Firebolt: Weak Supervision Under Weaker Assumptions

Zhaobin Kuang[†], Chidubem Arachie[‡], Bangyong Liang[†], Pradyumna Narayana[†], Giulia DeSalvo[†], Michael Quinn[†], Bert Huang[¶], Geoffrey Downs[†], and Yang Yang [†]
[†]Google, [‡]Virginia Tech (work performed at Google), and [¶]Tufts University

Abstract

Modern machine learning demands a large amount of training data. Weak supervision is a promising approach to meet this demand. It aggregates multiple labeling functions (LFs)—noisy, user-provided labeling heuristics—to rapidly and cheaply curate probabilistic labels for large-scale unlabeled data. However, standard assumptions in weak supervision—such as user-specified class balance, similar accuracy of an LF in classifying different classes, and full knowledge of LF dependency at inference time—might be undesirable in practice. In response, we present Firebolt, a new weak supervision framework that seeks to operate under weaker assumptions. In particular, Firebolt learns the class balance and class-specific accuracy of LFs jointly from unlabeled data. It carries out inference in an efficient and interpretable manner. We analyze the parameter estimation error of Firebolt and characterize its impact on downstream model performance. Furthermore, we show that on five publicly available datasets, Firebolt outperforms a state-of-the-art weak supervision method by up to 5.8 points in AUC. We also provide a case study in the production setting of a tech company, where a Firebolt-supervised model outperforms the existing weakly-supervised production model by 1.3 points in AUC and speeds up label model training and inference from one hour to three minutes.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 INTRODUCTION

In recent years, weak supervision has emerged as a promising approach to generate training labels for unlabeled data that feeds the training of data-hungry modern machine learning systems (Ratner et al., 2016, 2017; Arachie and Huang, 2019, 2020; Mazzetto et al., 2021b). In contrast to manual labeling, weak supervision relies on *labeling functions* (LFs)—user-provided labeling sources such as heuristic rules, crowdsourced labels, knowledge bases, among others—that can be combined in a weighted majority vote fashion to quickly and cheaply infer probabilistic training labels at large scale (Austen-Smith and Banks, 1996). As a result, weak supervision has been successfully deployed to power numerous real-world machine learning applications, from production at tech companies such as Google (Bach et al., 2019; Suri et al., 2020) and Apple (Ré et al., 2019), to fighting human trafficking (Ratner et al., 2017), and to identifying individuals with heart malformations (Fries et al., 2019).

Inferring training labels from unlabeled data via weak supervision can be challenging because of the lack of ground truth labels and due to complex behaviors among labeling functions. Without observing any ground truth labels, it is difficult to provide even a simple description such as the class balance of the data, let alone inferring labels for specific data points. While LFs partially label the data, they typically label different number of classes, produce labels of different quality, and may be dependent on or contradictory with each other. Modeling these complex behaviors among LFs is also difficult.

While weak supervision algorithms (Ratner et al., 2019; Fu et al., 2020; Chen et al., 2021) have made significant progress in tackling these aforementioned challenges, we observe that existing solutions may still operate under assumptions that can be difficult to meet in practice. To begin with, a weak supervision algorithm may assume that class balance of the data is user-provided, which can be a challenging burden for the users especially in imbalanced classification problems. Furthermore,

a weak supervision algorithm often assumes that a labeling function has similar accuracy in classifying examples from different classes, which is unlikely to hold given that LFs in practice do not even necessarily label the same number of classes. Finally, weak supervision algorithms also typically assume the full knowledge of dependencies among LFs to infer probabilistic labels, which could lead to exponential time complexity during inference and a lack of interpretability.

To relieve these potentially stringent assumptions, we propose Firebolt, a new weak supervision framework that operates under weaker assumptions. Firstly, Firebolt does not require the full specification of class balance. Upon knowing if the problem is imbalanced or not, Firebolt can jointly learn the class balance and the quality of the labeling functions directly from data. Secondly, Firebolt specifically models LFs that label different numbers of classes and is capable of learning class-specific accuracy for each LF. Finally, at inference time Firebolt does not require a full knowledge of dependencies among LFs. Instead, inference is achieved efficiently by solving a logistic regression problem with polynomial time complexity. It derives the contribution of each LF to the probabilistic label for interpretability.

The major contributions of this paper are summarized as follows:

- We present Firebolt, a new weak supervision algorithm that directly learns the class balance from data, models complex behaviors of labeling functions, and produces probabilistic labels in an efficient, interpretable manner at inference time.

- We analyze the parameter estimation error of Firebolt. Importantly, we characterize how this error can be influenced by dependency misspecification among LFs, and how this error influences downstream model performance.

- Firebolt outperforms existing weak supervision methods in a variety of settings. On five benchmark datasets, Firebolt outperforms a state-of-the-art weak supervision framework (Fu et al., 2020) by up to 5.8 points in AUC; in a real-world production setting of a tech company, the Firebolt-supervised model outperforms the existing weakly-supervised model by 1.3 points in AUC.

In the appendix, we discuss related work and present extended methodological, theoretical, and empirical results.

2 BACKGROUND

We provide an overview of the weak supervision workflow. We then describe different types of labeling functions and their representation. For the ease of exposition and for its high practical relevance, we focus our discussion on binary classification problems (see extension in Appendix B.9).

2.1 Weak Supervision in a Nutshell

Weak supervision quickly and cheaply produces massive amount of training labels for unlabeled data. A typical weak supervision workflow has three steps (Figure 1). First, users create labeling functions that programmatically produce noisy and incomplete labels of the dataset. For each data point, each LF can vote (label) positive (+1), vote negative (-1), or abstain (0) if it does not have enough information. These LFs can come from a variety of sources, such as heuristics rules (Safranchik et al., 2020), crowdsourcing (Karger et al., 2011), upstream classifiers (Bach et al., 2019), knowledge bases (Mintz et al., 2009). Secondly, a label model takes the votes from LFs as input and learns the accuracy and dependency among LFs, without observing the ground truth label. It then uses this learned information to produce probabilistic labels for unlabeled data. Finally, these probabilistic labels are used to train a supervised end model along with the associated data points.

2.2 Labeling Functions (LFs)

A key component of weak supervision is the abstraction provided by labeling functions: although we may have many labeling sources, weak supervision views them as black boxes producing noisy labels. We describe different types of LFs and how to represent them.

Types of LFs There are three common types of labeling functions for a binary classification problem. In reality, practitioners write a mix of these three types of LFs: *Unipolar LFs*: these include positive LFs and negative LFs. Unipolar LFs can take two actions: given a data point, positive (negative) LFs can either vote positive (negative) or abstain; *Bipolar LFs*: in contrast to unipolar LFs, a bipolar LF can take three actions on a data point: votes positive, votes negative, or abstains. Furthermore, when a bipolar LF abstains, it is assumed to provide no class information about the data point in question; *Binary LFs*: a binary LF can either vote positive or negative, it does not abstain.

LF Representation We describe a unified representation of these three types of LFs by mapping unipolar LFs and bipolar LFs to binary LFs. In detail, a bipolar LF can be mapped to a pair of positive LF and a

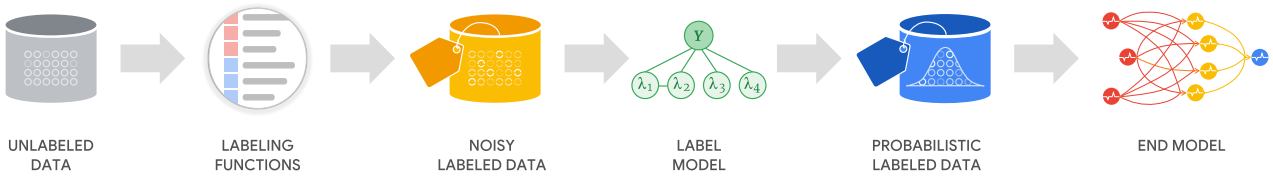


Figure 1: A weak supervision workflow: user-provided labeling functions are aggregated to produce probabilistic labels by a label model. These probabilistic labels are then used to train an end model.

negative LF. For the positive LF in this pair, it votes positive whenever the bipolar LF votes positive and abstains otherwise. For the negative LF in this pair, it votes negative whenever the bipolar LF votes negative and abstains otherwise. Furthermore, a positive LF can be converted to a binary LF by voting positive on the examples it labels and negative on the examples it abstains. Similarly, a negative LF can be converted to a binary LF by changing its abstention votes to positive votes. An important property preserved by this LF representation is the relationship of an LF to *random guessing*:

Definition 1. *Let $y \in \{-1, 1\}$ be the unobserved ground truth label. We say that a positive/bipolar/binary LF $\lambda_\star$ is better than random guessing if* $\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1)$*; and a negative LF λ_- is better than random guessing if* $\mathrm{P}(y = -1 \mid \lambda_- = -1) > \mathrm{P}(y = -1)$.

Given that we can represent both unipolar and bipolar LFs as binary LFs while preserving their relationships with random guessing, we will focus on weak supervision with binary LFs as input without loss of generality. We provide further discussion on random guessing, abstention, and LF representation in Appendix B.1–Appendix B.3.

3 WEAK SUPERVISION WITH FIREBOLT

We present the Firebolt algorithm for weak supervision. We first discuss the label model of Firebolt in Section 3.1. We then present the learning algorithm of the label model in Section 3.2 (Algorithm 1). With the learned label model, we show how Firebolt estimates probabilistic labels for unlabeled data in Section 3.3 (Algorithm 2). Finally, we discuss the implication and extension of Firebolt in Section 3.4.

3.1 Label Model

The label model in Firebolt treats the ground truth label y as a latent variable and uses LF votes to infer its value. It does so by learning an Ising model (Ravikumar et al., 2010) that characterizes the joint distribution between LFs and y. In what follows, we formulate this problem and provide necessary background on Ising models. We then describe a property crucial to Firebolt discovered by Jaffe et al. (2015) among triplets of conditionally independent LFs encoded by the Ising model. Finally, we discuss our assumptions.

Problem Formulation Formally, we consider a binary classification problem where the unobserved ground truth label is $y \in \{-1, 1\}$. Let x be the features associated with y and let there be p binary LFs $\lambda = \left[\lambda_1, \lambda_2, \cdots, \lambda_p\right]^\top$, where $\lambda \in \{-1, 1\}^p$. Let there be a dataset $\mathbb{X} = \left\{x^{(i)}\right\}_{i=1}^n$ with n data points. Weak supervision seeks to learn a label model with some parameters $\hat{\mu}$ for the joint distribution $\mathrm{P}_{\hat{\mu}}(y, \lambda)$. Afterwards, for a given i^{th} data point, the label model infers its probabilistic label as $\mathrm{P}_{\hat{\mu}}(y^{(i)} = 1 \mid \lambda^{(i)})$, where $\lambda^{(i)} = \lambda(x^{(i)})$. Once the probabilistic labels are given across $\mathbb{X}$, we use these probabilistic labels to train an end model along with their associated features.

Ising Models The primary role a label model is to characterize the joint distribution between y and λ. For binary variables, a typical modeling choice is an Ising model. In detail, let $G^* = (V^*, E^*)$ be an undirected graph with a node set $V^* = \{\lambda_1, \lambda_2, \cdots, \lambda_p, y\}$ and an edge set E^* that includes edges between each LF and y as well as some edges between LFs. The joint distribution between y and λ modeled by an Ising model associated with G^* is given as:

$$\mathrm{P}(y, \lambda) = \frac{1}{Z} \exp\left(\theta^*_{00} y + \sum_{j=1}^{p} \theta^*_{jj} \lambda_j + \sum_{j=1}^{p} \theta^*_{0j} \lambda_j y + \sum_{(\lambda_j, \lambda_k) \in E^*} \theta^*_{jk} \lambda_j \lambda_k\right), \tag{1}$$

where θ^*_{00}, θ^*_{jj}'s, θ^*_{0j}'s, and θ^*_{jk}'s can be collectively denoted by a real vector θ^* known as the *canonical parameters*. Specifically, θ^*_{00} influences the class balance of y; θ^*_{jj}'s, also known as the *external fields*, influence the class-specific accuracy of the LFs; θ^*_{0j}'s influence the overall accuracy of the LFs; and finally, θ^*_{jk}'s influence the direct dependency between two LFs as $\theta^*_{jk} \neq 0$ iff $(\lambda_j, \lambda_k) \in E^*$. In particular, for binary LFs representing one or more unipolar/bipolar LFs and their

dependencies, there is an edge in G^* between each pair of these binary LFs.

Associated with the canonical parameters are the *mean parameters*: $\mu^*_{00} = \mathbb{E}[y]$ (class balance); $\mu^*_{+} = \mathbb{E}[\lambda]$, with $(\mu_+)^*_j = \mu^*_{jj} = \mathbb{E}[\lambda_j]$ (prevalence); $\mu^*_{0+} = \mathbb{E}[y\lambda]$, with $(\mu^*_{0+})_j = \mu^*_{0j} = \mathbb{E}[y\lambda_j]$ (accuracy); and μ^*_{++} with $(\mu^*_{++})_{j'} = \mu^*_{jk} = \mathbb{E}[\lambda_j\lambda_k]$ (co-occurrence), where $j' = (j-1)p + k - \frac{j(j+1)}{2}$ and $1 \le j < k \le p$. Collectively, these mean parameters can be denoted by a vector μ^*. Furthermore, to measure class-specific accuracy of λ_j, we define sensitivity $\alpha_j^{+*} = \mathrm{P}(\lambda_j = 1 \mid y = 1)$ and specificity $\alpha_j^{-*} = \mathrm{P}(\lambda_j = -1 \mid y = -1)$. The mean of the two is known as balanced accuracy $\pi^*_j = \frac{1}{2}\left(\alpha_j^{+*} + \alpha_j^{-*}\right) \in [0,1]$, a metric that measures the overall quality of an LF. Indeed, λ_j is better than random guessing iff $\pi^*_j > 0.5$ (Appendix B.1). Finally, we use $\Sigma^* = \mathbb{E}[(\lambda - \mathbb{E}[\lambda])(\lambda - \mathbb{E}[\lambda])^\top]$ to represent the (two-way) covariance matrix among pairs of LFs and we use $T^* = \mathbb{E}[(\lambda - \mathbb{E}[\lambda]) \otimes (\lambda - \mathbb{E}[\lambda]) \otimes (\lambda - \mathbb{E}[\lambda])]$, where $\otimes$ is the tensor product, to represent the (three-way) covariance tensor among triplets of LFs.

Conditional Independence Ising models can encode rich conditional independence information among variables. We say that two LFs λ_j and λ_k in G^* are conditionally independent of each other upon y, i.e. $\lambda_j \perp \lambda_k \mid y$, if $\mathrm{P}(\lambda_j, \lambda_k \mid y) = \mathrm{P}(\lambda_j \mid y)\mathrm{P}(\lambda_k \mid y)$. In G^*, we can read off this conditional independence if there is no path between λ_j and λ_k after removing y and its associated edges. Conditional independence between λ_j and λ_k leads to an interesting relationship among class balance μ^*_{00}, balanced accuracy π^*_j and π^*_k, and the two-way covariance Σ^*_{jk} discovered by Parisi et al. (2014):

$$\Sigma^*_{jk} = (2\pi^*_j - 1)(2\pi^*_k - 1)(1 - \mu^{*2}_{00}). \tag{2}$$

Furthermore, conditional independence among a triplet of LFs $\{\lambda_j, \lambda_k, \lambda_l\}$ (i.e. $\mathrm{P}(\lambda_j, \lambda_k, \lambda_l \mid y) = \mathrm{P}(\lambda_j \mid y)\mathrm{P}(\lambda_k \mid y)\mathrm{P}(\lambda_l \mid y)$) leads to a similar relationship that involves the three-way covariance tensor T^*_{jkl} reported by Jaffe et al. (2015):

$$T^*_{jkl} = -2(2\pi^*_j - 1)(2\pi^*_k - 1)(2\pi^*_l - 1)\mu^*_{00}(1 - \mu^{*2}_{00}). \tag{3}$$

As we shall see in subsequent sections, (2) and (3) will become important for the algorithmic development of Firebolt.

Assumptions We make two assumptions on the data distribution, one on the quality of the LFs and the other on the class balance of the data distribution. We assume that all the LFs are better than random guessing (see Appendix C.2.1 for justification). Additionally, we assume that the users know if they are dealing with an imbalanced classification problem or not (i.e. $\mu^*_{00} = 0$ or not). If the problem is indeed imbalanced, we assume that the users know which class is the minority class and encode it as positive (i.e. $\mu^*_{00} < 0$). We discuss the setting where the class balance is known (e.g. balance classification) in Appendix B.5.

Algorithm 1 Label Model Learning

Input: The sample covariance matrix and tensor $\hat{\Sigma}$ and $\hat{T}$, the dependency graph $\hat{G}$, and the sample prevalence $\hat{\mu}_+$.

1: Form $\hat{M}$ and $\hat{q}$ with $\hat{\Sigma}$, $\hat{T}$, and $\hat{G}$ based on the description before (5) .
2: Solve the least squares problem of (5) for $\hat{t}$ with $\hat{M}$ as the design matrix and $\hat{q}$ as the response.
3: Get the estimated class balance and balanced accuracy parameters $\hat{\mu}_{00}$ and $\hat{\pi}$ using (6) with $\hat{t}$.
4: Get the estimated sensitivity and specificity parameters $\hat{\alpha}^{\pm}$ using (7) with $\hat{\mu}_{00}$ and $\hat{\mu}_+$.
5: Get the estimated accuracy parameters $\hat{\mu}_{0+}$ using (8) with $\hat{\mu}_{00}$ and $\hat{\alpha}^{\pm}$.

Output: The estimated class balance and accuracy mean parameters $\hat{\mu}_{00}$ and $\hat{\mu}_{0+}$.

3.2 Learning the Label Model

We describe parameter learning of the Firebolt label model. From Section 3.1, we can characterize this problem as learning partially observed Ising models with arbitrary external fields and complex dependencies, which is known to be a challenging problem (Goldberg and Jerrum, 2007; Goel, 2020). Indeed, existing methods in this vein make restrictive assumptions on the external fields (Parisi et al., 2014; Kuang et al., 2020; Fu et al., 2020) and dependencies (Jaffe et al., 2015; Bach et al., 2019; Boecking et al., 2020) as mitigation. In contrast, Firebolt tackles this problem in a weak supervision setting without these assumptions. The reward is that Firebolt can directly learn from data the class balance and class-specific accuracy among LFs with complex dependency, which is not possible for many existing weak supervision methods.

The goal of parameter estimation is to infer the class balance of the data $\hat{\mu}_{00}$ and the accuracy of the LFs $\hat{\mu}_{0+}$, without having access to labeled data (Algorithm 1). Firebolt achieves this goal in three steps. First, Firebolt provides joint estimates of the balanced accuracy $\hat{\pi}$ and class balance $\hat{\mu}_{00}$ from the covariance ($\hat{\Sigma}$ and $\hat{T}$) and the user-provided (Bach et al., 2017; Varma et al., 2019; Fu et al., 2020) dependency graph $\hat{G}$. It does so by solving a least squares problem associated with these quantities. Second, Firebolt provides estimated sensitivity and specificity $\hat{\alpha}^{\pm}$ analytically from $\hat{\pi}$, $\hat{\mu}_{00}$,

and the prevalence $\hat{\mu}_+$. Finally, Firebolt derives the estimated accuracy $\hat{\mu}_{0+}$ from $\hat{\alpha}^{\pm}$ and $\hat{\mu}_{00}$. Below we detail each of the three steps.

Learning Balanced Accuracy and Class Balance For an imbalanced classification problem, Firebolt takes $\hat{\Sigma}$, $\hat{T}$, and $\hat{G}$ as input and produces $\hat{\mu}_{00}$ and $\hat{\pi}$ as output (Line 1–Line 3 of Algorithm 1).

Example 1. *Given a conditionally independent triplet $(\lambda_j, \lambda_k, \lambda_l)$, we can use (2) and (3) to infer their balanced accuracy. In detail, let $t_0^* = \log \frac{2|\mu_{00}^*|}{\sigma_{00}^*}$, where σ_{00}^* is the standard deviation of y and hence $\sigma_{00}^{*2} = \Sigma_{00}^* = 1 - \mu_{00}^2$. Let $t_j^* = \log(2\pi_j^* - 1) + \log \sigma_{00}^*$ and define t_k^* and t_l^* similarly to t_j^*. From (2) and (3), we can derive a linear system of equations:*

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} t_0^* \\ t_j^* \\ t_k^* \\ t_l^* \end{bmatrix} = \begin{bmatrix} \log T_{jkl}^* \\ \log \Sigma_{jk}^* \\ \log \Sigma_{kl}^* \\ \log \Sigma_{jl}^* \end{bmatrix}, \tag{4}$$

where we have used $\mu_{00}^* < 0$ for an imbalanced classification problem, and π_j^*, π_k^*, $\pi_l^* > 0.5$, Σ_{jk}^*, Σ_{jl}^*, Σ_{kl}^*, and $T_{jkl}^* > 0$ because the LFs are better than random guessing. Since the binary matrix in (4) is full-rank, we can solve (4) uniquely for the parameters t_0^*, t_j^*, t_k^*, and t_l^* and further recover the balanced accuracy and positive rate from these parameters analytically.

In practice, by enumerating all the conditionally independent pairs and triplets in $\hat{G}$, one can solve a linear system of equations similar to (4) that is associated with the balanced accuracy of all the LFs as well as the class balance of the data: $\hat{M}t = \hat{q}$. Here, $t = \begin{bmatrix} t_0 & t_1 & t_2 & \cdots & t_p \end{bmatrix}^\top$, $\hat{q}$ is a vector that stacks up the logarithm of the positive entries of the two-way and three-way sample covariance associated with each conditionally independent pair and triplet, and $\hat{M}$ is the associated design matrix (Line 1 of Algorithm 1). Note that we use the sample covariance matrix $\hat{\Sigma}$ and tensor $\hat{T}$ since in reality, we do not have access to the ground truth Σ^* and T^*. We also solve the system of linear equations thorough least squares, yielding (Line 2 of Algorithm 1):

$$\hat{t} = \arg\min_t \frac{1}{2} \|\hat{q} - \hat{M}t\|_2^2. \tag{5}$$

Once $\hat{t}$ is estimated, we can estimate the class balance of the data and the balanced accuracy of the LFs (Line 3 of Algorithm 1). In particular, for $j \in \{1, 2, \cdots, p\}$:

$$\hat{\mu}_{00} = \frac{-\exp\left(\hat{t}_0\right)}{\sqrt{4 + \exp\left(2\hat{t}_0\right)}} \quad \hat{\pi}_j = \frac{1}{2}\left(\frac{\exp(\hat{t}_j)}{\sqrt{1 - \hat{\mu}_{00}^2}} + 1\right). \tag{6}$$

Algorithm 2 Inference

Input: estimated class balance and accuracy $\hat{\mu}_{00}$ and $\hat{\mu}_{0+}$ and LF votes $\mathbb{L} = \left\{\lambda^{(i)}\right\}_{i=1}^n$.

1: Solve for the estimated canonical parameters $\hat{\theta}_{0\cdot}$ using (10) with $\hat{\mu}_{00}$, $\hat{\mu}_{0+}$, and $\mathbb{L}$.
2: **for** $i \in [n]$ **do**
3: $\quad \tilde{y}^{(i)} \leftarrow \mathrm{P}(y = 1 \mid \lambda^{(i)}; \hat{\theta}_{00}, \hat{\theta}_{0+})$ using (9). ▷ Produce probabilistic label for each data point.
4: **end for**

Output: Probabilistic labels $\left\{\tilde{y}^{(i)}\right\}_{i=1}^n$.

Learning Sensitivity and Specificity In this step (Line 4 of Algorithm 1), Firebolt takes $\hat{\mu}_{00}$, $\hat{\mu}_+$, and $\hat{\pi}$ as input and produces $\hat{\alpha}^{\pm}$ as output. In particular, given a labeling function λ_j, Firebolt calculates $\hat{\alpha}_j^{\pm}$ analytically from the input as follows:

$$\hat{\alpha}_j^{\pm} = \frac{1}{2}\left(2\hat{\pi}_j \pm \hat{\mu}_{00} \mp 2\hat{\pi}_j\hat{\mu}_{00} \pm \hat{\mu}_{jj}\right). \tag{7}$$

Equation (7) is a direct consequence of the probability equalities: $\mathrm{P}(\lambda_j = 1 \mid y = 1)\mathrm{P}(y = 1) + \mathrm{P}(\lambda_j = 1 \mid y = -1)\mathrm{P}(y = -1) = \mathrm{P}(\lambda_j = 1)$, $\frac{1}{2}\mathrm{P}(\lambda_j = 1 \mid y = 1) + \frac{1}{2}\mathrm{P}(\lambda_j = -1 \mid y = -1) = \pi_j$, and $\mathrm{P}(\lambda_j = 1 \mid y = -1) + \mathrm{P}(\lambda_j = -1 \mid y = -1) = 1$.

Learning Accuracy Once the sensitivity and specificity of the labeling functions are known, calculating the mean accuracy parameter is straightforward (Line 5 of Algorithm 1). Indeed, given $\hat{\alpha}_j^+$, $\hat{\alpha}_j^-$, and $\hat{\mu}_{00}$, $\hat{\mu}_{0j}$ can be calculated as:

$$\hat{\mu}_{0j} = \frac{1}{2}(1+\hat{\mu}_{00})(2\hat{\alpha}_j^+ - 1) + \frac{1}{2}(1-\hat{\mu}_{00})(2\hat{\alpha}_j^- - 1). \tag{8}$$

(8) is a direct consequence of the probability equality: $\mathrm{P}(y = \lambda_j) = \mathrm{P}(\lambda_j = 1 \mid y = 1)\mathrm{P}(y = 1) + \mathrm{P}(\lambda_j = -1 \mid y = -1)\mathrm{P}(y = -1)$.

To sum up, upon the completion of parameter estimation, we have a full knowledge of the quality of all LFs in terms of their estimated sensitivity, specificity, and accuracy. Furthermore, we also have an estimated class balance of the dataset. These estimates will turn out to be useful to infer probabilistic labels for unlabeled data points.

3.3 Inference

In the inference step (Algorithm 2), Firebolt seeks to produce probabilistic labels for unlabeled data points, $\mathrm{P}(y = 1 \mid \lambda)$, using LF votes and the estimated mean parameters the LF votes produced in the label model learning step. Unlike many existing weak supervision algorithms (Ratner et al., 2019; Fu et al., 2020), inference

of Firebolt does not require $\hat{G}$ as an input[1]. Instead, Firebolt computes $\hat{\theta}_{00}$ and $\hat{\theta}_{0+} = \begin{bmatrix}\hat{\theta}_{01} & \hat{\theta}_{02} & \cdots \hat{\theta}_{0p}\end{bmatrix}^\top$ from $\hat{\mu}_{00}$, $\hat{\mu}_{0+}$, and the associated LF votes (Line 1 in Algorithm 2). It then uses $\hat{\theta}_{00}$ and $\hat{\theta}_{0+}$ to produce probabilistic labels as follows (Line 3 in Algorithm 2):

$$\mathrm{P}(y = 1 \mid \lambda; \hat{\theta}_{00}, \hat{\theta}_{0+}) = \text{sigmoid}\left(2\hat{\theta}_{00} + 2\hat{\theta}_{0+}^\top \lambda\right), \quad (9)$$

where $\text{sigmoid}(t) = \frac{1}{1+\exp(-t)}$. We use (9) to produce probabilistic labels because for the Ising model in (1), the conditional probability of y upon λ is given by $\mathrm{P}(y = 1 \mid \lambda; \theta^*_{00}, \theta^*_{0+}) = \text{sigmoid}\left(2\theta^*_{00} + 2\theta^{*\top}_{0+}\lambda\right)$.

