# Learning from Label Proportions: A Mutual Contamination Framework

Clayton Scott and Jianxin Zhang
Electrical Engineering and Computer Science
University of Michigan

June 15, 2020

**Abstract**

Learning from label proportions (LLP) is a weakly supervised setting for classification in which unlabeled training instances are grouped into bags, and each bag is annotated with the proportion of each class occurring in that bag. Prior work on LLP has yet to establish a consistent learning procedure, nor does there exist a theoretically justified, general purpose training criterion. In this work we address these two issues by posing LLP in terms of mutual contamination models (MCMs), which have recently been applied successfully to study various other weak supervision settings. In the process, we establish several novel technical results for MCMs, including unbiased losses and generalization error bounds under non-iid sampling plans. We also point out the limitations of a common experimental setting for LLP, and propose a new one based on our MCM framework.

## 1 Introduction

Learning from label proportions (LLP) is a weak supervision setting for classification. In this problem, training data come in the form of bags. Each bag contains unlabeled instances and is annotated with the proportion of instances arising from each class. Various methods for LLP have been developed, including those based on support vector machines and related models [32, 44, 43, 30, 9, 19, 36], Bayesian and graphical models [18, 14, 40, 29, 15], deep learning [21, 1, 12, 22, 41], clustering [7, 39], and random forests [37]. In addition, LLP has found various applications including image and video analysis [8, 19], high energy physics [10], vote prediction [40], remote sensing [21, 11], medical image analysis [5], activity recognition [29], and reproductive medicine [15].

Despite the emergence of LLP as a prominent weak learning paradigm, the theoretical underpinnings of LLP have been slow to develop. In particular, prior work has not established an algorithm for LLP that is consistent with respect to a classification performance measure. Furthermore, there does not even exist a general-purpose, theoretically grounded empirical objective for training LLP classifiers.

We propose a statistical framework for LLP based on mutual contamination models (MCMs), which have been used previously as models for classification with noisy labels and other weak supervision problems [34, 3, 25, 4, 17]. We use this framework to motivate a principled empirical objective for LLP, prove generalization error bounds associated to two bag generation models, and establish universal consistency with respect to the balanced error rate (BER). The MCM framework further motivates a novel experimental setting that overcomes a limitation of earlier experimental comparisons.

**Related Work.** Quadrianto et al. [31] study an exponential family model for labels given features, and show that the model is characterized by a certain “mean map" parameter that can be estimated in the LLP setting. They also provide Rademacher complexity bounds for the mean map and the associated log-posterior, but do not address a classification performance measure. Patrini et al. [28] extend the work of [31] in several ways, including a generalization error bound on the risk of a classifier. This bound is expressed in terms of an empirical LLP risk, a “bag Rademacher complexity," and a “label proportion complexity." The authors state that when bags are pure (LPs close to 0 or 1), the last of these terms is small, while for impure bags, the second term is small and the first term increases. While this bound motivates their algorithms, it is not clear how such a bound would imply consistency. Yu et al. [45] study the idea of minimizing the “empirical proportion risk" (EPR), which seeks a classifier that best reproduces the observed LPs. They develop a PAC-style bound on the accuracy of the resulting classifier under the assumption that all bags are very pure. Our work is the first to develop generalization error analysis and universal consistency for a classification performance measure, and we do so under a broadly applicable statistical model on bags.

The literature on LLP has so far yielded two general purpose training objectives that are usable across a variety of learning models. The first of these, the aforementioned EPR, minimizes the average discrepancy between observed and predicted LPs, where discrepancy is often measured by absolute or squared error in the binary case [45, 41, 10], and cross-entropy in the multiclass case [41, 12, 22, 5]. While [45] has been cited as theoretical support for this objective, that paper assumes the bags are very pure, and even provides examples of EPR minimization failure when bags are not sufficiently pure. We offer our own counterexample in an appendix. The second is the combinatorial objective introduced by [44] that incorporates the unknown labels as variables in the optimization, and jointly optimizes a conventional classification empirical risk together with a term (usually EPR) that encourages correctness of the imputed labels [44, 43, 21, 30, 9, 36, 37, 19, 12]. To our knowledge there is also no statistical theory supporting this objective. In contrast, we propose a theoretically grounded, general purpose criterion for training LLP models.

Finally, we note that an earlier version of this work approached LLP using so-called “label-flipping" or “class-conditional" noise models, as opposed to MCMs [35]. While that approach lead to the same algorithm described here, that setting is less natural for LLP, and the present version adds several more theoretical and experimental results.

**Notation.** Let $\mathcal{X}$ denote the feature space and $\{-1, 1\}$ the label space. For convenience we often abbreviate $-1$ and $+1$ by “-" and “+", and write $\{\pm\} = \{-, +\}$. A binary classification loss function, or *loss* for short, is a function $\ell : \mathbb{R} \times \{-1, 1\} \to \mathbb{R}$ (we allow losses to take negative values). For $\sigma \in \{\pm\}$, denote $\ell_\sigma(t) := \ell(t, \sigma)$. A loss $\ell$ is *Lipschitz (continuous)* if there exists $L$ such that for every $\sigma \in \{\pm\}$, and every $t, t' \in \mathbb{R}$, $|\ell_\sigma(t) - \ell_\sigma(t')| \leq L|t - t'|$. The smallest such $L$ for which this property holds is denoted $|\ell|$. Additionally, we define $|\ell|_0 := \max(|\ell_+(0)|, |\ell_-(0)|)$.

A decision function is a measurable function $f : \mathcal{X} \to \mathbb{R}$. The classifier induced by a decision function $f$ is the function $x \mapsto \operatorname{sign}(f(x))$. We will only consider classifiers induced by a decision function. In addition, we will often refer to a decision function as a classifier, in which case we mean the induced classifier. Let $P_+$ and $P_-$ be the class-conditional distributions of the feature vector $X$, and denote $P = (P_-, P_+)$. The performance measure considered in this work is the *balanced error rate* (BER) which, for a given loss $\ell$, and class conditional distributions $P = (P_+, P_-)$, is defined by $\mathcal{E}^\ell_P(f) := \frac{1}{2}\mathbb{E}_{X\sim P_+}[\ell_+(f(X))] + \frac{1}{2}\mathbb{E}_{X\sim P_-}[\ell_-(f(X))]$.

For an integer $n$, denote $[n] := \{1, 2, \ldots, n\}$. Given a sequence of numbers $(a_i)_{i\in[m]}$, denote the arithmetic and harmonic means by $\mathrm{AM}(a_i) := \frac{1}{m}\sum_{i\in[m]} a_i$ and $\mathrm{HM}(a_i) := (\frac{1}{m}\sum_{i\in[m]} a_i^{-1})^{-1}$. Finally, define the probability simplex $\Delta^N := \{w \in \mathbb{R}^N \mid w_i \geq 0 \,\forall i, \text{ and } \sum_i w_i = 1\}$.

## 2 Mutual Contamination Models

In this section we define MCMs and present new technical results for learning from MCMs that motivate our study of LLP in the next section, and which may also be of independent interest. We will consider collections of instances $X_1, \ldots, X_m \sim \gamma P_+ + (1 - \gamma)P_-$, where $\gamma \in [0, 1]$ and $m$ are fixed. Foreshadowing LLP, we refer to such collections of instances as *bags*.

We adopt the following assumption on bag data generation, with two cases depending on within-bag dependencies. Suppose there are $L$ total bags with sizes $n_i$, $i \in [L]$, proportions $\gamma_i \in [0, 1]$, and elements $X_{ij}$, $i \in [L], j \in [n_i]$. We assume

The distributions $P_+$ and $P_-$ are the same for all bags. $\gamma_i$ and $m_i$ may vary from bag to bag. If $i \neq r$, then $X_{ij}$ and $X_{rs}$ are independent $\forall j, s$. Furthermore, for all $i$,

**(IIM)** In the *independent instance model*, $X_{ij} \overset{iid}{\sim} \gamma_i P_+ + (1-\gamma_i)P_-$;

**(IBM)** In the *independent bag model*, the marginal distribution of $X_{ij}$ is $\gamma_i P_+ + (1-\gamma_i)P_-$.

**(IBM)** allows the instances within each bag to be dependent. Furthermore, any dependence structure, such as a covariance matrix, *may change from bag to bag.* **(IIM)** is a special case of **(IBM)** that allows us to quantify the impact of bag size $n_i$ on generalization error.

## 2.1 Mutual Contamination Models and Unbiased Losses

Recall that $P$ denotes the pair $(P_+, P_-)$. Let $\kappa = (\kappa^+, \kappa^-)$ be such that $\kappa^+ + \kappa^- < 1$. A *mutual contamination model* is the pair $P^\kappa := (P_+^\kappa, P_-^\kappa)$ where

$$P_+^\kappa := (1-\kappa^+)P_+ + \kappa^+ P_- \qquad \text{and} \qquad P_-^\kappa := (1-\kappa^-)P_- + \kappa^- P_+.$$

$P_+^\kappa$ and $P_-^\kappa$ may be thought of as noisy or contaminated versions of $P_+$ and $P_-$, respectively, where the contamination arises from the other distribution. MCMs are common models for label noise [34, 25, 4], where $\kappa^\sigma$ may be interpreted as the label noise rates $\mathbb{P}(Y = -\sigma | \tilde{Y} = \sigma)$, where $Y$ and $\tilde{Y}$ are the true and observed labels.

Given $\ell$ and $\kappa$ define the loss $\ell^\kappa$ by

$$\ell_\sigma^\kappa(t) := \frac{1-\kappa^{-\sigma}}{1-\kappa^- - \kappa^+}\ell_\sigma(t) - \frac{\kappa^{-\sigma}}{1-\kappa^- - \kappa^+}\ell_{-\sigma}(t), \qquad \sigma \in \{\pm\}.$$

This loss undoes the bias present in the mutual contamination model.

**Proposition 1.** *Consider any $P = (P_+, P_-)$, $\kappa = (\kappa^+, \kappa^-)$ with $\kappa^+ + \kappa^- < 1$, and loss $\ell$. For any $f$ such that all four of the quantities $\mathbb{E}_{X\sim P_\pm}\ell_\pm(f(X))$ exist and are finite, $\mathcal{E}_P^\ell(f) = \mathcal{E}_{P^\kappa}^{\ell^\kappa}(f)$.*

This result mirrors a similar result established by Natarajan et al. [26] under a label-flipping model for label noise, which is the other prominent models for random label noise besides the MCM. The proof simply matches coefficients of $\mathbb{E}_{X\sim P_\pm}\ell_\pm(f(X))$ on either side of the desired identity.

In an appendix we offer a sufficient condition for $\ell^\kappa$ to be convex. We also show (as an aside) that Prop. 1 enables a simple proof of a known result concerning symmetric losses, i.e., losses for which $\ell(t,1) + \ell(t,-1)$ is constant, such as the sigmoid loss. In particular, symmetric losses are immune to label noise under MCMs, meaning the original loss $\ell$ can be minimized on data drawn from the MCM and still optimize the clean BER [25, 42, 6].

The significance of Prop. 1 is that $\mathcal{E}_P^\ell(f)$ is the quantity we want to minimize, while $\mathcal{E}_{P^\kappa}^{\ell^\kappa}(f)$ can be estimated given data from an MCM. In particular, given bags $X_1^+, \ldots, X_{n^+}^+ \sim P_+^\kappa$ and $X_1^-, \ldots, X_{n^-}^- \sim P_-^\kappa$, Prop. 1 motivates minimizing the estimate of BER given by

$$\widehat{\mathcal{E}}(f) := \frac{1}{2n^+}\sum_{j=1}^{n^+}\ell_+^\kappa(f(X_j^+)) + \frac{1}{2n^-}\sum_{j=1}^{n^-}\ell_-^\kappa(f(X_j^-)) = \frac{1}{2}\sum_{\sigma\in\{\pm\}}\frac{1}{n^\sigma}\sum_{j=1}^{n^\sigma}\ell_\sigma^\kappa(f(X_j^\sigma))$$

over $f \in \mathcal{F}$, where $\mathcal{F}$ is some class of decision functions. We have

**Proposition 2.** *Under* **(IBM)** *, for any $f$ such that the quantities $\mathbb{E}_{X\sim P_\pm}\ell_\pm(f(X))$ exist and are finite, $\mathbb{E}[\widehat{\mathcal{E}}(f)] = \mathcal{E}_P^\ell(f)$.*

## 2.2 Learning from Multiple Mutual Contamination Models

In the next section we view LLP in terms of a more general problem that we now define. Suppose we are given $N$ different MCMs. Each has the same true class-conditional distributions $P_+$ and $P_-$, but possibly different contamination proportions $\kappa_i = (\kappa_i^+, \kappa_i^-)$, $i \in [N]$. Let $P^{\kappa_i} = (P_+^{\kappa_i}, P_-^{\kappa_i})$ denote the $i$th MCM, and assume $\kappa_i^+ + \kappa_i^- < 1$. Now suppose that for each $i \in [N]$, we observe

$$\begin{aligned} X_{i1}^+, \dots, X_{in_i^+}^+ &\sim P_+^{\kappa_i} := (1-\kappa_i^+)P_+ + \kappa_i^+ P_-, \\ X_{i1}^-, \dots, X_{in_i^-}^- &\sim P_-^{\kappa_i} := (1-\kappa_i^-)P_- + \kappa_i^- P_+. \end{aligned}$$

The problem of *learning from multiple mutual contamination models* (LMMCM) is to use all of the above data to design a single classifier that minimizes the clean BER $\mathcal{E}_P^\ell$.

A natural approach to this problem is to minimize the weighted empirical risk

$$\widehat{\mathcal{E}}_w(f) := \sum_{i=1}^N w_i \widehat{\mathcal{E}}_i(f) \quad \text{where} \quad \widehat{\mathcal{E}}_i(f) := \frac{1}{2n_i^+} \sum_{j=1}^{n_i^+} \ell_+^{\kappa_i}(f(X_{ij}^+)) + \frac{1}{2n_i^-} \sum_{j=1}^{n_i^-} \ell_-^{\kappa_i}(f(X_{ij}^-)),$$

where $w \in \Delta^N$. By Prop. 1, under **(IBM)** each $\widehat{\mathcal{E}}_i(f)$ is an unbiased estimate of $\mathcal{E}_P^\ell(f)$, and therefore so is $\widehat{\mathcal{E}}_w(f)$. This leads to the question of how best to set $w$. Intuitively, MCMs $P^{\kappa_i}$ with less corruption should receive larger weights. We confirm this intuition by choosing $w_i$ to optimize a generalization error bound (GEB). Our GEBs uses two weighted, multi-sample extensions of Rademacher complexity, corresponding to **(IIM)** and **(IBM)** , that we now introduce.

