, \tilde{P})$ satisfies the Bernstein condition with constant ηK.*

Proof.

$$\begin{aligned}\mathbb{E}_{\tilde{z}\sim\tilde{P}}(\ell_T(\tilde{z},a)-\ell_T(\tilde{z},a_P))^2 &= \mathbb{E}_{z\sim P}\mathbb{E}_{\tilde{z}\sim T(z)}(\ell_T(\tilde{z},a)-\ell_T(\tilde{z},a_P))^2\\ &\leq \eta\mathbb{E}_{z\sim P}(\ell(z,a)-\ell(z,a_P))^2\\ &\leq \eta K\mathbb{E}_{z\sim P}\ell(z,a)-\ell(z,a_P)\\ &= \eta K\mathbb{E}_{\tilde{z}\sim\tilde{P}}\ell_T(\tilde{z},a)-\ell_T(\tilde{z},a_P),\end{aligned}$$

where we have first used η-compatibility, then the fact that (ℓ, P) satisfies the Bernstein condition with constant K and finally the definition of ℓ_T.

□

While by no means the final line in fast corrupted learning, this theorem does allow one to prove interesting results in the binary classification setting.

Theorem 3.27. *Let T be label noise,* $T = \begin{pmatrix} 1-\sigma_{-1} & \sigma_1 \\ \sigma_{-1} & 1-\sigma_1 \end{pmatrix}$*, then the pair* (ℓ_{01}, T) *is* η*-compatible with* $\eta = \max(\left(\frac{1+\sigma_{-1}-\sigma_1}{1-\sigma_{-1}-\sigma_1}\right)^2, \left(\frac{1+\sigma_1-\sigma_{-1}}{1-\sigma_{-1}-\sigma_1}\right)^2)$.

Proof. Due to the symmetry of the left and right hand sides of the Bernstein condition, one only needs to check the case where $a_1 = 1$, $a_2 = -1$. Recall,

$$\begin{aligned}\ell_{01,T}(\tilde{y},a) &= \frac{(1-\sigma_{-y})\ell_{01}(\tilde{y},a)-\sigma_y\ell_{01}(-\tilde{y},a)}{1-\sigma_{-1}-\sigma_1}\\ &= \frac{(1-\sigma_{-y}+\sigma_y)\ell_{01}(\tilde{y},a)-\sigma_y}{1-\sigma_{-1}-\sigma_1}.\end{aligned}$$

For $y = 1$ it is easy to confirm $(\ell_{01}(1,1)-\ell_{01}(1,-1))^2 = 1$. We have,

$$\begin{aligned}\ell_{01,T}(\tilde{y},1)-\ell_{01,T}(\tilde{y},-1) &= \frac{(1-\sigma_{-y}+\sigma_y)(\ell_{01}(\tilde{y},1)-\ell_{01}(\tilde{y},-1))}{1-\sigma_{-1}-\sigma_1}\\ &= \frac{-\tilde{y}(1-\sigma_{-y}+\sigma_y)}{1-\sigma_{-1}-\sigma_1}.\end{aligned}$$

Squaring, taking maximums and finally expectations yields the desired result.

□

One very useful example of a pair (P, ℓ) satisfying the Bernstein condition with constant 1 is when P is separable, ℓ is 01 loss and the Bayes optimal classifier is in the function class. Theorem 3.27 guarantees that as long as $\sigma_{-1} + \sigma_1 \neq 0$ (i.e. it is possible to learn from noisy labels), one learns at a fast rate from noisy examples.

3.22 Corruption Invariant Loss Functions

We focus here on the problem of supervised learning. Given a label space Y, an instance space X, a distribution $P \in \mathbb{P}(X \times Y)$ and a loss $\ell : Y \times A \to \mathbb{R}$, the goal of the decision maker is to find a function $f \in \mathcal{F} \subset A^X$ with low expected loss,

$$\mathbb{E}_{(x,y)\sim P}\ell(y, f(x)).$$

By marginalising out over the instances, we can think of P and f as defining a distribution $Q \in \mathbb{P}(A \times Y)$. The loss ℓ allows the decision maker to *order* distributions, we say $Q_1 \leq_\ell Q_2$ if,

$$\mathbb{E}_{(a,y)\sim Q_1}\ell(y, a) \leq \mathbb{E}_{(a,y)\sim Q_2}\ell(y, a).$$

Abstractly, the problem of supervised learning reduces to finding the minimal Q in this ordering, where Q is specified by P and $f \in \mathcal{F}$. Given a corruption $T \in \mathbb{T}(Y, \tilde{Y})$, rather than the clean distribution Q, the decision maker works with a corrupted distribution $(a, \tilde{y}) \sim T(Q)$. To sample from $T(Q)$, first sample $(a, y) \sim Q$, and then sample $\tilde{y} \sim T(y)$. While the decision maker wants to compare clean Q, they can do so only by comparing $T(Q)$. If T is known, they can correct for the corruption by using the loss ℓ_T.

However, if all the decision maker knows is that $T \in \mathcal{C} \subset \mathbb{T}(Y, \tilde{Y})$, then assuming different corruptions may lead to different orderings. For $T, T' \in \mathcal{C}$, there is no guarantee that,

$$T(Q_1) \leq_{\ell_T} T(Q_2) \Leftrightarrow T(Q_1) \leq_{\ell_{T'}} T(Q_2).$$

In words, assuming the *wrong* corruption may lead to the *wrong* ordering. Corruption immune losses are precisely those where the ordering is *unaltered*.

Definition 3.28 (Order Equivalence)**.** *Let $\ell, \ell' : \tilde{Y} \times A \to \mathbb{R}$ be loss functions. ℓ is* order equivalent *to ℓ' if for all $\tilde{Q}_1, \tilde{Q}_2 \in \mathbb{P}(A \times \tilde{Y})$,*

$$\tilde{Q}_1 \leq_\ell \tilde{Q}_2 \Leftrightarrow \tilde{Q}_1 \leq_{\ell'} \tilde{Q}_2.$$

Definition 3.29. *Let $\mathcal{C} \subset \mathbb{T}(\mathcal{Z}, \mathcal{Z})$ be a set of reconstructible transitions. A loss ℓ is* immune to $\mathcal{C}$ *if for all $T, T' \in \mathcal{C}$, ℓ_T is order equivalent to $\ell_{T'}$.*

Lemma 3.30. *If ℓ is immune to $\mathcal{C}$, then for all $T, T' \in \mathcal{C}$ and for all $Q_1, Q_2 \in \mathbb{P}(A \times Y)$,*

$$Q_1 \leq_\ell Q_2 \Leftrightarrow T'(Q_1) \leq_{\ell_T} T'(Q_2).$$

Proof. Firstly, $Q_1 \leq_\ell Q_2 \Leftrightarrow T'(Q_1) \leq_{\ell'_T} T'(Q_2)$ from the definition of corruption correction. As ℓ_T is order equivalent to $\ell_{T'}$, $T'(Q_1) \leq_{\ell'_T} T'(Q_2) \Leftrightarrow T'(Q_1) \leq_{\ell_T} T'(Q_2)$. □

The converse of this lemma is also true in certain situations. For example if $Y = \tilde{Y}$ and $\mathrm{id}_Y \in \mathcal{C}$. Chapter 4 contains an example of a corruption immune loss. Order equivalence of loss functions is characterized by the following lemma.

Lemma 3.31. *ℓ is order equivalent to ℓ' if and only if there exists a constants $\alpha > 0$ and β such that,*

$$\ell(\tilde{y}, a) = \alpha\ell'(\tilde{y}, a) + \beta,\ \forall \tilde{y} \in \tilde{Y},\ \forall a \in A.$$

This lemma is theorem 2 in section 7.9 of [54]. This lemma allows one to test whether a loss is immune to $\mathcal{C}$.

Theorem 3.32. *Fix $T \in \mathcal{C}$. A loss ℓ is immune to $\mathcal{C}$ if and only if for all $T' \in \mathcal{C}$ and for all $a \in A$, the equation,*

$$(R^* - \alpha(R')^*)\ell(-, a) = \beta \mathbf{1}_{\tilde{Y}},$$

has a solution in α and β, with $\alpha > 0$.

Proof. By definition, ℓ_T must be order equivalent to $\ell_{T'}$ for all $T, T' \in \mathcal{C}$. As order equivalence is a transitive relationship, we can fix T. The rest of the theorem follows from lemma 3.31.

□

4

An Average Classification Algorithm

"While the truth is rarely pure, it can be simple."

- Oscar Wilde via Robert C. Williamson, *The Importance of Being Unhinged [121]*

In the problem of binary classification, the goal is to learn a classifier that accurately predicts an instance's corresponding label. Many cutting edge classification algorithms, such as the support vector machine, logistic regression, boosting (for a particular choice of weak learners) and so on, output a classifier of the form,

$$f(x) = \operatorname{sign}(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i)), \tag{4.1}$$

with $\alpha_i \geq 0$, $\sum \alpha_i = 1$ and $K(x, x')$ a function that measures the similarity of two instances x and x'. Algorithmically, optimizing these weights is a difficult problem that still attracts much research effort. Furthermore, *explaining* these methods to the uninitiated is a difficult task. Letting all the α_i be equal in 4.1 leads to a conceptually simpler classification rule, one that requires little effort to motivate or explain: the mean,

$$f(x) = \operatorname{sign}(\frac{1}{n}\sum_{i=i}^{n} y_i K(x_i, x)).$$

The above is a simple and intuitive classification rule. It classifies by the average similarity to the previously observed positive and negative instances, with the most similar class being the output of the classifier. It has been studied previously, for example in chapter one of [107] and further in [12; 56; 80; 108]. The main drawbacks of the mean classification rule are prohibitive storage and evaluation costs. In fact, this is the motivation given for the support vector machine (SVM) in [107]. Our goal here is to reinvigorate interest in this very average algorithm.

The chapter proceeds as follows:

- We argue for the mean classifier, showing it is the ERM solution for a classification calibrated loss function [16] (theorems 4.1 and 4.2).

- We explore the robustness properties of the mean classifier. We relate the noise tolerance of the mean algorithm to the *margin for error* in the solution (theorem 4.6). Finally we show, in a certain sense, the mean classifier is the *only* surrogate loss minimization method that is *immune* to the effects of symmetric label noise (theorem 4.14).

- Finally, we show how to *sparsely* approximate *any* kernel classifier through the use of kernel herding [8; 38; 128]. This produces a *simple*, understandable means of choosing representative points, with provable rates of convergence.

The result is a conceptually simple algorithm for learning classifiers, that is accurate, easily parallelized, robust and firmly grounded in theory.

4.1 Basic Notation

Denote by $\mathcal{H}$ an abstract Hilbert space, with inner product $\langle v_1, v_2 \rangle$. For any $v \in \mathcal{H}$, denote by $\hat{v}$ the unit vector in the direction of v. We work with loss functions $\ell : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}$.

4.2 Kernel Classifiers

Let X be the instance space and $Y = \{-1, 1\}$ the label space. A *classifier* is a bounded function $f \in \mathbb{R}^X$, with $f(x)$ the *score* assigned to the instance X and $\text{sign}(f(x))$ the predicted label. For a distribution $P \in \mathbb{P}(X \times Y)$, we define the *Bayes optimal* classifier to be the classifier $f_P(x) = 1$ if $P(Y = 1 | X = x) \geq \frac{1}{2}$ and -1 otherwise. We measure the distance between classifiers via the *supremum distance*,

$$\|f - f'\|_\infty = \sup_{x \in X} |f(x) - f'(x)|.$$

A *classification algorithm* is a function $\mathcal{A} : \cup_{n=1}^{\infty} (X \times Y)^n \to \mathbb{R}^X$, that given a training set S outputs a classifier. Define the misclassification loss $\ell_{01}(y, v) = [\![yv < 0]\!]$. Note, that $\ell_{01}(y, 0) = 1$ always. This non-standard form of misclassification loss will enhance the readability of many of the proofs. An output of zero can be viewed as *abstaining* from choosing a label. For any loss $\ell : Y \times \mathbb{R} \to \mathbb{R}$, we remind the reader of the *risk* and *sample risk* of f defined as,

$$\ell(P, f) := \mathbb{E}_{(x,y) \sim P} \ell(y, f(x)) \text{ and } \ell(S, f) := \frac{1}{|S|} \sum_{(x,y) \in S} \ell(y, f(x)),$$

respectively. Good classification algorithms should output classifiers with low misclassification risk. Many classification algorithms, such as the SVM, logistic regression, boosting (for a particular choice of weak learners) and so on, output a classifier of the form,

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i),$$

with $\alpha_i \geq 0$, $\sum \alpha_i = 1$ and $K(x, x') = \langle \phi(x), \phi(x') \rangle$ a *kernel* function, an inner product of feature vectors in a (possibly infinite dimensional) feature space. For simplicity, we use the mean,

$$f(x) = \frac{1}{n} \sum_{i=i}^{n} y_i K(x_i, x). \tag{4.2}$$

4.3 Why the Mean?

The mean is not only an intuitively appealing classification rule, it also arises as the optimal classifier for the linear loss, considered previously in [105; 113]. Let,

$$\ell_{\text{linear}}(y, v) = 1 - yv.$$

If $v \in \{-1, 1\}$, then $\ell_{01}(y, v) = \frac{1}{2} \ell_{\text{linear}}(y, v)$. Allowing $v \in [-1, 1]$ provides a *convexification* of misclassification loss. For $v \in [-1, 1]$, $\ell_{01}(y, v) \leq \ell_{\text{linear}}(y, v)$. Furthermore, we have the following surrogate regret bound.

Theorem 4.1 (Surrogate Regret Bound for Linear Loss)**.** *For all distributions P,*

$$f_P = \operatorname*{arg\,min}_{f \in [-1,1]^X} \ell_{\text{linear}}(P, f) \in \operatorname*{arg\,min}_{f \in [-1,1]^X} \ell_{01}(P, f).$$

Furthermore for all $f \in [-1, 1]^X$,

$$\ell_{01}(P, f) - \ell_{01}(P, f_P) \leq \ell_{\text{linear}}(P, f) - \ell_{\text{linear}}(P, f_P).$$

By theorem 4.1, linear loss is a suitable surrogate loss for learning classifiers much like the hinge, logistic and exponential loss functions [16]. As is usual, rather than minimizing over all bounded functions, to avoid overfitting the sample we work with a restricted function class. For a feature map $\phi : X \to \mathcal{H}$, define the linear function class,

$$\mathcal{F}_\phi := \{ f_\omega(x) = \langle \omega, \phi(x) \rangle : \omega \in \mathcal{H} \},$$

and the bounded linear function class,

$$\mathcal{F}_\phi^r := \{ f_\omega(x) = \langle \omega, \phi(x) \rangle : \omega \in \mathcal{H}, \|\omega\| \leq r \}.$$

We will assume throughout that the feature map is *bounded*, $\|\phi(x)\| \leq 1$ for all x. As shorthand we write $\ell(P, \omega) := \ell(P, f_\omega)$. By the Cauchy-Schwarz inequality $\mathcal{F}_\phi^r \subseteq [-r, r]^X$. As a surrogate to minimizing $\ell_{01}(P, f)$ over all functions, we will minimize $\ell_{\text{linear}}(S, f)$ over $f \in \mathcal{F}_\phi^1$.

For any sample $S \in \cup_{n=1}^{\infty}(X \times Y)^n$ define the *mean vector* $\omega_S = \frac{1}{|S|} \sum\limits_{(x,y)\in S} y\phi(x)$. The mean classifier (4.2) can be written as $f(x) = \langle \omega_S, \phi(x) \rangle$.

Theorem 4.2 (The Mean Classifier Minimizes Linear Loss)**.**

$$\hat{\omega}_S = \operatorname*{arg\,min}_{\omega:\|\omega\|\leq 1} \frac{1}{|S|} \sum_{(x,y)\in S} 1 - y\langle \omega, \phi(x) \rangle = \operatorname*{arg\,min}_{\omega:\|\omega\|\leq 1} 1 - \langle \omega, \omega_S \rangle$$

with minimum linear loss given by $1 - \|\omega_S\|$. *Furthermore classifying using* $\langle \hat{\omega}_S, \phi(x) \rangle$ *is equivalent to classifying according to equation (4.2).*

This has been noted in [113], we include it for completeness. The proof is a straight forward application of the Cauchy-Schwarz inequality. As $\hat{\omega}_S = \lambda \omega_S$, $\lambda > 0$, they both produce the same classifier. Changing the norm constraint to $\|\omega\| \leq r$ merely scales the classifier, and therefore does not change its misclassification performance. The quantity,

$$\|\omega_S\|^2 = \frac{1}{|S|^2} \sum_{(x,y)\in S} \sum_{(x',y')\in S} yy'K(x,x'),$$

can be thought of as the "self-similarity" of the sample. For a distribution P, define $\omega_P = \mathbb{E}_{(x,y)\sim P} y\phi(x)$. It is easily verified that,

$$\hat{\omega}_P = \operatorname*{arg\,min}_{\omega:\|\omega\|\leq 1} \mathbb{E}_{(x,y)\sim P} 1 - y\langle \omega, \phi(x) \rangle = \operatorname*{arg\,min}_{\omega:\|\omega\|\leq 1} 1 - \langle \omega, \omega_P \rangle.$$

Furthermore, we have the following generalization bound on the linear loss performance of ω_S.

Theorem 4.3. *For all distributions P and for all bounded feature maps* $\phi : X \to \mathcal{H}$, *with probability at least* $1 - \delta$ *on a draw* $S \sim P^n$,

$$\ell_{\text{linear}}(P, \omega_S) \leq \ell_{\text{linear}}(S, \omega_S) + \sqrt{\frac{2\left(1 + \log(\frac{1}{\delta})\right)}{n}}.$$

This theorem is a special case of a more general result, proved in the appendix to this chapter. In Smola et al. [112], bounds for the error in estimating the mean are presented.

Theorem 4.4. *For all distributions P and for all bounded feature maps* $\phi : X \to \mathcal{H}$, *with probability at least* $1 - \delta$ *on a draw* $S \sim P^n$,

$$\|\omega_P - \omega_S\| \leq \frac{2}{\sqrt{n}} + \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}.$$

The proof is via an appeal to standard Rademacher bounds.

4.3.1 Relation to Maximum Mean Discrepancy

Let $P_\pm \in \mathbb{P}(X)$ be the conditional distribution over instances given a positive or negative label respectively. Define the *maximum mean discrepancy* [66],

$$\mathrm{MMD}_\phi(P_+, P_-) := \max_{\omega : \|\omega\| \leq 1} \frac{1}{2} |\mathbb{E}_{x \sim P_+} \langle \omega, \phi(x) \rangle - \mathbb{E}_{x \sim P_-} \langle \omega, \phi(x) \rangle| = \frac{1}{2} \|\omega_{P_+} - \omega_{P_-}\| .$$

$\mathrm{MMD}_\phi(P_+, P_-)$ can be seen as a restricted variational divergence 2.29,

$$V(P_+, P_-) = \max_{f \in [-1,1]^X} \frac{1}{2} |\mathbb{E}_{x \sim P_+} f(x) - \mathbb{E}_{x \sim P_-} f(x)|,$$

a commonly used metric on probability distributions, where $f \in \mathcal{F}^1_\phi \subseteq [-1,1]^X$. Define the distribution $P \in \mathbb{P}(X \times Y)$ that first samples y uniformly from $\{-1, 1\}$ and then samples $x {\sim} P_y$. Then,

$$\mathrm{MMD}_\phi(P_+, P_-) = \max_{\omega : \|\omega\| \leq 1} |\mathbb{E}_{(x,y) \sim P} \langle \omega, y\phi(x) \rangle| = \|\omega_P\| .$$

Therefore, if we assume that positive and negative classes are equally likely, the mean classifier classifies using the ω that "witnesses" the MMD, i.e. it attains the max in the above.

4.3.2 Relation to the SVM

For a regularization parameter λ, the SVM solves the following convex objective,

$$\operatorname*{arg\,min}_{\omega \in \mathcal{H}} \frac{1}{|S|} \sum_{(x,y) \in S} [1 - y\langle \omega, \phi(x) \rangle]_+ + \frac{\lambda}{2} \|\omega\|^2 ,$$

where $[x]_+ = \max(x, 0)$. This is the Lagrange multiplier problem associated with,

$$\operatorname*{arg\,min}_{\omega : \|\omega\|^2 \leq c} \frac{1}{|S|} \sum_{(x,y) \in S} [1 - y\langle \omega, \phi(x) \rangle]_+ .$$

If we take $c = 1$, by Cauchy-Schwarz $[1 - y\langle \omega, \phi(x) \rangle]_+ = 1 - y\langle \omega, \phi(x) \rangle$ and the above objective is equivalent to that in theorem 4.2 . The mean classifier is the optimal solution to a highly regularized SVM, and is therefore *preferentially* optimizing the margin *over* the sample hinge loss. Prior evidence exists showing that feature normalisation (which is high regularization in disguise) increases the generalisation performance of SVM's [65].

4.3.3 Relation to Kernel Density Estimation

On the surface the mean classifier is a *discriminative* approach. Restricting to *positive* kernels, such as the Gaussian kernel, it can be seen as the following *generative* approach: estimate P with $\tilde{P}$, with class conditional distributions estimated by kernel density estimation. Letting $S_{\pm} = \{(x, \pm 1)\} \subseteq S$, take,

$$\tilde{P}(X = x | Y = \pm 1) \propto \frac{1}{|S_{\pm}|} \sum_{x' \in S_{\pm}} K(x, x')$$

and $\tilde{P}(Y = 1) = \frac{|S_+|}{|S|}$. To classify new instances, use the Bayes optimal classifier for $\tilde{P}$. This yields the same classification rule as (4.2). This is the "potential function rule" discussed in [56].