To compute $\hat{\theta}_{00}$ and $\hat{\theta}_{0+}$ from $\hat{\mu}_{00}$ and $\hat{\mu}_{0+}$, Firebolt solves the following logistic regression problem:

$$\hat{\theta}_{00}, \hat{\theta}_{0+} = \arg\min_{\theta_{00}, \theta_{0+}} -\theta_{00}\hat{\mu}_{00} - \theta_{0+}^\top \hat{\mu}_{0+} + \frac{1}{n}\sum_{i=1}^{n} \log\left[\exp(\theta_{00} + \theta_{0+}^\top \lambda^{(i)}) + \exp(-\theta_{00} - \theta_{0+}^\top \lambda^{(i)})\right]. \quad (10)$$

There are several benefits solving (10). First, (10) can be solved even though we do not directly observe y, nor do we need direct access to $\hat{G}$. This is because (10) only takes $\hat{\mu}_{00}$, $\hat{\mu}_{0+}$, and $\lambda^{(i)}$'s as input, and these quantities can either be directly observed or have been learned during parameter estimation. Second, (10) can be solved efficiently, as it is a logistic regression problem that only involves n data points and $p + 1$ parameters. In addition, since $\hat{\theta}_{0+}$ are the regression coefficients of logistic regression, they can be interpreted as the contribution of each individual LF to the prediction associated with the probabilistic labels, providing a way to better understand the importance of each LF. In Appendix B.7, we detail the derivation of the inference procedure. Furthermore, we showcase how to carry out efficient exact inference in closed form with graph structures commonly arise in weak supervision.

3.4 Implication and Extension

Other than binary classification, Firebolt has implications and extensions on a variety of topics of practical interests. These include understanding behaviors of different types of LFs (Appendix B.1-Appendix B.3), how to write LFs of good quality (Appendix B.4), alternative formulations of Firebolt(Appendix B.5 and Appendix B.6), failure mode of naive weak supervision methods for imbalanced classification (Appendix B.8), positive-only classification, and multi-class classification (Appendix B.9), among others.

[1]Note that similar to many existing weak supervision algorithms, $\hat{G}$ is still required by Firebolt in the label model learning step.

4 THEORY

In this section, we present theoretical analysis of the parameter estimation error of the Firebolt label model. Next, we further analyze how downstream generalization performance of weak supervision based on a noise-aware loss might be impacted by the error in label model parameter estimation.

Label Model Parameter Estimation Error We present a simplified version of our theorem to characterize label model parameter estimation error, under the assumptions made in Section 3.1 and assuming also $\hat{G} = G^*$. In Appendix C.2, we provide a careful analysis on how misspecified dependencies in $\hat{G}$ may influence the parameter estimation error.

Theorem 1. *Under the assumptions made in Section 3.1, the expected mean parameter estimation error of Firebolt learned from n unlabeled data points, p labeling functions, and G^* for an imbalanced classification problem can be upper bounded by:*

$$\mathbb{E}[\|\hat{\mu} - \mu^*\|_2] = \mathcal{O}\left(\frac{1}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*) + 1)\frac{p^5}{\sqrt{n}}\right),$$

where $n > n_0$ for some n_0 such that all the entries in $sign(\hat{\Sigma}) = sign(\Sigma^*)$ and $sign(\hat{T}) = sign(T^*)$, $\omega_{\min} > 0$ is a lower bound on the smallest positive entries of Σ^*, T^*, and Σ^*_{00}, and $\sigma_{\min}^{-1}(M^*)$ is the reciprocal of the smallest singular value of M^*.

In Theorem 1, we note that the error scales as $\mathcal{O}(1/\sqrt{n})$, matching the scaling in supervised learning. Furthermore, we note that the scaling with the number of LFs p is somewhat pessimistic. Indeed, our analysis suggests that the error is governed by the concentration of the tensor $\hat{T}$ to T^*, whose optimal rate is still an open statistical problem (Vershynin, 2020; Even and Massoulié, 2021). The scaling of $\mathcal{O}(p^5/\sqrt{n})$ is also due to jointly learning the class balance and the LF parameters from data. In Appendix C.3, we show that with known class balance, the error can scale as $\mathcal{O}(p^3/\sqrt{n})$. Finally, low correlations among LFs lead to a small $\omega_{\min}$ while too many direct dependency among LFs may lead to a high $\sigma_{\min}^{-1}(M^*)$, both increase the difficulty of parameter estimation.

End Model Generalization Error We analyze the generalization error of the end model trained with the probabilistic labels produced by the Firebolt label model. Formally, let $y = f_{w^*}(x)$ be an end model parametrized by w^* that we seek to learn from the dataset $\left\{(\mathrm{P}_{\hat{\theta}_{0\cdot}}(y \mid \lambda(x^{(i)})), x^{(i)})\right\}_{i=1}^n$, where $x^{(i)}$'s are drawn from the distribution $\mathcal{D}$, and $\hat{\theta}_{00}$ and $\hat{\theta}_{0+}$, collectively denoted as $\hat{\theta}_{0\cdot}$, are produced

by Algorithm 2. Let $l(y, x; w) \in [0, 1]$ be a loss function. We consider learning the end model $\hat{w} = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tilde{y}\sim \mathrm{P}_{\hat{\theta}_0.}(y|\lambda(x^{(i)}))}[l(\tilde{y}, x; w)]$ in a noise-aware fashion that samples labels for unlabeled data points based on $\mathrm{P}_{\hat{\theta}_0.}(y \mid \lambda(x^{(i)}))$ for end model training. We define the expected loss for a given w under $\mathcal{D}$ as $L(w) = \mathbb{E}_{\mathcal{D}}[l(x, y; w)]$. The generalization error of the end model $L(\hat{w}) - L(w^*)$ can be characterized by Theorem 2 as follows:

Theorem 2. *Under the assumptions made in Section 3.1, the generalization error of the end model learned from the probabilistic labels produced by Firebolt in a noise-aware fashion can be upper bounded by*

$$L(\hat{w}) - L(w^*) \leq \xi(n) + c\|\hat{\mu} - \mu^*\|_2 + \psi(\mathcal{D}, \mu^*),$$

where c is a constant related to the boundedness of the canonical parameters, $\psi(\mathcal{D}, \mu^*) = 2\sqrt{2 \cdot KL(\mathrm{P}_{\mathcal{D}}(y \mid x) \parallel \mathrm{P}_{\mu^*}(y \mid x))}$ is the divergence between $\mathrm{P}_{\mathcal{D}}(y \mid x)$ and $\mathrm{P}_{\mu^*}(y \mid x)$, $\xi(n)$ is a decreasing function of the sample size associated with empirical risk minimization.

Theorem 2 highlights two contentious factors influencing the generalization performance of the end model: parameter estimation error $\|\hat{\mu} - \mu^*\|_2$ and the divergence $\psi(\mathcal{D}, \mu^*)$ between the label and end model data generation process. Choosing a more complicated label model may decrease the divergence but increase the parameter estimation error under a given sample size. The tension between these two factors suggests the importance of choosing an appropriate label model to strike a good balance in practice.

5 EXPERIMENTS

We validate the practical utility of Firebolt on various settings including balanced/imbalanced classification, learning from unipolar/bipolar labeling functions, zero-shot multi-class learning and learning from complex dependencies. We evaluate the performance of Firebolt on publicly-available benchmark datasets, real-world production datasets, and synthetic data. We compare Firebolt with a variety of alternative methods, showing that Firebolt can achieve the same or better performance than state-of-the-art weak supervision frameworks and weak supervision models in production. In Appendix D, we report details of the experiment setup and we also report extended experimental results on synthetic data as well as experiments on Firebolt label model.

5.1 Evaluation on Benchmark Datasets

We demonstrate the performance of Firebolt using the labels it generates to train neural networks as end models. The end model performance is evaluated on a held out test set, and we report the results on four benchmark datasets.

Metric To measure performance, we use the area under curve of the receiver operating characteristic (AUC) and the average precision (AP). AUC and AP avoid the need for additional hyperparameter tuning due to binarizing probabilistic labels. AP is also particularly suitable for imbalanced classification problems (Davis and Goadrich, 2006).

Datasets We consider four publicly-available benchmark datasets. Three (`spam`, `crowdsourcing`, `spouse`) are benchmark weak supervision datasets and one is a text classification dataset (`IMDb`). In particular, `spouse` is an imbalanced dataset and the other three are balanced. Both `spouse` and `crowdsourcing` also include the use of bipolar LFs. There are up to $50,000$ samples in these datasets. Further details of the datasets are available in Appendix D.

Methods We compare Firebolt with four representative alternatives. (1) Majority vote: a baseline method; (2) Flyingsquid (Fu et al., 2020): a state-of-the-art weak supervision method; (3) a weak supervision pipeline in production (Bach et al., 2019; Suri et al., 2020); (4) Constrained Label Learning (CLL) (Arachie and Huang, 2020): a new, alternative constraint-based weak supervision method.

Protocol For each dataset, we split it into an unlabeled training set and a labeled test set. We use the LFs to produce votes on the training set and train label models on the training set. We then use the resulting probabilistic labels to train an end model. We evaluate the performance of the end model on the test set and report results over 5 trials. Some of the datasets also include validation sets, which are used in the end model training.

Results Table 1 shows the performance of Firebolt and alternative methods on four classification tasks. Firebolt either produces the best results or is in a (statistical) tie to be the best approach among all datasets.

On the `spam` classification task, where the dataset is relatively balanced and the labeling functions are independent, we see that majority voting is a strong baseline outperforming CLL and Flyingsquid. Firebolt is able to aggregate the labeling functions effectively, producing the best performance on the dataset (tied with the production model). On the imbalanced `spouse` classification task, Firebolt exceeds the next best performing method by about 5.8 point gain in AUC. Since the dataset is imbalanced, we also report AP for different

Dataset	Metric	Majority Vote	Production	CLL	Flyingsquid	Firebolt
`spam`	AUC	0.955±0.004	**0.975±0.001**	0.875±0.004	0.946±0.002	**0.975±0.001**
`crowdsourcing`	AUC	**0.768±0.010**	**0.756±0.011**	**0.779±0.003**	**0.740±0.012**	**0.771±0.012**
`spouse`	AUC	0.685±0.020	0.656±0.019	0.712±0.018	0.713±0.009	**0.771±0.002**
`spouse`	AP	0.179±0.034	0.266±0.012	0.215±0.033	0.231±0.019	**0.374±0.007**
`IMDb`	AUC	0.797±0.001	0.630±0.002	0.820±0.002	0.797±0.012	**0.832±0.001**

Table 1: Test performance of weakly supervised end models on four benchmark datasets over five trials (mean±s.d.).

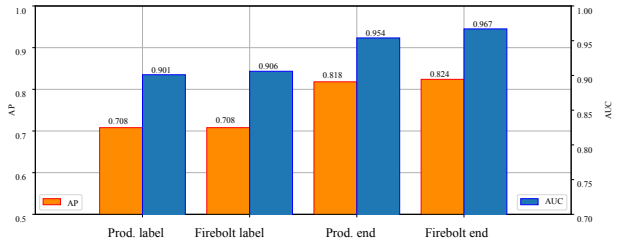


Figure 2: APs (orange bars) and AUCs (blue bars) of a Firebolt label and end model vs an existing weakly-supervised production pipeline with a label model and an end model.

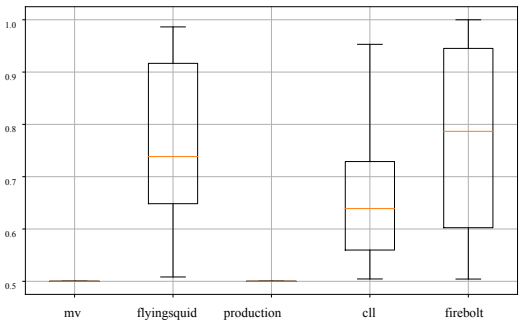
methods in Table 1. We observe that AP of different methods are better than random; the test positive rate on the dataset is 8%. Firebolt achieves a score that is 10.8 points greater than the second best performing method. For all tasks, Firebolt consistently outperforms Flyingsquid, a state-of-the-art weak supervision approach. All these results suggest the empirical advantage of Firebolt compared to alternative methods.

5.2 A Case Study in Industry Production

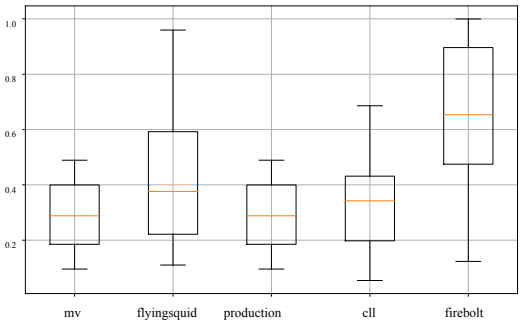
We further demonstrate the practical utility of Firebolt with a case study in a tech company. Here, we are faced with an imbalanced classification problem where we have access to 24 LFs. The goal is to train end models through weak supervision and evaluate their performance on a test set (1% positive rate). We use a linear model as the end model. The dataset has hundreds of millions of examples. Unlike benchmark experiments in Section 5.1, we only compare the performance of the end model trained by Firebolt against an existing weak supervision pipeline in production. We cannot compare to other methods, because they are not implemented in production. Furthermore, in this context, we also need to process larger datasets than those in Section 5.1. So, efficiency is also a key concern.

The comparison between the two approaches is given in Figure 2. Both the label model and the end model trained by Firebolt and the alternative deliver reasonable performance, with 0.9+ AUC and 0.7+ AP. Both end models generalize beyond the label models. Furthermore, the Firebolt label model has a slightly better AUC compared to its counterpart and tie in AP. However, Firebolt end model outperforms its counterpart by 1.3 points in AUC and 0.6 points in AP.

Finally, while the production label model takes an hour for training and inference, we observe that Firebolt speeds this up to only three minutes. These results demonstrate the real-world empirical gain of adopting Firebolt both in efficacy and in efficiency.



(a) AUC



(b) Average Precision

Figure 3: Performance of Firebolt and competing baselines on 45 image classification tasks for zero-shot learning.

5.3 Zero-Shot Multi-Class Learning with Firebolt

To demonstrate the utility of Firebolt beyond binary classifcation, we describe the use of Firebolt for a zero-shot learning problem.

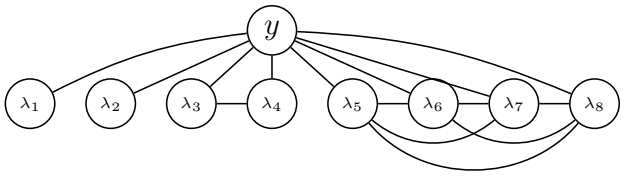


Figure 4: A label model with complex dependency: there are seven pairwise dependencies between eight LFs.

Protocol We use the Animals with Attributes 2 dataset (`AwA2`, Xian et al. 2018) to setup a zero-shot learning classification task. The dataset consists of 37,322 images of 50 animals classes that are split into 40 seen and ten unseen classes. We are interested in classifying objects from the ten unseen classes. While the problem can be viewed as a ten-class classification problem, we alternatively follow the same procedure in Mazzetto et al. (2021b) to perform binary classification on each of the pairs of unseen classes to create 45 binary classifcation problems. We used a classifier learned on the seen classes to generate weak supervision and make predictions on the unseen classes. We compare the performance of Firebolt on all 45 experiments to the other alternatives that are previously described using the same metrics.

Results Figure 3 shows the performance of the methods on the experiments. From the plots, we see that Firebolt outperforms other alternatives on both median AUC and average precision. On AUC described in Figure 3a, Firebolt obtains the best median score outperforming majority vote and production weak supervision method by over 20 percentage points. For AP described in Figure 3b, the median AP score obtained by Firebolt for all experiments is better than the score of the upper quartile of the next best performing method Flyingsquid. Additionally, the lower quartile score of Firebolt outperforms the upper quartile score of all other methods with exception of Flyingsquid. Considering that the weak supervision was generated on the seen classes, Firebolt's performance on the unseen classes demonstrates it's applicability for zero-shot learning tasks. While Firebolt improves upon existing methods in terms of the median scores, a sobering aspect of the result is that the minimum AUC and AP for each method are relatively close to each other and low, as indicated by the lower whiskers of the boxplots. This suggests the intrinsic difficulty of the zero-shot learning task in question. Further investigation may be desirable to understand if it is possible to improve the worst-case performance of weak supervision on this task.

Metric	MV	Prod	CLL	Flyingsquid	Firebolt
AUC	0.677	0.577	0.671	0.710	**0.717**
AP	0.245	0.209	0.242	0.345	**0.353**

Table 2: Performance of various methods on a dataset of 80,000 samples drawn from the label model in Figure 4.

5.4 Handling Complex Dependencies

We demonstrate the capacity of Firebolt to learn from complex dependencies among labeling functions. To this end, we use synthetic data, where we know the ground truth of the dependency graph among LFs. We sample $80,000$ data points from a label model (Figure 4) representing an imbalanced classification problem (positive rate 16.5%) with two unipolar LFs (λ_1 and λ_2) and three bipolar LFs (λ_3–λ_8), where the second (λ_5 and λ_6) and the third (λ_7 and λ_8) bipolar LFs are dependent on each other and the rest of the LFs are all independent of each other. All the labeling functions (λ_1–λ_8) are mildly predictive of y, with balanced accuracy between 0.51 and 0.6. Because we have access to the ground truth distribution, we compare the performance of Firebolt with other alternatives on the population-level distribution to avoid sampling error at test time. The results are given in Table 2, where Firebolt achieves an AUC of 0.717 and an AP of 0.353, outperforming the second best method Flyingsquid by 0.7 point in AUC and 0.8 point in AP. Furthermore, Firebolt also learns a positive rate of 13.0% from the training set, which coincides closely with the ground truth positive rate. These results suggest the practical utility of Firebolt in handling complex dependencies among noisy labeling functions for imbalanced classification problems.

6 CONCLUSION

We proposed Firebolt, a new weak supervision framework that operates under weaker assumptions than alternatives. In particular, Firebolt learns class balance directly from data, estimates class-specific accuracy and handles complex dependencies among LFs, and carries out inference efficiently and in an interpretable manner. We theoretically analyze the performance achieved by Firebolt and understand its empirical utility on various settings from publicly-available benchmark datasets, to real-world production data, and to synthetic data.

Acknowledgements

The authors would like to thank the anonymous reviewers and meta-reviewers of this paper, who provide valuable review and suggestions. The authors would also like to gratefully acknowledge Sugato Basu, Sinong Geng, Abhishek Gupta, Krista Holden, Shilpa Jain, Mathew Keegan, Marija Mikic, Brian Milch, Umut Oztok, Erik Racho, and Roman Zulauf for their feedback, help, and support during the development of this work.

References

Alberto, T. C., Lochter, J. V., and Almeida, T. A. (2015). Tubespam: Comment spam filtering on youtube. In *2015 IEEE 14th International conference on machine learning and applications (ICMLA)*, pages 138–143. IEEE.

Arachie, C. and Huang, B. (2019). Adversarial label learning. In *Proc. of the AAAI Conf. on Artif. Intelligence*, pages 3183–3190.

Arachie, C. and Huang, B. (2020). Constrained labeling for weakly supervised learning. *arXiv preprint arXiv:2009.07360*.

Arachie, C. and Huang, B. (2021). A general framework for adversarial label learning. *Journal of Machine Learning Research*, 22(118):1–33.

Austen-Smith, D. and Banks, J. S. (1996). Information aggregation, rationality, and the condorcet jury theorem. *American political science review*, 90(1):34–45.

Bach, S. H., He, B., Ratner, A., and Ré, C. (2017). Learning the structure of generative models without labeled data. In *International Conference on Machine Learning*, pages 273–282. PMLR.

Bach, S. H., Rodriguez, D., Liu, Y., Luo, C., Shao, H., Xia, C., Sen, S., Ratner, A., Hancock, B., and Alborzi, H. (2019). Snorkel DryBell: A case study in deploying weak supervision at industrial scale. In *Intl. Conf. on Manag. of Data*, pages 362–375.

Balsubramani, A. and Freund, Y. (2015a). Optimally combining classifiers using unlabeled data. *arXiv preprint arXiv:1503.01811*.

Balsubramani, A. and Freund, Y. (2015b). Scalable semi-supervised aggregation of classifiers. In *Advances in Neural Information Processing Systems*, pages 1351–1359.

Boecking, B., Neiswanger, W., Xing, E., and Dubrawski, A. (2020). Interactive weak supervision: Learning useful heuristics for data labeling. *arXiv preprint arXiv:2012.06046*.

Chen, M., Cohen-Wang, B., Mussmann, S., Sala, F., and Ré, C. (2021). Comparing the value of labeled and unlabeled data in method-of-moments latent variable estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 3286–3294. PMLR.

Corney, D. P., Albakour, D., Martinez-Alvarez, M., and Moussa, S. (2016). What do a million news articles look like? In *NewsIR@ ECIR*, pages 42–47.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. (2006). Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136.

Even, M. and Massoulié, L. (2021). Concentration of non-isotropic random tensors with applications to learning and empirical risk minimization. *arXiv preprint arXiv:2102.04259*.

Fries, J. A., Varma, P., Chen, V. S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M., et al. (2019). Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. *Nature communications*, 10(1):1–10.

Fu, D., Chen, M., Sala, F., Hooper, S., Fatahalian, K., and Ré, C. (2020). Fast and three-rious: Speeding up weak supervision with triplet methods. In *International Conference on Machine Learning*, pages 3280–3291. PMLR.

Goel, S. (2020). Learning ising and potts models with latent variables. In *International Conference on Artificial Intelligence and Statistics*, pages 3557–3566. PMLR.

Goldberg, L. A. and Jerrum, M. (2007). The complexity of ferromagnetic ising with local fields. *Combinatorics, Probability and Computing*, 16(1):43–61.

Honorio, J. (2012). Lipschitz parametrization of probabilistic graphical models. *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 347–354.

Hooper, S., Wornow, M., Seah, Y. H., Kellman, P., Xue, H., Sala, F., Langlotz, C., and Re, C. (2020). Cut out the annotator, keep the cutout: better segmentation with weak supervision. In *International Conference on Learning Representations*.

Jaffe, A., Nadler, B., and Kluger, Y. (2015). Estimating the accuracies of multiple classifiers without labeled data. In *Artificial Intelligence and Statistics*, pages 407–415. PMLR.

Karger, D. R., Oh, S., and Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. Neural Information Processing Systems.

Kuang, Z., Sala, F., Sohoni, N., Wu, S., Córdova-Palomera, A., Dunnmon, J., Priest, J., and Ré, C. (2020). Ivy: Instrumental variable synthesis for causal inference. In *International Conference on Artificial Intelligence and Statistics*, pages 398–410. PMLR.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

Mazzetto, A., Cousins, C., Sam, D., Bach, S. H., and Upfal, E. (2021a). Adversarial multiclass learning under weak supervision with performance guarantees. In *International Conference on Machine Learning (ICML)*.

Mazzetto, A., Sam, D., Park, A., Upfal, E., and Bach, S. (2021b). Semi-supervised aggregation of dependent weak supervision sources with performance guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 3196–3204. PMLR.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557.

Parisi, F., Strino, F., Nadler, B., and Kluger, Y. (2014). Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258.

Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., and Ré, C. (2019). Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4763–4771.

Ratner, A. J., Bach, S. H., Ehrenberg, H. R., and Ré, C. (2017). Snorkel: Fast training set generation for information extraction. In *Proceedings of the 2017 ACM Intl. Conf. on Management of Data*, pages 1683–1686. ACM.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. (2016). Data programming: Creating large training sets, quickly. In *Advances in Neural Info. Proc. Sys.*, pages 3567–3575.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Ré, C., Niu, F., Gudipati, P., and Srisuwananukorn, C. (2019). Overton: A data system for monitoring and improving machine-learned products. *arXiv preprint arXiv:1909.05372*.

Safranchik, E., Luo, S., and Bach, S. (2020). Weakly supervised sequence tagging from noisy rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5570–5578.

Sala, F., Varma, P., Fries, J., Fu, D. Y., Sagawa, S., Khattar, S., Ramamoorthy, A., Xiao, K., Fatahalian, K., Priest, J., et al. (2019). Multi-resolution weak supervision for sequential data. *arXiv preprint arXiv:1910.09505*.

Suri, S., Chanda, R., Bulut, N., Narayana, P., Zeng, Y., Bailis, P., Basu, S., Narlikar, G., Ré, C., and Sethi, A. (2020). Leveraging organizational resources to adapt models to new data modalities. *arXiv preprint arXiv:2008.09983*.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

Varma, P., Sala, F., He, A., Ratner, A., and Ré, C. (2019). Learning dependency structures for weak supervision models. In *International Conference on Machine Learning*, pages 6418–6427. PMLR.

Vershynin, R. (2020). Concentration inequalities for random tensors. *Bernoulli*, 26(4):3139–3162.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265.

Zhang, J., Yu, Y., Li, Y., Wang, Y., Yang, Y., Yang, M., and Ratner, A. (2021). Wrench: A comprehensive benchmark for weak supervision. *arXiv preprint arXiv:2109.11377*.

Appendix

The Appendix is organized as follows. In Appendix A, we present related work of Firebolt. In Appendix B, we provide more methodological details of Firebolt and its implication and extension. In Appendix C, we present proofs of our theoretical results. In Appendix D, we provide further experiment details as well as additional experiments. Finally, in Appendix E, we discuss limitation and societal impact of our work.

A Related Work

In terms of problem setting, Firebolt is concerned with the problem of programmatic weak supervision (a.k.a. Snorkel, Ratner et al. 2016, 2017; Bach et al. 2019; Ratner et al. 2019; Fu et al. 2020; Chen et al. 2021; Zhang et al. 2021). Related to the Snorkel-style weak supervision is the line of work on constraint-based weak supervision (Balsubramani and Freund, 2015a,b; Arachie and Huang, 2019, 2021, 2020; Mazzetto et al., 2021a,b). In terms of algorithmic framework, Firebolt draws inspiration from building aggregated classifiers without labeled data (Parisi et al., 2014; Jaffe et al., 2015), a line of research that predates Snorkel. Below we compare Firebolt with these three lines of research.

Comparison to Snorkel Style Weak Supervision Methods Snorkel-style programmatic weak supervision introduces machine learning systems and techniques that use multiple labeling functions to curate training labels for unlabeled data by learning a graphical model with the ground truth label as a latent variable. Weak supervision techniques emphasize the dependencies among labeling functions (Bach et al., 2017; Varma et al., 2019) as well as the ability that a labeling function can abstain from making a classification decision. These features make weak supervision suitable for practitioners to express their domain knowledge in a flexible and programmatic way to facilitate training data labeling.

However, most of these weak supervision approaches (Ratner et al., 2016, 2017; Bach et al., 2019; Fu et al., 2020; Chen et al., 2021) assume that class balance is user-provided. Furthermore, they may also assume that the accuracy of a labeling function in classifying different classes is more or less the same due to the fact that they do not handle arbitrary external fields. An exception in this line of work is Metal (Ratner et al., 2019), where it made attempt to learn the class balance directly from data. Furthermore, it also attempted to learn sensitivity and specificity of the labeling functions from data. Nonetheless, this approach is based on solving non-convex optimization problems through stochastic gradient descent, which could be slow and may not converge to a global optimal solution. Indeed, Metal is outperformed by Flyingsquid (Fu et al., 2020) on various tasks, a state-of-the-art weak supervision algorithm. Flyingsquid assumes that the class balance is user-provided. It can learn class-specific accuracy of labeling functions. Another advantage of Flyingsquid is its exact inference algorithm that produces probabilistic quantities needed to compute probabilistic labels by solving a system of linear equations. However, the inference algorithm may make assumptions on the external field of the graphical model. The system of linear equations needs to be solved for inference also scales exponentially with the clique size of the graphical model.