Let $S$ denote all the data $X_{ij}^\sigma$ from $N$ MCMs as described above.

**Definition 3.** *Let $\mathcal{F}$ be a class of decision functions. Assume that $\sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}} |f(x)| < \infty$. For any $c \in \mathbb{R}_{\geq 0}^N$, define*

$$\mathfrak{R}_c^I(\mathcal{F}) := \mathbb{E}_S \mathbb{E}_{(\epsilon_{ij}^\sigma)} \left[ \sup_{f\in\mathcal{F}} \sum_{i=1}^N c_i \sum_{\sigma\in\{\pm\}} \frac{1}{2n_i^\sigma} \sum_{j=1}^{n_i^\sigma} \epsilon_{ij}^\sigma f(X_{ij}^\sigma) \right], \tag{1}$$

*and*

$$\mathfrak{R}_c^B(\mathcal{F}) := \mathbb{E}_S \mathbb{E}_{((\sigma_i, X_i)\sim \widehat{P}^{\kappa_i})_{i\in[N]}} \mathbb{E}_{(\epsilon_i)} \left[ \sup_{f\in\mathcal{F}} \sum_{i=1}^N \epsilon_i c_i f(X_i) \right], \tag{2}$$

*where $\epsilon_{ij}^\sigma, \epsilon_i \overset{iid}{\sim} \text{unif}(\{-1,1\})$ are Rademacher random variables and $\widehat{P}^{\kappa_i}$ is the distribution that selects $\sigma_i \sim \text{unif}(\{-1,1\})$, and then draws $X_i$ uniformly from $X_{i,1}^\sigma, \dots, X_{i,n_i^\sigma}^\sigma$.*

The inner two summations in (1) reflect an adaptation of the usual Rademacher complexity to the BER, and the outer summation reflects the multiple MCMs. Eqn. (2) may be seen as a modification of (1) where the inner two sums are viewed as an empirical expectation that is pulled out of the supremum. If $\mathcal{F}$ satisfies the following, then $\mathfrak{R}_c^I(\mathcal{F})$ and $\mathfrak{R}_c^B(\mathcal{F})$ are bounded by tractable expressions.

**(SR)** There exist constants $A$ and $B$ such that $\sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}} |f(x)| \leq A$, and for all $M$, $x_1, \dots, x_M \in \mathcal{X}$, and $a \in \mathbb{R}_{\geq 0}^M$,

$$\mathbb{E}_{(\epsilon_i)} \left[ \sup_{f\in\mathcal{F}} \sum_{i=1}^M \epsilon_i a_i f(x_i) \right] \leq B \sqrt{\sum_{i=1}^M a_i^2}.$$

As one example of an $\mathcal{F}$ satisfying **(SR)** , let $k$ be a symmetric positive definite (SPD) kernel, bounded[1] by $K$, and let $\mathcal{H}$ be the associated reproducing kernel Hilbert space (RKHS). Let $\mathcal{F}_{K,R}^k$ denote the ball of radius $R$, centered at 0, in $\mathcal{H}$. As a second example, assume $\mathcal{X} \subset \mathbb{R}^d$ and $\|\mathcal{X}\|_2 := \sup_{x\in\mathcal{X}} \|x\|_2 < \infty$, where $\|\cdot\|_2$ is the Euclidean norm. Let $\alpha, \beta \in \mathbb{R}_+^M$ and denote $[x]_+ = \max(0, x)$. Define the class of two-layer neural networks with ReLU activation by

$$\mathcal{F}_{\alpha,\beta}^{\text{NN}} = \{f(x) = v^T [Ux]_+ : v \in \mathbb{R}^h, U \in \mathbb{R}^{h\times d}, |v_i| \leq \alpha_i, \|u_i\|_2 \leq \beta_i, i = 1, 2, \dots, h\}.$$

[1] An SPD kernel $k$ is bounded by $K$ if $\sqrt{k(x,x)} \leq K$ for all $x$. For example, the Gaussian kernel $k(x,x') = \exp(-\gamma\|x-x'\|^2)$ is bounded by $K = 1$.

**Proposition 4.** *$\mathcal{F}^k_{K,R}$ satisfies* **(SR)** *with $(A, B) = (RK, RK)$, and $\mathcal{F}^{NN}_{\alpha,\beta}$ satisfies* **(SR)** *with $(A, B) = (\|\alpha\|_2\|\beta\|_2\|\mathcal{X}\|_2, 2\langle\alpha, \beta\rangle\|\mathcal{X}\|_2)$.*

We emphasize that other classes $\mathcal{F}$ admit quantitative bounds on $\mathfrak{R}^I_c(\mathcal{F})$ and $\mathfrak{R}^B_c(\mathcal{F})$ that do not conform to **(SR)** , and that can also be leveraged as we do below. We focus on **(SR)** because the GEBs simplify considerably making it possible to derive closed form expressions for the optimal $w_i$. Below we write $\overset{(\mathbf{SR})}{\leq}$ to indicate an upper bound that holds provided **(SR)** is true.

Our first main result establishes GEBs for LMMCM under both **(IIM)** and **(IBM)** .

**Theorem 5.** *Let $S$ collect all the data $(X^\sigma_{ij})$ from $N$ MCMs with common base distributions $P_+, P_-$, and contamination proportions $\kappa_i = (\kappa^+_i, \kappa^-_i)$ satisfying $\kappa^-_i + \kappa^+_i < 1$. Let $\mathcal{F}$ be a class of decision functions such that $A = \sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}} |f(x)| < \infty$, let $\ell$ a Lipschitz loss, $w \in \Delta^N$, and $\delta > 0$. Under* **(IIM)** *, with probability $\geq 1 - \delta$ wrt the draw of $S$,*

$$\sup_{f\in\mathcal{F}} \left|\widehat{\mathcal{E}}_w(f) - \mathcal{E}(f)\right| \leq 2\mathfrak{R}^I_c(\mathcal{F}) + C\sqrt{\sum_{i=1}^N \frac{w_i^2}{\bar{n}_i(1-\kappa^-_i-\kappa^+_i)^2}} \overset{(\mathbf{SR})}{\leq} D\sqrt{\sum_{i=1}^N \frac{w_i^2}{\bar{n}_i(1-\kappa^-_i-\kappa^+_i)^2}} \tag{3}$$

*where $\bar{n}_i := \mathrm{HM}(n^-_i, n^+_i)$, $c_i = w_i|\ell|/(1 - \kappa^-_i - \kappa^+_i)$, $C = (1 + A|\ell|)\sqrt{\log(2/\delta)}$, and $D = 2B|\ell| + C$. Under* **(IBM)** *, the same statement holds after replacing $\mathfrak{R}^I_c(\mathcal{F}) \to \mathfrak{R}^B_c(\mathcal{F})$ and $\bar{n}_i \to 1$.*

Several remarks are in order. Under **(IIM)** , even in the special case $N = 1$ without noise ($\kappa^-_1 = \kappa^+_1 = 0$) the result appears new, and amounts to an adaptation of the standard Rademacher complexity bound to BER. The case $N = 1$ *with* noise can be used to prove consistency (with $\bar{n}_1 \to \infty$) of a discrimination rule for a single *MCM* given knowledge of, or consistent estimates of $\kappa^-_1, \kappa^+_1$. Previous results of this type have analyzed MCMs via label-flipping models which is less natural [4].

Because the result holds for any $w \in \Delta^N$, as long as the $\kappa_i$ are known a priori, we may set $w$ to optimize the rightmost expressions in (3). This leads to optimal weights $w_i \propto \bar{n}_i(1 - \kappa^-_i - \kappa^+_i)^2$ under **(IBM)** (here and below, replace $\bar{n}_i$ by 1 for **(IBM)** ), which supports our claim that MCMs with more information (larger samples, less noise) should receive more weight. With this choice of weights, the summation in the bound reduces to $\frac{1}{N}\mathrm{HM}(1/\bar{n}_i(1 - \kappa^-_i - \kappa^+_i)^2)$. In contrast, with uniform weights $w_i = 1/N$ the summation equals $\frac{1}{N}\mathrm{AM}(1/\bar{n}_i(1 - \kappa^-_i - \kappa^+_i)^2)$. The harmonic mean is much less sensitive to the presence of outliers, i.e., very noisy MCMs, than the arithmetic.

# 3 Learning from Label Proportions

In learning from label proportions with binary labels, the learner has access to $(b_1, \widehat{\gamma}_1), \ldots, (b_L, \widehat{\gamma}_L)$, where each $b_i$ is a bag of $n_i$ unlabeled instances, and each $\widehat{\gamma}_i \in [0, 1]$ is the proportion of instances from class 1 in the bag. The goal is to learn an accurate classifier as measured by some performance measure, which in our case we take to be the BER. This choice is already a departure from prior work on LLP, which typically looks at misclassification rate (MCR). The BER is defined without reference to a distribution of the label $Y$, and is thus invariant to changes in this distribution. In other words, BER is immune to shifts in class prevalence, and hence to shifts in the distribution of label proportions.

We adopt the following data generation model for bags. Each bag has a *true label proportion* $\gamma_i \in [0, 1]$. For each $i$, let $(X_{ij}, Y_{ij})$, $j \in [n_i]$, be random variables. The $i$th bag is formed from $(X_{ij})_{j\in[n_i]}$, and the *observed* or *empirical label proportion* is $\widehat{\gamma}_i = \frac{1}{n_i}\sum_j \frac{Y_{ij}+1}{2}$. Let $\boldsymbol{\gamma}, \boldsymbol{Y}$, and $\boldsymbol{X}$ be vectors collecting all of the values of $\gamma_i, Y_{ij}$, and $X_{ij}$, respectively. We assume

The distributions $P_+$ and $P_-$ are the same for all bags. The $\gamma_i$ may be random, and the sizes $n_i$ are nonrandom. Conditioned on $\boldsymbol{\gamma}$, if $i \neq r$, then $X_{ij}$ and $X_{rs}$ are independent $\forall j, s$. Furthermore, conditioned on $\boldsymbol{\gamma}$, for bag $i$

**(CIIM)** In the *conditionally independent instance model*, $\frac{Y_{ij}+1}{2} \overset{iid}{\sim}$ Bernoulli$(\gamma_i)$ and conditioned on $Y_{i1}, \ldots, Y_{in_i}$, $X_{i1}, \ldots, X_{in_i}$ are independent with $X_{ij} \sim P_{Y_{ij}}$.

**(CIBM)** In the *conditionally independent bag model*, $\mathbb{E}[\widehat{\gamma}_i] = \gamma_i$ and for each $j$, the distribution of $X_{ij} | Y_{i1}, \ldots, Y_{in_i}$ is $P_{Y_{ij}}$.

Under **(CIBM)** , conditioned on $\boldsymbol{\gamma}$, for bag $i$ the labels $Y_{i1}, \ldots, Y_{in_i}$ may be dependent, and given these labels the instances $X_{ij}$ may also be dependent. Furthermore, the dependence structure may change from bag to bag. This means that given its label, the distribution of an instance is still dependent on its bag, in contrast to prior work [31]. We also allow that the $\gamma_i$ may be dependent, so that without conditioning on $\boldsymbol{\gamma}$, the bags themselves may be dependent.

As in the previous section, the significance of our model is that it provides for (conditionally) unbiased estimates of BER as we describe below. Indeed, if we view $\boldsymbol{\gamma}$ as fixed, **(CIIM)** clearly implies **(IIM)** (in fact, the two independent instance models are equivalent). However, it is not the case that **(CIBM)** implies **(IBM)** – the introduction of the latent labels allows for a more general independent bag model while still ensuring unbiased BER estimates. A weakening of **(CIBM)** , namely

**(CIBM’)** For each $j$, $\mathbb{E}[\frac{Y_{ij}+1}{2}] = \gamma_i$ and the distribution of $X_{ij} | Y_{i1}, \ldots, Y_{in_i}$ is $P_{Y_{ij}}$

does imply **(IBM)** (still viewing $\boldsymbol{\gamma}$ as fixed), as we show in an appendix.

In this section we propose to reduce LLP to the setting of the previous section by pairing the bags, so that each pair of bags constitutes an MCM.

## 3.1 LLP when True Label Proportions are Known

We first consider the less realistic setting where the $\gamma_i$ are deterministic and *known*. In this situation we may reduce LLP to LMMCS by pairing bags. In particular, we re-index the bags and let $(b_i^-, \gamma_i^-)$ and $(b_i^+, \gamma_i^+)$ constitute the $i$th pair of bags, such that $\gamma_i^- < \gamma_i^+$. The bags may be paired in any way that depends on $\gamma_1, \ldots, \gamma_L$, subject to $\gamma_i^- < \gamma_i^+ \, \forall i$. We also assume the total number of bags is $L = 2N$, so that the number of bag pairs is $N$.

If we set $\kappa_i = (\kappa_i^+, \kappa_i^-) := (1 - \gamma_i^+, \gamma_i^-)$, then we are in the setting of LMMCM described in the previous setting. Furthermore, $1 - \kappa_i^- - \kappa_i^+ = \gamma_i^+ - \gamma_i^- > 0$. Therefore we may apply all of the theory developed in the previous section without modification. Since $\boldsymbol{\gamma}$ is deterministic, **(CIIM)** and **(CIBM)**’ imply **(IIM)** and **(IBM)** as discussed above, and we may simply apply Theorem 5 to obtain GEBs for LLP. Choosing weights $w_i$ to minimize the **(SR)** form yields final bounds proportional to the square root of $\frac{1}{N}\mathrm{HM}(1/(\bar{n}_i(\gamma_i^+ - \gamma_i^-)^2)) = (\sum_i \bar{n}_i(\gamma_i^+ - \gamma_i^-)^2)^{-1}$ (under **(CIBM’) replace** $\bar{n}_i \to 1$). In the LLP setting, we may further optimize this bound by optimizing the pairing of bags. This leads to an integer program known as the weighted matching problem for which exact and approximate algorithms are known. See appendices for details.