4.3.4 Extension to Multiple Kernels

To ensure the practical success of any kernel based method, it is important that the *correct* feature map be chosen. Thus far we have only considered the problem of learning with a single feature map, and not the problem of *learning the feature map*. Given k feature maps $\phi_i : X \to \mathcal{H}_i$, $i \in [1; k]$, multiple kernel learning [9; 43; 76; 84] considers learning over a function class that is the convex hull of the classes $\mathcal{F}^1_{\phi_i}$,

$$\mathcal{F} := \left\{ f(x) = \sum_{i=1}^{k} \alpha_i \langle \omega^i, \phi_i(x) \rangle : \left\| \omega^i \right\| \leq 1, \alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1 \right\}.$$

By an easy calculation,

$$\min_{f \in \mathcal{F}} \frac{1}{|S|} \sum_{(x,y) \in S} 1 - yf(x) = \min_{i \in [1;k]} (1 - \left\| \omega^i_S \right\|),$$

where ω^i_S is the sample mean in the i-th feature space. In words, we pick the feature space which minimizes $1 - \left\| \omega^i_S \right\|$. This is in contrast to usual multiple kernel learning techniques that do not in general pick out a *single* feature map. Furthermore, we have the following generalization bound.

Theorem 4.5. *For all distributions P and for all finite collections of bounded feature maps $\phi_i : X \to \mathcal{H}_i$, $i \in [1; k]$, with probability at least $1 - \delta$ on a draw $S \sim P^n$,*

$$\ell_{\text{linear}}(P, \omega^*_S) \leq \ell_{\text{linear}}(S, \omega^*_S) + \sqrt{\frac{2\left(1 + \log(k) + \log\left(\frac{1}{\delta}\right)\right)}{n}},$$

where ω^*_S corresponds to the mean that minimizes $1 - \left\| \omega^i_S \right\|$.

Like theorem 4.3, this is a specific case of a more refined bound presented in the appendix to this chapter.

4.4 The Robustness of the Mean Classifier

Here we detail the robustness of the mean classifier to perturbations of P. We do not consider the statistical issues of learning from a corrupted distribution. For detailed treatment of such problems the reader is directed to chapter 3 . We first show that the degree to which one can approximate a classifier without loss of performance is related to the *margin for error* of the classifier. We then discuss robustness properties of the mean classifier under the σ-contamination model of Huber [75]. Finally we show the immunity of the mean classifier to symmetric label noise.

The results of this section only pertain to *linear* function classes. In the following section we consider general function classes. We show that in this more general setting, linear loss is the *only* loss function that is robust to the effects of symmetric label noise.

4.4.1 Approximation Error and Margins

Define the *margin loss* at *margin* γ to be $\ell_\gamma(y,v) = [\![yv < \gamma]\!]$. The margin loss is an upper bound of misclassification loss. For $\gamma = 0$, $\ell_\gamma = \ell_{01}$. The margin loss is used in place of misclassification loss to produce tighter generalization bounds for minimizing misclassification loss [15; 109]. For a classifier f to have small margin loss it must not just accurately predict the label, it must do so with confidence. Maximizing the margin while forcing $\ell_\gamma(S,\omega) = 0$ is the original motivation for the hard margin SVM [44]. Here we relate the margin loss of a classifier f to the amount of slop allowed in approximating f.

Theorem 4.6 (Margins and Approximation). *$\ell_\epsilon(P,f) \leq \alpha$ if and only if $\ell_{01}(P,\tilde{f}) \leq \alpha$ for all $\tilde{f}$ with $\|f-\tilde{f}\|_\infty \leq \epsilon$.*

The *margin for error* on a distribution P of a classifier f is given by,

$$\Gamma(P,f) := \sup\{\gamma : \ell_\gamma(P,f) = \ell_{01}(P,f)\}.$$

For a sample S, setting $\epsilon < \Gamma(S,f)$ ensures,

$$\ell_{01}(S,\omega_S) = \ell_\epsilon(S,\omega_S) = \ell_{01}(S,\tilde{\omega}_S).$$

The margin therefore provides means of assessing the degree to which one can approximate a classifier; the larger the margin the greater error allowed in approximating the classifier.

4.4.2 Robustness under σ-contamination

Rather than samples from P, we assume the decision maker has access to samples from a perturbed distribution,

$$\tilde{P} = (1-\sigma)P + \sigma Q,$$

with Q the perturbation or corruption. We can view sampling from $\tilde{P}$ as sampling from P with probability $1-\sigma$ and from Q with probability σ. It is easy to show that $\omega_{\tilde{P}} = (1-\sigma)\omega_P + \sigma\omega_Q$. Furthermore,

$$\|\omega_P - \omega_{\tilde{P}}\| = \sigma \|\omega_P - \omega_Q\| .$$

Lemma 4.7. *If $\sigma \|\omega_P - \omega_Q\| < \Gamma(P, \omega_P)$ then $\ell_{01}(P, \omega_P) = \ell_{01}(P, \omega_{\tilde{P}})$.*

Hence the margin provides means to assess the immunity of the mean classifier to corruption. Furthermore, as $\|\omega_P - \omega_Q\| \leq 2$, if $\sigma < \frac{\Gamma(P,\omega_P)}{2}$ then the mean classifier is immune to the effects of *any* Q. We caution the reader that lemma 4.7 is a one way implication. For particular choices of Q, one can show greater robustness of the mean classifier.

4.4.3 Learning Under Symmetric Label Noise

Here we consider the problem of learning under symmetric label noise [4]. Rather than samples from P, the decision maker has access to samples from a corrupted distribution P_σ. To sample from P_σ, first draw $(x, y) \sim P$ and then flip the label with probability σ. Learning from P_σ can be understood as a corrupted learning problem of the sort studied in chapter 3. There the statistical effects of the corruption are quantified, in summary n corrupted samples are equivalent to $(1-2\sigma)n$ uncorrupted samples for the purpose of learning a classifier. Here we focus on robustness. This problem is of practical interest, particularly in situations where there are multiple labellers, each of which can be viewed as an "expert" labeller with added noises. We can decompose,

$$P_\sigma = (1-\sigma)P + \sigma P',$$

where P' is the "label flipped" version of P. It is easy to show $\omega_{P'} = -\omega_P$. Therefore $\omega_{P_\sigma} = (1-2\sigma)\omega_P$.

Theorem 4.8 (Symmetric Label Noise Immunity of the Mean Classifier)**.** *Let P_σ be P corrupted via symmetric label noise with label flip probability σ. Then for all $\sigma \in (0, \frac{1}{2})$, $\ell_{01}(P, \omega_P) = \ell_{01}(P, \omega_{P_\sigma})$.*

The proof comes from the simple observation that as ω_P and ω_{P_σ} are related by a positive constant, they produce the same classifier. This result extends previous results in [80; 108] on the symmetric label noise immunity of the mean classification algorithm, were it is assumed the marginal distribution over instances is uniform on the unit sphere in $\mathbb{R}^n$.

4.4.4 Other Approaches to Learning with Symmetric Label Noise

A large class of modern classification algorithms, such as logistic regression, the SVM and boosting, proceed by minimizing a *convex potential* or *margin loss* over a particular function class.

Definition 4.9. *A loss ℓ is a* convex potential *if their exists a convex function $\psi : \mathbb{R} \to \mathbb{R}$ with $\psi(v) \geq 0$, $\psi'(0) < 0$ and $\lim_{v \to \infty} \psi(v) = 0$, with,*

$$\ell(y, v) = \psi(yv).$$

The condition that $\psi'(0) < 0$ ensures that all convex potential losses are classification calibrated [16] (in fact this condition characterizes classification calibrated losses). Long and Servedio in [89] proved the following negative result on what is possible when learning under symmetric label noise: there exists a separable distribution P and function class $\mathcal{F}$ where, when the decision maker observes samples from P_σ with symmetric label noise of *any nonzero rate*, minimisation of *any convex potential* over $\mathcal{F}$ results in classification performance on P that is equivalent to random guessing. The example provided in [89] is far from esoteric, in fact it is a given by a distribution in $\mathbb{R}^2$ that is concentrated on three points with function class given by linear hyperplanes through the origin. We present their example in section 4.9.

Ostensibly, this result establishes that convex losses are not robust to symmetric label noise, and motivates using non-convex losses [55; 57; 94; 96; 114]. These approaches are computationally intensive and scale poorly to large data sets. We have seen in the previous that linear loss, with function class $\mathcal{F}_\phi$ (for any feature map ϕ), is immune to symmetric label noise. Furthermore, minimizing linear loss is easy. We show in the following section that linear loss minimization over *any* function class is immune to symmetric label noise.

An alternate means of circumventing the impossibility result in [89] is to use a rich function class, say by using a universal kernel, together with a standard classification calibrated loss. As the form of the Bayes optimal classifier is the same for both noisy and clean data, one can appeal to universality results such as those in [91]. While this approach is immune to label noise, performing the minimization is difficult. By theorem 4.1, for sufficiently rich function classes, using any of these other losses will produce the same result as using linear loss.

Finally, if the noise rate is known, one can use the method of unbiased estimators presented by Natarajan et al. [101] and correct for the corruption. The obvious drawback is in general, the noise rate is unknown. In the following section we explore the relationship between linear loss and the method of unbiased estimators. We show that linear loss is "unaffected" by this correction (in a sense to be made precise). Furthermore, linear loss is essentially the *only* convex loss with this property.

4.5 Symmetric Label Noise and Corruption Corrected Losses

The weakness of the analysis of section 4.4.3, was that it only considered *linear* function classes. Here we show that linear loss minimization over *general* function classes is unaffected by symmetric label noise, in the sense that for all $\sigma \in (0, \frac{1}{2})$ and for all function classes $\mathcal{F} \subseteq \mathbb{R}^X$,

$$\operatorname*{arg\,min}_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim P}\ell_{\text{linear}}(y, f(x)) = \operatorname*{arg\,min}_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim P_\sigma}\ell_{\text{linear}}(y, f(x)).$$

For the following section we work *directly* with distributions $Q \in \mathbb{P}(\mathbb{R} \times Y)$ over score, label pairs. Any distribution P and classifier f induces a distribution $Q(P, f)$ with,

$$\mathbb{E}_{(v,y)\sim Q(P,f)}\ell(y, v) = \mathbb{E}_{(x,y)\sim P}\ell(y, f(x)).$$

A loss ℓ provides means to *order* distributions. For two distributions Q, Q', we say $Q \leq_\ell Q'$ if,

$$\mathbb{E}_{(v,y)\sim Q}\ell(y, v) \leq \mathbb{E}_{(v,y)\sim Q'}\ell(y, v).$$

If $Q = Q(P, f_1)$ and $Q' = Q(P, f_2)$, the above is equivalent to,

$$\mathbb{E}_{(x,y)\sim P}\ell(y, f_1(x)) \leq \mathbb{E}_{(x,y)\sim P}\ell(y, f_2(x)),$$

i.e., the classifier f_1 has lower expected loss than f_2. The decision maker wants to find the distribution Q, in some restricted set, that is smallest in the ordering $\leq_\ell$. Denote by Q_σ, the distribution obtained from drawing pairs $(v, y)\sim Q$ and then flipping the label with probability σ. In light of Long and Servedio's example, there is no guarantee that,

$$Q \leq_\ell Q' \Leftrightarrow Q_\sigma \leq_\ell Q'_\sigma.$$

The noise might affect how distributions are ordered. To progress we seek loss functions that are *robust* to label noise.

Definition 4.10. *A loss ℓ is* robust to label noise *if for all $\sigma \in (0, \frac{1}{2})$,*

$$Q \leq_\ell Q' \Leftrightarrow Q_\sigma \leq_\ell Q'_\sigma.$$

In words, the decision maker correctly orders distributions if they assume no noise. Robustness to label noise easily implies,

$$\operatorname*{arg\,min}_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim P}\ell(y, f(x)) = \operatorname*{arg\,min}_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim P_\sigma}\ell(y, f(x)),$$

for all $\mathcal{F}$. Given any $\sigma \in (0, \frac{1}{2})$, Natarajan et al. showed in [101] how to correct for the corruption by associating with *any* loss, a corrected loss,

$$\ell_\sigma(y, v) = \frac{(1-\sigma)\ell(y, v) - \sigma\ell(-y, v)}{1 - 2\sigma}.$$

with the property,

$$\mathbb{E}_{(v,y)\sim Q}\ell(y, v) = \mathbb{E}_{(v,y)\sim Q_\sigma}\ell_\sigma(y, v), \ \forall Q \in \mathbb{P}(\mathbb{R} \times Y).$$

This is a specific instance of the corruption corrected losses considered in chapter 3.

Robustness to label noise can be characterized by the order equivalence of ℓ and ℓ_σ.

Definition 4.11 (Order Equivalence). *Two loss functions ℓ_1 and ℓ_2 are* order equivalent *if for all distributions $Q, Q' \in \mathbb{P}(\mathbb{R} \times Y)$,*

$$Q \leq_{\ell_1} Q' \Leftrightarrow Q \leq_{\ell_2} Q'.$$

Lemma 4.12. *ℓ is robust to label noise if and only if for all $\sigma \in (0, \frac{1}{2})$, ℓ and ℓ_σ are order equivalent.*

In words, the decision maker correctly orders distributions if they incorrectly assume noise. Following on from these insights, we now characterize when a loss is robust to label noise.

Theorem 4.13 (Characterization of Robustness). *Let ℓ be a loss with $\ell(-1, v) \neq \ell(1, v)$. Then ℓ is robust to label noise if and only if there exists a constant C such that,*

$$\ell(1, v) + \ell(-1, v) = C, \; \forall v \in \mathbb{R}.$$

Ghosh et al. in [63] prove a one way result. Misclassification loss satisfies the conditions for theorem 4.13, however it is difficult to minimize directly. For linear loss,

$$\ell(1, v) + \ell(-1, v) = 1 - v + 1 + v = 2.$$

Therefore linear loss is robust to label noise. Furthermore, up to equivalence, linear loss is the only convex function that satisfies 4.13.

Theorem 4.14 (Uniqueness of Linear Loss). *A loss ℓ is convex in its second argument and is robust to label noise if and only if there exists a constant λ and a function $g : Y \to \mathbb{R}$ such that,*

$$\ell(y, v) = \lambda y v + g(y).$$

4.5.1 Beyond Symmetric Label Noise

Thus far we have assumed that the noise on positive and negative labels is the same. A sensible generalization is label conditional noise, were the label $y \in \{-1, 1\}$, is flipped with a label dependent probability. Following Natarajen et al. [101], we can correct for class conditional label noise in the same way we can correct for symmetric label noise, and use the loss,

$$\ell_{\sigma_{-1},\sigma_1}(y, v) = \frac{(1 - \sigma_{-y})\ell(y, v) - \sigma_y \ell(-y, v)}{1 - \sigma_{-1} - \sigma_1}.$$

If the decision maker knows the ratio $\frac{\sigma_{-1}}{\sigma_1}$, then for a certain class of loss functions they can avoid estimating noise rates.

Theorem 4.15. *Let ℓ be a loss with $\sigma_1 \ell(-1, v) + \sigma_{-1}\ell(1, v) = C$ for all $v \in \mathbb{R}$. Then $\ell_{\sigma_{-1},\sigma_1}$ and ℓ are order equivalent.*

For linear loss,

$$\sigma_1(1+v) + \sigma_{-1}(1-v) = \sigma_1 + \sigma_{-1} + (\sigma_1 - \sigma_{-1})v,$$

which is not constant in v unless $\sigma_1 = \sigma_{-1}$. Linear (and similarly misclassification loss) are no longer robust under label conditional noise. This result also means there is no non trivial convex loss that is robust to label conditional noise for all noise rates $\sigma_{-1} + \sigma_1 < 1$, as linear loss would be a candidate for such a loss.

Progress can be made if one works with more general error measures, beyond expected loss. For a distribution $P \in \mathbb{P}(X \times Y)$, let $P_\pm \in \mathbb{P}(X)$ be the conditional distribution over instances given a positive or negative label respectively. The balanced error function is defined as,

$$\mathrm{BER}_\ell(P_+, P_-, f) = \frac{1}{2}\mathbb{E}_{x\sim P_+}\ell(1, f(x)) + \frac{1}{2}\mathbb{E}_{x\sim P_-}\ell(-1, f(x)).$$

If both labels are equally likely under P, then the balanced error is exactly the expected loss. The balanced error "balances" the two class, treating errors on positive and negative labels equally. Closely related to the problem of learning under label conditional noise, is the problem of learning under mutually contaminated distributions, presented in Menon et al. [100]. Rather than samples from the clean label conditional distributions, the decision maker has access to samples from corrupted distributions $\tilde{P}_\pm$, with,

$$\tilde{P}_+ = (1-\alpha)P_+ + \alpha P_- \text{ and } \tilde{P}_- = \beta P_+ + (1-\beta)P_-,\ \alpha + \beta < 1.$$

In words, the corrupted $\tilde{P}_y$ is a combination of the true P_y and the unwanted P_{-y}. We warn the reader that α and β are *not* the noise rates on the two classes. However, in section 2.3 of Menon et al. [100], they are shown to be related to $\sigma_{\pm 1}$ by an invertible transformation.

Theorem 4.16. *Let ℓ be robust to label noise. Then,*

$$\mathrm{BER}_\ell(\tilde{P}_+, \tilde{P}_-, f) = (1-\alpha-\beta)\mathrm{BER}_\ell(P_+, P_-, f) + \frac{(\alpha+\beta)}{2}C,$$

for some constant C.

This is a generalization of proposition 1 of Menon et al. [100], that restricts to misclassification loss. Taking argmins yields,

$$\operatorname*{arg\,min}_{f\in\mathcal{F}} \mathrm{BER}_\ell(\tilde{P}_+, \tilde{P}_-, f) = \operatorname*{arg\,min}_{f\in\mathcal{F}} \mathrm{BER}_\ell(P_+, P_-, f).$$

Thus balanced error can be optimized from corrupted distributions.

Going further beyond symmetric label noise, one can assume a general noise process with noise rates that depend both on the label and the observed instance. Define the noise function $\sigma : X \times Y \to [0, \frac{1}{2})$, with $\sigma(x,y)$ the probability that the instance label pair (x,y) has its label flipped. Rather than samples from P, the decision maker has samples from P_σ, where to sample from P_σ first sample $(x,y) \sim P$ and then flip the label with probability $\sigma(x,y)$. The recent work of Gosh et al. [63] proves the following theorem concerning the robustness properties of minimizing any loss that is robust to label noise.

Theorem 4.17. *For all distributions P, function classes $\mathcal{F}$, all noise functions $\sigma : X \times Y \to [0, \frac{1}{2})$ and all loss functions ℓ that are robust to label noise,*

$$\ell(P, f_\sigma^*) \leq \frac{\ell(P, f^*)}{\min_{(x,y)} 1 - 2\sigma(x,y)},$$

where f_σ^* and f^* are the minimizers over $\mathcal{F}$ of $\ell(P_\sigma, f)$ and $\ell(P, f)$ respectively.

Our proof of this theorem is a slight modification of the discussion that follows remark 1 in Ghosh et al. [63]. There they only consider variable noise rates that are functions of the instance. We include it for completeness. In particular, this theorem shows that if $\ell(P, f^*) = 0$ and $\max_{(x,y)} \sigma(x,y) < \frac{1}{2}$, then minimizing ℓ with samples from P_σ will also recover a classifier with zero loss against the clean P.

4.6 Herding for Sparse Approximation

```
Data: Distribution $P \in \mathbb{P}(Z)$, set of possible representative points $S \subseteq Z$,
      kernel function $\mathcal{K}$ and error tolerance $\epsilon$.
Result: Weighted set of representives $H = \{(\alpha_i, z_i)\}_{i=1}^{n}$ such that
      $\left\| \omega_P - \sum_{(\alpha,z)\in H} \alpha\psi(z) \right\| \leq \epsilon.$
Initialization: $z^* = \arg\max_{z'\in S} \mathbb{E}_{z\sim P}\mathcal{K}(z,z')$, $H = \{(1, z^*)\}$ ;
while $\left\| \omega_P - \sum_{(\alpha,z)\in H} \alpha\psi(z) \right\| > \epsilon$ do
    Let $z^* = \arg\max_{z'\in S} \mathbb{E}_{z\sim P}\mathcal{K}(z,z') - \sum_{(\alpha,z)\in H} \alpha\mathcal{K}(z,z')$ ;
    Set $\lambda^* = \arg\min_{\lambda\in[0,1]} \left\| \omega_P - \left( (1-\lambda) \sum_{(\alpha,z)\in H} \alpha\psi(z) + \lambda\psi(z) \right) \right\|$;
    Multiply all weights in $H$ by $1 - \lambda^*$ ;
    Add $(\lambda^*, z^*)$ to $H$
end
```

Algorithm 1: Pseudo-code specification of Herding.

The main problem classifying according to 4.2 is the dependence of the classifier on the *entire* sample. We show how to correct this. We first survey the technique of herding, before showing how it can be applied in estimating the classifier of 4.2.

For any set Z, mapping $\psi : Z \to \mathcal{H}$ and distribution $P \in \mathbb{P}(Z)$, define the mean $\omega_P = \mathbb{E}_{z \sim P}\psi(z)$. We recover our previous definition by taking $Z = X \times Y$ and $\psi(x,y) = y\phi(x)$. Given a set of examples $S = \{\psi(z_i)\}_{i=1}^{n}$, herding [8; 38; 128] is a method to sparsely approximate ω_P with a combination of the elements of S. In [8] it was shown that herding is an application of the Frank-Wolfe optimization algorithm to the convex problem,

$$\min_{\tilde{\omega} \in C} \|\omega_P - \tilde{\omega}\|^2 ,$$

where, $C = \text{co}(\{\psi(z) : z \in S\})$. Define the kernel $\mathcal{K}(z, z') = \langle \psi(z), \psi(z') \rangle$. Herding proceeds as in algorithm 1. Intuitively, herding begins by selecting the point in S that is most similar on average to draws from P, as measured by $\mathcal{K}$. When selecting a new representative, herding chooses the point in S that is most similar on average to draws from P *while being different from previously chosen points*. If herding runs for m iterations, then an approximation of ω_P with only m elements is obtained. One can also take $\lambda^* = \frac{1}{|R|+1}$, leading to uniform weights.

