Optimistic Rates for Learning from Label Proportions

Gene Li　　　　　　　　　　　　　　　　　　　　　　　　　GENE@TTIC.EDU
Toyota Technological Institute at Chicago

Lin Chen　　　　　　　　　　　　　　　　　　　　　　　LINCHE@GOOGLE.COM
Google Research

Adel Javanmard　　　　　　　　　　　　　　　　　　　AJAVANMA@USC.EDU
Marshall School of Business, University of Southern California
Google Research

Vahab Mirrokni　　　　　　　　　　　　　　　　　MIRROKNI@GOOGLE.COM
Google Research

Editors: Shipra Agrawal and Aaron Roth

Abstract

We consider a weakly supervised learning problem called Learning from Label Proportions (LLP), where examples are grouped into "bags" and only the average label within each bag is revealed to the learner. We study various learning rules for LLP that achieve PAC learning guarantees for classification loss. We establish that the classical Empirical Proportional Risk Minimization (EPRM) learning rule (Yu et al., 2014) achieves fast rates under realizability, but EPRM and similar proportion matching learning rules can fail in the agnostic setting. We also show that (1) a debiased proportional square loss, as well as (2) a recently proposed EasyLLP learning rule (Busa-Fekete et al., 2023) both achieve "optimistic rates" (Panchenko, 2002); in both the realizable and agnostic settings, their sample complexity is optimal (up to log factors) in terms of ε, δ, and VC dimension.

Keywords: weak supervision, learning from label proportions, PAC learning, sample complexity

1. Introduction

We study Learning from Label Proportions (LLP), which is a framework for weakly supervised learning. In the standard supervised learning framework, the learner has access to a dataset of n i.i.d. labeled examples $\{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, and the goal is to learn an accurate predictor $\widehat{f} : \mathcal{X} \mapsto \mathcal{Y}$. In the LLP framework, the learner does not get access to the true labels $\{y_i\}_{i=1}^n$; instead, training data is organized into "bags" which contain multiple unlabeled examples, and only the average (or "aggregated") label within the bag is provided to the learner.

The LLP framework has been studied in a long line of work, dating back to Kück and de Freitas (2005); Musicant et al. (2007). LLP is motivated by practical machine learning problems where individual labels are expensive to obtain or unavailable, see, e.g., applications in high energy physics (Dery et al., 2017), election prediction (Sun et al., 2017), and RADAR image classification (Ding et al., 2017). More recently, LLP was proposed as a mechanism to provide user privacy (Diemert et al., 2022); for example, in the Apple SKAN API (Apple, 2024) and Google Chrome's Private Aggregation API (Google, 2024), only aggregated labels are provided for ad conversion reporting.

Problem Formulation. Let $\mathcal{X}$ represent the instance space and $\mathcal{Y} = \{0, 1\}$ denote the label space. In LLP, we are given n bags of examples $\{(B_i, \alpha_i)\}_{i=1}^{n}$ where each bag $B_i = \{x_{i,j}\}_{j=1}^{k}$ contains k instances and $\alpha_i = \frac{1}{k} \sum_{j=1}^{k} y_{i,j}$ is the average label within the bag. We assume that each instance $(x_{i,j}, y_{i,j}) \sim \mathcal{D}$ is independently and identically distributed (i.i.d.) according to an unknown distribution $\mathcal{D}$. We also use $(B, \alpha) \sim \mathcal{D}$ to indicate that a bag is drawn from $\mathcal{D}$. Unlike supervised learning, in LLP the learner does not get individual labels $y_{i,j}$, but only the aggregated label α_i for a collection of k random instances.

We study the PAC learning objective of finding an accurate instance-level predictor from label proportion data that competes with the best predictor in specified function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$. Given parameters $\varepsilon, \delta \in (0, 1)$, we seek a predictor $\widehat{f} : \mathcal{X} \to \mathcal{Y}$ that satisfies the following condition with probability at least $1 - \delta$:

$$\mathcal{L}(\widehat{f}) \leq \inf_{f \in \mathcal{F}} \mathcal{L}(f) + \varepsilon, \quad \text{where} \quad \mathcal{L}(f) := \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell^{01}(y, f(x))], \tag{1}$$

and $\ell^{01} : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ is the classification loss, defined as $\ell^{01}(y, \widehat{y}) = \mathbb{1}\{y \neq \widehat{y}\}$. When the bag size $k = 1$, this becomes the classic binary classification setup, for which it is known precisely that the VC dimension of $\mathcal{F}$ characterizes the sample complexity of learning.

The fundamental question in LLP is to establish the sample complexity, i.e., the number of bags n required to guarantee Eq. (1), in terms of the VC dimension of $\mathcal{F}$, bag size k, and accuracy parameters ε, δ. Our paper investigates several proposed learning rules designed to directly minimize classification loss. We establish generalization bounds for these learning rules under both the realizable setting (where $\inf_{f \in \mathcal{F}} \mathcal{L}(f) = 0$ and fast $1/n$ rates are possible) and the agnostic setting (where $\inf_{f \in \mathcal{F}} \mathcal{L}(f)$ can be arbitrary, and one gets slow $1/\sqrt{n}$ rates). Specifically, we adopt the optimistic rates framework (Panchenko, 2002; Srebro et al., 2010) which uses localized uniform convergence bounds to show generalization guarantees that interpolate between the realizable and agnostic setting.

Notation. We denote the marginal label proportion $p = \mathbb{P}[y = 1]$. Whenever the function class $\mathcal{F}$ is clear from the context, we denote $d = \text{VC}(\mathcal{F})$. We adopt standard big-oh notation and use $\widetilde{O}(\cdot)$ to hide $\text{poly}(\log k, \log n)$ dependencies in our bounds.

1.1. Our Contributions

We obtain the following results on LLP for the classification objective (1).

Success and Failure of Empirical Proportional Risk Minimization (Section 2): We study the Empirical Proportional Risk Minimization (EPRM) learning rule (Yu et al., 2014), which is a natural extension of Empirical Risk Minimization (ERM) to the LLP setting. Concretely we prove that under realizability, $\widehat{f}_{\text{EPRM}} := \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\frac{1}{k} \sum_{j=1}^{k} f(x_{i,j}) \neq \alpha_t\}$ achieves the sample complexity guarantee

$$n = O\left(\frac{d \log k \cdot \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right).$$

However, in the agnostic setting we show that EPRM cannot attain polynomial (in k) sample complexity, and similar "folklore" learning rules based on minimization of proportional square or log losses even fail to return a predictor with constant suboptimality.

Optimistic Rates for Debiased Square Loss (Section 3): To address the failure of proportional loss learning rules in the agnostic setting, we consider a simple debiased variant of the proportional square loss. We show that the debiased square loss learning rule $\widehat{f}_{\text{DSQ}}$ achieves the optimistic rate:

$$\mathcal{L}(\widehat{f}_{\text{DSQ}}) \leq L^{\star} + \widetilde{O}\left(\frac{k^2(d + \log(1/\delta))}{n} + \sqrt{\frac{L^{\star} \cdot k^2(d + \log(1/\delta))}{n}}\right),$$

where we denote $L^{\star} = \min_{f \in \mathcal{F}} \mathcal{L}(f)$. Here, observe that under realizability (with $L^{\star} = 0$), $\widehat{f}_{\text{DSQ}}$ enjoys a fast $1/n$ rate, while in the agnostic setting we recover the $1/\sqrt{n}$ rate, both of which are optimal (up to log factors) in terms of d, $\log(1/\delta)$, and n.

Optimistic Rates for EasyLLP (Section 4): We study an alternative approach called EasyLLP, which was recently proposed by Busa-Fekete et al. (2023). Specialized to the classification setting, they show a $1/\sqrt{n}$ rate (hiding dependence on d, k, and δ). We improve upon their result to show an optimistic rate similar to the one achieved by the debiased square loss, showing that EasyLLP can indeed adapt to realizability. Our analysis reveals a curious phenomenon: in the realizable setting, EasyLLP exhibits a separation between loss estimation (which is necessarily $\Omega(1/\sqrt{n})$ even for the optimal predictor $f^{\star}$) and learning (which is $\widetilde{O}(1/n)$ by the optimistic rates guarantee).

Lower Bounds (Section 5): We investigate the optimal dependence on the bag size k, since our bounds are tight (up to log factors) in the other parameters. A trivial lower bound of $n = \Omega(1/k)$ follows because LLP with n bags is only harder than supervised learning with nk bags. It turns out this cannot be improved in general for all $\mathcal{F}$, but we give an explicit example of a function class $\mathcal{F}$ for which the minimax sample complexity has larger dependence on k.

Experiments (Appendix G): We empirically evaluate gradient-based versions of the learning rules considered herein on binary classification versions of MNIST and CIFAR10 studied in Busa-Fekete et al. (2023) for a wide range of NN architectures. We find that proportion matching and the debiased square loss perform the best; furthermore, we demonstrate that the debiased square loss enjoys faster optimization than proportion matching, as early in training the debiased square loss is a better estimate of the true instance-level loss.

1.2. Related Work

Learning from Label Proportions. The problem of LLP has been studied in a long line of work (Chen et al., 2006; Musicant et al., 2007; Kück and de Freitas, 2005; Quadrianto et al., 2008). Most works either assume some kind of distributional assumptions on bag/label generation (Kück and de Freitas, 2005; Quadrianto et al., 2008; Patrini et al., 2014; Scott and Zhang, 2020; Zhang et al., 2022), construct bags adaptively (Chen et al., 2023; Javanmard et al., 2024b), or study approaches to minimize a surrogate loss (Rueping, 2010; Yu et al., 2013; Qi et al., 2016; Shi et al., 2019; Dulac-Arnold et al., 2019; Javanmard et al., 2024a). On the computational side, Saket (2021, 2022) shows that even in the realizable setting learning linear thresholds for LLP is NP-hard and study SDP relaxations for this task. Thus, the aforementioned papers are not directly relevant to the goals of this work in providing distribution-free statistical guarantees on classification loss.

Several works provide guarantees on the instance-level classification loss. Yu et al. (2014) introduce the EPRM learning rule. While their main focus is proving guarantees on the proportional risk, they

show how to translate these to instance-level guarantees when the bags are "pure"—meaning the label proportions α_t are close to 0 or 1. They also give numerical bounds which indicate that EPRM achieves instance-level guarantees under realizability. Chen et al. (2023) introduce a similar debiased square loss learning rule and prove a $O(1/\sqrt{n})$ rate in the agnostic setting; while Busa-Fekete et al. (2023) introduce the EasyLLP framework and prove a $O(1/\sqrt{n})$ rate.

Optimistic Rates. Optimistic rates date back to seminal work of Vapnik and Chervonenkis (2015) and were expanded upon by Bousquet (2002); Koltchinskii and Panchenko (2000); Bartlett et al. (2005) using the technique of localized Rademacher complexities. Optimistic rates have been studied in various other contexts such as optimization with smooth losses (Srebro et al., 2010), multi-task learning (Yousefi et al., 2018; Watkins et al., 2023), vector-valued learning (Reeve and Kaban, 2020), and overparameterized regression (Zhou et al., 2021, 2022, 2023).

2. Empirical Proportional Risk Minimization

The most direct approach to finding a good predictor is the so-called *empirical proportional risk minimization* (EPRM) approach, which simply returns a predictor that matches the most label proportions (Yu et al., 2014). Concretely, we can consider the learning rule

$$\widehat{f}_{\mathrm{EPRM}} := \underset{f \in \mathcal{F}}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left\{ \frac{1}{k} \sum_{j=1}^{k} f(x_{i,j}) \neq \alpha_i \right\}. \tag{2}$$

Observe that in the setting of $k = 1$, EPRM recovers the classical empirical risk minimizer (ERM).

More generally, one can also define learning rules which minimize some other bag loss $\ell : [0, 1] \times [0, 1] \to \mathbb{R}$ between the predicted label proportions $\widehat{\alpha} = \frac{1}{k} \sum_i f(x_i)$ and the true label proportions α, i.e., the square loss $\ell_{\mathrm{SQ}}(\alpha, \widehat{\alpha}) = (\alpha - \widehat{\alpha})^2$ or the log loss $\ell_{\mathrm{LOG}}(\alpha, \widehat{\alpha}) = -\alpha \log \widehat{\alpha} - (1 - \alpha) \log(1 - \widehat{\alpha})$. In the literature, these are also called EPRM or proportion matching learning rules, and they are a "folklore" approach which, in conjunction with gradient-based methods, are competitive in practice (Busa-Fekete et al., 2023). To disambiguate, we exlusively refer to the learning rule (2) as EPRM and call the more general class of these learning rules as proportion matching.

In this section, we provide theoretical results which substantiate conventional wisdom surrounding EPRM. First, we show that under realizability, the EPRM learning rule attains fast rates for classification. However, in the agnostic setting, we illustrate that proportion matching can be ill-behaved, as we demonstrate an example for which minimizing the proportional risk gives no guarantees on the instance-level classification performance.

2.1. Fast Rates for EPRM Under Realizability

We show the following generalization guaranteee for EPRM under realizability.

Theorem 1 *Let $\mathcal{F}$ be a symmetric function class, i.e. if $f \in \mathcal{F}$, then $1 - f \in \mathcal{F}$. Let bag size $k \geq 11$, $\varepsilon \in (0, 1/(4k))$, and $\delta \in (0, 1)$. As long as $n = O(\frac{d \log k \cdot \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon})$, for any realizable distribution $\mathcal{D}$, with probability at least $1 - \delta$ over the draw of the sample, $\mathcal{L}(\widehat{f}_{\mathrm{EPRM}}) \leq \varepsilon$.*

Previous works have suggested that EPRM (or more generally, proportion matching learning rules) can succeed under realizability: Yu et al. (2014) give numerical evidence to show that a guarantee

4

on bag proportions can be translated to a guarantee on the instance-level classification error, and Busa-Fekete et al. (2023) show that under some conditions on the loss function, minimizers for a population proportion matching loss are also minimizers for the instance-level loss. However, to the best of our knowledge, Theorem 1 is the first result which provides a concrete generalization bound for bounded VC classes for the EPRM learning rule.

A few comments about Theorem 1 are in order.

- The assumption that $\mathcal{F}$ is symmetric is mild and due to technical reasons. Note that any nonsymmetric $\mathcal{F}$ can be enlarged to a symmetric one with VC dimension at most $2d + 1$.

- The bag size assumption is technically required, and we conjecture it can be removed. It is a mild assumption since if one has bags of size $k \leq 10$, then one can preprocess the dataset to combine sets of bags to form a larger bag of size at least 11, and then compute $\widehat{f}_{\mathrm{EPRM}}$ on the preprocessed dataset. This achieves the same guarantee (albeit with smaller range of ε).

Proof of Theorem 1. The proof has three steps. First, we show via standard uniform convergence arguments that the predictor $\widehat{f}_{\mathrm{EPRM}}$ must have small population proportional matching error. Next, we relate the proportion matching error to the classification error to show that $\widehat{f}_{\mathrm{EPRM}}$ must have classification error $\mathcal{L}(\widehat{f}_{\mathrm{EPRM}}) \notin [\varepsilon, 1 - \varepsilon]$. Finally, we show that $\widehat{f}_{\mathrm{EPRM}}$ must have classification error $\mathcal{L}(\widehat{f}_{\mathrm{EPRM}}) \leq \varepsilon$, as otherwise we would have selected the predictor $1 - \widehat{f}_{\mathrm{EPRM}}$.

Step 1. We rewrite our problem as a binary classification problem and apply standard uniform convergence guarantees. Define the function class $\mathcal{G} : \mathcal{X}^n \times [0, 1] \to \{0, 1\}$ as

$$\mathcal{G} = \left\{ g_f : (B, \alpha) \mapsto \mathbb{1}\left\{ \frac{1}{k} \sum_{j=1}^{k} f(x_j) \neq \alpha \right\} : f \in \mathcal{F} \right\}.$$

We claim that $\mathrm{VC}(\mathcal{G}) \leq O(d \log k)$. To show this, consider any $X = \{(B_1, \alpha_1), \cdots, (B_m, \alpha_m)\}$, and define the projection of $\mathcal{G}$ onto X as $\mathcal{G}_X := \{(g_f(B_1, \alpha_1), \cdots, g_f(B_m, \alpha_m)) : f \in \mathcal{F}\}$. It suffices to show that when $m = O(d \log k)$, the set of labellings for X is of size $|\mathcal{G}_X| < 2^m$. Observe that the labelling of X is determined by the labellings of $\mathcal{F}$ on the mk points, so by Sauer's lemma $|\mathcal{G}_X| \leq (emk/d)^d$. Therefore, when $m = O(d \log k)$ we have $|\mathcal{G}_X| < 2^m$.

The EPRM can be written as $\widehat{f}_{\mathrm{EPRM}} := \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell^{01}(0, g_f(B_i, \alpha_i))$. Applying the uniform convergence guarantee for VC classes (e.g., Shalev-Shwartz and Ben-David, 2014), we see that as long as $n = O(\frac{d \log k \cdot \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon})$, with probability at least $1 - \delta$ we have

$$\mathbb{P}_{(B,\alpha) \sim \mathcal{D}} \left[\frac{1}{k} \sum_{j=1}^{k} \widehat{f}_{\mathrm{EPRM}}(x_j) \neq \alpha \right] \leq \varepsilon. \tag{3}$$

Henceforth, we will condition on the event in Eq. (3) holding.

