Journal of Machine Learning Research 18 (2018) 1-50 Submitted 6/16; Revised 6/17; Published 7/18

A Theory of Learning with Corrupted Labels

Brendan van Rooyen BRENDAN.VANROOYEN@OUTLOOK.COM

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
The Australian National University and Data61
Canberra ACT 2601, Australia

Editor: Inderjit Dhillon

Abstract

It is usual in machine learning theory to assume that the training and testing sets comprise of draws from the same distribution. This is rarely, if ever, true and one must admit the presence of corruption. There are many different types of corruption that can arise and as of yet there is no general means to compare the relative ease of learning in these settings. Such results are necessary if we are to make informed economic decisions regarding the acquisition of data.

Here we begin to develop an abstract framework for tackling these problems. We present a generic method for learning from a fixed, known, *reconstructible* corruption, along with an analyses of its statistical properties. We demonstrate the utility of our framework via concrete novel results in solving supervised learning problems wherein the labels are corrupted, such as learning with noisy labels, semi-supervised learning and learning with partial labels.

Keywords: Supervised Learning, Generalized Supervision, Decision Theory, Minimax Bounds, Data Processing, Noise

1. Introduction

The goal of supervised learning is to find,

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

where $\mathcal{F}$ is a hypothesis class of functions and ℓ is loss function. Usually the decision maker is not given direct access to P, but rather a training set comprising n iid samples $\{(x_i, y_i)\}_{i=1}^n$ from P. There are many algorithms for solving this problem, for example empirical risk minimization (ERM), and this problem is well understood.

There are many other types of data one could learn from. For example in semi-supervised learning (Chapelle et al., 2010) the decision maker is given n instance label pairs and m instances devoid of labels. In learning with noisy labels (Angluin and Laird, 1988; Kearns, 1998; Natarajan et al., 2013), the decision maker observes instance label pairs where the observed labels have been corrupted by some noise process. There are many variants including, but not limited to, learning with label proportions (Quadrianto et al., 2009), learning with partial labels (Cour et al., 2011), multiple instance learning (Maron and Lozano-Pérez, 1998) as well as combinations of the above.

Abstractly, all of these problems can be understood as corrupted learning problems, in that the "ideal" clean training set is contaminated by a fixed and in general unknown noise process. 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.

To make progress, we consider abstract prediction problems that have been contaminated by a fixed, known corruption. We demonstrate a generic method for learning from corrupted data, if the corruption is *reconstructible*. We show the utility of this framework via concrete novel results in supervised learning problems wherein the label has been corrupted, as in learning with label noise, semi-supervised learning and learning with partial labels.

We do not explicitly consider corruption of the instance nor corruption of the label that depends on the instance. While certain examples of these problems are reconstructible, they are not our interest here. We also do not consider non-reconstructible problems such as multiple instance learning/learning with label proportions.

The concrete contributions of this paper are:

- A general method for learning from corrupted labels based on a generalization of the method of unbiased estimators presented in Natarajan et al. (2013) and implicit in the earlier work of Kearns (1998) (theorems 5 and 6).
- Upper and lower bounds on the risk of learning from combinations of corrupted labels, with some analyses of their tightness (theorems 7 and 16). Our results greatly extend the state of the art of Crammer et al. (2005), both in scope and in completeness.
- Demonstration of the computational feasibility of our approach via the preservation of convexity (theorem 29).

Elements of our framework have appeared elsewhere, for example in Crammer et al. (2005); Cid-Sueiro et al. (2014); Blanchard et al. (2016); Natarajan et al. (2013). While not the complete story for *all* problems, the contributions outlined above make progress toward the final goal of informed economic decisions. The end result is a collection of general tools that allow one to *learn from* and *compare* the usefulness of various corrupted labels.

2. Basic 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 $(\mathbb{R}^X)^*$. We take the general view that a probability distribution is an element of the dual.

Definition 1 *A* probability distribution *on a set X is an element of* $(\mathbb{R}^X)^*$*, i.e. a linear function* $\langle P, - \rangle : \mathbb{R}^X \to \mathbb{R}$*, such that:*

1. $\langle P, \mathbf{1}_X \rangle = 1$.

2. If $f(x) \leq g(x),\ \forall x \in X$ *then* $\langle P, f \rangle \leq \langle P, g \rangle$.

The linear function $\langle P, - \rangle$ 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 (Rudin, 1991). If X is finite, a distribution is nothing more than a vector.

At times we use the standard prefix notation, with $\mathbb{E}_P f = \langle P, f \rangle$. Define the set of all distributions on a set X to be $\mathbb{P}(X)$. For any $x \in X$ define the point mass distribution δ_x, with $\langle \delta_x, f \rangle = f(x)$ for all functions f. 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.

3. The General Decision Problem

Consider the problem faced by a scientist in a laboratory. In front of them is a beaker, containing an unknown substance. Available to them are a myriad of experiments that can be performed to ascertain its identity. The scientist could attempt to ignite it, mix a bit of it with a 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 substance's true identity. Therefore the scientist should focus their effort on the "most informative" experiments. Of course, what is informative depends 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 flavor, 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 section is the abstract formulation of this problem. We consider the problem of how a decision maker, or scientist, uses observations from experiments to inform their actions.

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), $\|L\|_\infty = \max_{\theta,a} |L(\theta, a)|$.

Unknown to the decision maker is the *exact* value of θ. To *discover* this, the decision maker is guided by experiments. Let $\mathcal{Z}$ be the set of possible outcomes of an experiment. We will assume that $\mathcal{Z}$ has enough mathematical structure so as to make sense of the theorems. As is standard in machine learning, the reader can assume that $\mathcal{Z}$ is of the form $X \times Y$ where X is a compact subset of $\mathbb{R}^d$ and Y is a finite set. All of the key ideas can be gleaned from assuming $\mathcal{Z}$ *finite*. This places no restrictions whatsoever one the usefulness of the theory in application to computer science problems. After all, computers work perfectly well with finite state spaces. 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 modeled by a *transition*.

3.1 Transitions

Definition 2 *A* transition *from a set X to a set Y is a linear map $T : (\mathbb{R}^X)^* \to (\mathbb{R}^Y)^*$.*

While abstract in appearance, we remark that when X and Y are *finite*, a transition is nothing more than a *matrix*. In general, a transition is an integral operator. Denote the set of all transitions from X to Y by $\mathbb{T}(X, Y)$. We call a transition *Markov* if $T(\mathbb{P}(X)) \subseteq \mathbb{P}(Y)$, ie it maps distributions over X to distributions over Y. When X and Y are finite, Markov transitions are represented by column stochastic matrices.

Markov transitions constitute a modern approach to conditional probability (Chang and Pollard, 1997; Torgersen, 1991; Le Cam, 1964; Chentsov, 1982). The distribution,

$$T(x) := T(\delta_x)$$

is how the decision maker summarizes their uncertainty about Y if the true value of X is x. In fact a transition is *completely determined* by its value on point masses. Every function $\phi \in Y^X$ defines a transition with,

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

Such a transition is called *deterministic*. Transitions can be combined in *series* and in *parallel*.

For transitions $T \in \mathbb{T}(X, Y)$ and $S \in \mathbb{T}(Y, Z)$ we can define $S \circ T \in \mathbb{T}(X, Z)$ by usual function composition. If X, Y and Z are finite, then this is just matrix multiplication. If T and S are Markov, this is just iterated expectation. Intuitively, this can be seen as "marginalizing" over Y in the Markov chain,

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

Combination of transitions in series models corruptions that are performed one after another.

Transitions can be combined in *parallel*. For $\alpha, \beta \in (\mathbb{R}^X)^*$, denote the product by $\alpha \otimes \beta \in (\mathbb{R}^{X \times X})^*$. 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 $\otimes_{i=1}^k T_i(x) = T_1(x_1) \otimes \cdots \otimes T_k(x_k)$, where × denotes the Cartesian product and ⊗ denotes products of duals. 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,

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

the n-fold product of $T(x)$.

Parallel composition of transitions models performing two different experiments as well as the repeated performance of the same experiment.

3.2 Experiments and Risk

An *experiment* is a Markov transition $e \in \mathbb{T}(\Theta, \mathcal{Z})$. We call $\mathcal{Z}$ the observation space of the experiment. The distribution $e(\theta)$ summarizes 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*.

In our language, a *learning algorithm* is a Markov transition $\mathcal{A} \in \mathbb{T}(\mathcal{Z}, A)$ [1]. $\mathcal{A}(z)$ summarizes the decision makers uncertainty in which action to choose. We define the *risk*,

$$\mathrm{Risk}_L(\theta, e, \mathcal{A}) := \mathbb{E}_{z\sim e(\theta)}\mathbb{E}_{a\sim\mathcal{A}(z)}L(\theta, a) = \langle \mathcal{A} \circ e(\theta), L(\theta, -)\rangle\,.$$

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 experiments directly we use the *minimax risk*,

$$\underline{\mathrm{Risk}}_L(e) := \inf_{\mathcal{A}} \sup_{\theta} \mathrm{Risk}_L(\theta, e, \mathcal{A}),$$

3.3 The Standard Prediction Problem

The preceding framework for defining and measuring the value of different experiments was largely conceived in the field of theoretical statistics, in the works of von Neumann and Morgenstern (1947); Blackwell (1951); DeGroot (1962) and Le Cam (1964). There is a perceived tension between the goals of statistics, that is to *discover* θ, versus the goals of machine learning, that is to *predict* the outcomes of the experiment e. As we now show, this distinction is only superficial. Let $\ell : \mathcal{Z} \times A \to \mathbb{R}$ be a "predictive" loss, that measures how well the action a predicts the outcome of the experiment e. Common examples include:

	Observations $\mathcal{Z}$	Actions A	Loss ℓ
Density Estimation	$X \subseteq \mathbb{R}^d$	Model $\Theta \subseteq P(X)$	$-\log(P_\theta(x))$
Classification	$X \times \{\pm 1\}$	Function class $\mathcal{F} \subseteq \{\pm 1\}^X$	$[\![y = f(x)]\!]$
Regression	$X \times \mathbb{R}$	Function class $\mathcal{F} \subseteq \mathbb{R}^X$	$(y - f(x))^2$
Supervised Learning	$X \times Y$	Function class $\mathcal{F} \subseteq \mathbb{P}(Y)^X$	$\ell(y, f(x))$

Due to the frequency with which we take expectations we overload our notation and define,

$$\ell(P, Q) := \mathbb{E}_{z\sim P}\mathbb{E}_{a\sim Q}\ell(z, a).$$

In learning theory, $\ell(P, Q)$ as defined above is refered to as the *risk*, here we reserve the term *risk* for its more classical statistical definition. Let $\Theta = \mathbb{P}(\mathcal{Z})$, the set of all possible distributions on the observation space. The loss function $L : \mathbb{P}(\mathcal{Z}) \times \mathbb{P}(A) \to \mathbb{R}$ of interest is the *regret*,

$$L(P, Q) := \ell(P, Q) - \inf_{a\in A} \ell(P, a),$$

Let e_n be the experiment that maps each $P \in \mathbb{P}(\mathcal{Z})$ to its n-fold product P^n. Then,

$$\underline{\mathrm{Risk}}_L(e_n) = \inf_{\mathcal{A}} \sup_{P} \mathbb{E}_{S\sim P^n}\ell(P, \mathcal{A}(S)) - \inf_{a\in A} \ell(P, a),$$

[1] We adopt the common terminology of algorithm, while not addressing computational issues.

is the minimax regret, a central object in learning theory. Therefore, the *prediction* problem of machine learning can be understood in the more general language of experiments. It is these problems that are of central interest in this paper. For the remainder L will represent the regret for a loss ℓ.