In contrast, Firebolt mitigates these aforementioned limitations by directly learning class balance and class-specific accuracy of labeling functions for binary imbalanced weak supervision problems. It also carries out inference in polynomial time complexity by solving a logistic regression problem that scales with the number of unique vote combinations of labeling functions and the number of labeling functions. Furthermore, for weak supervision algorithms (Ratner et al., 2019; Fu et al., 2020) that carry out inference using mean parameters, it may be difficult to interpret the individual contribution of each labeling function towards the probabilistic label under complex dependency. On the other hand, Firebolt carries out inference using canonical parameters through (9) in a logistic regression fashion, it therefore can provide interpretability for the individual contribution of each labeling function towards the probabilistic label.

Comparison to Constraint-Based Weak Supervision A related line of research—constraint-based methods—uses a different approach for solving weak supervision (Balsubramani and Freund, 2015a,b; Arachie and Huang, 2019, 2021, 2020; Mazzetto et al., 2021a,b). These methods use the labeling functions and user provided error rates or error bounds to constrain the possible space of the data labeling. They then solve an optimization problem to estimate probabilistic labels for the data. These methods have become popular in recent times since they do not make assumptions about the joint distribution of the true labels and the labeling functions. Firebolt

is similar to constraint-based approaches in that Firebolt estimates probabilistic labels for the training data. However, unlike constraint-based methods, Firebolt does not require user-defined estimates of the error rates of the labeling functions. It can be difficult for users to provide error estimates of the labeling functions in practice without having access to labeled data. Additionally, the Firebolt label models are more interpretable as it is able to accurately determine the contribution of each labeling function to the decisions made by the label models.

Comparison to Parisi et al. 2014; Jaffe et al. 2015 Firebolt is also closely related to a line of research that aggregates classifier output without labeled data. In particular, Parisi et al. (2014) discovers Equation (2), which is used in the Firebolt formulation of parameter learning of weak supervision. However, Parisi et al. (2014) did not estimate class balance, nor did it estimate class-specific accuracy. Subsequently, Jaffe et al. (2015) generalizes Parisi et al. (2014) to propose Equation (3), which is also used by Firebolt. It proposed a procedure to estimate both class balance and class-specific accuracy of classifiers.

Firebolt is distinctive from these two works in that Firebolt focuses on the weak supervision setting that does not require labeled data. In contrast, these two works aim at building ensemble classifiers from base classifiers, which require labeled data in the first place. Furthermore, labeling functions in weak supervision can abstain, which is not the case for these two works. Weak supervision algorithms also focus on addressing dependency between labeling functions, while these two works primarily focus on the conditional independent situation. While the Firebolt algorithm draws inspiration from these two works, it generalizes beyond them by providing fine-grained non-asymptotic theoretical guarantees for the parameter estimation error of the algorithm as well as its impact on downstream end model training performance, accounting for model misspecification. In contrast, these two works provided asymptotic guarantees for their algorithms.

B Extended Methodological Results

In this section, we present various methodological implications and extensions of Firebolt. First, we discuss topics related to labeling functions such as random guessing (Appendix B.1), abstention (Appendix B.2), LF representation(Appendix B.3), and how to write labeling functions of good quality (Appendix B.4). Next, we discuss topics related to the learning of label model such as learning the label model when the class balance is known (Appendix B.5), an alternative formulation of the Firebolt learning algorithm similar to that of Fu et al. (2020) (Appendix B.6), and details on the inference (Appendix B.7) algorithms. Finally, we discuss the implication of Firebolt on imbalanced classification problems (Appendix B.8) and its extension beyond binary classification (Appendix B.9).

B.1 Random Guessing

In this section, we discuss the relationship between a binary LF $\lambda_\star$ and random guessing. Furthermore, we characterize the precision and recall of $\lambda_\star$ using its relationship with random guessing. Finally, we will characterize if a labeling function is better than random guessing thorough balanced accuracy.

To begin with, in Definition 1, we consider $\lambda_\star$ to be better than random guessing if

$$\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1). \tag{11}$$

However, if (11) is not satisfied, we have that $\mathrm{P}(y = 1 \mid \lambda_\star = 1) \leq \mathrm{P}(y = 1)$. In this case, on the one hand, if $\mathrm{P}(y = 1 \mid \lambda_\star = 1) = \mathrm{P}(y = 1)$, it means $\lambda_\star$ is independent of y. That is to say, $\lambda_\star$ is equivalent to random guessing when it tries to infer the value of y. On the other hand, $\mathrm{P}(y = 1 \mid \lambda_\star = 1) < \mathrm{P}(y = 1)$ means that $\lambda_\star$ is worse than random guessing.

We are particularly interested in the precision $\mathrm{P}(y = 1 \mid \lambda_\star = 1)$ and recall $\mathrm{P}(\lambda_\star = 1 \mid y = 1)$ of $\lambda_\star$. To this end, we first show the following relationship between the precision and recall of a λ_* that is better than random guessing:

$$\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1) \Leftrightarrow \mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_* = 1). \tag{12}$$

This is because by Bayes theorem,

$$\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1) \Leftrightarrow \frac{\mathrm{P}(\lambda_\star = 1 \mid y = 1)\mathrm{P}(y = 1)}{\mathrm{P}(\lambda_\star = 1)} > \mathrm{P}(y = 1)$$

$$\Leftrightarrow \mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_* = 1).$$

Next, we show the following relationship related to the precision of $\lambda_\star$:

$$\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1) \Leftrightarrow \mathrm{P}(y = -1 \mid \lambda_\star = -1) > \mathrm{P}(y = -1). \tag{13}$$

This is because from (12),

$$\begin{aligned}
\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1) &\Leftrightarrow \mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_* = 1) \\
&\Leftrightarrow \mathrm{P}(\lambda_\star = -1 \mid y = 1) < \mathrm{P}(\lambda_* = -1) \\
&\Leftrightarrow \frac{\mathrm{P}(\lambda_\star = -1 \mid y = 1)\mathrm{P}(y = 1)}{\mathrm{P}(\lambda_* = -1)} < \mathrm{P}(y = 1) \\
&\Leftrightarrow \mathrm{P}(y = 1 \mid \lambda_\star = -1) < \mathrm{P}(y = 1) \\
&\Leftrightarrow (y = -1 \mid \lambda_\star = -1) > \mathrm{P}(y = -1).
\end{aligned}$$

Furthermore, we show the following relationship related to the recall of $\lambda_\star$:

$$\mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_* = 1) \Leftrightarrow \mathrm{P}(\lambda_\star = -1 \mid y = -1) > \mathrm{P}(\lambda_* = -1). \tag{14}$$

This is true because from (12) and (13),

$$\begin{aligned}
\mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_* = 1) &\Leftrightarrow \mathrm{P}(y = -1 \mid \lambda_\star = -1) > \mathrm{P}(y = -1) \\
&\Leftrightarrow \mathrm{P}(\lambda_\star = -1 \mid y = -1) > \mathrm{P}(\lambda_\star = -1),
\end{aligned}$$

where we can show the last equivalence to be true using (12) and a symmetric argument.

Note that for labeling functions that are equivalent to or worse than random guessing, we can also derive relationships similar to (12), (13), and (14) to characterize their precision and recall.

Finally, we show that the relationship between $\lambda_\star$ and random guessing can be characterized by balanced accuracy.

Proposition 1. *A binary labeling function $\lambda_\star$ is better than random guessing if and only if it has a balanced accuracy $\pi_\star > 0.5$.*

Proof. We first show sufficiency. This is obviously true because of (12) and (14) and the fact that $\mathrm{P}(\lambda_\star = 1) + \mathrm{P}(\lambda_\star = -1) = 1$. We then show necessity. Since $\pi_\star > 0.5$, we have that

$$\mathrm{P}(\lambda_\star = 1 \mid y = 1) + \mathrm{P}(\lambda_\star = -1 \mid y = -1) > 1 = \mathrm{P}(\lambda_\star = 1) + \mathrm{P}(\lambda_\star = -1).$$

If $\mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_\star = 1)$ and $\mathrm{P}(\lambda_\star = -1 \mid y = -1) > \mathrm{P}(\lambda_\star = -1)$, necessity will be true. So, it suffices to show that other configurations between $\mathrm{P}(\lambda_\star = 1 \mid y = 1)$, $\mathrm{P}(\lambda_\star = -1 \mid y = -1)$ and $\mathrm{P}(\lambda_\star = 1)$, $\mathrm{P}(\lambda_\star = -1)$ are not possible. Suppose that $\mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_\star = 1)$ but $\mathrm{P}(\lambda_\star = -1 \mid y = -1) \leq \mathrm{P}(\lambda_\star = -1)$. By (12), the former implies that $\lambda_\star$ is better than random guessing while the latter implies that $\lambda_\star$ is not better than random guessing, which is a contradiction. By a similar argument, we can show that $\mathrm{P}(\lambda_\star = 1 \mid y = 1) \leq \mathrm{P}(\lambda_\star = 1)$ and $\mathrm{P}(\lambda_\star = -1 \mid y = -1) > \mathrm{P}(\lambda_\star = -1)$ is not possible either. As a result, if $\mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_\star = 1)$, $\mathrm{P}(\lambda_\star = -1 \mid y = -1) > \mathrm{P}(\lambda_\star = -1)$ must hold, suggesting $\lambda_\star$ is better than random guessing. Furthermore, if $\mathrm{P}(\lambda_\star = 1 \mid y = 1) \leq \mathrm{P}(\lambda_\star = 1)$, $\mathrm{P}(\lambda_\star = -1 \mid y = -1) \leq \mathrm{P}(\lambda_\star = -1)$ must hold. But this will imply that $\pi \leq 0.5$, a contradiction. Therefore, we have proven necessity. □

B.2 Abstention

Different types of LFs may exhibit different behaviors. For example, one key distinction between unipolar LFs and bipolar LFs is their behaviors on abstention. While it is customary to assume that abstention in bipolar LFs does not provide additional class information (Fu et al., 2020), this is not the case for unipolar LFs. In fact,

Proposition 2. *Abstaining in a unipolar LF is equivalent to labeling the opposite class.*

Proposition 2 is a direct consequence of encoding unipolar LFs as binary LFs and a binary LF can either be better, equivalent, or worse than random guessing, as shown in Appendix B.1. As a result, even abstention in unipolar LFs provide additional class information, unlike bipolar LFs. Such distinction between different types of LFs suggests that one should model different types of LFs separately, which may introduce additional complexity. Indeed, for simplicity many existing weak supervision frameworks model after only one type of LFs (Ratner et al., 2019; Arachie and Huang, 2020; Chen et al., 2021). In contrast, Firebolt uses a unified LF representation by modeling unipolar and bipolar LFs through binary LFs.

B.3 LF Representation

An important property that an appropriate LF representation preserves is the relationship of an LF with random guessing. Indeed, the LF representation of Firebolt preserves this property:

Proposition 3. *Representing unipolar/bipolar LFs as binary LFs in Firebolt preserves their relationships with random guessing.*

Proof. From Proposition 2, it is obviously the case that the LF representation in Firebolt preserves the relationships of unipolar LFs to random guessing. It remains to show that this is also the case for bipolar LFs.

From Definition 1, we say that a bipolar LF $\lambda_\star$ is better than random guessing if $\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1)$. This intuitively makes sense because if a better than random guessing labeling function votes positive, we should have more confidence about the ground truth label being positive. Similarly, we say that $\lambda_\star$ performs worse than random guessing if $\mathrm{P}(y = 1 \mid \lambda_\star = 1) < \mathrm{P}(y = 1)$ and we say that $\lambda_\star$ performs equivalently to random guessing if $\mathrm{P}(y = 1 \mid \lambda_\star = 1) = \mathrm{P}(y = 1)$.

Furthermore, We use $\lambda_\star = 0$ to represent that $\lambda_\star$ abstains. Since semantically, abstention means that the labeling function lacks enough information to classify the data point one way or another, we can mathematically represent such a meaing as $\mathrm{P}(y \mid \lambda_\star = 0) = \mathrm{P}(y)$. This implies $\mathrm{P}(y = 1 \mid \lambda_\star = 0) = \mathrm{P}(y = 1)$ and $\mathrm{P}(y = -1 \mid \lambda_\star = 0) = \mathrm{P}(y = -1)$. Furthermore, using the Bayes theorem, we have that $\mathrm{P}(\lambda_\star = 0 \mid y = 1) = \mathrm{P}(\lambda_\star = 0 \mid y = -1) = \mathrm{P}(\lambda_\star = 0)$.

According to the LF representation scheme of Firebolt, a bipolar LF can be represented by a pair of positive LF λ_+ and negative LF λ_- (which in turn can be represented by binary LFs). It remains to show that if $\lambda_\star$ performs better than random guessing, both λ_+ and λ_- perform better than random guessing.

For this purpose, we first would like to show $\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1) \Rightarrow \mathrm{P}(y = -1 \mid \lambda_\star = -1) > \mathrm{P}(y = -1)$. Notice that $\mathrm{P}(y = 1 \mid \lambda_\star = 1) = \mathrm{P}(\lambda_\star = 1 \mid y = 1)\mathrm{P}(y = 1)/\mathrm{P}(\lambda_\star = 1) > \mathrm{P}(y = 1)$. This means $\mathrm{P}(\lambda_\star = 1 \mid y = 1) > \mathrm{P}(\lambda_\star = 1)$ and hence $\mathrm{P}(\lambda_\star \neq 1 \mid y = 1) < \mathrm{P}(\lambda_\star \neq 1)$. This implies that $\mathrm{P}(\lambda_\star = 0 \mid y = 1) + \mathrm{P}(\lambda_\star = -1 \mid y = 1) < \mathrm{P}(\lambda_\star = 0) + \mathrm{P}(\lambda_\star = -1)$ and hence $\mathrm{P}(\lambda_\star = -1 \mid y = 1) < \mathrm{P}(\lambda_\star = -1)$ because $\mathrm{P}(\lambda_\star = 0 \mid y = 1) = \mathrm{P}(\lambda_\star = 0)$ due to the property of abstention. Now, consider $\mathrm{P}(y = 1 \mid \lambda_\star = -1) = \mathrm{P}(\lambda_\star = -1 \mid y = 1)\mathrm{P}(y = 1)/\mathrm{P}(\lambda_\star = -1) < \mathrm{P}(y = 1)$, where we have used the fact that $\mathrm{P}(\lambda_\star = -1 \mid y = 1) < \mathrm{P}(\lambda_\star = -1)$. As a result, $\mathrm{P}(y = -1 \mid \lambda_\star = -1) > \mathrm{P}(y = -1)$.

Also notice that by definition of λ_+ and λ_-, we have that $\mathrm{P}(y = 1 \mid \lambda_\star = 1) = \mathrm{P}(y = 1 \mid \lambda_+ = 1)$ and $\mathrm{P}(y = -1 \mid \lambda_\star = -1) = \mathrm{P}(y = -1 \mid \lambda_- = -1)$. We therefore have that $\mathrm{P}(y = 1 \mid \lambda_+ = 1) > \mathrm{P}(y = 1)$ and $\mathrm{P}(y = -1 \mid \lambda_- = -1) > \mathrm{P}(y = -1)$ because we have shown previously that $\mathrm{P}(y = 1 \mid \lambda_\star = 1) > \mathrm{P}(y = 1)$ and $\mathrm{P}(y = -1 \mid \lambda_\star = -1) > \mathrm{P}(y = -1)$. Since both λ_+ and λ_- are unipolar labeling functions, $\mathrm{P}(y = 1 \mid \lambda_+ = 1) > \mathrm{P}(y = 1)$ and $\mathrm{P}(y = -1 \mid \lambda_- = -1) > \mathrm{P}(y = -1)$ mean λ_+ and λ_- are both better than random guessing.

Using similar arguments, we can also show that if a bipolar labeling function is worse than (equivalent to) random guessing, the associated pair of positive and negative labeling functions are also worse than (equivalent to) random guessing □

Given that we can represent both unipolar and bipolar LFs as binary LFs while preserving their relationships with random guessing, Firebolt can consider weak supervision problems with binary labeling functions as input without loss of generality.

B.4 Writing Labeling Functions

While weak supervision promises producing training labels without the need for manual labeling through the use of labeling functions, in reality, we observe that labeling functions of good quality are typically difficult to write. In response, we make the following observations on writing labeling functions:

- Good labeling functions tend to be positively correlated with each other, and one may aim at writing LFs that have positive covariance between each other.
- To encourage positive covariance among unipolar LFs, one may increase overlap between LFs of the same polarity and decrease conflict between LFs of the opposite polarity.
- High covariance between two LFs implies direct dependency between them.

It should be noticed that all these observations can be made without the aid of labeled data. While the aforementioned observations are made on population-level quantities, in practice, we can use sample-level quantities to estimate population-level quantities with a sufficient number of unlabeled data points. In what follows, we will further discus these observations.

LF Quality and Covariance We consider a set of LFs to be of good quality if they are better than random guessing and are conditionally independent of each other. From (2) and (3), we note that for such a set of LFs of good quality, the three way covariance among triplets of LFs T^*_{jkl}'s and the two way covariance between pairs of LFs Σ^*_{jk}'s will all be positive. Therefore, we can consider positive covariance entries of the population level a *necessary condition* for LFs to be better than random guessing and conditionally independent of each other. In practice, this observation implies that *one may write LFs that are positively correlated with each other* in order to meet this necessary condition.

Encourage Positive Covariance However, even though it may be desirable to write LFs that are positively correlated between each other, it is unclear what needs to be done to produce such labeling functions. To answer this question, we show that for unipolar LFs, *positive two-way covariance can be achieved by high overlap and low conflict between LFs*. We say that two labeling functions overlap with each other if the are of the same polarity and both of them vote for a given data point. We say that two labeling functions conflict with each other if they are of opposite polarity and both of them vote for a given data point. Below we show how overlap and conflict are associated with the covariance.

To proceed with our derivation, consider two binary labeling functions λ_j and λ_k. Without loss of generality, we use the decision made by λ_k as a reference. We then can define the true positive probability, true negative probability, false positive probability, and false negative probability respectively as: $tp = \mathrm{P}(\lambda_j = 1, \lambda_k = 1)$, $tn = \mathrm{P}(\lambda_j = -1, \lambda_k = -1)$, $fp = \mathrm{P}(\lambda_j = 1, \lambda_k = -1)$, and $fn = \mathrm{P}(\lambda_j = -1, \lambda_k = 1)$. Furthermore, we define the positive probability and the negative probability as $p = tp + fp = \mathrm{P}(\lambda_j = 1)$ and $n = tn + fn = \mathrm{P}(\lambda_j = -1)$. With these definitions, the covariance between λ_j and λ_k can be given as:

$$\Sigma^*_{jk} = (tp + tn - fp - fn) - (tp + fp - tn - fn)(tp + fn - tn - fp). \tag{15}$$

Using the fact that $fp = p - tp$ and $tn = (1 - p) - fn$, (15) can be written as:

$$\Sigma^*_{jk} = 4(n \cdot tp - p \cdot fn). \tag{16}$$

As a result, the sign of Σ^*_{jk} is determined by the sign of $(n \cdot tp - p \cdot fn)$, and $\Sigma^*_{jk} < 0$ is equivalent to $n/p < fn/tp$. Suppose that the two binary LFs represent two positive LFs, then fn is the proportion of data where λ_k votes positive but λ_j abstains, representing the scenario where λ_j and λ_k do not overlap. tp are the proportion of data where both λ_j and λ_k vote positive, representing the scenario where λ_j overlaps with λ_k. Therefore, the less frequent the two LFs overlap, the higher the ratio fn/tp, the more likely the covariance between the two LFs being negative.

(16) also holds when λ_j represents a negative labeling function and λ_k represents a positive labeling function. In this case, fn means λ_j votes negative and λ_k votes positive. It represents the situation where there are conflicts between the two labeling functions. On the other hand, tp means λ_j abstains and λ_k votes positive. It

represents the situation where the two labeling functions do not conflict with each other. As a result, the higher the proportion of conflicts among the two labeling functions, the more likely the covariance between the two labeling functions are negative.

Similarly, using the fact that $tp = p - fp$ and $fn = (1 - p) - tn$, we can write (15) as

$$\Sigma^*_{jk} = 4(p \cdot tn - n \cdot fp), \tag{17}$$

where $\Sigma^*_{jk} < 0$ if and only if $p/n < fp/tn$. Suppose the two binary LFs represent two two negative labeling functions, then fp means λ_j abstains and λ_k votes negative. It represents the situation where the two labeling functions do not overlap. On the other hand, tn means both λ_j and λ_k vote negative. It represents the situation where the two labeling functions overlap with each other. Therefore, if there is a high proportion of non-overlapping among the two labeling functions, we will observe negative covariance.

(17) also holds when λ_j is a positive labeling function and λ_k is a negative labeling function. In this case, fp means λ_j votes positive and λ_k votes negative. It represents conflicts between the two labeling functions. On the other hand, tn means λ_j abstains and λ_k votes negative. It represents no conflict between the two labeling functions.

High Covariance Implies Direct Dependency It should be noticed that just because two LFs have positive covariance between each other does not necessarily mean they are of good quality. For example, if the two LFs are both worse than random guessing, they could still have a positive covariance between each other. For another example, if one LF is a copy of the other LF, they will be perfectly correlated with each other. But using these two LFs together will not provide additional information to improve label quality. Indeed, we can show that if the magnitude of the covariance is too high between two LFs, the two LFs cannot be conditionally independent of each other. To this end, for two conditionally independent LFs λ_j and λ_k from (3) we have that

$$|\Sigma^*_{jk}| = |2\pi^*_j - 1||2\pi^*_k - 1|\sigma^{*2}_{00}.$$

Since $|2\pi^*_j - 1| \in [0, 1]$ and $|2\pi^*_k - 1| \in [0, 1]$, we have that $|\Sigma^*_{jk}| \leq \sigma^{*2}_{00}$. By a contrapositive argument, if we observe $|\Sigma^*_{jk}| > \sigma^{*2}_{00}$—that is, the magnitude of covariance between the two LFs are higher than the variance of the class balance of the ground truth label—the two LFs in question must not be conditionally independent of each other.

We make the following remarks for this procedure. To begin with, we have used σ^*_{00}. This information may or may not be available depending on whether we know the class balance of the dataset or not. Furthermore, if we know that the two LFs are better than random guessing but $\Sigma^*_{jk} < 0$, we can also conclude that the two LFs are not conditionally independent of each other. Finally, we can also determine whether a triplet of LFs are conditionally independent of each other through the use of $|T^*_{jkl}|$ with (3) in a similar fashion to the use of two-way covariance.

While these aforementioned observations by no mean provide a comprehensive guideline on writing labeling functions, it nonetheless offers some actionable insights to debug labeling functions and understand the quality and dependency among them. Furthermore, these observations do not require the use of labeled data, making them applicable to more practical situations where labeled data are not easy to collect.

B.5 Learning with Known Class Balance

In this section, we describe the parameter estimation procedure when the class balance of the data distribution is known. In this case, we do not need to make use of the covariance tensor to estimate the balanced accuracy of the LFs. Indeed, given a triplet of conditionally independent LFs λ_j, λ_k, and λ_l, we can directly compute the balanced accuracy of each of the three LFs by solving the following system of equations associated with the covariance, the balanced accuracy, and the given μ_{00} due to known class balance according to (2):

$$\begin{aligned}
\Sigma_{jk} &= (2\pi_j - 1)(2\pi_k - 1)(1 - \mu^2_{00}) \\
\Sigma_{kl} &= (2\pi_k - 1)(2\pi_l - 1)(1 - \mu^2_{00}) \\
\Sigma_{jl} &= (2\pi_j - 1)(2\pi_l - 1)(1 - \mu^2_{00}).
\end{aligned} \tag{18}$$

In practice, we typically have more than three LFs and do not observe the population-level covariance matrix. We can generalize from (18) and make use of the sample covariance matrix $\hat{\Sigma}$, the estimated dependency graph

$\hat{G}$, and the known class balance to estimate the balanced accuracy of each labeling function. In detail, we solve the following least squares problem:

$$\hat{l} = \arg\min_{l} \|\hat{M}l - \hat{q}\|_2^2,$$

where $\hat{M}$ is the incidence matrix of the inverse graph of $\hat{G}$, l is a $p \times 1$ parameter vector, and $\hat{q}$ is a vector constructed by stacking up $\log(\hat{\Sigma}_{jk}/(1-\mu_{00}^2))$'s, with (j,k)'s correspond to the rows of the incidence matrix representing the edge (j,k) in the inverse graph of $\hat{G}$. Upon knowing $\hat{l}$, we can estimate the balanced accuracy as $\hat{\pi} = \frac{\exp(\hat{l})+1}{2}$, where $\exp(\cdot)$ acts component-wise on $\hat{l}$. Once we know $\hat{\pi}$, the rest of estimation proceed as the same as previously described.

B.6 Firebolt Triplet Method Formulation

Fu et al. (2020) showcases the advantage of the triplet method for parameter estimation of the label model in weak supervision, where the parameters of labeling functions can be computed and aggregated efficiently in a closed-form manner by solving a series of systems of equations that involve triplets of conditionally independent labeling functions. Using (2) and (3), we can also directly derive a triplet formulation of Firebolt for binary classification problem. In detail, suppose that λ_j, λ_k, and λ_l is a triplet of conditionally independent LFs, and without loss of generality, $\pi_j^* > \frac{1}{2}$, $\pi_k^* > \frac{1}{2}$, $\pi_l^* > \frac{1}{2}$, $\mu_{00}^* \leq 0$, $\hat{\Sigma}_{jk} > 0$, $\hat{\Sigma}_{jl} > 0$, $\hat{\Sigma}_{kl} > 0$, and $\hat{T}_{jkl} > 0$, we can use (2) and (3) to analytically derive estimators:

$$\hat{\pi}_j = \frac{1}{2}\left(\sqrt{\frac{\hat{\Sigma}_{jk}\hat{\Sigma}_{jl}}{\hat{\Sigma}_{kl}}} + 1\right),\ \hat{\pi}_k = \frac{1}{2}\left(\sqrt{\frac{\hat{\Sigma}_{jk}\hat{\Sigma}_{kl}}{\hat{\Sigma}_{jl}}} + 1\right),\ \hat{\pi}_l = \frac{1}{2}\left(\sqrt{\frac{\hat{\Sigma}_{jl}\hat{\Sigma}_{kl}}{\hat{\Sigma}_{jk}}} + 1\right),\ \hat{\mu}_{00} = -\sqrt{\frac{\hat{T}_{jkl}^2}{\hat{T}_{jkl}^2 + 4\hat{\Sigma}_{jk}\hat{\Sigma}_{jl}\hat{\Sigma}_{kl}}}. \tag{19}$$

If we have multiple conditionally independent triplets to yield multiple estimators in (19), we can aggregate these estimators in a robust fashion by taking the median among the estimators (Chen et al., 2021). The resultant $\hat{\pi}$ and $\hat{\mu}_{00}$ can then be used downstream in the same way as we described in Section 3.

Compared to Fu et al. (2020), the triplet method formulation of Firebolt can solve the parameter estimation problem of Ising models with external field in one pass, instead of the two-pass procedure described in Section C.2 of Fu et al. (2020). Using the inference procedure of Firebolt, we can also carry out inference over the Ising model with arbitrary external fields learned by the Firebolt triplet method efficiently. This is also in contrast to Fu et al. (2020), where exact inference is carried out efficiently when making assumptions on the external fields of the Ising model.