If $\boldsymbol{\gamma}$ is random, and the $\gamma_i$ are distinct (which occurs w. p. 1, e.g., if $\boldsymbol{\gamma}$ is jointly continuous), Theorem 5 still holds conditioned on $\boldsymbol{\gamma}$, and therefore unconditionally by the law of total expectation.

Although the $\gamma_i$ are typically unknown in practice, the above discussion still yields a useful algorithm: simply “plug in” $\widehat{\gamma}_i$ for $\gamma_i$ and proceed to minimize $\widehat{\mathcal{E}}_w(f)$ (with optimally paired bags and optimized weights) over $\mathcal{F}$. A description of the learning procedure, which we use in our experiments, is presented in Algorithm 1.

**Algorithm 1** Plug-in approach to LLP via LMMCM (outline)

```
Input: $(b_1, \widehat{\gamma}_1), \ldots, (b_{2N}, \widehat{\gamma}_{2N})$, model class $\mathcal{F}$, loss $\ell$, tuning parameters
procedure LLP-LMMCM
    Solve weighted matching problem to find pairings maximizing $\sum_i (\widehat{\gamma}_i^+ - \widehat{\gamma}_i^-)^2$ (see supp.)
    Set $\kappa_i = (1 - \widehat{\gamma}_i^+, \widehat{\gamma}_i^-)$ and optimal weights $w_i \propto (\widehat{\gamma}_i^+ - \widehat{\gamma}_i^-)^2$
    Minimize $\widehat{\mathcal{E}}_w(f)$ over $\mathcal{F}$, perhaps with regularization
```

## 3.2 Consistent Learning from Label Proportions

When the true label proportions are not known, as is usually the case in practice, it is difficult to establish consistency of the plug-in approach without restrictive assumptions. This is because the $\widehat{\gamma}_i$ are random, and so there is always some nonnegligible probability that in each pair, the bag with larger $\gamma_i$ will be misidentified. This problem is especially pronounced for very small bag sizes. For example, if two bags with $\gamma_1 = .45$ and $\gamma_2 = .55$ are paired, and the bag sizes are 8 with independent labels, the probability that $\widehat{\gamma}_2 < \widehat{\gamma}_1$ is .26. One approach to overcoming this issue is to have the bag sizes $n_i^\sigma$ tend to $\infty$ asymptotically, in which case $\widehat{\gamma}_i \overset{a.s.}{\to} \gamma_i$. This is a less interesting setting, however, because the learner can discard all but one pair of bags and still achieve consistency using existing techniques for learning in MCMs [4]. Furthermore, the bag size is often fixed in applications.

We propose an approach based on merging the original "small bags" to form "big bags," and then applying the approach of Section 3.1. For convenience assume all original (small) bags have the same size $n_i = n$ moving forward. Let $K$ be an integer and assume $N$ is a multiple of $K$ for convenience, $N = MK$. As before, let $(b_i, \widehat{\gamma}_i)$, $i \in [2N]$, be the original, unpaired bags of size $n$. We refer to a *$K$-merging scheme* as any procedure that takes the original unpaired bags of size $n$ and combines them, using knowledge of the $\widehat{\gamma}_i$, to form paired bags of size $nK$. Let the paired bags be denoted $(B_i^+, \widehat{\Gamma}_i^+)$ and $(B_i^-, \widehat{\Gamma}_i^-)$, $i \in [M]$. Let $I_i^\sigma$ denote the original indices of the small bags comprising $B_i^\sigma$, so that $B_i^\sigma = \cup_{j \in I_i^+} b_i$ and $\widehat{\Gamma}_i^\sigma = \frac{1}{K}\sum_{j \in I_i^\sigma} \widehat{\gamma}_j^\sigma$.

We offer two examples of $K$-merging schemes. The first, called the *blockwise-pairwise (BP) scheme*, simply takes the original small bags in their given order. The $i$th block of 2K consecutive small bags are used to form the $i$th pair of big bags. This is done by considering consecutive, nonoverlapping pairs of small bags and assigning the small bag with larger $\widehat{\gamma}_i$ to $B_i^+$. Using notation, we define $I_i^+ = \{j \in [2K(i-1)+1 : 2Ki] \,|\, j \text{ is odd and } \widehat{\gamma}_j \geq \widehat{\gamma}_{j+1} \text{ or } j \text{ is even and } \widehat{\gamma}_j \geq \widehat{\gamma}_{j-1}\}$ and $I_i^- = [2K(i-1)+1 : 2Ki] \backslash I_i^+$ (ties may be broken arbitrarily). The *blockwise-max (BM) scheme* is like BP, except that for each block of $2K$ small bags, the $K$ small bags with largest $\widehat{\gamma}_j$ are assigned to the positive bag. One can imagine more elaborate schemes that are not blockwise. We say that scheme 1 *dominates* scheme 2 if, with probability 1, for every $i$, $\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^-$ for scheme 1 is at least as large as it is for scheme 2. For example, BM dominates BP.

Next, we form the modified weighted empirical risk. For each $i \in [M]$ and $\sigma \in \{\pm\}$, let $(X_{ij}^\sigma)$, $j \in [nK]$, denote the elements of $B_i^\sigma$, and $(Y_{ij}^\sigma)$ the associated labels. Also set $\widehat{\kappa}_i = (1 - \widehat{\Gamma}_i^+, \widehat{\Gamma}_i^-)$. Let $w \in \Delta^M$ such that $w_i \propto (\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^-)^2$, and define

$$\tilde{\mathcal{E}}(f) := \sum_{i=1}^M w_i \tilde{\mathcal{E}}_i(f) \qquad \text{where} \qquad \tilde{\mathcal{E}}_i(f) := \left[\frac{1}{2n}\sum_{\sigma \in \{\pm\}} \sum_{j=1}^{nK} \ell_\sigma^{\widehat{\kappa}_i}(f(X_{ij}^\sigma))\right].$$

In the proof of Thm. 6, we show that under **(CIBM)** , with high probability, $\tilde{\mathcal{E}}_i(f)$ is an unbiased estimate for $\mathcal{E}_P^\ell(f)$ when conditioned on $\boldsymbol{\gamma}$ and $\boldsymbol{Y}$.

To state our main result we adopt the following assumption on the distribution of label proportions.

**(LP)** There exist $\Delta, \tau > 0$ such that the sequence of random variables $Z_j = \mathbf{1}_{\{|\gamma_j - \gamma_{j+1}| < \Delta\}}$ satisfies the following. For every $J \subseteq [2N-1]$, $\mathbb{P}(\prod_{j \in J} Z_j = 1) \leq \tau^{|J|}$.

This condition is satisfied if the $\gamma_i$ are iid draws from any non-constant distribution. However, it also allows for the $\gamma_i$ to be correlated. As one example, let $(w_j)$ be iid random variables with support $\supseteq [-1,1]$. **(LP)** is satisfied if $\gamma_{j+1} = \gamma_j + \underline{w}_j$, where $\underline{w}_j$ is the truncation of $w_j$ to $[-\gamma_j, 1-\gamma_j]$. The point of **(LP)** is that it offers a dependence setting where a one-sided version of Hoeffding's inequality holds, which allows us to conclude that with high probability, for all odd $j \in [2N]$, $|\gamma_j - \gamma_{j+1}| \geq \Delta$ for approximately $N(1-\tau)$ of the original pairs of small bags [27].

We now state our main result. Define $\Gamma_i^+ = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\gamma}}[\widehat{\Gamma}_i^+]$ and $\Gamma_i^- = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\gamma}}[\widehat{\Gamma}_i^-]$.

**Theorem 6.** *Let* **(LP)** *hold. Let* $\epsilon_0 \in (0, \Delta(1-\tau))$. *Let* $\mathcal{F}$ *satisfy* $\sup_{x \in \mathcal{X}, f \in \mathcal{F}} |f(x)| \leq A < \infty$ *and let* $\ell$ *be a Lipschitz loss. Let* $\epsilon \in (0, \frac{\Delta(1-\tau)-\epsilon_0}{1+\Delta}]$ *and* $\delta \in (0,1]$. *For the BP merging scheme, under* **(CIIM)** *, with* *probability at least* $1-\delta-2\frac{N}{K}e^{-2K\epsilon^2}$ *with respect to the draw of* $\boldsymbol{\gamma}, \boldsymbol{Y}, \boldsymbol{X}$*,*

$$\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- \geq \Gamma_i^+ - \Gamma_i^- - \epsilon \geq \epsilon_0$$

*and*

$$\sup_{f\in\mathcal{F}} \left|\tilde{\mathcal{E}}(f) - \mathcal{E}(f)\right| \leq 2\mathfrak{R}_c^I(\mathcal{F}) + C\sqrt{\frac{\mathrm{HM}((\Gamma_i^+ - \Gamma_i^- - \epsilon)^{-2})}{2(N/K)n}} \overset{\textbf{(SR)}}{\leq} D\sqrt{\frac{\mathrm{HM}((\Gamma_i^+ - \Gamma_i^- - \epsilon)^{-2})}{2(N/K)n}}, \tag{4}$$

*where* $c_i = w_i|\ell|/(\Gamma_i^+ - \Gamma_i^- - \epsilon)$*,* $C = (1 + A|\ell|)\sqrt{\log(2/\delta)}$*, and* $D = 2B|\ell| + C$*. Under* **(CIBM)***, the same bounds hold with the same probability if we substitute* $\mathfrak{R}_c^I(\mathcal{F}) \to \mathfrak{R}_c^B(\mathcal{F})$ *and* $n \to 1$*.*

This result states that BP achieves essentially the same bound (modulo $\epsilon$) as if we applied LMMCM to the big bags with *known* $\Gamma_i^+, \Gamma_i^-$. We also note that there is no restriction on bag size $n$. A corollary of this result also applies to any scheme that dominates BP, as we explain in an appendix.

Theorem 6 implies a consistent learning algorithm for LLP under both **(CIIM)** and **(CIBM)** , using any merging scheme that dominates BP. To achieve consistency the bound should tend to zero while the confidence tends to 1, as $N \to \infty$. Even with $n$ fixed, this is true provided $K \to \infty$ and $N/K \to \infty$ as $N \to \infty$, such that $N = O(K^\beta)$ for some $\beta > 0$. Beyond that, standard arguments may be applied to arrive at a formal consistency result. In an appendix we state such a result for completeness. Here the consistency is *universal* in that it makes no assumptions on $P_-$ or $P_+$.

# 4 Experiments

The vast majority of LLP methodology papers simulate data for LLP by taking a classification data set, randomly shuffling the data, and sectioning off the data into bags of a certain size. This implies that the expected label proportions for all bags are the same, and as bag size increases, all label proportions converge to the class prior probabilities. The case where all LPs are the same is precisely the setting where LLP becomes intractable, and hence these papers report decreasing performance with increasing bag size.

We propose an alternate sampling scheme inspired by our MCM framework. Each experiment is based on a classification data set, a distribution of LPs, and the bag size $n$. For each dataset, the total number of training instances $T$ is fixed, so that the number of bags is $T/n$. We consider the Adult ($T = 8192$) and MAGIC Gamma Ray Telescope ($T = 6144$) datasets (both available from the UCI repository[2]), LPs that are iid uniform on $[0, \frac{1}{2}]$ and on $[\frac{1}{2}, 1]$, and bag sizes $n \in \{8, 32, 128, 512\}$. The total number of experimental settings is thus $2 \times 2 \times 4 = 16$. The numerical features in both datasets are standardized to have 0 mean and unit variance, the categorical features are one-hot encoded.

We implement a method based on our general approach (see Algorithm 1) by taking $\ell$ to be the logistic loss, $\mathcal{F}$ to be the RKHS associated to a Gaussian kernel $k$, and selecting $f \in \mathcal{F}$ by minimizing $\widehat{\mathcal{E}}_w(f) + \lambda\|f\|_{\mathcal{F}}^2$. By the representer theorem [33], the minimizer of this objective has the form $f(x) = \sum_i \alpha_i k(x, x_i)$ where $\alpha_i \in \mathbb{R}$ and $x_i$ ranges over all training instances. Our Python implementation uses SciPy's L-BFGS routine to find the optimal $\alpha_i$. The kernel parameter is computed by $\frac{1}{d*Var(X)}$ where $d$ is the number of features and $Var(X)$ is the variance of the data matrix, and the parameter $\lambda \in \{1, 10^{-1}, 10^{-2}, \ldots, 10^{-5}\}$ is chosen by 5-fold cross validation. We tried the EPR as a criterion for model selection but found our own criterion to be better. For each dataset, our implementation runs all 8 settings in roughly 50 minutes using 48 cores.

We compare against InvCal [32] and alter-∝SVM [44], the two most common reference methods in LLP, using Matlab implementations provided by the authors of [44]. Those methods are designed to optimize accuracy, whereas ours is designed to optimize BER. For a fair comparison, for each method we shift the decision function's threshold to generate an ROC curve and evaluate the area under the curve (AUC) using all data that was not used for training. For each experimental setting, the reported AUC and standard

[2] http://archive.ics.uci.edu/ml

deviation reflect the average results over 5 randomized trials. Additional experimental details are found in an appendix.

The results are reported in Table 1. Bold numbers indicate that a method's mean AUC was the largest for that experimental setting. We see that for the smallest bag size, the methods all perform comparably, while for larger bag sizes, LMMCM exhibits far less degradation in performance. Using the Wilcoxon signed-rank test, we find that LMMCM outperforms InvCal with p-value $< 0.005$.