4. 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 Markov transition $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$.

Definition 3 *A* corruption *is a Markov transition* $\mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$.

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 access to the expert's data, we are given data from one of their apprentices. Here T models the hypothesized link between the expert's and apprentice's data. The goal of predicting as well as the expert remains.

More concretely, we are interested in supervised learning problems wherein,

$$\mathcal{Z} = X \times Y \text{ and } \tilde{\mathcal{Z}} = X \times \tilde{Y},$$

with T corrupting only the label Y and acting as the identity on the instance X,

$$T = id_X \otimes T_Y.$$

Prominent examples include:

- **Learning with Label Noise**:

$$T_Y = \begin{pmatrix} 1-\sigma_{-1} & \sigma_1 \\ \sigma_{-1} & 1-\sigma_1 \end{pmatrix}.$$

- **Semi Supervised Learning**:

$$T_Y = \begin{pmatrix} \sigma_{-1} & 0 \\ 0 & \sigma_1 \\ 1-\sigma_{-1} & 1-\sigma_1 \end{pmatrix}$$

- **Learning with partial labels**: 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.

$$T_Y = \begin{pmatrix} (1-\sigma)^2 & 0 & 0 \\ 0 & (1-\sigma)^2 & 0 \\ 0 & 0 & (1-\sigma)^2 \\ (1-\sigma)\sigma & (1-\sigma)\sigma & 0 \\ (1-\sigma)\sigma & 0 & (1-\sigma)\sigma \\ 0 & (1-\sigma)\sigma & (1-\sigma)\sigma \\ \sigma^2 & \sigma^2 & \sigma^2 \end{pmatrix}.$$

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

$$\underline{\mathrm{Risk}}_L(\tilde{e}_n) := \min_{\mathcal{A}} \max_{P} \mathrm{Risk}_L(P, \tilde{e}_n, \mathcal{A}).$$

Note that the domain of $\mathcal{A}$ is now $\tilde{\mathcal{Z}}^n$. Ideally we wish to compare $\underline{\mathrm{Risk}}_L(\tilde{e}_n)$ with $\underline{\mathrm{Risk}}_L(e_n)$, the minimum risk of the corrupted and the clean experiments. By the general data processing theorem, $\underline{\mathrm{Risk}}_L(\tilde{e}_n) \geq \underline{\mathrm{Risk}}_L(e_n)$, 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 $\underline{\mathrm{Risk}}_L(e_n)$ exactly, let alone $\underline{\mathrm{Risk}}_L(\tilde{e}_n)$ for general corruptions. Consequently our effort for the remaining turns to upper and lower bounds of $\underline{\mathrm{Risk}}_L(\tilde{e}_n)$.

4.1 Corruption Corrected Losses

When convenient we use the shorthand $T(P) = \tilde{P}$. Natarajan et al. (2013) 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 that a transition $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is a linear map $T : (\mathbb{R}^{\mathcal{Z}})^* \to (\mathbb{R}^{\tilde{\mathcal{Z}}})^*$. Associated with *any* transition is a *dual* or *adjoint* linear map $T^* : \mathbb{R}^{\tilde{\mathcal{Z}}} \to \mathbb{R}^{\mathcal{Z}}$ with,

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

In words, T^* “pulls back” functions of the corrupted sample to functions of the clean sample. When T is a matrix, T^* is the *transpose* of T. We wish to go in the other direction, to *transfer* functions of clean samples to those of corrupted samples.

Definition 4 *A transition $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is* reconstructible *if T has a left inverse; that is there exists a transition $R \in \mathbb{T}(\tilde{\mathcal{Z}}, \mathcal{Z})$ 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. Note that R need not be Markov for Markov T. We denote the set of all reconstructible transitions by $\mathbb{T}_{\leftarrow}(\mathcal{Z}, \tilde{\mathcal{Z}})$.

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. Appendix A contains several worked examples.

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

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

the Moore-Penrose pseudo inverse of T. For *invertible* T, R is given by the standard inverse. In general it will be useful to consider other reconstructions. In particular, for the proof of theorem 29, we use reconstructions with,

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

ie those which preserve the *constant* loss function. This condition is *always* satisfied when T is invertible. More broadly, there is no loss in generality working with such 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 5 (Corruption Corrected Loss) *For all reconstructible corruptions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and loss functions $\ell : \mathcal{Z} \times A \to \mathbb{R}$ define the* corruption corrected *loss $\ell_R : \tilde{\mathcal{Z}} \times A \to \mathbb{R}$ with,*

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

Then for all distributions $P \in \mathbb{P}(\mathcal{Z})$, $\mathbb{E}_{z\sim P}\ell(z, a) = \mathbb{E}_{\tilde{z}\sim\tilde{P}}\ell_R(\tilde{z}, a)$.

4.1.1 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 $y \in Y$ from $x \in X$ with low expected loss. Given a suitable function class $\mathcal{F} \subseteq A^X$ and a loss $\ell : Y \times A \to \mathbb{R}$ one attempts to find,

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

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

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

4.1.2 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 modeled by the following corruption 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* Markov, as some of the entries of R are negative. For a loss $\ell : \{-1, 1\} \times A \to \mathbb{R}$ we have,

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

or more compactly,

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

This is equivalent to the "method of unbiased estimators" of Natarajan et al. (2013). Several examples of corruption corrected losses are given in appendix A.

4.1.3 WHEN TO APPEAL TO A SURROGATE LOSS

Many loss functions in machine learning are non convex, the zero one loss a prime example. Rather than minimizing one of these losses directly, it is common to instead minimize a convex surrogate loss function, such as the hinge loss. When learning in the presence of corruption we face a dilemma; when does one appeal to a surrogate? The method of unbiased estimators *first* appeals to a surrogate loss before correcting for noise. The method of label dependent costs (Natarajan et al., 2013) first corrects for noise before appealing to a surrogate.

4.2 Similarities with Other Frameworks

The framework here shares common points with two other broad directions in the literature, in the line of working of Cid-Sueiro et al. (2014) and Blanchard et al. (2016).

Cid-Sueiro et al. (2014); Cid-Sueiro (2012) consider the problem of learning when the corruption is only partially known. They also use transitions (stochastic matrices) and their reconstructions to construct loss functions that "correct for" noise in the labeling process. Their development closely mirrors that of section 11, with subtle differences (see section 11.3).

Blanchard et al. (2016) consider the closely related problem of learning under mutually contaminated distributions. For example, in binary classification many performance measures can be optimized with access to samples from $P_1, P_{-1} \in \mathbb{P}(X)$, the distribution over instances given a positive or negative label respectively. In practice however the samples can be *mixed*, yielding distributions,

$$\tilde{P}_{\pm 1} = (1 - \pi_{\pm 1})P_1 + \pi_{\pm 1}P_{-1}.$$

In general the mixing constants $\pi_{\pm 1}$ are unknown. Under the assumption that P_1 and P_{-1} are *mutually irreducible*, Blanchard et al. (2016) shows that one can reconstruct the clean distributions together with the mixing coefficients, providing means to estimate the corruption. Their framework has been extended to other problems (Katz-Samuels and Scott, 2016).

5. Upper Bounds for Corrupted Learning

In this section we develop upper bounds on the risk of any algorithm that learns from a corrupted sample, in terms of the *sample risk*, $\tilde{S} \sim \tilde{P}^n$. For simplicity we assume that A is finite. This assumption can be removed by PAC-Bayesian bounds (as we do in the appendix), via covering number arguments (Bartlett et al., 1997) or via more refined bounds from empirical process theory (Bartlett and Mendelson, 2006).

By an application of the PAC-Bayes bound (Zhang, 2006), one has for all algorithms $\mathcal{A} \in \mathbb{T}(\tilde{\mathcal{Z}}^n, A)$ and distributions $P \in \mathbb{P}(\mathcal{Z})$,

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

By the construction of ℓ_R, $\ell_R(\tilde{P}, \mathcal{A}(\tilde{S})) = \ell(P, \mathcal{A}(\tilde{S}))$, and the above bound yields the following theorem.

Theorem 6 *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,*

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

A similar result also holds with high probability on draws from $\tilde{P}^n$. This bound motivates the following algorithm for learning from corrupted data,

$$\mathcal{A}_{ERM}(\tilde{S}) = \operatorname*{arg\,min}_{a\in A} \ell_R(\tilde{S}, a).$$

As this algorithm minimizes the loss on the sample,

$$\mathbb{E}_{\tilde{S}\sim\tilde{P}^n} \ell_R(\tilde{S}, \mathcal{A}_{ERM}(\tilde{S})) \leq \mathbb{E}_{\tilde{S}\sim\tilde{P}^n} \ell_R(\tilde{S}, a) = \ell(P, a),\ \forall a \in A.$$

Together with theorem 6 we have,

$$\mathbb{E}_{\tilde{S}\sim\tilde{P}^n} \ell(P, \mathcal{A}_{ERM}(\tilde{S})) \leq \inf_a \ell(P, a) + \|\ell_R\|_\infty \sqrt{\frac{2\log(|A|)}{n}},\ \forall P$$

yielding,

$$\underline{\text{Risk}}_L(\tilde{e}_n) \leq \|\ell_R\|_\infty \sqrt{\frac{2\log(|A|)}{n}}.$$

Similarly,

$$\underline{\text{Risk}}_L(e_n) \leq \|\ell\|_\infty \sqrt{\frac{2\log(|A|)}{n}}.$$

Therefore, the ratio $\frac{\|\ell_R\|_\infty}{\|\ell\|_\infty}$ measures the relative difficulty of corrupted versus clean learning, as judged solely by our upper bound.

5.1 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 7 *Let $T_i \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_i)$ be a collection of k reconstructible corruptions. 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,*

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

where $K = \sqrt{\sum\limits_{i=1}^k r_i \|\ell_{R_i}\|_\infty^2}$.

A similar result also holds with high probability on draws from $\tilde{P}$. Theorem 7 is a generalization of the final bound appearing in Crammer et al. (2005) that only pertains to symmetric label noise and binary classification. Theorem 7 suggests the following means of choosing n_i examples from each of the corrupted experiments. Let c_i be the cost of acquiring data corrupted by T_i. First, choose data from the T_i with lowest $c_i \|\ell_{R_i}\|_\infty^2$ 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.

6. 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 (see Yu (1997) for a more detailed account), 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 Massart and Nédélec (2006)). In recent times it has been used to establish lower bounds for: differentially private learning (Duchi et al., 2013), learning in a distributed set up (Zhang et al., 2013), function evaluations required in convex optimization (Agarwal et al., 2012) as well as generic lower bounds in statistical estimation problems (Yang and Barron, 1999). We show how this method can be extended using the strong data processing theorem (Cohen and Kempermann, 1998) to provide a general tool for lower bounding the possible performance attainable in corrupted prediction problems.