Herding can also be viewed as minimizing $\text{MMD}_\psi(P, Q)$, where the approximating distribution Q is concentrated on S [38]. Originally, herding was motivated as means to produce "super samples" from a distribution P. Standard monte-carlo techniques lead to convergence at rate $\frac{1}{\sqrt{m}}$ of the square error $\|\omega_P - \hat{\omega}\| \to 0$. Using herding, under certain conditions faster rates can be achieved. For our application, P is the empirical distribution over the set S, $\frac{1}{|S|}\sum_{z \in S} \delta_z$, or equivalently $\omega_S = \frac{1}{|S|}\sum_{z \in S} \psi(z)$. As we will see, herding converges rapidly: $O(\log(\frac{1}{\epsilon}))$ iterations gives an approximation of accuracy ϵ.

The expression for the optimal λ^* is available in closed form, see section 4 of Bach et al. [8]. More exotic forms of the Frank Wolfe algorithm exist, see [77] for a fantastic review. In fully corrective methods, the line search over λ is replaced with a full optimization over *all* current points in the herd. The minimum norm point algorithm replaces the minimization over the convex hull of the current representative points with a minimization over the affine hull together with a line search step. Away step methods consider both adding a new member to the herd as well as deleting a current member. These more involved methods can be used in place of algorithm 1.

4.6.1 Rates of Convergence for Herding

Let $\tilde{\omega}_m$ be the approximation to ω_P obtained from running herding for m iterations. As discussed previously, herding can be used as a means of sampling from a distribution, with rates of convergence $\|\omega_P - \tilde{\omega}\| \to 0$ faster than that for random sampling. While in the worst case, one can not do better than a $\frac{1}{\sqrt{m}}$ rate, if $\omega_P \in C$ faster rates can be obtained [8]. Let D be the diameter of C and d the distance from ω_P to the boundary of C. For herding with line search as in algorithm 1,

$$\|\omega_P - \tilde{\omega}_m\| \leq \|\omega_P - \tilde{\omega}_1\| \, e^{-\alpha m},$$

where $\alpha = \frac{d}{2(\|\omega_P\|+D)}$ [19]. For our application $\omega_P = \omega_S = \frac{1}{S}\sum_{z \in S} \psi(z)$ which is clearly in C. If $d > 0$ the herded approximation converges quickly to ω_S.

Recent work [83] has shown that some of the more exotic varieties of the Frank-Wolfe algorithm exhibit fast convergences even when $d = 0$. They show for the fully corrective, minimum norm point and away step alternatives,

$$\|\omega_P - \tilde{\omega}_m\| \leq \left(1 - \frac{1}{4}\left(\frac{\delta}{D}\right)^2\right)^m \|\omega_P - \tilde{\omega}_1\|,$$

where δ is the *pyramidal width* of the convex hull of the samples feature vectors. While their analysis does overcome issues concerning distance to the boundary, it is based on a worst case analysis of steps of the Frank-Wolfe algorithm, leading to unexpected sets having the best constant (the largest ratio $\frac{\delta}{D}$ is given by the unit simplex which is very unlike most samples seen in practice). Furthermore the constant δ can be difficulty to calculate. More work is required to better understand the convergence properties of the herding algorithm.

4.6.2 Computational Analysis of Herding

The main bottleneck of the herding algorithm is the population of the kernel matrix, which runs in time of order n^2. Like most greedy algorithms, to calculate each iteration of the herding algorithm, only knowledge of the previously added point is required. Therefore, each iteration runs in order n. One can avoid calculating the entire kernel matrix by estimating $\mathbb{E}_{z \sim P}\mathcal{K}(z, z')$. This reduces the initialization time to order n, at the cost of extra time per iteration required to calculate the kernel between the newly added point and all the elements of the sample.

There exists many tricks to speed up the training of SVM's [118]. In section 4.6.6 we show how these methods can be applied to herding.

4.6.3 Parallel Extension

It is very easy to parallelize the herding algorithm. Rewriting the mean as a "mean of means", one has,

$$\frac{1}{n}\sum_{i=1}^{n} \psi(z_i) = \sum_{i=1}^{m} \frac{n_i}{n}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} \psi(z_{ij})\right),$$

where we have split the n data points into m disjoint groups with z_{ij} the j-th element of the i-th group. We can use herding to approximate each sub mean $\frac{1}{n_i}\sum_{j=1}^{n_i} \psi(z_{ij})$ separately. Furthermore, if we approximate each sub mean to tolerance ϵ, combining the approximations yields an approximation to the total mean with tolerance ϵ.

Lemma 4.18 (Parallel Means). *Let $\omega = \sum \lambda_i \omega_i$ with $\lambda_i \geq 0$ and $\sum \lambda_i = 1$. Suppose that for each i there is an approximation $\tilde{\omega}_i$ with $\|\omega_i - \tilde{\omega}_i\| \leq \epsilon$. Then $\|\omega - \sum \lambda_i \tilde{\omega}_i\| \leq \epsilon$.*

The proof is a simple application of the triangle inequality and the homogeneity of norms. Lemma 4.18 allows one to use a map-reduce algorithm to herd large sets of data. One splits the data into M groups, herds each group in parallel and then combines the groups, possibly herding the result.

4.6.4 Discriminative Herding for Approximating Rule 4.2

Our goal is to approximate equation (4.2), which in turn means approximating ω_S. To this end, we run herding on the sample S. Let $\psi : X \times Y \to \mathcal{H}$, with $\psi(x,y) = y\phi(x)$ and corresponding kernel $\mathcal{K}((x,y),(x',y')) = yy'K(x,x')$. We take,

$$\tilde{\omega}_S = \sum_{(\alpha,(x,y)) \in H} \alpha y \phi(x),$$

where H is the representative set (or herd) of instance, label pairs obtained from herding S to tolerance ϵ. Our approximate classifier is $\tilde{f}(x) = \langle \tilde{\omega}_S, \phi(x) \rangle$. We have by a simple application of the Cauchy-Schwarz inequality,

$$\|f - \tilde{f}\|_\infty = \sup_x |\langle \omega_S - \tilde{\omega}_S, \phi(x) \rangle| \leq \epsilon.$$

Hence the tolerance used in the herding algorithm directly controls the approximation accuracy.

4.6.5 Comparisons with Previous Work

Herding has appeared under a different name in the field of statistics, in the work of Jones [78]. There an algorithm closely related to the Frank-Wolfe algorithm (projection pursuit) is considered, and rates of convergence of $\frac{1}{\sqrt{m}}$ for the general case when $\omega_P \notin C$ are proved. The appendix of [73] features a theoretical discussion of sparse approximations. Herbrich and Williamson in [73] show the existence of a m-sparse approximation with $\|\omega - \tilde{\omega}\| \leq \frac{\sqrt{2}\epsilon_{\frac{m}{2}}(S)}{\sqrt{m}}$, with $\epsilon_{\frac{m}{2}}(S)$ the *entropy numbers* of the set S. We further explore the connections to their approach in the appendix to this chapter.

4.6.6 Comparing Herding to Sparse SVM Solvers

Recall that the SVM solves the following convex objective,

$$\operatorname*{arg\,min}_{\omega \in \mathcal{H}} \frac{1}{|S|} \sum_{(x,y) \in S} [1 - y\langle \omega, \phi(x) \rangle]_+ + \frac{\lambda}{2} \|\omega\|^2 . \tag{4.3}$$

There are many approximate, "greedy", methods to attack this problem. These methods are deeply related to Frank Wolfe algorithms [3; 40; 118]. Here we show the connection of these methods to kernel herding. It is well known that the optimal solution to the SVM objective 4.3 is of the form,

$$\omega = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i),\ \alpha_i \geq 0\ \forall i \in [1;n].$$

Let $C = \mathrm{co}(\{y\phi(x) : (x,y) \in S\})$. If we normalize the α_i, i.e. take $\sum_{i=1}^{n} \alpha_i = 1$ (which does not change the outputted classifier), then $\omega \in C$. For all $\omega \in C$, $\|\omega\| \leq 1$. Therefore, via an application of the Cauchy Schwarz inequality, the SVM objective 4.3 is equivalent to,

$$\operatorname*{arg\,min}_{\omega \in C} 1 - \langle \omega, \omega_S \rangle + \frac{\lambda}{2} \|\omega\|^2 .$$

Setting $\lambda = 1$ gives optimal solution $\omega^* = \omega_S$. Furthermore, for $\lambda = 1$,

$$\underbrace{-\langle \omega, \omega_S \rangle + \frac{1}{2}\|\omega\|^2}_{\text{SVM objective}} = \frac{1}{2}\|\omega - \omega_S\|^2 - \underbrace{\frac{1}{2}\|\omega_S\|^2}_{\text{Independent of } \omega} .$$

Therefore the SVM objective 4.3 reduces to the herding objective,

$$\operatorname*{arg\,min}_{\omega \in C} \|\omega - \omega_S\|^2 .$$

Herding can thus be understood as the application of "greedy" algorithms presented in [3; 40; 118] to a sufficiently regularized SVM objective.

4.6.7 Sparsity Inducing Objectives versus Sparsity Inducing Algorithms

Much of practical machine learning can be understood as solving regularized empirical loss problems,

$$\operatorname*{arg\,min}_{\omega \in \mathcal{H}} \frac{1}{|S|} \sum_{(x,y) \in S} \ell(y, \langle \omega, \phi(x)\rangle) + \Omega(\omega),$$

with ℓ a loss and Ω a regularizer. It is desirable for the evaluation speed of the outputted classifier that ω be as sparse as possible. For example, the linear loss objective does not return a sparse solution. There are two main approaches to this problem.

One can understand objectives that promote sparsity, via sparsity inducing losses or sparsity inducing regularizers. For example in the LASSO, the L1 regularizer $\Omega(\omega) = \lambda \sum_{i=1}^{n} |\omega_i|$ is used [115]. Alternately, Bartlett and Tewari in [18] use the standard square norm regularizer, $\Omega(\omega) = \frac{\lambda}{2}\|\omega\|^2$, and vary the loss. They show there is an inherit trade off between sparse solutions, and solutions that give calibrated probability estimates. We point out that this is for this *particular* choice of regularizer. In the objective based approach, properties of the *actual* minimizer are deduced from the KKT conditions of the relevant optimization objective.

In practice, one rarely if ever returns the *exact* minimizer. Therefore, the search of *objectives* that have sparse minimizers does not tell the full story. The approach taken here, and in [3; 40; 118], is to use an optimization algorithm that provides sparsity for free.

In the context of learning with symmetric label noise, this further highlights the importance of strong robustness. What is important is how the objective *orders* solutions, and not necessarily what the *exact* minimizer of the objective is.

4.7 Conclusion

We have taken a simple classifier, given by the sample mean, and have placed it on a firm theoretical grounding. We have shown its relation to maximum mean discrepancy, highly regularized support vector machines and finally to kernel density estimation. We have proven a surrogate regret bound highlighting its usefulness in learning classifiers, as well as generalization bounds for single and multiple feature maps. We have analysed the robustness properties of the mean classifier, and have shown that linear loss is the *only* convex loss function that is robust to symmetric label noise. Finally, we have shown how herding can be used to speed up its evaluation. The result is a conceptually clear, theoretically justified means of learning classifiers.

Appendix to Chapter 4

4.8 Proof of Concept Experiment

Here we include a proof of concept experiment, highlighting the performance of herding as a means of compressing data sets. Keeping up with the current fashion, we consider classifying 3's versus 8's from the MNIST data set, comprising 11982 training examples and 1984 test examples. We normalize all pixel values to lie in the interval $[0, 1]$ and use a Gaussian kernel with bandwidth 1. We plot the test set performance of the learned classifier as a function of the percentage of the training set used in the herd. To produce the dashed curve, we recursively herd with an allowed error of 0.01 (i.e. we herd the data set, and then the herd and so on). To produce the dotted curve, we recursively use parallel herding with an allowed error of 0.025 and a maximum number of 200 data points in each sub division. Each large dot signifies a herd. For both curves, we recurse until there are only 100 data points in the herd. As a baseline (in red), we plot the performance of the mean of the entire training set.

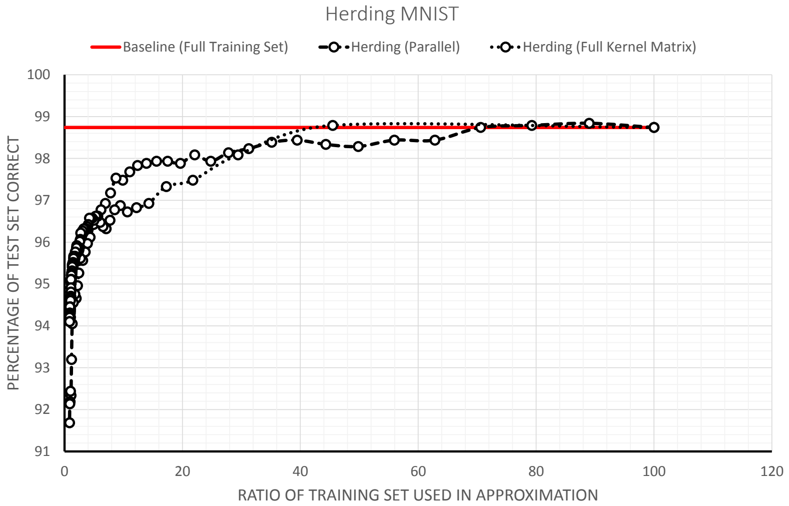


Fig. 4.1: Experiment on the MNIST data set highlighting herding's ability to compress data sets. Curves are produced by recursively running the herding algorithm (herding the data and then the herd and so on), see text (best viewed in colour).

The baseline method achieves test set performance of 98.74%. Firstly, the curves for parallel and non-parallel herding are qualitatively the same. We comment on the non-parallel herding. We see that with little as 1% of the training set, an accuracy of over 94% is obtained. The performance of the herded samples rapidly approaches that of the full mean. Less than 20% of the training set affords an accuracy of over 97%.

4.9 Long and Servedio Example

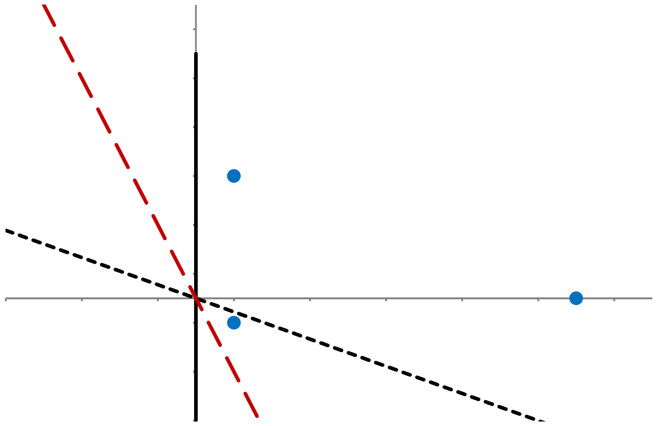


Fig. 4.2: Long and Servedio's example highlighting the non-robustness to label noise of hinge loss minimization. See text.

Figure 4.2 details Long and Servedio's example highlighting the non-robustness to label noise of hinge loss minimization. The distribution P is concentrated on the blue points, with each point deterministically labelled positive. The southern most point is chosen with probability $\frac{1}{2}$, and the other two points are chosen with probability $\frac{1}{4}$. The function class considered is hyper planes through the origin. Solving for,

$$\operatorname*{arg\,min}_{\omega \in \mathbb{R}^2} \mathbb{E}_{(x,y)\sim P}[1 - \langle \omega, x \rangle]_+,$$

yields the solid black hyperplane, which correctly classifies all points. Solving for,

$$\operatorname*{arg\,min}_{\omega \in \mathbb{R}^2} \mathbb{E}_{(x,y)\sim P_\sigma}[1 - \langle \omega, x \rangle]_+,$$

for sufficiently large σ, yields the dashed black hyperplane, which incorrectly classifies the southern most point. As this point is chosen with probability $\frac{1}{2}$, this classifier performs as well as random guessing. The scale of the data set can be chosen so that this occurs for σ arbitrarily small. In contrast the mean solution provides the red hyperplane, which correctly classifies all data points.

4.10 Proof of Theorem 4.1

Proof. It is well known that $f_P \in \arg\min_{f\in[-1,1]^X} \ell_{01}(P,f)$. From P define P_X to be the marginal distribution over instances and $\eta(x) = P(Y = 1|X = x)$. Then,

$$\begin{aligned}\ell_{\text{linear}}(P,f) &= \mathbb{E}_{(x,y)\sim P}1 - yf(x)\\ &= \mathbb{E}_{x\sim P_X}1 + (1-2\eta(x))f(x).\end{aligned}$$

Minimizing over $f \in [-1,1]^X$ gives $f(x) = -1$ if $1 - 2\eta(x) \geq 0$ i.e. when $\eta(x) < \frac{1}{2}$ and $f(x) = 1$ otherwise. This proves the first claim. We have,

$$\ell_{\text{linear}}(P,f_P) = \mathbb{E}_{x\sim P_X}1 - |(1-2\eta(x))|.$$

Therefore,

$$\begin{aligned}\ell_{\text{linear}}(P,f) - \ell_{\text{linear}}(P,f_P) &= \mathbb{E}_{x\sim P_X}(1-2\eta(x))f(x) + |(1-2\eta(x))|\\ &= \mathbb{E}_{x\sim P_X}|(1-2\eta(x))| - \operatorname{sign}(2\eta(x)-1)\,|(1-2\eta(x))|\,f(x)\\ &= \mathbb{E}_{x\sim P_X}|(1-2\eta(x))|\,(1-\operatorname{sign}(2\eta(x)-1)f(x)).\end{aligned}$$

It is well known that,

$$\ell_{01}(P,f) - \ell_{01}(P,f_P) = \mathbb{E}_{x\sim P_X}|(1-2\eta(x))|\,[\![\operatorname{sign}(2\eta(x)-1)f(x) \leq 0]\!].$$

We complete the proof by noting $[\![v \leq 0]\!] \leq 1 - v$ for $v \in [-1,1]$.

□

4.11 PAC-Bayesian Bounds for Linear Loss

Here we develop general bounds for learning with linear loss. Theorems 4.3 and 4.5 are recovered as special cases. For the following, ℓ will denote linear loss.

Let $\mathcal{F} \subseteq \mathbb{R}^X$. Denote the expected linear loss of $f \in \mathcal{F}$ by $\ell(P,f)$. We consider *randomized* algorithms $\mathcal{A} : \cup_{n=1}^{\infty}(X \times Y)^n \to \mathbb{P}(\mathcal{F})$. For any algorithm $\mathcal{A}$, define the mean function $\bar{\mathcal{A}} : \cup_{n=1}^{\infty}(X \times Y)^n \to \mathbb{R}^X$,

$$\bar{\mathcal{A}}(S)(x) = \mathbb{E}_{f\sim\mathcal{A}(S)}f(x).$$

For a distribution over functions $Q \in \mathbb{P}(\mathcal{F})$, define the *doubly annealed* loss,

$$\ell_{\beta\beta}(P,Q) = -\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{f\sim Q}e^{-\beta(1-yf(x))}).$$

Theorem 4.19 (PAC-Bayes Linear Loss theorem). *For all distributions P, $\phi : X \to \mathcal{H}$, priors π, randomized algorithms $\mathcal{A}$ and $\beta > 0$,*

$$\mathbb{E}_{S\sim P^n}\ell_{\beta\beta}(P,\mathcal{A}(S)) \leq \mathbb{E}_{S\sim P^n}\left[\ell(S,\bar{\mathcal{A}}(S)) + \frac{D_{KL}(\mathcal{A}(S),\pi)}{\beta n}\right].$$

Furthermore, with probability at least $1-\delta$ on a draw from $S\sim P^n$ with $\mathcal{A}$, π and β fixed before the draw,

$$\ell_{\beta\beta}(P,\mathcal{A}(S)) \leq \ell(S,\bar{\mathcal{A}}(S)) + \frac{D_{KL}(\mathcal{A}(S),\pi) + \log(\frac{1}{\delta})}{\beta n}.$$

Proof. This is theorem A.6 of the appendix for linear loss, coupled with the convexity of $-\log$.

□

We call π the prior and $\mathcal{A}(S)$ the posterior. The decision maker is lucky (has a tighter bound), if $D_{KL}(\mathcal{A}(S),\pi)$ is small. For linear function classes we identify $f_\omega \in \mathcal{F}_\phi$ with its weight vector ω. We take $\mathcal{A}(S) \in \mathbb{P}(\mathcal{H})$ and with a slight abuse of notation define $\bar{\mathcal{A}}(S) = \mathbb{E}_{\omega\sim\mathcal{A}(S)}\omega$. We have,

$$\bar{\mathcal{A}}(S)(x) = \mathbb{E}_{\omega\sim\mathcal{A}(S)}\langle\omega,\phi(x)\rangle = \langle\bar{\mathcal{A}}(S),\phi(x)\rangle \in \mathcal{F}_\phi.$$

The sample risk of the posterior distribution is determined by its mean. To exploit this, we focus on posteriors and priors of simple form, allowing exact calculation of the annealed loss and the KL divergence term. We assume $\pi = \mathcal{N}(\omega_\pi, \mathbb{1})$ and $\mathcal{A}(S) = \mathcal{N}(\bar{\mathcal{A}}(S),\mathbb{1})$. In words, priors and posteriors are normal distributions with identity covariance. This restriction and the following theorem lead to theorem 4.2.