Step 2. Now we show that a proportional risk guarantee of the form Eq. (3) translates to a guarantee on the instance-level loss $\mathcal{L}(\cdot)$. Let $f^\star \in \mathcal{F}$ be the optimal predictor that achieves $\mathcal{L}(f^\star) = 0$, and for any $f \in \mathcal{F}$ define $\mathrm{dis}(f, f^\star) := \{x \in \mathcal{X} : f(x) \neq f^\star(x)\}$ to be the disagreement region on which f and $f^\star$ disagree. By definition, $\mathcal{L}(\widehat{f}_{\mathrm{EPRM}}) = \mathbb{P}_{x \sim \mathcal{D}}[x \in \mathrm{dis}(\widehat{f}_{\mathrm{EPRM}}, f^\star)]$. We will show that $\mathcal{L}(\widehat{f}_{\mathrm{EPRM}}) \notin [\varepsilon, 1 - \varepsilon]$.

To do so, we bound the probability that $\widehat{f}_{\text{EPRM}}$ does not match the proportion on a freshly sampled bag. We already have an upper bound on this from Eq. (3). Now we compute a lower bound.

$$\mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\frac{1}{k}\sum_{j=1}^{k}\widehat{f}_{\text{EPRM}}(x_j) \neq \alpha\right] \geq \mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\sum_{j=1}^{k}\mathbb{1}\left\{x_j \in \text{dis}(\widehat{f}_{\text{EPRM}}, f^\star)\right\} \text{ is odd}\right]$$

$$= \frac{1}{2} - \frac{1}{2}\left(1 - 2\mathcal{L}(\widehat{f}_{\text{EPRM}})\right)^k. \tag{4}$$

The first inequality follows because if the bag contains an odd number of points in the disagreement set, then it is impossible for $\widehat{f}_{\text{EPRM}}$ and $f^\star$ to have the same proportional label.

For sake of contradiction, suppose that $\mathcal{L}(\widehat{f}_{\text{EPRM}}) \in [\varepsilon, 1-\varepsilon]$. Then we know $\frac{1}{2} - \frac{1}{2}(1 - 2\mathcal{L}(\widehat{f}_{\text{EPRM}}))^k \geq \frac{1}{2} - \frac{1}{2}(1 - 2\varepsilon)^k$. However, if $\varepsilon \in (0, 1/2)$, we arrive at a contradiction, since $\varepsilon \in (0, 1/2)$ implies that for any bag size $k \geq 2$, we have $1 - 2\varepsilon > (1 - 2\varepsilon)^k$, which implies that $\varepsilon < \frac{1}{2} - \frac{1}{2}(1 - 2\varepsilon)^k$, so Eqs. (3) and (4) cannot simultaneously hold. Therefore we must have $\mathcal{L}(\widehat{f}_{\text{EPRM}}) \notin [\varepsilon, 1-\varepsilon]$.

Step 3. Now we establish that $\mathcal{L}(\widehat{f}_{\text{EPRM}}) \notin [1-\varepsilon, 1]$. We claim the following: for any near optimal predictor $\widetilde{f} \in \mathcal{F}_\varepsilon := \{f \in \mathcal{F} : \mathcal{L}(f) \leq \varepsilon\}$, we must have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{\frac{1}{k}\sum_{j=1}^{k}\widetilde{f}(x_{i,j}) \neq \alpha_i\right\} < \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{\frac{1}{k}\sum_{j=1}^{k}1 - \widetilde{f}(x_{i,j}) \neq \alpha_i\right\}. \tag{5}$$

Call this event $\mathcal{E}(\widetilde{f})$. From here, the result that $\mathcal{L}(\widehat{f}_{\text{EPRM}}) \notin [1-\varepsilon, 1]$ follows because if $\mathcal{L}(\widehat{f}_{\text{EPRM}}) \in [1-\varepsilon, 1]$, then it could not have been the EPRM, as the predictor $1 - \widehat{f}_{\text{EPRM}}$ has strictly better empirical proportional risk and also lies in the class $\mathcal{F}$ by the symmetric assumption.

We now prove Eq. (5). Consider any predictor $\widetilde{f} \in \mathcal{F}_\varepsilon$. Define the indicator variable $Z_i \in \{0, 1\}$ as $Z_i = \mathbb{1}\{\frac{1}{k}\sum_{j=1}^{k}\widetilde{f}(x_{i,j}) = \alpha_i\} \cdot \mathbb{1}\{\frac{1}{k}\sum_{j=1}^{k}1 - \widetilde{f}(x_{i,j}) \neq \alpha_i\}$. We see that $\{\frac{1}{n}\sum_i Z_i > 1/2\} \subseteq \mathcal{E}(\widetilde{f})$. We bound the expectation of Z_i as:

$$\mathbb{E}[Z_i] \geq \mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\forall x_j \in B : x_j \notin \text{dis}(\widetilde{f}, f^\star) \text{ and } \alpha \neq \frac{1}{2}\right].$$

To lower bound this, we can bound the two events separately. First we have

$$\mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\forall x_j \in B : x_j \notin \text{dis}(\widetilde{f}, f^\star)\right] \geq (1 - \varepsilon)^k \geq \frac{3}{4},$$

where the last inequality is true whenever $\varepsilon \leq 1/(4k)$. In addition,

$$\mathbb{P}_{(B,\alpha)\sim\mathcal{D}}[\alpha \neq 1/2] = 1 - \mathbb{P}_{(B,\alpha)\sim\mathcal{D}}[\alpha = 1/2] \geq 1 - 2^{-k}\binom{k}{k/2} \geq 1 - \frac{1}{\sqrt{3k/2 + 1/2}},$$

by Stirling's approximation. Using the law of total probability we get

$$1 \geq \mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\forall x_j \in B : x_j \notin \text{dis}(\widehat{f}, f^\star) \text{ or } \alpha \neq \frac{1}{2}\right]$$

$$= \mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\forall x_j \in B : x_j \notin \mathrm{dis}(\widehat{f}, f^\star)\right] + \mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\alpha \neq \frac{1}{2}\right]$$

$$- \mathbb{P}_B\left[\forall x_j \in B : x_j \notin \mathrm{dis}(\widehat{f}, f^\star) \text{ and } \alpha \neq \frac{1}{2}\right]$$

$$\geq \frac{3}{4} + 1 - \frac{1}{\sqrt{3k/2 + 1/2}} - \mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\forall x_j \in B : x_j \notin \mathrm{dis}(\widehat{f}, f^\star) \text{ and } \alpha \neq \frac{1}{2}\right],$$

so therefore

$$\mathbb{P}_{(B,\alpha)\sim\mathcal{D}}\left[\forall x_j \in B : x_j \notin \mathrm{dis}(\widehat{f}, f^\star) \text{ and } \alpha \neq \frac{1}{2}\right] \geq \frac{3}{4} - \frac{1}{\sqrt{3k/2 + 1/2}}.$$

Whenever $k \geq 11$ the RHS is at least 0.507. As a consequence by Hoeffding's inequality, we have $\mathbb{P}[\mathcal{E}(\widetilde{f})^c] \leq \mathbb{P}[\frac{1}{n}\sum_i Z_i \leq 1/2] \leq \exp(-2n \cdot 0.07^2)$. By union bound, we have $\mathbb{P}[\exists f \in \mathcal{F}_\varepsilon : \mathcal{E}(\widetilde{f})^c] \leq \Gamma_\mathcal{F}(nk) \cdot \exp(-2n \cdot 0.07^2)$, where $\Gamma_\mathcal{F} : \mathbb{N} \to \mathbb{N}$ is the growth function for $\mathcal{F}$. Setting the RHS to δ and using Sauer's lemma we get that as long as $n = O(d\log k + \log(1/\delta))$, the event $\mathcal{E}(\widetilde{f})$ holds for all $\widetilde{f} \in \mathcal{F}_\varepsilon$.

Putting it together. Therefore, with probability at least $1 - 2\delta$, $\mathcal{L}(\widehat{f}_{\mathrm{EPRM}}) \leq \varepsilon$ as long as $\varepsilon \leq 1/(4k)$ and

$$n = O\left(\frac{d\log k \cdot \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right).$$

After rescaling δ, this concludes the proof of Theorem 1. ∎

2.2. Proportion Matching Fails in Agnostic Setting

In the agnostic setting, we illustrate how proportion matching can perform quite poorly.

Example: EPRM may require $\Omega(2^k)$ sample complexity. Fix any $\varepsilon \in (0, 1/2)$ and consider the input space $\mathcal{X} = \{x\}$, with $\mathcal{D}$ given by $(x, 1)$ with probability $1/2 + \varepsilon$ and $(x, 0)$ with probability $1/2 - \varepsilon$. Let $\mathcal{F}$ consist of two functions $f_0(x) = 0$ and $f_1(x) = 1$. The optimal predictor within the class $\mathcal{F}$ is f_1. However, observe that the loss estimates for proportion matching are $\widehat{L}(f_0) := 1 - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\alpha_i = 0\}$ and $\widehat{L}(f_1) := 1 - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\alpha_i = 1\}$. For small ε, unless the number of bags is exponential in k, with constant probability we do not see any "pure" bags (with $\alpha_i = 0$ or $\alpha_i = 1$), so we have no way of distinguishing which of f_0 and f_1 achieves smaller loss.

Proportion matching can fail. One may object that the previous failure mode is due to the fact that we are using a noncontinuous measure of discrepancy between the predicted and true label proportion, and such issues can be resolved if we minimize a continuous measure of discrepancy. We show that this does not help, as proportion matching approaches can return a predictor with constant suboptimality. This is because in the agnostic setting, the predictor which matches the bag-level proportions may not be the optimal instance-level predictor.

Consider the learning rule that minimizes the proportional square loss:

$$\widehat{f}_{\mathrm{SQ}} := \underset{f\in\mathcal{F}}{\mathrm{argmin}}\ \widehat{L}_{\mathrm{SQ}}(f) = \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{k}\sum_{j=1}^k f(x_{i,j}) - \alpha_i\right)^2. \tag{6}$$

We show that in general, minimizing the proportional square loss can fail in the agnostic setting.

Proposition 2 *There exists a $\mathcal{F}$ with $\mathrm{VC}(\mathcal{F}) = 1$ and distribution $\mathcal{D}$ such that for any $\delta \in (0, 1)$, bag size $k \geq 7$, and sample size $n = \Omega(\log(1/\delta))$, with probability at least $1 - \delta$, the learning rule $\widehat{f}_{\mathrm{SQ}}$ is $1/3$-suboptimal.*

The proportional square loss learning rule, as well as the proportional log loss learning rule

$$\widehat{f}_{\mathrm{LOG}} := \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} -\alpha_i \cdot \log\left(\frac{1}{k} \sum_{j=1}^{k} f(x_{i,j})\right) - (1 - \alpha_i) \cdot \log\left(1 - \frac{1}{k} \sum_{j=1}^{k} f(x_{i,j})\right) \quad (7)$$

are regarded as folklore learning rules, and they were evaluated in the context of gradient-based learning (Busa-Fekete et al., 2023). Busa-Fekete et al. show that while gradient-based minimization of either the proportional square or log loss performs well in practice, it can fail in synthetic experimental settings. Proposition 2 demonstrates a simple theoretical failure mode for $\widehat{f}_{\mathrm{SQ}}$. For completeness, in Appendix A we provide a similar result for the failure of $\widehat{f}_{\mathrm{LOG}}$ on the same construction, but note that even in the standard classification setting ($k = 1$) it is well known that minimizing surrogate losses like the log loss do not necessarily give guarantees on the classification error in the agnostic setting (Ben-David et al., 2012). Lastly, we remark that Appendix A of Scott and Zhang (2020) shows an example of similar flavor that in the limit as the bag size $k \to \infty$, proportion matching learning rules can suffer constant suboptimality.

Proof of Proposition 2. Let $\mathcal{X} = \{x^{(1)}, x^{(2)}\}$. Let $\mathcal{F} = \{f_1, f_2\}$ where $f_1(x) = \mathbb{1}\{x = x^{(1)}\}$ and $f_2(x) = \mathbb{1}\{x = x^{(2)}\}$. The distribution $\mathcal{D}$ is $(x, y) \sim \mathrm{Unif}(\{(x^{(1)}, 1), (x^{(1)}, 0), (x^{(2)}, 1)\})$. We can calculate that $\mathcal{L}(f_1) = 2/3$ and $\mathcal{L}(f_2) = 1/3$. However we also have $\mathbb{E} f_1 = 2/3$ while $\mathbb{E} f_2 = 1/3$, and $p = 2/3$. While f_1 in expectation matches the marginal label proportion, it is actually $1/3$-suboptimal compared to f_2.

We compute the expectations of the bag-level losses for f_1 and f_2. For f_1 we have

$$\mathbb{E}\left[\left(\frac{1}{k} \sum_j f_1(x_j) - \alpha\right)^2\right] = \frac{1}{k^2} \mathbb{E}\left[\left(\sum_j f_1(x_j) - y_i\right)^2\right] = \frac{1}{k^2} \mathbb{E}\left[\sum_j (f_1(x_j) - y_j)^2\right] = \frac{2}{3k}.$$

For f_2 we have

$$\mathbb{E}\left[\left(\frac{1}{k} \sum_j f_2(x_j) - \alpha\right)^2\right] = \frac{1}{k^2} \mathbb{E}\left[\sum_j (f_2(x_j) - y_j)^2\right] + \frac{1}{k^2} \mathbb{E}\left[\sum_{j \neq j'} (f_2(x_j) - y_j)(f_2(x_{j'}) - y_{j'})\right]$$

$$= \frac{1}{3k} + \frac{k-1}{9k} = \frac{k+2}{9k}.$$

Fix any $k \geq 7$. Then the expectation of the bag-level loss of f_1 is at most $2/21$ while the expectation of f_2 is at least $1/9$. By Hoeffding's inequality, we have with probability at least $1 - \delta$, that $\widehat{\mathcal{L}}(f_1) \leq 2/21 + \sqrt{2\log(1/\delta)/n}$ and $\widehat{\mathcal{L}}(f_2) \geq 1/9 - \sqrt{2\log(1/\delta)/n}$. So as long as $n = \Omega(\log(1/\delta))$, the proportional square loss learning rule will return the wrong predictor. ∎

3. Debiased Square Loss

In this section, we show that a simple debiasing of the square loss $\widehat{L}_{\mathrm{SQ}}(\cdot)$ results in a learning rule that achieves optimal rates in both the realizable and the agnostic settings. Computing the expectation of the square loss, for any predictor f,

$$
\begin{aligned}
\mathbb{E}\left[\widehat{L}_{\mathrm{SQ}}(f)\right] &= \frac{1}{k^2} \cdot \mathbb{E}_{(B,\alpha)\sim\mathcal{D}}\left[\sum_j (f(x_j) - y_j)^2 + \sum_{j \neq j'} (f(x_j) - y_j)(f(x_{j'}) - y_{j'})\right] \\
&\overset{(i)}{=} \frac{1}{k} \cdot \mathcal{L}(f) + \frac{k-1}{k} \cdot \mathbb{E}_{(x,y),(x',y')\sim\mathcal{D}}\left[(f(x) - y)(f(x') - y')\right] \\
&\overset{(ii)}{=} \frac{1}{k} \cdot \mathcal{L}(f) + \frac{k-1}{k} \cdot (\mathbb{E}\,f - p)^2.
\end{aligned}
\tag{8}
$$

Equality (i) uses the fact that for any j, $(f(x_j) - y_j)^2 = \mathbb{1}\{f(x_j) \neq y_j\}$. Equality (ii) uses independence of the samples.

Rearranging Eq. (8), we can see that the quantity $k \cdot \widehat{L}_{\mathrm{SQ}}(f) - (k-1) \cdot (\mathbb{E}\,f - p)^2$ is an unbiased estimate of $\mathcal{L}(f)$. Of course, the caveat is that we do not have access to the values $\mathbb{E}\,f$ and p, but we can replace them with their empirical counterparts, giving us the debiased square loss learning rule:

$$
\widehat{f}_{\mathrm{DSQ}} := \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{L}_{\mathrm{DSQ}}(f) = \frac{1}{n}\sum_{i=1}^{n} k \cdot \left(\frac{1}{k}\sum_{j=1}^{k} f(x_{i,j}) - \alpha_i\right)^2 - (k-1)\left(\widehat{\mathbb{E}}f - \widehat{p}\right)^2,
$$

where $\widehat{\mathbb{E}}f = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} f(x_{i,j})$ and $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i$ are empirical estimates of $\mathbb{E}\,f$ and p respectively. Strictly speaking, $\widehat{L}_{\mathrm{DSQ}}(f)$ is not an unbiased estimate of $\mathcal{L}(f)$, but if n is large enough, the second term approximates $(\mathbb{E}\,f - p)^2$ closely.

A similar debiasing idea for the proportional square loss was proposed by Chen et al. (2023), where in their Theorem 2 they prove the $1/\sqrt{n}$ rate in the agnostic setting; due to difference in the bag generation assumption (they consider generating bags by resampling a dataset without replacement), our debiasing term takes a slightly different form.