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

6.1 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 $L : \Theta \times A \to \mathbb{R}$ a loss. We assume further that $\inf_a L(\theta, a) = 0$ for all $\theta \in \Theta$. Define the *separation* $\rho : \Theta \times \Theta \to \mathbb{R}$,

$$\rho(\theta_1, \theta_2) := \inf_a L(\theta_1, a) + L(\theta_2, a).$$

The separation measures how hard it is to act well against both θ_1 and θ_2 simultaneously. Furthermore define the *variational divergence*,

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

The variational divergence measures how hard to is to distinguish two distributions P and Q and is deeply related to binary classification.

Lemma 8 *For all experiments $e \in \mathbb{T}(\Theta, \mathcal{Z})$ and loss functions L,*

$$\underline{\mathrm{Risk}}_L(e) \geq \frac{1}{4} \sup_{\theta_1, \theta_2 \in \Theta} \rho(\theta_1, \theta_2)\left(1 - V\left(e(\theta_1), e(\theta_2)\right)\right).$$

A compact proof is given in appendix B. 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).

Lemma 8 suggests that to *rank* the difficulty of various experiments, one should work with their *Le Cam function*.

Definition 9 *Let e be an experiment and L a loss. The* Le Cam *function of e is,*

$$\mathrm{Le\ Cam}_L\left(e, \gamma\right) = \frac{1}{4} \sup_{\theta_1, \theta_2 \in \Theta} \rho(\theta_1, \theta_2)\left(1 - \gamma V\left(e(\theta_1), e(\theta_2)\right)\right), \ \forall \gamma \geq 0.$$

Lemma 8 can be restated as,

$$\underline{\mathrm{Risk}}_L(e) \geq \mathrm{Le\ Cam}_L\left(e, 1\right).$$

Replication and Rates

We wish to lower bound how the risk decreases as n, the number of times the experiment is replicated, grows. The following lemma provides a simple way to do this.

Lemma 10 *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 8 and lemma 10 yield the following.

Lemma 11 *For all experiments $e \in \mathbb{T}(\Theta, \mathcal{Z})$, loss functions L and n,*

$$\underline{\mathrm{Risk}}_L(e_n) \geq \mathrm{Le\ Cam}_L\left(e_n, 1\right) \geq \mathrm{Le\ Cam}_L\left(e, n\right).$$

Other Methods for Obtaining Minimax Lower Bounds

There are many other techniques for constructing lower bounds in terms of functions of pairwise KL divergences (Yu, 1997) as well as functions of pairwise f-divergences (Guntuboyina, 2011). Ultimately these methods replace the Variational divergence and separation in lemma 11 with a more general function of the experiment. While such methods are often required to get tighter lower bounds, all of what follows can be applied to these more intricate techniques. For the sake of conceptual clarity we proceed with Le Cam's method.

6.2 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 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*.

Definition 12 *The* Clarity *of a Markov transitions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is,*

$$\text{Clarity}(T) := \sup_{P,Q \in \mathbb{P}(\mathcal{Z})} \frac{V(T(P), T(Q))}{V(P, Q)}.$$

One has,

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

$\text{Clarity}(T)$ measures how much T corrupts. For example, if T is constant and maps all P to the same distribution, then $\text{Clarity}(T) = 0$. If T is an invertible function, then $\text{Clarity}(T) = 1$. When $\mathcal{Z}$ and $\tilde{\mathcal{Z}}$ are finite,

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

where $\Omega = \{v : \sum v_i = 0, v \neq 0\}$. Hence $\text{Clarity}(T)$ is the operator 1-norm of T when restricted to Ω. Clarity behaves as expected under composition.

Lemma 13 *For all transitions $T_1 \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_1)$ and $T_2 \in \mathbb{T}(\tilde{\mathcal{Z}}_1, \tilde{\mathcal{Z}}_2)$,*

$$\text{Clarity}(T_2 \circ T_1) \leq \text{Clarity}(T_2)\text{Clarity}(T_1) \leq \min(\text{Clarity}(T_2), \text{Clarity}(T_1)).$$

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

$\text{Clarity}(T)$ was first used by Dobrushin (1956), where it is called the coefficient of ergodicity and is used to prove rates of convergence of Markov chains to their stationary distribution.

6.3 Lower bounds Relative to the Amount of Corruption

Together with lemma 11, the Clarity leads to meaningful lower bounds that allow the comparison of different corrupted experiments.

Theorem 14 *For all experiments e, loss functions L, corruptions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and $\gamma \geq 0$,*

$$\text{Le Cam}_L\left(T \circ e, \gamma\right) \geq \text{Le Cam}_L\left(e, \text{Clarity}(T)\gamma\right).$$

The proof is a simple application of the strong data processing theorem. Together with lemma 11, the above lemma yields the following corollary.

Corollary 15 *For all experiments e, loss functions L, corruptions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and n,*

$$\underline{\text{Risk}}_L(\tilde{e}_n) \geq \text{Le Cam}_L\left(e, \text{Clarity}(T)n\right)$$

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 15 yields a lower bound of $\frac{C}{\sqrt{\mathrm{Clarity}(T)n}}$ for the corrupted problem. The *rate* at which one learns is unaffected, only the constants. A penalty of $\mathrm{Clarity}(T)$ is unavoidable no matter what learning algorithm is used.

6.4 Lower Bounds for Combinations of Corrupted Data

As in section 5.1 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 16 *Let $T_i \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_i)$, $i \in [1; k]$, be k corruptions. Let $T = \otimes_{i=1}^{k} T_i^{n_i}$ with $n = \sum_{i=1}^{k} n_k$. Then,*

$$\text{Le Cam}_L\,(T \circ e_n, \gamma) \geq \text{Le Cam}_L\left(e, \left(\sum_{i=1}^{n} n_i \mathrm{Clarity}(T_i)\right)\gamma\right).$$

Furthermore,

$$\underline{\mathrm{Risk}}_L(T \circ e_n) \geq \text{Le Cam}_L\left(e, \left(\sum_{i=1}^{n} n_i \mathrm{Clarity}(T_i)\right)\right).$$

As in section 5.1, this bound suggests means of choosing data sets via the following integer program,

$$\underset{n_1, n_2 \ldots n_k}{\arg\max} \sum_{i=1}^{k} \mathrm{Clarity}(T_i) n_i \;\text{ subject to} \sum_{i=1}^{k} c_i n_i \leq C,$$

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 (Dantzig, 1957) which admits the following near optimal greedy algorithm. First, choose data from the T_i with highest $\frac{\mathrm{Clarity}(T_i)}{c_i}$ until picking more violates the constraints. Then pick from the second highest and so on.

6.5 The Generality of Clarity

In light of section 6.1, it is often the case that more complicated lower bounding techniques based on pairwise f-divergences are required to produce *tight* lower bounds. Recall the definition of an f-divergence (Ali and Silvey, 1966).

Definition 17 *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,*

$$D_f(P, Q) := \mathbb{E}_P f\left(\frac{dQ}{dP}\right),$$

if P and Q are absolutely continuous*, and is infinite otherwise.*

Both the variational and KL divergence are examples of f divergences. Following on from the reasoning of section 6.2, we seek a $\mathrm{Clarity}_f(T)$ such that,

$$D_f(T(P), T(Q)) \leq \mathrm{Clarity}_f(T) D_f(P, Q)\; \forall P, Q, f.$$

On the surface the choice of f matters. However, the clarity as we have defined is *generic*.

Theorem 18 (Strong Data Processing(Theorem 4.1 of Cohen et al. (1993))) *Let $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ be a Markov transition. Then for all P, Q, f,*

$$D_f(T(P), T(Q)) \leq \text{Clarity}(T) D_f(P, Q).$$

7. Measuring the Tightness of the Upper Bounds and Lower Bounds

In the previous sections we have shown upper bounds that depend on $\|\ell_R\|_\infty$ as well as lower bounds that depend on $\text{Clarity}(T)$. Here we compare these bounds.

Recall from theorem 5, $\ell_R(-, a) = R^*(\ell(-, a))$. The worst case ratio $\frac{\|\ell_R\|_\infty}{\|\ell\|_\infty}$ 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 (Bernstein, 2009). Hence $\|R\|_1 = \|R^*\|_\infty$.

Lemma 19 *For all losses ℓ, $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ and reconstructions R, $\frac{\|\ell_R\|_\infty}{\|\ell\|_\infty} \leq \|R^*\|_\infty$.*

Lemma 20 *If $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is reconstructible, with reconstruction R, then,*

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

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 21 *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,*

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

Proof The first statement is obvious. For the first inequality simply use lemma 20 followed by lemma 13. The second inequality is an easy to prove property of operator norms. ■

7.1 Comparing Theorems 6 and 15

We have shown the following implication,

$$\frac{C_1}{\sqrt{n}} \leq \underline{\text{Risk}}_L(e_n) \leq \frac{C_2 \|\ell\|_\infty}{\sqrt{n}} \Rightarrow \frac{C_1}{\sqrt{\text{Clarity}(T) n}} \leq \underline{\text{Risk}}_L(\tilde{e}_n) \leq \frac{C_2 \|\ell_R\|_\infty}{\sqrt{n}},$$

for all reconstructible T. By lemma 20, in the worst case $\|\ell_R\|_\infty \geq \frac{\|\ell\|_\infty}{\text{Clarity}(T)}$. 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 $\sqrt{\mathrm{Clarity}(T)}$ 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 4.1.2, with misclassification loss ℓ_{01},

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

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 $\frac{4}{\sqrt{15}} \approx 1.03$.

7.2 Comparing Theorems 7 and 16

Let $\mathrm{Cost}(T)$ be the cost of acquiring data corrupted by T. Theorem 7 prefers corruptions with low $\|\ell_R\|_\infty^2 \mathrm{Cost}(T)$, or equivalently those with high,

$$\frac{1}{\|\ell_R\|_\infty \mathrm{Cost}(T)}$$

Theorem 16 prefers corruptions with high,

$$\frac{\mathrm{Clarity}(T)}{\mathrm{Cost}(T)} \approx \frac{\|\ell\|_\infty}{\|\ell_R\|_\infty \mathrm{Cost}(T)}.$$

In theorems 16 and 7 we have, respectively, best case and a worst case loss specific method for choosing data sets. Theorem 7 combined with 1emma 19 provides a worst case loss insensitive method for choosing data sets.

8. Corrupted Learning when Clean Learning is Fast

The contents of this paper largely solve the problem of learning from data with corrupted labels, 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 Bartlett and Mendelson (2006); van Erven et al. (2012).

Definition 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 for empirical risk minimization (ERM).

Theorem 23 *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,

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

While our lower bounding techniques will turn a lower bound of $\frac{1}{n}$ for clean learning to one of $\frac{1}{\alpha(T)n}$ for corrupted learning, it may be the case that this bound is too optimistic, there may be no algorithm that gives a $\frac{1}{n}$ rate of convergence.