Table 1: AUC. Column header indicates bag size.

| Data set, LP dist | Method | 8 | 32 | 128 | 512 |
|---|---|---|---|---|---|
| Adult, $\left[0, \frac{1}{2}\right]$ | InvCal | 0.8720 ± 0.0035 | 0.8672 ± 0.0067 | 0.8537 ± 0.0101 | 0.7256 ± 0.0159 |
| | alter-$\propto$SVM | 0.8586 ± 0.0185 | 0.7394 ± 0.0686 | 0.7260 ± 0.0953 | 0.6876 ± 0.1219 |
| | LMMCM | **0.8728 ± 0.0019** | **0.8693 ± 0.0047** | **0.8669 ± 0.0041** | **0.8674 ± 0.0040** |
| Adult, $\left[\frac{1}{2}, 1\right]$ | InvCal | **0.8680 ± 0.0021** | 0.8598 ± 0.0073 | 0.8284 ± 0.0093 | 0.7480 ± 0.0500 |
| | alter-$\propto$SVM | 0.8587 ± 0.0097 | 0.7429 ± 0.1473 | 0.8204 ± 0.0318 | 0.7602 ± 0.1215 |
| | LMMCM | 0.8584 ± 0.0164 | **0.8644 ± 0.0052** | **0.8601 ± 0.0045** | **0.8500 ± 0.0186** |
| MAGIC, $\left[0, \frac{1}{2}\right]$ | InvCal | **0.8918 ± 0.0076** | 0.8574 ± 0.0079 | 0.8295 ± 0.0139 | 0.8133 ± 0.0109 |
| | alter-$\propto$SVM | 0.8701 ± 0.0026 | 0.7704 ± 0.0818 | 0.7753 ± 0.0207 | 0.6851 ± 0.1580 |
| | LMMCM | 0.8909 ± 0.0077 | **0.8799 ± 0.0113** | **0.8753 ± 0.0157** | **0.8734 ± 0.0092** |
| MAGIC, $\left[\frac{1}{2}, 1\right]$ | InvCal | **0.8936 ± 0.0066** | 0.8612 ± 0.0056 | 0.8180 ± 0.0092 | 0.8215 ± 0.0136 |
| | alter-$\propto$SVM | 0.8689 ± 0.0135 | 0.8219 ± 0.0218 | 0.8179 ± 0.0487 | 0.7949 ± 0.0478 |
| | LMMCM | 0.8911 ± 0.0083 | **0.8790 ± 0.0091** | **0.8684 ± 0.0046** | **0.8567 ± 0.0292** |

We performed an additional set of experiments where the number of bags $N$ remains fixed. For Adult dataset, the total number of bags is 16, and for MAGIC, it is 12. For each method, we generate an ROC curve and evaluate the area under the curve (AUC) using the test data. The average AUCs and the standard deviations over 5 random trials are reported in Table 2. Bold numbers indicate that a method's mean AUC was the largest for that experimental setting. We observe that LMMCM exhibits excellent performance in this setting as well.

Table 2: AUC. Column header indicates bag size.

| Data set, LP dist | Method | 8 | 32 | 128 | 512 |
|---|---|---|---|---|---|
| Adult, $\left[0, \frac{1}{2}\right]$ | InvCal | 0.6427 ± 0.0922 | 0.6545 ± 0.0643 | 0.6518 ± 0.0139 | 0.7230 ± 0.0253 |
| | alter-$\propto$SVM | 0.6525 ± 0.0817 | 0.5959 ± 0.1145 | 0.6199 ± 0.1267 | 0.6419 ± 0.0997 |
| | LMMCM | **0.7299 ± 0.0796** | **0.7765 ± 0.0590** | **0.8329 ± 0.0166** | **0.8456 ± 0.0213** |
| Adult, $\left[\frac{1}{2}, 1\right]$ | InvCal | 0.5973 ± 0.0740 | 0.6634 ± 0.0864 | 0.6408 ± 0.0216 | 0.7218 ± 0.0170 |
| | alter-$\propto$SVM | 0.6035 ± 0.1626 | **0.7774 ± 0.0443** | 0.5863 ± 0.2775 | 0.7106 ± 0.2193 |
| | LMMCM | **0.7228 ± 0.1048** | 0.7674 ± 0.0586 | **0.8428 ± 0.0101** | **0.8588 ± 0.0091** |
| MAGIC, $\left[0, \frac{1}{2}\right]$ | InvCal | **0.7381 ± 0.0439** | 0.7828 ± 0.0212 | 0.7936 ± 0.0371 | 0.8196 ± 0.0231 |
| | alter-$\propto$SVM | 0.5997 ± 0.1163 | 0.5376 ± 0.1671 | 0.6859 ± 0.0371 | 0.7193 ± 0.1278 |
| | LMMCM | 0.7180 ± 0.0450 | **0.7852 ± 0.7828** | **0.8140 ± 0.0463** | **0.8630 ± 0.0275** |
| MAGIC, $\left[\frac{1}{2}, 1\right]$ | InvCal | 0.6741 ± 0.0673 | 0.7405 ± 0.0433 | 0.7876 ± 0.0249 | 0.8135 ± 0.0132 |
| | alter-$\propto$SVM | 0.6589 ± 0.1029 | 0.6330 ± 0.1254 | 0.6790 ± 0.1072 | 0.7965 ± 0.0708 |
| | LMMCM | **0.6807 ± 0.0779** | **0.7639 ± 0.0335** | **0.7905 ± 0.0258** | **0.8491 ± 0.0245** |

## 5 Conclusion

We have introduced a principled framework for LLP based on MCMs. We have developed several novel results for MCMs, and used them to develop a statistically consistent procedure and an effective practical algorithm for LLP. The most natural direction for future work is to extend to multiclass.

# A Failure Case for Empirical Proportion Risk Minimization

We offer a simple example where minimizing the empirical proportion risk leads to suboptimal performance. Let $P_-$ be uniform on $[0,1]$, with density $p_-(x) = \mathbf{1}_{\{x\in[0,1]\}}$, and let $P_+$ have the triangular density function $p_+(x) = 2x\mathbf{1}_{\{x\in[0,1]\}}$. Suppose there is a single bag, and that the label proportion is $\gamma = \frac{1}{2}$. Also suppose $\mathcal{F}$ consists of threshold classifiers $f_t(x) = \text{sign}(x - t)$, $t \in [0,1]$. This class contains the optimal BER classifier (define wrt 0-1 loss) corresponding to $t^* = \frac{1}{2}$. Now suppose we are in the infinite bag-size limit (which only makes the problem easier), so that the observed label proportion $\widehat{\gamma}$ is simply $\gamma = \frac{1}{2}$. Then we seek the threshold $t'$ that minimizes

$$\text{EPR}(t) := \left|\mathbb{P}(f_t(X) = 1) - \frac{1}{2}\right|^p.$$

For any $p > 0$, $t'$ is the median of the marginal distribution of $X$, $\frac{1}{2}P_- + \frac{1}{2}P_+$, which equals $(\sqrt{5}-1)/2 \approx 0.62 \neq t^*$. Thus, minimizing EPR does not yield an optimal classifier for BER or for misclassification rate, which agrees with BER in this setting where the two classes are equally likely.

Now suppose there are $N$ bags, with label proportions $\gamma_1, \ldots, \gamma_N$ drawn iid from a distribution whose (population) mean and median are $\frac{1}{2}$, such as the uniform distribution on $[0,1]$. The optimal BER classifier remains the same, with threshold $t^* = \frac{1}{2}$. The optimal classifier wrt misclassification rate is also the same, assuming we view $\mathbb{E}[\gamma_i] = \frac{1}{2}$ as the class prior. In the infinite bag-size limit, EPR would seek the threshold $t'$ that minimizes

$$\text{EPR}_N(t) := \frac{1}{N}\sum_{i=1}^{N} \left|\mathbb{P}(f_t(X) = 1) - \gamma_i\right|^p.$$

For $p = 1$, EPR minimization selects $t'$ such that $\mathbb{P}(f_{t'}(X) = 1)$ is the empirical median of $\gamma_1, \ldots, \gamma_N$, which will be near $\frac{1}{2}$, which means $t'$ will be near 0.62. For $p = 2$, EPR minimization selects $t'$ such that $\mathbb{P}(f_{t'}(X) = 1)$ is the empirical mean of $\gamma_1, \ldots, \gamma_N$, which will again be near $\frac{1}{2}$, which again means $t'$ will be near 0.62.

More generally, based on the above example, EPR seems likely to fail whenever $P_+$ and $P_-$ are not sufficiently “symmetric."

# B Proofs of Results From Main Document

This section contains the proofs.

## B.1 Proof of Proposition 1

Consider the loss function $\tilde{\ell}$ given by

$$\begin{aligned}\tilde{\ell}_+(t) &= A\ell_+(t) - B\ell_-(t),\\ \tilde{\ell}_-(t) &= C\ell_-(t) - D\ell_+(t).\end{aligned}$$

Equating $\mathcal{E}^{\tilde{\ell}}_{P^\kappa}(f)$ to $\mathcal{E}^{\ell}_{P}(f)$ yields four equations in the four unknowns A, B, C, and D, corresponding to the coefficients of $\mathbb{E}_{X\sim P_\pm}\ell_\pm(f(X))$. The unique solution to this system is $\tilde{\ell} = \ell^\kappa$.

## B.2 Proof of Proposition 4

We begin with $\mathcal{F}^k_{R,K}$. For any $R > 0$, $f \in \mathcal{F}^k_{R,K}$, and $x \in \mathcal{X}$,

$$|f(x)| = |\langle f, k(\cdot, x)\rangle| \leq \|f\|_{\mathcal{H}}\|k(\cdot,x)\|_{\mathcal{H}} = RK.$$

by the reproducing property and Cauchy-Schwarz. Thus $A = RK$.

For the second part, the expectation may be bounded by a modification of the standard bound of Rademacher complexity for kernel classes. Thus,

$$\begin{aligned}
\mathbb{E}_{(\epsilon_i)}\left[\sup_{f\in\mathcal{F}^k_{R,K}}\sum_i a_i\epsilon_i f(x_i)\right] &= \mathbb{E}_{(\epsilon_i)}\left[\sup_{f\in\mathcal{F}^k_{R,K}}\sum_i a_i\epsilon_i\langle f, k(\cdot,x_i)\rangle\right] && (5)\\
&= \mathbb{E}_{(\epsilon_i)}\left[\sup_{f\in\mathcal{F}^k_{R,K}}\left\langle f, \sum_i a_i\epsilon_i k(\cdot,x_i)\right\rangle\right]\\
&= \mathbb{E}_{(\epsilon_i)}\left[\left\langle R\frac{\sum_i a_i\epsilon_i k(\cdot,x_i)}{\|\sum_i a_i\epsilon_i k(\cdot,x_i)\|}, \sum_i a_i\epsilon_i k(\cdot,x_i)\right\rangle\right] && (6)\\
&= R\mathbb{E}_{(\epsilon_i)}\left[\sqrt{\left\|\sum_i a_i\epsilon_i k(\cdot,x_i)\right\|^2}\right]\\
&\le R\sqrt{\mathbb{E}_{(\epsilon_i)}\left[\left\|\sum_i a_i\epsilon_i k(\cdot,x_i)\right\|^2\right]} && (7)\\
&= R\sqrt{\sum_i a_i^2\|k(\cdot,x_i)\|^2} && (8)\\
&\le RK\sqrt{\sum_{i=1}^M a_i^2}, && (9)
\end{aligned}$$

where (5) uses the reproducing property, (6) is the condition for equality in Cauchy-Schwarz, (7) is Jensen's inequality, (8) follows from independence of the Rademacher random variables, and (9) follows from the reproducing property and the bound on the kernel.

Next, consider $\mathcal{F}^{\text{NN}}_{\alpha,\beta}$. For the first part we have for any $f\in\mathcal{F}^{\text{NN}}_{\alpha,\beta}$ and $x\in\mathcal{X}$,

$$\begin{aligned}
|f(x)| &= |\langle v, [Ux]_+\rangle|\\
&\le \|v\|\|[Ux]_+\|\\
&\le \|\alpha\|\|[Ux]_+\|\\
&\le \|\alpha\|\|Ux\|\\
&= \|\alpha\|\sqrt{\sum_j |\langle u_j, x\rangle|^2}\\
&\le \|\alpha\|\sqrt{\sum_j \|u_j\|^2\|x\|^2}\\
&\le \|\mathcal{X}\|\|\alpha\|\sqrt{\sum_j \|u_j\|^2}\\
&\le \|\mathcal{X}\|\|\alpha\|\|\beta_j\|.
\end{aligned}$$

For the second part, observe

$$\begin{aligned}
\mathbb{E}_{(\epsilon_k)}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{M}\epsilon_k a_k f(x_k)\right] &= \mathbb{E}_{(\epsilon_k)}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{M}\epsilon_k a_k\sum_{j=1}^{h}v_j\left[\langle u_j, x_k\rangle\right]_+\right]\\
&= \mathbb{E}_{(\epsilon_k)}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{M}\epsilon_k\sum_{j=1}^{h}v_j\left[\langle u_j, a_k x_k\rangle\right]_+\right]\\
&= \mathbb{E}_{(\epsilon_k)}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{h}v_j\sum_{k=1}^{M}\epsilon_k\left[\langle u_j, a_k x_k\rangle\right]_+\right]\\
&\le \mathbb{E}_{(\epsilon_k)}\left[\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^{h}v_j\sum_{k=1}^{M}\epsilon_k\left[\langle u_j, a_k x_k\rangle\right]_+\right|\right]\\
&\le \mathbb{E}_{(\epsilon_k)}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{h}\alpha_j\left|\sum_{k=1}^{M}\epsilon_k\left[\langle u_j, a_k x_k\rangle\right]_+\right|\right]\\
&\le \sum_{j=1}^{h}\alpha_j\mathbb{E}_{(\epsilon_k)}\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{M}\epsilon_k\left[\langle u_j, a_k x_k\rangle\right]_+\right|.\\
&= \sum_{j=1}^{h}\alpha_j\mathbb{E}_{(\epsilon_k)}\sup_{u_j:\|u_j\|\le\beta_j}\left|\sum_{k=1}^{M}\epsilon_k\left[\langle u_j, a_k x_k\rangle\right]_+\right|. \qquad (10)
\end{aligned}$$

We bound the expectations in (10) using Ledoux-Talagrand contraction [20, Theorem 4.12].