3.1. Main Result: Optimistic Rates for Debiased Square Loss

Theorem 3 (Sample Complexity Bound for $\widehat{f}_{\mathrm{DSQ}}$) *Let $\delta \in (0,1)$. Fix any distribution $\mathcal{D}$ and any function class $\mathcal{F}$, and let $L^\star = \inf_{f \in \mathcal{F}} \mathcal{L}(f)$. With probability at least $1 - \delta$, we have*

$$
\mathcal{L}(\widehat{f}_{\mathrm{DSQ}}) \leq L^\star + \widetilde{O}\left(\frac{k^2(d + \log(1/\delta))}{n} + \sqrt{\frac{L^\star \cdot k^2(d + \log(1/\delta))}{n}}\right).
$$

Theorem 3 shows that under realizability (with $L^\star = 0$), $\widehat{f}_{\mathrm{DSQ}}$ enjoys a fast $1/n$ rate, while in the agnostic setting it achieves the $1/\sqrt{n}$ rate, both of which are optimal (up to log factors) in terms of $d, \log(1/\delta)$, and n. The dependence on k is certainly loose, as under realizability, $\widehat{f}_{\mathrm{DSQ}}$ and $\widehat{f}_{\mathrm{EPRM}}$ are identical learning rules, and Theorem 1 only has a $\log k$ dependence. We leave sharpening the dependence on k to future work.

The proof of Theorem 3 can be found in Appendix C. At a high level, we separately show uniform convergence bounds for both the square loss and the bias correction term to their expectations, and

then we combine the guarantees to get a optimistic rate bound for $\widehat{L}_{\mathrm{DSQ}}(\cdot)$, which is an unbiased estimate of $\mathcal{L}(f)$, giving us the final guarantee.

Revisiting what happens if we minimize $\widehat{L}_{\mathrm{SQ}}$. Our analysis provides an answer to the question raised by Busa-Fekete et al. (2023) on understanding when proportion matching is consistent. Suppose the learner minimizes the square loss of Eq. (6), $\widehat{L}_{\mathrm{SQ}}(f) = \frac{1}{n}\sum_{i=1}^{n} k \cdot (\frac{1}{k}\sum_{j=1}^{k} f(x_{i,j}) - \alpha_i)^2$. (For sake of comparison, we multiply the loss by a factor of k.) One can show an optimistic-rate style guarantee for $\widehat{f}_{\mathrm{SQ}}$, albeit one that is weaker than Theorem 3. Let $L_{\mathrm{SQ}}(\cdot)$ denote the expectation of $\widehat{L}_{\mathrm{SQ}}$. In the proof of Theorem 3, we get the following guarantee on $\widehat{f}_{\mathrm{SQ}}$:

$$L_{\mathrm{SQ}}(\widehat{f}_{\mathrm{SQ}}) \leq \inf_{f \in \mathcal{F}} \left\{ L_{\mathrm{SQ}}(f) + \widetilde{O}\left(\frac{kd + k^2 \log(1/\delta)}{n} + \sqrt{\frac{L_{\mathrm{SQ}}(f) \cdot k(d + \log(1/\delta))}{n}} \right) \right\}.$$

This bound essentially replaces the instance-level classification error $\mathcal{L}(f)$ with the (larger) $L_{\mathrm{SQ}}(f)$ in Theorem 3. Applying the substitution $L_{\mathrm{SQ}}(f) = \mathcal{L}(f) + B(f)$, where $B(f) := (k-1)(\mathbb{E} f - p)^2$, we see that this bound implies

$$\mathcal{L}(\widehat{f}_{\mathrm{SQ}}) \leq \inf_{f \in \mathcal{F}} \left\{ \mathcal{L}(f) + B(f) + \widetilde{O}\left(\frac{kd + k^2 \log(1/\delta)}{n} + \sqrt{\frac{(\mathcal{L}(f) + B(f)) \cdot k(d + \log(1/\delta))}{n}} \right) \right\}. \quad (9)$$

When is Eq. (9) a useful bound? Under realizability, we have $\mathcal{L}(f^\star) = 0$ and $B(f^\star) = 0$, and since any minimizer $\widehat{f}_{\mathrm{EPRM}}$ is also a minimizer of $\widehat{L}_{\mathrm{SQ}}(\cdot)$ and vice versa, Eq. (9) recovers the guarantee of Theorem 1, albeit with worse dependence on k. More generally, in the agnostic setting, as long as it is possible to achieve a good trade-off between $\mathcal{L}(f)$ and $B(f)$, we expect $\widehat{f}_{\mathrm{SQ}}$ to generalize well. This reasoning suggests why proportion matching learning rules perform well in practice (Busa-Fekete et al., 2023), despite the negative result of Proposition 2. In the modern over-parameterized regime, where the function class at hand (i.e., neural networks) are nearly realizable for the data distribution, Eq. (9) delivers strong guarantees.

Lastly, observe that there is no contradiction between the guarantee of Eq. (9) and the lower bound in Proposition 2: in Proposition 2, there is no predictor for which the two terms $\mathcal{L}(f)$ and $B(f)$ are both small, as the suboptimal predictor has larger classification loss but satisfies $B(f) = 0$.

4. EasyLLP Learning Rule

Recently, Busa-Fekete et al. (2023) proposed EasyLLP, a general recipe for constructing unbiased estimators of any instance-level loss function $\ell_{\mathrm{ins}} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$. Given predictor $f \in \mathcal{F}$ and bag (B, α), the EasyLLP loss estimate of ℓ_{ins} can be written as

$$\ell_{\mathrm{EZ}}(f, (B, \alpha)) := \frac{1}{k}\sum_{j=1}^{k} (k(\alpha - p) + p)\ell_{\mathrm{ins}}(1, f(x_j)) + (k(p - \alpha) + (1-p))\ell_{\mathrm{ins}}(0, f(x_j)) \quad (10)$$

Proposition 4.2 of Busa-Fekete et al. (2023) shows that for any loss ℓ_{ins}, $\ell_{\mathrm{EZ}}(f, (B, \alpha))$ is an unbiased estimate of the population loss, e.g., $\mathbb{E}_{(B,\alpha)\sim\mathcal{D}}[\ell_{\mathrm{EZ}}(f, (B, \alpha))] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathrm{ins}}(y, f(x))]$. In this work, we consider EasyLLP instantiated with the classification loss $\ell_{\mathrm{ins}} = \ell^{01}$. The EasyLLP

estimate takes the following form:

$$\ell_{\text{EZ}}(f,(B,\alpha)) = (k(\alpha - p) + p)\big(1 - \frac{1}{k}\sum_{j=1}^{k} f(x_j)\big) + (k(p - \alpha) + (1 - p))\big(\frac{1}{k}\sum_{j=1}^{k} f(x_j)\big).$$
(11)

The EasyLLP learning rule $\widehat{f}_{\text{EZ}} := \operatorname{argmin}_{f \in \mathcal{F}} \widehat{L}_{\text{EZ}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{EZ}}(f,(B_i,\alpha_i))$ attains the following guarantee (Theorem 5.2 of Busa-Fekete et al., 2023): with probability at least $1 - \delta$,

$$\mathcal{L}(\widehat{f}_{\text{EZ}}) \leq \inf_{f \in \mathcal{F}} \mathcal{L}(f) + \widetilde{O}\bigg(\sqrt{\frac{d + k\log(1/\delta)}{n}}\bigg).$$
(12)

However, the question of whether EasyLLP can adapt to realizability to achieve fast rates remained open. The standard observation which enables fast rates under realizability is that loss estimates for (nearly) optimal predictors converge to their expectations at a rate of $O(1/n)$. As a consequence, in supervised learning, one can show that any predictor $\widehat{f}$ which achieves training error 0 (i.e., any ERM) has generalization error at most $\mathcal{L}(\widehat{f}) \leq \widetilde{O}(d/n)$.

In contrast, the EasyLLP loss estimate does not satisfy this property. Even for the optimal predictor $f^\star$, the EasyLLP loss estimate of Eq. (11) is a *random* quantity, since

$$\ell_{\text{EZ}}(f^\star,(B,\alpha)) = (k(\alpha - p) + p) \cdot (1 - \alpha) + (k(p - \alpha) + (1 - p)) \cdot \alpha$$

takes values depending on the bag label proportion α which itself is distributed as $1/k \cdot \text{Bin}(k, p)$. As the following proposition shows, the loss estimate only concentrates at a $\Theta(1/\sqrt{n})$ rate.

Proposition 4 *There exists a realizable distribution $\mathcal{D}$ such that for any $\varepsilon \in (0,1)$, the EasyLLP loss estimate of $f^\star$ with bag size $k = 2$ requires $\Theta(1/\varepsilon^2)$ samples in order for $\widehat{L}_{\text{EZ}}(f^\star) \leq \varepsilon$ with constant probability.*

Proof. Consider the setting $\mathcal{X} = \{x_0, x_1\}$ with $\mathcal{D}$ that returns $(x_0, 0)$ with probability $1/2$ and $(x_1, 1)$ with probability $1/2$. Let $f^\star(x) = \mathbb{1}\{x = x_1\}$ be the optimal predictor achieving $\mathcal{L}(f^\star) = 0$. For any bag (B, α) the EasyLLP estimate of the loss can be written as

$$\ell_{\text{EZ}}(f^\star,(B,\alpha)) = (k(\alpha - 1/2) + 1/2) \cdot (1 - \alpha) + (k(1/2 - \alpha) + 1/2) \cdot \alpha$$
$$= -2k\alpha^2 + 2\alpha k - k/2 + 1/2.$$

For bag size $k = 2$, we have $\ell_{\text{EZ}}(f^\star,(B,\alpha)) \sim 1/2 \cdot \text{Rad}(1/2)$. From here, we can apply standard anti-concentration bounds for sums of Rademacher random variables. We let $x_j = \ell_{\text{EZ}}(f^\star,(B_i,\alpha_i))$, and we can observe that by Paley-Zygmund (Theorem 20) that

$$\mathbb{P}\bigg[\frac{1}{n}\sum_{i=1}^{n} x_j \geq \frac{0.1}{\sqrt{n}}\bigg] = \frac{1}{2}\mathbb{P}\bigg[\bigg|\frac{1}{n}\sum_{i=1}^{n} x_j\bigg| \geq \frac{0.1}{\sqrt{n}}\bigg] \geq \Omega(1).$$

Thus, if $n = c/\varepsilon^2$ for sufficiently small $c > 0$ then with constant probability we have $\widehat{L}_{\text{EZ}}(f^\star) \geq \varepsilon$. On the flip side, we know that by Hoeffding's inequality, $n = O(1/\varepsilon^2)$ suffices for $\widehat{L}_{\text{EZ}}(f^\star) \leq \varepsilon$ with constant probability. This proves the proposition. ∎

4.1. Main Result: Optimistic Rates for EasyLLP

Despite the fact that the EasyLLP loss estimates only concentrate to their expectations at a rate of $O(1/\sqrt{n})$, it is possible improve upon the guarantee in Eq. (12) and show that the EasyLLP learning rule instantiated with the classification loss attains optimistic rates.

Theorem 5 (Sample Complexity Bound for $\widehat{f}_{\mathrm{EZ}}$) *Let $\delta \in (0,1)$. Fix any distribution $\mathcal{D}$ and function class $\mathcal{F}$, and let $L^\star := \inf_{f \in \mathcal{F}} \mathcal{L}(f)$. With probability at least $1 - \delta$ we have*

$$\mathcal{L}(\widehat{f}_{\mathrm{EZ}}) \leq L^\star + \widetilde{O}\left(\frac{k^2(d + \log(1/\delta))}{n} + \sqrt{\frac{L^\star \cdot k^2(d + \log(1/\delta))}{n}} \right).$$

The bound for EasyLLP is order-wise identical to that shown for the debiased square loss (Theorem 3). In the agnostic setting, the guarantee in Theorem 5 is worse in terms of dependence on k compared to Eq. (12); we leave sharpening the dependence on k to future work. In the realizable setting, Proposition 4 and Theorem 5 together show that there is a separation between the rate of estimation (which is necessarily $\Omega(1/\sqrt{n})$) and the rate of learning (which is $\widetilde{O}(1/n)$).

We sketch the proof ideas for Theorem 5, and we defer the full proof to Appendix D.

Proof Sketch. There is no hope for us to prove Theorem 5 through the usual route of showing uniform convergence bound like the following:

$$\text{for all } f \in \mathcal{F}, |\mathcal{L}(f) - \widehat{L}_{\mathrm{EZ}}(f)| \leq \widetilde{O}\left(\frac{k^2(d + \log(1/\delta))}{n} + \sqrt{\frac{\mathcal{L}(f) \cdot k^2(d + \log(1/\delta))}{n}} \right),$$

since the previous display directly contradicts Proposition 4 for $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}(f)$.

Instead, we make the critical observation that the *offset* empirical losses concentrate at the optimistic rate. Even though the learner does not know the identity of $f^\star$, it is still true that minimizing $\widehat{L}_{\mathrm{EZ}}(\cdot)$ is the same as minimizing the offset loss $\widehat{\Gamma}(\cdot, f^\star) := \widehat{L}_{\mathrm{EZ}}(\cdot) - \widehat{L}_{\mathrm{EZ}}(f^\star)$, so we can equivalently think of the learning rule as minimizing $\widehat{\Gamma}(\cdot, f^\star)$. The following lemma shows that one can prove an optimistic rate for the offset empirical losses. In the lemma, we use $\Gamma(f, f^\star) := \mathcal{L}(f) - \mathcal{L}(f^\star)$ to denote the expected difference in classification error.

Lemma 6 *Let $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}(f)$. Then with probability at least $1 - \delta$ we have for all $f \in \mathcal{F}$*

$$\left| \widehat{\Gamma}(f, f^\star) - \Gamma(f, f^\star) \right| \leq \widetilde{O}\left(\frac{kd + k^2 \log(1/\delta)}{n} + \sqrt{\frac{\mathcal{L}(f) \cdot k^2(d + \log(1/\delta))}{n}} \right).$$

Lemma 6 says that while the EasyLLP empirical estimate $\widehat{L}_{\mathrm{EZ}}$ is only $O(1/\sqrt{n})$ close to the true classification error, the estimate of the *difference* in classification error with that of $f^\star$ is more accurate, as much of the fluctuations in the bag estimates $\widehat{L}_{\mathrm{EZ}}(\cdot)$ gets canceled out by subtraction.

In light of Lemma 6, Theorem 5 follows from standard approach of translating uniform convergence bounds to guarantees on the returned predictor $\widehat{f}_{\mathrm{EZ}}$, see Appendix D.3 for more details.

Estimating the Marginal Label Proportion. The EasyLLP learning rule requires knowledge of the marginal label proportion $p = \mathbb{P}[y = 1]$. We demonstrate that the optimistic rate guarantee in Theorem 5 can be attained without knowledge of p, instead splitting the dataset into half to estimate $\widehat{p}$ from a separate dataset (this was claimed without proof by Busa-Fekete et al. (2023)).

Corollary 7 (Sample Complexity Bound for $\widehat{f}_{\mathrm{EZ}}$ with Sample Splitting) *Let $\delta \in (0, 1)$. Fix any distribution $\mathcal{D}$ and function class $\mathcal{F}$, and let $L^\star := \inf_{f \in \mathcal{F}} \mathcal{L}(f)$. Then with probability at least $1 - \delta$ the EasyLLP learning rule with sample splitting satisfies*

$$\mathcal{L}(\widehat{f}_{\mathrm{EZ}}) \leq L^\star + \widetilde{O}\left(\frac{k^2(d + \log(1/\delta))}{n} + \sqrt{\frac{L^\star \cdot k^2(d + \log(1/\delta))}{n}} \right).$$

The details of the sample splitting procedure and the proof of Corollary 7 are shown in Appendix D.4. In contrast to the debiased square loss learning rule, where we could estimate square loss and bias terms from the same dataset, here our analysis requires a separate dataset in order to estimate $\widehat{p}$. However, we conjecture that sample splitting is not required for the guarantee in Corollary 7.

5. Lower Bounds

The sample complexity bounds we prove in Theorem 1, 3, and 5 are optimal (up to log factors) in terms of the dependence on d, n, and $\log(1/\delta)$. However, the question remains of resolving the optimal dependence on the bag size k.

For every function class $\mathcal{F}$, we have the trivial lower bound on the minimax sample complexity of $n = \Omega(d \log(1/\delta)/(k\varepsilon))$ for the realizable setting and $n = \Omega\big(d \log(1/\delta)/(k\varepsilon^2)\big)$ for the agnostic setting, since LLP with n bags is only harder than supervised learning with nk examples. In general, the $1/k$ dependence in the lower bound cannot be improved, as there are function classes for which it is tight. For example, consider the function class $\mathcal{F} = \{f_0, f_1\}$ where $f_i(x) = i$ for $i \in \{0, 1\}$. For this class $\mathcal{F}$, observing the label proportion α allows us to compute the average instance-level classification loss over the bag for both f_0 and f_1. Therefore, LLP with n bags is no harder than supervised learning with nk labeled examples, so $\mathcal{F}$ can be PAC learned with $\widetilde{O}(1/(k\varepsilon))$ samples in the realizable setting and $\widetilde{O}(1/(k\varepsilon^2))$ in the agnostic setting.

For *specific* function classes $\mathcal{F}$, it is possible to improve the $1/k$ lower bound.