A natural question to ask is when using the ERM algorithm for the loss ℓ_R converges quickly from samples drawn from $\tilde{P}$. Here we ask the simpler question: when does $(\ell_R, \tilde{P})$ satisfy the Bernstein condition?

Lemma 24 *If $(\ell_R, \tilde{P})$ satisfies the Bernstein condition with constant K then (ℓ, P) also satisfies the Bernstein condition with the same constant.*

This theorem (almost) rules out pathological behavior where ERM learns quickly from corrupted data and yet slowly for clean data. The converse of lemma 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 25 *Let $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$be a corruption and ℓ a loss. A pair (ℓ, T) are η-*compatible *if for all $z \in \mathcal{Z}$ and $a_1, a_2 \in A$,*

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

Theorem 26 *If the pair (ℓ, P) satisfies the Bernstein condition with constant K and the pair (ℓ, T) are η-compatible then $(\ell_R, \tilde{P})$ satisfies the Bernstein condition with constant ηK.*

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

Theorem 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 $\eta = \left(\frac{1+|\sigma_{-1}-\sigma_1|}{1-\sigma_{-1}-\sigma_1}\right)^2$.*

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 27 guarantees that as long as $\sigma_{-1} + \sigma_1 \neq 1$ (i.e. it is possible to learn from noisy labels), one learns at a fast rate from noisy examples.

9. Canonical Losses and Convexity

Thus far we have focused on the statistical properties of corrected loss minimization procedures. However, performing our correction may incur a computational penalty. Even if the loss ℓ is convex, there is no guarantee that the corrected loss remains so. Here we show that a large and useful class of loss functions remain convex when corrected.

Recall the constant function, $\mathbf{1}_{\mathcal{Z}}(z)$ and define,

$$\mathbf{1}_{\mathcal{Z}}^{\perp} := \left\{ v \in \mathbb{R}^{\mathcal{Z}} : \sum_{z \in \mathcal{Z}} v(z) = 0 \right\},$$

those functions that are orthogonal to $\mathbf{1}_{\mathcal{Z}}$. Proper losses and the closely related canonical losses form a large class of loss functions that provide means of attacking prediction problems.

Definition 28 *A loss $\mathcal{L} : \mathcal{Z} \times A \to \mathbb{R}$ is* canonical *if A is a convex subset of $\mathbf{1}_{\mathcal{Z}}^{\perp}$ and,*

$$\mathcal{L}(z, v) = -\langle \delta_z, v \rangle + \Psi(v),$$

for some convex function $\Psi : A \to \mathbb{R}$.

Intuitively, minimizing a canonical loss reduces to "lining up" with the average observed label, with "over confident" predictions penalized by the function Ψ. In appendix C we show that all losses are essentially canonical losses in disguise (theorem 49). Note that canonical losses split into two terms, one convex in v and unaffected by the observation, and a term linear in v. These losses are particular easy to correct for corruption, as only the linear term needs to be corrected.

Theorem 29 (Correcting Canonical Losses) *Let $\mathcal{L} : \mathcal{Z} \times A \to \mathbb{R}$ be a canonical loss. For all reconstructible corruptions $T \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ there exists a reconstruction R such that,*

$$\mathcal{L}_R(\tilde{z}, v) = -\langle \delta_{\tilde{z}}, R^*(v) \rangle + \Psi(v)$$

Furthermore R can be calculated efficiently via Gaussian elimination.

We develop the necessary representation results needed to prove this theorem in appendix C.

Theorem 29 extends results presented in Natarajan et al. (2013) concerning when losses used for learning binary classifiers remain convex when corrected for asymmetric label noise. Therefore we need not abandon the framework of convex surrogates when the corruption is known.

9.1 Comparison with Natarajan et al. (2013)

Recall in the problem of learning with label noise,

$$T = \begin{pmatrix} 1 - \sigma_{-1} & \sigma_1 \\ \sigma_{-1} & 1 - \sigma_1 \end{pmatrix},$$

where σ_y is the probability of label y being flipped. For this problem, the corrected loss is,

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

If $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$ is convex, smooth and satisfies,

$$\ell''(y, a) = \ell''(-y, a),$$

then lemma 4 of Natarajan et al. (2013) guarantees that ℓ_R is convex. Integrating twice yields,

$$\ell(y, a) = b_y + c_y v + g(a),$$

For some convex function g. Rearranging one has,

$$\ell(y, a) = \left(\frac{b_{-1} - b_1 + (c_1 - c_{-1})\, a}{2}\right) y + \underbrace{\frac{b_1 + b_{-1} + (c_1 + c_{-1})\, a}{2} + g(a)}_{\text{convex in } a}.$$

For canonical losses, corollary 50 states that any binary canonical loss is of the form $\mathcal{L} : \{\pm 1\} \times I \to \mathbb{R}$,

$$\mathcal{L}(y, v) = -yv + \Psi(v),\ v \in I,$$

where I is a convex subset of $\mathbb{R}$ and $\Psi : I \to \mathbb{R}$ a convex function. Letting,

$$\nu = -\left(\frac{b_{-1} - b_1 + (c_1 - c_{-1})\, a}{2}\right),$$

one can see that the losses considered in Natarajan et al. (2013) are affine re-parametrizations of canonical losses. Theorem 29 is therefore a generalization of lemma 4 of Natarajan et al. (2013).

9.2 Comparison with Cid-Sueiro et al. (2014)

Section 5 of Cid-Sueiro et al. (2014) provides some discussion of the convexity of corruption corrected losses. They state, without proof, that a requirement for the preservation of convexity is that the reconstruction have *non negative* entries. Theorem 29 shows that this is not the case, one can *always* preserve convexity as long as the correct reconstruction and parametrization of the loss function are used.

10. 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 lies in a subset $\mathcal{C} \subset \mathbb{T}_{\leftarrow}(\mathcal{Z}, \tilde{\mathcal{Z}})$. For example when learning classifiers under symmetric label noise (Angluin and Laird, 1988), the corruption is of the form,

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

where $\sigma \neq \frac{1}{2}$. There are three ways in which one can proceed.

If we assume access to a "gold standard" sample $S{\sim}P^n$ as well as a corrupted sample $\tilde{S}$, we can use methods akin to those in Kearns (1998). One covers the set $\mathcal{C}$ 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_{R_i} and the corrupted sample. Finally, choose the a_i that best predicts the gold standard sample. Using theorem 6, 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 and mutual irreducibility), Menon et al. (2015) surveys methods for estimating T for the problem of learning under asymmetric label noise. Blanchard et al. (2016) gives theory for these estimators. There is a growing literature extending these estimates to other problems as well as new, computational efficient estimates (Blanchard and Scott, 2014; Katz-Samuels and Scott, 2016; Ramaswamy et al., 2016). We believe these methods can be extended to general corruptions.

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

Lemma 30 *Let $T, T' \in \mathbb{T}_{\leftarrow}(\mathcal{Z}, \tilde{\mathcal{Z}})$. Then,*

$$\left\|\ell_R - \ell_{R'}\right\|_\infty \leq \left\|R - R'\right\|_1 \left\|\ell\right\|_\infty .$$

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 "invariant" to $\mathcal{C}$. This approach is explored further in section 11 and in van Rooyen et al. (2017).

11. Corruption Invariant Loss Functions

While exact knowledge of $T \in \mathcal{C} \subset \mathbb{T}_{\leftarrow}(\mathcal{Z}, \tilde{\mathcal{Z}})$ is required to estimate the *expected loss* from a corrupted distribution, in certain situations this is unnecessary for estimating optimal actions and *any* reconstruction will suffice. In this section we formalize this notion and *characterize* when exact knowledge of T is unnecessary. We assume for each $T \in \mathcal{C}$ an explicit reconstruction R, ie a function,

$$\mathrm{Rec} : \mathcal{C} \to \mathbb{T}(\tilde{\mathcal{Z}}, \mathcal{Z}),$$

with $\mathrm{Rec}(T)T = \mathrm{id}_{\mathcal{Z}}$. As in theorem 29, we will assume that,

$$\mathrm{Rec}(T)^*(\mathbf{1}_{\mathcal{Z}}) = \mathbf{1}_{\tilde{\mathcal{Z}}},\ \forall T \in \mathcal{C}.$$

We focus on the problem of prediction. Let $\ell : \mathcal{Z} \times A \to \mathbb{R}$ be a loss. ℓ provides an *ordering* on $\mathbb{P}(\mathcal{Z} \times A)$. $P_1 \leq_\ell P_2$ if,

$$\mathbb{E}_{(z,a)\sim P_1} \ell(z, a) \leq \mathbb{E}_{(z,a)\sim P_2} \ell(z, a).$$

In words, P_1 is making better decisions than P_2. One can think of P_1 and P_2 being the output of a learning algorithm, although exactly how they came to be is of no concern in this section.

Let $T \in \mathcal{C}$ be the true corruption and $T_0 \in \mathcal{C}$ the assumed corruption. By properties of adjoints,

$$\langle T(P), \mathrm{Rec}(T_0)^* \ell(-, a)\rangle = \langle P, (\mathrm{Rec}(T_0)T)^* \ell(-, a)\rangle .$$

When using the wrong reconstruction the decision maker is effectively using the loss,

$$\tilde{\ell}_T(-,a) := (\mathrm{Rec}(T_0)T)^*\ell(-,a),$$

in place of ℓ. In general there is no guarantee,

$$P_1 \leq_\ell P_2 \Leftrightarrow P_1 \leq_{\tilde{\ell}_T} P_2.$$

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

The reader may wonder why we focus on preserving *order* rather than preserving only the optimal action. Invariably in machine learning we make approximations, be it via working with a finite simple, or a restricted function class. What is key therefore is preserving what is *better* and not only what is *best*.

Definition 31 (Order Equivalence) *Let $\ell, \ell' : \mathcal{Z} \times A \to \mathbb{R}$ be loss functions. ℓ is* order equivalent *to ℓ' if for all $P_1, P_2 \in \mathbb{P}(A \times \mathcal{Z})$,*

$$P_1 \leq_\ell P_2 \Leftrightarrow P_1 \leq_{\ell'} P_2.$$

The lemma below *characterizes* when losses are order equivalent.

Lemma 32 (Theorem 2, Section 7.9 of DeGroot (1962)) *ℓ is order equivalent to ℓ' if and only if there exists a constants $\alpha > 0$ and β such that,*

$$\ell(z,a) = \alpha\ell'(z,a) + \beta,\ \forall z \in \mathcal{Z},\ \forall a \in A.$$

Definition 33 *Let $\mathcal{C} \subset \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}})$ be a set of reconstructible transitions. A loss ℓ is* immune to $\mathcal{C}$ *if for all $T \in \mathcal{C}$, ℓ is order equivalent to $\tilde{\ell}_T$.*

Losses that are immune to $\mathcal{C}$ are precisely those where exact knowledge of the corruption is unnecessary, assuming *any* $T \in \mathcal{C}$ and using its corresponding reconstruction is sufficient.