**Theorem 7** (Ledoux-Talagrand contraction)**.** *Let $F:\mathbb{R}_+\to\mathbb{R}_+$ be convex and increasing. Further let $\varphi_i$, $i\in[M]$ be 1-Lipschitz functions such that $\varphi(0)=0$. Then, for any bounded subset $T\subset\mathbb{R}^M$,*

$$\mathbb{E}_{(\epsilon_i)}F\left(\frac{1}{2}\sup_{t\in T}\left|\sum_{i=1}^{M}\epsilon_i\varphi_i(t_i)\right|\right)\le\mathbb{E}_{(\epsilon_i)}F\left(\sup_{t\in T}\left|\sum_{i=1}^{M}\epsilon_i t_i\right|\right).$$

To apply this result, for each $j$ notice that

$$\mathbb{E}_{(\epsilon_k)}\sup_{u_j:\|u_j\|\le\beta_j}\left|\sum_{k=1}^{M}\epsilon_k\left[\langle u_j, a_k x_k\rangle\right]_+\right| = \mathbb{E}_{(\epsilon_k)}\sup_{t\in T_j}\left|\sum_{k=1}^{M}\epsilon_k\left[t_k\right]_+\right|$$

where $t=(t_1,t_2,\ldots,t_M)^T$ and

$$T_j=\left\{t=(\langle u_j,a_1x_1\rangle,\langle u_j,a_2x_2\rangle,\ldots,\langle u_j,a_Mx_M\rangle)^T\in\mathbb{R}^M:\|u_j\|\le\beta_j\right\}$$

which is clearly bounded. Now taking $F$ to be the identity and $\varphi_i = [\cdot]_+$, we have

$$\begin{aligned}
\mathbb{E}_{(\epsilon_k)} \sup_{u_j : \|u_j\| \le \beta_j} \left| \sum_{k=1}^{M} \epsilon_k \left[\langle u_j, a_k x_k \rangle\right]_+ \right| &\le 2\mathbb{E}_{(\epsilon_k)} \sup_{u_j : \|u_j\| \le \beta_j} \left| \sum_{k=1}^{M} \epsilon_k \langle u_j, a_k x_k \rangle \right| \\
&= 2\mathbb{E}_{(\epsilon_k)} \sup_{u_j : \|u_j\| \le \beta_j} \left| \left\langle u_j, \sum_{k=1}^{M} \epsilon_k a_k x_k \right\rangle \right| \\
&= 2\mathbb{E}_{(\epsilon_k)} \left\langle \beta_j \frac{\sum_{k=1}^{M} \epsilon_k a_k x_k}{\| \sum_{k=1}^{M} \epsilon_k a_k x_k \|}, \sum_{k=1}^{M} \epsilon_k a_k x_k \right\rangle \\
&= 2\beta_j \mathbb{E}_{(\epsilon_k)} \sqrt{\left\| \sum_{k=1}^{M} \epsilon_k a_k x_k \right\|^2} \qquad (11) \\
&\le 2\beta_j \sqrt{\mathbb{E}_{(\epsilon_k)} \left\| \sum_{k=1}^{M} \epsilon_k a_k x_k \right\|^2} \qquad (12) \\
&\le 2\beta_j \sqrt{\sum_{k=1}^{M} a_k^2 \|x_k\|^2} \qquad (13) \\
&\le 2\|\mathcal{X}\|_2 \beta_j \sqrt{\sum_{k=1}^{M} a_k^2}, \qquad (14)
\end{aligned}$$

where (11) uses the condition for equality in Cauchy-Schartz, (12) uses Jensen's inequality, and (13) uses independence of the $\epsilon_k$. The result now follows from (10) and (14).

## B.3 Proof of Theorem 5

We first review the following properties of the supremum which are easily verified.

P1 For any real-valued functions $f_1, f_2 : \mathcal{X} \to \mathbb{R}$,

$$\sup_x f_1(x) - \sup_x f_2(x) \le \sup_x (f_1(x) - f_2(x)).$$

P2 For any real-valued functions $f_1, f_2 : \mathcal{X} \to \mathbb{R}$,

$$\sup_x (f_1(x) + f_2(x)) \le \sup_x f_1(x) + \sup_x f_2(x).$$

P3 $\sup(\cdot)$ is a convex function, i.e., if $(x_\lambda)_{\lambda \in \Lambda}$ and $(x'_\lambda)_{\lambda \in \Lambda}$ are two sequences (where $\Lambda$ is possibly uncountable), then $\forall \alpha \in [0, 1]$,

$$\sup_{\lambda \in \Lambda} (\alpha x_\lambda + (1 - \alpha) x'_\lambda) \le \alpha \sup_{\lambda \in \Lambda} x_\lambda + (1 - \alpha) \sup_{\lambda \in \Lambda} x'_\lambda.$$

Introduce the variable $S$ to denote all realizations $X_{ij}^\sigma$, $1 \in [N], \sigma \in \{-, +\}, j \in [n_i^\sigma]$. We would like to bound

$$\xi(S) := \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{N} w_i \left( \frac{1}{2} \sum_{\sigma \in \{\pm 1\}} \left[ \frac{1}{n_i^\sigma} \sum_{j=1}^{n_i^\sigma} \ell_\sigma^\kappa (f(X_{ij}^\sigma)) \right] - \mathcal{E}(f) \right) \right|.$$

Introduce

$$\xi^+(S) := \sup_{f\in\mathcal{F}} \sum_{i=1}^{N} w_i \left( \frac{1}{2} \sum_{\sigma\in\{\pm 1\}} \left[ \frac{1}{n_i^\sigma} \sum_{j=1}^{n_i^\sigma} \ell_\sigma^\kappa(f(X_{ij}^\sigma)) \right] - \mathcal{E}(f) \right),$$
$$\xi^-(S) := \sup_{f\in\mathcal{F}} - \sum_{i=1}^{N} w_i \left( \frac{1}{2} \sum_{\sigma\in\{\pm 1\}} \left[ \frac{1}{n_i^\sigma} \sum_{j=1}^{n_i^\sigma} \ell_\sigma^\kappa(f(X_{ij}^\sigma)) \right] - \mathcal{E}(f) \right).$$

Assume **(IIM)** holds. Since the realizations $X_{ij}^\sigma$ are independent, we can apply the Azuma-McDiarmid bounded difference inequality [23] to $\xi^+$ and to $\xi^-$. We will show that the same bound on $\xi^+$ and $\xi^-$ holds with probability at least $1-\delta/2$. Combining these bounds gives the desired bound on $\xi$. We consider $\xi^+$ below, with the analysis for $\xi^-$ being identical.

**Definition 8.** *Let $A$ be some set and $\phi : A^n \to R$. We say $\phi$ satisfies the bounded difference assumption if $\exists c_1, \dots, c_n \geqslant 0$ s.t. $\forall i, 1 \leqslant i \leqslant n$*

$$\sup_{x_1,\dots,x_n,x_i'\in A} |\phi(x_1,\dots,x_i,\dots,x_n) - \phi(x_1,\dots,x_i',\dots,x_n)| \leqslant c_i$$

*That is, if we substitute $x_i$ to $x_i'$, while keeping other $x_j$ fixed, $\phi$ changes by at most $c_i$.*

**Lemma 9** (Bounded Difference Inequality)**.** *Let $X_1, \dots, X_n$ be arbitrary independent random variables on set $A$ and $\phi : A^n \to R$ satisfy the bounded difference assumption. Then $\forall t > 0$*

$$\Pr\{\phi(X_1,\dots,X_n) - \mathbb{E}[\phi(X_1,\dots,X_n)] \geqslant t\} \leqslant e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}.$$

To apply this result to $\xi^+$, first note that for any $f \in \mathcal{F}, x \in \mathcal{X}$, and $y \in \{-1, 1\}$,

$$\begin{aligned} |\ell^{\kappa_i}(f(x),y)| &\le |\ell^{\kappa_i}(0,y)| + |\ell^{\kappa_i}(f(x),y) - \ell^{\kappa_i}(0,y)| \\ &\le |\ell^{\kappa_i}|_0 + |\ell^{\kappa_i}||f(x)| \\ &\le |\ell^{\kappa_i}|_0 + |\ell^{\kappa_i}|A. \end{aligned}$$

If we modify $S$ by replacing some $X_{ij}^\sigma$ with another $X'$, while leaving all other values in $S$ fixed, then (by P1) $\xi^+$ changes by at most $2\frac{w_i(|\ell^{\kappa_i}|_0 + |\ell^{\kappa_i}|A)}{2n_i^\sigma}$, and we obtain that with probability at least $1-\delta/2$ over the draw of $S_1, \dots, S_N$,

$$\begin{aligned} \xi^+ - \mathbb{E}\left[\xi^+\right] &\le 2\sqrt{\frac{1}{2}\sum_{i=1}^{N} \frac{w_i^2(|\ell^{\kappa_i}|_0 + |\ell^{\kappa_i}|A)^2}{\bar{n}_i} \frac{\log(2/\delta)}{2}} \\ &\le 2(1 + A|\ell|)\sqrt{\frac{1}{2}\sum_{i=1}^{N} \frac{w_i^2}{\bar{n}_i(1-\kappa_i^- - \kappa_i^+)^2} \frac{\log(2/\delta)}{2}}, \end{aligned}$$

where we have used $|\ell^{\kappa_i}|_0 \le 1/(1-\kappa_i^- - \kappa_i^+)$ and $|\ell^{\kappa_i}| \le |\ell|/(1-\kappa_i^- - \kappa_i^+)$.

To bound $\mathbb{E}\left[\xi^+\right]$ we will use ideas from Rademacher complexity theory. Thus let $S'$ denote a separate (ghost) sample of corrupted data $(\underline{X}_{ij}^\sigma) \overset{iid}{\sim} \tilde{P}_\sigma^{\kappa_i}$, $i = 1, \dots, N$, $\sigma \in \{\pm\}$, $j = 1, \dots, n_i^\sigma$, independent of the realizations in $S$. Let $\widehat{\mathbb{E}}_S[f]$ be shorthand for $\sum_i w_i \sum_{\sigma\in\{\pm\}} \frac{1}{2n_i^\sigma} \sum_j \ell_\sigma^{\kappa_i}(f(X_{ij}^\sigma))$. Denote by $(\epsilon_{ij}^\sigma)$ $i \in [N], \sigma \in \{\pm\}, j \in [n_i^\sigma]$, iid Rademacher variables (independent from everything else), and let $\mathbb{E}_{(\epsilon_{ij}^\sigma)}$ denote the expectation with respect to all of these variables. We have

$$
\begin{aligned}
\mathbb{E}\left[\xi^+\right] &= \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N} w_i\left(\left[\sum_{\sigma\in\{\pm\}}\frac{1}{2n_i^\sigma}\sum_{j=1}^{n_i^\sigma}\ell_\sigma^{\kappa_i}(f(X_{ij}^\sigma))\right]-\mathcal{E}_P^\ell(f)\right)\right]\\
&= \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left(\widehat{\mathbb{E}}_S[f]-\mathbb{E}_{S'}\left[\widehat{\mathbb{E}}_{S'}[f]\right]\right)\right]\\
&\qquad \text{(by writing } \mathcal{E}_P^\ell(f)=\sum w_i\mathcal{E}_{P^{\kappa_i}}^{\ell^{\kappa_i}}(f) \text{ and applying Prop. 1 for each } i\text{)}\\
&\le \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\widehat{\mathbb{E}}_S[f]-\widehat{\mathbb{E}}_{S'}[f]\right)\right]\\
&\qquad \text{(by P3 and Jensen's inequality)}\\
&= \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{N} w_i\sum_{\sigma\in\{\pm\}}\frac{1}{2n_i^\sigma}\sum_{j=1}^{n_i^\sigma}\ell_\sigma^{\kappa_i}(f(X_{ij}^\sigma))-\ell_\sigma^{\kappa_i}(f(\underline{X}_{ij}^\sigma))\right)\right]\\
&= \mathbb{E}_{S,S',(\epsilon_{ij}^\sigma)}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{N} w_i\sum_{\sigma\in\{\pm\}}\frac{1}{2n_i^\sigma}\sum_{j=1}^{n_i^\sigma}\epsilon_{ij}^\sigma\Big(\ell_\sigma^{\kappa_i}(f(X_{ij}^\sigma))-\ell_\sigma^{\kappa_i}(f(\underline{X}_{ij}^\sigma))\Big)\right)\right]\\
&\qquad \text{(for all } i,\sigma,j,\ X_{ij}^\sigma \text{ and } \underline{X}_{ij}^\sigma \text{ are iid, and } \epsilon_{ij}^\sigma \text{ are symmetric)}\\
&\le \mathbb{E}_{S,S',(\epsilon_{ij}^\sigma)}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N} w_i\sum_{\sigma\in\{\pm\}}\frac{1}{2n_i^\sigma}\sum_{j=1}^{n_i^\sigma}\epsilon_{ij}^\sigma\ell_\sigma^{\kappa_i}(f(X_{ij}^\sigma))\right]\\
&\qquad + \mathbb{E}_{S,S',(\epsilon_{ij}^\sigma)}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N} w_i\sum_{\sigma\in\{\pm\}}\frac{1}{2n_i^\sigma}\sum_{j=1}^{n_i^\sigma}(-\epsilon_{ij}^\sigma)\ell_\sigma^{\kappa_i}(f(\underline{X}_{ij}^\sigma))\right]\\
&\qquad \text{(by P2)}\\
&= 2\mathbb{E}_S\mathbb{E}_{(\epsilon_{ij}^\sigma)}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N} w_i\sum_{\sigma\in\{\pm\}}\frac{1}{2n_i^\sigma}\sum_{j=1}^{n_i^\sigma}\epsilon_{ij}^\sigma\ell_\sigma^{\kappa_i}(f(X_{ij}))\right].
\end{aligned}
$$

To bound the innermost expectation we use the following result from Meir and Zhang [24].