Theorem 8 *For any $d \geq 3$, there exists a function class $\mathcal{F}$ with $\mathrm{VC}(\mathcal{F}) = d$ such that any learning rule for LLP that PAC learns $\mathcal{F}$ for bag size $k \leq O(2^d/\log d)$ with $\varepsilon \leq 1/16$, and $\delta \leq 1/15$ requires $\Omega\left(\frac{d}{\log k}\right)$ samples in the realizable setting and $\Omega\left(\max\left(\frac{d}{\log k}, \frac{d}{\sqrt{k}\varepsilon^2}\right)\right)$ samples in the agnostic setting.*

The proof of Theorem 8 can be found in Appendix E. In the realizable setting, Theorem 8 gives a stronger lower bound when the accuracy parameter ε is a constant; however it is open to show a lower bound which dominates the trivial one of $\Omega(d/(k\varepsilon))$ for $\varepsilon \to 0$. In the agnostic setting, the lower bound in Theorem 8 dominates the trivial one for all $\varepsilon \leq 1/16$.

6. Discussion

Our work studies various learning rules for minimizing classification loss in the LLP framework. We show that EPRM attains fast rates under realizability, but EPRM and other proportion matching approaches can fail in the agnostic setting. For the debiased square loss and EasyLLP learning rules, we prove optimistic rate sample complexity bounds which are optimal (up to log factors) in terms of the dependence on d, n, and $\log(1/\delta)$ in both the realizable and agnostic settings. In addition, we investigate the optimal dependence on k from the lower bound side. We also compare the empirical performance of gradient-based versions of these learning rules and demonstrate the benefits of debiasing for optimization (deferred to Appendix G due to space considerations).

For clarity of exposition, we focus on binary classification, but we note that both the debiased square loss and EasyLLP learning rules can be extended to the multi-label multi-class setting (as also observed in Chen et al. (2023); Busa-Fekete et al. (2023)) by using one-hot encoding to write the label as a binary vector. In this way the learning task decouples into multiple binary classification tasks, and it would be straightforward to extend our analysis to this more general setting.

Our work leaves open several future directions. The most immediate one is to resolve the optimal dependence on k. On the upper bound side, we believe that the dependence on k can be improved in our analysis, and leave this to future work. Our optimistic rate results (Theorem 3 and 5) are stated for a more general setting, and their proofs do not use the combinatorial structure of the LLP problem in the way that the proof of Theorem 1 does, thus hinting at a source of looseness. On the lower bound side, there is room to improve upon the construction in Theorem 8. Other directions for future work include understanding the debiased square loss and EasyLLP learning rules in a more unified manner, designing a debiased variant of log loss for LLP, and studying the role of optimization in gradient-based algorithms for LLP.

Acknowledgments

References

Apple. Apple storekit ad network. https://developer.apple.com/documentation/storekit/skadnetwork/, 2024.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. 2005.

Shai Ben-David, David Loker, Nathan Srebro, and Karthik Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. *arXiv preprint arXiv:1206.6442*, 2012.

Olivier Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, École Polytechnique: Department of Applied Mathematics Paris, France, 2002.

Robert Istvan Busa-Fekete, Heejin Choi, Travis Dick, Claudio Gentile, et al. Easy learning from label proportions. *arXiv preprint arXiv:2302.03115*, 2023.

Bee-Chung Chen, Lei Chen, Raghu Ramakrishnan, and David R Musicant. Learning from aggregate views. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 3–3. IEEE, 2006.

Lin Chen, Thomas Fu, Amin Karbasi, and Vahab Mirrokni. Learning from aggregated data: Curated bags versus random bags. *arXiv preprint arXiv:2305.09557*, 2023.

Lucio Mwinmaarong Dery, Benjamin Nachman, Francesco Rubbo, and Ariel Schwartzman. Weakly supervised classification in high energy physics. *Journal of High Energy Physics*, 2017(5):1–11, 2017.

Eustache Diemert, Romain Fabre, Alexandre Gilotte, Fei Jia, Basile Leparmentier, Jérémie Mary, Zhonghua Qu, Ugo Tanielian, and Hui Yang. Lessons from the AdKDD'21 privacy-preserving ML challenge. In *Proceedings of the ACM Web Conference 2022*, pages 2026–2035, 2022.

Yongke Ding, Yuanxiang Li, and Wenxian Yu. Learning from label proportions for sar image classification. *Eurasip Journal on Advances in Signal Processing*, 2017:1–12, 2017.

Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. Deep multi-class learning from label proportions. *arXiv preprint arXiv:1905.12909*, 2019.

Google. Private aggregation api of chrome privacy sandbox. https://developer.chrome.com/docs/privacy-sandbox/aggregation-service/, 2024.

Adel Javanmard, Lin Chen, Vahab Mirrokni, Ashwinkumar Badanidiyuru, and Gang Fu. Learning from aggregate responses: Instance level versus bag level loss functions. In *Twelfth International Conference on Learning Representations (arXiv preprint arXiv:2401.11081)*, 2024a.

Adel Javanmard, Matthew Fahrbach, and Vahab Mirrokni. Priorboost: An adaptive algorithm for learning from aggregate responses. In *The Forty-first International Conference on Machine Learning (arXiv preprint arXiv:2402.04987)*, 2024b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

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

Shahar Mendelson. A few notes on statistical learning theory. In *Advanced Lectures on Machine Learning: Machine Learning Summer School 2002 Canberra, Australia, February 11–22, 2002 Revised Lectures*, pages 1–40. Springer, 2003.

David R Musicant, Janara M Christensen, and Jamie F Olson. Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 252–261. IEEE, 2007.

Dmitriy Panchenko. Some extensions of an inequality of vapnik and chervonenkis. 2002.

Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (almost) no label no cry. *Advances in Neural Information Processing Systems*, 27, 2014.

Zhiquan Qi, Bo Wang, Fan Meng, and Lingfeng Niu. Learning with label proportions via npsvm. *IEEE transactions on cybernetics*, 47(10):3293–3305, 2016.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. In *Proceedings of the 25th International Conference on Machine learning*, pages 776–783, 2008.

Henry Reeve and Ata Kaban. Optimistic bounds for multi-output learning. In *International Conference on Machine Learning*, pages 8030–8040. PMLR, 2020.

Stefan Rueping. Svm classifier estimation from group probabilities. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 911–918, 2010.

Rishi Saket. Learnability of linear thresholds from label proportions. *Advances in Neural Information Processing Systems*, 34:6555–6566, 2021.

Rishi Saket. Algorithms and hardness for learning linear thresholds from label proportions. *Advances in Neural Information Processing Systems*, 35:1267–1279, 2022.

Clayton Scott and Jianxin Zhang. Learning from label proportions: A mutual contamination framework. *Advances in neural information processing systems*, 33:22256–22267, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Yong Shi, Limeng Cui, Zhensong Chen, and Zhiquan Qi. Learning from label proportions with pinball loss. *International Journal of Machine Learning and Cybernetics*, 10(1):187–205, 2019.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

Tao Sun, Dan Sheldon, and Brendan 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. IEEE, 2017.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

Austin Watkins, Enayat Ullah, Thanh Nguyen-Tang, and Raman Arora. Optimistic rates for multi-task representation learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Niloofar Yousefi, Yunwen Lei, Marius Kloft, Mansooreh Mollaghasemi, and Georgios C Anagnostopoulos. Local rademacher complexity-based learning guarantees for multi-task learning. *The Journal of Machine Learning Research*, 19(1):1385–1431, 2018.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

Felix Yu, Dong Liu, Sanjiv Kumar, Jebara Tony, and Shih-Fu Chang. ∝SVM for learning with label proportions. In *International conference on machine learning*, pages 504–512. PMLR, 2013.

Felix X Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv preprint arXiv:1402.5902*, 2014.

Jianxin Zhang, Yutong Wang, and Clay Scott. Learning from label proportions by learning with label noise. *Advances in Neural Information Processing Systems*, 35:26933–26942, 2022.

Lijia Zhou, Frederic Koehler, Danica J Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *arXiv preprint arXiv:2112.04470*, 2021.

Lijia Zhou, Frederic Koehler, Pragya Sur, Danica J Sutherland, and Nati Srebro. A non-asymptotic moreau envelope theory for high-dimensional generalized linear models. *Advances in Neural Information Processing Systems*, 35:21286–21299, 2022.

Lijia Zhou, Zhen Dai, Frederic Koehler, and Nathan Srebro. Uniform convergence with square-root lipschitz loss. *arXiv preprint arXiv:2306.13188*, 2023.

Appendix A. Failure of Proportional Log Loss in Agnostic Setting

Proposition 9 *There exists a $\mathcal{F}$ with $\mathrm{VC}(\mathcal{F}) = 1$ and distribution $\mathcal{D}$ such that as long as n, k, and δ satisfy the relationship $k \geq 18 \log(2n/\delta)$, with probability at least $1 - \delta$, the learning rule $\widehat{f}_{\mathrm{LOG}}$ is $1/3$-suboptimal.*

Therefore, unless n is exponentially large in the bag size k, the proportional log loss learning rule will return a suboptimal predictor.

Proof. We will use the same construction as in the proof of Proposition 2. For every bag (B_i, α_i) we will calculate the difference in proportional log loss for f_1 and f_2. For any $i \in [n]$, let $\beta_i := \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}\{x_{i,j} = x^{(1)}\}$. Then we have

$$\ell_{\mathrm{LOG}}(f_1, (B_i, \alpha_i)) = -\alpha_i \log \beta_i - (1 - \alpha_i) \log(1 - \beta_i)$$
$$\ell_{\mathrm{LOG}}(f_2, (B_i, \alpha_i)) = -\alpha_i \log(1 - \beta_i) - (1 - \alpha_i) \log \beta_i.$$

Therefore the difference between the two losses is

$$\ell_{\mathrm{LOG}}(f_2, (B_i, \alpha_i)) - \ell_{\mathrm{LOG}}(f_1, (B_i, \alpha_i)) = (2\alpha_i - 1) \log \frac{\beta_i}{1 - \beta_i}.$$

We now show that with high probability, for all $i \in [n]$ we have $\ell_{\mathrm{LOG}}(f_2, (B_i, \alpha_i)) - \ell_{\mathrm{LOG}}(f_1, (B_i, \alpha_i)) \geq 0$. By Hoeffding's inequality (Theorem 19) and union bound we have

$$\mathbb{P}[\exists i \in [n] : \alpha_i < 1/2 \text{ or } \beta_i < 1/2] \leq 2n \exp\left(-\frac{k}{18}\right).$$

Thus as long as $k \geq 18 \log(2n/\delta)$, with probability at least $1 - \delta$ we have $\ell_{\mathrm{LOG}}(f_2, (B_i, \alpha_i)) - \ell_{\mathrm{LOG}}(f_1, (B_i, \alpha_i)) \geq 0$ for all $i \in [n]$. This implies that $\widehat{f}_{\mathrm{LOG}} = f_1$, the predictor which is $1/3$-suboptimal. ∎

Appendix B. Technical Background for Optimistic Rates

B.1. Uniform Convergence via Local Rademacher Complexity

In this section, we establish technical several results on uniform convergence using local Rademacher complexities. Recall the worst-case Rademacher for a function class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}}$ for any $n \in \mathbb{N}$

$$\mathfrak{R}_n(\mathcal{G}) = \sup_{x_1, \cdots, x_n \in \mathcal{X}^n} \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_j) \right],$$

where $\{\sigma_i\}_{i=1}^{n}$ are i.i.d. Rademacher random variables. For any g, we denote $\mathbb{E}[g]$ to denote its expectation and $\widehat{\mathbb{E}}_n[g]$ to denote the empirical average over a sample $\{x_i\}_{i=1}^{n}$ where x_i are i.i.d. drawn.

Lemma 10 (Modified Version of Lemma 6.2 in Bousquet (2002)) *Let $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}}$ be a class of functions such that $\|g\|_\infty \leq b$ for all $g \in \mathcal{G}$ and let $(\mathcal{G}_k)_{n \in \mathbb{N}}$ be a sequence of subsets of $\mathcal{G}$ such that $\sup_{g \in \mathcal{G}_k} \mathbb{E}[g^2] \leq A + B\gamma_k$, where $\gamma_k = b/2^k$ and $A, B > 0$ are constants.*

Then for all $\delta \in (0, 1)$ with probability at least $1 - \delta$, for all $k \geq 0$ and $g \in \mathcal{G}_k$:

$$|\mathbb{E}[g] - \widehat{\mathbb{E}}_n[g]| \leq 6\mathfrak{R}_n(\mathcal{G}_k) + \sqrt{\frac{2(A + B\gamma_k)\left(\log \frac{1}{\delta} + c \log \log \frac{b}{\gamma_k}\right)}{n}} + \frac{6b\left(\log \frac{1}{\delta} + c \log \log \frac{b}{\gamma_k}\right)}{n}.$$

Here, $c > 0$ is an absolute constant.

Lemma 10 is a trivial modification of Lemma 6.2 in Bousquet (2002), so we omit the proof details.

Definition 11 (Sub-Root Function) *The function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is said to be a sub-root function if ϕ is non-negative, non-decreasing, not identically zero, and $\phi(r)/\sqrt{r}$ is non-increasing.*

Assumption 1 (Variance-Expectation Bound) *The function class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}}$ satisfies the following properties: (1) $\mathbb{E}[g] \geq 0$ for all $g \in \mathcal{G}$ and (2) there exists constants $A, B \geq 0$ such that for all $g \in \mathcal{G}$, $\mathbb{E}[g^2] \leq A + B \mathbb{E}[g]$.*

Assumption 1 is a slight generalization of the variance-expectation bound stated in Assumption 1.4 from Bousquet (2002); the difference is that we allow an additional constant offset $A \geq 0$.

Theorem 12 (Modified Version of Theorem 6.2 in Bousquet (2002)) *Let $\mathcal{G}$ be a class of functions such that for all $g \in \mathcal{G}$, $\|g\|_\infty \leq b$ and $\mathcal{G}$ satisfies Assumption 1 with parameters (A, B).*

Let ϕ_n be a sub-root function such that

$$\mathbb{E}_\sigma \left[\sup_{g : \widehat{\mathbb{E}}_n[g^2] \leq r} \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i) \right] \leq \phi_n(r).$$

Define $r_n^\star$ to be the largest solution of $\phi_n(r) = r$. For any $\delta > 0$, we have with probability at least $1 - \delta$ for all $g \in \mathcal{G}$:

$$|\mathbb{E}[g] - \widehat{\mathbb{E}}_n[g]| \leq C \left(br_n^\star + \sqrt{r_n^\star(A + B\mathbb{E}[g])} + \sqrt{\frac{B\mathbb{E}[g]\left(\log \frac{1}{\delta} + c \log \log n\right)}{n}} + r_0 \right).$$

where $C > 0$ is an absolute numerical constant and $r_0 := \sqrt{\frac{2b^2 A\left(\log \frac{1}{\delta} + c \log \log n\right)}{n}} + \frac{22b^2\left(\log \frac{1}{\delta} + c \log \log n\right)}{n}$.

Proof. For all $k \in \mathbb{N}$ define $\gamma_k = b/2^k$. We define $\mathcal{G}_k := \{g \in \mathcal{G} : \gamma_{k+1} < \mathbb{E}[g] \leq \gamma_k\}$, so that $\mathcal{G} = \cup_{k \geq 0} \mathcal{G}_k$.

By Lemma 10 and Assumption 1, we have with probability at least $1 - \delta$ for all $k \geq 0$ and $g \in \mathcal{G}_k$:

$$|\widehat{\mathbb{E}}_n[g] - \mathbb{E}[g]| \leq 8\mathfrak{R}_n(\mathcal{G}_k) + \sqrt{\frac{2(A + B\gamma_k)\left(\log \frac{1}{\delta} + c \log \log \frac{b}{\gamma_k}\right)}{n}} + \frac{20b\left(\log \frac{1}{\delta} + c \log \log \frac{b}{\gamma_k}\right)}{n}. \tag{13}$$

In addition, we can apply Lemma 10 to the squares of $g \in \mathcal{G}_k$ to get that with probability at least $1 - \delta$ for all $k \geq 0$ and every $g \in \mathcal{G}_k$:

$$|\widehat{\mathbb{E}}_n[g^2] - \mathbb{E}[g^2]| \leq 8\mathfrak{R}_n(\mathcal{G}_k^2) + \sqrt{\frac{2b^2(A + B\gamma_k)\left(\log \frac{1}{\delta} + c \log \log \frac{b}{\gamma_k}\right)}{n}} + \frac{20b^2\left(\log \frac{1}{\delta} + c \log \log \frac{b}{\gamma_k}\right)}{n}$$

19

$$\leq 16b\mathfrak{R}_n(\mathcal{G}_k) + \sqrt{\frac{2b^2(A + B\gamma_k)\left(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\right)}{n}} + \frac{20b^2\left(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\right)}{n},$$
(14)

where the last inequality uses the fact that $x \mapsto x^2$ is $2b$-Lipschitz and centered at 0.