Moreover, the triplet formulation of Firebolt also removes the assumption that the users need to know if the classification problem in question is a balanced classification problem or not. Nonetheless, it still assumes that users can encode the minority class as positive. We anticipate that similar theoretical guarantees can be derived following the techniques developed in Appendix C and Fu et al. (2020) for the Firebolt triplet method, but we leave such results as future work.

B.7 Inference

In this section, we describe the inference algorithm used in Firebolt. For label model with general dependency, we derive an inference algorithm (Appendix B.7.1) based on solving a logistic regression problem that has been described in Section 3.3. We then describe closed-form exact inference algorithms for conditionally independent unipolar (Appendix B.7.2) and bipolar labeling functions (Appendix B.7.3).

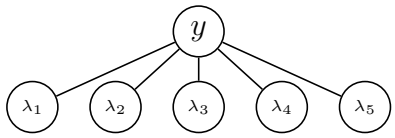
B.7.1 General Graph

We first describe how to compute probabilistic labels $\mathrm{P}(y = 1 \mid \lambda)$ given that the joint distribution between y and λ is modeled by (1). The result is shown in Proposition 4.

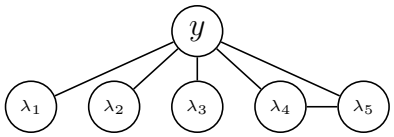
Proposition 4. *Let y and λ follow the joint distribution given in* (1)*, the probabilistic label of y given λ is given as:*

$$\mathrm{P}(y = 1 \mid \lambda) = \mathrm{sigmoid}\left(2\theta_{00}^* + 2\theta_{0+}^{*\top}\lambda\right), \tag{20}$$

where $\mathrm{sigmoid}(t) = \frac{1}{1+\exp(-t)}$.



(a) A label model representing five conditionally independent unipolar labeling functions and the ground truth label.



(b) A label model representing four conditionally independent labeling functions and the ground truth label. λ_1-λ_3 represent three unipolar LFs. λ_4 and λ_5 jointly represent a bipolar LF.

Figure 5: Conditionally independent label models.

Proof. The proof of Theorem 4 follows similarly to the arguments in Ravikumar et al. (2010). In detail, by the Bayes theorem,

$$\begin{aligned}
\mathrm{P}(y=1 \mid \lambda) =& \frac{\mathrm{P}(y=1,\lambda)}{\mathrm{P}(\lambda)} \\
=& \frac{\mathrm{P}(y=1,\lambda)}{\mathrm{P}(y=1,\lambda)+\mathrm{P}(y=-1,\lambda)} \\
=& \frac{1}{1+\frac{\mathrm{P}(y=-1,\lambda)}{\mathrm{P}(y=1,\lambda)}} \\
=& \frac{1}{1+\frac{\frac{1}{Z}\exp\left(-\theta^*_{00}+\sum_{j=1}^p \theta^*_{jj}\lambda_j-\theta^*_{0j}\lambda_j+\sum_{(\lambda_j,\lambda_k)\in E^*}\theta^*_{jk}\lambda_j\lambda_k\right)}{\frac{1}{Z}\exp\left(\theta^*_{00}+\sum_{j=1}^p \theta^*_{jj}\lambda_j+\theta^*_{0j}\lambda_j+\sum_{(\lambda_j,\lambda_k)\in E^*}\theta^*_{jk}\lambda_j\lambda_k\right)}} \\
=& \frac{1}{1+\frac{\exp\left(-\theta^*_{00}-\sum_{j=1}^p \theta^*_{0j}\lambda_j\right)}{\exp\left(\theta^*_{00}+\sum_{j=1}^p \theta^*_{0j}\lambda_j\right)}} \\
=& \frac{1}{1+\exp(-2\theta^*_{00}-2\sum_{j=1}^p \theta^*_{0j}\lambda_j)} \\
=& \text{sigmoid}\left(2\theta^*_{00}+2\theta^{*\top}_{0+}\lambda\right).
\end{aligned}$$

□

When y is observed, Ravikumar et al. (2010) also suggests that one can provide an estimate of $\theta^*_{0\cdot}$ called $\hat{\theta}_{0\cdot}$ by solving a logistic regression problem that regress y on λ. Such a logistic regression problem can be written as (10), where $\hat{\mu}_{0\cdot}$ can be directly obtained from the dataset when y is observed. However, since y is not observed in a weak supervision setting, we used the parameters estimated by Firebolt $\hat{\mu}_{0\cdot}$ in lieu of their counterparts directly derived from the dataset in a fully observed setting. As a result, we have derived the inference algorithm that we described in Section 3.3. In Appendix C, we provide theoretical guarantees for downstream model performance using the probabilistic labels produced by our inference procedure.

B.7.2 Conditionally Independent Unipolar LFs

One common situation arises in practice is when the Ising model represents the joint distribution between y and unipolar LFs that are conditionally independent with each other upon y, such as the model described in Figure 5a. In this case, exact inference can be carried out in closed-form using the sensitivity and specificity parameters of each labeling functions as well as the class balance. Indeed,

Proposition 5. *Let y and λ follow the following joint distribution given by the conditionally independent Ising model*

$$\mathrm{P}(y,\lambda) = \frac{1}{Z}\exp\left(\theta^*_{00}y + \sum_{j=1}^p \theta^*_{jj}\lambda_j + \theta^*_{0j}\lambda_j y\right).$$

Let $\alpha^{\pm *}$ be the sensitivity and specificity parameters associated with the labeling functions and let μ^*_{00} be the class

balance parameter. We have that

$$\mathrm{P}(y=1 \mid \lambda) = \text{sigmoid}\left(\log \frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \sum_{j=1}^{p} \log \frac{\mathrm{P}(\lambda_j \mid y=1)}{\mathrm{P}(\lambda_j \mid y=-1)}\right). \tag{21}$$

Furthermore, for all $j \in \{1, 2, \cdots, p\}$,

$$\begin{aligned} \theta_{00}^* =& \frac{1}{2}\log\frac{1+\mu_{00}^*}{1-\mu_{00}^*} + \frac{1}{4}\sum_{j=1}^{p}\log\frac{\alpha_j^{+*}(1-\alpha_j^{+*})}{\alpha_j^{-*}(1-\alpha_j^{-*})}, \\ \theta_{0j}^* =& \frac{1}{4}\left(\log\frac{\alpha_j^{+*}}{1-\alpha_j^{+*}} + \log\frac{\alpha_j^{-*}}{1-\alpha_j^{-*}}\right). \end{aligned} \tag{22}$$

Proof. (21) is true because

$$\begin{aligned} \mathrm{P}(y=1 \mid \lambda) =& \frac{\mathrm{P}(\lambda \mid y=1)\mathrm{P}(y=1)}{\mathrm{P}(\lambda \mid y=1)\mathrm{P}(y=1) + \mathrm{P}(\lambda \mid y=-1)\mathrm{P}(y=-1)} \\ =& \frac{1}{1 + \frac{\mathrm{P}(\lambda|y=-1)\mathrm{P}(y=-1)}{\mathrm{P}(\lambda|y=1)\mathrm{P}(y=1)}} \\ =& \frac{1}{1 + \frac{\prod_{j=1}^{p}\mathrm{P}(\lambda_j|y=-1)\mathrm{P}(y=-1)}{\prod_{j=1}^{p}\mathrm{P}(\lambda_j|y=1)\mathrm{P}(y=1)}} \\ =& \frac{1}{1+\exp\left(-\log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} - \sum_{j=1}^{p}\log\frac{\mathrm{P}(\lambda_j|y=1)}{\mathrm{P}(\lambda_j|y=-1)}\right)} \\ =& \text{sigmoid}\left(\log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \sum_{j=1}^{p}\log\frac{\mathrm{P}(\lambda_j \mid y=1)}{\mathrm{P}(\lambda_j \mid y=-1)}\right), \end{aligned} \tag{23}$$

where for the third equality, we have used the fact that $\mathrm{P}(\lambda \mid y) = \prod_{j=1}^{p}\mathrm{P}(\lambda_j \mid y)$ because of the conditional independence among the labeling functions.

Next, we show (22) is True. For this purpose, we compare (20) with (21). In particular, consider $\lambda_j = 1$ and $\lambda_j = -1$ for all $j \in \{1, 2, \cdots, p\}$, we have that

$$\begin{cases} 2\theta_{00}^* + 2\sum_{j=1}^{p}\theta_{0j}^* = \log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \sum_{j=1}^{p}\log\frac{\alpha_j^{+*}}{1-\alpha_j^{-*}} \\ 2\theta_{00}^* - 2\sum_{j=1}^{p}\theta_{0j}^* = \log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \sum_{j=1}^{p}\log\frac{1-\alpha_j^{+*}}{\alpha_j^{-*}} \end{cases} \Rightarrow \theta_{00}^* = \frac{1}{2}\log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \frac{1}{4}\sum_{j=1}^{p}\log\frac{\alpha_j^{+*}(1-\alpha_j^{+*})}{\alpha_j^{-*}(1-\alpha_j^{-*})}.$$

Furthermore, consider $\mathrm{P}(y=1 \mid \lambda_j = 1, \lambda_{-j})$ and $\mathrm{P}(y=1 \mid \lambda_j = -1, \lambda_{-j})$, where λ_{-j} represents a column vector that consists of all the entries in λ but λ_j, we have that

$$\begin{aligned} &\begin{cases} 2\theta_{00}^* + 2\theta_{0j}^* + \sum_{j'\neq j} 2\theta_{0j}^*\lambda_{j'} = \log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \log\frac{\alpha_j^{+*}}{1-\alpha_j^{-*}} + \sum_{j'\neq j}\log\frac{\mathrm{P}(\lambda_{j'}|y=1)}{\mathrm{P}(\lambda_{j'}|y=-1)} \\ 2\theta_{00}^* - 2\theta_{0j}^* + \sum_{j'\neq j} 2\theta_{0j}^*\lambda_{j'} = \log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \log\frac{1-\alpha_j^{+*}}{\alpha_j^{-*}} + \sum_{j'\neq j}\log\frac{\mathrm{P}(\lambda_{j'}|y=1)}{\mathrm{P}(\lambda_{j'}|y=-1)} \end{cases} \\ \Rightarrow &\theta_{0j}^* = \frac{1}{4}\left(\log\frac{\alpha_j^{+*}}{1-\alpha_j^{-*}} - \log\frac{1-\alpha_j^{+*}}{\alpha_j^{-*}}\right) = \frac{1}{4}\left(\log\frac{\alpha_j^{+*}}{1-\alpha_j^{+*}} + \log\frac{\alpha_j^{-*}}{1-\alpha_j^{-*}}\right). \end{aligned}$$

□

B.7.3 Conditionally Independent Unipolar/Bipolar LFs

Another common situation is when we are dealing with a mix of conditionally independent unipolar and bipolar LFs. Such a situation arises, for example, when we produce some LFs by bucketizing the probability

scores produced by various classifiers that are conditionally independent of each other into bipolar LFs. After transforming the original LFs into binary LFs λ, the joint distribution between λ and y can be represented by an Ising model where there is at most one edge between one labeling function and another for any given labeling function. Figure 5b describes such a label model, where inference can still be carried out using (23), excepts that $\mathrm{P}(\lambda \mid y) = \mathrm{P}(\lambda_4, \lambda_5 \mid y) \prod_{j=1}^{3} \mathrm{P}(\lambda_k \mid y)$, where we need to compute the quantity $\mathrm{P}(\lambda_4, \lambda_5 \mid y)$. In fact, it is not difficult to see, for conditionally independent label model with a mixed of unipolar and bipolar LFs, we have that

$$\mathrm{P}(\lambda \mid y) = \prod_{j=1}^{p} \mathrm{P}(\lambda_k \mid y) \prod_{(\lambda_k, \lambda_l) \in E^*} \frac{\mathrm{P}(\lambda_k, \lambda_l \mid y)}{\mathrm{P}(\lambda_k \mid y)\mathrm{P}(\lambda_l \mid y)}. \tag{24}$$

It remains to show how to compute $\mathrm{P}(\lambda_k, \lambda_l \mid y)$ for $(\lambda_k, \lambda_l) \in E^*$, which is given in Theorem 6.

Proposition 6. *Suppose that λ_k and λ_l jointly represent a bipolar labeling function $\lambda_\star$ that is conditionally independent of other labeling functions, with λ_k representing the associated positive LF and λ_l representing the associated negative LF. We have that:*

$$\begin{aligned}
\mathrm{P}(\lambda_\star = 1 \mid y = 1) =& \mathrm{P}(\lambda_k = 1, \lambda_l = 1 \mid y = 1) = \alpha_k^{+*}, \\
\mathrm{P}(\lambda_\star = 0 \mid y = 1) =& \mathrm{P}(\lambda_k = -1, \lambda_l = 1 \mid y = 1) = \alpha_l^{+*} - \alpha_k^{+*}, \\
\mathrm{P}(\lambda_\star = -1 \mid y = 1) =& \mathrm{P}(\lambda_k = -1, \lambda_l = -1 \mid y = 1) = 1 - \alpha_l^{+*}, \\
\mathrm{P}(\lambda_\star = 1 \mid y = -1) =& \mathrm{P}(\lambda_k = 1, \lambda_l = 1 \mid y = -1) = 1 - \alpha_k^{-*}, \\
\mathrm{P}(\lambda_\star = 0 \mid y = -1) =& \mathrm{P}(\lambda_k = -1, \lambda_l = 1 \mid y = -1) = \alpha_k^{-*} - \alpha_l^{-*}, \\
\mathrm{P}(\lambda_\star = -1 \mid y = -1) =& \mathrm{P}(\lambda_k = -1, \lambda_l = -1 \mid y = -1) = \alpha_l^{-*}.
\end{aligned}$$

Proof. To compute $\mathrm{P}(\lambda_\star = 1|y = 1)$, $\mathrm{P}(\lambda_\star = 0|y = 1)$, $\mathrm{P}(\lambda_\star = -1|y = 1)$, $\mathrm{P}(\lambda_\star = 1|y = -1)$, $\mathrm{P}(\lambda_\star = 0|y = -1)$, and $\mathrm{P}(\lambda_\star = -1|y = -1)$, it suffices to focus on $\mathrm{P}(\lambda_\star = 1|y = 1)$, $\mathrm{P}(\lambda_\star = 0|y = 1)$, $\mathrm{P}(\lambda_\star = 0|y = -1)$, and $\mathrm{P}(\lambda_\star = -1|y = -1)$. To this end, notice that through parameter learning we can provide estimates for $\alpha_k^{+*} = \mathrm{P}(\lambda_k = 1|y = 1)$, $\alpha_k^{-*} = \mathrm{P}(\lambda_k = -1|y = -1)$, $\alpha_l^{-*} = \mathrm{P}(\lambda_l = -1|y = -1)$, and $\alpha_l^{+*} = \mathrm{P}(\lambda_l = 1|y = 1)$. Furthermore,

$$\begin{aligned}
\alpha_k^{+*} =& \mathrm{P}(\lambda_k = 1, \lambda_l = 1 \mid y = 1) + \mathrm{P}(\lambda_k = 1, \lambda_l = -1 \mid y = 1) \\
\Rightarrow& \mathrm{P}(\lambda_\star = 1 \mid y = 1) = \alpha_k^{+*}, \\
\alpha_l^{+*} =& \mathrm{P}(\lambda_k = 1, \lambda_l = 1 \mid y = 1) + \mathrm{P}(\lambda_k = -1, \lambda_l = 1 \mid y = 1) \\
\Rightarrow& \mathrm{P}(\lambda_\star = 0 \mid y = 1) = \alpha_l^{+*} - \alpha_k^{+*}, \\
\alpha_l^{-*} =& \mathrm{P}(\lambda_k = 1, \lambda_l = -1 \mid y = -1) + \mathrm{P}(\lambda_k = -1, \lambda_l = -1 \mid y = -1) \\
\Rightarrow& \mathrm{P}(\lambda_\star = -1 \mid y = -1) = \alpha_l^{-*}, \\
\alpha_k^{-*} =& \mathrm{P}(\lambda_k = -1, \lambda_l = 1 \mid y = -1) + \mathrm{P}(\lambda_k = -1, \lambda_l = -1 \mid y = -1) \\
\Rightarrow& \mathrm{P}(\lambda_\star = 0 \mid y = -1) = \alpha_k^{-*} - \alpha_l^{-*}.
\end{aligned}$$

As a result,

$$\mathrm{P}(\lambda_\star = 1 \mid y = -1) = 1 - \alpha_k^{-*}, \quad \mathrm{P}(\lambda_\star = -1 \mid y = 1) = 1 - \alpha_l^{+*}.$$

We have completed the proof. □

We now present a generalization of Proposition 5 for a mix of conditionally independent unipolar and bipolar labeling functions.

Proposition 7. *Let y and λ follow the following joint distribution given by the Ising model representing a mix conditionally independent of unipolar and bipolar labeling functions,*

$$\mathrm{P}(y, \lambda) = \frac{1}{Z} \exp\left(\theta_{00}^* y + \sum_{j=1}^{p} \theta_{jj}^* \lambda_j + \sum_{\lambda_{j'} \in U^*} \theta_{0j'}^* \lambda_{j'} y + \sum_{(\lambda_k, \lambda_l) \in E^*} \theta_{kl}^* \lambda_k \lambda_l \right),$$

where U^* is the set of binary labeling functions that represent unipolar labeling functions in G^*. Let $\alpha^{\pm *}$ be the sensitivity and specificity parameters associated with the labeling functions and let μ_{00}^* be the class balance parameter. We have that

$$\mathrm{P}(y=1 \mid \lambda) = \mathrm{sigmoid}\left(\log \frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \sum_{\lambda_j \in U^*} \log \frac{\mathrm{P}(\lambda_j \mid y=1)}{\mathrm{P}(\lambda_j \mid y=-1)} + \sum_{(\lambda_k,\lambda_l)\in E^*} \log \frac{\mathrm{P}(\lambda_k,\lambda_l \mid y=1)}{\mathrm{P}(\lambda_k,\lambda_l \mid y=-1)}\right). \quad (25)$$

Furthermore, for all $\lambda_j \in U^*$,

$$\theta_{0j}^* = \frac{1}{4}\left(\log \frac{\alpha_j^{+*}}{1-\alpha_j^{+*}} + \log \frac{\alpha_j^{-*}}{1-\alpha_j^{-*}}\right).$$

For all $(\lambda_k, \lambda_l) \in E^*$, where λ_k representing the positive LF and λ_l representing the negative LF in a bipolar LF representation,

$$\theta_{0k}^* = \frac{1}{4}\left(\log \frac{\alpha_k^{+*}}{1-\alpha_k^{-*}} - \log \frac{\alpha_l^{+*}-\alpha_k^{+*}}{\alpha_k^{-*}-\alpha_l^{-*}}\right), \quad \theta_{0l}^* = \frac{1}{4}\left(\log \frac{\alpha_l^{+*}-\alpha_k^{+*}}{\alpha_k^{-*}-\alpha_l^{-*}} - \log \frac{1-\alpha_l^{+*}}{\alpha_l^{-*}}\right).$$

Finally,

$$\theta_{00}^* = \frac{1}{2}\log \frac{1+\mu_{00}^*}{1-\mu_{00}^*} + \frac{1}{4}\sum_{\lambda_j \in U^*} \log \frac{\alpha_j^{+*}(1-\alpha_j^{+*})}{\alpha_j^{-*}(1-\alpha_j^{-*})} + \frac{1}{4}\sum_{(\lambda_k,\lambda_l)\in E^*} \log \frac{\alpha_k^{+*}(1-\alpha_l^{+*})}{\alpha_l^{-*}(1-\alpha_k^{-*})}.$$

.

Proof. Using U^*, we can rewrite (24) as:

$$\mathrm{P}(\lambda \mid y) = \prod_{j\in U^*} \mathrm{P}(\lambda_j \mid y) \prod_{(\lambda_k,\lambda_l)\in E^*} \mathrm{P}(\lambda_k,\lambda_l \mid y). \quad (26)$$

Using (23) and (26), we have that

$$\mathrm{P}(y=1 \mid \lambda) = \mathrm{sigmoid}\left(\log \frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \sum_{\lambda_j \in U^*} \log \frac{\mathrm{P}(\lambda_j \mid y=1)}{\mathrm{P}(\lambda_j \mid y=-1)} + \sum_{(\lambda_k,\lambda_l)\in E^*} \log \frac{\mathrm{P}(\lambda_k,\lambda_l \mid y=1)}{\mathrm{P}(\lambda_k,\lambda_l \mid y=-1)}\right).$$

Next, we compute the canonical parameters. To begin with, suppose that $\lambda_j \in U^*$. Consider $\mathrm{P}(y=1 \mid \lambda_j=1, \lambda_{-j})$ and $\mathrm{P}(y=1 \mid \lambda_j=-1, \lambda_{-j})$, we have that

$$\begin{aligned}
2\theta_{00}^* + 2\theta_{0j}^* + \sum_{j'\neq j} 2\theta_{0j}^*\lambda_{j'} =& \log \frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \log \frac{\alpha_j^{+*}}{1-\alpha_j^{-*}} \\
&+ \sum_{j'\neq j, j'\in U^*} \log \frac{\mathrm{P}(\lambda_{j'} \mid y=1)}{\mathrm{P}(\lambda_{j'} \mid y=-1)} + \sum_{(\lambda_k,\lambda_l)\in E^*} \log \frac{\mathrm{P}(\lambda_j,\lambda_k \mid y=1)}{\mathrm{P}(\lambda_j,\lambda_k \mid y=-1)}, \\
2\theta_{00}^* - 2\theta_{0j}^* + \sum_{j'\neq j} 2\theta_{0j}^*\lambda_{j'} =& \log \frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)} + \log \frac{1-\alpha_j^{+*}}{\alpha_j^{-*}} \\
&+ \sum_{j'\neq j, j'\in U^*} \log \frac{\mathrm{P}(\lambda_{j'} \mid y=1)}{\mathrm{P}(\lambda_{j'} \mid y=-1)} + \sum_{(\lambda_k,\lambda_l)\in E^*} \log \frac{\mathrm{P}(\lambda_k,\lambda_l \mid y=1)}{\mathrm{P}(\lambda_k,\lambda_l \mid y=-1)}.
\end{aligned}$$

This means

$$\theta_{0j}^* = \frac{1}{4}\left(\log \frac{\alpha_j^{+*}}{1-\alpha_j^{+*}} + \log \frac{\alpha_j^{-*}}{1-\alpha_j^{-*}}\right)$$

is true for $\lambda_j \in U^*$. On the other hand, for a pair of λ_k and λ_l such that $(\lambda_k, \lambda_l) \in E^*$ with λ_k representing the positive LF and λ_l representing the negative LF. Let $\lambda_k = \pm 1$ and $\lambda_l = 1$, from Proposition 6 we have that

$$\log \frac{\mathrm{P}(\lambda_k=1,\lambda_l=1 \mid y=1)}{\mathrm{P}(\lambda_k=1,\lambda_l=1 \mid y=-1)} = \log \frac{\alpha_k^{+*}}{1-\alpha_k^{-*}}, \quad \log \frac{\mathrm{P}(\lambda_k=-1,\lambda_l=1 \mid y=1)}{\mathrm{P}(\lambda_k=-1,\lambda_l=1 \mid y=-1)} = \log \frac{\alpha_l^{+*}-\alpha_k^{+*}}{\alpha_k^{-*}-\alpha_l^{-*}}.$$

Write and compare $\mathrm{P}(y=1 \mid \lambda_k=1, \lambda_l=1, \lambda_{-\{k,l\}})$ and $\mathrm{P}(y=1 \mid \lambda_k=-1, \lambda_l=1, \lambda_{-\{k,l\}})$ with (20) and (25), we have that

$$\begin{cases} 2\theta^*_{00}+2\theta^*_{0k}+2\theta^*_{0l}+\sum_{j\notin\{k,l\}}2\theta^*_{0j}\lambda_j=\log\frac{\alpha_k^{+*}}{1-\alpha_k^{+*}}+f(\lambda_{-\{k,l\}}), \\ 2\theta^*_{00}-2\theta^*_{0k}+2\theta^*_{0l}+\sum_{j\notin\{k,l\}}2\theta^*_{0j}\lambda_j=\log\frac{\alpha_l^{+*}-\alpha_k^{+*}}{\alpha_k^{-*}-\alpha_l^{-*}}+f(\lambda_{-\{k,l\}}). \end{cases}$$

As a result,

$$\theta^*_{0k}=\frac{1}{4}\left(\log\frac{\alpha_k^{+*}}{1-\alpha_k^{-*}}-\log\frac{\alpha_l^{+*}-\alpha_k^{+*}}{\alpha_k^{-*}-\alpha_l^{-*}}\right).$$

Similarly,

$$\log\frac{\mathrm{P}(\lambda_k=-1,\lambda_l=-1\mid y=1)}{\mathrm{P}(\lambda_k=-1,\lambda_l=-1\mid y=-1)}=\log\frac{1-\alpha_l^{+*}}{\alpha_l^{-*}}.$$

Write and compare $\mathrm{P}(y=1 \mid \lambda_k=-1, \lambda_l=-1, \lambda_{-\{k,l\}})$ and $\mathrm{P}(y=1 \mid \lambda_k=-1, \lambda_l=1, \lambda_{-\{k,l\}})$ with (20) and (25), we have that

$$\begin{cases} 2\theta^*_{00}-2\theta^*_{0k}-2\theta^*_{0l}+\sum_{j\notin\{k,l\}}2\theta^*_{0j}\lambda_j=\log\frac{1-\alpha_l^{+*}}{\alpha_l^{-*}}+f(\lambda_{-\{k,l\}}), \\ 2\theta^*_{00}-2\theta^*_{0k}+2\theta^*_{0l}+\sum_{j\notin\{k,l\}}2\theta^*_{0j}\lambda_j=\log\frac{\alpha_l^{+*}-\alpha_k^{+*}}{\alpha_k^{-*}-\alpha_l^{-*}}+f(\lambda_{-\{k,l\}}). \end{cases}$$

As a result,

$$\theta^*_{0l}=\frac{1}{4}\left(\log\frac{\alpha_l^{+*}-\alpha_k^{+*}}{\alpha_k^{-*}-\alpha_l^{-*}}-\log\frac{1-\alpha_l^{+*}}{\alpha_l^{-*}}\right).$$

Finally, we determine θ^*_{00}. To this end, consider $\lambda_j=1$ and $\lambda_j=-1$ for all $j\in\{1,2,\cdots,p\}$, we have that

$$\begin{cases} 2\theta^*_{00}+\sum_{j=1}^p 2\theta^*_{0j}=\log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)}+\sum_{\lambda_j\in U^*}\log\frac{\alpha_j^{+*}}{1-\alpha_j^{-*}}+\sum_{(\lambda_k,\lambda_l)\in E^*}\log\frac{\alpha_k^{+*}}{1-\alpha_k^{-*}}, \\ 2\theta^*_{00}-\sum_{j=1}^p 2\theta^*_{0j}=\log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)}+\sum_{\lambda_j\in U^*}\log\frac{1-\alpha_j^{+*}}{\alpha_j^{-*}}+\sum_{(\lambda_k,\lambda_l)\in E^*}\log\frac{1-\alpha_l^{+*}}{\alpha_l^{-*}}. \end{cases}$$

As a result,

$$\theta^*_{00}=\frac{1}{2}\log\frac{\mathrm{P}(y=1)}{\mathrm{P}(y=-1)}+\frac{1}{4}\sum_{\lambda_j\in U^*}\log\frac{\alpha_j^{+*}(1-\alpha_j^{+*})}{\alpha_j^{-*}(1-\alpha_j^{-*})}+\frac{1}{4}\sum_{(\lambda_k,\lambda_l)\in E^*}\log\frac{\alpha_k^{+*}(1-\alpha_l^{+*})}{\alpha_l^{-*}(1-\alpha_k^{-*})}.$$

□

B.7.4 Choice of Inference Algorithms

It should be noticed that the aforementioned inference algorithms have their own advantages and disadvantages. Caution should be taken when deciding which inference algorithm to use in practice. On the one hand, when we have precise prior knowledge of the complex dependency between LFs, using the logistic regression based algorithm described in Appendix B.7.1 can be advantageous compared to exact inference algorithms. In this case, it is desirable to provide good estimators $\hat{\mu}_{00}$ and $\hat{\mu}_{0+}$ to deliver a mapping of good quality to $\hat{\theta}_{00}$ and $\hat{\theta}_{0+}$. On the other hand, when we do not have access to the dependency graph, one may assume that the LFs are conditionally independent between each other. In this case, exact inference algorithms described in Appendix B.7.2 and Appendix B.7.3 are more efficient compared to the logistic regression algorithm in Appendix B.7.1 because of their closed-form solutions.