Theorem 4.20. *For all distributions P, feature maps ϕ, prior vectors $\omega_\pi \in \mathcal{H}$, sample dependent weight vectors $\bar{\mathcal{A}} : (X\times Y)^n \to \mathcal{H}$ and $\beta > 0$ such that $\|\phi(x)\| \leq 1$ $\forall x$ and $\|\bar{\mathcal{A}}(S)\| \leq 1$ $\forall S$,*

$$\mathbb{E}_{S\sim P^n}\ell(P,\bar{\mathcal{A}}(S)) \leq \mathbb{E}_{S\sim P^n}\left[\ell(S,\bar{\mathcal{A}}(S)) + \frac{\|\bar{\mathcal{A}}(S)-\omega_\pi\|^2}{\beta n}\right] + \beta.$$

Furthermore, with probability at least $1-\delta$ on a draw from $S\sim P^n$ with $\bar{\mathcal{A}}$, ω_π and β fixed before the draw,

$$\ell(P,\bar{\mathcal{A}}(S)) \leq \ell(S,\bar{\mathcal{A}}(S)) + \frac{\|\bar{\mathcal{A}}(S)-\omega_\pi\|^2 + \log(\frac{1}{\delta})}{\beta n} + \beta.$$

Proof. We begin with theorem 4.19 and the function class $\mathcal{F}_\phi$. For priors and posteriors given by normal distributions,

$$D_{KL}(\mathcal{A}(S),\pi) = \|\bar{\mathcal{A}}(S)-\omega_\pi\|^2.$$

For the left hand side of the bound,

$$\begin{aligned}
&-\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{\omega\sim\mathcal{A}(S)}e^{-\beta(1-\langle\omega,y\phi(x)\rangle)})\\
=&-\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{\omega\sim\mathcal{N}(\bar{\mathcal{A}}(S),\mathbb{1})}e^{-\beta(1-\langle\omega,y\phi(x)\rangle)})\\
=&-\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}e^{-\beta(1-\langle\bar{\mathcal{A}}(S),y\phi(x)\rangle)+\frac{\beta^2}{2}\|\phi(x)\|^2}),
\end{aligned}$$

where the final line follows from standard results on the moment generating function of normal distributions. We can lower bound this quantity as follows,

$$\begin{aligned}
&-\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}e^{-\beta(1-\langle\bar{\mathcal{A}}(S),y\phi(x)\rangle)+\frac{\beta^2}{2}\|\phi(x)\|^2})\\
\geq&-\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}e^{-\beta(1-\langle\bar{\mathcal{A}}(S),y\phi(x)\rangle)})-\frac{\beta}{2}\\
\geq&\mathbb{E}_{(x,y)\sim P}1-\langle\bar{\mathcal{A}}(S),y\phi(x)\rangle-\beta\\
=&1-\langle\bar{\mathcal{A}}(S),\omega_P\rangle-\beta,
\end{aligned}$$

where the first line follows as $-\log$ is a decreasing function and $\|\phi(x)\|\leq 1$, and the second follows from lemma A.8 of the appendix, which can be applied as by Cauchy-Schwarz,

$$|1-\langle\bar{\mathcal{A}}(S),y\phi(x)\rangle|\in[0,2].$$

By theorem 4.19 we have,

$$\mathbb{E}_{S\sim P^n}1-\langle\bar{\mathcal{A}}(S),\omega_P\rangle-\beta\leq\mathbb{E}_{S\sim P^n}\left[1-\langle\bar{\mathcal{A}}(S),\omega_S\rangle+\frac{\|\bar{\mathcal{A}}(S)-\omega_\pi\|^2}{\beta n}\right],$$

with a corresponding high probability version. □

To recover theorem 4.3, consider the algorithm,

$$\mathcal{A}(S)=\mathcal{N}(\omega_S,\mathbf{1}),$$

with prior $\omega_\pi=0$. Upper bounding $\|\bar{\mathcal{A}}(S)-\omega_\pi\|^2\leq 1$ yields,

$$\ell(P,\omega_S)\leq\ell(S,\omega_S)+\frac{1+\log\left(\frac{1}{\delta}\right)}{\beta n}+\beta.$$

Finally, optimize over β.

PAC-Bayesian Bounds for Learning over Multiple Feature Maps

It is common for the decision maker to have access to several feature maps $\phi_i : X \to \mathcal{H}_i$, for i in a (possibly infinite) index set $\mathcal{I}$. Define,

$$\mathcal{F}_{\mathcal{I}} = \cup_{i \in \mathcal{I}} \mathcal{F}_{\phi_i},$$

the disjoint union of the function classes $\mathcal{F}_{\phi_i}$. Rather than priors and posteriors on a single $\mathcal{F}_{\phi_i}$, we consider distributions on $\mathcal{F}_{\mathcal{I}}$ that are *mixtures* of normals,

$$\begin{aligned}&\mathcal{A}(S) = i{\sim}\alpha(S),\ \omega^i{\sim}\mathcal{N}(\bar{\mathcal{A}}^i(S), \mathbf{1})\\ &\pi = i{\sim}\alpha_\pi,\ \omega^i{\sim}\mathcal{N}(\omega^i_\pi, \mathbf{1}),\end{aligned}$$

where $\pi^i_\omega, \bar{\mathcal{A}}^i(S) \in \mathcal{H}_i$ and $\alpha_\pi, \alpha(S) \in \mathbb{P}(\mathcal{I})$. These distributions first pick a tag i and then generate a weight vector $\omega^i \in \mathcal{H}_i$.

Theorem 4.21. *For all distributions P, collections of feature maps ϕ_i, prior weights $\alpha_\pi \in \mathbb{P}(\mathcal{I})$, prior vectors $\omega^i_\pi \in \mathcal{H}_i$, sample dependent weights $\alpha(S) \in \mathbb{P}(\mathcal{I})$, sample dependent weight vectors $\bar{\mathcal{A}}^i(S) \in \mathcal{H}_i$ and $\beta > 0$ such that $\|\phi(x)\| \leq 1\ \forall x$ and $\|\bar{\mathcal{A}}^i(S)\| \leq 1\ \forall S$,*

$$\begin{aligned}&\mathbb{E}_{S\sim P^n}\mathbb{E}_{i\sim\alpha(S)}\ell(P, \bar{\mathcal{A}}^i(S))\\ &\leq \mathbb{E}_{S\sim P^n}\left[\mathbb{E}_{i\sim\alpha(S)}\ell(S, \bar{\mathcal{A}}^i(S)) + \frac{D_{KL}(\alpha(S), \alpha_\pi) + \mathbb{E}_{i\sim\alpha(S)}\|\bar{\mathcal{A}}^i(S) - \omega_\pi\|^2}{\beta n}\right] + \beta.\end{aligned}$$

Furthermore, with probability at least $1 - \delta$ on a draw from $S{\sim}P^n$ with $\bar{\mathcal{A}}^i$, ω^i_π and β fixed before the draw,

$$\begin{aligned}&\mathbb{E}_{i\sim\alpha(S)}\ell(P, \bar{\mathcal{A}}^i(S))\\ &\leq \mathbb{E}_{i\sim\alpha(S)}\ell(S, \bar{\mathcal{A}}^i(S)) + \frac{D_{KL}(\alpha(S), \alpha_\pi) + \mathbb{E}_{i\sim\alpha(S)}\|\bar{\mathcal{A}}^i(S) - \omega_\pi\|^2}{\beta n} + \beta.\end{aligned}$$

Proof. The proof proceeds in very similar fashion to that of the previous theorem. We begin with theorem 4.19 and the function class $\mathcal{F}_{\mathcal{I}}$. By simple properties of the KL divergence [46], for priors and posteriors given by mixtures of normal distributions,

$$D_{KL}(\mathcal{A}(S), \pi) = D_{KL}(\alpha(S), \alpha_\pi) + \mathbb{E}_{i\sim\alpha(S)}\|\bar{\mathcal{A}}^i(S) - \omega_\pi\|^2.$$

For the left hand side of the bound,

$$\begin{aligned}&-\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{\omega\sim\mathcal{A}(S)}e^{-\beta(1-\langle\omega, y\phi(x)\rangle)})\\ &= -\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{i\sim\alpha(S)}\mathbb{E}_{\omega\sim\mathcal{N}(\bar{\mathcal{A}}^i(S),\mathbb{1})}e^{-\beta(1-\langle\omega^i, y\phi_i(x)\rangle)})\\ &= -\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{i\sim\alpha(S)}e^{-\beta(1-\langle\bar{\mathcal{A}}^i(S), y\phi_i(x)\rangle)+\frac{\beta^2}{2}\|\phi_i(x)\|^2}),\end{aligned}$$

where the final line follows from standard results on the moment generating function of normal distributions. We can lower bound this quantity as follows,

$$\begin{aligned}
&-\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{i\sim\alpha(S)}e^{-\beta(1-\langle\bar{\mathcal{A}}^i(S),y\phi_i(x)\rangle)+\frac{\beta^2}{2}\|\phi_i(x)\|^2})\\
\geq&-\frac{1}{\beta}\log(\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{i\sim\alpha(S)}e^{-\beta(1-\langle\bar{\mathcal{A}}^i(S),y\phi_i(x)\rangle)})-\frac{\beta}{2}\\
\geq&\mathbb{E}_{(x,y)\sim P}\mathbb{E}_{i\sim\alpha(S)}1-\langle\bar{\mathcal{A}}^i(S),y\phi(x)\rangle-\beta\\
=&\mathbb{E}_{i\sim\alpha(S)}1-\langle\bar{\mathcal{A}}(S),\omega_P\rangle-\beta,
\end{aligned}$$

where the first line follows as $-\log$ is a decreasing function and $\|\phi(x)\|\leq 1$, and the second follows from lemma A.8 of the appendix, which can be applied as by Cauchy-Schwarz,

$$|1-\langle\bar{\mathcal{A}}(S),y\phi(x)\rangle|\in[0,2].$$

By theorem 4.19 we have,

$$\begin{aligned}
&\mathbb{E}_{S\sim P^n}\mathbb{E}_{i\sim\alpha(S)}1-\langle\bar{\mathcal{A}}^i(S),\omega_P\rangle-\beta\\
\leq\ &\mathbb{E}_{S\sim P^n}\left[\mathbb{E}_{i\sim\alpha(S)}1-\langle\bar{\mathcal{A}}^i(S),\omega_S\rangle+\frac{D_{KL}(\alpha(S),\alpha_\pi)+\mathbb{E}_{i\sim\alpha(S)}\|\bar{\mathcal{A}}^i(S)-\omega_\pi\|^2}{\beta n}\right],
\end{aligned}$$

with a corresponding high probability version.

□

To recover theorem 4.5, consider the algorithm with,

$$\mathcal{A}^i(S)=\mathcal{N}(\omega_S^i,\mathbf{1}),$$

and $\alpha(S)$ placing all mass on the feature map with minimum $1-\|\omega_S^i\|$. Using prior, $\omega_\pi^i=0$ and α_π the uniform distribution of $[1;k]$ and upper bounding,

$$\|\bar{\mathcal{A}}^i(S)-\omega_\pi^i\|^2\leq 1 \text{ and } D_{KL}(\alpha(S),\alpha_\pi)\leq\log(k),$$

yields,

$$\ell(P,\omega_S^*)\leq\ell(S,\omega_S^*)+\frac{1+\log(k)+\log\left(\frac{1}{\delta}\right)}{\beta n}+\beta.$$

Finally, optimise over β.

4.12 Proof of Theorem 4.6

Before the proof we prove the following simple lemma.

Lemma. *Let $v,\tilde{v}\in\mathbb{R}$ with $|v-\tilde{v}|\leq\epsilon$. Then $\tilde{v}<0$ implies $v<\epsilon$.*

Proof. We have $v - \epsilon \leq \tilde{v} \leq v + \epsilon$. If $\tilde{v} < 0$, then $v - \epsilon < 0$.

□

We now prove the theorem.

Proof. First we prove the forward implication. By the conditions of the theorem, $|f(x) - \tilde{f}(x)| \leq \epsilon$ for all $x \in X$, meaning $|yf(x) - y\tilde{f}(x)| \leq \epsilon$ for all pairs (x, y). By the previous lemma, $y\tilde{f}(x) < 0$ implies $yf(x) < \epsilon$. This means,

$$[\![y\tilde{f}(x) < 0]\!] \leq [\![yf(x) < \epsilon]\!].$$

Averaging over P yields the desired result. For the reverse implication, define the function,

$$\tilde{f}(x) = \begin{cases} 0 & : |f(x)| \leq \epsilon \\ f(x) & : |f(x)| > \epsilon \end{cases}$$

By simple calculation $\left\|f - \tilde{f}\right\|_\infty \leq \epsilon$ and $\ell_{01}(P, \tilde{f}) = \ell_\epsilon(P, f)$. By assumption, $\ell_{01}(P, \tilde{f}) \leq \alpha$. Therefore $\ell_\epsilon(P, f) \leq \alpha$.

□

4.13 Comparison with Makovoz's Theorem

We call $\omega \in \text{co}(S)$ *m-sparse* if it is a combination of only m elements of S. Makovoz's theorem [93] is an existential result concerning the degree to which one can approximate *any* $\omega \in \text{co}(S)$, with an m-sparse approximation $\tilde{\omega}_m$. Let $\{B(z_i, \epsilon)\}_{i=1}^n$ be a collection of n balls in $\mathcal{H}$. We say such a collection of balls *covers* S if $S \subseteq \cup_{i=1}^n B(z_i, \epsilon)$. We call ϵ the *radius* of the cover. Define the nth entropy number of S as,

$$\epsilon_m(S) := \inf\{\epsilon : \exists \text{ a cover of } S \text{ with radius } \epsilon \text{ and } n \leq m\}.$$

The entropy number of S is a fine grained means to assess its complexity. Intuitively, the simpler S is the faster $\epsilon_n(S)$ decays as $n \to \infty$. The following is theorem 27 in [73].

Theorem 4.22. *Let $\mathcal{H}$ be a Hilbert space of dimension d. Then for all finite $S \subseteq \mathcal{H}$, for all $\omega \in \text{co}(S)$, and for all even $m \leq |S|$ there exists an m-sparse $\tilde{\omega} \in \text{co}(S)$ such that,*

$$\|\omega - \tilde{\omega}\| \leq \frac{\sqrt{2}\epsilon_{\frac{m}{2}}(S)}{\sqrt{m}}.$$

Theorem 4.22 has an advantage over the analysis in section 4.6.1. It includes more information about the sample than just the diameter of S and the distance from the sample mean to the boundary of S in the form of the entropy numbers of S. It is known for S the d-dimensional unit ball, $m^{-\frac{1}{d}} \leq \epsilon_m(S) \leq 4m^{-\frac{1}{d}}$ (see equation 1.1.10

of [33]). Naively, this means theorem 4.22 gives rates of convergence,

$$\|\omega - \tilde{\omega}\| \leq \frac{4\sqrt{2}}{m^{\frac{1}{2}+\frac{1}{d}}},$$

where d can be replaced by $|S|$ for infinite dimensional problems. This suggests that herding outperforms the bound in theorem 4.22. Ideally one wants a version of equation 2 that has *direct* reference to the entropy numbers of S. This will be the subject of future work.

4.14 Proof of Lemma 4.12 and Theorem 4.13

Before the proofs we require the following lemma.

Lemma 4.23. *Let ℓ_1 and ℓ_2 be loss functions. ℓ_1 and ℓ_2 are order equivalent if and only if there exists constants $\alpha > 0$ and β such that,*

$$\ell_2(y, v) = \alpha \ell_1(y, v) + \beta.$$

This is theorem 2 of section 7.9 in [54]. We now prove lemma 4.12.

Proof. We begin with the reverse implication. Since,

$$\mathbb{E}_{(v,y)\sim Q}\ell(y, v) = \mathbb{E}_{(v,y)\sim Q_\sigma}\ell_\sigma(y, v),\ \forall Q, Q',$$

we have $Q \leq_\ell Q' \Leftrightarrow Q_\sigma \leq_{\ell_\sigma} Q'_\sigma$. As we assume ℓ and ℓ_σ are order equivalent, $Q_\sigma \leq_{\ell_\sigma} Q'_\sigma \Leftrightarrow Q_\sigma \leq_\ell Q'_\sigma$. Therefore,

$$Q \leq_\ell Q' \Leftrightarrow Q_\sigma \leq_\ell Q'_\sigma.$$

For the forward implication, define the loss ℓ' with,

$$\begin{pmatrix} \ell'(-1, v) \\ \ell'(1, v) \end{pmatrix} = \begin{pmatrix} 1-\sigma & \sigma \\ \sigma & 1-\sigma \end{pmatrix} \begin{pmatrix} \ell(-1, v) \\ \ell(1, v) \end{pmatrix},\ \forall v \in \mathbb{R}.$$

It is easily verified that $\ell'_\sigma = \ell$. This means,

$$\mathbb{E}_{(v,y)\sim Q}\ell'(y, v) = \mathbb{E}_{(v,y)\sim Q_\sigma}\ell(y, v),\ \forall Q, Q',$$

but as $Q \leq_\ell Q' \Leftrightarrow Q_\sigma \leq_\ell Q'_\sigma$, we have,

$$Q \leq_\ell Q' \Leftrightarrow Q \leq_{\ell'} Q'.$$

Therefore ℓ and ℓ' are order equivalent. Invoking lemma 4.23 and the definition of ℓ'

yields,

$$\begin{pmatrix} 1-\sigma & \sigma \\ \sigma & 1-\sigma \end{pmatrix}\begin{pmatrix} \ell(-1,v) \\ \ell(1,v) \end{pmatrix} = \alpha \begin{pmatrix} \ell(-1,v) \\ \ell(1,v) \end{pmatrix} + \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ \forall v \in \mathbb{R},$$

for $\alpha > 0$. This yields,

$$\begin{pmatrix} \ell(-1,v) \\ \ell(1,v) \end{pmatrix} = \alpha \underbrace{\left(\frac{1}{1-2\sigma} \begin{pmatrix} 1-\sigma & -\sigma \\ -\sigma & 1-\sigma \end{pmatrix}\begin{pmatrix} \ell(-1,v) \\ \ell(1,v) \end{pmatrix}\right)}_{\ell_\sigma} + \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ \forall v \in \mathbb{R}.$$

Therefore ℓ is order equivalent to ℓ_σ.

□

We now prove theorem 4.13.

Proof. As ℓ and ℓ_σ are order equivalent, by the lemma 4.23, $\ell_\sigma(y,v) = \alpha\ell(y,v) + \beta$. Combined with the definition of ℓ_σ yields,

$$\frac{(1-\sigma)\ell(y,v) - \sigma\ell(-y,v)}{1-2\sigma} = \alpha\ell(y,v) + \beta.$$

Setting $y = \pm 1$ yields the following two equations,

$$(1-\sigma)\ell(1,v) - \sigma\ell(-1,v) = (1-2\sigma)(\alpha\ell(1,v) + \beta) \quad (4.4)$$

$$(1-\sigma)\ell(-1,v) - \sigma\ell(1,v) = (1-2\sigma)(\alpha\ell(-1,v) + \beta). \quad (4.5)$$

Adding these two equations together and dividing through by $1-2\sigma$ yields,

$$\ell(1,v) + \ell(-1,v) = \alpha(\ell(1,v) + \ell(-1,v)) + 2\beta. \quad (4.6)$$

If $\alpha \neq 1$, $\ell(1,v) + \ell(-1,v) = \frac{2\beta}{1-\alpha} = C$ and the proof is complete. If $\alpha = 1$, $\beta = 0$ by 4.6. Inserting these values into 4.4 yields,

$$(1-\sigma)\ell(1,v) - \sigma\ell(-1,v) = (1-2\sigma)\ell(1,v).$$

Thus $\ell(1,v) = \ell(-1,v)$, an excluded pathological case. For the converse, if $\ell(y,v) + \ell(-y,v) = C$ then $\ell(-y,v) = C - \ell(y,v)$. This means,

$$\begin{aligned} \ell_\sigma(y,v) &= \frac{(1-\sigma)\ell(y,v) - \sigma\ell(-y,v)}{1-2\sigma} \\ &= \frac{(1-\sigma)\ell(y,v) - \sigma(C - \ell(y,v))}{1-2\sigma} \\ &= \frac{1}{1-2\sigma}\ell(y,v) - \frac{\sigma C}{1-2\sigma}, \end{aligned}$$

and thus by the above lemma, ℓ and ℓ_σ are order equivalent.

□

4.15 Proof of Theorem 4.14

Proof. We begin with the forward implication. We have $\ell(y,v)$ is convex in v, furthermore $\ell(y,v)+\ell(-y,v)=C$. This means $\ell(y,v)=C-\ell(-y,v)$, hence $-\ell(-y,v)$ is convex. Thus as $\ell(y,v)$ and $-\ell(y,v)$ are convex, $\ell(y,v)=\alpha_y v+g(y)$. But,

$$\begin{aligned}\ell(y,v)+\ell(-y,v) &= \alpha_y v+g(y)+\alpha_{-y}v+g(-y)\\ &= (\alpha_y+\alpha_{-y})v+g(y)+g(-y)\\ &= C.\end{aligned}$$

Therefore $\alpha_{-y}=-\alpha_y=\lambda$ and $\ell(y,v)=\lambda yv+g(y)$. For the converse, if $\ell(y,v)=\lambda yv+g(y)$, then,

$$\ell(y,v)+\ell(-y,v)=g(y)+g(-y)=C.$$

□

4.16 Proof of Theorem 4.15

Proof. If $\sigma_1\ell(-1,v)+\sigma_{-1}\ell(1,v)=C$, this means $\sigma_{-y}\ell(y,v)+\sigma_y\ell(-y,v)=C$ for all y. This yields,

$$\begin{aligned}\ell_{\sigma_{-1},\sigma_1}(y,v) &= \frac{(1-\sigma_{-y})\ell(y,v)-\sigma_y\ell(-y,v)}{1-\sigma_{-1}-\sigma_1}\\ &= \frac{(1-\sigma_{-y})\ell(y,v)-(C-\sigma_{-y}\ell(y,v))}{1-\sigma_{-1}-\sigma_1}\\ &= \frac{1}{1-\sigma_{-1}-\sigma_1}\ell(y,v)-\frac{C}{1-\sigma_{-1}-\sigma_1},\end{aligned}$$

where the first line is the definition of $\ell_{\sigma_{-1},\sigma_1}(y,v)$ and the second is by assumption. By lemma 4.23, $\ell_{\sigma_{-1},\sigma_1}$ and ℓ are order equivalent.