Now we condition on Eq. (13) and (14), which happens with probability at least $1 - 2\delta$. By Eq. (14) and Assumption 1, for all $g \in \mathcal{G}_k$,

$$\widehat{\mathbb{E}}_n[g^2] \leq (A + B\gamma_k) + 16b\mathfrak{R}_n(\mathcal{G}_k) + \sqrt{\frac{2b^2(A + B\gamma_k)\left(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\right)}{n}} + \frac{20b^2\left(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\right)}{n}.$$

Define the RHS of the previous display to be U_k. By definition of ϕ_n, we know that

$$\mathfrak{R}_n(\mathcal{G}_k) \leq \sup_{x_1,\cdots,x_n} \mathbb{E}\,\sigma\left[\sup_{g:\widehat{\mathbb{E}}_n[g^2]\leq U_k} \frac{1}{n}\sum_{i=1}^n \sigma_i g(x_i)\right] \leq \phi_n(U_k),$$

so therefore

$$U_k \leq (A + B\gamma_k) + 16b\phi_n(U_k) + \sqrt{\frac{2b^2(A + B\gamma_k)\left(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\right)}{n}} + \frac{20b^2\left(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\right)}{n}.$$

Let us define k_0 to be the largest value such that $\gamma_{k_0+1} \geq \frac{b}{n}$. For all $k \leq k_0$, we have $c\log\log\frac{b}{\gamma_k} \leq c\log\log n$. Therefore

$$U_k \leq A + B\gamma_k + 16b\phi_n(U_k) + \sqrt{\frac{2b^2(A + B\gamma_k)\left(\log\frac{1}{\delta} + c\log\log n\right)}{n}} + \frac{20b^2\left(\log\frac{1}{\delta} + c\log\log n\right)}{n}$$

$$\leq A + 2B\gamma_k + 16b\phi_n(U_k) + \underbrace{\sqrt{\frac{2b^2 A\left(\log\frac{1}{\delta} + c\log\log n\right)}{n}} + \frac{22b^2\left(\log\frac{1}{\delta} + c\log\log n\right)}{n}}_{=:r_0}.$$

If $U_k \geq r_n^\star$, then by definition of the sub-root function $\phi_n(U_k)/\sqrt{U_k} \leq \phi_n(r_n^\star)/\sqrt{r_n^\star} = \sqrt{r_n^\star}$, so

$$U_k \leq 16b\sqrt{U_k r_n^\star} + A + 2B\gamma_k + r_0 \leq C\left(b^2 r_n^\star + A + B\gamma_k + r_0\right) =: r_n(\gamma_k)$$

for some absolute constant $C > 0$. The last inequality holds by Fact 1. In addition if $U_k < r_n^\star$, the conclusion of the previous display trivially holds.

By Eq. (13), we also know that for any $g \in \mathcal{G}_k$

$$\mathbb{E}[g] \leq \widehat{\mathbb{E}}_n[g] + 8\mathfrak{R}_n(\mathcal{G}_k) + \sqrt{\frac{2B\gamma_k\left(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\right)}{n}} + r_0$$

$$\leq \widehat{\mathbb{E}}_n[g] + 8\phi_n(r_n(\gamma_k)) + \sqrt{\frac{2B\gamma_k\left(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\right)}{n}} + r_0.$$

20

By definition of $\mathcal{G}_k$ and γ_k, we know that $\gamma_k \leq 2\,\mathbb{E}[g]$ so that

$$\mathbb{E}[g] \leq \widehat{\mathbb{E}}_n[g] + 8 \cdot \phi_n(r_n(2\,\mathbb{E}[g])) + \sqrt{\frac{4B\,\mathbb{E}[g]\Big(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\Big)}{n}} + r_0$$

$$\leq \widehat{\mathbb{E}}_n[g] + 8\sqrt{r_n^\star} \cdot \sqrt{C(b^2 r_n^\star + A + B\,\mathbb{E}[g] + r_0)} + \sqrt{\frac{4B\,\mathbb{E}[g]\Big(\log\frac{1}{\delta} + c\log\log\frac{b}{\gamma_k}\Big)}{n}} + r_0$$

$$\leq \widehat{\mathbb{E}}_n[g] + C'\Big(br_n^\star + \sqrt{r_n^\star(A + B\,\mathbb{E}[g])}\Big) + \sqrt{\frac{4B\,\mathbb{E}[g]\big(\log\frac{1}{\delta} + c\log\log n\big)}{n}} + 2r_0,$$

for some absolute numerical constant $C' > 0$. When $k \geq k_0$, we have $\gamma_k \leq \frac{b}{n}$, so the previous display also trivially holds.

Lastly, we can repeat the same argument with the function class $\mathcal{G}' = \{-g : g \in \mathcal{G}\}$ to get the two-sided bound. This concludes the proof of Theorem 12. ∎

B.2. Calculating the Complexity Radius for LLP Losses

Now we present results which allow us to calculate the complexity radius $r_n^\star$ for function classes with bounded VC dimension.

Lemma 13 (Refined Dudley's Inequality, Lemma A.1 from Srebro et al. (2010)) *For any function class $\mathcal{F} : \mathcal{X} \to \mathbb{R}$,*

$$\mathfrak{R}_n(\mathcal{F}) \leq \inf_{\eta > 0}\left\{4\eta + 12\int_\eta^{\sup_{f \in \mathcal{F}} \sqrt{\widehat{\mathbb{E}}_n[f^2]}} \sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \varepsilon, n)}{n}}\, d\varepsilon\right\}.$$

Lemma 14 (Theorem 2.14 in Mendelson (2003)) *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ with VC dimension d. Then*

$$\log \mathcal{N}_2(\mathcal{F}, \varepsilon, n) \leq d\log\left(4e^2 \log \frac{2e^2}{\varepsilon}\right) + (2d) \cdot \log\left(\frac{1}{\varepsilon}\right).$$

We next present a lemma which allows us to calculate the local Rademacher complexity for various loss functions for LLP. Fix any $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ with VC dimension d. Consider any loss $\ell : [0,1] \times [0,1] \to \mathbb{R}$. Define the constrained loss class

$$\mathcal{L}_\ell(r) := \left\{(B, \alpha) \mapsto \ell\left(\frac{1}{k}\sum_{j=1}^k f(x_j), \alpha\right) : f \in \mathcal{F}, \widehat{\mathbb{E}}_n\left[\ell\left(\frac{1}{k}\sum_{j=1}^k f(x_j), \alpha\right)^2\right] \leq r\right\}.$$

We also let $\mathcal{L}_\ell$ denote the unrestricted loss class which contains all bag-level losses for $f \in \mathcal{F}$.

Lemma 15 *Let $\ell : [0,1] \times [0,1] \to \mathbb{R}$ be any λ-Lipschitz (in the first argument) bag loss. For any $n > d$ and any $r > 0$, we have*

$$\mathfrak{R}_n(\mathcal{L}_\ell(r)) \leq C\sqrt{\frac{rd\log\left(\frac{\lambda n}{r}\right)}{n}}$$

for some absolute numerical constant $C > 0$. Furthermore, if we denote the sub-root function $\phi_n(r) := C\sqrt{\frac{rd\log(\frac{\lambda n}{r})}{n}}$, and let $r_n^\star$ be the largest number such that $\phi_n(r) = r$, we have

$$r_n^\star \leq O\left(\frac{d\log(\lambda n)}{n}\right).$$

Proof. We use Lemma 13 applied to $\mathcal{L}_\ell(r)$. This gives

$$\mathfrak{R}_n(\mathcal{L}_\ell(r)) \leq \inf_{\eta>0}\left\{4\eta + 12\int_\eta^{\sqrt{r}}\sqrt{\frac{\log\mathcal{N}_2(\mathcal{L}_\ell(r),\varepsilon,n)}{n}}\,d\varepsilon\right\}. \tag{15}$$

From here we bound the covering numbers of the loss class in terms of the function class as

$$\log\mathcal{N}_2(\mathcal{L}_\ell(r),\varepsilon,n) \leq \log\mathcal{N}_2(\mathcal{L}_\ell,\varepsilon,n) \leq \log\mathcal{N}_2(\mathcal{F},\varepsilon/\lambda,nk).$$

The last inequality follows because for any f, f_ε and $(B_1,\alpha_1),\cdots,(B_n,\alpha_n)$:

$$\sqrt{\frac{1}{n}\sum_{i=1}^n\left(\ell\left(\frac{1}{k}\sum_{j=1}^k f(x_{i,j}),\alpha\right) - \ell\left(\frac{1}{k}\sum_{j=1}^k f_\varepsilon(x_{i,j}),\alpha\right)\right)^2}$$

$$\leq \lambda\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{k}\sum_i f(x_{i,j}) - f_\varepsilon(x_{i,j})\right)^2} \leq \lambda\sqrt{\frac{1}{nk}\sum_{i=1}^n\sum_{j=1}^k(f(x_{i,j}) - f_\varepsilon(x_{i,j}))^2},$$

where we use the λ-Lipschitz property of ℓ as well as Jensen's inequality. Thus, an empirical ℓ_2 cover of $\mathcal{F}$ at scale ε/λ for nk points implies an empirical ℓ_2 cover of $\mathcal{L}$ at scale ε for n points.

From here, we can apply the covering number bound of Lemma 14 to get

$$\log\mathcal{N}_2(\mathcal{F},\varepsilon/\lambda,nk) \leq d\log\left(4e^2\log\frac{2e^2\lambda}{\varepsilon}\right) + (2d)\cdot\log\left(\frac{\lambda}{\varepsilon}\right) \leq Cd\log\left(\frac{\lambda}{\varepsilon}\right),$$

for some absolute numerical constant $C > 0$.

Now we can plug the covering number bound back into Eq. (15) to get

$$\mathfrak{R}_n(\mathcal{L}_\ell(r)) \leq \inf_{\eta>0}\left\{4\eta + 12\sqrt{C}\int_\eta^{\sqrt{r}}\sqrt{\frac{d\log\frac{\lambda}{\varepsilon}}{n}}\,d\varepsilon\right\}.$$

Choosing $\eta = \Theta(\sqrt{rd/n})$ gives

$$\mathfrak{R}_n(\mathcal{L}_\ell(r)) \leq C'\sqrt{\frac{rd\log\left(\frac{\lambda n}{r}\right)}{n}},$$

for some absolute numerical constant $C' > 0$. For the proof of the second part, it is easy to see that the solution to the equation $C'\sqrt{\frac{r_n^\star d\log(\frac{\lambda n}{r_n^\star})}{n}} = r$ must satisfy $r_n^\star \leq C''\cdot\frac{d\log(\lambda n)}{n}$ for some absolute constant $C'' > 0$. This concludes the proof of Lemma 15. ■

Appendix C. Proof of Theorem 3

C.1. Notation and Preliminaries

We define several quantities which will be used in the proof. The loss function $\widehat{L}_{\mathrm{DSQ}}(f)$ is the difference between a proportion matching term and a debiasing term. For a given bag (B, α) and function f, we let $\ell_{\mathrm{SQ}}(f, (B, \alpha)) := k \cdot (\frac{1}{k} \sum_{j=1}^{k} f(x_j) - \alpha)^2$ and let

$$\widehat{L}_{\mathrm{SQ}}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{SQ}}(f, (B_i, \alpha_i)), \quad \text{and} \quad L_{\mathrm{SQ}}(f) := \mathbb{E}_B[\ell_{\mathrm{SQ}}(f, (B, \alpha))].$$

In addition, we let

$$\widehat{B}(f) := (k-1) \cdot (\widehat{\mathbb{E}}f - \widehat{p})^2, \quad \text{and} \quad B(f) := (k-1) \cdot (\mathbb{E}f - p)^2.$$

Written in this notation, we have $\widehat{L}_{\mathrm{DSQ}}(f) := \widehat{L}_{\mathrm{SQ}}(f) - \widehat{B}(f)$. In addition, recall that we showed in Section 3 that for any predictor f we have $\mathcal{L}(f) = L_{\mathrm{SQ}}(f) - B(f)$.

We also establish several elementary facts about $\widehat{B}(f)$, $\widehat{L}_{\mathrm{SQ}}$, and $\widehat{L}_{\mathrm{DSQ}}$.

Lemma 16 *The following statements are true for any predictor f and dataset $\{(B_i, \alpha_i)\}_{i=1}^{n}$.*

1. $\widehat{B}(f) \leq \frac{k-1}{k} \cdot \widehat{L}_{\mathrm{SQ}}(f)$.

2. $\widehat{L}_{\mathrm{DSQ}}(f) \geq 0$.

3. $\widehat{L}_{\mathrm{SQ}}(f) \leq k \cdot \widehat{L}_{\mathrm{DSQ}}(f)$.

Proof. Fix any predictor f. For the first statement, using Jensen's inequality we can compute that

$$\widehat{B}(f) = (k-1) \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{k} f(x_{i,j}) - \alpha_i \right)^2$$

$$\leq (k-1) \cdot \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{k} \sum_{j=1}^{k} f(x_{i,j}) - \alpha_i \right)^2 = \frac{k-1}{k} \widehat{L}_{\mathrm{SQ}}(f).$$

The second statement is a simple consequence of the first statement. For the third statement we have

$$\widehat{L}_{\mathrm{SQ}}(f) = \widehat{L}_{\mathrm{DSQ}}(f) + \widehat{B}(f) \leq \widehat{L}_{\mathrm{DSQ}}(f) + \frac{k-1}{k} \cdot \widehat{L}_{\mathrm{SQ}}(f),$$

and rearranging yields the statement. This proves Lemma 16. ∎

C.2. Showing Optimistic Rates

To prove Theorem 3, we separately prove optimistic rates for the square loss and debiasing terms, and then combine the guarantees. Consider the function class

$$\mathcal{G} := \left\{ (B, \alpha) \mapsto k \cdot \left(\frac{1}{k} \sum_{j=1}^{k} f(x_j) - \alpha \right)^2 : f \in \mathcal{F} \right\}.$$

For any function $g \in \mathcal{G}$, we know that $\|g\| \leq k$ and that g is k-Lipschitz in the first argument; furthermore, by nonnegativity, $\mathcal{G}$ satisfies Assumption 1 with parameters $A = 0$, $B = k$. Thus we apply Theorem 12 to $\mathcal{G}$ to get that with probability at least $1 - \delta$ for all $g \in \mathcal{G}$,

$$\left|\mathbb{E}[g] - \widehat{\mathbb{E}}_n[g]\right| \leq C\left(kr_n^\star + \sqrt{r_n^\star(k\,\mathbb{E}[g])} + \sqrt{\frac{k\,\mathbb{E}[g]\left(\log\frac{1}{\delta} + c\log\log n\right)}{n}} + \frac{k^2\left(\log\frac{1}{\delta} + c\log\log n\right)}{n} \right).$$

where $r_n^\star$ is the critical radius and $C > 0$ is an absolute numerical constant. Applying Lemma 15, we can also calculate the critical radius as

$$r_n^\star \leq O\left(\frac{d\log(kn)}{n}\right),$$

so therefore we have with probability at least $1 - \delta$ for any $f \in \mathcal{F}$:

$$\left|\widehat{L}_{\mathrm{SQ}}(f) - L_{\mathrm{SQ}}(f)\right| \leq \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{L_{\mathrm{SQ}}(f)\cdot k\left(d + \log\frac{1}{\delta}\right)}{n}} \right). \tag{16}$$

Next, we show a uniform convergence bound which relates $\widehat{B}(f)$ to $B(f)$. Consider the partition

$$\mathcal{G}^+ := \{(x,y) \mapsto f(x) - y : f \in \mathcal{F}, \mathbb{E}\,f \geq p\},$$
$$\mathcal{G}^- := \{(x,y) \mapsto y - f(x) : f \in \mathcal{F}, \mathbb{E}\,f < p\}.$$

The function class $\mathcal{G}^+$ is 1-Lipschitz, bounded in $[-1, +1]$, and satisfies Assumption 1 with $A = 0$, $B = 1$. Applying Theorem 12 to $\mathcal{G}^+$ and using Lemma 15 we get that with probability at least $1 - \delta$ for any $f \in \mathcal{F}$ such that $\mathbb{E}\,f \geq p$:

$$\left|\,|\widehat{\mathbb{E}}f - \widehat{p}| - |\mathbb{E}\,f - p|\,\right| \leq \widetilde{O}\left(\sqrt{\frac{(\mathbb{E}\,f - p)\left(d + \log\frac{1}{\delta}\right)}{nk}} + \frac{d + \log\frac{1}{\delta}}{nk} \right). \tag{17}$$

This implies

$$(\widehat{\mathbb{E}}f - \widehat{p})^2 \leq (\mathbb{E}\,f - p)^2 + \widetilde{O}\left(\sqrt{\frac{(\mathbb{E}\,f - p)^3\left(d + \log\frac{1}{\delta}\right)}{nk}} + \frac{d + \log\frac{1}{\delta}}{nk} \right)$$

$$\Rightarrow \quad \widehat{B}(f) \leq B(f) + \widetilde{O}\left(\sqrt{\frac{B(f)\left(d + \log\frac{1}{\delta}\right)}{n}} + \frac{d + \log\frac{1}{\delta}}{n} \right). \tag{18}$$

Likewise, for the reverse inequality we can also get that

$$(\mathbb{E}\,f - p)^2 \leq (\widehat{\mathbb{E}}f - \widehat{p})^2 + \widetilde{O}\left(|\widehat{\mathbb{E}}f - \widehat{p}| \cdot \sqrt{\frac{(\mathbb{E}\,f - p)\left(d + \log\frac{1}{\delta}\right)}{nk}} + \frac{d + \log\frac{1}{\delta}}{nk} \right)$$

$$\leq (\widehat{\mathbb{E}}f - \widehat{p})^2 + \widetilde{O}\left(\sqrt{\frac{(\mathbb{E}\,f - p)^3\left(d + \log\frac{1}{\delta}\right)}{nk}} + \frac{d + \log\frac{1}{\delta}}{nk} \right)$$

24

$$\Rightarrow \quad B(f) \leq \widehat{B}(f) + \widetilde{O}\left(\sqrt{\frac{B(f)\big(d + \log\frac{1}{\delta}\big)}{n}} + \frac{d + \log\frac{1}{\delta}}{n} \right), \tag{19}$$

where the second inequality uses the bound in Eq. (17).