Theorem 34 *Fix $T_0 \in \mathcal{C}$. A loss ℓ is* immune to $\mathcal{C} \subset \mathbb{T}_{\leftarrow}(\mathcal{Z}, \tilde{\mathcal{Z}})$ *if and only if for all $T \in \mathcal{C}$.*

$$(\mathrm{Rec}(T_0)T)^*\ell(-,a) = \alpha(T)\ell(-,a) + \beta(T)\mathbf{1}_{\mathcal{Z}},\ \forall a \in A,$$

for functions $\alpha : \mathcal{C} \to \mathbb{R}_+$ and $\beta : \mathcal{C} \to \mathbb{R}$.

The proof follows directly from the definition and lemma 32. The operator $(\mathrm{Rec}(T_0)T)$ measures the effect of reconstructing incorrectly.

11.1 Constructing Corruption Immune Loss Functions

Theorem 34 provides means to *test* when a loss function is immune to $\mathcal{C}$. Here we show how to *construct* such losses. Immune losses arise from the *persistent eigenvectors* of $(\mathrm{Rec}(T_0)T)^*$.

Let $S \subset \mathbb{T}(\mathcal{Z}, \mathcal{Z})$. We call v a *persistent eigenvector* of S if,

$$Mv = \lambda(M)v, \ \forall M \in S,$$

ie, v is an eigenvector *for all* $M \in S$, albeit with differing eigenvalue. We call the function $\lambda : S \to \mathbb{R}$ a *persistent eigenvalue*. Much like normal eigenspaces, we define *persistent eigenspaces* as subspaces of $\mathbb{R}^{\mathcal{Z}}$ with the same persistent eigenvalue,

$$E_{\lambda,S} = \left\{v \in \mathbb{R}^{\mathcal{Z}} : Mv = \lambda(M)v, \ \forall T \in S\right\}.$$

Persistent eigenspaces provide an alternative statement of theorem 34.

Corollary 35 *Let $\mathcal{C} \subset \mathbb{T}_{\leftarrow}(\mathcal{Z}, \tilde{\mathcal{Z}})$, with $T_0 \in \mathcal{C}$. Let,*

$$S = \{(\mathrm{Rec}(T_0)T)^* : T \in \mathcal{C}\}.$$

and let $\alpha : S \to \mathbb{R}_+$ be a persistent eigenvalue of S. Any loss ℓ of the form,

$$\ell(-, a) = v(a) + \gamma \mathbf{1}_{\mathcal{Z}},$$

where $v(a) \in E_{\alpha,S}$ $\forall a \in A$ and $\gamma \in \mathbb{R}$ is immune to $\mathcal{C}$.

Therefore the search for losses that are immune to $\mathcal{C}$ reduces to the calculation of the persistent eigenspaces of $\{(\mathrm{Rec}(T_0)T)^* : T \in \mathcal{C}\}$.

11.2 Examples of Corruption Immune Losses

Here we show how to apply corollary 35 to find losses that are immune to families of corruptions.

11.2.1 Symmetric Label Noise and the Linear Loss

We proceed as in van Rooyen et al. (2017). In the problem of symmetric label noise $\mathcal{Z} = \{-1, +1\}$ with,

$$T = \begin{pmatrix} 1-\sigma & \sigma \\ \sigma & 1-\sigma \end{pmatrix} \text{ and } R = \frac{1}{1-2\sigma}\begin{pmatrix} 1-\sigma & -\sigma \\ -\sigma & 1-\sigma \end{pmatrix},$$

for $\sigma \neq \frac{1}{2}$. Let $\mathcal{C}$ be all such T with $\sigma \neq \frac{1}{2}$. Define the *linear* loss function,

$$\ell(y, v) = -yv, \ v \in \mathbb{R},$$

or in partial form,

$$\ell(-, v) = \begin{pmatrix} v \\ -v \end{pmatrix}.$$

We can quickly verify that the linear loss is immune to $\mathcal{C}$. Let $T_0 = \mathrm{id}_{\mathcal{Z}}$. As $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ is an eigenvector of T^* with eigenvalue of $1 - 2\sigma$ we have,

$$T^* \ell(-, v) = (1 - 2\sigma)\, \ell(-, v),\ \forall v \in \mathbb{R},\ \forall \sigma \neq \frac{1}{2}.$$

Therefore the linear loss is immune to $\mathcal{C}$.

11.2.2 MULTI-CLASS IMMUNE LOSSES

Here we consider the three-class generalization of linear loss with $\mathcal{Z} = \{1, 2, 3\}$ and,

$$\ell(-, v) = v, v \in \mathbf{1}_{\mathcal{Z}}^{\perp},$$

with,

$$v = v_1 \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} + v_2 \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix},\ v_1, v_2 \in \mathbb{R}.$$

We consider two classes of corruption, firstly three-class symmetric noise, secondly symmetric partial label noise.

For three-class symmetric noise,

$$T = \begin{pmatrix} 1-\sigma & \frac{\sigma}{2} & \frac{\sigma}{2} \\ \frac{\sigma}{2} & 1-\sigma & \frac{\sigma}{2} \\ \frac{\sigma}{2} & \frac{\sigma}{2} & 1-\sigma \end{pmatrix} \text{ and } R = \begin{pmatrix} \frac{2-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} \\ \frac{-\sigma}{2-3\sigma} & \frac{2-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} \\ \frac{-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} & \frac{2-\sigma}{2-3\sigma} \end{pmatrix},$$

for $\sigma \neq \frac{2}{3}$. Fix $T_0 = \mathrm{id}_{\mathcal{Z}}$. Both,

$$v = v_1 \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \text{ and } \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix},$$

are eigenvalues of T^* with eigenvalue $\frac{2-3\sigma}{2}$. Therefore,

$$T^* \ell(-, v) = \frac{2 - 3\sigma}{2} \ell(-, v),\ \forall v \in \mathbf{1}_{\mathcal{Z}}^{\perp},\ \forall \sigma \neq \frac{2}{3},$$

Hence linear loss is immune to three-class symmetric noise. Note that general losses are *not* immune to symmetric three-class label noise.

In learning under symmetric partial label noise,

$$T = \begin{pmatrix} (1-\sigma)^2 & 0 & 0 \\ 0 & (1-\sigma)^2 & 0 \\ 0 & 0 & (1-\sigma)^2 \\ (1-\sigma)\sigma & (1-\sigma)\sigma & 0 \\ (1-\sigma)\sigma & 0 & (1-\sigma)\sigma \\ 0 & (1-\sigma)\sigma & (1-\sigma)\sigma \\ \sigma^2 & \sigma^2 & \sigma^2 \end{pmatrix} \text{ and } R = \begin{pmatrix} 1 & 0 & 0 & \frac{3-2\sigma}{3-3\sigma} & \frac{3-2\sigma}{3-3\sigma} & -\frac{\sigma-3}{3(\sigma-1)} & \frac{1}{3} \\ 0 & 1 & 0 & \frac{3-2\sigma}{3-3\sigma} & -\frac{\sigma-3}{3(\sigma-1)} & \frac{3-2\sigma}{3-3\sigma} & \frac{1}{3} \\ 0 & 0 & 1 & -\frac{\sigma-3}{3(\sigma-1)} & \frac{3-2\sigma}{3-3\sigma} & \frac{3-2\sigma}{3-3\sigma} & \frac{1}{3} \end{pmatrix},$$

for $\sigma \neq 1$. Taking,

$$T_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

ie $\sigma = 0$, one has,

$$(RT_1)^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \forall \sigma \neq 1.$$

This means that *all* losses are immune to symmetric partial label noise.

11.3 Comparison with Cid-Sueiro et al. (2014)

The development here closely mirrors that of Cid-Sueiro et al. (2014) with one key exception. While we look for *losses* with,

$$(\mathrm{Rec}(T_0)T)^*\ell(-,a) \cong \ell(-,a),\ \forall T \in \mathcal{C},$$

Cid-Sueiro et al. (2014) looks for classes of losses $\ell : \mathcal{Z} \times \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}$ that are closed under the action of $(\mathrm{Rec}(T_0)T)$. In particular they consider strictly proper and classification calibrated (Bartlett et al., 2006) loss functions. They show (theorem 2) that if,

$$(\mathrm{Rec}(T_0)T)^* = \alpha(T)\mathrm{id}_{\mathcal{Z}},\ \forall T \in \mathcal{C},$$

for $\alpha(T) > 0$, then proper losses are mapped to proper losses by $(\mathrm{Rec}(T_0)T)^*$. By theorem 34 this would guarantee,

$$(\mathrm{Rec}(T_0)T)^*\ell(-,a) = \alpha(T)\ell(-,a),\ \forall T \in \mathcal{C},\ \forall a \in A,\ \forall \ell,$$

ie *all* losses are immune to $\mathcal{C}$. We saw an example of this in section 11.2.2. They also show (theorem 5) that if,

$$(\mathrm{Rec}(T_0)T)^* = \alpha(T)\mathrm{id}_{\mathcal{Z}} + v(T) \otimes \mu,\ \forall T \in \mathcal{C},$$

for $\alpha(T) > 0$ and $v(T) \in \mathbb{R}^{\mathcal{Z}}$, where μ is the uniform distribution over the outcomes, then classification calibrated losses are mapped to classification calibrated losses. This yields,

$$(\mathrm{Rec}(T_0)T)^*\ell(-,a) = \alpha(T)\ell(-,a) + \left(\sum_{z\in\mathcal{Z}} \ell(z,a)\right) v(T),\ \forall T \in \mathcal{C},\ \forall a \in A,\ \forall \ell,$$

or in expectation,

$$\langle P, (\mathrm{Rec}(T_0)T)^*\ell(-,a)\rangle = \alpha(T)\ell(P,a) + \langle P, v(T)\rangle\, \ell(\mu,a),$$

ie the effect of reconstructing incorrectly is to add a distribution dependent multiple of $\ell(\mu,a)$ to the loss.

The virtue of our approach is the identification of loss functions with *stronger* robustness properties over standard classification calibrated losses. We direct the reader to van Rooyen et al. (2017) for an in depth discussion.

12. Summary and a Guide to the Practitioner

Real world data sets are amalgamations of data of variable quality and type. Understanding how to *learn from* and *compare* different corrupted data sets is therefore a problem of great practical importance. Theorem 7, Appendix C and theorem 29 yield powerful means to do this. In concert, they provide the following framework to the practitioner:

1. Identify the appropriate loss function for the problem at hand.
2. Acquire data in accordance with theorem 7, and any relevant financial constraints.
3. Correct for any noise present in the data, and solve for,

$$\operatorname*{arg\,min}_{f\in\mathcal{F}} \sum_{i=1}^{n} \ell_{R_i}(y_i, f(x_i)).$$

As long as the decision maker is willing to work with *linear* and *kernelized* function classes, this is a convex problem with *many* available algorithmic solutions.

Acknowledgments

This work was completed while Brendan van Rooyen was with NICTA, The Australian National University, ACEMS and The Queensland University of Technology. The work was partially supported by the Australian Research Council. Many thanks go to Aditya Menon for many interesting conversations and providing some of the initial impetus for the paper. Thanks to Clayton Scott for providing detailed feedback.

Appendix A. Examples

Here we show examples of common corrupted machine learning problems.