□

4.17 Proof of Theorem 4.16

Proof. Recall the balanced error,

$$\mathrm{BER}_\ell(P_+,P_-,f)=\frac{1}{2}\mathbb{E}_{x\sim P_+}\ell(1,f(x))+\frac{1}{2}\mathbb{E}_{x\sim P_-}\ell(-1,f(x)).$$

Remember that,

$$\tilde{P}_+=(1-\alpha)P_++\alpha P_- \text{ and } \tilde{P}_-=\beta P_++(1-\beta)P_-.$$

This means for all classifiers f,

$$\begin{aligned}\mathbb{E}_{x\sim\tilde{P}_+}\ell(1,f(x)) &= (1-\alpha)\mathbb{E}_{x\sim P_+}\ell(1,f(x)) + \alpha\mathbb{E}_{x\sim P_-}\ell(1,f(x)) \\ &= (1-\alpha)\mathbb{E}_{x\sim P_+}\ell(1,f(x)) - \alpha\mathbb{E}_{x\sim P_-}\ell(-1,f(x)) + C\alpha,\end{aligned}$$

where in the second line we have used the fact that $\ell(1,v) = C - \ell(-1,v)$. Similarly,

$$\mathbb{E}_{x\sim\tilde{P}_-}\ell(-1,f(x)) = -\beta\mathbb{E}_{x\sim P_+}\ell(1,f(x)) + (1-\beta)\mathbb{E}_{x\sim P_-}\ell(-1,f(x)) + C\beta.$$

Taking the average of these two equations yields,

$$\mathrm{BER}_\ell(\tilde{P}_+,\tilde{P}_-,f) = (1-\alpha-\beta)\mathrm{BER}_\ell(P_+,P_-,f) + \frac{(\alpha+\beta)}{2}C.$$

□

4.18 Proof of Theorem 4.17

Proof. Firstly, for all classifiers f,

$$\begin{aligned}\ell(P_\sigma,f) &= \mathbb{E}_{(x,y)\sim P}(1-\sigma(x,y))\ell(y,f(x)) + \sigma(x,y)\ell(-y,f(x)) \\ &= \mathbb{E}_{(x,y)\sim P}(1-\sigma(x,y))\ell(y,f(x)) + \sigma(x,y)(C-\ell(y,f(x))) \\ &= \mathbb{E}_{(x,y)\sim P}(1-2\sigma(x,y))\ell(y,f(x)) + C\mathbb{E}_{(x,y)\sim P}\sigma(x,y),\end{aligned}$$

where in the second line we have used the fact that $\ell(1,v)+\ell(-1,v) = C$. Now let,

$$f^*_\sigma = \operatorname*{arg\,min}_{f\in\mathcal{F}} L(P_\sigma,f) \text{ and } f^* = \operatorname*{arg\,min}_{f\in\mathcal{F}} L(P,f),$$

respectively. By definition, $\ell(P_\sigma,f^*_\sigma) \leq \ell(P_\sigma,f^*)$. Combined with the above this yields,

$$\mathbb{E}_{(x,y)\sim P}(1-2\sigma(x,y))\ell(y,f^*_\sigma(x)) \leq \mathbb{E}_{(x,y)\sim P}(1-2\sigma(x,y))\ell(y,f^*(x)).$$

From the assumption that $\sigma(x,y) < \frac{1}{2}$ for all $(x,y)\in X\times Y$,

$$\min_{(x,y)} 1-2\sigma(x,y) \leq 1-2\sigma(x,y) \leq 1,\ \forall(x,y)\in X\times Y.$$

This yields,

$$\left(\min_{(x,y)} 1-2\sigma(x,y)\right)\mathbb{E}_{(x,y)\sim P}\ell(y,f^*_\sigma(x)) \leq \mathbb{E}_{(x,y)\sim P}\ell(y,f^*(x)),$$

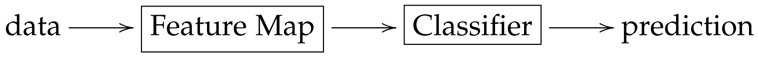
and the proof is complete.

□

5

Feature Learning via Transitions

Machine Learning methods are only as good as the features they learn from. This simple observation has led to a plethora of feature learning methods. From methods that aim to learn features and a linear classifier in one go such as neural networks and predictive sparse coding [28; 74; 92], to methods based on conditional independence tests [14; 61; 116; 127], to unsupervised feature learning methods [21; 69; 74; 102; 123] and of course good old fashion hand engineered features. While there exist many heuristic justifications for these methods, what is lacking is a general theory of feature learning.



We are all familiar with the above flow chart. Many methods exist to optimize each of the above components. For a real application we are interested in measuring the predictive performance of the combined system. For the sake of *modularity* we seek means to measure the quality of each *component* separately. We seek a measure of the quality of a feature map that is *independent* from the rest of the system, as well as a means to *combine* this with the generalization performance of a classification algorithm to provide bounds on the entire system.

To this end we review both supervised and unsupervised feature learning schemes, presenting a novel supervised feature learning objective (section 5.2.1) as well as novel means to measure the quality of features learnt independently from the supervised task they are used in (theorem 5.4). We draw inspiration from both rate distortion theory [46] as well as the comparison of statistical experiments [86; 117]. We provide to our knowledge the *first* framework from which to understand feature learning as well as a *characterization* (theorem 5.4) of when unsupervised feature learning is possible within our framework. We show that some existing feature learning schemes can be realized as solving surrogates to the objective presented in theorem 5.4.

5.1 Notation and Preliminaries

Throughout the chapter, Y, X and Z will denote the label, instance and feature spaces respectively. For simplicity, we work with proper losses $\ell : Y \times \mathbb{P}(Y) \to \mathbb{R}$, defined in section 2.4. When working with a general loss, we can take ℓ to be its canonical proper loss, as discussed in section 2.4. Recall the notion of entropy,

$$\underline{L}(P) = \inf_{Q \in \mathbb{P}(Y)} \mathbb{E}_{y \sim P} \ell(y, Q) = \mathbb{E}_{y \sim P} \ell(y, P).$$

We assume that the labels are distributed according to $\pi \in \mathbb{P}(Y)$ and that the relationship between labels Y and instances X is modelled by an experiment $e \in \mathbb{T}(Y, X)$. For any algorithm $\mathcal{A}_X \in \mathbb{T}(X, \mathbb{P}(Y))$ recall the risk,

$$\mathcal{R}_\ell^\pi(e, \mathcal{A}) := \mathbb{E}_{y \sim \pi} \mathbb{E}_{x \sim e(y)} \mathbb{E}_{Q \sim \mathcal{A}_X(x)} \ell(y, Q),$$

minimum Bayesian risks and Bayes optimal algorithm by,

$$\underline{\mathcal{R}}_\ell^\pi(e) = \min_{\mathcal{A}_X} \mathcal{R}_\ell^\pi(e, \mathcal{A}_X) \text{ and } \mathcal{A}_X^* = \arg\min_{\mathcal{A}_X} \mathcal{R}_\ell^\pi(e, \mathcal{A}_X),$$

the minimum Bayesian risk and Bayes optimal algorithm respectively. Denote by π_X, the marginal distribution over X. Let $\eta_X(x)$ be the conditional distribution of Y given a particular $x \in X$. By standard manipulations,

$$\underline{\mathcal{R}}_\ell^\pi(e) = \mathbb{E}_{x \sim \pi_X} \underline{L}(\eta_X(x)) = \mathbb{E}_{x \sim \pi_X} \mathbb{E}_{y \sim \eta_X(x)} \ell(y, \eta_X(x))$$

and $\mathcal{A}_X^*(x) = \eta_X(x)$, the true conditional distribution of Y given $x \in X$. Finally recall that the regret for a proper loss has the following simple expression,

$$\Delta\ell(P, Q) = \mathbb{E}_{y \sim P} \left[\ell(y, Q) - \ell(y, P) \right].$$

5.2 Supervised Feature Learning

For a multitude of reasons including but not limited to, computation, storage, the curse of dimensionality, increased classification performance, knowledge discovery and so on we may wish to process the instances through a (possibly randomized) feature map $T \in \mathbb{T}(X, Z)$. For a given feature map, learning follows the protocol: First, nature draws $y \sim \pi$ and $x \sim e(y)$. Second, the decision maker observes $z \sim T(x)$ and chooses a distribution Q via an algorithm $\mathcal{A}_Z$. Finally, the decision maker incurs loss $\ell(y, Q)$. Diagrammatically,

$$\begin{tikzcd} y \sim \pi \arrow[r] & \boxed{e} \arrow[r, "x"] & \boxed{T} \arrow[r, "z"] & \boxed{\mathcal{A}_Z} \arrow[r, "Q"] & \ell(y, Q). \end{tikzcd}$$

Ideally T should contain all the relevant information in x for predicting Y, The *feature gap*,

$$\Delta\underline{\mathcal{R}}_\ell^\pi(e,T) := \underline{\mathcal{R}}_\ell^\pi(T \circ e) - \underline{\mathcal{R}}_\ell^\pi(e),$$

should be small. Denote by $\eta_Z(z)$, the conditional distribution of Y given a particular $z \in Z$.

Theorem 5.1. *For all priors π, experiments e, feature maps T and loss functions ℓ,*

$$\Delta\underline{\mathcal{R}}_\ell^\pi(e,T) = \mathbb{E}_{x\sim\pi_X}\mathbb{E}_{z\sim T(x)}\Delta\ell(\eta_X(x),\eta_Z(z))$$

The feature gap is the expected regret suffered in predicting Y with the "cruder" η_Z versus the clean η_X. Theorem 5.1 can be seen to underpin several algorithms for supervised feature learning. One picks a proper loss ℓ and minimizes,

$$\Delta\underline{\mathcal{R}}_\ell^\pi(e,T) = \mathbb{E}_{x\sim\pi_X}\mathbb{E}_{z\sim T(x)}\Delta\ell(\eta_X(x),\eta_Z(z))).$$

If ℓ is strictly proper, i.e. $\Delta\ell(P,Q) = 0$ if and only if $P = Q$, and the feature gap is zero for this loss, it will be zero for all other losses. If,

$$\Delta\underline{\mathcal{R}}_\ell^\pi(e,T) = 0, \ \forall\ell,$$

then by the Blackwell-Sherman Stein theorem 2.24, $T \circ e \mid e$, i.e. there is a reconstruction $R \in \mathbb{T}(Z,X)$ with $e = R \circ T \circ e$. In this case, the features are a sufficient statistic for e. This means X and Y are conditionally independent given the features Z. Methods that minimize the objective in theorem 5.1 can therefore be understood as finding features that are approximately sufficient for the experiment e.

Using log loss gives $\Delta\ell(P,Q) = D_{KL}(P,Q)$ which leads to the information bottleneck of Tishby et al [116]. More general Bregman divergences lead to clustering with Bregman divergences [14]. Banerjee et al. in [14] present a meta algorithm for minimizing the objective in theorem 5.1. This method is closely related to the Blahut-Arimoto algorithm of rate distortion theory [46].

In practice, one might not know the exact loss function to use. Care must be taken in choosing a suitable surrogate or set of surrogates. We show in section 5.4 that the loss function can greatly influence the choice of feature map. This should be of no surprise as the loss function *defines* the relevant information contained in X for predicting Y [105].

5.2.1 Link to Deficiency

If the loss is not known one can perform a worst case analysis, and endeavour to minimize over T,

$$\sup_{\ell,\|\ell\|_\infty\leq 1} \Delta\underline{\mathcal{R}}_\ell^\pi(e,T).$$

By theorem 2.32 of chapter 2,

$$\sup_{\ell, \|\ell\|_\infty \leq 1} \Delta\underline{\mathcal{R}}^\pi_\ell(e, T) = \Xi^\pi(e, T \circ e).$$

Furthermore, as $e \mid T \circ e$,

$$\Xi^\pi(e, T \circ e) = \xi^\pi(T \circ e, e) = \min_{R \in \mathbb{T}(Z,X)} \mathbb{E}_{y \sim \pi_Y} V(R \circ T \circ e(y), e(y)).$$

In words, for T to provide good features no matter what the loss, we must be able to reconstruct the experiment e from the experiment $T \circ e$. Unlike in chapter 3, R *is* a transition.

This suggests a means to construct features when the loss function is not known, by minimizing the deficiency. While this may appear difficult, one can exploit properties of the variational divergence that make calculating the deficiency a linear minimization problem (see lemma 2.35). As long as the sets X, Y and Z are finite, fast methods exist to solve this problem. One can obtain features by finding,

$$\arg\min_{R \in \mathbb{T}(Z,X), T \in \mathbb{T}(X,Z)} \mathbb{E}_{y \sim \pi_Y} V(R \circ T \circ e(y), e(y)),$$

and then using T as the feature map. This can be solved approximately through an alternating scheme of linear minimization problems (see section 5.11). Examples of how this method behaves on some toy problems are given in section 5.4.

5.3 Unsupervised Feature Learning

One drawback of supervised feature learning methods is that they require knowledge of the joint distribution of instances and labels, and possibly of the loss function of interest. These methods consider a single supervised task in isolation. They extract the information in X that is relevant to predicting Y. In many problems of interest the decision maker has access to a large data set of unlabelled samples drawn from π_X, however they may have limited knowledge of the tasks that X will be *used for*. They seek a feature map that loses little of the information contained in X, no matter what task X is used in.

Here we make the assumption that we have enough data to form an accurate estimate of π_X, the marginal distribution over instances, and ask the following question. Under what conditions can we guarantee that a feature map does not lose more than ϵ information about Y no matter what the relation between X and Y or the loss function? The only restriction we place on the possible joint distributions on instance label pairs is that the marginal distribution over instances is π_X.

Theorem 5.2. *For all feature maps T, $\Delta\underline{\mathcal{R}}^\pi_\ell(e, T) \leq \epsilon\|\ell\|$ for all label spaces Y, loss func-*

tions ℓ and label priors π and experiments e with $e(\pi) = \pi_X$, if and only if there exists a reconstruction $R \in \mathbb{T}(Z, X)$ such that,

$$\mathbb{E}_{x \sim \pi_X} \mathbb{E}_{x' \sim R \circ T(x)} [\![x' \neq x]\!] \leq \epsilon.$$

In order to minimize the worst case information loss, one needs to be able to reconstruct X from Z with high probability. We show in the next section that under some of the heuristic justifications of deep learning techniques, like the autoencoder [123], one is solving a surrogate to this problem.

Theorem 5.2 makes no use of any geometric structure on the instance space X. It is required that the instance be reconstructed *exactly*. For a particular supervised task, the conditional distribution η_X and the loss define a geometry on X. If we make certain smoothness assumptions about this geometry, then we are no longer required to reconstruct the features exactly.

Definition 5.3. *For all conditional distributions $\eta_X \in \mathbb{T}(X, Y)$ and proper losses $\ell : Y \times \mathbb{P}(Y) \to \mathbb{R}$ the* reconstruction regret *is,*

$$D_{\ell,\eta}(x', x) = \Delta\ell(\eta_X(x'), \eta_X(x)).$$

The reconstruction regret is the regret suffered in choosing actions based on a nearby x' when in fact one should have used x. Let $d : X \times X \to \mathbb{R}$ be a dissimilarity function on X, i.e. a positive function with $d(x', x) = 0$ if and only if $x = x'$. If we assume that the supervised tasks that X is to be used in are "smooth" with respect to d, then we no longer need to reconstruct the instances *exactly*. Rather, we only require reconstructing well according to d.

Theorem 5.4. *Let $d : X \times X \to \mathbb{R}$ be a dissimilarity function on X. For all feature maps T the following are equivalent,*

1. *$\exists R \in \mathbb{T}(Z, X)$ such that $\mathbb{E}_{x \sim \pi_X} \mathbb{E}_{x' \sim R \circ T(x)} d(x', x) \leq \epsilon$.*
2. *For all η_X and loss functions ℓ with $D_{\ell,\eta}(x', x) \leq \lambda d(x', x)$ $\forall x, x'$,*
   $\Delta\underline{\mathcal{R}}_\ell^\pi(e, T) \leq \epsilon\lambda$.

Theorem 5.2 follows by taking d to be the discrete metric on X, i.e. $d(x', x) = 0$ if $x = x'$ and 1 otherwise.

5.3.1 Surrogate Approaches Motivated by Theorem 5.2

Theorem 5.2 requires that the instances can be reconstructed from the features with high probability. Many existing feature learning methods are motivated through an appeal to the Infomax principle [88], which requires that features should be chosen to maximize the mutual information between features and raw data, or equivalently to minimize the conditional entropy. Here we show the Infomax principle is a surrogate to theorem 5.2. Let $\nu(z)$ be the conditional probability of X given $z \in Z$. Recall the entropy and conditional entropy, see for example [46],

$$H(X) = \mathbb{E}_{x \sim \pi_X} - \log(\pi_X(x)) \text{ and } H(X|Z) = \mathbb{E}_{x \sim \pi_X} \mathbb{E}_{z \sim T(x)} - \log(\nu(z)),$$

Theorem 5.5 (Hellman-Raviv [72]). *Let X and Z be finite spaces. For all feature maps T and marginals over instances π_X,*

$$\inf_R \mathbb{E}_{x \sim \pi_X} \mathbb{E}_{x' \sim R \circ T(x)} [\![x' \neq x]\!] \leq \frac{1}{2} H(X|Z).$$

The conditional entropy bounds the smallest probability of error possible when one attempts to reconstruct X from the features. One can view the Infomax principle as a surrogate to reconstruction error. By exploiting various representations of $H(X|Z)$, many other surrogates to reconstructing with high probability can be obtained [20; 123]. By the properness of log loss,

$$H(X|Z) = \inf_{\tilde{\nu} \in \mathbb{T}(Z,X)} \mathbb{E}_{x \sim \pi_X} \mathbb{E}_{z \sim T(x)} - \log(\tilde{\nu}(x)).$$

For example if $X = \mathbb{R}^n$, and we restrict the possible $\tilde{\nu}$ to distributions of the form $\tilde{\nu}(z) = \mathcal{N}(f(z), \sigma^2)$, i.e. normal distributions with mean $f(z)$ and standard deviation σ, we obtain,

$$H(X|Z) \leq \inf_{f,\sigma} \mathbb{E}_{x \sim \pi_X} \mathbb{E}_{z \sim T(x)} \frac{1}{2\sigma^2} (x - f(z))^2 + \log(\sqrt{2\pi}\sigma).$$

If we restrict the possible feature maps to deterministic functions $g \in Z^X$, then the autoencoder is obtained,

$$\arg\min_g H(X|Z) = \arg\min_{f,g} \mathbb{E}_{x \sim \pi_X} \frac{1}{2} (x - f \circ g(x))^2.$$

Hence the autoencoder can be seen as a surrogate approach motivated by theorem 5.2. Furthermore, the autoencoder can be motivated via theorem 5.4, with $X = \mathbb{R}^n$ and $d(x', x) = \|x' - x\|^2$, the squared euclidean distance.

Many feature learning methods such as K-means and principle component analysis can be seen as specific instances of the autoencoder. For example if f and g are restricted to linear functions, PCA is recovered. If Z is finite set, then the autoencoder becomes K-means. We summarize this in the table below.

The Autoencoder in its Different Forms, $X = \mathbb{R}^n$

K-Means	$Z = [1;k]$, f and g arbitrary.
PCA	$Z = \mathbb{R}^m$, with $m < n$. f and g linear.
Deep Autoencoder	$Z = \mathbb{R}^m$, with $m < n$. f and g deep neural networks.

5.3.2 Rate Distortion Theory

Theorems 5.2 and 5.4 provide upper bounds for the feature gap in terms of the reconstruction error. These upper bounds can be calculated without knowledge of the particular task that the features are used in. Rate distortion theory provides lower bounds.

Recall the mutual information $I(X;Z) = H(X) - H(X|Z)$. Rate-distortion theory [46], provides lower bounds on the *distortion*, or in our terminology $\underline{\mathcal{R}}_\ell^\pi(T \circ e)$, in terms of the *rate*, the mutual information between instances and features $I(X;Z)$. Let $\ell : Y \times A \to \mathbb{R}$ be a general loss. The rate distortion function for ℓ is given by,

$$RD_\ell(\sigma) = \inf\{I(Y;A) : \mathbb{E}_{y\sim\pi}\mathbb{E}_{a\sim\mathcal{A}(y)}\ell(x,a) \leq \sigma\}.$$

In words, the rate distortion function is the minimum mutual information required between Y and the actions chosen by the algorithm to ensure expected loss less than σ. It is a non-increasing function of σ. The Blahut-Arimoto algorithm is a fast iterative scheme for calculating this function for an arbitrary loss, see for example [46]. The decision maker is not allowed to use *any* transition from Y to A, rather they are restricted to those of the form $\mathcal{A} = \mathcal{A}_Z \circ T \circ e$ as in the diagram below,

$$\begin{tikzcd} Y \arrow[r, "e"] & X \arrow[r, "T"] & Z \arrow[r, "\mathcal{A}_Z"] & A. \end{tikzcd}$$

By a form of the data processing theorem, presented in Cover and Thomas [46], $I(Y;A) \leq I(X;Z)$. The distortion must satisfy $\sigma \geq RD_\ell^{-1}(I(Y;A))$. Combining these two facts yields,

$$RD_\ell^{-1}(I(X;Z)) \leq \underline{\mathcal{R}}_\ell^\pi(T \circ e).$$

The end to end performance of the complete system is captured in the rate distortion function, the quality of the feature map by $I(X;Z)$. Combined with theorem 5.5, one obtains bounds of the form,

$$RD_\ell^{-1}(I(X;Z)) \leq \underline{\mathcal{R}}_\ell^\pi(T \circ e) \leq \underline{\mathcal{R}}_\ell^\pi(e) + \frac{1}{2}H(X|Z)\|\ell\|_\infty,\ \forall \ell.$$

Figure 5.1 contains rate distortion curves for two loss functions $\ell : \{-1,1\} \times [0,1] \to \mathbb{R}$. In blue is the curve for the Brier or quadratic loss, and in orange is the curve for a tilted brier loss that is more biased to errors made on the first class. The curves show that more mutual information is required for small expected tilted Brier loss than for small expected Brier loss. The process of tilting a loss is explained in section 5.9.