B.8 Imbalanced Classification

In this section, we are going to show why being able to learn from Ising models with arbitrary external fields is important for weak supervision algorithms in an imbalanced classification setting. To this end, consider a conditional independent Ising model that represents the joint distribution of unipolar labeling functions and the ground truth label. Without loss of generality, suppose that λ_j is a labeling function that is not associated with an external field, that is $\theta^*_{jj}=0$. Because λ_j is only connected to y in G^*, following an argument similar to the proof of Proposition 4, we can show that $\mathrm{P}(\lambda_j=1 \mid y)=\mathrm{sigmoid}(\theta^*_{0j}y)$. As a result, both the sensitivity

and specificity of λ_j is the same and $\alpha_j^{\pm *} = \text{sigmoid}(\theta_{0j}^*)$. This relationship shows why the lack of external fields may imply that a labeling function has similar accuracy for different classes. Such a restriction may also lead to undesirable consequences in imbalanced classification problems.

Consider the typical imbalanced classification problem where the negative class is the majority class. For illustration purposes, let's consider a random guessing labeling function λ_j with $\alpha_j^+ + \alpha_j^- = 1$ (Proposition 1) and $\alpha_j^- \gg \alpha_j^+$. In extreme, if $\alpha_j^+ = 0$ and $\alpha_j^- = 1$, this corresponds to the constant labeling function that always labels everything as negative. Let $\alpha_j = P(\lambda_j = y)$ be the overall accuracy of labeling function λ_j in guessing y. For a negative constant labeling function, $\alpha_j = P(y = -1)$. Therefore, in our imbalanced classification problem, α_j is high. Our hope is that the inference process will exclude such random guessing labeling functions.

We first show that if the weak supervision model fails to identify that the sensitivity of the labeling function is different from its specificity, the influence of random guessing labeling functions may be arbitrarily bad during inference. Under this scenario suppose we know the accuracy of λ_j to be α_j, we will have that $\alpha_j = \alpha_j^+ = \alpha_j^-$, which means the labeling function will contribute $\log \frac{\alpha_j}{1-\alpha_j}\lambda_j$ according to (20) and (22). In this case, as the class imbalance exacerbates, $\alpha_j \to 1$ and the weight $\log \frac{\alpha_j}{1-\alpha_j} \to \infty$. As $\lambda_j = -1$, it will drive $P(y = 1|\lambda) \to 0$. This implies that λ_j will overwhelm other labeling functions with finite contribution and the data point will be classified as negative as a result. However, λ_j is a constant labeling function and does not have any predictive power and hence it should not contribute to the decision process at all.

On the other hand, if the weak supervision model can identify that the sensitivity of the labeling function is different from its specificity, λ_j will be safely excluded during inference. Specifically, by (21), when $\lambda_j = 1$, the labeling function will contribute $\log \frac{\alpha_j^+}{1-\alpha_j^-}$. When $\lambda_j = -1$, the labeling function will contribute $\log \frac{1-\alpha_j^+}{\alpha_j^-}$. Notice that Firebolt can identify α_j^+ and α_j^- and as a result can make use of the relationship that $\alpha_j^+ + \alpha_j^- = 1$ for a random guessing labeling function. The weights $\log \frac{\alpha_j^+}{1-\alpha_j^-} = \log \frac{1-\alpha_j^+}{\alpha_j^-} = \log 1 = 0$. Therefore, when the sensitivity and specificity are correctly identified, λ_j will contribute nothing to the decision process as desired.

B.9 Beyond Binary Classification

Multi-class Classification Handling abstention in bipolar labeling functions can be viewed as a multi-class classification problem. In general, we can carry out a one-vs-all or a one-vs-one reduction of multi-class classification problems into multiple binary classification problems and apply Firebolt. In Appendix D, we demonstrate such an example on a ten-class classification problem in using the animal attribute dataset.

Positive-Only Classification Since both positive and negative labeling functions can be represented as binary labeling functions, Firebolt can carry out weak supervision with only positive labeling functions and without negative labeling functions. Such a formulation is particularly useful in many problem domains where users find positive labeling functions are easier to write compared to writing negative labeling functions.

C Extended Theoretical Results

In this section, we present extended theoretical results of Firebolt. First, we present the auxiliary lemmas useful in our proof (Appendix C.1). Next, we analyze the parameter estimation error of Firebolt when the class balance is unknown (Appendix C.2). Furthermore, we also present corresponding results when the class balance is known (Appendix C.3). Finally, we analyze generalization error of downstream end models learned with Firebolt-produced probabilistic labels (Appendix C.4).

C.1 Auxiliary Lemmas

Definition 2 (Lipschitz Continuity, Paraphrase of Definition 4 of Honorio 2012)**.** *Let $f(x)$ be a function such that $f :\in \mathbb{R}^m \to \mathbb{R}$. We say that $f(x)$ is Lipschitz continuous with a Lipschitz constant K with respect to the ℓ_p norm if*

$$|f(x_1) - f(x_2)| \leq K\|x_1 - x_2\|_p.$$

Lemma 1 (Lipschitz Continuity for Differentiable Functions, Paraphrase of Definition 4 of Honorio 2012)**.** *Let $f(x)$ be a differentiable function such that $f :\in \mathbb{R}^m \to \mathbb{R}$. Let $\nabla f(x)$ be the gradient of $f(x)$. $f(x)$ is K-Lipschitz continuous if with respect to the ℓ_p norm for all $x \in \mathbb{R}^m$,*

$$\|\nabla f(x)\|_p \leq K.$$

Lemma 2 (Matrix Hoeffding's Inequality)**.** *Consider a finite sequence $\left\{X^{(i)}\right\}$ of independent, random, self-adjoint matrices with dimension d, and let $\{A_i\}$ be a sequence of fixed self-adjoint matrices. Assume that each random matrix satisfies*

$$\mathbb{E}X^{(i)} = 0 \quad \textit{and} \quad X^{(i)2} \preceq A^{(i)2} \quad \textit{almost surely.}$$

Then, for all $t \geq 0$,

$$\mathrm{P}\left(\left\|\sum_i X^{(i)}\right\|_2 \geq t\right) \leq d \cdot \exp\left(-\frac{t^2}{8\sigma^2}\right), \tag{27}$$

where $\sigma^2 = \left\|\sum_i A^{(i)2}\right\|_2$. Furthermore,

$$\mathbb{E}\left[\left\|\sum_i X^{(i)}\right\|_2\right] \leq \sqrt{2\pi} d\sigma. \tag{28}$$

Proof. (27) is shown in Theorem 1.3 of Tropp (2012). (28) can be derived by integrating both side of (27) on $t \in [0, \infty)$. □

Let $\mathbb{L} = \left\{\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(n)}\right\}$ be the set of votes of the labelling functions over n data points. We now apply the matrix Hoeffding's inequality to demonstrate the concentration of the sample covariance matrix $\hat{\Sigma}$ as well as the sample second moment matrix $\mathbb{E}_{\mathbb{L}}[\lambda\lambda^\top]$. The results are summarized as Lemma 3 and Lemma 6 as follows.

Lemma 3 (Concentration of the Sample Covariance Matrix)**.** *The sample covariance matrix $\hat{\Sigma}$ of p LFs of a dataset of n data points concentrates around the population level covariance matrix Σ^* as follows:*

$$\mathbb{E}\|\hat{\Sigma} - \Sigma^*\|_2 \leq \frac{2\sqrt{2\pi}p^2}{\sqrt{n}}.$$

Proof. Let $\mathbb{E}_{\mathbb{L}}[\lambda] = \frac{1}{n}\sum_{i=1}^n \lambda^{(i)}$ and let $w^{(i)} = \lambda^{(i)} - \mathbb{E}_{\mathbb{L}}[\lambda]$. The sample covariance matrix is given as:

$$\begin{aligned}
\hat{\Sigma} =& \mathbb{E}_{\mathbb{L}}(\lambda - \mathbb{E}_{\mathbb{L}}[\lambda])(\lambda - \mathbb{E}_{\mathbb{L}}[\lambda])^\top \\
=& \frac{1}{n}\sum_{i=1}^n (\lambda^{(i)} - \mathbb{E}_{\mathbb{L}}[\lambda])(\lambda^{(i)} - \mathbb{E}_{\mathbb{L}}[\lambda])^\top \\
=& \frac{1}{n}\sum_{i=1}^n w^{(i)} w^{(i)\top}.
\end{aligned}$$

To apply Lemma 2, we set $X^{(i)} = \frac{1}{n}(w^{(i)}w^{(i)\top} - \Sigma^*)$. Apparently, $\mathbb{E}[X^{(i)}] = 0$. $X^{(i)}$'s are also symmetric and statistically independent of each other. Next, we determine $A^{(i)2}$'s. We make use of the relationship that $\|v_2\|_2^2 I \succeq v_2 v_2^\top$ because for all v_1, we have that $v_1^\top(\|v_2\|_2^2 I - v_2 v_2^\top)v_1 = \|v_2\|_2^2\|v_1\|_2^2 - \|v_1^\top v_2\|_2^2 \geq 0$ by Cauchy-Schwartz inequality. Furthermore, because $\Sigma^* \succeq 0$ and $w^{(i)}w^{(i)\top} \succeq 0$, we have that $(\Sigma^* + w^{(i)}w^{(i)\top})^2 \succeq 0$. Also,

$$\begin{aligned}
(nX^{(i)})^2 =& (\Sigma^* - w^{(i)}w^{(i)})^2 \\
\preceq& (\Sigma^* - w^{(i)}w^{(i)})^2 + (\Sigma^* + w^{(i)}w^{(i)})^2 \\
=& 2((w^{(i)}w^{(i)})^2 + \Sigma^{*2}) \\
\preceq& 2(p^2 I + \Sigma^{*2}),
\end{aligned}$$

where in the last step, we have used the fact that $w^{(i)} \in [-1, 1]^p$ and

$$p^2 I \succeq \|w^{(i)}\|_2^4 I = \|w^{(i)}\|_2^2(\|w^{(i)}\|_2^2 I) \succeq \|w^{(i)}\|_2^2 w^{(i)}w^{(i)\top} = (w^{(i)}w^{(i)\top})^2.$$

As a result, we can set

$$A^{(i)2} = \frac{2}{n^2}(p^2 I + \Sigma^2).$$

It remains to compute σ^2:

$$\sigma^2 = \left\|\sum_i^n A^{(i)2}\right\|_2 \leq \sum_{i=1}^n \left\|A^{(i)2}\right\|_2 \leq \frac{2}{n^2}\sum_{i=1}^n \left(\|p^2 I\|_2 + \|\Sigma^2\|_2\right) \leq \frac{2}{n^2}\sum_{i=1}^n \left(p^2 + p^2\right) = \frac{4p^2}{n}.$$

Plug our choice of $X^{(i)}$'s and σ^2 into Lemma 2 yields the desired bound. □

Lemma 4 (Concentration of Second Moments)**.** *The sample second moment matrix $\mathbb{E}_{\mathbb{L}}[\lambda\lambda^\top]$ of the LFs of a dataset of n data points concentrates around population level second moment matrix $\mathbb{E}[\lambda\lambda^\top]$ as follows:*

$$\mathbb{E}[\left\|\mathbb{E}_{\mathbb{L}}[\lambda\lambda^\top] - \mathbb{E}[\lambda\lambda^\top]\right\|_2] \leq \frac{2\sqrt{2\pi}p^2}{\sqrt{n}}.$$

Proof. The proof is very similar to the proof of Lemma 3. The sample second moment matrix can be written as:

$$\mathbb{E}_{\mathbb{L}}[\lambda\lambda^\top] = \frac{1}{n}\sum_{i=1}^n \lambda^{(i)}\lambda^{(i)\top}.$$

To apply Lemma 2, we set $X^{(i)} = \frac{1}{n}(\lambda^{(i)}\lambda^{(i)\top} - \mathbb{E}[\lambda\lambda^\top])$. Apparently, $\mathbb{E}[X^{(i)}] = 0$. $X^{(i)}$'s are also symmetric and statistically independent of each other. Next, we determine $A^{(i)2}$'s. To this end, consider:

$$\begin{aligned}
(nX^{(i)})^2 =& (\mathbb{E}[\lambda\lambda^\top] - \lambda^{(i)}\lambda^{(i)})^2 \\
\preceq& (\mathbb{E}[\lambda\lambda^\top] - \lambda^{(i)}\lambda^{(i)})^2 + (\mathbb{E}[\lambda\lambda^\top] + \lambda^{(i)}\lambda^{(i)})^2 \\
=& 2((\lambda^{(i)}\lambda^{(i)})^2 + \mathbb{E}[\lambda\lambda^\top]^2) \\
\preceq& 2(p^2 I + \mathbb{E}[\lambda\lambda^\top]^2),
\end{aligned}$$

where we have used the fact that $(\mathbb{E}[\lambda\lambda^\top] + \lambda^{(i)}\lambda^{(i)})^2 \succeq 0$ and in the last step, we have used the fact that $\lambda^{(i)} \in \{-1, 1\}^p$ and

$$p^2 I = \|\lambda^{(i)}\|_2^4 I = \|\lambda^{(i)}\|_2^2(\|\lambda^{(i)}\|_2^2 I) \succeq \|\lambda^{(i)}\|_2^2 \lambda^{(i)}\lambda^{(i)\top} = (\lambda^{(i)}\lambda^{(i)\top})^2.$$

As a result, we can set

$$A^{(i)2} = \frac{2}{n^2}(p^2 I + \mathbb{E}[\lambda\lambda^\top]^2).$$

It remains to compute σ^2:

$$\sigma^2 = \left\|\sum_i^n A^{(i)2}\right\|_2 \leq \sum_{i=1}^n \left\|A^{(i)2}\right\|_2 \leq \frac{2}{n^2}\sum_{i=1}^n \left(\|p^2 I\|_2 + \|\mathbb{E}[\lambda\lambda^\top]^2\|_2\right) \leq \frac{2}{n^2}\sum_{i=1}^n \left(p^2 + p^2\right) = \frac{4p^2}{n}.$$

Plug our choice of $X^{(i)}$'s and σ^2 into Lemma 2 yields the desired bound. □

Lemma 5 (Concentration of the Mean Vector)**.** *The sample mean vector $\hat{\mu}_+$ concentrates around its population level counterpart μ_+ as follows:*

$$\mathbb{E}[\|\hat{\mu}_+ - \mu_+^*\|_\infty] \leq \frac{4\sqrt{2\pi}p}{n}.$$

Proof. To show the concentration of $\hat{\mu}_+$ around μ_+, our approach is to consider a diagonal matrix whose entries are $\hat{\mu}_+$. We then can use matrix concentration bound to analyze the concentration of $\hat{\mu}_+$. In detail, we let $D(\cdot)$ represents the construction of a diagonal matrix such that for a $p \times 1$ vector a, we have that

$$D(a)_{jk} = \begin{cases} a_j, & j = k; \\ 0, & j \neq k; \end{cases}. \tag{29}$$

We also note that $\hat{\mu}_+ = \frac{1}{n}\sum_{i=1}^n \lambda^{(i)}$. To apply Lemma 2, we construct $X^{(i)} = \frac{1}{n}\left(D(\lambda^{(i)}) - D(\mu_+^*)\right)$. In this way, $\mathbb{E}[X^{(i)}] = 0$ and $X^{(i)}$'s are also symmetric and statically independent of each other. Next, we determine $A^{(i)2}$'s. To this end, consider

$$(nX^{(i)})^2 = (D(\lambda^{(i)}) - D(\mu_+^*))^2 \preceq 4I,$$

where we have used the fact that $\lambda^{(i)} \in \{-1, 1\}^p$, $\mu_+^* \in [-1, 1]^p$, and $X^{(i)}$ is a diagonal matrix. Therefore, we can choose $A^{(i)2} = \frac{4}{n^2}I$. It remains to determine σ^2:

$$\sigma^2 = \left\|\sum_{i=1}^n A^{(i)2}\right\|_2 \leq \sum_{i=1}^n \|A^{(i)2}\|_2 = \frac{4}{n^2}\sum_{i=1}^n \|I\|_2 = \frac{4}{n}.$$

Plug our choice of $X^{(i)}$'s and σ^2 into Lemma 2 yields

$$\mathbb{E}[\|\hat{\mu}_+ - \mu_+^*\|_\infty] = \mathbb{E}[\left\|D(\hat{\mu}_+) - D(\mu_+^*)\right\|_2] \leq \frac{4\sqrt{2\pi}p}{n}.$$

□

Let τ be an $m \times 1$ vector that is constructed by stacking up the entries in the population covariance tensor T_{jkl}'s, where $1 \leq j < k < l \leq p$, and λ_j, λ_k, and λ_l are conditionally independent given y according to G. Let $\hat{\tau}$ be constructed similarly from $\hat{T}$.

Lemma 6 (Concentration of the Sample Covariance Tensor)**.** *The sample covariance tensor vector $\hat{\tau}$ of the LFs of a dataset of n data points concentrates around the population level covariance tensor vector τ as follows:*

$$\mathbb{E}[\|\hat{\tau} - \tau^*\|_\infty] \leq \frac{4\sqrt{2\pi}m}{n},$$

where $m = \frac{1}{6}p(p-1)(p-2)$.

Proof. It should be noticed that the sample covariance tensor

$$\hat{T} = \frac{1}{n}\sum_{i=1}^n \left(\lambda^{(i)} - \mathbb{E}_{\mathbb{L}}[\lambda]\right) \otimes \left(\lambda^{(i)} - \mathbb{E}_{\mathbb{L}}[\lambda]\right) \otimes \left(\lambda^{(i)} - \mathbb{E}_{\mathbb{L}}[\lambda]\right) \tag{30}$$

can be viewed as a summation of third-order, symmetric, rank-one, independent, random tensor. In principle, we can make use of tensor concentration inequality to bound the 2-norm between $\hat{T}$ and T^*. However, optimal concentration inequalities for symmetric rank-one tensors are still an open question in statistics (Vershynin, 2020; Even and Massoulié, 2021), and addressing this open question is beyond the scope of this paper. Instead of tackling the covariance tensors directly, our proof resorts to bounding the difference between their vectorization $\|\hat{\tau} - \tau^*\|_2$ through the matrix Hoeffding's inequality, using arguments similar to those presented in Lemma 5. Specifically, from (30), let $w^{(i)} = \lambda^{(i)} - \mathbb{E}_{\mathbb{L}}[\lambda]$, for all $j' \in \{1, 2, \cdots, m\}$, where $m = \frac{1}{6}p(p-1)(p-2)$, we can write $\hat{\tau}_{j'} = \frac{1}{n}\sum_{i=1}^n \hat{\phi}_{j'}^{(i)}$, where $\hat{\phi}_{j'}^{(i)} = \hat{w}_j^{(i)}\hat{w}_k^{(i)}\hat{w}_l^{(i)}$ for some (j, k, l) associated with j'. As a result, we can construct

$$X^{(i)} = \frac{1}{n}\left(D(\hat{\phi}^{(i)}) - D(\tau^*)\right),$$

with $\mathbb{E}[X^{(i)}] = 0$, $X^{(i)}$ are all symmetric and statistically independent of each other, and $D(\cdot)$ constructs a diagonal matrix as specified in (29). We determine $A^{(i)2}$ next. For this purpose, we consider

$$(nX^{(i)})^2 = \left(D(\hat{\phi}^{(i)}) - D(\tau^*)\right)^2 \preceq 4I,$$

where we have used the fact that $\phi^{(i)} \in [-1, 1]^m$, $\tau^* \in [-1, 1]^m$, and $X^{(i)}$ is a diagonal matrix. Therefore, we can choose $A^{(i)2} = \frac{4}{n^2}I$ and $\sigma^2 = \frac{4}{n}$, which can be computed in the exact same way as in the proof of Lemma 5. Plug our choice of $X^{(i)}$'s and σ^2 into Lemma 2 yields

$$\mathbb{E}[\|\hat{\tau} - \tau^*\|_\infty] = \mathbb{E}[\|D(\hat{\tau}) - D(\tau^*)\|_2] \leq \frac{4\sqrt{2\pi}m}{n}.$$

□

C.2 Parameter Estimation with Unknown Class Balance

In this section, we provide theoretical characterization of the parameter estimation error achieved by Firebolt. We start with the assumptions that we make to establish theoretical guarantees (Appendix C.2.1). We then discuss identifiability of the parameters (Appendix C.2.2). Next, we decompose the parameter estimation error into a sampling error term and a model misspecification error term in our proof (Appendix C.2.3). This allows us to analyze the two error terms in turn to complete the proof (Appendix C.2.4 and Appendix C.2.5).

Notation By default when a vector is given, we consider it as a column vector. We use $\sigma_{\min}(A)$ to denote the smallest singular value of A and hence $\sigma_{\min}^{-1}(A) = \frac{1}{\sigma_{\min}(A)} = \|A^\dagger\|_2$. We use $\hat{G}$ to represent a dependency graph among the labeling functions that one has access to in practice and we use G^* to represent the ground truth dependency graph. We use the hat notation (e.g. $\hat{l}$) to represent a sample-level quantity that is associated with (or learned from) $\hat{G}$. We use the tilde notation to represent a sample-level quantity (e.g. $\tilde{l}$) that is associated with (or learned from) G^*. We use the star notation to represent a population-level quantity (e.g. l^*) that is associated with (or learned from) G^*.

C.2.1 Assumptions

We present the following three assumptions made in our proof and comment on whether they are hard to meet in practice.

First, we assume that all the LFs are are better than random guessing. This is similar to the assumption that we have full knowledge about whether the LFs are better than random guessing or not, a standard assumption made in analyzing weak supervision algorithms (Fu et al., 2020). Our assumption is without loss of generality in comparison. This is because in practice, if we speculate an LF is worse than random guessing, the preferable option is to improve it instead of directly accounting for it in the label model. If we indeed want to include the information provided by such an LF, we can flip the decision made by it so that it is better than random guessing. Fu et al. (2020) also points out that if we have access to G^*, we can make use of the conditional independence relationships encoded in G^* to determine if the labeling functions are better than random guessing or not. In reality, when we do not have access to G^*, we can estimate it through structure learning (Bach et al., 2017; Varma et al., 2019). These procedures together may allow robustness against up to half of the labeling functions being worse than random guessing. Firebolt can benefit from all these procedures to meet this assumption in practice.

Secondly, we assume that we are dealing with an imbalanced classification problem ($\mu_{00}^* < 0$) where the minority class is encoded as positive. For theoretical results when the class balance is known (e.g. balanced classification $\mu_{00}^* = 0$), see Appendix C.3. In reality, practitioners typically have a good sense about whether the dataset in question is a balanced dataset or not. They can also distinguish between minority class and majority class easily.

Thirdly, we assume that the covariance entries are bounded away from zero. That is, there exists a constant $\omega_{\min} > 0$ such that $\min_{(j,k)\in\{1,2,\cdots,p\}^2} |\Sigma_{jk}^*| \geq \omega_{\min}$ and that $\min_{(j,k,l)\in\{1,2,\cdots,p\}^3} |T_{jkl}^*| \geq \omega_{\min}$. Furthermore, we assume that there are enough samples $n > n_0$ for some n_0 such that $\text{sign}(\hat{\Sigma}) = \text{sign}(\Sigma)$ and $\text{sign}(\hat{T}) = \text{sign}(T)$, where $\text{sign}(\cdot)$ is the sign function that is applied entry-wise, and that $\min_{(j,k)\in\{1,2,\cdots,p\}^2} |\hat{\Sigma}_{jk}| \geq \omega_{\min}$ and $\min_{(j,k,l)\in\{1,2,\cdots,p\}^3} |\hat{T}_{jkl}| \geq \omega_{\min}$. Finally, we assume that $\Sigma_{00}^* > \omega_{\min}$ and $\tilde{\Sigma}_{00} > \omega_{\min}$. Note that the samples here are unlabeled samples, which are abundant in a weak supervision setting. Therefore, it is reasonable to assume having enough samples in practice to meet this assumption.

C.2.2 Identifiability

We would like to establish identifiability of the label model parameters. For this purpose, it is sufficient to establish identifiability of l^* given q^*. Such a result is relatively straightforward to derive as we only need to study the theoretical guarantee of the linear system (5). Obviously, as long as the ground truth value of q^* is given and finite and M^* is full-rank, then l^* can be identified by solving the system of linear equations in (5). Because q^* needs to be finite, both T_{jkl}^*'s and Σ_{jk}'s contributing to q^* need to be larger than zero. These constraints can be met by the first two assumptions described in Appendix C.2.1.

Next, we discuss how $\hat{l}$ learned from data may differ from the ground truth l^*. Our analysis considers distortion due to both model misspecification and sampling noise. On the one hand, in terms of model misspecification, we may start from a graph $\hat{G}$ that is different from the ground truth graph G^*. Compared to G^*, two types of model

misspecification are possible: there are redundant edges in $\hat{G}$ or there are missing edges in $\hat{G}$. Let us denote the incidence matrix of the inverse graph of $\hat{G}$ as $\hat{M}$. Inheriting the assumption from our discussion of identifiability that M^* is full-rank, a few redundant edges in $\hat{G}$ will still result in a full-rank $\hat{M}$. However, the situation is more intricate when we are dealing with missing edges as there are additional rows in $\hat{M}$ that are not present in M^*. Therefore, in what follows, without loss of generality, we shall analyze the impact of model misspecification when there are missing edges in $\hat{G}$. To this end, one may define a 0-1 selection matrix S such that $S\hat{M} = M^*$. Define $\hat{q}$ as the statistics used to learn $\hat{l}$. Note that under model misspecification, the dimension of $\hat{q}$ might not necessarily be equal to the dimension of q^*. Furthermore, we define $\tilde{l}$ as the estimation produced with sampling noise but without model misspecification, and correspondingly $\tilde{q}$ the statistics used to produce $\tilde{l}$. In this way, $\tilde{q}$ is of the same dimension with q^*.

C.2.3 Non-Asymptotic Characterization of Parameter Estimation Error

In this section, we provide and prove Theorem 3, which is an extended version of Theorem 1 that accounts for both the sampling error and the model misspecification error due to the difference between $\hat{G}$ and G^* during the parameter learning of a Firebolt label model.

Theorem 3. *Under the assumptions made in Appendix C.2.1, the expected mean parameter estimation error of Firebolt learned from n unlabeled data points, p labeling functions, and G^* for an imbalanced classification problem can be upper bounded by:*

$$\mathbb{E}[\|\hat{\mu} - \mu^*\|_2] \leq \frac{233}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*) + 1)\frac{p^5}{\sqrt{n}} + \frac{18p^{\frac{7}{2}}}{\omega_{\min}}\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s + \sigma_{\min}^{-1}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^{\top}\right\|_2 |\log \omega_{\min}|,$$

where S is a selection matrix such that $S\hat{M} = M^*$, s is the number of missing edges and triplets in $\hat{G}$ compared to G^*, $U_{\hat{M}}$ is the left unitary matrix of the singular value decomposition of $\hat{M}$, $U_{\hat{M}}^{\perp}$ is the orthogonal complement of $U_{\hat{M}}$, and $u_{\max}$ is the largest norm among the rows in $U_{\hat{M}}$.