A.1 Noisy Labels

We consider the problem of learning from noisy binary labels (Angluin and Laird, 1988; Natarajan et al., 2013). Here σ_i is the probability that class i is has its label flipped. We have,

$$T = \begin{pmatrix} 1-\sigma_{-1} & \sigma_1 \\ \sigma_{-1} & 1-\sigma_1 \end{pmatrix} \text{ and } R = \frac{1}{1-\sigma_{-1}-\sigma_1} \begin{pmatrix} 1-\sigma_1 & -\sigma_1 \\ -\sigma_{-1} & 1-\sigma_{-1} \end{pmatrix},$$

This yields,

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

The above loss is lemma 1 in Natarajan et al. (2013). Interestingly, even if ℓ is positive, ℓ_R can be negative. If the noise is symmetric with $\sigma_{-1} = \sigma_1 = \sigma$ and ℓ is 01 loss then,

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

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 (Audibert and Tsybakov, 2007). 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 and Nédélec (2006), can be transferred to the setting of learning with symmetric label noise. Our lower bound reproduces the results of Massart and Nédélec (2006).

Below is a table of the relevant parameters for learning with noisy binary labels. These results directly extend those presented in Kearns (1998) 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^*	$\frac{1}{1-\sigma_{-1}-\sigma_1} \begin{pmatrix} 1-\sigma_1 & -\sigma_1 \\ -\sigma_{-1} & 1-\sigma_{-1} \end{pmatrix}$
Clarity(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. $\|R^*\|_\infty$ tends to be marginally greater than $\frac{1}{\alpha(T)}$, with $\|\ell_{01,T}\|_\infty$ less than both. In the symmetric case our lower bound reproduces that of Aslam and Decatur (1996).

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

$$\|R_\sigma - R_{\sigma'}\|_1 = \frac{2|\sigma - \sigma'|}{|1 - 2\sigma||1 - 2\sigma'|}.$$

For fixed true noise rate σ, the presence of a factor $|1-2\sigma'|$ 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 10, more focus should be given to higher noise rates than to lower.

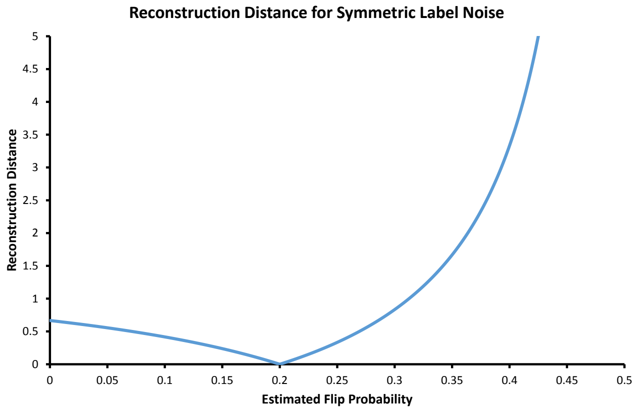

Figure 1: Plot of $\|R_\sigma - R_{\sigma'}\|_1$ for $\sigma = 0.2$. $\|R_\sigma - R_{\sigma'}\|_1$ is a measure of how far apart two corruptions are. This distance measure can be used when constructing estimators for the corruption process T. See text.

A.2 Semi-Supervised Learning

We consider the problem of semi-supervised learning (Chapelle et al., 2010). Here $1 - \sigma_i$ is the probability class i has a missing label. We consider the 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^*	$\begin{pmatrix} \frac{1-2\sigma+2\sigma^2}{1-3\sigma+5\sigma^2-3\sigma^3} & \frac{-\sigma^2}{1-3\sigma+5\sigma^2-3\sigma^3} \\ \frac{-\sigma^2}{1-3\sigma+5\sigma^2-3\sigma^3} & \frac{1-2\sigma+2\sigma^2}{1-3\sigma+5\sigma^2-3\sigma^3} \\ \frac{\sigma}{1-2\sigma+3\sigma^2} & \frac{\sigma}{1-2\sigma+3\sigma^2} \end{pmatrix}$
$\mathrm{Clarity}(T)$	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 $\|\ell_{01,T}\|_\infty \leq \frac{1}{\mathrm{Clarity}(T)}$. Our lower bound confirms that in general unlabelled data does not help (Balcan and Blum, 2010). 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 (Balcan and Blum, 2010), 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.

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

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

A.4 Partial Labels

Here we follow Cour et al. (2011) 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	$\begin{pmatrix} (1-\sigma)^2 & 0 & 0 \\ 0 & (1-\sigma)^2 & 0 \\ 0 & 0 & (1-\sigma)^2 \\ (1-\sigma)\sigma & (1-\sigma)\sigma & 0 \\ (1-\sigma)\sigma & 0 & (1-\sigma)\sigma \\ 0 & (1-\sigma)\sigma & (1-\sigma)\sigma \\ \sigma^2 & \sigma^2 & \sigma^2 \end{pmatrix}$
R	$\begin{pmatrix} 1 & 0 & 0 & \frac{3-2\sigma}{3-3\sigma} & \frac{3-2\sigma}{3-3\sigma} & -\frac{\sigma-3}{3(\sigma-1)} & \frac{1}{3} \\ 0 & 1 & 0 & \frac{3-2\sigma}{3-3\sigma} & -\frac{\sigma-3}{3(\sigma-1)} & \frac{3-2\sigma}{3-3\sigma} & \frac{1}{3} \\ 0 & 0 & 1 & -\frac{\sigma-3}{3(\sigma-1)} & \frac{3-2\sigma}{3-3\sigma} & \frac{3-2\sigma}{3-3\sigma} & \frac{1}{3} \end{pmatrix}$
$\mathrm{Clarity}(T)$	$1-\sigma$
$\lVert R^* \rVert_\infty$	$\frac{\lvert 3-\sigma\rvert+2\lvert 3-2\sigma\rvert}{3\lvert 1-\sigma\rvert}$
$\lVert \ell_{01,T} \rVert_\infty$	$\frac{6-4\sigma}{3-3\sigma}$

We see that as long as $\sigma \neq 1$, T is reconstructible. In this case $\|\ell_{01,T}\|_\infty$ and $\|R^*\|_\infty$ 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.

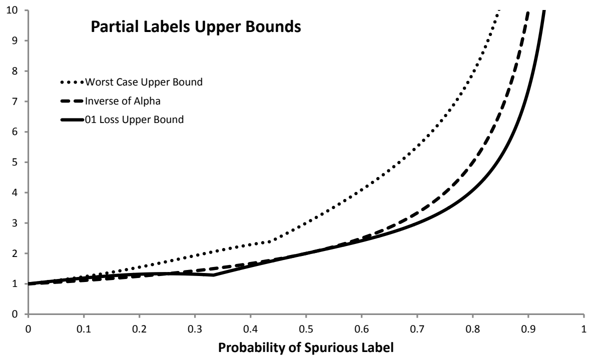

Figure 2: Upper and lower bounds for the problem of learning from partial labels, see text.

Appendix B. Le Cam's Method and Minimax Lower Bounds

The development here closely follows Duchi et al. (2013) with some streamlining. We consider a general decision problem with unknowns Θ, observation space $\mathcal{Z}$ and loss $L : \Theta \times A \to \mathbb{R}$, with $\inf_a L(\theta, a) = 0$. For any learning algorithm $\mathcal{A} \in \mathbb{T}(\mathcal{Z}, A)$ we wish to lower bound the minimax risk,

$$\underline{\text{Risk}}_L(e) = \inf_{\mathcal{A}} \sup_{\theta} \mathbb{E}_{z\sim e(\theta)} L(\theta, \mathcal{A}(z)).$$

The method proceeds by reducing a general decision problem to an easier binary classification problem, by considering a supremum of the risk 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,

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

Recall the *separation* $\rho : \Theta\times\Theta \to \mathbb{R}$, $\rho(\theta_1, \theta_2) = \inf_a L(\theta_1, a) + L(\theta_2, a)$. 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 $L(\theta_i, a) < \delta$ and error otherwise. This function is well defined as if there exists an action a with $L(\theta_1, a) < \delta$ and $L(\theta_2, a) < \delta$ 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,

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

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 (see Reid and Williamson (2011) for a modern treatment). Taking $\delta = \frac{\rho(\theta_1,\theta_2)}{2}$ yields lemma 8.

B.1 Extension of Le Cam's Method to Bayesian Risk

Rather than lower bounding $\sup_\theta \mathbb{E}_{z\sim e(\theta)} L(\theta, \mathcal{A}(z))$, 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,

$$\underline{\text{Risk}}_L^\pi(e) := \inf_{\mathcal{A}} \mathbb{E}_{\theta\sim\pi} \mathbb{E}_{z\sim e(\theta)} L(\theta, \mathcal{A}(z)).$$

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

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

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,

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

This insight leads to a Bayesian version of lemma 8.

Lemma 36 *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,*

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

Using this in place of lemma 8 leads to Bayesian versions of theorems 15 and 16.

Appendix C. Canonical Loss Functions and their Convexification

Here we develop *general* representations of loss functions $L : \Theta \times A \to \mathbb{R}$. We assume that the set Θ is finite.

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 37 *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.*

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 (Brier, 1950; Grünwald and Dawid, 2004; Zhang, 2004; Gneiting and Raftery, 2007; Dawid, 2007; Reid and Williamson, 2009a; Dawid, 2007; Ávila Pires et al., 2013).

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 (Boyd and Vandenberghe, 2004; Lucchetti, 2006) to aid in calculating optimal actions.

C.1 Entropy from Loss

Rather than working with probability distributions, we take the route of Williamson (2014) and work with un-normalized distributions. Denote the set of all unnormalized distributions on Θ by $\mathbb{P}^+(\Theta)$. 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* (DeGroot, 1962; Williamson, 2014). It is concave and 1-homogeneous.

Definition 38 *A function $f : \mathbb{P}^+(\Theta) \to \mathbb{R}$ is 1-homogeneous if for all $x \in \mathbb{P}^+(\Theta)$ and for all $\lambda > 0$,*

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

C.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.

C.2.1 Bayes Actions and Super-gradients

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

$$A_P = \operatorname*{arg\,min}_{a \in A} \langle P, L(-, a) \rangle .$$

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

Definition 39 (Super-gradient of a concave function) *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 (Lucchetti, 2006). 1-homogeneous functions afford a very simple representation via their super-gradients.

Theorem 40 (Generalized Euler's Homogeneous Function Theorem) *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 41 *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 42 *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 non-Bayes 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 43 (Canonical Loss (Preliminary)) *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, $\mathcal{L}(\theta, \zeta) = \zeta(\theta)$.*

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.

C.2.2 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. However, they gain no benefit from these extra actions. From any entropy define the *Bayes super prediction set*,

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

By the definition,

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

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

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

with equality holding for at least one μ.

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

Canonical losses use *all* super gradients of $\underline{L}$. Proper losses use some.

Corollary 45 (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.