Combining Eq. (18) and (19), we get the two-sided bound for all $f \in \mathcal{F}$ such that $\mathbb{E} f \geq p$:

$$|B(f) - \widehat{B}(f)| \leq \widetilde{O}\left(\sqrt{\frac{B(f)\big(d + \log\frac{1}{\delta}\big)}{n}} + \frac{d + \log\frac{1}{\delta}}{n} \right) \tag{20}$$

Following the same approach, we can show that the Eq. (20) also holds for any $f \in \mathcal{F}$ such that $\mathbb{E} f < p$, as the function class $\mathcal{G}^-$ also is 1-Lipschitz, bounded in $[-1, +1]$ and satisfies Assumption 1 with $A = 0, B = 1$. Therefore we conclude that with probability at least $1 - \delta$, Eq. (20) holds uniformly for all $f \in \mathcal{F}$.

Now we use Eqs. (16) and (20) to get

$$
\begin{aligned}
\mathcal{L}(\widehat{f}_{\mathrm{DSQ}}) &= L_{\mathrm{SQ}}(\widehat{f}_{\mathrm{DSQ}}) - B(\widehat{f}_{\mathrm{DSQ}}) \\
&\leq \widehat{L}_{\mathrm{SQ}}(\widehat{f}_{\mathrm{DSQ}}) - \widehat{B}(\widehat{f}_{\mathrm{DSQ}}) \\
&\quad + \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{L_{\mathrm{SQ}}(\widehat{f}_{\mathrm{DSQ}}) \cdot k\big(d + \log\frac{1}{\delta}\big)}{n}} + \sqrt{\frac{B(\widehat{f}_{\mathrm{DSQ}})\big(d + \log\frac{1}{\delta}\big)}{n}} \right) \\
&\leq \widehat{L}_{\mathrm{SQ}}(\widehat{f}_{\mathrm{DSQ}}) - \widehat{B}(\widehat{f}_{\mathrm{DSQ}}) + \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(\widehat{f}_{\mathrm{DSQ}}) \cdot k^2\big(d + \log\frac{1}{\delta}\big)}{n}} \right) \\
&\leq \widehat{L}_{\mathrm{SQ}}(f^\star) - \widehat{B}(f^\star) + \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(\widehat{f}_{\mathrm{DSQ}}) \cdot k^2\big(d + \log\frac{1}{\delta}\big)}{n}} \right) \\
&\leq \mathcal{L}(f^\star) + \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(\widehat{f}_{\mathrm{DSQ}}) \cdot k^2\big(d + \log\frac{1}{\delta}\big)}{n}} \right).
\end{aligned}
$$

The third inequality uses the fact that for any predictor f, we have $kL_{\mathrm{SQ}}(f) + B(f) = k\mathcal{L}(f) + (k+1)B(f) \leq (k^2 + k - 1)\mathcal{L}(f)$. The fourth and fifth inequalities uses the optimality of $\widehat{f}_{\mathrm{DSQ}}$ and $f^\star$ for the empirical and population minimization problems respectively.

Finally, using the inequality Fact 1 we get that

$$
\begin{aligned}
\mathcal{L}(\widehat{f}_{\mathrm{DSQ}}) &\leq \mathcal{L}(f^\star) + \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(\widehat{f}_{\mathrm{DSQ}}) \cdot k^2\big(d + \log\frac{1}{\delta}\big)}{n}} \right) \\
&\leq \mathcal{L}(f^\star) + \widetilde{O}\left(\frac{k^2\big(d + \log\frac{1}{\delta}\big)}{n} + \sqrt{\mathcal{L}(f^\star) + \frac{kd + k^2\log\frac{1}{\delta}}{n}} \cdot \sqrt{\frac{k^2\big(d + \log\frac{1}{\delta}\big)}{n}} \right)
\end{aligned}
$$

$$= \mathcal{L}(f^\star) + \widetilde{O}\left(\frac{k^2\left(d + \log\frac{1}{\delta}\right)}{n} + \sqrt{\frac{\mathcal{L}(f^\star) \cdot k^2\left(d + \log\frac{1}{\delta}\right)}{n}}\right)$$

This concludes the proof of Theorem 3.

Appendix D. Proofs for Section 4

D.1. Offset Loss Class

To analyze the performance of $\widehat{f}_{\mathrm{EZ}}$, we consider an *offset* version of the EasyLLP loss estimate. Specifically, let $f^\star := \inf_{f \in \mathcal{F}} \mathcal{L}(f)$. We define the offset loss

$$\widehat{\Gamma}(f, f^\star) := \widehat{L}_{\mathrm{EZ}}(f) - \widehat{L}_{\mathrm{EZ}}(f^\star) = \frac{1}{n}\sum_{i=1}^{n}(k(2\alpha_i - 2p) + (2p - 1)) \cdot \left(\frac{1}{k}\sum_{j=1}^{k} f^\star(x_{i,j}) - f(x_{i,j})\right)$$

Moreover, we use $\Gamma(f, f^\star)$ to denote its expectation, and we have $\mathbb{E}[\widehat{\Gamma}(f, f^\star)] = \mathcal{L}(f) - \mathcal{L}(f^\star)$. Clearly, minimizing the original EasyLLP loss is equivalent to minimizing $\widehat{\Gamma}(f, f^\star)$.

We also define the (empirical) second moment of the Γ function as

$$\widehat{\Gamma}^2(f, f^\star) := \frac{1}{n}\sum_{i=1}^{n}(k(2\alpha_i - 2p) + (2p - 1))^2 \cdot \left(\frac{1}{k}\sum_{j=1}^{k} f^\star(x_{i,j}) - f(x_{i,j})\right)^2,$$

and use $\Gamma^2(f, f^\star)$ to denote $\mathbb{E}\left[\widehat{\Gamma}^2(f, f^\star)\right]$.

We show that the offset loss class

$$\mathcal{G} := \left\{(B, \alpha) \mapsto (k(2\alpha - 2p) + (2p - 1)) \cdot \left(\frac{1}{k}\sum_{j=1}^{k} f^\star(x_j) - f(x_j)\right) : f \in \mathcal{F}\right\}$$

satisfies Assumption 1.

Lemma 17 *The function class $\mathcal{G}$ satisfies Assumption 1 with $A = 8k^2\mathcal{L}(f^\star)$ and $B = 4k^2$.*

Proof. First we observe that because $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}(f)$, we have $\Gamma(f, f^\star) \geq 0$ for all $f \in \mathcal{F}$. For the variance bound we can compute that

$$\Gamma^2(f, f^\star) \leq 4 \cdot \mathbb{E}\left[\left(\sum_{j=1}^{k} f^\star(x_j) - f(x_j)\right)^2\right] \leq 4k^2\,\mathbb{E}\left[(f^\star(x) - f(x))^2\right]$$

$$= 4k^2\,\mathbb{E}\left[(f^\star(x) - y + y - f(x))^2\right] \leq 4k^2(\mathcal{L}(f) + \mathcal{L}(f^\star))$$

$$= 8k^2\mathcal{L}(f^\star) + 4k^2\Gamma(f, f^\star).$$

The first inequality uses the fact that $(k(2\alpha - 2p) + (2p - 1)) \in [-2k + 1, 2k - 1]$. The second inequality uses the independence of the $\{x_j\}$ as well as Jensen's inequality. The third inequality uses the fact that $(f(x) - y)^2 = \mathbb{1}\{f(x) \neq y\}$, and that the cross terms satisfy $(f^\star(x) - y)(y - f(x)) \leq 0$.

26

This concludes the proof of Lemma 17. ■

D.2. Uniform Convergence for Offset Loss Class

Now we are ready to prove a uniform convergence bound for the function class $\mathcal{G}$.

Lemma 18 (Lemma 6, restated) *Let $f^\star = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}(f)$. Then with probability at least $1 - \delta$ we have for all $f \in \mathcal{F}$*

$$\left| \widehat{\Gamma}(f, f^\star) - \Gamma(f, f^\star) \right| \leq \widetilde{O} \left(\frac{kd + k^2 \log \frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(f) \cdot k^2 \left(d + \log \frac{1}{\delta} \right)}{n}} \right).$$

Proof. By Lemma 17, we know that the function class $\mathcal{G}$ satisfies Assumption 1 with $A = 8k^2 \mathcal{L}(f^\star)$ and $B = 4k^2$. We also know that $\|g\|_\infty \leq k$ for all $g \in \mathcal{G}$. Therefore, we can apply Theorem 12 to get that with probability at least $1 - \delta$ for all $f \in \mathcal{F}$,

$$\left| \widehat{\Gamma}(f, f^\star) - \Gamma(f, f^\star) \right|$$

$$\leq C \left(kr_n^\star + \sqrt{r_n^\star \cdot 2k^2 \mathcal{L}(f)} + \sqrt{\frac{k^2 \mathcal{L}(f) \left(\log \frac{1}{\delta} + c \log \log n \right)}{n}} + \frac{k^2 \left(\log \frac{1}{\delta} + c \log \log n \right)}{n} \right). \quad (21)$$

where $r_n^\star$ is the critical radius and $C > 0$ is an absolute numerical constant.

Since the function $(B, \alpha) \mapsto (k(2\alpha - 2p) + (2p - 1)) \cdot \left(\frac{1}{k} \sum_{j=1}^{k} f^\star(x_j) - f(x_j) \right)$ is $2k$-Lipschitz, we have by Lemma 15 that

$$r_n^\star = O \left(\frac{d \log(kn)}{n} \right).$$

Plugging this into Eq. (21) we get the conclusion of Lemma 6. ■

D.3. Proof of Theorem 5

Now we will prove the final generalization bound for EasyLLP. Using Lemma 6, we get that

$$\Gamma(\widehat{f}_{\mathrm{EZ}}, f^\star) \leq \widehat{\Gamma}(\widehat{f}_{\mathrm{EZ}}, f^\star) + \widetilde{O} \left(\frac{kd + k^2 \log \frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(\widehat{f}_{\mathrm{EZ}}) \cdot k^2 \left(d + \log \frac{1}{\delta} \right)}{n}} \right)$$

$$= \widehat{\Gamma}(\widehat{f}_{\mathrm{EZ}}, f^\star) + \widetilde{O} \left(\frac{kd + k^2 \log \frac{1}{\delta}}{n} + \sqrt{\frac{\left(\Gamma(\widehat{f}_{\mathrm{EZ}}, f^\star) + \mathcal{L}(f^\star) \right) \cdot k^2 \left(d + \log \frac{1}{\delta} \right)}{n}} \right)$$

In fact, we know that since $f^\star \in \mathcal{F}$, we must have $\widehat{\Gamma}(\widehat{f}_{\mathrm{EZ}}, f^\star) \le \widehat{\Gamma}(f^\star, f^\star) = 0$, so using Fact 1 we can further upper bound this as

$$\Gamma(\widehat{f}_{\mathrm{EZ}}, f^\star) \le \widetilde{O}\left(\frac{k^2\left(d + \log\frac{1}{\delta}\right)}{n} + \sqrt{\frac{\mathcal{L}(f^\star) \cdot k^2\left(d + \log\frac{1}{\delta}\right)}{n}}\right).$$

Plugging in the definition of Γ proves the bound.

D.4. Proof of Corollary 7

We describe the sample splitting version of the EasyLLP learning rule, which allows us to use an estimate $\widehat{p}$ instead of the true marginal label proportion p. We assume that we are given a dataset of size $2n$, denoted $S = \{(B_i, \alpha_i)\}_{i=1}^{2n}$. We split it randomly into two equally-sized parts S and S'. In the proof, we will use $i \in [n]$ to index bags in S and $i' \in [n]$ to index bags in S'.

1. Using S, estimate marginal label proportion $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n}\alpha_i$.

2. Return $\widehat{f}_{\mathrm{EZ}} := \operatorname{argmin}_{f\in\mathcal{F}}\frac{1}{n}\sum_{i'=1}^{n}\ell_{\mathrm{EZ}}(f,(B_{i'},\alpha_{i'}))$ using S', where $\ell_{\mathrm{EZ}}(\cdot,\cdot)$ is defined with the plug-in estimate $\widehat{p}$ instead of p:

$$\ell_{\mathrm{EZ}}(f,(B,\alpha)) = (k(\alpha - \widehat{p}) + \widehat{p}) \cdot \left(1 - \frac{1}{k}\sum_{j=1}^{k}f(x_j)\right) + (k(\widehat{p} - \alpha) + (1 - \widehat{p})) \cdot \left(\frac{1}{k}\sum_{j=1}^{k}f(x_j)\right).$$

For any $q \in [0,1]$ let us define the quantity

$$\widehat{\Gamma}_q(f, f^\star) := \frac{1}{n}\sum_{i'=1}^{n}(k(2\alpha_{i'} - 2q) + (2q - 1)) \cdot \left(\frac{1}{k}\sum_{j=1}^{k}f^\star(x_{i',j}) - f(x_{i',j})\right),$$

which represents the empirical offset loss estimated on S' if we plugged in the value q for the marginal label proportion.

By Lemma 6 we know that for any $f \in \mathcal{F}$

$$\Gamma_p(f, f^\star) \le \widehat{\Gamma}_p(f, f^\star) + \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(f) \cdot k^2\left(d + \log\frac{1}{\delta}\right)}{n}}\right).$$

From here, we need to relate $\widehat{\Gamma}_p(f, f^\star)$ to $\widehat{\Gamma}_{\widehat{p}}(f, f^\star)$. We can bound the difference as:

$$\left|\widehat{\Gamma}_p(f, f^\star) - \widehat{\Gamma}_{\widehat{p}}(f, f^\star)\right| \le (2k - 2) \cdot |\widehat{p} - p| \cdot \underbrace{\left|\frac{1}{n}\sum_{i'=1}^{n}\frac{1}{k}\sum_{j=1}^{k}f^\star(x_{i',j}) - f(x_{i',j})\right|}_{=:\Xi(f)}.$$

By Hoeffding's inequality (Theorem 19), with probability at least $1 - \delta$ over S, we have $|\widehat{p} - p| \le \sqrt{\frac{2\log(2/\delta)}{nk}}$. Henceforth we condition on this event holding over S.

Now we prove a uniform convergence bound on $|\Xi(\cdot)|$ for all $f \in \mathcal{F}$. For any $f \in \mathcal{F}$ we write that

$$|\Xi(f)| = \left|\frac{1}{n}\sum_{i'=1}^{n}\frac{1}{k}\sum_{j=1}^{k}f^{\star}(x_{i',j}) - f(x_{i',j})\right| \leq \sqrt{\frac{1}{nk}} \cdot \sqrt{\sum_{i'=1}^{n}\sum_{j=1}^{k}\left(f^{\star}(x_{i',j}) - f(x_{i',j})\right)^2}$$

$$\leq \sqrt{\frac{1}{nk}} \cdot \sqrt{\sum_{i'=1}^{n}\sum_{j=1}^{k}\left(f^{\star}(x_{i',j}) - y_{i',j}\right)^2 + (y_{i',j} - f(x_{i',j}))^2} = \sqrt{\widehat{L}(f^{\star}) + \widehat{L}(f)}.$$

Here, we use $\widehat{L}(\cdot)$ to denote the empirical classification loss on S'. The first inequality follows by Cauchy-Schwarz. The second inequality uses the fact that for $a, b, c \in \{0, 1\}$ we have $(a - b)(b - c) \leq 0$. The last equality follows because $(f(x) - y)^2 = \mathbb{1}\{f(x) \neq y\}$. Now we use the standard uniform convergence guarantee: with probability at least $1 - \delta$ over S', for any $f \in \mathcal{F}$:

$$|\mathcal{L}(f) - \widehat{L}(f)| \leq \widetilde{O}\left(\mathcal{L}(f) + \frac{d + \log\frac{1}{\delta}}{nk}\right).$$

So therefore with probability at least $1 - \delta$ over S' we have for all $f \in \mathcal{F}$:

$$|\Xi(f)| \leq \sqrt{\widehat{L}(f^{\star}) + \widehat{L}(f)} \leq \widetilde{O}\left(\sqrt{\mathcal{L}(f) + \frac{d + \log\frac{1}{\delta}}{nk}}\right).$$

Thus with probability at least $1 - 2\delta$ over the draws of S_1 and S_2, we have for all $f \in \mathcal{F}$:

$$\left|\widehat{\Gamma}_p(f, f^{\star}) - \widehat{\Gamma}_{\widehat{p}}(f, f^{\star})\right| \leq \widetilde{O}\left(\frac{d + \log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(f) \cdot k\log\frac{1}{\delta}}{n}}\right) \tag{22}$$

Now we are ready to prove the final guarantee. Similar to the proof of [Theorem 5](#) we compute that

$$\Gamma_p(\widehat{f}_{\text{EZ}}, f^{\star}) \leq \widehat{\Gamma}_p(\widehat{f}_{\text{EZ}}, f^{\star}) + \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(\widehat{f}_{\text{EZ}}) \cdot k^2(d + \log\frac{1}{\delta})}{n}}\right)$$

$$\leq \widehat{\Gamma}_{\widehat{p}}(\widehat{f}_{\text{EZ}}, f^{\star}) + \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(\widehat{f}_{\text{EZ}}) \cdot k^2(d + \log\frac{1}{\delta})}{n}}\right)$$

$$\leq \widetilde{O}\left(\frac{kd + k^2\log\frac{1}{\delta}}{n} + \sqrt{\frac{\mathcal{L}(\widehat{f}_{\text{EZ}}) \cdot k^2(d + \log\frac{1}{\delta})}{n}}\right),$$

where the second line uses Eq. [(22)](#) and the last line follows from the fact that $\widehat{\Gamma}_{\widehat{p}}(\widehat{f}_{\text{EZ}}, f^{\star}) \leq \widehat{\Gamma}_{\widehat{p}}(f^{\star}, f^{\star}) = 0$. From here, the proof concludes similarly as the proof of [Theorem 5](#) in [Appendix D.3](#) by using [Fact 1](#) and plugging in the definition of $\Gamma_p(\cdot, f^{\star})$.