Proof. Let $\hat{\mu}$ be the estimated mean parameters provided by Firebolt with $\hat{G}$ and $\mathbb{L}$ and let μ^* be the ground truth mean parameters (associated with G^*). Our goal is to upper bound $\mathbb{E}[\|\hat{\mu} - \mu^*\|_2]$. To this end, observe that

$$\mathbb{E}[\|\hat{\mu} - \mu^*\|_2] \leq \mathbb{E}[\|\hat{\mu}_{00} - \mu_{00}^*\|_2] + \mathbb{E}[\|\hat{\mu}_{0+} - \mu_{0+}^*\|_2] + \mathbb{E}[\|\hat{\mu}_{+} - \mu_{+}^*\|_2] + \mathbb{E}[\|\hat{\mu}_{++} - \mu_{++}^*\|_2], \tag{31}$$

where we have used the fact that $\sqrt{a^2 + b^2} \leq \sqrt{a^2 + b^2 + 2ab} \leq a + b$, for a, $b \geq 0$. It suffices to bound the four terms on the right hand side of (31) in turn. Since $\hat{\mu}_+$ and $\hat{\mu}_{++}$ are directly observed, we can bound the last two terms in (31) through Lemma 5 and Lemma 4, respectively. It remains to bound the first two terms. We first bounds $\mathbb{E}[\|\hat{\mu}_{0+} - \mu_{0+}^*\|_2]$. Recall that $\tilde{\mu}_{0+}$ is the estimator produced by Firebolt using G^* and $\mathbb{L}$. With $\tilde{\mu}_{0+}$, we consider the following inequality:

$$\mathbb{E}[\|\hat{\mu}_{0+} - \mu_{0+}^*\|_2] \leq \mathbb{E}[\|\hat{\mu}_{0+} - \tilde{\mu}_{0+}\|_2] + \mathbb{E}[\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_2].$$

As a result, it suffices to bound $\mathbb{E}[\|\hat{\mu}_{0+} - \tilde{\mu}_{0+}\|_2]$ and $\mathbb{E}[\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_2]$ respectively, where the first term can be viewed as error introduced by model misspecification while the second term can be viewed as sampling error when the ground truth graph structure G^* is given. Similarly for $\mathbb{E}[\|\hat{\mu}_{00} - \mu_{00}^*\|_2]$, we have that

$$\mathbb{E}[\|\hat{\mu}_{00} - \mu_{00}^*\|_2] \leq \mathbb{E}[\|\hat{\mu}_{00} - \tilde{\mu}_{00}\|_2] + \mathbb{E}[\|\tilde{\mu}_{0+} - \mu_{00}^*\|_2].$$

The bounds for $\mathbb{E}[\|\tilde{\mu}_{00} - \mu_{00}^*\|_2]$ and $\mathbb{E}[\|\hat{\mu}_{0+} - \tilde{\mu}_{0+}\|_2]$ are given in Lemma 7 while the bounds for $\mathbb{E}[\|\hat{\mu}_{00} - \tilde{\mu}_{00}\|_2]$ and $\mathbb{E}[\|\hat{\mu}_{0+} - \tilde{\mu}_{0+}\|_2]$ are given in Lemma 8.

We have described all the ingredients needed in order to construct our bounds. In what follows, we will first bound the parameter estimation error of Firebolt when we know the ground truth dependency graph G^*, we then derive the bound under $\hat{G}$.

When G^* is used as $\hat{G}$, we have that $\hat{\mu} = \tilde{\mu}$. As a result,

$$\begin{aligned}\mathbb{E}[\|\hat{\mu} - \mu^*\|_2] \leq& \mathbb{E}[\|\tilde{\mu}_{00} - \mu_{00}^*\|_2] + \mathbb{E}[\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_2] + \mathbb{E}[\|\hat{\mu}_{+} - \mu_{+}^*\|_2] + \mathbb{E}[\|\hat{\mu}_{++} - \mu_{++}^*\|_2] \\ \leq& \frac{8\sigma_{\min}^{-1}(M^*)}{\omega_{\min}}\frac{p^{\frac{9}{2}}}{\sqrt{n}} + \frac{208}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*) + 1)\frac{p^5}{\sqrt{n}} + \frac{4\sqrt{2\pi}p^{\frac{3}{2}}}{n} + \frac{2\sqrt{2\pi}p^{\frac{5}{2}}}{\sqrt{n}}\end{aligned}$$

$$\leq \frac{8}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*)+1)\frac{p^5}{\sqrt{n}} + \frac{208}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*)+1)\frac{p^5}{\sqrt{n}} + \frac{4\sqrt{2\pi}}{\omega_{\min}^2}\frac{p^5}{\sqrt{n}} + \frac{2\sqrt{2\pi}}{\omega_{\min}^2}\frac{p^5}{\sqrt{n}}$$
$$\leq \frac{233}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*)+1)\frac{p^5}{\sqrt{n}}.$$

This corresponds to Theorem 1. On the other hand when $\hat{G}$ is used, we have that

$$\begin{aligned}\mathbb{E}[\|\hat{\mu}-\mu^*\|_2] \leq& \mathbb{E}[\|\hat{\mu}-\tilde{\mu}\|_2] + \mathbb{E}[\|\tilde{\mu}-\mu^*\|_2] \\ \leq& \mathbb{E}[\|\tilde{\mu}-\mu^*\|_2] + \mathbb{E}[\|\hat{\mu}_{00}-\tilde{\mu}_{00}\|_2] + \mathbb{E}[\|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2] \\ \leq& \frac{233}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*)+1)\frac{p^5}{\sqrt{n}} + \frac{18p^{\frac{7}{2}}}{\omega_{\min}}\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s + \sigma_{\min}^{-1}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^\top\right\|_2 |\log\omega_{\min}|.\end{aligned}$$

□

C.2.4 Analyzing the Sampling Error

Lemma 7. *$\mathbb{E}[\|\tilde{\mu}_{00}-\mu_{00}^*\|_2]$ and $\mathbb{E}[\|\tilde{\mu}_{0+}-\mu_{0+}^*\|_2]$ can be upper bounded as follows:*

$$\mathbb{E}[\|\tilde{\mu}_{00}-\mu_{00}^*\|_2] \leq \frac{8\sigma_{\min}^{-1}(M^*)}{\omega_{\min}}\frac{p^{\frac{9}{2}}}{\sqrt{n}},$$
$$\mathbb{E}[\|\tilde{\mu}_{0+}-\mu_{0+}^*\|_2] \leq \frac{208}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*)+1)\frac{p^5}{\sqrt{n}}.$$

Proof. We briefly describe the roadmap of our proof. Loosely speaking, we can bound $\mathbb{E}[\|\tilde{\mu}_{0+}-\mu_{0+}^*\|_2]$ in the following four steps: for all $j \in \{1,2,\cdots,p\}$, (1) we show that $\|\tilde{\mu}_{0j}-\mu_{0j}^*\|_2$ can be upper bounded by the difference in balanced accuracy $\|\tilde{\pi}_j-\pi_j^*\|_2$; (2) we show $\|\tilde{\pi}_j-\pi_j^*\|_2$ can be upper bounded by a quantity associated with the solution of the least squares problem $\|\tilde{t}-t^*\|_2$; (3) we show $\|\tilde{t}-t^*\|_2$ can be upper bounded by a quantity associated with the response vector of the least squares problem $\|\tilde{q}-q^*\|_2$; and finally (4) we can show $\|\tilde{q}-q^*\|_2$ can be bounded through the concentration of the covariance matrix. Meanwhile, we can bound $\mathbb{E}[\|\tilde{\mu}_{00}-\mu_{00}^*\|_2]$ in a similar, yet more simplified fashion.

Bounding $\|\tilde{\mu}_{0j}-\mu_{0j}^*\|_2$ Here we show how to upper bound $\|\tilde{\mu}_{0j}-\mu_{0j}^*\|_2$ with $\|\tilde{\pi}_j-\pi_j^*\|_2$. In detail, given λ_j, we make use of the following facts that connect sensitivity, specificity, balanced accuracy, and the mean accuracy parameter:

$$\begin{aligned}\alpha_j^{+*} =& \frac{1}{2}\left(2\pi_j^* + \mu_{00}^* - 2\pi_j^*\mu_{00}^* + \mu_{jj}^*\right), \\ \alpha_j^{-*} =& \frac{1}{2}\left(2\pi_j^* - \mu_{00}^* + 2\pi_j^*\mu_{00}^* - \mu_{jj}^*\right), \\ \pi_j^* =& \frac{1}{2}\alpha_j^{+*} + \frac{1}{2}\alpha_j^{-*}, \\ \mu_{0j}^* =& \frac{1+\mu_{00}^*}{2}\alpha_j^{+*} + \frac{1-\mu_{00}^*}{2}\alpha_j^{-*}.\end{aligned}$$

As a result, we have that

$$\mu_{0j}^* = 2\Sigma_{00}^*\pi_j^* + (\mu_{00}^*+\mu_{jj}^*)\mu_{jj}^* - 1, \tag{32}$$

where

$$\Sigma_{00}^* = 1-\mu_{00}^{*2} \tag{33}$$

represents the covariance of y. Similarly, using the fact that we know μ_{00} we have the following equation for $\tilde{\mu}_{0j}$:

$$\tilde{\mu}_{0j} = 2\tilde{\Sigma}_{00}\tilde{\pi}_j + (\tilde{\mu}_{00}+\hat{\mu}_{jj})\hat{\mu}_{jj} - 1. \tag{34}$$

(34) - (32) and using

$$\tilde{\Sigma}_{00} = \Sigma_{00}^* + \tilde{\Delta}_{00}, \quad \tilde{\mu}_{00} = \mu_{00}^* + \tilde{\delta}_{00}, \quad \tilde{\Delta}_{00} = -\tilde{\delta}_{00}^2 - 2\tilde{\delta}_{00}\mu_{00}^*, \quad \hat{\mu}_{jj} = \mu_{jj}^* + \delta_{jj},$$

we have that

$$\tilde{\mu}_{0j} - \mu^*_{0j} = 2\tilde{\Sigma}^*_{00}\tilde{\pi}^*_j - 2\Sigma_{00}\pi_j + (\tilde{\mu}_{00} + \hat{\mu}_{jj})\hat{\mu}_{jj} - (\mu^*_{00} + \mu^*_{jj})\mu^*_{jj}.$$

We would like to bound $\|\tilde{\mu}_{0j} - \mu^*_{0j}\|_2$. To do so, we make use of

$$\|\tilde{\mu}_{0j} - \mu^*_{0j}\|_2 \leq 2\|\tilde{\Sigma}_{00}\tilde{\pi}_j - \Sigma^*_{00}\pi^*_j\|_2 + \|(\tilde{\mu}_{00} + \hat{\mu}_{jj})\hat{\mu}_{jj} - (\mu^*_{00} + \mu^*_{jj})\mu^*_{jj}\|_2. \tag{35}$$

Therefore, it suffices to bound the two terms on the right hand side of the aforementioned inequality respectively. On the one hand,

$$\begin{aligned}
\tilde{\Sigma}_{00}\tilde{\pi}_j - \Sigma^*_{00}\pi^*_j =& (\Sigma^*_{00} + \tilde{\Delta}_{00})\tilde{\pi}_j - \Sigma^*_{00}\pi^*_j \\
=& \Sigma^*_{00}\tilde{\pi}_j - \Sigma^*_{00}\pi^*_j + \tilde{\Delta}_{00}\tilde{\pi}_j \\
=& \Sigma^*_{00}(\tilde{\pi}_j - \pi^*_j) + \tilde{\Delta}_{00}(\tilde{\pi}_j - \pi^*_j) + \tilde{\Delta}_{00}\pi^*_j \\
=& (\Sigma^*_{00} + \tilde{\Delta}_{00})(\tilde{\pi}_j - \pi^*_j) + \tilde{\Delta}_{00}\pi^*_j.
\end{aligned}$$

As a result,

$$\begin{aligned}
\|\tilde{\Sigma}_{00}\tilde{\pi}_j - \Sigma^*_{00}\pi^*_j\|_2 =& \|(\Sigma^*_{00} + \tilde{\Delta}_{00})(\tilde{\pi}_j - \pi^*_j) + \tilde{\Delta}_{00}\pi^*_j\|_2 \\
\leq& \|(\Sigma^*_{00} + \tilde{\Delta}_{00})(\tilde{\pi}_j - \pi^*_j)\|_2 + |\tilde{\Delta}_{00}|\|\pi^*_j\|_2 \\
\leq& \Sigma^*_{00}\|\tilde{\pi}_j - \pi^*_j\|_2 + |\tilde{\Delta}_{00}|\|\tilde{\pi}_j - \pi^*_j\|_2 + |\tilde{\Delta}_{00}|\|\pi^*_j\|_2 \\
\leq& \|\tilde{\pi}_j - \pi^*_j\|_2 + 2|\tilde{\Delta}_{00}| \\
\leq& \|\tilde{\pi}_j - \pi^*_j\|_2 + 8|\tilde{\delta}_{00}|.
\end{aligned} \tag{36}$$

where for the penultimate inequality with have used the fact that $\|\tilde{\pi}_j - \pi_j\|_2^* \leq 1$, $\|\pi_j\|_2^* \leq 1$, and $0 \leq \Sigma^*_{00} \leq 1$, and for the last inequality we have used the fact

$$\begin{aligned}
|\tilde{\Delta}_{00}| =& |\tilde{\delta}^2_{00} + 2\tilde{\delta}_{00}\mu^*_{00}| \\
\leq& |\tilde{\delta}_{00}|^2 + 2|\tilde{\delta}_{00}||\mu^*_{00}| \\
\leq& 2|\tilde{\delta}_{00}|(1 + |\mu^*_{00}|) \\
\leq& 4|\tilde{\delta}_{00}|,
\end{aligned} \tag{37}$$

which is due to the fact that $|\tilde{\delta}_{00}| \leq 2$ and $|\mu^*_{00}| \leq 1$. On the other hand,

$$\begin{aligned}
(\tilde{\mu}_{00} + \hat{\mu}_{jj})\hat{\mu}_{jj} - (\mu^*_{00} + \mu^*_{jj})\mu^*_{jj} =& (\mu^*_{00} + \tilde{\delta}_{00} + \mu^*_{jj} + \delta_{jj})(\mu^*_{jj} + \delta_{jj}) - (\mu^*_{00} + \mu^*_{jj})\mu^*_{jj} \\
=& \tilde{\delta}_{00}\delta_{jj} + \tilde{\delta}_{00}\mu^*_{jj} + \delta^2_{jj} + \delta_{jj}(\mu^*_{00} + 2\mu^*_{jj}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|(\tilde{\mu}_{00} + \hat{\mu}_{jj})\hat{\mu}_{jj} - (\mu^*_{00} + \mu^*_{jj})\mu^*_{jj}\|_2 =& \|\tilde{\delta}_{00}\delta_{jj} + \tilde{\delta}_{00}\mu^*_{jj} + \delta^2_{jj} + \delta_{jj}(\mu^*_{00} + 2\mu^*_{jj})\|_2 \\
\leq& |\tilde{\delta}_{00}|\|\delta_{jj}\|_2 + |\tilde{\delta}_{00}|\|\mu^*_{jj}\|_2 + \|\delta_{jj}\|_2^2 + \|\delta_{jj}\|_2\|\mu^*_{00} + 2\mu^*_{jj}\|_2 \\
\leq& 2\|\delta_{jj}\|_2 + |\tilde{\delta}_{00}|\|\mu^*_{jj}\|_2 + 2\|\delta_{jj}\|_2 + \|\delta_{jj}\|_2\|\mu^*_{00} + 2\mu^*_{jj}\|_2 \\
=& |\tilde{\delta}_{00}|\|\mu^*_{jj}\|_2 + \|\delta_{jj}\|_2(4 + \|\mu^*_{00} + 2\mu^*_{jj}\|_2) \\
\leq& |\tilde{\delta}_{00}| + 7\|\delta_{jj}\|_2.
\end{aligned} \tag{38}$$

where we have used $|\tilde{\delta}_{00}| \leq 2$ and $\|\delta_{jj}\|_2 \leq 2$ for the last inequality. Finally, using (35), (36), and (38), we have that

$$\begin{aligned}
\|\tilde{\mu}_{0j} - \mu_{0j}\|_2 \leq& 2\|\tilde{\pi}_j - \pi^*_j\|_2 + 16|\tilde{\delta}_{00}| + |\tilde{\delta}_{00}| + 7\|\delta_{jj}\|_2 \\
=& 2\|\tilde{\pi}_j - \pi^*_j\|_2 + 17|\tilde{\delta}_{00}| + 7\|\delta_{jj}\|_2.
\end{aligned} \tag{39}$$

Bounding $\|\tilde{\pi}_j - \pi_j^*\|_2$ We define $b^* = \log \sigma_{00}^*$ and $\tilde{b} = \log \tilde{\sigma}_{00}$. We will show that $\|\tilde{\pi}_j - \pi_j^*\|_2$ can be upper bounded by $\|\tilde{t} - t^*\|_2 + \|\tilde{b} - b^*\|_2$. To this end, notice that $t_j^* = \log(2\pi_j^* - 1) + b^*$ and $\tilde{t}_j = \log(2\tilde{\pi}_j - 1) + \tilde{b}$. We consider

$$\begin{aligned}
\|\tilde{\pi}_j - \pi_j^*\|_2 \leq & \left\| \frac{\exp(\tilde{t}_j - \tilde{b}) + 1}{2} - \frac{\exp(t_j^* - b^*) + 1}{2} \right\|_2 \\
\leq & \frac{1}{2} \|\exp(\tilde{t}_j - \tilde{b}) - \exp(t_j^* - b^*)\|_2 \\
= & \frac{1}{2} \|\exp(t_j^* - b^*)(\exp((\tilde{t}_j - \tilde{b}) - (t_j^* - b^*)) - 1)\|_2 \\
\leq & \frac{1}{2} \|\exp(t_j^* - b^*)\|_2 \|\exp((\tilde{t}_j - \tilde{b}) - (t_j^* - b^*)) - 1\|_2 \\
\leq & \frac{1}{2} \|2\pi_j - 1\|_2 \|\exp((\tilde{t}_j - \tilde{b}) - (t_j^* - b^*)) - 1\|_2 \\
\leq & \frac{1}{2} \|\exp((\tilde{t}_j - \tilde{b}) - (t_j^* - b^*)) - 1\|_2.
\end{aligned}$$

Using the fact that for $x \leq 1 \Rightarrow \exp(x) - 1 \leq 2x$, we have that

$$\begin{aligned}
\|\tilde{\pi}_j - \pi_j^*\|_2 \leq & \|(\tilde{t}_j - \tilde{b}) - (t_j^* - b^*)\|_2 \\
\leq & \|\tilde{b} - b^*\|_2 + \|\tilde{t}_j - t_j^*\|_2 \\
\leq & \frac{1}{\Sigma_{00}^*} \|\tilde{t}_0 - t_0^*\|_2 + \|\tilde{t}_j - t_j^*\|_2 \\
\leq & \frac{2}{\Sigma_{00}^*} \|\tilde{t} - t^*\|_\infty,
\end{aligned} \tag{40}$$

where the first inequality trivially holds when $\|(\tilde{t}_j - \tilde{b}) - (t_j^* - b^*)\|_2 > 1$ because $\|\tilde{\pi}_j - \pi_j^*\|_2 \leq 1$, and the penultimate inequality is due to the bound for $\|\tilde{b} - b^*\|_2$ and $|\tilde{\delta}_{00}|$ that we present as follows:

Bounding $\|\tilde{b} - b^*\|_2$ To bound $\|\tilde{b} - b^*\|_2$, notice that

$$\begin{aligned}
\|\tilde{b} - b\|_2 = & \|\log \tilde{\sigma}_{00} - \log \sigma_{00^*}\|_2 \\
= & \frac{1}{2} \|\log \tilde{\sigma}_{00}^2 - \log \sigma_{00}^{*2}\|_2 \\
= & \frac{1}{2} \|\log \tilde{\Sigma}_{00} - \log \Sigma_{00}\|_2^* \\
= & \frac{1}{2} \left\| \log(\Sigma_{00}^* + \tilde{\Delta}_{00}) - \log \Sigma_{00}^* \right\|_2 \\
= & \frac{1}{2} \left\| \log \left(1 + \frac{\tilde{\Delta}_{00}}{\Sigma_{00}^*} \right) \right\|_2 \\
\leq & \frac{1}{2} \left\| \frac{\tilde{\Delta}_{00}}{\Sigma_{00}^*} \right\|_2 \\
\leq & \frac{2}{\Sigma_{00}^*} |\tilde{\delta}_{00}|,
\end{aligned}$$

where for the penultimate inequality we have use the fact that $\log(1 + x) \leq x$ and for the last inequality we have used (37).

Bounding $|\tilde{\delta}_{00}|$ Recall that $\tilde{\delta}_{00} = \tilde{\mu}_{00} - \mu_{00}^*$, where $\tilde{\mu}_{00}^* = -\frac{\exp(\tilde{t}_0)}{\sqrt{4 + \exp(2\tilde{t}_0)}} = f(\tilde{t}_0)$, we have that

$$\begin{aligned}
|\tilde{\delta}_{00}| = & |\tilde{\mu}_{00} - \mu_{00}^*| \\
= & |f(\tilde{t}_0) - f(t_0^*)|
\end{aligned}$$

$$\leq \frac{1}{2}\|\tilde{t}_0 - t_0^*\|_2, \tag{41}$$

where for the last inequality we have used the fact that for $z \in \mathbb{R}$,

$$f'(z) = \frac{4\exp(z)}{(4+\exp(2z))^{\frac{3}{2}}} \leq \frac{4}{4^{\frac{3}{2}}} = 4^{-\frac{1}{2}} = \frac{1}{2},$$

and hence $f(z)$ is $\frac{1}{2}$-Lipschitz continuous.

Bounding $\|\tilde{t}-t^*\|_2$ Bounding $\|\tilde{t}-t^*\|_2$ with $\|\tilde{q}-q^*\|_2$ is straightforward. Indeed, because $\tilde{t} = \arg\min_l \frac{1}{2}\|M^*t - \tilde{q}\|_2^2$ and $t^* = \arg\min_t \frac{1}{2}\|M^*t - q^*\|_2^2$, we have that

$$\|\tilde{t} - t^*\|_2 = \|M^{*\dagger}\tilde{q} - M^{*\dagger}q^*\|_2 \leq \|M^{*\dagger}\|_2\|\tilde{q} - q^*\|_2, \tag{42}$$

where $M^{*\dagger}$ is the pseudo-inverse of M^*.

Bounding $\|\tilde{q} - q^*\|_2$ It should be noticed that the entries of q^* consist of both the logarithm of the entries of the two-way and three-way covariance. That is, $q_j^* = \log\omega_j^*$, where ω_j^* is an entry in T^* or Σ^*. Let $\hat{\omega}_j = \omega_j^* + \epsilon_j$, we have that

$$\begin{aligned}
\|\tilde{q}_j - q_j\|_2 &= \left\|\log\hat{\omega}_j - \log\omega_j^*\right\|_2 \\
&= \left\|\log\hat{\omega}_j - \log\omega_j^*\right\|_2 \\
&= \left\|\log(\omega_j^* + \epsilon_j) - \log\omega_j^*\right\|_2 \\
&= \left\|\log\left(1 + \frac{\epsilon_j}{\omega_j^*}\right)\right\|_2 \\
&\leq \left\|\frac{\epsilon_j}{\omega_j^*}\right\|_2 \\
&\leq \frac{1}{\omega_{\min}}\|\epsilon_j\|_2,
\end{aligned} \tag{43}$$

where we have use the fact that $\log(1+x) \leq x$ and $\omega_{\min}$ is the smallest positive entry in T^* and Σ^*.

Assembling Bounds for $\mathbb{E}[\|\tilde{\mu}_{00} - \mu_{00}^*\|_2]$ and $\mathbb{E}[\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_2]$ Putting (39), (40), and (41) together, we have that for all $j \in \{1, 2, \cdots, p\}$,

$$\begin{aligned}
\|\tilde{\mu}_{0j} - \mu_{0j}^*\|_2 &\leq 2\|\tilde{\pi}_j - \pi_j^*\|_2 + 17|\tilde{\delta}_{00}| + 7\|\delta_{jj}\|_2 \\
&\leq \frac{4}{\Sigma_{00}^*}\|\tilde{t} - t^*\|_\infty + \frac{17}{2}\|\tilde{t} - t^*\|_\infty + 7\|\delta_{jj}\|_2 \\
&\leq \frac{13}{\Sigma_{00}^*}\|\tilde{t} - t^*\|_\infty + 7\|\delta_+\|_\infty,
\end{aligned}$$

implying

$$\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_\infty \leq \frac{13}{\Sigma_{00}^*}\|\tilde{t} - t^*\|_\infty + 7\|\delta_+\|_\infty.$$

Using (42), (43), and the fact that for a $p \times 1$ vector a, $\|a\|_\infty \leq \|a\|_2 \leq \sqrt{p}\|a\|_\infty$, we further have that

$$\begin{aligned}
\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_2 &\leq \frac{13\sqrt{p}}{\Sigma_{00}^*}\|\tilde{t} - t^*\|_2 + 7\sqrt{p}\|\delta_+\|_\infty \\
&\leq \frac{13\sqrt{p}}{\Sigma_{00}^*}\|M^{*\dagger}\|_2\|\tilde{q} - q^*\|_2 + 7\sqrt{p}\|\delta_+\|_\infty \\
&\leq \frac{13\sqrt{p}}{\omega_{\min}^2}\|M^{*\dagger}\|_2\|\epsilon\|_2 + 7\sqrt{p}\|\delta_+\|_\infty
\end{aligned}$$

$$\leq \frac{13\sqrt{p}}{\omega_{\min}^2}\|M^{*\dagger}\|_2\sqrt{\|\hat{\tau}-\tau\|_2^2+\|\hat{\Sigma}-\Sigma\|_{\mathrm{F}}^2}+7\sqrt{p}\|\delta_+\|_\infty$$
$$\leq \frac{13\sqrt{p}}{\omega_{\min}^2}\|M^{*\dagger}\|_2(\|\hat{\tau}-\tau^*\|_2+\|\hat{\Sigma}-\Sigma^*\|_F)+7\sqrt{p}\|\hat{\mu}_+-\mu_+^*\|_\infty,$$

where for the last inequality we have used the fact that $\sqrt{a^2+b^2}\leq\sqrt{a^2+b^2+2ab}=a+b$ for $a,b\geq 0$. Taking the expectation on both side ends of the foregoing inequality and applying Lemma 3, 5, and 6, we have that

$$\begin{aligned}\mathbb{E}[\|\tilde{\mu}_{0+}-\mu_{0+}^*\|_2] \leq& \frac{13\sqrt{p}}{\omega_{\min}^2}\|M^{*\dagger}\|_2\left(\frac{4\sqrt{2\pi}m^{\frac{3}{2}}}{n}+\frac{2\sqrt{2\pi}p^{\frac{5}{2}}}{\sqrt{n}}\right)+\frac{28\sqrt{2\pi}p^{\frac{3}{2}}}{n}\\ \leq& \frac{13\sqrt{p}}{\omega_{\min}^2}\|M^{*\dagger}\|_2\frac{16p^{\frac{9}{2}}}{\sqrt{n}}+\frac{71p^{\frac{9}{2}}}{\sqrt{n}}\\ \leq& \frac{208}{\omega_{\min}^2}\|M^{*\dagger}\|_2\frac{p^5}{\sqrt{n}}+\frac{208p^{\frac{9}{2}}}{\sqrt{n}}\\ \leq& \frac{208}{\omega_{\min}^2}(\sigma_{\min}^{-1}(M^*)+1)\frac{p^5}{\sqrt{n}},\end{aligned}$$

where we have also used the fact that for a symmetric matrix A, $\|A\|_2\leq\sqrt{p}\|A\|_{\mathrm{F}}$. Finally, we would like to bound $\mathbb{E}[\|\tilde{\mu}_{00}-\mu_{00}^*\|_2]$. To this end, we use (41), (42), and (43) to yield

$$\begin{aligned}\|\tilde{\mu}_{00}-\mu_{00}^*\|_2 \leq& \frac{1}{2}\|\tilde{t}_0-t_0^*\|_2\\ \leq& \frac{1}{2}\|\tilde{t}-t^*\|_2\\ \leq& \frac{1}{2}\|M^{*\dagger}\|_2\|\tilde{q}-q^*\|_2\\ \leq& \frac{1}{2\omega_{\min}}\|M^{*\dagger}\|_2(\|\hat{\tau}-\tau^*\|_2+\|\hat{\Sigma}-\Sigma^*\|_F).\end{aligned}$$

Taking expectation on both end of the foregoing inequality and applying Lemma 3 and 6 yields

$$\begin{aligned}\mathbb{E}[\|\tilde{\mu}_{00}-\mu_{00}^*\|_2] \leq& \frac{\|M^{*\dagger}\|_2}{2\omega_{\min}}\left(\frac{4\sqrt{2\pi}m^{\frac{3}{2}}}{n}+\frac{2\sqrt{2\pi}p^{\frac{5}{2}}}{\sqrt{n}}\right)\\ \leq& \frac{6\sqrt{2\pi}\|M^{*\dagger}\|_2}{2\omega_{\min}}\frac{p^{\frac{9}{2}}}{\sqrt{n}}\\ \leq& \frac{8\sigma_{\min}^{-1}(M^*)}{\omega_{\min}}\frac{p^{\frac{9}{2}}}{\sqrt{n}}.\end{aligned}$$

□

C.2.5 Analyzing the Model Misspecification Error

Lemma 8. *$\mathbb{E}[\|\hat{\mu}_{00}-\tilde{\mu}_{00}\|_2]$ and $\mathbb{E}[\|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2]$ can be bounded as follows.*

$$\mathbb{E}[\|\hat{\mu}_{00}-\tilde{\mu}_{00}\|_2] \leq \frac{p^3}{2}\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s+\sigma_{\min}^{-1}(SU_{\hat{M}})-\sigma_{\min}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^\top\right\|_2|\log\omega_{\min}|,$$
$$\mathbb{E}[\|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2] \leq \frac{9p^{\frac{7}{2}}}{\omega_{\min}}\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s+\sigma_{\min}^{-1}(SU_{\hat{M}})-\sigma_{\min}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^\top\right\|_2|\log\omega_{\min}|,$$

where S is a selection matrix such that $S\hat{M}=M^*$, s is the number of missing edges and triplets in $\hat{G}$ compared to G^*, $U_{\hat{M}}$ is the left unitary matrix of the singular value decomposition of $\hat{M}$, $U_{\hat{M}}^{\perp}$ is the orthogonal complement of $U_{\hat{M}}$, and $u_{\max}$ is the largest norm among the rows in $U_{\hat{M}}$.