The mutual information provides *one* surrogate means of measuring the information lost by a feature map. Are there better surrogates? If we know the loss function can we do better than mutual information for providing performance bounds? At least in the case of the lower bound the answer is yes. In [132], Ziv and Zakai consider a large class of generalized information measures. For each of these information measures a rate-distortion theorem is obtained, and in some cases using one of these other measures produce *tighter* lower bounds than mutual information.

Definition 5.6. *For convex $f : \mathbb{R}^+ \to \mathbb{R}$ with $f(1) = 0$, the f-*information *of a joint distribution P_{XY} is given by,*

$$I_f(P_{XY}) = \mathbb{E}_{P_{XY}} f(\frac{d(P_X \otimes P_Y)}{dP_{XY}}).$$

The f-information is the f-divergence between P_{XY} and the product distributions of its two marginals over X and Y. As an example of the different bounds one can obtain using f-informations, we consider a simple example where $Y = \{-1, 1\}$ and the loss is a cost sensitive misclassification loss with $\ell(-1, 1) = 1$ and $\ell(1, -1) = 4$. We consider the feature map,

$$T = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.4 & 0.5 \end{pmatrix}$$

given as a row stochastic matrix with uniform prior π_X. We consider $f(x) = (\sqrt{x} - 1)^2$ resulting in Hellinger information, as well as the standard rate distortion curves obtained from using mutual information. In figure 5.2, we plot the rate distortion curves for both mutual information (red) and Hellinger information (blue) as well as the two informations of the channel (the dashed horizontal lines). The black vertical line represents the lower bound on the distortion. For this channel Hellinger information gives a tighter (higher) lower bound.

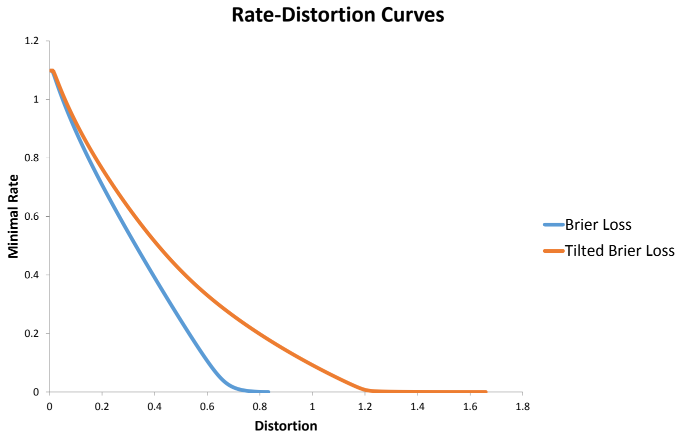


Fig. 5.1: Rate-Distortion plots using mutual information showing the performance of a channel for two different loss functions. The rate-distortion curve summarizes the trade off between rate (mutual information of the channel) and the distortion (expected loss sending messages across the channel). See text.

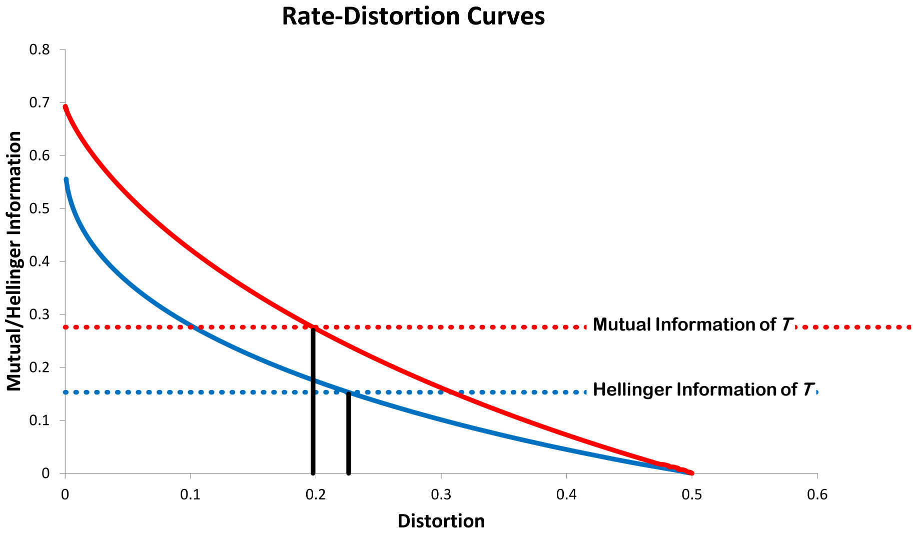


Fig. 5.2: Generalized Rate-Distortion Plots providing lower bounds on the quality of features via mutual and Hellinger information. In this example Hellinger information (blue) presents a tighter lower bound than mutual information (red). See text.

5.3.3 Hierarchical Learning of Features

One of the main tenets of the deep learning paradigm is that features should be learnt in a hierarchical fashion. Rather than learning a single feature map, one learns a chain,

$$\begin{tikzcd} X = Z_0 \arrow[r, bend left, "T_1"] & Z_1 \arrow[l, bend left, "R_1"] \arrow[r, bend left, "T_2"] & Z_2 \arrow[l, bend left, "R_2"] \arrow[r, bend left, "T_3"] & \dots \arrow[l, bend left, "R_3"] \arrow[r, bend left, "T_n"] & Z_n \arrow[l, bend left, "R_n"] \end{tikzcd}$$

with final feature map $T = T_n \circ \cdots \circ T_1$ the composition of all the feature maps in the chain, and final reconstruction given by $T = R_1 \circ \cdots \circ R_n$. The layers of such a system can be learned in a greedy fashion. We can understand this procedure as a surrogate approach motivated by theorem 5.2.

Theorem 5.7. *For all chains of feature maps and reconstruction functions,*

$$\begin{tikzcd} X = Z_0 \arrow[r, bend left, "T_1"] & Z_1 \arrow[l, bend left, "R_1"] \arrow[r, bend left, "T_2"] & Z_2 \arrow[l, bend left, "R_2"] \arrow[r, bend left, "T_3"] & \dots \arrow[l, bend left, "R_3"] \arrow[r, bend left, "T_n"] & Z_n, \arrow[l, bend left, "R_n"] \end{tikzcd}$$

the probability of reconstruction error for the entire chain is bounded by the the sum of the *reconstruction errors for each layer,*

$$\boldsymbol{P}(x' \neq x) \leq \sum_{i=0}^{n-1} \mathbb{E}_{z_i \sim \pi_{Z_i}} \mathbb{E}_{z_i' \sim R_i \circ T_i} [\![z_i \neq z_i']\!].$$

5.4 Illustrations

In this section we present simple examples of how the different feature learning schemes discussed operate in practice. We also give examples of when one can learn sufficient features for a particular experiment as well as when it is possible to learn generic features.

Experiment Specific Features. Let $Y = \mathbb{R}$ with $X = \mathbb{R}^n$ and e given by the product of n normal distributions with mean y and variance 1. It is easy to verify that the sample mean is a sufficient statistic meaning that at least for this experiment we can greatly compress the information contained in X. However, if we take as a prior for Y a normal distribution of mean 0 and variance 1, then the marginal distribution π_X will not be concentrated on a set of smaller dimension nor have any particularly interesting structure (it will be concentrated on all of $\mathbb{R}$). Hence we can not find interesting generic features in this case.

Experiment and Loss Specific Features. Let $Y = \{-1, 1\}$ with $e = \mathcal{N}(y, 1)$. For this experiment, misclassification loss and π uniform on the two labels, the Bayes optimal f is given by $f(x) = 1$ if $x > 0$ as $P(-1|x) > \frac{1}{2}$ and $f(x) = -1$ otherwise as $P(-1|x) \leq \frac{1}{2}$. It is easy to show that $\Delta\underline{\mathcal{R}}_{\ell_{01}}(e, f) = 0$, all we need is the output of f. However if we change the loss to a cost sensitive loss ℓ_c where misclassifying a positive example is more costly than a negative example, we no longer have $\Delta\underline{\mathcal{R}}_{\ell}(e, f) = 0$. This is because the optimal f will no longer threshold at 0. However, if there was a jump discontinuity in η_X, i.e. it jumped from say 0.4 to 0.6 as x crossed over $x = 0$ then the feature gap would be zero for a broader range of cost sensitive losses. Once again there are not generic features of interest.

Loss Sensitive versus Loss Insensitive Features. Let $Y = \{1, 2, 3\}$ with π uniform and e given by the normal distributions in the figure below. Consider the feature space $Z = \{1, 2\}$. Figure 5.3 shows a plot of the features learnt by two different feature learning schemes.

The first feature map does not use knowledge of the loss, and is learnt by minimizing deficiency (see section 5.2.1). The second features make use of a particular loss function, where misclassifying a 2 is more costly than misclassifying one of the other classes. This feature map is learnt via clustering with Bregman divergences. A proper loss of this form is achieved by tilting the standard Brier loss [32] toward class 2. Tilting is explained in section 5.9. The green regions are those x that are mapped to the feature 1, the blue are those mapped to 2.

We can see even in this simple example that the loss function matters when determining features. While the first feature map divides X into regions that allow good reconstruction of all the class conditionals, the second focuses on separating the conditional for 2 as dictated by the loss function. The deficiency of each of these feature maps is similar. $\xi^\pi(T \circ e, e) = 0.629$ for the first feature map and $\xi^\pi(T \circ e, e) = 0.698$ for the second. Thus from a *worst case* perspective the two feature maps are very similar. However, for the particular loss we have used to construct the second feature map, difference of the two feature maps is more pronounced. $\Delta \underline{\mathcal{R}}_\ell^\pi(e, T) = 1.075$ for the first feature map versus $\Delta \underline{\mathcal{R}}_\ell^\pi(e, T) = 0.325$ for the second.

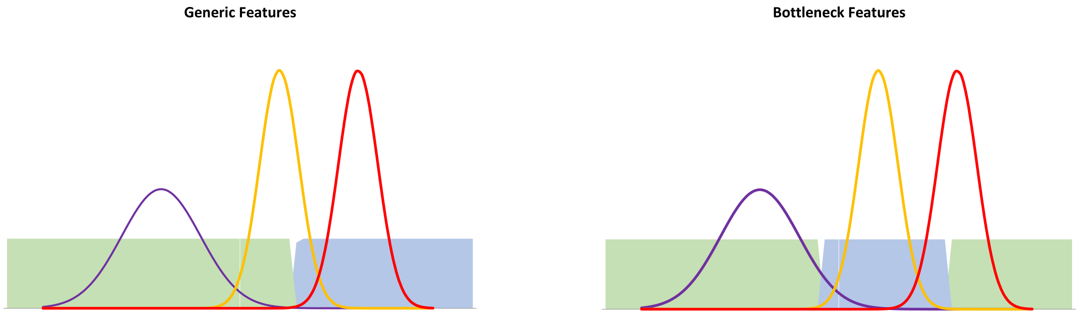


Fig. 5.3: Loss Sensitive versus Loss Insensitive Features. The curves denote the probability distribution of instances given labels, the squares denote regions with the same feature map. The plot shows that the loss really matters for choosing features. See text.

Learning Generic Features. All previous examples have considered a *fixed* experiment. When learning features in an unsupervised fashion, one wishes to find features that work for all experiments that use X. There are many examples of when this is possible, and they all boil down to some sort of manifold assumption. If π_X is concentrated on some lower dimensional subset of X, then one can find generic features. The challenges here are primarily computational, rather than information theoretic.

5.5 Conclusion

Automated feature learning methods have produced remarkable empirical results, however little theory exists explaining their performance. This chapter provides some direction. To this end, we have placed several supervised feature learning methods in a general framework, provided a novel loss insensitive objective for learning features as well as providing novel means quantifying the quality of features learnt by unsupervised methods. Finally, we have shown the usefulness of rate-distortion theory and its under utilized generalizations in ascertaining the quality of learnt features.

Appendix to Chapter 5

5.6 Proof of Theorem 5.1

Proof.

$$\begin{aligned}\underline{\mathcal{R}}_\ell^\pi(T\circ e)-\underline{\mathcal{R}}_\ell^\pi(e)&=\mathbb{E}_{y\sim\pi}\mathbb{E}_{x\sim e(y)}\mathbb{E}_{z\sim T(x)}\ell(y,\eta_Z(z))-\mathbb{E}_{y\sim\pi}\mathbb{E}_{x\sim e(y)}\ell(y,\eta_X(x))\\&=\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{y\sim\eta_X(x)}\mathbb{E}_{z\sim T(x)}\ell(y,\eta_Z(z))-\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{y\sim\eta_X(x)}\ell(y,\eta_X(x))\\&=\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{z\sim T(x)}\left[\mathbb{E}_{y\sim\eta_X(x)}\ell(y,\eta_Z(z))-\ell(y,\eta_X(x))\right]\\&=\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{z\sim T(x)}\Delta\ell(\eta_Z(z),\eta_X(x)),\end{aligned}$$

where the first line is by definition, the second is by rewriting $\mathbb{E}_{y\sim\pi}\mathbb{E}_{x\sim e(y)}$ as an expectation $\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{y\sim\eta_X(x)}$, the third uses Fubini's theorem to change the order of $\mathbb{E}_{y\sim\eta_X(x)}\mathbb{E}_{z\sim T(x)}$ to $\mathbb{E}_{z\sim T(x)}\mathbb{E}_{y\sim\eta_X(x)}$, and finally we have used the definition of regret. □

5.7 Proof of Theorem 5.4

Proof. We first prove the forward implication. By assumption, there exits a reconstruction R with,

$$\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{x'\sim R\circ T(x)}d(x'\neq x)\leq\epsilon.$$

Now consider the algorithm $\eta_X\circ R$, that first reconstructs the instance and then uses the optimal learning algorithm for the instances, η_X. We have,

$$\begin{aligned}\underline{\mathcal{R}}_\ell^\pi(T\circ e)-\underline{\mathcal{R}}_\ell^\pi(e)&\leq\mathcal{R}_\ell^\pi(T\circ e,\eta_X\circ R)-\underline{\mathcal{R}}_\ell^\pi(e)\\&=\mathbb{E}_{y\sim\pi}\mathbb{E}_{x\sim e(y)}\mathbb{E}_{x'\sim R\circ T(x)}\ell(y,\eta_X(x'))-\mathbb{E}_{y\sim\pi}\mathbb{E}_{x\sim e(y)}\ell(y,\eta_X(x)).\end{aligned}$$

Rearranging the expectations using Bayes rule and Fubini's theorem yields,

$$\begin{aligned}\underline{\mathcal{R}}_\ell^\pi(T\circ e)-\underline{\mathcal{R}}_\ell^\pi(e)&\leq\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{x'\sim R\circ T(x)}\left[\mathbb{E}_{y\sim\eta_X}\ell(y,\eta_X(x'))-\ell(y,\eta_X(x))\right]\\&=\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{x'\sim R\circ T(x)}\Delta\ell(\eta_X(x'),\eta_X(x))\\&=\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{x'\sim R\circ T(x)}D_{\ell,\eta}(x',x)\\&\leq\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{x'\sim R\circ T(x)}\lambda d(x',x)\\&\leq\lambda\epsilon,\end{aligned}$$

where the second and third line follows by definition, the forth follows by the smoothness assumption on the reconstruction regret, and finally the last line follows from our assumptions on R.

For the converse, take $Y = X$ with $e = \mathrm{id}_X$, the most informative experiment on X and loss function given by the metric itself. In this case $\eta_X(x)$ is a point mass on x and the reconstruction regret is d,

$$\Delta d(\eta_X(x'), \eta_X(x)) = d(x', x).$$

We have $\underline{\mathcal{R}}_d^\pi(\mathrm{id}_X) = 0$, which means by assumption there exists an algorithm R with,

$$\underline{\mathcal{R}}_d^\pi(T) = \mathbb{E}_{x \sim \pi_X} \mathbb{E}_{x' \sim R \circ T(x)} d(x', x) \leq \epsilon.$$

□

5.8 Proof of Theorem 5.7

Proof. Let $(z_0, \dots, z_n)$ be the "true" elements at each level of the chain and $(z_0', \dots, z_{n-1}')$ their reconstructions. Consider the joint distribution $\boldsymbol{P}$ with,

$$\boldsymbol{P}(z_0, z_1, \dots, z_n, z_0', z_1', \dots, z_{n-1}') = P(z_0)P(z_1|z_0)\dots P(z_n|z_n - 1)P(z_{n-1}'|z_n)\dots P(z_0'|z_1'),$$

where the conditional distributions are specified by the feature maps T_i and the reconstructions R_i. Under this joint distribution,

$$\begin{aligned} \boldsymbol{P}(z_0 \neq z_0') &= \boldsymbol{P}(z_0 \neq z_0' \cap z_1 = z_1') + \boldsymbol{P}(z_0 \neq z_0' \cap z_1 \neq z_1') \\ &\leq \boldsymbol{P}(z_0 \neq z_0' \cap z_1 = z_1') + \boldsymbol{P}(z_1 \neq z_1'). \end{aligned}$$

To complete the proof, note that $\boldsymbol{P}(z_0 \neq z_0' \cap z_1 = z_1') = \mathbb{E}_{z_0 \sim \pi_{Z_0}} \mathbb{E}_{z_0' \sim R_1 \circ T_1(z_0)} [\![z_0 \neq z_0']\!]$ and proceed inductively.

□

5.9 Tilted Loss Functions

In producing figures 5.1 and 5.3, we made use of the "tilted" Breir loss. Here we explain how to "tilt" any loss function, producing an asymmetric loss from a symmetric loss.

Recall from section 2.4, that a canonical loss can be represented by its entropy, $\underline{L} : \mathbb{P}(Y)^+ \to \mathbb{R}$, that assigns an uncertainty to each weight μ. For two weights $\mu_1, \mu_2 \in \mathbb{P}^+(Y)$, denote by $\mu_1 \odot \mu_2$ the weight,

$$\mu_1 \odot \mu_2(y) = \mu_1(y)\mu_2(y),$$

the element wise product of the two weights. Intuitively, this can be thought of as altering the weights of μ_2 with the "importances" μ_1. For a "tilting" weight ω, define the tilted entropy,

$$\underline{L}_\omega(\mu) = \underline{L}(\omega \odot \mu).$$

Intuitively, this entropy is $\underline{L}$ tilted toward the $y \in Y$ that are riskier. This tilted entropy can be used to produce a skewed loss function. We show this through an example.

Let $Y = \{-1, 1\}$ and let $\underline{L}$ be the Brier (or quadratic) entropy,

$$\underline{L}(\mu) = \frac{2\mu_1\mu_{-1}}{\mu_{-1} + \mu_1},$$

where $\mu_{\pm 1}$ is the weight assigned to the positive and negative labels respectively. Let $\omega = (1+\alpha, 1-\alpha)$, i.e. ω increases the weight on the negative labels and decreases the weight on positive labels.

$$\underline{L}_\omega(\mu) = \frac{2(1-\alpha)(\alpha+1)\mu_1\mu_-1}{(1+\alpha)\mu_{-1} + (1-\alpha)\mu_1}.$$

We plot Breir and tilted Breir entropies (for $\alpha = \frac{1}{2}$) as a function of the probability of negative labels in figure 5.4.

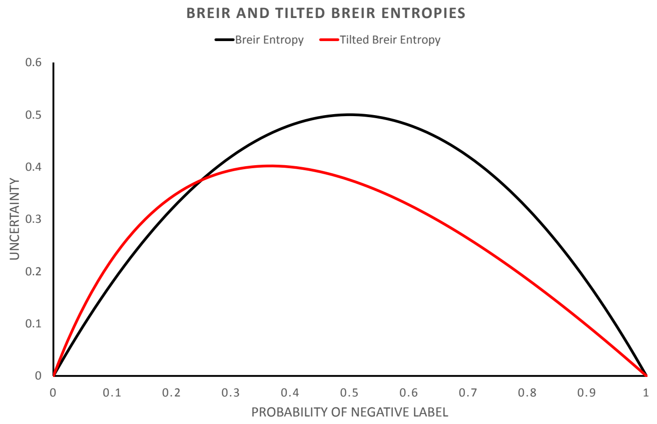


Fig. 5.4: Breir and Tilted Breir entropies. Tilting provides means to skew an entropy toward a particular class. See text.

We can see that tilting has made predicting the positive class riskier. Recall from section 2.4, that taking super gradients of an entropy produces a canonical loss. In figure 5.5, we plot the partial loss functions for Breir and tilted Breir entropies (for $\alpha = \frac{1}{2}$). As can be seen, the tilting process has made classifying negatives correctly more important.