Appendix E. Proof of [Theorem 8](#)

First, we describe the construction, then separately prove the lower bound for both the realizable and agnostic settings.

E.1. Construction

We define the instance space $\mathcal{X} = \{0,1\}^{2^d}$ and let $\mathcal{F} = \{f_i : i \in [2^d]\}$ where the function f_i is defined as $f_i(x) = x[i]$. It is clear that $\mathrm{VC}(\mathcal{F}) \leq d$. Now we define a family of distributions $\mathcal{D}_i$ for $i \in [2^d]$. For some parameter choice $\gamma \in [0, 1/2]$, each $\mathcal{D}_i$ is defined as follows: the example $x \sim \mathrm{Unif}(\{0,1\}^{2^d})$ and $y = f_i(x)$ with probability $1/2 + \gamma$, $y = 1 - f_i(x)$ with probability $1/2 - \gamma$. For any predictor f the classification loss for distribution $\mathcal{D}_i$ can be written as

$$\mathcal{L}_{\mathcal{D}_i}(f) - \inf_{f' \in \mathcal{F}} \mathcal{L}_{\mathcal{D}_i}(f') = 2\gamma \cdot \mathbb{E}_{x \sim \mathrm{Unif}(\{0,1\}^{2^d})}[\mathbb{1}\{f(x) \neq x[i]\}].$$

Furthermore, for any predictor f, as well as distributions $\mathcal{D}_i$ and $\mathcal{D}_j$ we have the separation condition

$$\mathcal{L}_{\mathcal{D}_i}(f) - \inf_{f' \in \mathcal{F}} \mathcal{L}_{\mathcal{D}_i}(f') + \mathcal{L}_{\mathcal{D}_j}(f) - \inf_{f' \in \mathcal{F}} \mathcal{L}_{\mathcal{D}_j}(f')$$

$$= 2\gamma \cdot \mathbb{E}_{x \sim \mathrm{Unif}(\{0,1\}^{2^d})}[\mathbb{1}\{f(x) \neq x[i]\} + \mathbb{1}\{f(x) \neq x[j]\}] \geq \gamma. \tag{23}$$

E.2. Realizable Setting

For the realizable setting result, we use the construction with $\gamma = 1/2$. We claim that for any learning rule for LLP that PAC learns $\mathcal{F}$, there exists a distribution $\mathcal{D}_i$ for which it requires $n = \Omega(d/\log k)$ samples in expectation. Let us define $\bar{\mathcal{D}}$ to be the averaged distribution where one first draws $i \sim \mathrm{Unif}([2^d])$ then samples the bag $(B, \alpha) \sim \mathcal{D}_i$. Using the separation condition of Eq. (23), we invoke Fano's inequality (e.g., Lemma 3 of Yu (1997)) to get

$$\inf_{\hat{f}} \sup_{i \in [2^d]} \mathbb{E}_{\mathcal{D}_i}\left[\mathcal{L}_{\mathcal{D}_i}\left(\hat{f}\right)\right] \geq \frac{1}{4}\left(1 - \frac{n \cdot \frac{1}{2^d} \sum_{i=1}^{2^d} \mathrm{KL}(\mathcal{D}_i \| \bar{\mathcal{D}}) + \log 2}{d}\right) = \frac{1}{4}\left(1 - \frac{n \cdot \mathrm{KL}(\mathcal{D}_1 \| \bar{\mathcal{D}}) + \log 2}{d}\right),$$
$$\tag{24}$$

where the equality follows by symmetry of the distributions $\mathcal{D}_i$.

From here we need to estimate $\mathrm{KL}(\mathcal{D}_1 \| \bar{\mathcal{D}})$. By chain rule for KL divergence, we see that

$$\mathrm{KL}(\mathcal{D}_1 \| \bar{\mathcal{D}}) = \mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[B] \| \mathbb{P}_{\bar{\mathcal{D}}}[B]) + \mathbb{E}_{B \sim \mathcal{D}_1}[\mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\bar{\mathcal{D}}}[\alpha \mid B])]$$
$$= \mathbb{E}_{B \sim \mathcal{D}_1}[\mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\bar{\mathcal{D}}}[\alpha \mid B])]$$
$$= \mathbb{E}_{B \sim \mathrm{Unif}(\{0,1\}^{2^d})}[\mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\bar{\mathcal{D}}}[\alpha \mid B])],$$

since all of the $\mathcal{D}_i$ have the same marginal over $\mathcal{X}$. Fix a bag $B = \{x_1, \cdots, x_k\}$, and let us define the vector $z = \frac{1}{k} \sum_{j=1}^k x_j \in [0, 1]^{2^d}$. We calculate that

$$\mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\bar{\mathcal{D}}}[\alpha \mid B]) = \sum_{\alpha \in \{0, \frac{1}{k}, \cdots, 1\}} \mathbb{P}_{\alpha \sim \mathcal{D}_1}[\alpha \mid B] \cdot \log \frac{\mathbb{P}_{\alpha \sim \mathcal{D}_1}[\alpha \mid B]}{\mathbb{P}_{\alpha \sim \bar{\mathcal{D}}}[\alpha \mid B]}$$
$$= \log \frac{1}{\mathbb{P}_{\alpha \sim \bar{\mathcal{D}}}[\alpha = z[1] \mid B]}$$
$$= \log \frac{1}{\frac{1}{2^d} + \frac{1}{2^d} \sum_{i>1} \mathbb{P}_{\alpha \sim \mathcal{D}_i}[\alpha = z[1] \mid B]}$$

30

$$= \log \frac{1}{\frac{1}{2^d} + \frac{1}{2^d} \sum_{i>1} \mathbb{1}\{z[i] = z[1]\}}$$

$$\leq \min\left\{ d, \log \frac{2^d}{\sum_{i>1} \mathbb{1}\{z[i] = z[1]\}} \right\}.$$

The second line follows from the fact that once we fix B, the value of $\alpha = z[1]$ is deterministic under $\mathcal{D}_1$. Putting it together we get that, we get

$$\mathrm{KL}\big(\mathcal{D}_1 \| \bar{\mathcal{D}}\big) \leq \mathbb{E}_{B \sim \mathrm{Unif}(\{0,1\}^{2^d})}\left[\min\left\{ d, \log \frac{2^d}{\sum_{i>1} \mathbb{1}\{z[i] = z[1]\}} \right\} \right].$$

Observe that $k \cdot z[i]$ is distributed as independent $\mathrm{Bin}(k, 1/2)$ variables for all $i \in [2^d]$. Using Lemma 21, we have for all $i > 1$ that $\mathbb{P}_B[z[i] = z[1]] \geq 1/(k+1)$, since the random variable $z[i]$ takes at most $k + 1$ values. Thus, applying Chernoff bounds we have for any $\delta \in (0, 1)$,

$$\mathbb{P}_B\left[\sum_{i>1} \mathbb{1}\{z[i] = z[1]\} \geq (1 - \delta)\frac{2^d - 1}{k+1} \right] \geq 1 - \exp\left(-\frac{\delta^2}{2} \cdot \frac{2^d - 1}{k+1} \right).$$

Let us pick $\delta = \sqrt{2(k+1)\log d/(2^d - 1)}$; by our assumption on k we have $\delta \leq 1/2$. This guarantees that the above event, call it $\mathcal{E}$, happens with probability at least $1 - 1/d$. Therefore, we get

$$\mathrm{KL}\big(\mathcal{D}_1 \| \bar{\mathcal{D}}\big) \leq d \cdot \frac{1}{d} + \log \frac{2^d \cdot 2(k+1)}{2^d - 1} \leq 2 + \log(k+1). \tag{25}$$

Therefore, plugging in Eq. (25) into Eq. (24) we see that

$$\inf_{\widehat{f}} \sup_{i \in [2^d]} \mathbb{E}_{\mathcal{D}_i}\left[\mathcal{L}_{\mathcal{D}_i}\big(\widehat{f}\big) \right] \geq \frac{1}{4}\left(1 - \frac{n(2 + \log(k+1)) + \log 2}{d} \right).$$

For $n = C'd/(\log(k+1))$ where $C' > 0$ is a sufficiently small constant we have $\inf_{\widehat{f}} \sup_{i \in [2^d]} \mathbb{E}_{\mathcal{D}_i}\left[\mathcal{L}_{\mathcal{D}_i}\big(\widehat{f}\big) \right] \geq \frac{1}{8}$. Fix any learning rule $\widehat{f}$, and let $\mathcal{D}_{i^\star}$ be the distribution which witnesses the supremum. We have

$$\frac{1}{8} \leq \mathbb{E}_{\mathcal{D}_{i^\star}}\left[\mathcal{L}_{\mathcal{D}_{i^\star}}\big(\widehat{f}\big) \right] \leq \mathbb{P}_{\mathcal{D}_{i^\star}}\left[\mathcal{L}_{\mathcal{D}_{i^\star}}\big(\widehat{f}\big) > \frac{1}{16} \right] + \frac{1}{16} \cdot \mathbb{P}_{\mathcal{D}_{i^\star}}\left[\mathcal{L}_{\mathcal{D}_{i^\star}}\big(\widehat{f}\big) \leq \frac{1}{16} \right]$$

which implies that $\mathbb{P}_{\mathcal{D}_{i^\star}}\left[\mathcal{L}_{\mathcal{D}_{i^\star}}\big(\widehat{f}\big) > \frac{1}{16} \right] \geq 1/15$. We conclude that any learning rule which PAC learns $\mathcal{F}$ with parameters $(\varepsilon, \delta) = (1/16, 1/15)$ requires $n = \Omega(d/\log k)$ samples.

E.3. Agnostic Setting

For the agnostic setting result we use the construction with $\gamma = \varepsilon$. Again, by Fano's inequality we get

$$\inf_{\widehat{f}} \sup_{i \in [2^d]} \mathbb{E}_{\mathcal{D}_i}\left[\mathcal{L}_{\mathcal{D}_i}\big(\widehat{f}\big) - \inf_{f' \in \mathcal{F}} \mathcal{L}_{\mathcal{D}_i}(f') \right] \geq \frac{\varepsilon}{2}\left(1 - \frac{n \cdot \frac{1}{2^d} \sum_{i=1}^{2^d} \mathrm{KL}\big(\mathcal{D}_i \| \bar{\mathcal{D}}\big) + \log 2}{d} \right)$$

$$= \frac{\varepsilon}{2}\left(1 - \frac{n \cdot \mathrm{KL}\big(\mathcal{D}_1 \| \bar{\mathcal{D}}\big) + \log 2}{d}\right), \tag{26}$$

where the equality follows by symmetry of the distributions $\mathcal{D}_i$.

From here we need to estimate $\mathrm{KL}\big(\mathcal{D}_1 \| \bar{\mathcal{D}}\big)$. By chain rule for KL divergence, we see that

$$\mathrm{KL}\big(\mathcal{D}_1 \| \bar{\mathcal{D}}\big) = \mathbb{E}_{B \sim \mathrm{Unif}(\{0,1\}^{2^d})}[\mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\bar{\mathcal{D}}}[\alpha \mid B])].$$

From here we can bound this in two ways. The first way is to use the data-processing inequality for KL divergence. Observe that in distribution $\mathcal{D}_i$, the label proportion α is distributed as $\mathrm{Bin}(kz[i], 1/2 + \varepsilon) + \mathrm{Bin}(k - kz[i], 1/2 - \varepsilon)$. Therefore, by the data processing inequality

$$\mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\bar{\mathcal{D}}}[\alpha \mid B]) \le \mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha_{\mathrm{clean}} \mid B] \| \mathbb{P}_{\bar{\mathcal{D}}}[\alpha_{\mathrm{clean}} \mid B])$$

$$\le \min\left\{d, \log \frac{2^d}{\sum_{i>1} \mathbb{1}\{z[i] = z[1]\}}\right\},$$

where $\alpha_{\mathrm{clean}} = z[i]$ in distribution $\mathcal{D}_i$. From here, the proof proceeds similarly as in the realizable setting. We get the lower bound of $\Omega(d/\log k)$, with no dependence on ε.

Alternatively, we can directly calculate the bound on the KL divergence:

$$\mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\bar{\mathcal{D}}}[\alpha \mid B])$$

$$\le \frac{1}{2^d} \sum_{i=1}^{2^d} \mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\mathcal{D}_i}[\alpha \mid B])$$

$$= \frac{2^d - 1}{2^d} \cdot \mathrm{KL}(\mathbb{P}_{\mathcal{D}_1}[\alpha \mid B] \| \mathbb{P}_{\mathcal{D}_2}[\alpha \mid B])$$

$$\le \max\{kz[1] - kz[2], 0\} \cdot \mathrm{kl}(1/2 + \varepsilon \| 1/2 - \varepsilon) + \max\{kz[2] - kz[1], 0\} \cdot \mathrm{kl}(1/2 - \varepsilon \| 1/2 + \varepsilon)$$

$$\le |z[1] - z[2]| \cdot O(k\varepsilon^2).$$

The first line uses the convexity of KL. The second line uses the symmetry of the distributions $\mathcal{D}_i$. The third line follows because under $\mathcal{D}_i$, $k\alpha \sim \mathrm{Bin}(kz[i], 1/2 + \varepsilon) + \mathrm{Bin}(k - kz[i], 1/2 - \varepsilon)$, then applying chain rule for KL divergence. The last line applies the bound on the KL divergence between two Bernoulli random variables.

Now we investigate the expected difference between $z[1]$ and $z[2]$. Note that both of these are independently distributed as $\mathrm{Bin}(k, 1/2)/k$. By Hoeffding's inequality, we see that for any $i \in [2^d]$,

$$\mathbb{P}_B\left[\left|z[i] - \frac{1}{2}\right| \ge \sqrt{\frac{\log(2k)}{2k}}\right] \le \frac{1}{k},$$

so by union bound, with probability at least $1 - 2/k$ we have $|z[1] - z[2]| \le \sqrt{\frac{2\log(2k)}{k}}$. Using this, we can compute the bound that

$$\mathrm{KL}\big(\mathcal{D}_1 \| \bar{\mathcal{D}}\big) \le O(k\varepsilon^2) \cdot \mathbb{E}_B[|z[1] - z[2]|]$$

$$\leq O(k\varepsilon^2) \cdot \left(\frac{2}{k} + \sqrt{\frac{2\log(2k)}{k}} \right)$$

$$\leq O(\varepsilon^2) \cdot \left(2 + \sqrt{2k\log(2k)} \right).$$

Plugging the previous display into Eq. (26), we see that if $n = C'd/(\sqrt{k}\varepsilon^2)$ where $C' > 0$ is a sufficiently small constant we have $\inf_{\widehat{f}} \sup_{i \in [2^d]} \mathbb{E}_{\mathcal{D}_i} \left[\mathcal{L}_{\mathcal{D}_i}\left(\widehat{f}\right) - \inf_{f' \in \mathcal{F}} \mathcal{L}_{\mathcal{D}_i}(f') \right] \geq \frac{\varepsilon}{4}$. As with the realizable setting proof, we can translate this to a lower bound for PAC learning; we omit the details.

Appendix F. Technical Lemmas

Fact 1 *For any $A, B, C \geq 0$ if $A \leq B + C\sqrt{A}$ then $A \leq B + C^2 + \sqrt{B}C \leq 2B + 2C^2$.*

Theorem 19 (Hoeffding's Inequality) *Let $Z_1, \cdots, Z_n$ be independent bounded random variables with $Z_i \in [a, b]$ for all $i \in [n]$. Then*

$$\mathbb{P}\left[\left| \frac{1}{n} \sum_{t=1}^{n} Z_i - \mathbb{E}[Z_i] \right| \geq t \right] \leq 2\exp\left(\frac{-2nt^2}{(b-a)^2} \right).$$

Theorem 20 (Paley-Zygmund) *Let $Z \geq 0$ be a random variable with finite variance. For any $\theta \in [0, 1]$,*

$$\mathbb{P}[Z \geq \theta \cdot \mathbb{E}\, Z] \geq (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

Lemma 21 *Let X and Y be two i.i.d. discrete random variables which are supported on a set of size k. Then $\mathbb{P}[X = Y] \geq 1/k$.*

Let Ω be the support. The probability that $X = Y$ can be calculated as

$$\mathbb{P}[X = Y] = \sum_{x \in \Omega} \mathbb{P}[X = x]^2 \geq \frac{1}{k} \cdot \left(\sum_{x \in \Omega} \mathbb{P}[X = x] \right)^2 = \frac{1}{k}.$$

The inequality uses the fact that $\|v\|_1 \leq \sqrt{k}\|v\|_2$ for any $v \in \mathbb{R}^k$. ∎

Appendix G. Experiments

In this section we empirically evaluate the performance of the learning rules discussed herein and present results which illustrate the differences of the learning rules in practical implementations.