Proof. In Appendix C.2.4, we have provided an upper bound for $\mathbb{E}[\|\tilde{\mu}_{0+}-\mu_{0+}^*\|_2]$. Next, we show that $\mathbb{E}[\|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2]$ can be upper bounded similarly in the following three steps: for all $j\in\{1,2,\cdots,p\}$, (1) we show that

$\|\hat{\mu}_{0j} - \tilde{\mu}_{0j}\|_2$ can be upper bounded by a quantity associated with the balance accuracy $\|\hat{\pi}_j - \tilde{\pi}_j\|_2$; (2) we show that $\|\hat{\pi}_j - \tilde{\pi}_j\|_2$ can be upper bounded by a quantity associated with the balance accuracy $\|\hat{t} - \tilde{t}\|_2$; (3) we show that $\|\hat{t} - \tilde{t}\|_2$ can be characterized by terms associated with the number of missing edges of $\hat{G}$ and the residuals of the least squares. Meanwhile, we can bound $\mathbb{E}[\|\hat{\mu}_{00} - \tilde{\mu}_{00}\|_2]$ in a similar, yet more simplified fashion.

Bounding $\|\hat{\mu}_{0j} - \tilde{\mu}_{0j}\|_2$ We describe how to upper bound $\|\hat{\mu}_{0j} - \tilde{\mu}_{0j}\|_2$ with $\|\hat{\pi}_j - \tilde{\pi}_j\|_2$. Similar to (32) and (34), we have that

$$\hat{\mu}_{0j} = 2\hat{\Sigma}_{00}\hat{\pi}_j + (\hat{\mu}_{00} + \hat{\mu}_{jj})\hat{\mu}_{jj} - 1. \tag{44}$$

(44) - (34) and using

$$\hat{\Sigma}_{00} = \tilde{\Sigma}_{00} + \Delta_{00}, \quad \hat{\mu}_{00} = \tilde{\mu}_{00} + \delta_{00}, \quad \Delta_{00} = -\delta_{00}^2 - 2\delta_{00}\tilde{\mu}_{00}, \tag{45}$$

we have that

$$\begin{aligned}\hat{\mu}_{0j} - \tilde{\mu}_{0j} =& 2\hat{\Sigma}_{00}\hat{\pi}_j - 2\tilde{\Sigma}_{00}\tilde{\pi}_j + \hat{\mu}_{00}\hat{\mu}_{jj} - \tilde{\mu}_{00}\hat{\mu}_{jj} \\ =& 2\tilde{\Sigma}_{00}(\hat{\pi}_j - \tilde{\pi}_j) + \hat{\mu}_{jj}\delta_{00} + 2\hat{\pi}_j\Delta_{00}.\end{aligned}$$

As a result,

$$\|\hat{\mu}_{0j} - \tilde{\mu}_{0j}\|_2 \le 2\|\hat{\pi}_j - \tilde{\pi}_j\|_2 + \|\delta_{00}\|_2 + 2\|\Delta_{00}\|_2. \tag{46}$$

Bounding $\|\hat{\pi}_j - \tilde{\pi}_j\|_2$ Following the same rationale to bound $\|\tilde{\pi}_j - \pi_j^*\|_2$ with $\|\tilde{t}_j - t_j^*\|_2$, we can bound $\|\hat{\pi}_j - \tilde{\pi}_j\|_2$ with $\|\hat{t}_j - \tilde{t}_j\|_2$ as:

$$\begin{aligned}\|\hat{\pi}_j - \tilde{\pi}_j\|_2 \le& \|\hat{b}_j - \tilde{b}_j\|_2 + \|\hat{t}_j - \tilde{t}_j\|_2 \\ \le& \frac{2}{\tilde{\Sigma}_{00}}|\delta_{00}| + \|\hat{t}_j - \tilde{t}_j\|_2 \\ \le& \frac{1}{\tilde{\Sigma}_{00}}\|\hat{t}_0 - \tilde{t}_0\|_2 + \|\hat{t}_j - \tilde{t}_j\|_2 \\ \le& \frac{2}{\tilde{\Sigma}_{00}}\|\hat{t} - \tilde{t}\|_\infty \\ \le& \frac{2}{\omega_{\min}}\|\hat{t} - \tilde{t}\|_\infty\end{aligned} \tag{47}$$

where $\hat{t}_j = \log(2\hat{\pi}_j - 1) + \hat{b}$, $\hat{b} = \log\hat{\sigma}_{00}$, and $\hat{\sigma}_{00}^2 = \hat{\Sigma}_{00}$. Furthermore, for the second and the third inequality, we have used

$$\|\hat{b}_j - \tilde{b}_j\|_2 \le \frac{2}{\tilde{\Sigma}_{00}}|\delta_{00}| \quad \text{and} \quad |\delta_{00}| \le \frac{1}{2}\|\hat{t}_0 - \tilde{t}_0\|_2, \tag{48}$$

respectively. They are derived in a fashion similar to the bound for $\|\tilde{b}_j - b_j^*\|_2$ and $|\tilde{\delta}_{00}|$. Finally, for the last inequality we have used the third assumption in Appendix C.2.1.

Bounding $\|\hat{t} - \tilde{t}\|_2$ We now show how to bound $\|\hat{t} - \tilde{t}\|_2$ with quantities associated with model misspecification and residuals. Note that if $\hat{G} = G^*$, we have that $\|\hat{t} - \tilde{t}\|_2 = 0$. Otherwise, we detail our proof of bounding $\|\hat{t} - \tilde{t}\|_2$ below. Our proof follows the arguments similar to those of Drineas et al. (2006) and Kuang et al. (2020). In particular,

$$\begin{aligned}\hat{t} - \tilde{t} =& \hat{M}^\dagger\hat{q} - M^{*\dagger}\tilde{q} \\ =& \hat{M}^\dagger\hat{q} - (S\hat{M})^\dagger(S\hat{q}) \\ =& V_{\hat{M}}D_{\hat{M}}^{-1}U_{\hat{M}}^\top\hat{q} - (SU_{\hat{M}}D_{\hat{M}}V_{\hat{M}}^\top)^\dagger S\hat{q} \\ =& V_{\hat{M}}D_{\hat{M}}^{-1}U_{\hat{M}}^\top\hat{q} - V_{\hat{M}}D_{\hat{M}}^{-1}(SU_{\hat{M}})^\dagger S\hat{q} \\ =& V_{\hat{M}}D_{\hat{M}}^{-1}U_{\hat{M}}^\top\hat{q} - V_{\hat{M}}D_{\hat{M}}^{-1}(SU_{\hat{M}})^\dagger SI\hat{q} \\ =& V_{\hat{M}}D_{\hat{M}}^{-1}U_{\hat{M}}^\top\hat{q} - V_{\hat{M}}D_{\hat{M}}^{-1}(SU_{\hat{M}})^\dagger S(U_{\hat{M}}^\perp(U_{\hat{M}}^\perp)^\top + U_{\hat{M}}U_{\hat{M}}^\top)\hat{q} \\ =& V_{\hat{M}}D_{\hat{M}}^{-1}U_{\hat{M}}^\top\hat{q} - V_{\hat{M}}D_{\hat{M}}^{-1}(SU_{\hat{M}})^\dagger SU_{\hat{M}}^\perp(U_{\hat{M}}^\perp)^\top\hat{q} - V_{\hat{M}}D_{\hat{M}}^{-1}(SU_{\hat{M}})^\dagger SU_{\hat{M}}U_{\hat{M}}^\top\hat{q}\end{aligned}$$

$$
\begin{aligned}
=&V_{\hat{M}} D_{\hat{M}}^{-1} U_{\hat{M}}^\top \hat{q} - V_{\hat{M}} D_{\hat{M}}^{-1} (SU_{\hat{M}})^\dagger SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} - V_{\hat{M}} D_{\hat{M}}^{-1} U_{\hat{M}}^\top \hat{q} \\
=&- V_{\hat{M}} D_{\hat{M}}^{-1} (SU_{\hat{M}})^\dagger SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q}.
\end{aligned}
$$

For the first equality, we have used the definition of the least square solution. For the second equality, we have used the definition that $M^* = S\hat{M}$ and $\tilde{q} = S\hat{q}$. For the third equality, we have carried an SVD for $\hat{M}$ and $S\hat{M}$. For the sixth equality, we use the fact that $I = (U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top + U_{\hat{M}} U_{\hat{M}}^\top)$. Now, let $\Gamma = (SU_{\hat{M}})^\dagger - (SU_{\hat{M}})^\top$. We have that,

$$
\begin{aligned}
\hat{t} - \tilde{t} =& - V_{\hat{M}} D_{\hat{M}}^{-1} (SU_{\hat{M}})^\dagger SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \\
=& - V_{\hat{M}} D_{\hat{M}}^{-1} ((SU_{\hat{M}})^\top + \Gamma) SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q}.
\end{aligned}
$$

Furthermore, $\Gamma = (SU_{\hat{M}})^\dagger - (SU_{\hat{M}})^\top \Rightarrow \|\Gamma\|_2 = \left\| D_{SU_{\hat{M}}}^{-1} - D_{SU_{\hat{M}}} \right\|_2 = \sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}})$. As a result,

$$
\begin{aligned}
\|\hat{t} - \tilde{t}\|_2 =& \left\| V_{\hat{M}} D_{\hat{M}}^{-1} ((SU_{\hat{M}})^\top + \Gamma) SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \\
=& \left\| D_{\hat{M}}^{-1} ((SU_{\hat{M}})^\top + \Gamma) SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \\
\leq& \left\| D_{\hat{M}}^{-1} (SU_{\hat{M}})^\top SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 + \left\| D_{\hat{M}}^{-1} \Gamma SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \\
\leq& \sigma_{\min}^{-1}(\hat{M}) \left(\left\| (SU_{\hat{M}})^\top SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 + \left\| \Gamma SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \right) \\
\leq& \sigma_{\min}^{-1}(\hat{M}) \left(\left\| U_{\hat{M}}^\top S^\top SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 + \|\Gamma\|_2 \left\| SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \right) \\
=& \sigma_{\min}^{-1}(\hat{M}) \left(\left\| U_{\hat{M}}^\top (I - \Xi) U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 + (\sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}})) \left\| SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \right) \\
=& \sigma_{\min}^{-1}(\hat{M}) \left(\left\| U_{\hat{M}}^\top \Xi U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 + (\sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}})) \left\| SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \right),
\end{aligned}
$$

where we have used the fact that $S^\top S = I - \Xi$ as $S^\top S$ is a diagonal matrix with the diagonal elements corresponding to the missing edges/triplets being zero and one otherwise, and Ξ is a diagonal matrix with the diagonal elements corresponding to the missing edges/triplets being one and zero otherwise. Subsequently,

$$
\begin{aligned}
\|\hat{t} - \tilde{t}\|_2 =& \sigma_{\min}^{-1}(\hat{M}) \left(\left\| U_{\hat{M}}^\top \Xi U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 + (\sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}})) \left\| SU_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \right) \\
\leq& \sigma_{\min}^{-1}(\hat{M}) \left(\left\| U_{\hat{M}}^\top \Xi \right\|_2 \left\| U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 + \left(\sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}}) \right) \right) \left\| U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \\
=& \sigma_{\min}^{-1}(\hat{M}) \left(\left\| U_{\hat{M}}^\top \Xi \right\|_2 + \left(\sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}}) \right) \right) \left\| U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \hat{q} \right\|_2 \\
\leq& \sigma_{\min}^{-1}(\hat{M}) \left(\left\| U_{\hat{M}}^\top \Xi \right\|_{\mathrm{F}} + \sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}}) \right) \left\| U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \right\|_2 \|\hat{q}\|_2 \\
\leq& \frac{1}{6} p(p^2 - 1) \sigma_{\min}^{-1}(\hat{M}) \left(u_{\max} s + \sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}}) \right) \left\| U_{\hat{M}}^\perp (U_{\hat{M}}^\perp)^\top \right\|_2 |\log \omega_{\min}|, \qquad (49)
\end{aligned}
$$

where for the first inequality, we have used the fact that S has a maximum singular value of 1 and for the last inequality, s is the number of misspecification, $u_{\max}$ is the largest norm among the rows in $U_{\hat{M}}$, and we have used the third assumption in Appendix C.2.1 to get $\|\hat{q}\|_2 \leq \frac{1}{6} p(p^2 - 1) |\log \omega_{\min}|$.

Assembling Bounds for $\mathbb{E}[\|\hat{\mu}_{00} - \tilde{\mu}_{00}\|_2]$ and $\mathbb{E}[\|\hat{\mu}_{0+} - \tilde{\mu}_{0+}\|_2]$ Combining (46), (47), and (48), we have that for $j \in \{1, 2, \cdots, p\}$,

$$
\begin{aligned}
\|\hat{\mu}_{0j} - \tilde{\mu}_{0j}\|_2 \leq& 2\|\hat{\pi}_j - \tilde{\pi}_j\|_2 + \|\delta_{00}\|_2 + 2\|\Delta_{00}\|_2 \\
\leq& \frac{4}{\omega_{\min}} \|\hat{t} - \tilde{t}\|_\infty + 9\|\delta_{00}\|_2 \\
\leq& \frac{4}{\omega_{\min}} \|\hat{t} - \tilde{t}\|_\infty + \frac{9}{2} \|\hat{t} - \tilde{t}\|_\infty \\
\leq& \frac{9}{\omega_{\min}} \|\hat{t} - \tilde{t}\|_\infty,
\end{aligned}
$$

where for the second inequality we have used $\|\Delta_{00}\|_2 \leq 4\|\delta_{00}\|_2$ derived similarly to (37). This means

$$\begin{aligned}
\|\hat{\mu}_{0+} - \tilde{\mu}_{0+}\|_\infty \leq & \frac{9}{\omega_{\min}}\|\hat{t} - \tilde{t}\|_\infty \\
\Rightarrow \|\hat{\mu}_{0+} - \tilde{\mu}_{0+}\|_2 \leq & \frac{9\sqrt{p}}{\omega_{\min}}\|\hat{t} - \tilde{t}\|_\infty \\
\leq & \frac{9\sqrt{p}}{\omega_{\min}}\|\hat{t} - \tilde{t}\|_2 \\
\leq & \frac{9p^{\frac{7}{2}}}{\omega_{\min}}\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s + \sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^\top\right\|_2 |\log\omega_{\min}|, \qquad (50)
\end{aligned}$$

where we have used (49) for the last inequality. On the other hand, from (45), (48), and (49), we have that

$$\begin{aligned}
\|\hat{\mu}_{00} - \tilde{\mu}_{00}\|_2 \leq & \frac{1}{2}\|\hat{t}_0 - \tilde{t}_0\|_2 \\
\leq & \frac{1}{2}\|\hat{t} - \tilde{t}\|_2 \\
\leq & \frac{p^3}{2}\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s + \sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^\top\right\|_2 |\log\omega_{\min}|. \qquad (51)
\end{aligned}$$

Because given $\hat{G}$, the right hand side of (50) and (51) are both constant, taking the expectation of both side of (50) and (51) and using the fact that $\sigma_{\min}(SU_{\hat{M}}) \geq 0$ yield the result. □

C.3 Parameter Estimation with Known Class Balance

Since μ_{00}^* is known, we provide a characterization of $\mathbb{E}[\|\hat{\mu}_{0+} - \mu_{0+}^*\|_2]$. $\mathbb{E}[\|\hat{\mu}_{+} - \mu_{+}^*\|_2]$ and $\mathbb{E}[\|\hat{\mu}_{++} - \mu_{++}^*\|_2]$ can be characterized in the same way as in Theorem 3.

Theorem 4. *Under the assumptions made in Appendix C.2.1, the expected mean accuracy parameter estimation error of Firebolt learned from n unlabeled data points, p labeling functions, and $\hat{G}$ for a binary classification problem with class balance μ_{00}^* can be upper bounded by:*

$$\mathbb{E}[\|\hat{\mu}_{0+} - \mu_{0+}^*\|_2] \leq \frac{12\sqrt{2\pi}}{\omega_{\min}} \cdot (\sigma_{\min}^{-1}(M^*) + 1) \cdot \frac{p^3}{\sqrt{n}} + p^2\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s + \sigma_{\min}^{-1}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^\top\right\|_2 |\log\omega_{\min}|,$$

where S is a selection matrix such that $S\hat{M} = M^*$, s is the number of missing edges in $\hat{G}$ compared to G^*, $U_{\hat{M}}$ is the left unitary matrix of the singular value decomposition of $\hat{M}$, $U_{\hat{M}}^{\perp}$ is the orthogonal complement of $U_{\hat{M}}$, and $u_{\max}$ is the largest norm among the rows in $U_{\hat{M}}$.

Proof. The proof of Theorem 4 is similar to that of Theorem 3. Following the decomposition in (31), it is sufficient to analyze the sampling error $\mathbb{E}[\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_2]$ and the model misspecification error $\mathbb{E}[\|\hat{\mu}_{0+} - \tilde{\mu}_{0+}\|_2]$.

Bounding $\|\tilde{\mu}_{0j} - \mu_{0j}^*\|_2$ Similar to (34), we have that

$$\tilde{\mu}_{0j} = 2\Sigma_{00}^*\tilde{\pi}_j + (\mu_{00}^* + \hat{\mu}_{jj})\hat{\mu}_{jj} - 1. \qquad (52)$$

(52) - (32) and using $\hat{\mu}_{jj} = \mu_{jj}^* + \delta_{jj}$ yields,

$$\begin{aligned}
\tilde{\mu}_{0j} - \mu_{0j}^* =& 2\Sigma_{00}^*\tilde{\pi}_j - 2\Sigma_{00}^*\pi_j^* + (\mu_{00}^* + \hat{\mu}_{jj})\hat{\mu}_{jj} - (\mu_{00}^* + \mu_{jj}^*)\mu_{jj}^* \\
=& 2\Sigma_{00}^*(\tilde{\pi}_j - \pi_j^*) + \mu_{00}^*(\hat{\mu}_{jj} - \mu_{jj}^*) + \hat{\mu}_{jj}^2 - \mu_{jj}^{*2} \\
=& 2\Sigma_{00}^*(\tilde{\pi}_j - \pi_j^*) + \mu_{00}^*\delta_{jj} + (\hat{\mu}_{jj} + \mu_{jj}^*)\delta_{jj}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\tilde{\mu}_{0j} - \mu_{0j}^*\|_2 \leq & 2\Sigma_{00}^*\|\tilde{\pi}_j - \pi_j^*\|_2 + \|\mu_{00}^* + \hat{\mu}_{jj} + \mu_{jj}^*\|_2\|\delta_{jj}\|_2 \\
\leq & 2\|\tilde{\pi}_j - \pi_j^*\|_2 + 3\|\delta_{jj}\|_2. \qquad (53)
\end{aligned}$$

Bounding $\|\tilde{\pi}_j - \pi_j^*\|_2$ Here, we are interested in bounding $\|\tilde{\pi}_j - \pi_j^*\|_2$ with $\|\tilde{l}_j - l_j^*\|_2$. To this end, we consider

$$\begin{aligned}
\|\tilde{\pi}_j - \pi_j^*\|_2 \leq & \left\| \frac{\exp(\tilde{l}_j) + 1}{2} - \frac{\exp(l_j^*) + 1}{2} \right\|_2 \\
\leq & \frac{1}{2} \|\exp(\tilde{l}_j) - \exp(l_j^*)\|_2 \\
= & \frac{1}{2} \|\exp(l_j^*)(\exp(\tilde{l}_j - l) - 1)\|_2 \\
\leq & \frac{1}{2} \|\exp(l_j^*)\|_2 \|\exp(\tilde{l}_j - l_j^*) - 1\|_2 \\
\leq & \frac{1}{2} \|2\pi_j^* - 1\|_2 \|\exp(\tilde{l}_j - l_j^*) - 1\|_2 \\
\leq & \frac{1}{2} \|\exp(\tilde{l}_j - l_j^*) - 1\|_2.
\end{aligned}$$

Using the fact that for $x \leq 1 \Rightarrow \exp(x) - 1 \leq 2x$, we have that

$$\|\tilde{\pi}_j - \pi_j^*\|_2 \leq \|\tilde{l}_j - l_j^*\|_2 \leq \|\tilde{l} - l^*\|_\infty. \tag{54}$$

Note that $\|\tilde{\pi}_j - \pi_j^*\|_2 \leq \|\tilde{l}_j - l_j^*\|_2$ in (54) trivially holds when $\|\tilde{l}_j - l_j^*\|_2 > 1$ because $\|\tilde{\pi}_j - \pi_j^*\|_2 \leq 1$.

Bounding $\|\tilde{l} - l^*\|_2$ Bounding $\|\tilde{l} - l^*\|_2$ with $\|\tilde{q} - q^*\|_2$ is straightforward. Indeed, because $\tilde{l} = \arg\min_l \frac{1}{2}\|M^* l - \tilde{q}\|_2^2$ and $l^* = \arg\min_l \frac{1}{2}\|M^* l - q^*\|_2^2$, we have that

$$\|\tilde{l} - l^*\|_2 = \|M^{*\dagger}\tilde{q} - M^{*\dagger} q^*\|_2 \leq \|M^{*\dagger}\|_2 \|\tilde{q} - q^*\|_2, \tag{55}$$

where $M^{*\dagger}$ is the pseudo-inverse of M^*.

Bounding $\|\tilde{q} - q^*\|_2$ Here we bound $\|\tilde{q} - q^*\|_2$. In particular, using $\hat{\Sigma}_{jk} = \Sigma_{jk}^* + \Delta_{jk}$, we have that

$$\begin{aligned}
\left\|\tilde{q}_{jk} - q_{jk}^*\right\|_2 = & \left\| \log \frac{\hat{\Sigma}_{jk}}{\Sigma_{00}^*} - \log \frac{\Sigma_{jk}^*}{\Sigma_{00}^*} \right\|_2 \\
= & \left\| \log \hat{\Sigma}_{jk} - \log \Sigma_{jk}^* \right\|_2 \\
= & \left\| \log(\Sigma_{jk}^* + \Delta_{jk}) - \log \Sigma_{jk}^* \right\|_2 \\
= & \left\| \log \left(1 + \frac{\Delta_{jk}}{\Sigma_{jk}^*} \right) \right\|_2 \\
\leq & \left\| \frac{\Delta_{jk}}{\Sigma_{jk}^*} \right\|_2 \\
\leq & \frac{1}{\omega_{\min}} \|\Delta_{jk}\|_2,
\end{aligned} \tag{56}$$

where we have use the fact that $\log(1 + x) \leq x$, and $\omega_{\min}$ is the smallest positive entry of Σ^*.

Assembling Bounds for $\mathbb{E}[\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_2]$ Putting (53), (54), (55), and (56) together, we have that for all $j \in \{1, 2, \cdots, p\}$,

$$\begin{aligned}
\|\tilde{\mu}_{0j} - \mu_{0j}^*\|_2 \leq & 2\|\tilde{\pi}_j - \pi_j^*\|_2 + 3\|\delta_{jj}\|_2 \\
\leq & 2\|\tilde{l} - l^*\|_\infty + 3\sqrt{p}\|\delta_+\|_\infty.
\end{aligned}$$

As a result,

$$\|\tilde{\mu}_{0+} - \mu_{0+}^*\|_2 \leq 2\sqrt{p}\|\tilde{l} - l^*\|_2 + 3\|\delta_+\|_\infty$$

$$\leq 2\sqrt{p}\|M^{*\dagger}\|_2\|\tilde{q}-q^*\|_2 + 3\sqrt{p}\|\delta_+\|_\infty$$
$$\leq \frac{2\sqrt{p}\|M^{*\dagger}\|_2}{\omega_{\min}}\|\hat{\Sigma}-\Sigma^*\|_{\mathrm{F}} + 3\sqrt{p}\|\hat{\mu}_+ - \mu_+^*\|_\infty.$$

Applying Lemma 3 and Lemma 5, we have that

$$\mathbb{E}[\|\tilde{\mu}_{0+}-\mu_{0+}^*\|_2] \leq \frac{2\sqrt{p}\|M^{*\dagger}\|_2}{\omega_{\min}}\cdot\frac{2\sqrt{2\pi}p^{\frac{5}{2}}}{\sqrt{n}} + 3\sqrt{p}\cdot\frac{4\sqrt{2\pi}p}{n}$$
$$\leq \frac{12\sqrt{2\pi}}{\omega_{\min}}\cdot(\|M^{*\dagger}\|_2+1)\cdot\frac{p^3}{\sqrt{n}}.$$

Bounding $\|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2$ We describe how to upper bound $\|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2$ with $\|\hat{\pi}-\tilde{\pi}\|_2$. Similar to (44), we have that

$$\hat{\mu}_{0j} = 2\Sigma_{00}^*\hat{\pi}_j + (\mu_{00}^* + \hat{\mu}_{jj})\hat{\mu}_{jj} - 1. \tag{57}$$

(57) - (52) yields,

$$\hat{\mu}_{0j} - \tilde{\mu}_{0j} = 2\Sigma_{00}^*(\hat{\pi}_j - \tilde{\pi}_j)$$
$$\Rightarrow \|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2 = 2\Sigma_{00}^*\|\hat{\pi}-\tilde{\pi}\|_2. \tag{58}$$

Bounding $\|\hat{\pi}-\tilde{\pi}\|_2$ Following the same rationale to bound $\|\tilde{\pi}-\pi\|_2$ with $\|\tilde{l}-l\|_2$, we can bound $\|\hat{\pi}-\tilde{\pi}\|_2$ with $\|\hat{l}-\tilde{l}\|_2$ as:

$$\|\hat{\pi}-\tilde{\pi}\|_2 \leq \|\hat{l}-\tilde{l}\|_2. \tag{59}$$

Bounding $\|\hat{l}-\tilde{l}\|_2$ Using an argument similar to that of bounding $\|\hat{t}-\tilde{t}\|_2$, we have that

$$\|\hat{l}-\tilde{l}\|_2 \leq \frac{1}{2}p(p-1)\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s + \sigma_{\min}^{-1}(SU_{\hat{M}}) - \sigma_{\min}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^\top\right\|_2|\log\omega_{\min}|, \tag{60}$$

where s is the number of missing edges in $\hat{G}$ compared to G^*.