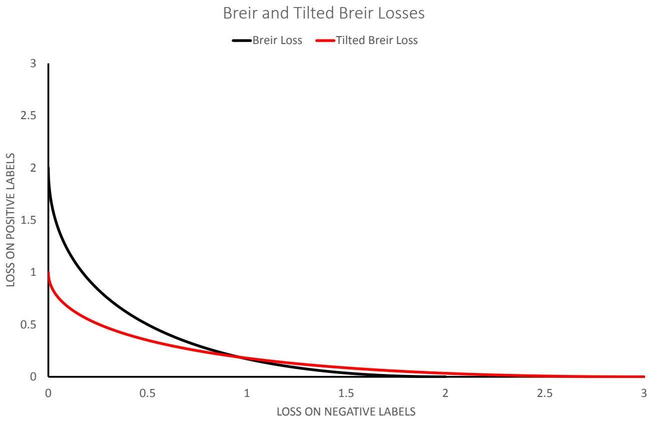


Fig. 5.5: Plot of Breir and Tilted Breir loss for $\alpha = \frac{1}{2}$. Tilting has made misclassifying negative classes more risky. See text.

5.10 The Information Bottleneck/ Clustering with Bregman Divergences

For a given experiment e and loss function ℓ, the information bottleneck/ clustering with Bregman divergences attempt to find a feature map by solving,

$$\inf_T \beta \Delta \underline{\mathcal{R}}_\ell^\pi(e,T) + I(X;Z),$$

i.e. a regularized feature gap, with the mutual information $I(X;Z)$ serving as the regularizer. Intuitively, the features should maintain information relevant for predicting Y, while throwing away as much information from X as possible. This problem can be solved by an alternating algorithm presented in both Tishby et al [116] (the original bottleneck for log loss) and Banerjee et al [14] (a generalization to Bregman divergences). Here we review the derivation of this algorithm.

Theorem 5.8 (Clustering with Bregman Divergences)**.** *For all experiments e and proper losses ℓ,*

$$\begin{aligned}&\min_T \beta \Delta \underline{\mathcal{R}}_\ell^\pi(e,T) + I(X;Z)\\ =&\min_T \min_{\tilde{\eta}_Z} \min_{\tilde{\pi}_Z} \beta \mathbb{E}_{x\sim\pi_X} \mathbb{E}_{z\sim T(x)} \Delta\ell(\eta_X(x), \tilde{\eta}_Z(z)) + \mathbb{E}_{x\sim\pi_X} D_{KL}(T(x), \tilde{\pi}_Z).\end{aligned}$$

This theorem replaces a single minimization over T, that is difficult to calculate directly, by three separate minimizations. The difficulty stems from the fact that $\eta_Z(x)$ and π_Z, the conditional distribution of labels given features and the marginal distribution over features, both depend on T. The above theorem provides means to circumvent this.

To proceed, one fixes two of the quantities and performs a minimization of the third. Each of these separate problems is easy to solve, and is the driving force behind the information bottleneck and clustering with Bregman divergences.

For the proof we require the following lemma from Banerjee et al [14] on the minimizers of Bregman divergences. Bregman divergences [29] are a large class of distance like quantities, that include the *KL* divergence and the regret (see [105] and section 2.4.2).

Recall for a concave function $\Phi : C \to \mathbb{R}$, with C a convex set, the *Bregman divergence* is given by,

$$D_\Phi(x,y) = \Phi(x) + \langle y - x, \nabla \underline{L}(x) \rangle - \Phi(y),$$

where $\nabla \underline{L}(x)$ is a super-gradient of Φ at the point x.

Lemma 5.9. *For all concave $\phi : C \to \mathbb{R}$ and distributions $P \in \mathbb{P}(C)$,*

$$\mathbb{E}_{x\sim P} x \in \operatorname*{arg\,min}_{x \in C} \mathbb{E}_{y\sim P} D_\phi(y,x).$$

The mean is the expected Bregman divergence minimizer.

This is proposition 1 of [14]. We can now prove the theorem.

Proof. Firstly,

$$I(X;Y) = \mathbb{E}_{x\sim \pi_X} D_{KL}(T(x), \pi_Z) = \min_{\tilde{\pi}_Z} \mathbb{E}_{x\sim\pi_X} D_{KL}(T(x), \tilde{\pi}_Z),$$

as $\mathbb{E}_{x\sim\pi_X} T(x) = \pi_Z$. Secondly, let $\nu(z)$ be the conditional distribution of X given a particular $z \in Z$. Then,

$$\begin{aligned}\mathbb{E}_{x\sim\pi_X}\mathbb{E}_{z\sim T(x)} \Delta\ell(\eta_X(x), \eta_Z(z)) &= \mathbb{E}_{z\sim\pi_Z}\mathbb{E}_{x\sim\nu(z)} \Delta\ell(\eta_X(x), \eta_Z(z)) \\ &= \mathbb{E}_{z\sim\pi_Z} \min_{\tilde{\eta}_Z(z)} \mathbb{E}_{x\sim\nu(z)} \Delta\ell(\eta_X(x), \tilde{\eta}_Z(z)) \\ &= \mathbb{E}_{z\sim\pi_Z} \min_{\tilde{\eta}_Z(z)} \mathbb{E}_{x\sim\nu(z)} \Delta\ell(\eta_X(x), \tilde{\eta}_Z(z)),\end{aligned}$$

as $\mathbb{E}_{x\sim\nu(z)} \eta_X(x) = \eta_Z(z)$, as this is just marginalizing over X in the Markov chain

$$Z \xrightarrow{\nu} X \xrightarrow{\eta_X} Y.$$

Combing these two results yields the theorem.

□

Theorem 5.8 allows one to (at least approximately) find loss specific features.

5.11 Loss Insensitive Feature Learning

Recall that loss insensitive feature learning seeks to find a feature map T and reconstruction R that minimize,

$$\min_{R \in \mathbb{T}(Z,X), T \in \mathbb{T}(X,Z)} \mathbb{E}_{y \sim \pi_Y} V(R \circ T \circ e(y), e(y)).$$

In lemma 2.35, we showed how deficiency can be calculated via linear programming. Here we show how the above objective can be minimized via an alternating pair of linear programs. Assuming that X, Y, Z are all finite sets, e, T and R can be represented by column stochastic matrices E, T, R respectively, with composition represented as matrix multiplication. Furthermore the marginal over labels π can be represented by a probability vector. Fixing T and minimizing over R means solving the following linear program,

$$\begin{aligned} &\inf_{M_{ij}, R_{ij}} \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} M_{ij} \\ \text{subject to } & M_{i,j}, R_{i,j} \geq 0 \ \forall i, j \\ & \sum_{i=1}^{|X|} R_{i,j} = 1 \ \forall j \\ & |\pi_i E_{ij} - \pi_i \left[RTE\right]_{ij}| \leq M_{ij} \ \forall i, j. \end{aligned}$$

The final constraint can be written as a pair of linear constraints. Fixing R and minimizing over T means solving the following linear program,

$$\begin{aligned} &\inf_{M_{ij}, T_{ij}} \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} M_{ij} \\ \text{subject to } & M_{i,j}, T_{i,j} \geq 0 \ \forall i, j \\ & \sum_{i=1}^{|Z|} T_{i,j} = 1 \ \forall j \\ & |\pi_i T_{ij} - \pi_i \left[RTE\right]_{ij}| \leq M_{ij} \ \forall i, j. \end{aligned}$$

Alternating these two minimizations provides means to find loss insensitive features.

6

Conclusion

Beginning with the notion of a transition, this thesis has sought to provide a foundational language for machine learning as a whole. The presentation is abstract. We argue that this abstraction allows greater connections to be seen. An example of this is theorem 3.19 of chapter 3. To prove this theorem, and more importantly to understand why it should be proved, many results concerning corruption corrected losses and representation of losses need to be in place. Without the abstract development and presentation of these results, it would be hard to understand why theorem 3.19 is important, or why you should look for it in the first place.

The major contributions of this thesis are:

- An increased understanding of corrupted learning problems, including new methods to learn from corrupted data as well as means to measure the *relative* difficulty of these problems (theorems 3.2, 3.4 and 3.15). Furthermore, theorem 3.19 shows that we need not abandon the framework of convex risk minimization to attack corrupted learning problems.

- A conceptually simple, easily parallelized and robust classification algorithm (chapter 4).

- An increased understanding of when it is possible to learn generically good features from data together with justification of various techniques used in practice as surrogate approaches to this problem (theorems 5.2, 5.4 and 5.5).

The contributions of this thesis would not have been possible without taking a sufficiently abstract approach. While there is never going to be one single language for all of machine learning, seeking unification is something we should all take more seriously. Often in machine learning research, the focus is on "solving problems" by producing fancy new algorithms. Theorems are replaced by tables of experimental results.

A great example of this is the field of learning with noisy labels. Experimentation reveals that the standard methods, such as the support vector machine, are not robust in this setting. Rather than clearly formulating the problem and attacking it via theorems, many have tried to use intuition and tables of results. Methods of exceptional computational complexity (surveyed in section 4.4.4) have been suggested for solving this problem, methods that in fact are not robust to label noise. As we have seen in chapter 4, by firmly grounding the problem of learning with noisy labels in clear definitions, and seeking theorems, we have discovered a classification algorithm that is *simpler* than the standard methods.

There are many directions open to explore. In order of increasing difficulty:

1. Explore the algorithmic consequences of the mean classification algorithm. There are many techniques for reducing more complicated problems to those of binary classification. For example, the problem of multi-class classification can be reduced to a collection of binary classification problems, via one versus all and one versus rest approaches. It will be worthwhile understanding how the mean classifier fits in with all of this.

2. Explore methods to learn when the corruption process is not known (see section 3.7). This could proceed in two directions; firstly one could identify more corruption invariant losses akin to the loss explored in chapter 4. Alternately, one could develop methods for *estimating* the corruption process like those presented in [100]. Generalized forms of these estimators do not exist, nor does any theory explaining why those that do exist work.

3. Develop a better understanding of how to design unsupervised feature learning schemes for use in a *particular* supervised learning algorithm. In chapter 5, we assumed the decision maker would would use the *best* algorithm for the features. Perhaps the insights gleaned in chapter 4 could prove useful.

4. Develop theory for a "restricted deficiency", when one considers a *subset* of possible losses. By theorem 2.32, the deficiency distance is a supremum of the difference in risk for *all* losses. The risk uses a *single* loss. It would be useful to use subsets of possible losses. In chapter 5, feature learning algorithms were motivated via the risk and the deficiency, i.e. we either work with one loss or all of them. A "restricted deficiency" would fill in this gap. Little work exists in this direction beyond a few comments in section 10.2 of Torgersen [117].

Transitions will serve as a guiding light in this future work.

APPENDIX

A

PAC-Bayesian Generalization Bounds

Here we present the PAC-Bayesian generalization and risk bounds used in proofs in chapters 3 and 4.

Machine Learning can be understood as the study of *prediction*. A decision maker is required to predict observations of nature. We assume the decision maker's observations lie in a set $\mathcal{Z}$, their available predictions/actions in a set A and the quality of their actions is assessed by a loss function $\ell : \mathcal{Z} \times A \to \mathbb{R}$. The decision maker summarizes their uncertainty in their observations through a probability distribution $P \in \mathbb{P}(\mathcal{Z})$. Ultimately they aim to minimize their expected loss, $\mathbb{E}_{z \sim P} \ell(z, a)$. In order to make the correct decision, the decision maker must infer the correct P. They do this by observing the phenomena in question. Define the expected loss,

$$\ell(P, Q) = \mathbb{E}_{z \sim P} \mathbb{E}_{a \sim Q} \ell(z, a).$$

In a Bayesian analysis the decision maker restricts themselves to a subset $\Theta \subseteq \mathbb{P}(\mathcal{Z})$, together with a prior $\pi \in \mathbb{P}(\Theta)$. They seek an algorithm $\mathcal{A} : \mathcal{Z} \to \mathbb{P}(A)$ that minimizes the *expected risk* under the prior distribution,

$$\mathbb{E}_{\theta \sim \pi} \mathbb{E}_{z \sim P_\theta} \ell(P_\theta, \mathcal{A}(z)).$$

It is easy to show [67] that the algorithm that minimizes the above *risk*, first calculates the Bayesian predictive distribution, and then calculates the best action to play against this distribution. This approach is overly reliant on the restriction, or "model", Θ. If the decision maker faces a P outside of this model, the Bayesian approach may result in a substantially suboptimal algorithm. The focus here is on algorithms with good worst case performance. To do this, one controls the risk with bounds of the form,

$$\mathbb{E}_{z \sim P} \ell(P, \mathcal{A}(z)) \leq \mathbb{E}_{z \sim P} \ell(z, \mathcal{A}(z)) + \psi(\mathcal{A}),$$

that link the risk of an algorithm via its ability to predict the observation. Such bounds allow the design of generic learning algorithms. Also of use are *generalization* *bounds* of the form,

$$\ell(P, \mathcal{A}(z)) \leq \ell(z, \mathcal{A}(z)) + \psi(\mathcal{A}),$$

with high probability on a draw $z \sim P$. Such bounds allow the assessment of the performance of the output of algorithms via its predictive accuracy on the observation. The term $\psi(\mathcal{A})$ punishes complicated algorithms that *over fit* the observation. One particular technique for producing these bounds provides methods that are very similar to the Bayesian approach, with a key difference. Rather than priors and posteriors on Θ, the focus is priors and posteriors on A. Rather than guessing which P they are likely to face via a model, the decision maker guesses which *actions* are more likely to be appropriate apriori. PAC-Bayesian bounds [6; 34; 65; 98; 110; 134] provide means to assess the generalization performance of learning algorithms. Here we present the version of the bounds presented in [134].

A.1 Information Exponential Inequality and Annealed Loss

All of the following bounds can be derived from the following simple lemmas.

Lemma A.1 (The Information Exponential Inequality). *For all sample spaces Ω, functions $f \in \mathbb{R}^\Omega$ and distributions $P, Q \in \mathbb{P}(\Omega)$,*

$$\mathbb{E}_P f \leq \log \mathbb{E}_Q e^f + D_{KL}(P, Q).$$

Proof. Consider the distribution $Q' \propto e^f dQ$. Then,

$$\begin{aligned} D_{KL}(P, Q') &\geq 0 \\ \mathbb{E}_P \log\left(\frac{\mathbb{E}_Q e^f}{e^f} \frac{dP}{dQ}\right) &\geq 0 \\ \log(\mathbb{E}_Q e^f) + \mathbb{E}_P \log\left(\frac{dP}{dQ}\right) - \mathbb{E}_P f &\geq 0, \end{aligned}$$

which upon rearranging gives the desired result.

□

One can understand the information exponential inequality as an application of Fenchel duality [13]. To gain high probability guarantees on the output of our algorithms, we utilize the Chernoff bound.

Lemma A.2 (Chernoff Bound). *Let $f \in \mathbb{R}^\Omega$. Then for all $\alpha \in \mathbb{R}$ and for all $P \in \mathbb{P}(\Omega)$,*

$$\mathbf{P}_{x \sim P}(f(x) \leq \alpha) \geq 1 - e^{-\alpha} \mathbb{E}_P e^f.$$

The proof of this lemma is a simple application of Markov's inequality and can be found in any standard text on probability theory.

Annealed Losses

To construct generalization bounds from the previous lemmas, one needs to choose the right f in the information exponential inequality. Naively one would think to use the *generalization gap,*

$$f(z,a) = \ell(P,a) - \ell(z,a).$$

Tighter bounds can be obtained with a different choice of f. To this end, we consider *annealed* loss functions. For any loss function ℓ and $\beta > 0$, define the *annealed* and *doubly annealed* loss,

$$\ell_\beta(P,Q) = -\frac{1}{\beta}\mathbb{E}_{a\sim Q}\log(\mathbb{E}_{z\sim P}e^{-\beta\ell(z,a)}),\ \ell_{\beta\beta}(P,Q) = -\frac{1}{\beta}\log(\mathbb{E}_{z\sim P}\mathbb{E}_{a\sim Q}e^{-\beta\ell(z,a)}).$$

The larger β, the more large losses are suppressed. By the convexity of $-\log$,

$$\ell_{\beta\beta}(P,Q) \leq \ell_\beta(P,Q) \leq \ell(P,Q).$$

Furthermore, as $\beta \to 0_+$, $\ell_{\beta\beta}(P,Q), \ell_\beta(P,Q) \to \ell(P,Q)$. The annealed loss can also be understood as the expected cummulant generating function of the random variable $\ell(-,a)$, and the doubly annealed loss the cumulant generating function of ℓ. Finally the annealed generalisation gap,

$$\beta\left(\ell_\beta(P,a) - \ell(z,a)\right),$$

can be understood as the log probability of observing z under the probability distribution $P(\mathcal{Z} = z) \propto e^{-\beta\ell(z,a)}$. The bounds that follow link $\ell_\beta(P,\mathcal{A}(z))$ to $\ell(z,\mathcal{A}(z))$ for $z{\sim}P$.

A.2 The Main Theorems

For a fixed prior $\pi \in \mathbb{P}(A)$ and algorithm $\mathcal{A} : \mathcal{Z} \to \mathbb{P}(A)$, consider the two joint distributions $P\otimes\mathcal{A}, P\otimes\pi \in \mathbb{P}(\mathcal{Z}\times A)$. The distribution $P\otimes\mathcal{A}$ *adapts* to the observation z. To sample from $P\otimes\mathcal{A}$, first sample $z{\sim}P$ and then $a{\sim}\mathcal{A}(z)$. The other samples actions from a *fixed* "prior" distribution π irrespective of the observed z.

Theorem A.3 (PAC-Bayes Expectation)**.** *For all distributions P, priors π, algorithms $\mathcal{A}$, loss functions ℓ and $\beta > 0$,*

$$\mathbb{E}_{z\sim P}\ell_\beta(P,\mathcal{A}(z)) \leq \mathbb{E}_{z\sim P}\left[\ell(z,\mathcal{A}(z)) + \frac{D_{KL}(\mathcal{A}(z),\pi)}{\beta}\right].$$

Proof. Let,

$$\begin{aligned} f(z,a) &= \beta\left(\ell_\beta(P,a) - \ell(z,a)\right) \\ &= -\log(\mathbb{E}_{z'\sim P}e^{-\beta(\ell(z',a)-\ell(z,a))}), \end{aligned}$$

the annealed generalization gap. The proof is a simple application of lemma 1 using the f and distributions defined previously. We have by the lemma,

$$\mathbb{E}_{P\otimes\mathcal{A}} f \leq \log \mathbb{E}_{P\otimes\pi} e^f + D_{KL}(P\otimes\mathcal{A}, P\otimes\pi).$$

Firstly,

$$\begin{aligned}\mathbb{E}_{P\otimes\pi} e^f &= \mathbb{E}_{P\otimes\pi} \frac{e^{-\beta\ell(z,a)}}{\mathbb{E}_{z'\sim P} e^{-\beta\ell(z',a)}} \\ &= \mathbb{E}_{a\sim\pi}\mathbb{E}_{z\sim P} \frac{e^{-\beta\ell(z,a)}}{\mathbb{E}_{z'\sim P} e^{-\beta\ell(z',a)}} \\ &= 1.\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}_{P\otimes\mathcal{A}} f &= \mathbb{E}_{z\sim P}\mathbb{E}_{a\sim\mathcal{A}(z)} \left[-\log(\mathbb{E}_{z'\sim P} e^{-\beta\ell(z',a)}) - \beta\ell(z,a)\right] \\ D_{KL}(P\otimes\mathcal{A}, P\otimes\pi) &= \mathbb{E}_{z\sim P} D_{KL}(\mathcal{A}(z),\pi),\end{aligned}$$

yielding,

$$\mathbb{E}_{z\sim P}\mathbb{E}_{a\sim\mathcal{A}(z)} - \log(\mathbb{E}_{z'\sim P} e^{-\beta\ell(z',a)}) \leq \mathbb{E}_{z\sim P}\left[\beta\ell(z,\mathcal{A}(z)) + D_{KL}(\mathcal{A}(z),\pi)\right].$$

Finally, divide both sides by β.

□

Theorem A.4 (PAC-Bayes High Probability). *With probability at least $1-\delta$ on a draw $z\sim P$ with $\mathcal{A}$, π and β fixed before the draw,*

$$\ell_\beta(P,\mathcal{A}(z)) \leq \ell(z,\mathcal{A}(z)) + \frac{D_{KL}(\mathcal{A}(z),\pi) + \log\left(\frac{1}{\delta}\right)}{\beta}.$$

Proof. Once again we invoke lemma 1 with f as in the previous theorem, yielding for all z,

$$\begin{aligned}\mathbb{E}_{a\sim\mathcal{A}(z)} f(z,a) - D_{KL}(\mathcal{A}(z),\pi) &\leq \log \mathbb{E}_{a\sim\pi} e^{f(z,a)} \\ \exp(\mathbb{E}_{a\sim\mathcal{A}(z)} f(z,a) - D_{KL}(\mathcal{A}(z),\pi)) &\leq \mathbb{E}_{a\sim\pi} e^{f(z,a)} \\ \mathbb{E}_{z\sim P}\exp(\mathbb{E}_{a\sim\mathcal{A}(z)} f(z,a) - D_{KL}(\mathcal{A}(z),\pi)) &\leq \mathbb{E}_{z\sim P}\mathbb{E}_{a\sim\pi} e^{f(z,a)} \\ &= 1,\end{aligned}$$

where the last line was shown in the previous theorem. We obtain the required result from Chernoff's bound. □

The great utility of these theorems is their remarkable generality. They apply to all negatively bounded loss functions.

A.3 Replication and Rates

Rather than one realization of $\mathcal{Z}$, it is usual to repeat an experiment several times, obtaining a sample S of n independent observations. For a sample $S \in \mathcal{Z}^n$ define,

$$\ell(S,a) = \frac{1}{|S|} \sum_{z \in S} \ell(z,a),$$

the average loss of a on the sample and $\ell(P^n,a) = \ell(P,a)$ the expected loss on the sample. $\ell(S,a)$ can also be understood as the expectation of the loss under the uniform distribution over the sample.