C.3 Convexification of Losses in Canonical Form

The preceding shows how to *construct* losses, we begin with a concave 1-homogeneous function and 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. We closely follow Dawid (2007), with a focus on canonical losses. This streamlines the development. For example, for some proper losses lemma 47 fails to hold. Furthermore, our result on convexification of canonical losses (theorem 49), 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 46 *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 $\mathbf{1}_\Theta^\perp$. Define,

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

By lemma 46 $\hat{\Gamma}_{\underline{L}}$ is in 1-1 correspondence with $\partial \underline{L}$.

Lemma 47 *$\hat{\Gamma}_{\underline{L}}$ is a convex set.*

Proof To show $\hat{\Gamma}_{\underline{L}}$ is convex, we are required to show that for all $\zeta_1, \zeta_2 \in \partial \underline{L}$ and all $\lambda \in [0, 1]$ there is a constant γ such that,

$$\lambda \zeta_1 + (1 - \lambda) \zeta_2 - \gamma \mathbf{1}_\Theta \in \partial \underline{L}.$$

By lemma 44, this is equivalent to,

$$\underbrace{\lambda \langle P, \zeta_1 \rangle + (1 - \lambda) \langle P, \zeta_2 \rangle - \underline{L}(P)}_{\gamma(P)} - \gamma = \gamma(P) - \gamma \geq 0,\ \forall P \in \mathbb{P}(\Theta),$$

with equality holding for one P. Let $\gamma^* = \min_P \gamma(P)$, with P^* the distribution that achieves the minimum. Clearly $\gamma(P) - \gamma^* \geq 0$. Therefore,

$$\lambda \langle P, \zeta_1 \rangle + (1 - \lambda) \langle P, \zeta_2 \rangle - \gamma^* \geq \underline{L}(P),\ \forall P \in \mathbb{P}(\Theta),$$

with equality for P^*. Therefore by lemma 44, $\lambda \zeta_1 + (1 - \lambda) \zeta_2 - \gamma^* \mathbf{1}_\Theta \in \partial \underline{L}$.

■

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 46, Ψ is well defined.

Lemma 48 *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,

$$\begin{aligned}
\zeta_1 &= v_1 + \Psi(v_1)\mathbf{1}_\Theta \\
\zeta_2 &= v_2 + \Psi(v_2)\mathbf{1}_\Theta \\
\zeta_\lambda &= \lambda v_1 + (1-\lambda)v_2 + \Psi(\lambda v_1 + (1-\lambda)v_2)\mathbf{1}_\Theta,
\end{aligned}$$

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

$$\langle P_\lambda, \zeta_\lambda \rangle \leq \langle P_\lambda, \zeta \rangle, \ \forall \zeta \in \partial \underline{L}.$$

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,

$$\langle P_{\lambda^*}, \lambda^* \zeta_1 + (1-\lambda^*)\zeta_2 \rangle < \langle P_{\lambda^*}, \zeta_{\lambda^*} \rangle,$$

a contradiction. ■

This gives the following representation theorem for canonical losses.

Theorem 49 (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,*

$$\mathcal{L}(\theta, v) = -\langle \delta_\theta, v \rangle + \Psi(v),$$

for a convex function Ψ.

The situation is particularly simple for binary problems.

Corollary 50 *Let $\Theta = \{\pm 1\}$. Then all canonical losses can be written as $\mathcal{L} : \{\pm 1\} \times I \to \mathbb{R}$,*

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

where I is a convex subset of $\mathbb{R}$ and $\Psi : I \to \mathbb{R}$ a convex function.

The proof comes from the observation that,

$$\mathbf{1}_\Theta^\perp = \operatorname{Span}((-1, 1)).$$

C.4 Illustrative Example: Binary Decisions and Square Loss

In binary problems, $\Theta = \{\pm 1\}$. An often used loss is the square or Brier loss. $L(\theta, \hat{\theta}) = \left(\theta - \hat{\theta}\right)^2$, where $\hat{\theta} \in [-1, 1]$. We plot this loss in figure 3. The partial losses are given by the red curve, the super prediction set in gray. The loss on negatives is plotted on the x-axis. In figure 4 we show geometrically how to produce canonical coordinates.

We seek to decompose,

$$L(-, \hat{\theta}) = \left((-1-\hat{\theta})^2, (1-\hat{\theta})^2\right) = \alpha \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Solving for α and β in terms of $\hat{\theta}$ gives,

$$\alpha = 2\hat{\theta} \text{ and } \beta = 1 + \hat{\theta}^2.$$

These equations can be easily solved for $\hat{\theta}$ giving,

$$\beta = \Psi(\alpha) = 1 + \frac{\alpha^2}{4}.$$

Therefore the canonical form of square loss is,

$$\mathcal{L}(\theta, \alpha) = -\theta\alpha + 1 + \frac{\alpha^2}{4}.$$

Notice that the only dependence on θ is via the "linear" term $-\theta\alpha$.

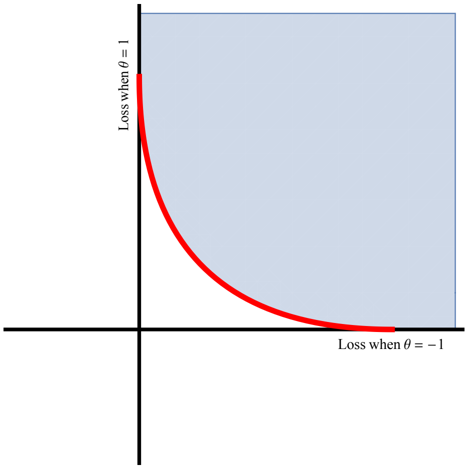

Figure 3: Plot of super prediction set and its lower boundary for square loss, see text.

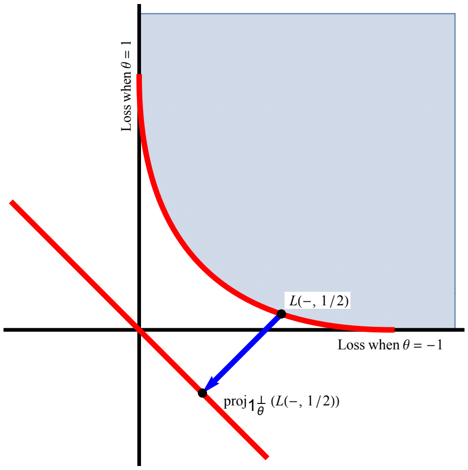

Figure 4: Construction of canonical coordinates for square loss, see text.

C.5 Illustrative Example: Hinge Loss

Hinge loss $L : \{-1, 1\} \times \mathbb{R} \to \mathbb{R}_+$,

$$L(\theta, \nu) = \max(0, 1 - \theta\nu),$$

is an often used, non-differentiable, loss function. Its minimization forms the basis of the support vector machine learning algorithm (Cortes and Vapnik, 1995; Schölkopf and Smola, 2002; Steinwart and Christmann, 2008), a popular method for learning classifiers. Here we show that the canonical form of hinge loss is the *linear loss*,

$$\mathcal{L}(\theta, \nu) = -\theta\nu,\ \nu \in [-1, 1].$$

Figure 5 is a plot of hinge loss and its super prediction set. Shown in orange are actions with $|\nu| \geq 1$. These actions are *inadmissible* and therefore not Bayes for any distribution over Θ. Figure 6 shows the projection of the Bayes actions for hinge loss unto $\mathbf{1}_\Theta^\perp$. As can be seen from the figure, linear loss is the result of "canonizing" hinge loss.

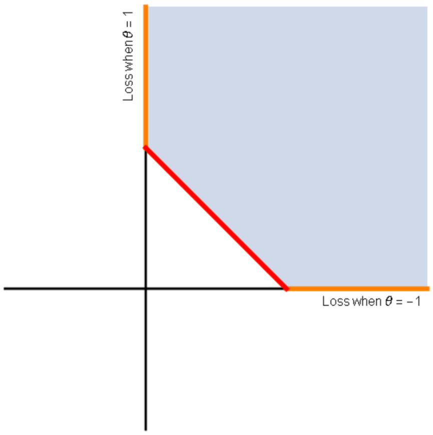

Figure 5: Plot of super prediction set and its lower boundary for hinge loss, see text (best viewed in color).

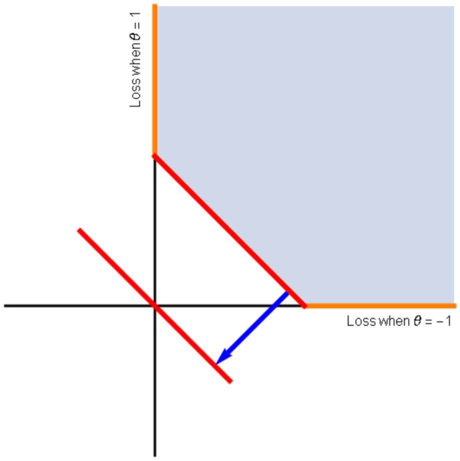

Figure 6: Construction of canonical coordinates for square loss, see text.

C.6 Proof of Theorem 29

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,

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

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 $v \in \mathbf{1}^{\perp}_{\mathcal{Z}}$,

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

Therefore $T(\mathbf{1}^{\perp}_{\mathcal{Z}}) \subseteq \mathbf{1}^{\perp}_{\tilde{\mathcal{Z}}}$. As left inverses are not unique, we can further restrict R to those with $R(\mathbf{1}^{\perp}_{\tilde{\mathcal{Z}}}) \subseteq \mathbf{1}^{\perp}_{\mathcal{Z}}$, 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 $\mathbf{1}^{\perp}_{\mathcal{Z}}$ is also left invertible. Furthermore, $T(\mathbf{1}_{\mathcal{Z}}) \notin \mathbf{1}^{\perp}_{\tilde{\mathcal{Z}}}$, as $T(\mathbf{1}_{\mathcal{Z}})$ nonnegative entries. Therefore, we can take R restricted to $\mathbf{1}^{\perp}_{\tilde{\mathcal{Z}}}$ to be the left inverse of T restricted to $\mathbf{1}^{\perp}_{\mathcal{Z}}$, 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, $\mathcal{L}_R(\tilde{z}, v) = \langle \delta_{\tilde{z}}, \mathcal{L}_R(-, v) \rangle$, yielding,

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

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.