**Lemma 10.** *Suppose $\{\phi_t\},\{\psi_t\}, t=1,\dots,T$, are two sets of functions on a set $\Theta$ such that for each $t$ and $\theta,\theta'\in\Theta, |\phi_t(\theta)-\phi_t(\theta')|\le|\psi_t(\theta)-\psi_t(\theta')|$. Then for all functions $c:\Theta\to\mathbb{R}$,*

$$
\mathbb{E}_{(\epsilon_t)}\left[\sup_\theta\left\{c(\theta)+\sum_{t=1}^{T}\epsilon_t\phi_t(\theta)\right\}\right]\le\mathbb{E}_{(\epsilon_t)}\left[\sup_\theta\left\{c(\theta)+\sum_{t=1}^{T}\epsilon_t\psi_t(\theta)\right\}\right].
$$

Switching from the single index $t$ to our three indices $i$, $\sigma$, and $j$, we apply the lemma with $\Theta=\mathcal{F}$, $\theta=f$, $c(\theta)=0$, $\phi_{ij}^\sigma(\theta)=\frac{w_i}{2n_i^\sigma}\ell_\sigma^{\kappa_i}(f(X_{ij}^\sigma))$, and $\psi_{ij}^\sigma(\theta)=\frac{w_i|\ell|}{2n_i^\sigma(1-\kappa_i^--\kappa_i^+)}f(X_{ij}^\sigma)$, where we use $|\ell_\sigma^{\kappa_i}|\le|\ell|/(1-\kappa_i^--\kappa_i^+)$. This yields

$$
\begin{aligned}
\mathbb{E}\left[\xi^+\right] &\le 2\mathbb{E}_S\mathbb{E}_{(\epsilon_{ij}^\sigma)}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{N}\frac{w_i|\ell|}{1-\kappa_i^--\kappa_i^+}\sum_{\sigma\in\{\pm\}}\frac{1}{2n_i^\sigma}\sum_{j=1}^{n_i^\sigma}\epsilon_{ij}^\sigma f(X_{ij}^\sigma)\right]\\
&= 2\mathfrak{R}_c^I(\mathcal{F}),
\end{aligned}
$$

To see the second inequality in (3), by **(SR)** we have

$$
\begin{aligned}
2\mathfrak{R}^I_c(\mathcal{F}) &\le 2B|\ell|\sqrt{\sum_{i,\sigma,j}\left(\frac{w_i}{2n_i^\sigma(1-\kappa_i^- - \kappa_i^+)}\right)^2} \\
&= 2B|\ell|\sqrt{\sum_i \frac{w_i^2}{4(1-\kappa_i^- - \kappa_i^+)^2}\sum_\sigma \frac{1}{n_i^\sigma}} \\
&= 2B|\ell|\sqrt{\sum_i \frac{w_i^2}{2\bar{n}_i(1-\kappa_i^- - \kappa_i^+)^2}} \\
&= \sqrt{2}B|\ell|\sqrt{\sum_i \frac{w_i^2}{\bar{n}_i(1-\kappa_i^- - \kappa_i^+)^2}},
\end{aligned}
$$

This concludes the proof in the **(IIM)** case.

Now assume **(IBM)** holds. The idea is to apply the bounded difference inequality at the MCM level. If we modify $S$ by replacing $X_{ij}^\sigma$ (with $i$ fixed, $j, \sigma$ variable) with other values $(X_{ij}^\sigma)'$, while leaving all other values in $S$ fixed, then (by P1) $\xi^+$ changes by at most $2w_i(|\ell^{\kappa_i}|_0 + |\ell^{\kappa_i}|A)$, and we obtain that with probability at least $1-\delta/2$ over the draw of $S$,

$$
\begin{aligned}
\xi^+ - \mathbb{E}\left[\xi^+\right] &\le \sqrt{\sum_{i=1}^N w_i^2(|\ell^{\kappa_i}|_0 + |\ell^{\kappa_i}|A)^2\frac{\log(2/\delta)}{2}} \\
&\le (1+A|\ell|)\sqrt{\frac{\log(2/\delta)}{2}}\sqrt{\sum_{i=1}^N \frac{w_i^2}{(1-\kappa_i^- - \kappa_i^+)^2}}.
\end{aligned}
$$

To bound $\mathbb{E}\left[\xi^+\right]$, we use the same reasoning as in the **(IIM)** case to arrive at

$$
\mathbb{E}\left[\xi^+\right] \le 2\mathbb{E}_S\mathbb{E}_{(\epsilon_i)}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^N w_i\epsilon_i \sum_{\sigma\in\{\pm\}}\frac{1}{2n_i^\sigma}\sum_{j=1}^{n_i^\sigma}\ell_\sigma^{\kappa_i}(f(X_{ij}))\right],
$$

where now there is a Rademacher variable for every bag. The inner two summations may be expressed

$$
\mathbb{E}_{(\sigma,X)\sim\widehat{P}^{\kappa_i}}\left[\ell_\sigma^{\kappa_i}(f(X))\right]
$$

and so by Jensen's inequality and Lemma 10 we have

$$
\begin{aligned}
\mathbb{E}\left[\xi^+\right] &\le 2\mathbb{E}_S\mathbb{E}_{(\epsilon_i)}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^N w_i\mathbb{E}_{(\sigma,X)\sim\widehat{P}^{\kappa_i}}\left[\ell_\sigma^{\kappa_i}(f(X))\right]\right] \\
&\le 2\mathbb{E}_S\mathbb{E}_{((\sigma_i,X_i)\sim\widehat{P}^{\kappa_i})_{i\in[N]}}\mathbb{E}_{(\epsilon_i)}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^N \epsilon_i w_i \ell_{\sigma_i}^{\kappa_i}(f(X_i))\right] \\
&\le 2\mathbb{E}_S\mathbb{E}_{((\sigma_i,X_i)\sim\widehat{P}^{\kappa_i})_{i\in[N]}}\mathbb{E}_{(\epsilon_i)}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^N \epsilon_i \frac{w_i|\ell|}{1-\kappa_i^- - \kappa_i^+}f(X_i)\right] \\
&= 2\mathfrak{R}^B_c(\mathcal{F})
\end{aligned}
$$

This proves the first inequality. To prove the second, by **(SR)** we have

$$
2\mathfrak{R}^B_c(\mathcal{F}) \le 2B|\ell|\sqrt{\sum_i \frac{w_i^2}{(1-\kappa_i^- - \kappa_i^+)^2}}.
$$

This concludes the proof.

## B.4 Proof of Theorem 6

We begin by stating a generalization of Chernoff's bound to correlated binary random variables [27, 16].

**Lemma 11.** *Let $Z_1, \ldots, Z_m$ be binary random variables. Suppose there exists $0 \le \tau \le 1$ such that for all $I \subset [m]$, $\mathbb{P}(\prod_{i \in I} Z_i = 1) \le \tau^{|I|}$. Then for any $\epsilon \ge 0$, $\mathbb{P}(\sum_{i=1}^m Z_i \ge m(\tau + \epsilon)) \le e^{-2m\epsilon^2}$.*

We will first prove the theorem for BP. The result for dominating schemes will then follow easily. Thus, assume the $K$-merging scheme is BP. For now assume **(CIBM)** , which is implied by **(CIIM)** .

Let $\widehat{\gamma}_{ik}^+$ be the larger of the two *empirical* label proportions within the $k$th pair of small bags within the $i$th pair of big bags, and similarly let $\widehat{\gamma}_{ik}^-$ be the smaller. Also let $\gamma_{ik}^+$ be the larger of the two *true* label proportions within the $k$th pair of small bags within the $i$th pair of big bags, and similarly let $\gamma_{ik}^-$ be the smaller.

Let $\epsilon_0 \in (0, \Delta(1-\tau))$ and let $\epsilon \in (0, \frac{\Delta(1-\tau)-\epsilon_0}{1+\Delta}]$. For $i \in [M]$, let $K_i$ be the number of original pairs in the $i$th block (the $i$th pair of big bags) for which $|\gamma_{ik}^+ - \gamma_{ik}^-| \ge \Delta$, $k \in [K]$ and define $\Omega_{\boldsymbol{\gamma},i}$ to be the event that $K_i \ge K(1-\tau-\epsilon)$. By Lemma 11 and **(LP)** , we have $\Pr_{\boldsymbol{\gamma}}(\Omega_{\boldsymbol{\gamma},i}^c) \le e^{-2K\epsilon^2}$.

Also define $\Omega_{\boldsymbol{Y},i}$ to be the event that $\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- \ge \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\gamma}}[\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^-] - \epsilon = \Gamma_i^+ - \Gamma_i^- - \epsilon$. Note that conditioned on $\boldsymbol{\gamma}$, $\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- = \frac{1}{K}\sum_{k=1}^K (\widehat{\gamma}_{ik}^+ - \widehat{\gamma}_{ik}^-)$ is the sum of $K$ independent random variables with range $[0,1]$ (here we use the definition of BP and conditional independence of the small bags under **(CIBM)** ). By Hoeffding's inequality, $\mathbb{P}_{\boldsymbol{Y}|\boldsymbol{\gamma}}(\Omega_{\boldsymbol{Y},i}^c) \le e^{-2K\epsilon^2}$.

Now define $\Omega_{\boldsymbol{\gamma}} := \bigcap_{i=1}^M \Omega_{\boldsymbol{\gamma},i}$ and $\Omega_{\boldsymbol{Y}} := \bigcap_{i=1}^M \Omega_{\boldsymbol{Y},i}$. Also define $\Theta$ to be the event that the first inequality in (4) does not hold. Then

$$
\begin{aligned}
\mathbb{P}(\Theta) &\le \mathbb{P}(\Theta|\Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}) + \mathbb{P}((\Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}})^c) \\
&\le \mathbb{P}(\Theta|\Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}) + \mathbb{P}(\Omega_{\boldsymbol{\gamma}}^c) + \mathbb{P}(\Omega_{\boldsymbol{Y}}^c) \\
&\le \mathbb{P}(\Theta|\Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}) + \frac{N}{K}e^{-2K\epsilon^2} + \mathbb{E}_{\boldsymbol{\gamma}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\gamma}}\left[\mathbf{1}_{\{\Omega_{\boldsymbol{Y}}^c\}}\right] \\
&\le \mathbb{P}(\Theta|\Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}) + \frac{2N}{K}e^{-2K\epsilon^2} \\
&= \mathbb{E}_{\boldsymbol{\gamma},\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{Y}}\left[\mathbf{1}_{\{\Theta\}}|\boldsymbol{\gamma},\boldsymbol{Y}\right]|\Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}\ \right] + \frac{2N}{K}e^{-2K\epsilon^2}.
\end{aligned}
$$

We next bound the inner expectation of the last line above, which is the conditional probability of $\Theta$ given fixed values of $(\boldsymbol{\gamma}, \boldsymbol{Y}) \in \Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}$. We will bound this probability the same argument as in the proof of Thm. 5. To apply that argument, we first need to confirm two things: Conditioned on $\boldsymbol{\gamma}, \boldsymbol{Y}$, (1) for each $i$, $\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- > 0$, and (2) the empirical error $\tilde{\mathcal{E}}(f)$ is an unbiased estimate of $\mathcal{E}_P^\ell$. The first property is given by the following.

**Lemma 12.** *Conditioned on $(\boldsymbol{\gamma}, \boldsymbol{Y}) \in \Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}$, for all $i \in [M]$*

$$
\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- \ge \Gamma_i^+ - \Gamma_i^- - \epsilon \ge \epsilon_0.
$$

*Proof.* Fix $(\boldsymbol{\gamma}, \boldsymbol{Y}) \in \Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}$. Let $i \in [M]$. By definition of $\Omega_{\boldsymbol{Y}}$,

$$
\begin{aligned}
\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- &\ge \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\gamma}}[\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^-] - \epsilon \\
&= \left(\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\gamma}}[\widehat{\gamma}_{ik}^+ - \widehat{\gamma}_{ik}^-]\right) - \epsilon \\
&\ge \left(\frac{1}{K}\sum_{k=1}^K \gamma_{ik}^+ - \gamma_{ik}^-\right) - \epsilon.
\end{aligned}
$$

To see the last step, let $U$ and $V$ be random variables with means $p$ and $q$. Then $\mathbb{E}[\max(U,V)-\min(U,V)] = \mathbb{E}[|U-V|] \geq |\mathbb{E}[U-V]| = |p-q| = \max(p,q)-\min(p,q)$, by Jensen's inequality. Here we have again used the definitions of BP and **(CIBM)** .

By definition of $\Omega_{\boldsymbol{\gamma}}$, $\gamma_{ik}^+ - \gamma_{ik}^- \geq \Delta$ for $K_i \geq K(1-\tau-\epsilon)$ values of $k \in [K]$. From this we conclude that $\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- \geq \Delta(1-\tau-\epsilon) - \epsilon \geq \epsilon_0$, where the last step follows from $\epsilon \leq \frac{\Delta(1-\tau)-\epsilon_0}{1+\Delta}$. □

For the second property, recall $\tilde{\mathcal{E}}(f) = \sum_i w_i \tilde{\mathcal{E}}_i(f)$ with $w \in \Delta^M$ and $w_i \propto (\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^-)^2$. We note that $\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{Y}\in\Omega_{\boldsymbol{\gamma}}\cap\Omega_{\boldsymbol{Y}}}\left[\tilde{\mathcal{E}}_i(f)\right]$ is well defined because $|\ell^{\widehat{\kappa}_i}(f(x))|$ is bounded for $x \in \mathcal{X}$. This follows from the assumption $\sup_{f\in\mathcal{F},x\in\mathcal{X}} |f(x)| \leq A < \infty$, the fact that $\ell^{\widehat{\kappa}_i}$ is Lipschitz continuous on $\Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}$ by Lemma 12, and the observation $|\ell^{\widehat{\kappa}_i}(f(x))| \leq |\ell^{\widehat{\kappa}_i}|_0 + |\ell^{\widehat{\kappa}_i}|A$.

**Lemma 13.** *For all* $f \in \mathcal{F}$, $\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{Y}\in\Omega_{\boldsymbol{\gamma}}\cap\Omega_{\boldsymbol{Y}}}\left[\tilde{\mathcal{E}}_i(f)\right] = \mathcal{E}_P^\ell(f)$.