Lemma A.5. *For all distributions P, loss functions ℓ, $\beta > 0$ and sample sizes n,*

$$\ell_{\beta n}(P^n, Q) = \ell_\beta(P,Q).$$

Proof.

$$\begin{aligned}
\ell_{\beta n}(P^n, Q) &= -\frac{1}{\beta n} \mathbb{E}_{a \sim Q} \log(\mathbb{E}_{S \sim P^n} e^{-\beta n \ell(S,a)}) \\
&= -\frac{1}{\beta n} \mathbb{E}_{a \sim Q} \log(\mathbb{E}_{z^n \sim P^n} e^{-\beta \sum_{i=1}^n \ell(z_i,a)}) \\
&= -\frac{1}{\beta n} \mathbb{E}_{a \sim Q} \log(\prod_{i=1}^n \mathbb{E}_{z_i \sim P} e^{-\beta \ell(z_i,a)}) \\
&= -\frac{1}{\beta} \mathbb{E}_{a \sim Q} \log(\mathbb{E}_{z \sim P} e^{-\beta \ell(z,a)}),
\end{aligned}$$

where the third and fourth line follows as the z_i are iid random variables with distribution P.

□

Theorem A.6 (Replicated PAC-Bayes Theorem). *For all distributions P, priors π, algorithms $\mathcal{A}$, loss functions ℓ and $\beta > 0$,*

$$\mathbb{E}_{S \sim P^n} \ell_\beta(P, \mathcal{A}(S)) \leq \mathbb{E}_{S \sim P^n} \left[\ell(S, \mathcal{A}(S)) + \frac{D_{KL}(\mathcal{A}(S), \pi)}{\beta n} \right].$$

Furthermore, with probability at least $1-\delta$ on a draw $S \sim P^n$ with $\mathcal{A}$, π and β fixed before the draw,

$$\ell_\beta(P, \mathcal{A}(S)) \leq \ell(S, \mathcal{A}(S)) + \frac{D_{KL}(\mathcal{A}(S), \pi) + \log\left(\frac{1}{\delta}\right)}{\beta n}.$$

Proof. Use theorems A.3 and A.4 with the above lemma. □

A.4 Relationship to Union Bounds

Let A be finite with π uniform on A. We assume further that all algorithms are deterministic. By standard results in information theory (see Cover and Thomas [46]), we have $D_{KL}(\mathcal{A}(S), \pi) = \log(|A|)$, and theorem A.6 yields,

$$\ell_\beta(P, \mathcal{A}(S)) \leq \ell(S, \mathcal{A}(S)) + \frac{\log(|A|) + \log\left(\frac{1}{\delta}\right)}{\beta n},$$

with probability at least $1 - \delta$. We show how this can be realized as Chernoff's bound with a union bound. Fix an action a and let,

$$f_a(S) = \log\left(\frac{e^{-n\beta\ell(S,a)}}{\mathbb{E}_{S \sim P^n} e^{-n\beta\ell(S,a)}}\right),$$

which can be interpreted as the log probability of a distribution over samples, where samples with low loss $\ell(S, a)$, are more likely. Chernoff's bound yields,

$$\mathbf{P}_{S \sim P^n}(f_a(S) > \alpha) \leq e^{-\alpha}.$$

This bound applies for *one* action. To obtain a bound that applies for all actions simultaneously, we invoke a union bound yielding,

$$\mathbf{P}_{S \sim P^n}(f_a(S) > \alpha,\ \forall a \in A) \leq |A| e^{-\alpha}.$$

Setting $\delta = |A|e^{-\alpha}$, and rearranging yields,

$$\ell_\beta(P, a) \leq \ell(S, a) + \frac{\log(|A|) + \log\left(\frac{1}{\delta}\right)}{\beta n},\ \forall a \in A,$$

with probability at least $1 - \delta$. PAC-Bayesian bounds can therefore be understood as continuous union bounds, as pointed out by Tim van Erwen in [119]. The Chernoff bound is a vital ingredient in the proof of many concentration results, such as Hoefding's and Bernstein's inequalities. We show in the following section how these techniques can be used with PAC-Bayesian bounds.

A.5 Bounds for Bounded Losses

Here we assume the losses are *positive* and *bounded*,

$$\|\ell\|_\infty = \max_{z,a} |\ell(z, a)| \leq 1.$$

The previous theorems bound $\ell_\beta(P, \mathcal{A}(S))$ in terms of $\ell(S, \mathcal{A}(S))$. Ideally, we wish to bound $\ell(P, \mathcal{A}(S))$. The following lemmas allow this. Their proofs appear in appendix A.1 of [35]. They are used in the proofs of the standard Hoeffding and Bernstein Inequalities.

Lemma A.7. *Let $f : \Omega \to [0,1]$. For all $\beta \in \mathbb{R}$ and all $P \in \mathbb{P}(\Omega)$,*

$$(1 - e^{-\beta})\mathbb{E}_P f \leq -\log(\mathbb{E}_P e^{-\beta f}).$$

Lemma A.8. *Let $f : \Omega \to [0,1]$. For all $\beta \in \mathbb{R}$ and all $P \in \mathbb{P}(\Omega)$,*

$$\beta \mathbb{E}_P f - \frac{\beta^2}{8} \leq -\log(\mathbb{E}_P e^{-\beta f}).$$

Lemma A.9. *Let $f : \Omega \to [-1, \infty)$, then for all $\beta > 0$ and all $P \in \mathbb{P}(\Omega)$,*

$$\beta \mathbb{E}_P f - (e^\beta - 1 - \beta)\mathbb{E}_P f^2 \leq -\log(\mathbb{E}_P e^{-\beta f}).$$

Combined with the PAC-Bayes theorem these inequalities yield the follow theorems for bounded losses. For brevity we only state the high probability result.

Theorem A.10 (PAC-Bayes Multiplicative Hoeffding)**.** *For all distributions P, with probability at least $1 - \delta$ on a draw $S \sim P^n$ with $\mathcal{A}$, π and β fixed before the draw,*

$$\ell(P, \mathcal{A}(S)) \leq \frac{\beta}{1 - e^{-\beta}} \ell(S, \mathcal{A}(S)) + \frac{1}{1 - e^{-\beta}} \frac{D_{KL}(\mathcal{A}(S), \pi) + \log\left(\frac{1}{\delta}\right)}{n}.$$

Theorem A.11 (PAC-Bayes Additive Hoeffding)**.** *For all distributions P, with probability at least $1 - \delta$ on a draw $S \sim P^n$ with $\mathcal{A}$, π and β fixed before the draw,*

$$\ell(P, \mathcal{A}(S)) \leq \ell(S, \mathcal{A}(S)) + \frac{D_{KL}(\mathcal{A}(S), \pi) + \log\left(\frac{1}{\delta}\right)}{\beta n} + \frac{\beta}{8}.$$

Theorem A.12 (PAC-Bayes Bernstein)**.** *Let $\gamma = \frac{(e^\beta - 1 - \beta)}{\beta}$. For all P, with probability at least $1 - \delta$ on a draw $S \sim P^n$ with $\mathcal{A}$, π and β fixed before the draw,*

$$\ell(P, \mathcal{A}(S)) \leq \ell(S, \mathcal{A}(S)) + \frac{D_{KL}(\mathcal{A}(S), \pi) + \log\left(\frac{1}{\delta}\right)}{\beta n} + \gamma \ell^2(P, \mathcal{A}(S)).$$

For a fixed learning algorithm and β, these three bounds all hold *simultaneously*. The multiplicative Hoeffding bound is appropriate for low loss settings where the lack of *tightness* is no issue. The additive Hoeffding bound is tight however we pay with extra additive slack. Finally, the Bernstein bound is useful when we can control the variance term,

$$\ell^2(P, \mathcal{A}(S)) \leq \kappa \ell(P, \mathcal{A}(S)), \ \forall S.$$

This condition can hold in more general scenarios than the low noise or "realizable" condition needed for fast rates in the multiplicative Hoeffding bound.

A.6 The ERM Principle

The generalization bounds presented in the previous section show that to produce algorithms with low risk, one can use the following procedure. First pick a prior π, regularization parameter β and class of possible posterior distributions $\mathcal{Q} \subseteq \mathbb{P}(A)$ and take,

$$\mathcal{A}(S) = \operatorname*{arg\,min}_{Q \in \mathcal{Q}} \ell(S, Q) + \frac{D_{KL}(Q, \pi)}{\beta n}.$$

If A is finite and π is uniform, this "generalized" ERM principle reduces to the more standard ERM principle [122],

$$\mathcal{A}_{\text{ERM}}(S) = \operatorname*{arg\,min}_{a \in A} \ell(S, a).$$

If ℓ is bounded then with probability at least $1 - \delta$,

$$\ell(P, \mathcal{A}_{\text{ERM}}(S)) \leq \ell(S, \mathcal{A}_{\text{ERM}}(S)) + \frac{\log(|A|) + \log\left(\frac{1}{\delta}\right)}{\beta n} + \frac{\beta}{8}.$$

Optimizing over β yields a rate of convergence to the optimal at rate $\frac{1}{\sqrt{n}}$. More generally, the best randomized algorithm is given by,

$$\mathcal{A}(S) \propto e^{-\beta n \ell(S, -)} d\pi.$$

Actions with low loss on the sample are up-weighted accordingly (as one would expect). If ℓ is log loss, then the above distribution is the standard Bayesian posterior distribution.

It is advantageous to work with deterministic algorithms. On means of "de-randomizing" an algorithm is by taking the expected action. Suppose A is convex, ℓ is convex in a and let $\bar{Q} = \mathbb{E}_{a \sim Q} a$. Then,

$$\ell(P, \bar{Q}) \leq \ell(P, Q).$$

Therefore averaging over the posterior distribution provides a deterministic algorithm with lower loss than randomizing. If ℓ is log loss, $\bar{Q}$ is the standard Bayesian predictive distribution.

A.7 Bias Variance Trade Off

To motivate algorithms, we fixed β and π and minimized the bound. Alternately, we can fix $\mathcal{A}$ and β and find the prior that minimizes the upper bound

Theorem A.13 ([99]). *For any algorithm $\mathcal{A}$ and a distribution P, let $\pi_{\mathcal{A}}$ be the marginal distribution over actions. Then,*

$$\pi_{\mathcal{A}} = \operatorname*{arg\,min}_{\pi} \mathbb{E}_{S \sim P^n} D_{KL}(\mathcal{A}(S), \pi).$$

Furthermore, $\mathbb{E}_{S\sim P^n} D_{KL}(\mathcal{A}(S), \pi_{\mathcal{A}})$ *is the mutual information [46] between the sample and the action chosen by the algorithm.*

Algorithms with low risk are precisely those that predict the training sample well (have high bias) while using as little information from the sample as possible (low variance). One means to produce algorithms with lower variance is to treat the sample S as if it was P, sample from S with replacement generating a "bootstrapped" sample S' [30]. One repeats this procedure k times yielding k actions $a_i = \mathcal{A}(S'_i)$. Finally one aggregates these functions, perhaps by averaging, by randomizing or in the case of classification by taking a majority vote. Such *ensemble* methods have had large practical success.

Bibliography

1. Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization. *IEEE Transactions on Information Theory*, 58(5):3235, 2012.

2. Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

3. Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *The Proceedings of the 19th Annual Conference on Learning Theory (COLT06)*, pages 139–153. Springer, 2006.

4. Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

5. Javed A. Aslam and Scott E. Decatur. On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195, 1996.

6. Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

7. Bernardo Ávila Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1391–1399, 2013.

8. Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the Equivalence between Herding and Conditional Gradient Algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1359–1366, 2012.

9. Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

10. John C. Baez, Tobias Fritz, and Tom Leinster. A characterization of entropy in terms of information loss. *Entropy*, 13(11):1945–1957, 2011.

11. Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM*, 57(3):19, 2010.

12. Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

13. Arindam Banerjee. On Bayesian Bounds. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 81–88. ACM, 2006.

14. Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman Divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

15. Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *Information Theory, IEEE Transactions on*, 44(2):525–536, 1998.

16. Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

17. Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

18. Peter L. Bartlett and Ambuj Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. *The Journal of Machine Learning Research*, 8:775–790, 2007.

19. Amir Beck and Marc Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.

20. Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013.

21. Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

22. Dennis S. Bernstein. *Matrix mathematics: Theory, Facts and Formulas*. Princeton University Press, 2009.

23. David Blackwell. Comparison of experiments. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 93–102, 1951.

24. David H. Blackwell and Meyer A. Girshick. *Theory of Games and Statistical Decisions*. John Wiley and Sons, Inc., New York, 1954.

25. Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004.

26. Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

27. Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 33–42, 1998.

28. David M. Bradley and J. Andrew Bagnell. Differentiable sparse coding. *Advances in Neural Information Processing Systems*, 21:113–120, 2008.

29. Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

30. Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

31. Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

32. Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

33. Bernd Carl and Irmtraud Stephani. *Entropy, compactness, and the approximation of operators*. Cambridge University Press, 1990.

34. Olivier Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007.

35. Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning and Games*. Cambridge University Press Cambridge, 2006.

36. J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, November 1997.

37. Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2010.

38. Yutian Chen, Max Welling, and Alexander J. Smola. Super Samples from Kernel Herding. In *Uncertainty in Artificial Inteligence (UAI)*, 2010.

39. Nikolai Nikolaevich Chentsov. *Statistical decision rules and optimal inference*. American Mathematical Society, 1982.

40. Kenneth L. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms*, 6(4):63, 2010.

41. Joel E. Cohen and J. H. B. Kempermann. *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population*. Springer, 1998.

42. Alain Connes. *Noncommutative geometry year 2000*. Springer, 2010.

43. Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local Rademacher complexity. In *Advances in Neural Information Processing Systems*, pages 2760–2768, 2013.

44. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

45. Timothee Cour, Benn Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

46. Thomas M. Cover and Jay A. Thomas. *Elements of Information Theory*. Wiley, 2012.

47. Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from data of variable quality. In *Advances in Neural Information Processing Systems*, 2005.

48. I. Csisz'ar. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

49. Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

50. George B. Dantzig. Discrete-variable extremum problems. *Operations research*, 5(2):266–288, 1957.

51. A. Philip Dawid, Steffen Lauritzen, and Matthew Parry. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608, 2012.

52. A Phillip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, (April 2006):77–93, 2007.

53. Morris H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

54. Morris H. DeGroot. *Optimal statistical decisions*. John Wiley & Sons, 1970.

55. Vasil Denchev, Nan Ding, Hartmut Neven, and S. V. N. Vishwanathan. Robust Classification with Adiabatic Quantum Optimization. In *International Conference on Machine Learning (ICML)*, pages 863–870, 2012.

56. Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996.

57. Nan Ding and S. V. N. Vishwanathan. t-Logistic regression. In *Advances in Neural Information Processing Systems*, pages 514–522, 2010.

58. Roland L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability and its Applications*, 1(1):65–80, 1956.

59. John C. Duchi, Michael Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *IEEE Symposium on the Foundations of Computer Science (FOCS)*, pages 429–438. IEEE, 2013.

60. Thomas S. Ferguson. *Mathematical statistics: A decision theoretic approach*. Academic Press New York, 1967.

61. Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' Rule. In *Advances in Neural Information Processing Systems*, 2011.

62. Dario Garcia-Garcia and Robert C. Williamson. Divergences and Risks for Multiclass Experiments. In *The Proceedings of the Annual Conference on Learning Theory (COLT)*, 2012.

63. Aritra Ghosh, Naresh Manwani, and P. S. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

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

65. Thore Graepel and Ralf Herbrich. A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work. In *Advances in Neural Information Processing Systems*, volume 13, pages 224–230. MIT Press, 2001.

66. Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-sample Test. *Journal of Machine Learning Research*, 13:723–773, March 2012.

67. Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

68. Adityanand Guntuboyina. *Minimax Lower Bounds*. PhD thesis, Yale, 2011.

69. Isabelle Guyon, Ulrike Von Luxburg, and Robert C. Williamson. Clustering: Science or Art. In *NIPS 2009 Workshop on Clustering Theory*, 2009.

70. Joseph Y. Halpern. *Reasoning about uncertainty*, volume 21. MIT press Cambridge, 2003.

71. Peter Harremoës and Naftali Tishby. The information bottleneck revisited or how to choose a good distortion measure. In *IEEE International Symposium on Information Theory*, pages 566–570. IEEE, 2007.

72. Martin Hellman and Josef Raviv. Probability of error, equivocation, and the Chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970.

73. Ralf Herbrich and Robert C. Williamson. Algorithmic luckiness. *The Journal of Machine Learning Research*, 3:175–212, 2003.

74. Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

75. Peter J Huber. *Robust Statistics*. John Wiley & Sons, 1981.

76. Zakria Hussain and John Shawe-Taylor. Improved loss bounds for multiple kernel learning. In *International Conference on Artificial Intelligence and Statistics*, pages 370–377, 2011.

77. Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435, 2013.

78. Lee K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 1(20):608–613, 1992.

79. Shizuo Kakutani. Concrete representation of abstract (m)-spaces (a characterization of the space of continuous functions). *Annals of Mathematics*, pages 994–1024, 1941.

80. Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

81. Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

82. Hidetoshi Komiya. Elementary proof for Sion's minimax theorem. *Kodai mathematical journal*, 11(1):5–7, 1988.

83. Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504, 2015.

84. Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

85. Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419–1455, 1964.

86. Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer London, 2011.

87. Yann LeCun. Learning Representations: A Challenge for Learning Theory, Plenary talk COLT 2013, 2013.

88. Ralph Linsker. An Application of the Principle of Maximum Information Preservation to Linear Systems. In *Advances in Neural Information Processing Systems*. NIPS, 1989.

89. Philip M. Long and Rocco A. Servedio. Random classification noise defeats all convex potential boosters. In *Proceedings of the 25th International Conference on Machine Learningternational conference on Machine learning*, pages 608–615, 2008.

90. Roberto Lucchetti. *Convexity and well-posed problems*. Springer, 2006.

91. Gábor Lugosi and Nicolas Vayatis. On the bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, pages 30–55, 2004.

92. Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions onPattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

93. Yuly Makovoz. Random approximants and neural networks. *Journal of Approximation Theory*, 85(1):98–109, 1996.

94. Naresh Manwani and P. S. Sastry. Noise Tolerance Under Risk Minimization. *IEEE Transactions on Cybernetics*, 43(3):1146–1151, June 2013.

95. Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 1998.

96. Hamed Masnadi-Shirazi, Vijay Mahadevan, and Nuno Vasconcelos. On the design of robust classifiers for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

97. Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34:2326–2366, 2006.

98. David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, 1998.

99. David A. McAllester. A PAC-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013.

100. Aditya Menon, Brendan van Rooyen, Cheng Soon Ong, and Robert C. Williamson. Learning from Corrupted Binary Labels via Class-Probability Estimation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 125–134, 2015.

101. Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

102. Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997.

103. Novi Quadrianto and Alex J. Smola. Estimating labels from label proportions. *The Journal of Machine Learning Research*, 10:2349–2374, 2009.

104. Mark D. Reid and Robert C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

105. Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *The Journal of Machine Learning Research*, 12:731–817, 2011.

106. MD Reid and RC Williamson. Surrogate regret bounds for proper losses. *Proceedings of the 26th Annual International ...*, (Theorem 3):897–904, 2009.

107. Bernhard Schölkopf and Alex J. Smola. *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

108. Rocco A. Servedio. On PAC learning using Winnow, Perceptron, and a Perceptron-like algorithm. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 296–307, 1999.

109. John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *Information Theory, IEEE Transactions on*, 44(5):1926–1940, 1998.

110. John Shawe-Taylor and John Langford. PAC-Bayes & margins. *Advances in Neural Information Processing Systems*, 15:439, 2003.

111. David Simmons. Conditional measures and conditional expectation; Rohlin's Disintegration Theorem. *Discrete and Continuous Dynamical Systems*, 32(7):2565–2582, March 2012.

112. Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

113. Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert R. G. Lanckriet, and Bernhard Schölkopf. Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions. In *In Neural Information Processing Systems (NIPS) 2009*, pages 1750–1758, 2009.

114. Guillaume Stempfel and Liva Ralaivola. Learning SVMs from Sloppily Labeled Data. In *International Conference on Artificial Neural Networks*, volume 5768, pages 884–893. 2009.

115. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

116. Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

117. Erik Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991.

118. Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core vector machines: Fast SVM training on very large data sets. In *Journal of Machine Learning Research*, pages 363–392, 2005.

119. Tim van Erven. PAC-Bayes Mini-tutorial: A Continuous Union Bound. *arXiv preprint arXiv:1405.1580*, 2014.

120. Tim van Erven, Peter Grünwald, Mark D. Reid, and Robert C. Williamson. Mixability in statistical learning. In *Advances in Neural Information Processing Systems*, pages 1691–1699, 2012.

121. Brendan van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. Learning with Symmetric Label Noise: The Importance of Being Unhinged. *Advances in Neural Information Processing Systems*, pages 10–18, 2015.

122. Vladimir N. Vapnik. *Statistical learning theory*. Wiley, 1998.

123. Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103. ACM, 2008.

124. Dan-Virgil Voiculescu, K. J. Dykema, and Alexandru Nica. *Free Random Variables: A Noncommutative Probability Approach to Free Products with Applications to Random Matrices, Operator Algebras, and Harmonic Analysis on Free Groups*. American Mathematical Society, 1992.

125. John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

126. Abraham Wald. Statistical decision functions. *The Annals of Mathematical Statistics*, pages 165–205, 1949.

127. Meihong Wang, Fei Sha, and Michael I. Jordan. Unsupervised Kernel Dimension Reduction. In *Advances in Neural Information Processing Systems*, 2010.

128. Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

129. Robert C. Williamson. Geometry of Losses. *Proceedings of the 27th Annual Conference on Learning Theory*, pages 1078–1108, 2014.

130. Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

131. Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

132. Moshe Zakai and Jacob Ziv. A generalization of the rate-distortion theory and application. *Information Theory, New Trends and Open Problems*, pages 87–123, 1975.

133. Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *The Journal of Machine Learning Research*, 5:1225–1251, 2004.

134. Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

135. Yuchen Zhang, John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.