■

C.7 Illustrative Example: Multi-class Logistic Loss and its Corrections

Here we show by example how to apply theorem 29 to construct losses for multiclass corrupted learning problems. We take $\Theta = \{1, 2, 3\}$, with L the log loss,

$$L(-, P) = (-\log(P_1), -\log(P_2), -\log(P_3))\,.$$

We express $v \in \mathbf{1}_\Theta^\perp$ as,

$$v = v_1 \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} + v_2 \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix}.$$

Note that these basis vectors are the projections onto $\mathrm{id}_\Theta^\perp$ of the partial losses,

$$\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},$$

respectively. We have,

$$P = \left(\frac{e^{v_1}}{1 + e^{v_1} + e^{v_2}}, \frac{e^{v_2}}{1 + e^{v_1} + e^{v_2}}, \frac{1}{1 + e^{v_1} + e^{v_2}} \right).$$

with,

$$\begin{aligned} \mathcal{L}(-, v) &= \begin{pmatrix} -v_1 + \log(1 + e^{v_1} + e^{v_2}) \\ -v_2 + \log(1 + e^{v_1} + e^{v_2}) \\ \log(1 + e^{v_1} + e^{v_2}) \end{pmatrix} && (1) \\ &= v + \underbrace{\left(-\frac{1}{3}(v1 + v2) + \log(1 + e^{v_1} + e^{v_2}) \right)}_{\Psi(v)} \mathbf{1}_\Theta. \end{aligned}$$

(1) is the usual form of multiclass logistic loss. For the case of symmetric noise,

$$T = \begin{pmatrix} 1-\sigma & \frac{\sigma}{2} & \frac{\sigma}{2} \\ \frac{\sigma}{2} & 1-\sigma & \frac{\sigma}{2} \\ \frac{\sigma}{2} & \frac{\sigma}{2} & 1-\sigma \end{pmatrix} \text{ and } R = \begin{pmatrix} \frac{2-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} \\ \frac{-\sigma}{2-3\sigma} & \frac{2-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} \\ \frac{-\sigma}{2-3\sigma} & \frac{-\sigma}{2-3\sigma} & \frac{2-\sigma}{2-3\sigma} \end{pmatrix},$$

respectively. All $v \in \mathbf{1}_\Theta^\perp$ are eigenvalues of R, with eigenvalue $\frac{2}{2-3\sigma}$, giving a corruption corrected loss of,

$$\mathcal{L}_R(-, v) = \frac{2}{2 - 3\sigma} v + \Psi(v)\mathrm{id}_\Theta.$$

Recall for the problem of learning partial labels $\tilde{\Theta} = \{0, 1\}^\Theta$ the set of all partial labellings. Under the assumption that the partial label always includes the correct underlying label, and that spurious labels are added with probability σ,

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

For this problem, there are several different alternatives for R. The reconstruction,

$$R = \begin{pmatrix} 1 & 0 & 0 & \frac{3-2\sigma}{3-3\sigma} & \frac{3-2\sigma}{3-3\sigma} & \frac{\sigma-3}{3(1-\sigma)} & \frac{1}{3} \\ 0 & 1 & 0 & \frac{3-2\sigma}{3-3\sigma} & \frac{\sigma-3}{3(1-\sigma)} & \frac{3-2\sigma}{3-3\sigma} & \frac{1}{3} \\ 0 & 0 & 1 & \frac{\sigma-3}{3(1-\sigma)} & \frac{3-2\sigma}{3-3\sigma} & \frac{3-2\sigma}{3-3\sigma} & \frac{1}{3} \end{pmatrix}, \ \sigma \neq 1,$$

is a left inverse for T and satisfies the requirements of theorem 29. For this reconstruction one has,

$$R^* \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{\sigma-2}{3(1-\sigma)} \\ \frac{\sigma-2}{3(1-\sigma)} \\ \frac{2(2-\sigma)}{3(1-\sigma)} \\ 0 \end{pmatrix} \text{ and } R^* \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \\ \frac{\sigma-2}{3(1-\sigma)} \\ \frac{2(2-\sigma)}{3(1-\sigma)} \\ \frac{\sigma-2}{3(1-\sigma)} \\ 0 \end{pmatrix},$$

leading to the loss,

$$\mathcal{L}_R(-, v) = v_1 \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{\sigma-2}{3(1-\sigma)} \\ \frac{\sigma-2}{3(1-\sigma)} \\ \frac{2(2-\sigma)}{3(1-\sigma)} \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \\ \frac{\sigma-2}{3(1-\sigma)} \\ \frac{2(2-\sigma)}{3(1-\sigma)} \\ \frac{\sigma-2}{3(1-\sigma)} \\ 0 \end{pmatrix} + \Psi(v)\mathrm{id}_{\tilde{\Theta}}.$$

Appendix D. Proofs of Theorems in Main Text

D.1 Proof of Theorem 7

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

Theorem 51 *Let $T_i \in \mathbb{T}(\mathcal{Z}, \tilde{\mathcal{Z}}_i)$ be a collection of k reconstructible corruptions. 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,*

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

where $K = \sqrt{\sum\limits_{i=1}^k r_i \|\ell_{R_i}\|_\infty^2}$.

Proof Define $L(\tilde{S}, a) = \sum_{i=1}^k \sum_{\tilde{z}\in\tilde{S}_i} \ell_{R_i}(z_i, a)$, the sum of the corrupted losses on the sample. By theorem 2.1 of Zhang (2006) for $\beta > 0$,

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

where the first line follows from the theorem and the second from properties of the cumulant generating function. Invoking lemma A.1 of Cesa-Bianchi and Lugosi (2006) yields,

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

As the T_i are reconstructible,

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

Optimizing over β yields the desired result. ■

Theorem 7 is recovered by taking A finite, π uniform on A and upper bounding $D_{KL}(\mathcal{A}(S), \pi) \leq \log(|A|)$.

D.2 Proof of Lemma 10

Proof Firstly V is a *metric* on $\mathbb{P}(\times_{n=1}^k \mathcal{Z}_i)$ (Reid and Williamson, 2009b). 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. ■

D.3 Proof of Theorem 16

Proof Let

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

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 10,

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

Rearranging gives,

$$\rho(\theta_1, \theta_2) \left(1 - \gamma V(T(e_n(\theta_1), T(e_n(\theta_2)))\right) \geq \rho(\theta_1, \theta_2) \left(1 - K\gamma V(T(e(\theta_1), T(e(\theta_2)))\right), \ \forall \gamma > 0.$$

Taking supremum's over θ_1, θ_2 yields,

$$\text{Le Cam}_L \left(T \circ e_n, \gamma\right) \geq \text{Le Cam}_L \left(e, K\gamma\right)$$

Applying lemma 11 yields the result. ■

D.4 Proof of Lemma 13

Proof

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

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 $\text{Clarity}(T) \leq 1$. ■

D.5 Proof of Lemma 19

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 $\|R^*\|_\infty$. ■

D.6 Proof of Lemma 20

Proof Firstly $\|R\|_1 = \|R^*\|_\infty$ (Bernstein, 2009). 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,

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

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

■

D.7 Proof of Theorem 23

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 2.1 of Zhang (2006), together with a lower bound presented in the appendix, 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.

■

D.8 Proof of Lemma 24

Proof

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

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

■

D.9 Proof of Theorem 26

Proof

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

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_R.

■

D.10 Proof of Theorem 27

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 maxima and finally expectations yields the desired result. ■

D.11 Proof of Corollary 35

Proof Let,

$$\ell(-, a) = v(a) + \gamma \mathbf{1}_{\mathcal{Z}},$$

where $v(a) \in E_{\alpha,S}\ \forall a \in A$ and $\gamma \in \mathbb{R}$. Therefore for all $M \in S$ and $\forall a \in A$,

$$\begin{aligned}M\ell(-, a) &= Mv(a) + \gamma M\mathbf{1}_{\mathcal{Z}} \\ &= \alpha(M)v(a) + \gamma \mathbf{1}_{\mathcal{Z}} \\ &= \underbrace{\alpha(M)\left(v(a) + \gamma \mathbf{1}_{\mathcal{Z}}\right)}_{\alpha(M)\ell(-,a)} + \underbrace{\left(\gamma - \gamma\alpha(M)\right)\mathbf{1}_{\mathcal{Z}}}_{\beta(M)\mathbf{1}_{\mathcal{Z}}},\end{aligned}$$

where the second line follows as by assumption $M\mathbf{1}_{\mathcal{Z}} = \mathbf{1}_{\mathcal{Z}}$ for all $M \in S$. ■

References

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.

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.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Javed A. Aslam and Scott E. Decatur. On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195, 1996.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

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.

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM*, 57(3):19, 2010.

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

Peter L. Bartlett, Sanjeev R. Kulkarni, and S. Eli Posner. Covering numbers for real-valued function classes. *IEEE transactions on information theory*, 43(5):1721–1724, 1997.

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.

Dennis S. Bernstein. *Matrix mathematics: Theory, Facts and Formulas*. Princeton University Press, 2009.

David Blackwell. Comparison of experiments. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 93–102, 1951.

Gilles Blanchard and Clayton Scott. Decontamination of Multually Contaminated Models. In *AISTATS*, 2014.

Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10 (2):2780–2824, 2016.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning and Games*. Cambridge University Press Cambridge, 2006.

Joseph T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3): 287–317, November 1997.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2010.

Nikolai Nikolaevich Chentsov. *Statistical decision rules and optimal inference*. American Mathematical Society, 1982.

J Cid-Sueiro. Proper losses for learning from partial labels. *Advances in Neural Information Processing System*, pages 1–9, 2012.

Jesús Cid-Sueiro, Darío García-García, and Raúl Santos-Rodríguez. Consistency of Losses for Learning from Weak Labels. In *Machine Learning and Knowledge Discovery in Databases*, pages 197–210. Springer, 2014.

Joel E. Cohen and Johannes H. B. Kempermann. *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population Sciences*. Springer, 1998.

Joel E Cohen, Yoh Iwasa, Gh Rautu, Mary Beth Ruskai, Eugene Seneta, and Gh Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear algebra and its applications*, 179:211–235, 1993.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Timothee Cour, Benn Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from data of variable quality. In *Advances in Neural Information Processing Systems*, pages 219–226, 2005.

George B. Dantzig. Discrete-variable extremum problems. *Operations research*, 5(2):266–288, 1957.

A. Phillip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, (April 2006):77–93, 2007.

Morris H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

Roland L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability and its Applications*, 1(1):65–80, 1956.

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.

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

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.

Adityanand Guntuboyina. *Minimax Lower Bounds*. PhD thesis, Yale, 2011.

Julian Katz-Samuels and Clayton Scott. A mutual contamination analysis of mixed membership and partial label models. *arXiv preprint arXiv:1602.06235*, 2016.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45 (6):983–1006, 1998.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419–1455, 1964.

Roberto Lucchetti. *Convexity and well-posed problems*. Springer, 2006.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576, 1998.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34:2326–2366, 2006.

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.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep D Ravikumar, and Ambuj Tewari. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1196–1204, 2013.

Novi Quadrianto, Alex J Smola, Tibério S Caetano, and Quoc V Le. Estimating Labels from Label Proportions. *Journal of Machine Learning Research*, 10:2349–2374, December 2009.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International Conference on Machine Learning*, pages 2052–2060, 2016.

Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 897–904, 2009a.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *The Journal of Machine Learning Research*, 12:731–817, 2011.

MD Reid and RC Williamson. Generalised pinsker inequalities. *arXiv preprint arXiv:0906.1244*, 2009b. URL http://arxiv.org/abs/0906.1244.

Walter Rudin. *Functional Analysis*. McGraw-Hill, 1991.

Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels*, volume 129. MIT Press, 2002.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Erik Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991.

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.

Brendan van Rooyen, Aditya K. Menon, and Bob Williamson. An average classification algorithm. http://www.mlunhinged.online/Content/AnAverageClassificationAlgorithm.pdf, 2017.

John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

Robert C. Williamson. Geometry of Losses. *Proceedings of the 27th Annual Conference on Learning Theory*, pages 1078–1108, 2014.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *The Journal of Machine Learning Research*, 5:1225–1251, 2004.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

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.