Learning Rule Implementations. Since minimizing the 0-1 loss is computationally intractable, we consider algorithmic variants of LLP learning rules which minimize a surrogate loss using mini-batch stochastic gradient descent (SGD). Fix a parameterized function class $\mathcal{F} = \{f_\theta : \theta \in \Theta\} \subseteq [0,1]^{\mathcal{X}}$. For EasyLLP, we minimize the loss (10) with $\ell_{\text{ins}} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ instantiated to be the square loss EZ.Sq or the log loss EZ.Log. Directly minimizing the debiased square loss via SGD is

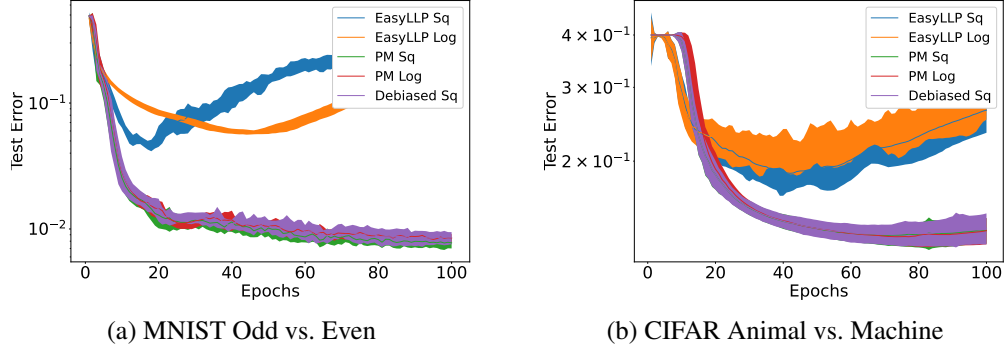(a) MNIST Odd vs. Even   (b) CIFAR Animal vs. Machine

Figure 1: Training curves of various algorithms for LLP, using the large CNN architecture and bag size $k = 100$. One standard deviation confidence bands are plotted.

not straightforward: for every gradient update, the second term in the loss requires computing $\widehat{\mathbb{E}} f_\theta$ (i.e., the average prediction of f_θ over the entire dataset). We approximate it with an exponential moving average $\widehat{v}$ with parameter $\beta \in (0,1)$. Given a minibatch $S = \{(B_1, \alpha_1), \cdots, (B_\ell, \alpha_\ell)\}$, we perform the updates

$$\widehat{v}^+ = \beta \cdot \widehat{v} + (1 - \beta) \cdot \frac{1}{\ell k} \sum_{i=1}^{\ell} \sum_{j=1}^{k} f_\theta(x_{i,j}), \tag{27}$$

$$\theta^+ = \theta - \eta \cdot \nabla_\theta \left(\frac{1}{\ell} \sum_{i=1}^{\ell} k \cdot \left(\frac{1}{k} \sum_{j=1}^{k} f_\theta(x_{i,j}) - \alpha_i \right)^2 - (k-1)\left(\widehat{v}^+ - \widehat{p}\right)^2 \right). \tag{28}$$

When performing backpropagation for the gradients on the debiasing term $(k-1)(\widehat{v}^+ - \widehat{p})^2$, the backpropagation does not go through $\widehat{v}$, only the summation over $f_\theta(x_{i,j})$. We refer to this as DebiasedSq. For DebiasedSq, EZ.Log, and EZ.Sq, the label proportion $\widehat{p}$ is estimated as the average of the training set labels. Lastly, we also consider algorithms that minimize the proportion matching square loss (6) and log loss (7), which we refer to as PM.Sq and PM.Log respectively.

G.1. Comparison of LLP Learning Rules

The goal of this experiment is to understand how these learning rules perform under various bag sizes and function classes.

Experimental Setup. We adopt a similar experiment setup as in (Busa-Fekete et al., 2023). We run LLP algorithms on the MNIST odd vs. even task and the CIFAR10 animal vs. machine task (binary classification versions of the MNIST and CIFAR10). We consider 5 architectures: a linear model, small two layer NN (with 100 hidden units), a large two layer NN (with 1000 hidden units), a small CNN, and a large CNN. We experiment with bag sizes $k \in \{10, 100, 1000\}$. For each dataset, model, bag size, and algorithm, we run 10 trials with different random seeds. To select the learning rate, we report the best average achieved test error for learning rates in $\{0.01, 0.005, 0.001, 0.0005, 0.0001\}$. In all of our experiments we use the Adam optimizer (Kingma and Ba, 2014) with minibatches of size 1000 and train for 100 epochs. For DebiasedSq, we use the approximate version in Eqs. (27)-(28) with $\beta = 0.99$. Further details can be found in Appendix G.3.

34

(a) Loss Estimates Tracking (b) EZ.Sq Loss Estimates (c) DebiasedSq Loss Estimates
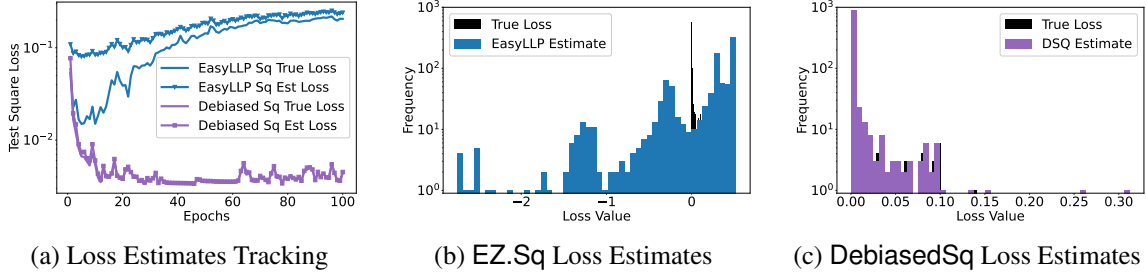
Figure 2: Left: Loss estimates throughout training. We run a single trial of EZ.Sq and DebiasedSq on MNIST Odd vs. Even using the large CNN architecture, bag size $k = 10$, and optimally chosen learning rate. Using the test set, we plot the averaged true square loss $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \frac{1}{k} \sum_{j=1}^{k} (f_\theta(x_{i,j}) - y_{i,j})^2$ vs. the estimated square loss $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \widehat{\ell}_{\text{est}}(B_i, \alpha_i)$, where $\widehat{\ell}_{\text{est}}$ is either the EZ.Sq/DebiasedSq loss estimate. Middle and Right: Histogram of true per-bag square losses $\{\frac{1}{k} \sum_{j=1}^{k} (f_\theta(x_{i,j}) - y_{i,j})^2\}_{i=1}^{n_{\text{test}}}$ and per-bag loss estimates $\{\widehat{\ell}_{\text{est}}(B_i, \alpha_i)\}_{i=1}^{n_{\text{test}}}$ for EZ.Sq/DebiasedSq loss estimates at epoch 10.

Discussion of Results. In Figure 1 as well as Table 1 and 2 (in the Appendix), we see that the best algorithm is a tossup between DebiasedSq, PM.Sq, and PM.Log.[1] In light of our discussion in Section 3.1, this may be not that surprising, since the model classes we work with are expressive enough to minimize both the proportion matching loss and the bias, so we do not have the failure mode shown in Proposition 2. Furthermore, in Figure 1 we see that both versions of EasyLLP exhibits overfitting as training progresses, thus necessitating early stopping; this was also observed in (Busa-Fekete et al., 2023). We observe that DebiasedSq, PM.Sq, and PM.Log also sometimes exhibits overfitting (see Figure 1b), but to a much less degree.

One hypothesis for why DebiasedSq performs better than EZ.Sq is that is that the debiased square loss is a more accurate estimate of the true loss than the EasyLLP loss (see Proposition 4 and the discussion before it). In Figure 2, we compare how well the estimated losses track the true square loss on the test set. Although we plot a single trial, we found that the behavior was similar across different random seeds. In Figure 2a, the DebiasedSq loss closely tracks the true square loss, but EZ.Sq is consistently an *overestimate* of the true square loss. (Interestingly, we observe that the "shape" of the loss curve is still preserved). An explanation for this phenomenon can be found in Figure 2b and Figure 2c, where we plot the histogram of ground-truth per-bag square losses as well as the corresponding per-bag square loss estimates in the test set. The histogram of per-bag loss estimates for DebiasedSq is quite similar to the histogram of ground-truth per-bag losses. On the other hand for EZ.Sq, the histogram of per-bag loss estimates has large variance and a heavy left tail. Since the randomness in the loss estimates is due to the bag generation procedure, it is likely that the bags with highly negative EZ.Sq loss estimates (i.e., taking values < -2) were not generated, so the sample mean overestimates the true square loss.

---

1. The results in (Busa-Fekete et al., 2023) suggest that EasyLLP and the proportion matching baselines have similar performance; however, note that they only train for 20 epochs. We find that EasyLLP is competitive in early stages of training but eventually DebiasedSq, PM.Sq, and PM.Log outperform it.
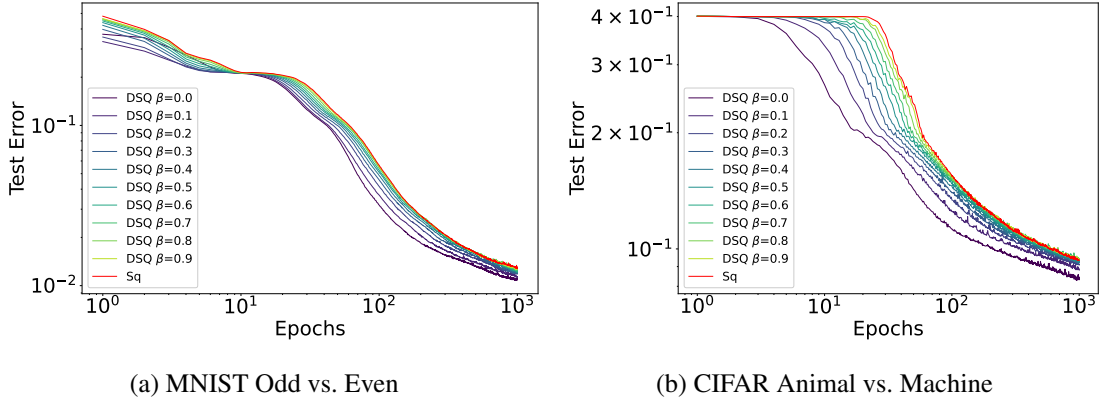
(a) MNIST Odd vs. Even

(b) CIFAR Animal vs. Machine

Figure 3: Training curves for PM.Sq and DebiasedSq with various β for bag size $k = 10$ on the small CNN architecture. We use a fixed learning rate of 0.001 and run full-batch GD for 1000 epochs. Each line is an average over 10 trials with different random seeds.

G.2. Benefits of Debiasing for Optimization

The previous results suggest that the performance of PM.Sq and DebiasedSq is similar, and that there might not be a need for debiasing in practice, as the failure mode in Proposition 2 is arguably pathological. On the contrary, we present evidence that (accurate) debiasing can improve optimization. In Figure 3, we compare the performance of PM.Sq and DebiasedSq with varying β for full-batch GD. (We also investigated large batch SGD, but we found this effect was more striking with larger batches, so we only report the full-batch GD results.) Here, we can compute $\widehat{\mathbb{E}} f_\theta$ in every epoch at no additional cost, so the debiasing term can be exactly computed for every gradient update.

We find that in early stages of training, debiasing results in faster optimization. For CIFAR Animal vs. Machine, GD on PM.Sq seems to get stuck at a bad local minimum near initialization and only escapes after ~ 25 epochs, but GD on DebiasedSq is able to escape the local minimum much faster. Early in training, the model f_θ does not fit the label proportion $\widehat{p}$ well, so the debiasing term $B(f_\theta)$ is large, and therefore the PM.Sq loss will *overestimate* the true instance-level square loss. On the contrary, DebiasedSq uses more accurate estimates of the instance-level square loss, which is why we see benefits of debiasing. Later in training, as $B(f_\theta) \to 0$, the PM.Sq/DebiasedSq loss estimates are more similar, explaining why the different lines seem to converge to similar test errors.

G.3. Implementation Details

Our code can be found on GitHub at `https://github.com/GXLI97/llp_experiments`. All experiments were run on an NVIDIA RTX A6000 GPU using Tensorflow and Keras.

We elaborate on the architectures used in our experiments.

- The linear model has a single dense output layer with 1 unit and sigmoid activation.

- The small two layer NN comprises of a dense layer with 100 units and ReLU activation, followed by a dense output layer with 1 unit and sigmoid activation.

- The large two layer NN comprises of a dense layer with 1000 units and ReLU activation, followed by a dense output layer with 1 unit and sigmoid activation.

- The small CNN is the same architecture in (Busa-Fekete et al., 2023):

    - Convolutional layer with 32 kernels of size 3×3 and ReLU activation.

    - Max pooling layer with pool 2×2.

    - Convolutional layer with 64 kernels of size 3×3 and ReLU activation.

    - Max pooling layer with pool size 2×2.

    - Flatten layer.

    - Dropout layer with drop rate 0.5.

    - Dense output layer with 1 unit and sigmoid activation.

- The large CNN is the same architecture in (Busa-Fekete et al., 2023):

    - Convolutional layer with 32 kernels of size 3×3 and ReLU activation.

    - Convolutional layer with 32 kernels of size 3×3 and ReLU activation.

    - Max pooling layer with pool 2×2.

    - Dropout layer with drop rate 0.25.

    - Convolutional layer with 64 kernels of size 3×3 and ReLU activation.

    - Convolutional layer with 64 kernels of size 3×3 and ReLU activation.

    - Max pooling layer with pool 2×2.

    - Dropout layer with drop rate 0.25.

    - Flatten layer.

    - Dense layer with 512 units and ReLU activation.

    - Dropout layer with drop rate 0.5.

    - Dense output layer with 1 unit and sigmoid activation.

G.4. Results Summary

		EZ.Log	EZ.Sq	PM.Log	PM.Sq	DebiasedSq
$k = 10$	Linear	10.9	10.2	**9.47**	9.51	9.50
	Two Layer Small	5.32	3.93	**1.56**	**1.56**	**1.56**
	Two Layer Large	4.56	3.50	1.28	**1.18**	1.19
	CNN Small	3.89	2.47	1.06	1.08	**1.02**
	CNN Large	3.54	1.58	0.417	**0.416**	0.422
$k = 100$	Linear	14.2	12.9	**11.8**	**11.8**	**11.8**
	Two Layer Small	8.72	8.33	**3.77**	3.84	3.80
	Two Layer Large	9.29	8.50	3.21	**2.80**	**2.80**
	CNN Small	6.89	5.82	2.11	2.12	**2.08**
	CNN Large	5.86	4.63	0.826	**0.774**	0.841
$k = 1000$	Linear	21.3	21.5	**19.9**	**19.9**	**19.9**
	Two Layer Small	20.5	20.5	**20.1**	20.2	20.2
	Two Layer Large	19.6	21.0	**18.2**	18.5	18.4
	CNN Small	16.1	16.3	**14.1**	14.4	14.3
	CNN Large	15.3	15.2	**13.1**	13.7	13.7

Table 1: MNIST Test Error (%) for LLP Algorithms. Best error is reported in **bold.**

		EZ.Log	EZ.Sq	PM.Log	PM.Sq	DebiasedSq
$k = 10$	Linear	19.2	18.7	**18.4**	**18.4**	**18.4**
	Two Layer Small	16.3	15.8	**14.6**	14.7	14.7
	Two Layer Large	18.0	16.5	**13.2**	13.4	13.4
	CNN Small	10.5	10.2	8.30	**8.09**	8.23
	CNN Large	10.5	10.2	8.19	**8.11**	8.18
$k = 100$	Linear	22.1	21.0	20.5	**20.4**	**20.4**
	Two Layer Small	20.1	20.1	**19.3**	**19.3**	**19.3**
	Two Layer Large	20.3	24.5	**19.5**	19.7	19.7
	CNN Small	17.4	17.3	13.0	13.0	**12.9**
	CNN Large	20.4	18.7	**13.2**	13.3	13.4
$k = 1000$	Linear	24.0	45.7	**28.5**	**28.5**	**28.5**
	Two Layer Small	37.2	49.8	36.8	**36.7**	**36.7**
	Two Layer Large	27.1	54.0	**35.1**	37.9	37.4
	CNN Small	33.5	33.6	35.7	35.8	**35.6**
	CNN Large	36.1	36.1	35.9	36.8	**36.6**

Table 2: CIFAR Test Error (%) for LLP Algorithms. Best error is reported in **bold.**