*Proof.* Recall that $X_{mj}$ denotes the $j$th instance in the $m$th original (pre-merging) small bag, $m \in [2N]$, $j \in [n]$, and that $Y_{mj}$ denotes the corresponding label. We have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{Y}\in\Omega_{\boldsymbol{\gamma}}\cap\Omega_{\boldsymbol{Y}}}\left[\tilde{\mathcal{E}}_i(f)\right] \\
&\qquad = \frac{1}{2}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{Y}\in\Omega_{\boldsymbol{\gamma}}\cap\Omega_{\boldsymbol{Y}}}\left[\frac{1}{nK}\sum_{m\in I_i^+}\sum_{j=1}^n \ell_+^{\widehat{\kappa}_i}(f(X_{mj})) + \frac{1}{nK}\sum_{m\in I_i^-}\sum_{j=1}^n \ell_-^{\widehat{\kappa}_i}(f(X_{mj}))\right] \\
&\qquad = \frac{1}{2}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{\gamma},\boldsymbol{Y}\in\Omega_{\boldsymbol{\gamma}}\cap\Omega_{\boldsymbol{Y}}}\Bigg[\widehat{\Gamma}_i^+ \frac{1}{nK\widehat{\Gamma}_i^+}\sum_{m\in I_i^+}\sum_{j:Y_{mj}=1} \ell_+^{\widehat{\kappa}_i}(f(X_{mj})) \\
&\qquad\quad + (1-\widehat{\Gamma}_i^+)\frac{1}{nK(1-\widehat{\Gamma}_i^+)}\sum_{m\in I_i^+}\sum_{j:Y_{mj}=-1} \ell_+^{\widehat{\kappa}_i}(f(X_{mj})) \\
&\qquad\quad + \widehat{\Gamma}_i^- \frac{1}{nK\widehat{\Gamma}_i^-}\sum_{m\in I_i^-}\sum_{j:Y_{mj}=1} \ell_-^{\widehat{\kappa}_i}(f(X_{mj})) \\
&\qquad\quad + (1-\widehat{\Gamma}_i^-)\frac{1}{nK(1-\widehat{\Gamma}_i^-)}\sum_{m\in I_i^-}\sum_{Y_{mj}=-1} \ell_-^{\widehat{\kappa}_i}(f(X_{mj}))\Bigg] \\
&\qquad = \frac{1}{2}\Big\{\widehat{\Gamma}_i^+ \mathbb{E}_{X\sim P_+}\left[\ell_+^{\widehat{\kappa}_i}(f(X))\right] + (1-\widehat{\Gamma}_i^+)\mathbb{E}_{X\sim P_-}\left[\ell_+^{\widehat{\kappa}_i}(f(X))\right] \\
&\qquad\quad + \widehat{\Gamma}_i^- \mathbb{E}_{X\sim P_+}\left[\ell_-^{\widehat{\kappa}_i}(f(X))\right] + (1-\widehat{\Gamma}_i^-)\mathbb{E}_{X\sim P_-}\left[\ell_-^{\widehat{\kappa}_i}(f(X))\right]\Big\} \\
&\qquad = \frac{1}{2}\left\{\mathbb{E}_{X\sim P_+^{\widehat{\kappa}_i}}\left[\ell_+^{\widehat{\kappa}_i}(f(X))\right] + \mathbb{E}_{X\sim P_-^{\widehat{\kappa}_i}}\left[\ell_-^{\widehat{\kappa}_i}(f(X))\right]\right\} \\
&\qquad = \mathcal{E}_P^\ell(f)
\end{aligned}
$$

where the third step uses the definition of **(CIBM)** , and the last step uses Prop. 1 and Lemma 12. □

By Lemmas 12 and Lemma 13, we can apply the argument in the proof of Theorem 5, conditioned on $(\boldsymbol{\gamma},\boldsymbol{Y}) \in \Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}$, with the estimator $\tilde{\mathcal{E}}$ instead of $\widehat{\mathcal{E}}_w$. The only other changes are that in the application of Lemma 10, we use the bound

$$|\ell^{\widehat{\kappa}_i}| \leq \frac{|\ell|}{\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^-} \leq \frac{|\ell|}{\Gamma_i^+ - \Gamma_i^- - \epsilon},$$

and in the final bounds, we upper bound $(\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^-)^{-1}$ by $(\Gamma_i^+ - \Gamma_i^- - \epsilon)^{-1}$.

# C Symmetric Losses

A loss is said to by *symmetric* if there exists a constant $K$ such that for all $t$, $\ell(t,1)+\ell(t,-1)=K$. Examples include the 0-1, sigmoid, and ramp losses. For a symmetric loss, $\ell^\kappa$ simplifies to

$$\ell^\kappa(t,y)=\frac{1}{1-\kappa^+-\kappa^-}\ell(t,y)-\frac{K}{1-\kappa^+-\kappa^-}(\kappa^-\mathbf{1}_{\{y=1\}}+\kappa^+\mathbf{1}_{\{y=-1\}}).$$

Combined with Proposition 1, this yields

$$\mathcal{E}^\ell_{P^\kappa}(f)=(1-\kappa^+-\kappa^-)\mathcal{E}^\ell_P(f)+K\Big(\frac{\kappa^++\kappa^-}{2}\Big).$$

Therefore, the two sides have the same minimizer which implies that the BER is *immune* to label noise under a mutual contamination model. That is, training on the contaminated data without modifying the loss still minimizes the clean BER. This result has been previously observed for the 0/1 loss [25] and general symmetric losses [42, 6]. The above argument gives a simple derivation from Prop. 1.

# D Convexity

We say that the loss $\ell$ is *convex* if, for each $\sigma$, $\ell_\sigma(t)$ is a convex function of $t$. Let $\ell''_\sigma$ denote the second derivative of $\ell$ with respect to its first variable. The condition in (15) below was used by Natarajan et al. [26] to prove a convexity result an unbiased loss in the class-conditional noise setting. Here we prove a version for MCMs.

**Proposition 14.** *Suppose $\kappa^-+\kappa^+<1$ and let $\ell$ be a convex, twice differentiable loss satisfying*

$$\ell''_+(t)=\ell''_-(t). \tag{15}$$

*If $\kappa^\sigma<\frac{1}{2}$ for $\sigma\in\{\pm\}$, then $\ell^\kappa$ is convex.*

Examples of losses satisfying the second order condition include the logistic, Huber, and squared error losses. The result is proved by simply observing

$$\begin{aligned}(\ell^\kappa_\sigma)''(t)&=\ell''_+(t)\frac{1-2\kappa^{-\sigma}}{1-\kappa^--\kappa^+}\\&\geq 0.\end{aligned}$$

The statement about $\widehat{\mathcal{E}}_i(f)$ being convex when $f$ is linear was a holdover from an earlier draft and should be disregarded. In the infinite bag size limit, $\widehat{\mathcal{E}}_i(f)$ converges to $\mathcal{E}^\ell_P(f)$, which is convex in the output of $f$ provided $\ell$ is convex. Sufficient conditions for the convexity of $\widehat{\mathcal{E}}_i(f)$ or $\widehat{\mathcal{E}}_w(f)$ for small bag sizes is an interesting open question.

# E (CIBM') implies (IBM)

Assume that **(CIBM')** holds. To show **(IBM)** , we need to show that for a fixed bag $i$, and for all $j\in[n_i]$, the marginal distribution of $X_{ij}$, conditioned on the bag, is $\gamma_iP_++(1-\gamma_i)P_-$. Thus let $A$ be an arbitrary event. Also let $p_i$ be the joint pmf of $Y_{i1},\ldots,Y_{in_i}$, conditioned on the bag. Without loss of generality let

$j = 1$. We have

$$\begin{aligned}
\mathbb{P}(X_{i1} \in A) &= \mathbb{E}_X[\mathbf{1}_{\{X_{i1}\in A\}}] \\
&= \mathbb{E}_{Y_{i1},\ldots,Y_{in_i}} \mathbb{E}_{X_{i1}|Y_{i1},\ldots,Y_{in_i}} \left[\mathbf{1}_{\{X_{i1}\in A\}}\right] \\
&= \mathbb{E}_{Y_{i1},\ldots,Y_{in_i}} \mathbb{P}_{Y_{i1}}(X_{i1} \in A) \qquad (16) \\
&= \sum_{(y_1,\ldots,y_{n_i})\in\{-1,1\}^{n_i}} \mathbb{P}_{y_1}(X_{i1} \in A) p_i(y_1,\ldots,y_{n_i}) \\
&= P_+(A) \sum_{(y_2,\ldots,y_{n_i})\in\{-1,1\}^{n_i-1}} p_i(1, y_2, \ldots, y_{n_i}) \\
&\quad + P_-(A) \sum_{(y_2,\ldots,y_{n_i})\in\{-1,1\}^{n_i-1}} p_i(-1, y_2, \ldots, y_{n_i}) \\
&= \gamma_i P_+(A) + (1-\gamma_i) P_-(A), \qquad (17)
\end{aligned}$$

where (16) and (17) use **(CIMB')**.

# F Optimal Bag Matching

The bound is minimized by selecting weights

$$w_i \propto \bar{n}_i(\gamma_i^+ - \gamma_i^-)^2,$$

which gives preference to pairs of bags where one bag is mostly $+1$'s (large $\gamma_i^+$) and the other is mostly -1's (small $\gamma_i^-$). With these weights, the **(SR)** bound is proportional to under **(CIIM)**

$$\sqrt{\left(\sum_{i=1}^{N} \bar{n}_i(\gamma_i^+ - \gamma_i^-)^2\right)^{-1}}.$$

Here and below, under **(CIBM')**' substitute $\bar{n}_i \to 1$.

We can optimize the pairing of bags by further optimizing the bound. Consider the unpaired bags $(B_i, \gamma_i)$, $i = 1, \ldots, 2N$. Recall that $\bar{n}_i = \mathrm{HM}(n_i^+, n_i^-)$. We would like to pair each bag to a different bag, forming pairs $(\gamma_i^+, \gamma_i^-)$, such that

$$\sum_{i=1}^{N} \bar{n}_i(\gamma_i^+ - \gamma_i^-)^2$$

is maximized. For each $i < j$, let $u_{ij}$ be a binary variable, with $u_{ij} = 1$ indicating that the $i$th and $j$th bags are paired. The optimal pairing of bags is given by the solution to the following integer program:

$$\begin{aligned}
\max_u \quad & \sum_{1\le i<2N} \sum_{i<j\le 2N} \mathrm{HM}(n_i, n_j)(\gamma_i - \gamma_j)^2 u_{ij} \qquad (18) \\
\text{s.t.} \quad & u_{ij} \in \{0,1\}, \forall i, j \\
& \sum_{i<j} u_{ij} + \sum_{j<i} u_{ji} = 1, \forall i
\end{aligned}$$

The equality constraint ensures that every bag is paired with precisely one other distinct bag. This problem is known as the "maximum weighted (perfect) matching" problem. An exact algorithm to solve it was given by Edmonds [13], and several approximate algorithms also exist for large scale problems.

When $n_i^\sigma = n$ for all $i$ and $\sigma$, the solution to this integer program is very simple.

**Proposition 15.** *If $n_i^\sigma = n$ for all $i$ and $\sigma$, then the solution to* (18) *is to match the largest $\gamma_i$ with the smallest, the second largest $\gamma_i$ with the second smallest, and so on.*

*Proof.* Suppose the statement is false. Then there exists an optimal solution, and $i$ and $j$, such that $\gamma_i^+ > \gamma_j^+$ and $\gamma_i^- > \gamma_j^-$. Now consider the matching obtained by swapping the bags associated to $\gamma_i^-$ and $\gamma_j^-$. Then the objective function increases by

$$(\gamma_i^+ - \gamma_j^-)^2 + (\gamma_j^+ - \gamma_i^-)^2 - (\gamma_i^+ - \gamma_i^-)^2 - (\gamma_j^+ - \gamma_j^-)^2 = 2(\gamma_i^+ - \gamma_j^+)(\gamma_i^- - \gamma_j^-) > 0.$$

This contradicts the assumed optimality. □

# G Merging Schemes that Dominate Blockwise-Pairwise

Let $\underline{\Gamma}_i^+$ and $\underline{\Gamma}_i^-$ denote the quantities $\Gamma_i^+$ and $\Gamma_i^-$ when the merging scheme is BP, and let $\Gamma_i^+$ and $\Gamma_i^-$ refer to any other merging scheme under consideration. Similarly, let $\widehat{\underline{\Gamma}}_i^+$ and $\widehat{\underline{\Gamma}}_i^-$ denote the quantities $\widehat{\Gamma}_i^+$ and $\widehat{\Gamma}_i^-$ when the merging scheme is BP, and let $\widehat{\Gamma}_i^+$ and $\widehat{\Gamma}_i^-$ refer to any other merging scheme under consideration.

For a $K$-merging scheme that dominates BP, we still have $\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- \geq \underline{\Gamma}_i^+ - \underline{\Gamma}_i^- - \epsilon \geq \epsilon_0 > 0$ on $\Omega_{\boldsymbol{\gamma}} \cap \Omega_{\boldsymbol{Y}}$ by definition of dominating. Hence the same proof goes through in this case, and we may state the following.

**Theorem 16.** *Let* **(LP)** *hold. Let $\epsilon_0 \in (0, \Delta(1-\tau))$. Let $\ell$ be a Lipschitz loss and let $\mathcal{F}$ satisfy $\sup_{x\in\mathcal{X}, f\in\mathcal{F}} |f(x)| \leq A < \infty$. Let $\epsilon \in (0, \frac{\Delta(1-\tau)-\epsilon_0}{1+\Delta}]$ and $\delta \in (0,1]$. For any $K$-merging scheme that dominates BP, under* **(CIIM)** *, with probability at least $1 - \delta - 2\frac{N}{K}e^{-2K\epsilon^2}$ with respect to the draw of $\boldsymbol{\gamma}, \boldsymbol{Y}, \boldsymbol{X}$,*

$$\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- \geq \underline{\Gamma}_i^+ - \underline{\Gamma}_i^- - \epsilon \geq \epsilon_0$$

*and*

$$\sup_{f\in\mathcal{F}} \left|\tilde{\mathcal{E}}(f) - \mathcal{E}(f)\right| \leq 2\mathfrak{R}_c^I(\mathcal{F}) + C\sqrt{\frac{\mathrm{HM}((\underline{\Gamma}_i^+ - \underline{\Gamma}_i^- - \epsilon)^{-2})}{(N/K)n}} \overset{\textbf{(SR)}}{\leq} D\sqrt{\frac{\mathrm{HM}((\underline{\Gamma}_i^+ - \underline{\Gamma}_i^- - \epsilon)^{-2})}{(N/K)n}}, \tag{19}$$

*where $c_i = w_i|\ell|/(\underline{\Gamma}_i^+ - \underline{\Gamma}_i^- - \epsilon)$, $C = (1 + A|\ell|)\sqrt{\log(2/\delta)}$, and $D = 2B|\ell| + C$. Under* **(CIBM)***, the same bounds hold with the same probability if we substitute $\mathfrak{R}_c^I(\mathcal{F}) \to \mathfrak{R}_c^B(\mathcal{F})$ and $n \to 1$.*

We conjecture that it is possible to improve the bound for dominating schemes. Using the current proof technique, this would require proving that

$$\widehat{\Gamma}_i^+ - \widehat{\Gamma}_i^- \geq \Gamma_i^+ - \Gamma_i^- - \epsilon$$

with high probability. For example, with BM, this would require a one-sided tail inequality for how the difference between the average of the larger half and the average of the smaller half of $2K$ independent random variables deviates from its mean. The BP scheme was selected as a reference because it is straightforward to prove such a bound for BP using Hoeffding's inequality.

# H Consistency

A discrimination rule $\widehat{f}$ is (weakly) consistent if $\mathcal{E}_P^\ell(\widehat{f}) \to \inf_f \mathcal{E}_P^\ell(f)$ in probability as $N \to \infty$, where the infimum is over all decision functions.

We first note that if we desire consistency wrt the BER defined with 0-1 loss, it suffices to prove consistency wrt the BER defined with a loss $\ell$ that is "classification calibrated" [2]. This is because the BER corresponds to a special case of the usual misclassification risk when the class probabilities are equal. Thus, let $\ell$ be Lipschitz and classification calibrated, such as the logistic loss.

We state our consistency result for the discrimination rule

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} J(f) := \tilde{\mathcal{E}}(f) + \lambda \|f\|^2_{\mathcal{F}_k},$$

where $\mathcal{F}_k$ is the reproducing kernel Hilbert space associated to a symmetric, positive definite kernel, and $\lambda > 0$.

**Theorem 17.** *Let $\mathcal{X}$ be compact and let $k$ be a bounded, universal kernel on $\mathcal{X}$. Let $K \to \infty$ such that $N/K \to \infty$ and $N = O(K^\beta)$ for some $\beta > 0$, as $N \to \infty$. Let $\lambda$ be such that $\lambda \to 0$ and $\lambda(N/K)/\log(N/K) \to \infty$ as $N \to \infty$. Let* **(LP)** *and* **(CIBM)** *hold. Then for any merging scheme that dominates BP,*

$$\mathcal{E}(\widehat{f}) \to \inf_f \mathcal{E}^\ell_P(f) \tag{20}$$

*in probability as $N \to \infty$.*

*Proof.* Let $B$ denote the bound on the kernel. By Proposition 4 and by Theorem 16 applied to $\mathcal{F}^k_{B,R}$, for all $\epsilon_0 \in (0, \Delta(1-\tau))$, $\epsilon \in (0, \frac{\Delta(1-\tau)-\epsilon_0}{1+\Delta}]$, and $\delta \in (0,1]$, with probability at least $1 - \delta - \frac{N}{K} e^{-2K\epsilon^2}$,

$$\sup_{f \in B_k(R)} \left| \tilde{\mathcal{E}}(f) - \mathcal{E}^\ell_P(f) \right| \leq \frac{D}{\epsilon_0} \sqrt{\frac{K}{N}}$$

where $D = (1 + RB|\ell|)\sqrt{\log(2/\delta)} + 2RB|\ell|$.

Observe that $J(\widehat{f}) \leq J(0) \leq \frac{|\ell|_0}{\epsilon_0}$. Therefore $\lambda \|\widehat{f}\|^2 \leq \frac{|\ell|_0}{\epsilon_0} - \tilde{\mathcal{E}}(\widehat{f}) \leq \frac{2|\ell|_0}{\epsilon_0}$ and so $\|\widehat{f}\|^2 \leq \frac{2|\ell|_0}{\epsilon_0 \lambda}$.

Set $R = \sqrt{\frac{2|\ell|_0}{\epsilon_0 \lambda}}$. Note that $R$ grows asymptotically because $\lambda$ shrinks. We just saw that $\widehat{f} \in B_k(R)$.

Let $\epsilon > 0$. Fix $f_\epsilon \in \mathcal{F}_k$ s.t. $\mathcal{E}^\ell_P(f_\epsilon) \leq \inf_f \mathcal{E}^\ell_P + \epsilon/2$, possible since $k$ is universal [38]. Note that $f_\epsilon \in B_k(R)$ for $N$ sufficiently large. In this case the generalization error bound implies that with probability $\geq 1 - \delta - \frac{N}{K} e^{-2K\epsilon^2}$,

$$\begin{aligned}
\mathcal{E}^\ell_P(\widehat{f}) &\leq \tilde{\mathcal{E}}(\widehat{f}) + \frac{D}{\epsilon_0}\sqrt{\frac{K}{N}} \\
&\leq \tilde{\mathcal{E}}(f_\epsilon) + \lambda\|f_\epsilon\|^2 - \lambda\|\widehat{f}\|^2 + \frac{D}{\epsilon_0}\sqrt{\frac{K}{N}} \\
&\leq \tilde{\mathcal{E}}(f_\epsilon) + \lambda\|f_\epsilon\|^2 + \frac{D}{\epsilon_0}\sqrt{\frac{K}{N}} \\
&\leq \mathcal{E}^\ell_P(f_\epsilon) + \lambda\|f_\epsilon\|^2 + \frac{2D}{\epsilon_0}\sqrt{\frac{K}{N}}.
\end{aligned}$$

Taking $\delta = K/N$, the result now follows. □

# I Experimental Details

The parameters of InvCal [32] and alter-∝SVM [44] are tuned by five-fold cross validation. We only consider the RBF kernel. Following [44], the parameters for both methods were set as follows. The kernel bandwidth $\gamma$ of the RBF kernel is chosen from $\{0.01, 0.1, 1\}$. For InvCal, the parameters are tuned from $C_p \in \{0.1, 1, 10\}$, and $\epsilon \in \{0, 0.01, 0.1\}$. For alter-∝SVM, the parameters are tuned from $C \in \{0.1, 1, 10\}$, and $C_p \in \{1, 10, 100\}$.

A Matlab implementation of both InvCal and alter-∝SVM was obtained online.[3] These implementations rely on LIBSVM[4] and CVX[5]. We modified the code to preform parameter tuning with cross validation as

[3] https://github.com/felixyu/pSVM

[4] https://www.csie.ntu.edu.tw/ cjlin/libsvm/

[5] http://cvxr.com/cvx/

described above. LIBSVM contains its own random number generator that was unfortunately not seeded and hence the results for alter-$\propto$SVM are not reproducible.

For the MAGIC dataset, InvCal takes roughly 30 minutes on 36 cores to complete the experiments for all bag sizes. For the Adult dataset, InvCal takes roughly 60 minutes on 36 cores. For alter-$\propto$-SVM, the approximated runtime on MAGIC dataset is 70 minutes on 144 cores. On Adult dataset, it is 100 minutes on 144 cores.

All three algorithms require random initialization. Yu et al. [44] randomly initialize their algorithm ten times and take the result with smallest objective value. This was deemed to be computationally excessive, and hence we only consider one random initialization for each method. This could account for the relatively poor performance of alter-$\propto$SVM compared to past reported performance.

We found that in some cases, the code for alter-$\propto$-SVM wouldn't create a variable 'support_v', which is used to predict the test label. This resulted from LIBSVM not returning any support vectors. If 'support_v' did not exist for a given fold, we excluded that fold from the cross-validation error estimate.

For bag size 8, in the experiments with fixed number of bags, on a handful of occasions there are only two bags in the validation data within a given fold of cross-validation, and both bags have the same label proportion. When this occurs, we cannot compute our criterion, and exclude such folds.

# References

[1] Ehsan M. Ardehaly and Aron Culotta. Co-training for demographic classification using deep learning from label proportions. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1017–1024, 2017.

[2] P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *J. American Statistical Association*, 101(473):138–156, 2006.

[3] G. Blanchard and C. Scott. Decontamination of mutually contaminated models. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[4] G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10:2780–2824, 2016.

[5] Gerda Bortsova, Florian Dubost, Silas Ørting, Ioannis Katramados, Laurens Hogeweg, Laura Thomsen, Mathilde Wille, and Marleen de Bruijne. Deep learning from label proportions for emphysema quantification. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 768–776, 2018.

[6] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 961–970, 2019.

[7] S. Chen, B. Liu, M. Qian, and C. Zhang. Kernel k-means based framework for aggregate outputs classification. In *2009 IEEE International Conference on Data Mining Workshops*, pages 356–361, 2009.

[8] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *Proceedings of the 22nd ACM International Conference on Multimedia*, page 367–376, 2014.

[9] Zhensong Chen, Zhiquan Qi, Bo Wang, Limeng Cui, Fan Meng, and Yong Shi. Learning with label proportions based on nonparallel support vector machines. *Knowledge-Based Systems*, 119:126 – 141, 2017.

[10] Lucio Mwinmaarong Dery, Benjamin Nachman, Francesco Rubbo, and Ariel Schwartzman. Weakly supervised classification for high energy physics. *Journal of Physics: Conference Series*, 1085(4), 2018.

[11] Yongke Ding, Yuanxiang Li, and Wenxian Yu. Learning from label proportions for SAR image classification. *EURASIP Journal on Advances in Signal Processing*, 2017.

[12] Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. Deep multi-class learning from label proportions. *ArXiv*, abs/1905.12909, 2019.

[13] Jack Edmonds. Maximum matching and a polyhedron with 0, 1-vertices. *Journal of Research of the National Bureau of Standards B*, 69:125–130, 1965.

[14] J. Hernández-González, I. Inza, and J. A. Lozano. Learning Bayesian network classifiers from label proportions. *Pattern Recognition*, 46(12):3425 – 3440, 2013.

[15] J. Hernández-González, I. Inza, Lorena Crisol-Ortíz, M. A. Guembe, M. J. Iñarra, and J. A. Lozano. Fitting the data from embryo implantation prediction: Learning from label proportions. *Statistical Methods in Medical Research*, 27(4):1056–1066, 2018.

[16] Russell Impagliazzo and Valentine Kabanets. Constructive proofs of concentration bounds. In Maria Serna, Ronen Shaltiel, Klaus Jansen, and José Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 617–631, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[17] Julian Katz-Samuels, Gilles Blanchard, and Clayton Scott. Decontamination of mutual contamination models. *Journal of Machine Learning Research*, 20(41):1–57, 2019. URL `http://jmlr.org/papers/v20/17-576.html`.

[18] Hendrik Kück and Nando de Freitas. Learning about individuals from group statistics. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, page 332–339, 2005.

[19] K. Lai, F. X. Yu, M. Chen, and S. Chang. Video event detection by inferring temporal instance labels. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2251–2258, 2014.

[20] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer, 1991.

[21] Fan Li and Graham Taylor. Alter-CNN: An approach to learning from label proportions with application to ice-water classification. In *Neural Information Processing Systems Workshops (NIPSW) on Learning and privacy with incomplete data and weak supervision*, 2015.

[22] Jiabin Liu, Bo Wang, Zhiquan Qi, YingJie Tian, and Yong Shi. Learning from label proportions with generative adversarial networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7169–7179. 2019.

[23] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141:148–188, 1989.

[24] R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

[25] A. Menon, B. Van Rooyen, C. S. Ong, and R. Williamson. Learning from corrupted binary labels via class-probability estimation. In F. Bach and D. Blei, editors, *Proc. 32th Int. Conf. Machine Learning (ICML)*, Lille, France, 2015.

[26] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Cost-sensitive learning with noisy labels. *Journal of Machine Learning Research*, 18(155):1–33, 2018. URL `http://jmlr.org/papers/v18/15-226.html`.

[27] Alessandro Panconesi and Aravind Srinivasan. Randomized distributed edge coloring via an extension of the Chernoff–Hoeffding bounds. *SIAM J. Comput.*, 26(2):350–368, 1997.

[28] Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (Almost) No label no cry. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 190–198. 2014.

[29] R. Poyiadzi, R. Santos-Rodriguez, and N. Twomey. Label propagation for learning with label proportions. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2018.

[30] Zhiquan Qi, Bo Wang, Fan Meng, and Lingfeng Niu. Learning with label proportions via NPSVM. *IEEE Transactions on Cybernetics*, 47:3293–3305, 2017.

[31] 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, 2009.

[32] Stefan Rueping. SVM classifier estimation from group probabilities. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 911–918, 2010.

[33] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 416–426. Springer Berlin Heidelberg, 2001.

[34] C. Scott, G. Blanchard, and G. Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proc. Conf. on Learning Theory,* JMLR W&CP, volume 30, pages 489–511. 2013.

[35] Clayton Scott and Jianxin Zhang. Learning from multiple corrupted sources, with application to learning from label proportions. *ArXiv*, abs/1910.04665v1, 2019.

[36] Yong Shi, Limeng Cui, Zhensong Chen, and Zhiquan Qi. Learning from label proportions with pinball loss. *International Journal of Machine Learning and Cybernetics*, 10:187–205, 2017.

[37] Yong Shi, Jiabin Liu, Zhiquan Qi, and Bo Wang. Learning from label proportions on high-dimensional data. *Neural Networks*, 103:9 – 18, 2018.

[38] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, 2008.

[39] Marco Stolpe and Katharina Morik. Learning from label proportions by optimizing cluster model selection. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 349–364, 2011.

[40] T. Sun, D. Sheldon, and B. O'Connor. A probabilistic approach for learning with label proportions applied to the us presidential election. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 445–454, 2017.

[41] Kuen-Han Tsai and Hsuan-Tien Lin. Learning from label proportions with consistency regularization, 2020. URL `https://openreview.net/forum?id=SyecdJSKvr`.

[42] Brendan van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. An average classification algorithm. Technical Report arXiv:1506.01520, 2015.

[43] B. Wang, Z. Chen, and Z. Qi. Linear twin SVM for learning from label proportions. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 56–59, 2015.

[44] Felix X. Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. $\propto$SVM for learning with label proportions. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, page III–504–III–512, 2013.

[45] Felix X. Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. Technical Report arXiv:1402.5902, 2015.