Assembling Bounds for $\mathbb{E}[\|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2]$ Combining (58), (59), and (60), and taking the expectation yields:

$$\mathbb{E}[\|\hat{\mu}_{0+}-\tilde{\mu}_{0+}\|_2] \leq p^2\sigma_{\min}^{-1}(\hat{M})\left(u_{\max}s + \sigma_{\min}^{-1}(SU_{\hat{M}})\right)\left\|U_{\hat{M}}^{\perp}(U_{\hat{M}}^{\perp})^\top\right\|_2|\log\omega_{\min}|.$$

□

C.4 Generalization Error

We first provide problem setup and some definitions. Let $y = f_w(x)$ be an end model parameterized by a given w that we seek to learn from the dataset $\mathbb{X} = \left\{x^{(i)}\right\}_{i=1}^n$, where $y^{(i)}$'s are unobserved, drawn from the distribution $\mathcal{D}$. Let $l(y,x;w)\in[0,1]$ be a loss function that takes value between 0 and 1, without loss of generality. Let

$$L(w) = \mathbb{E}_{\mathcal{D}}[l(x,y;w)] \tag{61}$$

be the expected loss of the ends model under $\mathcal{D}$ parameterized by w. Ideally, we will seek $w^* = \arg\min L(w)$. However, since we do not observe y in a weak supervision setting, we will need to consider optimizing an alternative expected loss. In detail, let μ be the population-level mean parameters of the label model under the ground truth dependency graph G. With the label model parametrized by μ, the end model learned from weak supervision has the following expected noise-aware loss of $L_\mu(w)$:

$$L_\mu(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{E}_{(x,\tilde{y})\sim \mathrm{P}_\mu(\cdot|\lambda(x))}[l(x,\tilde{y};w)]]. \tag{62}$$

Furthermore, in pratice, instead of having access to μ, it is the case that we estimate μ from the dataset $\mathbb{L}$ with a dependency graph $\hat{G}$. This yields $\hat{\mu}$. The empirical loss of w associated with $\hat{\mu}$ can be written as:

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{(x,\tilde{y})\sim \mathrm{P}_\mu(\cdot|\lambda(x))}[l(x,\tilde{y};w)].$$

With these definitions, we now introduce the following lemma that link $L(w)$ and $L_\mu(w)$ through the KL divergence between the conditional distribution of y given x of the ground truth and that governed by μ.

Lemma 9. *The following inequality holds between the expected loss defined in* (61) *and the noise-aware expected loss defined in* (62)*:*

$$L_\mu(w) - \sqrt{2 \cdot KL(P_\mathcal{D}(y \mid x) \parallel P_\mu(y \mid x))} \leq L(w) \leq L_\mu(w) + \sqrt{2 \cdot KL(P_\mathcal{D}(y \mid x) \parallel P_\mu(y \mid x))}.$$

Proof. The proof of this theorem can be found in Fu et al. (2020). □

Lemma 10. *Let μ^* and μ^{**} be the mean parameters of two label models of the same set of LFs. The difference between the probabilistic labels produced by the two label models $|P_{\mu^*}(y \mid \lambda) - P_{\mu^{**}}(y \mid \lambda)|$ can be bounded by:*

$$|P_{\mu^*}(y \mid \lambda) - P_{\mu^{**}}(y \mid \lambda)| \leq \frac{1}{2}\|\theta^* - \theta^{**}\|_2,$$

where θ^* and θ^{**} are the canonical parameters associated with μ^* and μ^{**}, respectively.

Proof. Let θ^* be the canonical parameters associated with μ^* and let θ^{**} be the canonical parameters associated with μ^{**}. By (20),

$$\begin{aligned}
|P_{\mu_1}(y \mid \lambda) - P_{\mu_2}(y \mid \lambda)| =& |\text{sigmoid}(2\theta^*_{00} + 2\theta^{*\top}_{0+}\lambda) - \text{sigmoid}(2\theta^{**}_{00} + 2\theta^{**\top}_{0+}\lambda)| \\
\leq& \frac{1}{2}\left\|\theta^*_{0\cdot} - \theta^{**}_{0\cdot}\right\|_\infty \\
\leq& \frac{1}{2}\|\theta^* - \theta^{**}\|_2.
\end{aligned}$$

where for the first inequality we have used Lipschitzness (Definition 4 of Honorio 2012) and the fact that

$$\left|\frac{\partial \mathrm{P}_{\theta_{0\cdot}}(y \mid \lambda)}{\partial \theta_{0j}}\right| = \left|2\lambda_j \cdot (1 - \text{sigmoid}(2\theta_{00} + 2\theta^\top_{0+}\lambda)) \cdot \text{sigmoid}(2\theta_{00} + 2\theta^\top_{0+}\lambda)\right| \leq \frac{1}{2},$$

for all $j \in \{0, 1, 2, \cdots, p\}$, and viewing $\lambda_0 = 1$. □

Lemma 11 (Paraphrase of Lemma 8 of Fu et al. 2020)**.** *Let μ^* and μ^{**} be the mean parameters of two Ising models for the joint distribution of y and λ. Let θ^* and θ^{**} be the associated canonical parameters. For some constant $c_1 > 0$, we have that*

$$\|\theta_1 - \theta_2\|_2 \leq \frac{1}{c_1}\|\mu_1 - \mu_2\|_2.$$

Proof of Theorem 2 With the foregoing preparation, we are now ready to prove Theorem 2.

Proof. We use Lemma 9 to upper bound $L(\hat{w})$ and lower bound $L(w^*)$, this yields

$$L(\hat{w}) - L(w^*) \leq L_{\theta^*_{0\cdot}}(\hat{w}) - L_{\theta^*_{0\cdot}}(w^*) + 2\sqrt{2 \cdot KL(P_\mathcal{D}(y \mid x) \parallel P_\mu(y \mid x))}.$$

We further bound $L_{\theta^*_{0\cdot}}(\hat{w}) - L_{\theta^*_{0\cdot}}(w^*)$, which follows a similar rationale in Ratner et al. (2019). In detail, let $\tilde{w} = \arg\min L_{\hat{\theta}_{0\cdot}}(w)$, we consider

$$\begin{aligned}
L_{\theta^*_{0\cdot}}(\hat{w}) - L_{\theta^*_{0\cdot}}(w^*) =& L_{\theta^*_{0\cdot}}(\hat{w}) + L_{\hat{\theta}_{0\cdot}}(\hat{w}) - L_{\hat{\theta}_{0\cdot}}(\hat{w}) + L_{\hat{\theta}_{0\cdot}}(\tilde{w}) - L_{\hat{\theta}_{0\cdot}}(\tilde{w}) - L_{\theta^*_{0\cdot}}(w^*) \\
\leq& L_{\theta^*_{0\cdot}}(\hat{w}) + L_{\hat{\theta}_{0\cdot}}(\hat{w}) - L_{\hat{\theta}_{0\cdot}}(\hat{w}) + L_{\hat{\theta}_{0\cdot}}(w^*) - L_{\hat{\theta}_{0\cdot}}(\tilde{w}) - L_{\theta^*_{0\cdot}}(w^*) \\
\leq& |L_{\hat{\theta}_{0\cdot}}(\hat{w}) - L_{\hat{\theta}_{0\cdot}}(\tilde{w})| + |L_{\theta^*_{0\cdot}}(\hat{w}) - L_{\hat{\theta}_{0\cdot}}(\hat{w})| + |L_{\hat{\theta}_{0\cdot}}(w^*) - L_{\theta^*_{0\cdot}}(w^*)| \\
\leq& \xi_1(n) + 2|L_{\theta^*_{0\cdot}}(w^\dagger) - L_{\hat{\theta}_{0\cdot}}(w^\dagger)|,
\end{aligned}$$

where for the first inequality we have used the fact that $L_{\hat{\theta}_{0\cdot}}(\tilde{w}) \leq L_{\hat{\theta}_{0\cdot}}(w^*)$, $\xi_1(n)$ is the sampling error, and $w^\dagger = \arg\max_{w \in \{\hat{w}, w^*\}} |L_{\hat{\theta}_{0\cdot}}(w) - L_{\theta^*_{0\cdot}}(w)|$. It remains to bound $|L_{\theta^*_{0\cdot}}(w^\dagger) - L_{\hat{\theta}_{0\cdot}}(w^\dagger)|$. In detail,

$$|L_{\theta^*_{0\cdot}}(w^\dagger) - L_{\hat{\theta}_{0\cdot}}(w^\dagger)| = |\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{E}_{(x,\tilde{y})\sim \mathrm{P}_{\theta^*_{0\cdot}}(\cdot|\lambda(x))}[l(x,\tilde{y};w^\dagger)] - \mathbb{E}_{(x,\tilde{y})\sim \mathrm{P}_{\hat{\theta}_{0\cdot}}(\cdot|\lambda(x))}[l(x,\tilde{y};w^\dagger)]]|$$

$$
\begin{aligned}
=&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{y'\in\{-1,1\}} l(x,y';w^\dagger)\left(\mathrm{P}_{\theta^*_{0\cdot}}(y'\mid\lambda(x))-\mathrm{P}_{\hat{\theta}_{0\cdot}}(y'\mid\lambda(x))\right)\right]\\
\leq&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{y'\in\{-1,1\}} |\mathrm{P}_{\theta^*_{0\cdot}}(y'\mid\lambda(x))-\mathrm{P}_{\hat{\theta}_{0\cdot}}(y'\mid\lambda(x))|\right]\\
=&\sum_{y'\in\{-1,1\}}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[|\mathrm{P}_{\theta^*_{0\cdot}}(y'\mid\lambda(x))-\mathrm{P}_{\hat{\theta}_{0\cdot}}(y'\mid\lambda(x))|\right]\\
\leq&2\max_{y'}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[|\mathrm{P}_{\theta^*_{0\cdot}}(y'\mid\lambda(x))-\mathrm{P}_{\hat{\theta}_{0\cdot}}(y'\mid\lambda(x))|\right]\\
\leq&\|\hat{\theta}_{0\cdot}-\theta^*_{0\cdot}\|_2,
\end{aligned}
$$

where for the last inequality we have used Lemma 10. It remains to bound $\|\hat{\theta}_{0\cdot}-\theta^*_{0\cdot}\|_2$ with $\|\hat{\mu}-\mu^*\|_2$. This is not straightforward because $\hat{\theta}_{0\cdot}$ is learned from $\hat{\mu}$ via logistic regression according to (10) instead of through the use of mean-canonical mapping. Therefore, the mean parameters associated with $\hat{\theta}_{0\cdot}$ are not $\hat{\mu}_{0\cdot}$. As a result, we cannot bound $\|\hat{\theta}_{0\cdot}-\theta^*_{0\cdot}\|_2$ with $\|\hat{\mu}-\mu^*\|_2$ through a direct application of Lemma 11[2]. This motivates the following arguments. In detail, we define a loss function

$$
\ell(\mu_{0\cdot},\lambda;\theta_{0\cdot})=-\theta_{0\cdot}^\top\mu_{0\cdot}+\log[\exp(\theta_{00}+\theta_{0+}^\top\lambda)+\exp(-\theta_{00}-\theta_{0+}^\top\lambda)].
$$

In this way, (10) can still be written as the following empirical risk minimization problem: $\frac{1}{n}\sum_{i=1}^n\ell(\hat{\mu},\lambda^{(i)};\theta)$. Furthermore, for a data generation process governed by θ^*, we can write the population-level risk of a parameter θ as:

$$
\mathcal{L}_{\theta^*}(\theta)=\mathbb{E}[\ell(\mu^*_{0\cdot},\lambda;\theta_{0\cdot})].
$$

That is, to compute $\mathcal{L}_{\theta^*}(\theta)$, we compute $\mu^*_{0\cdot}$ from θ^*, we then view $\mu^*_{0\cdot}$ as a constant input for $\ell(\cdot)$ and sample λ (infinitely many times) for the computation of the the population level loss. It should also be noticed that $\mathcal{L}_{\theta^*}(\theta_{0\cdot})$ is also the population level loss for a logistic regression problem parameterized by $\theta_{0\cdot}$ under the data generation process governed by $\theta^*_{0\cdot}$. Therefore, under standard regularity conditions (Negahban et al., 2012), we have that $\mathcal{L}_{\theta^*}(\theta_{0\cdot})$ is strongly convex. This implies that for some strongly-convex constant γ,

$$
\|\hat{\theta}_{0\cdot}-\theta^*_{0\cdot}\|_2\leq\frac{2}{\gamma}\left(\mathcal{L}_{\theta^*}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\theta^*}(\theta^*_{0\cdot})\right),
$$

where we have also used the fact that $\theta^*_{0\cdot}$ is a global minimizer for $\mathcal{L}_{\theta^*}(\theta_{0\cdot})$. It remains to bound $\mathcal{L}_{\theta^*}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\theta^*}(\theta^*_{0\cdot})$ with $\|\hat{\mu}-\mu^*\|_2$, which we achieved through an argument similar to that of bounding $L_{\theta^*}(\hat{w})-L_{\theta^*}(w^*)$ for the end model. Specifically, define the canonical parameters associated with $\hat{\mu}$ achieved through mean-canonical parameter mapping as $\hat{\tilde{\theta}}$, and further define $\tilde{\theta}_{0\cdot}=\arg\min_\theta\mathcal{L}_{\hat{\tilde{\theta}}}(\theta_{0\cdot})$, we have that

$$
\begin{aligned}
\mathcal{L}_{\theta^*}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\theta^*}(\theta^*_{0\cdot})=&\mathcal{L}_{\theta^*}(\hat{\theta}_{0\cdot})+\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})+\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta}_{0\cdot})-\mathcal{L}_{\theta^*}(\theta^*_{0\cdot})\\
\leq&\mathcal{L}_{\theta^*}(\hat{\theta}_{0\cdot})+\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})+\mathcal{L}_{\hat{\tilde{\theta}}}(\theta^*_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta}_{0\cdot})-\mathcal{L}_{\theta^*}(\theta^*_{0\cdot})\\
\leq&|\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta}_{0\cdot})|+|\mathcal{L}_{\theta^*}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})|+|\mathcal{L}_{\hat{\tilde{\theta}}}(\theta^*_{0\cdot})-\mathcal{L}_{\theta^*}(\theta^*_{0\cdot})|\\
\leq&|\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta}_{0\cdot})|+2|\mathcal{L}_{\theta^*}(\theta^\dagger_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\theta^\dagger_{0\cdot})|,
\end{aligned}
$$

where $\theta^\dagger_{0\cdot}=\arg\max_{\theta_{0\cdot}\in\{\hat{\theta}_{0\cdot},\theta^*_{0\cdot}\}}|\mathcal{L}_{\hat{\tilde{\theta}}}(\theta_{0\cdot})-\mathcal{L}_{\theta^*}(\theta_{0\cdot})|$. It remains to analyze $|\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta}_{0\cdot})|$ and $|\mathcal{L}_{\theta^*}(\theta^\dagger_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\theta^\dagger_{0\cdot})|$ for the last inequality. For $|\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta}_{0\cdot})|$, we notice that $\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})=\mathbb{E}[\ell(\hat{\mu},\lambda;\hat{\theta}_{0\cdot})]$ and $\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta})=\mathbb{E}[\ell(\hat{\mu},\lambda;\tilde{\theta}_{0\cdot})]$. Because $\hat{\theta}_{0\cdot}$ is learned from a dataset of n samples $\{\lambda^{(i)}\}_{i=1}^n$ and $\hat{\mu}$ while $\tilde{\theta}_{0\cdot}$ is learned from the corresponding population distribution of λ and $\hat{\mu}$, we can view $|\mathcal{L}_{\hat{\tilde{\theta}}}(\hat{\theta}_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\tilde{\theta}_{0\cdot})|$ as sampling error of empirical risk minimization bounded by some decreasing function of $\xi_2(n)$. For $|\mathcal{L}_{\theta^*}(\theta^\dagger)-\mathcal{L}_{\hat{\tilde{\theta}}}(\theta^\dagger)|$, we observe

$$
|\mathcal{L}_{\theta^*}(\theta^\dagger_{0\cdot})-\mathcal{L}_{\hat{\tilde{\theta}}}(\theta^\dagger_{0\cdot})|=|\mathbb{E}[\ell(\mu^*,\lambda;\theta^\dagger_{0\cdot})]-\mathbb{E}[\ell(\hat{\mu},\lambda;\theta^\dagger_{0\cdot})]|
$$

[2]However, for the exact inference algorithms mentioned in Appendix B.7.2 and Appendix B.7.3 for conditionally independent labeling functions, we can apply Lemma 11 directly to conclude the proof.

Name	# Samples	# Non-abstain Samples	# LFs	# Bipolar LFs	Test positive rate
`spam`	1,486	1,381	10	0	47.2%
`crowdsourcing`	187	187	190	88	56%
`spouse`	22,254	5,734	9	1	8.07%
`IMDB`	50,000	0	16	0	50.25%

Table 3: Summary of four benchmark datasets.

Dataset	Majority Vote	Production	CLL	Flyingsquid	Firebolt
`spam`	0.908	0.943	0.879	0.889	**0.948**
`spouse`	0.767	0.784	0.655	0.783	**0.797**
`IMDb`	0.738	0.636	0.701	0.756	**0.777**

Table 4: AUCs of the label models on various benchmark datasets.

$$\leq c_2 \|\hat{\mu} - \mu^*\|_2,$$

where we have used the Lipschitzness of $\mathbb{E}[\ell(\mu, \lambda; \theta^\dagger_{0.})]$ about μ for a bounded space of $\theta^\dagger_{0.}$. □

D Extended Experiments

In this section, we describe experiment setup and present extended experimental results. First, we provide further details about the setup and results on four benchmark weak supervision datasets (Appendix D.1). Next, we describe the experiments for zero-shot learning through the use of 45 datasets derived from the Animal with Attributes 2 (`AwA2`) dataset (Xian et al., 2018; Mazzetto et al., 2021b) in Section 5.3. Finally, we seek to better understand the characteristics of Firebolt empirically via experiments on synthetic data (Appendix D.2).

D.1 Experiments on Benchmark Datasets

Datasets We further describe the setup for each of the benchmark datasets we used in the main paper. For all datasets, only data points that receive at least one vote from the labeling functions are used in the training set. Since we do not have access to the dependency graph between labeling functions in these datasets, we further assume that the labeling functions are conditionally independent of each other. Table 3 shows a summary of the datasets.

- `spam`: The `spam` dataset is a balanced dataset of a binary classification problem that compares whether Youtube comments are spam or non-spam (Alberto et al., 2015). We use this dataset to demonstrate the utility of Firebolt tackling weak supervision problems when the class balance is known. To do this, we assume that the class balance is known and set the positive rate of the dataset manually to 0.5. For the end model, we used a `countvectorizor` to produce features and feed those features to a one hidden layer neural network with 100 hidden units. The use of the `countvectorizor` based end model follows the feature representation used in the Snorkel tutorial of weak supervision from which this dataset is derived.

- `crowdsourcing`: The `crowdsourcing` dataset is yet another balanced dataset. We also use this dataset to demonstrate the utility of Firebolt tackling balanced weak supervision problems. Similar to `spam` dataset, we set the positive rate of the dataset manually to 0.5. For the end model, we use the same end model as `spam`.

- `spouse`: The `spouse` dataset seeks to identify mentions of spouse relationships in a set of news articles Corney et al. (2016). The dataset is highly imbalanced with a test set positive rate of about 8%. We use Firebolt to estimate this positive rate and then run Firebolt using the estimated positive rate. We use the same end model from the Snorkel tutorial from which this dataset is derived as our end model.

- `IMDb`: The `IMDb` dataset seeks to distinguish positive user sentiments from negative sentiments in movie reviews (Maas et al., 2011). The dataset is also balanced with a positive rate of 50.25% on the test set. The dataset contains 50,000 examples and we use 40,000 for training, 8,000 for test, and 2,000 for validation. We provide a

Dataset	Majority Vote	Production	CLL	Flyingsquid	Firebolt
spam	0.884	0.935	0.869	0.896	**0.940**
spouse	0.267	**0.381**	0.262	0.371	0.377
IMDb	0.679	0.606	0.636	0.729	**0.753**

Table 5: APs of the label models on three benchmark datasets.

Dataset	Majority Vote	Production	CLL	Flyingsquid	Firebolt
spam	0.969±0.002	**0.982±0.001**	0.923±0.002	0.961±0.002	**0.982±0.001**
crowdsourcing	**0.838±0.009**	0.819±0.006	**0.845±0.004**	0.806±0.005	**0.835±0.010**
spouse	0.179±0.034	0.266±0.012	0.215±0.033	0.231±0.019	**0.374±0.007**
IMDb	0.798±0.002	0.624±0.001	**0.820±0.002**	0.788±0.001	**0.822±0.001**

Table 6: Test APs of weakly supervised end models on four benchmark datasets over five trials (mean±s.d.).

0.5 positive rate to Firebolt for this dataset. The end model we used is a one-hidden-layer relu neural network and the features are trainable embeddings from `https://tfhub.dev/google/nnlm-en-dim50/2`. The hidden layer has 512 nodes. We use word embedding in our end model because we experience out-of-memory issues using the `countvectorizor` for this dataset.

We generate labeling functions for the dataset by using a simple heuristic that either votes or abstain on the data examples. We define a set of words where each word represents a positive or negative sentiment. For words that have positive sentiments, we obtain labeling functions from them by voting positive if the word is present in the review else we abstain. Similarly, for words with negative sentiments we vote negative when the words are present in the review else we abstain. The positive sentiment words we use are: `like, love, good, great, best, excellent, amazing` while the negative sentiment words are: `could, awful, better, bad, terrible, worst, horrible, sucks`.

Protocol We follow the same protocol as in the main paper and learn the labels on the training data then evaluate the end model on the test data. When validation sets are available, we also use the validation sets during the training of the end models. We report mean of the results over 5 trials and also show the standard deviation of the runs.

Results We present results on both the label model and the end model. Table 4 and Table 5 present the AUC and average precision (AP) of various label models on three datasets, where we do not report the results of the `crowdsourcing` dataset because the LFs in this dataset abstain in the test set. As can be seen, Firebolt outperforms alternative methods in AUC over all three datasets. Nonetheless, the production label model slightly outperforms Firebolt in AP on the `spouse` dataset. Table 6 shows the performance of Firebolt and alternative methods on the four benchmark classification tasks. Other than the AUC reported in Table 1, here we also report the AP and from the Table we see that Firebolt either produces the best results or is in a (statistical) tie to be the best approach among all datasets. Note that for the imbalanced `spouse` dataset, Firebolt outperforms the next best performing method by 10.8 percentage points in AP.

D.2 Synthetic Data

We present a series of synthetic experiments to better understand the characteristics of Firebolt. Specifically, we show that

- Firebolt can tackle imbalanced classification with conditionally independent unipolar LFs (Appendix D.2.1).
- Firebolt can learn from a mix of conditionally independent unipolar and bipolar LFs (Appendix D.2.2).
- Firebolt can learn from LFs that have complex dependency among them (Appendix D.2.3).
- Firebolt is robust against worse-than-random LFs and misspecified dependencies (Appendix D.2.4).

It should be noticed that the distinction and discussion of the first three settings (Appendix D.2.1–Appendix D.2.3) are meaningful because Firebolt can use different inference algorithms in these three settings. In particular,

Method	Majority Vote	Production	CLL	Flyingsquid	Firebolt
AUC	0.565	0.916	0.400	0.962	**0.967**
AP	0.007	0.186	0.007	0.504	**0.554**

Table 7: Test performance of label models learned from an imbalanced dataset of 80,000 samples of conditionally independent positive labeling functions.

Appendix D.2.1 uses the inference algorithm described in Appendix B.7.2; Appendix D.2.2 uses the inference algorithm described in Appendix B.7.3; and Appendix D.2.3 uses the inference algorithm described in Appendix B.7.1.

D.2.1 Imbalanced Classification with Unipolar LFs

We empirically evaluate the performance of Firebolt for imbalanced classification by learning from a set of conditionally independent unipolar labeling functions. To this end, we carry out experiments on synthetic data using a ground truth label model of 10 conditionally independent positive labeling functions of varying accuracy and prevalence. The ground truth positive rate of label is 0.608%. We want to understand the predictive performance, the parameter learning performance, and the interpretability of Firebolt.

Metrics We use the AUC and AP on the test set to understand the predictive performance of Firebolt. We use the parameter estimation error of the mean parameters $\|\hat{\mu}_{0+} - \mu^*_{0+}\|_2$ and $\|\hat{\mu}_{00} - \mu^*_{00}\|_2$ as metrics for parameter learning performance (including class balance). Finally, we use the the canonical parameter estimation error $\|\hat{\theta}_{0+} - \theta^*_{0+}\|_2$ and $\|\hat{\theta}_{00} - \theta^*_{00}\|_2$ to understand the interpretability of Firebolt.

Protocol Using the ground truth label model, we produce two datasets of $80,000$ and $160,000$ data points, respectively. For predictive performance, we use the dataset of $80,000$ data points to train a Firebolt model as well as four alternative methods. We then use the AUC and AP of the label model learned by the various methods on the ground truth test distribution to evaluate the predictive performance of each method. Inference is carried out based on Appendix B.7.2. For parameter learning and interpretability, we train Firebolt on the two datasets to see how the mean and canonical parameter estimation error change with increasing sample size.

Expected Results We anticipate that Firebolt can deliver reasonable test AUC and AP for predictive performance. Furthermore, we anticipate that the the mean and canonical parameter estimation error will decrease by learning from dataset of increased sample size.

Sample Size	$\|\hat{\mu}_{0+} - \mu^*_{0+}\|_2$	$\|\hat{\mu}_{00} - \mu^*_{00}\|_2$	$\|\hat{\theta}_{0+} - \theta^*_{0+}\|_2$	$\|\hat{\theta}_{00} - \theta^*_{00}\|_2$
80,000	0.016	0.011	0.285	0.264
160,000	0.004	0.002	0.176	0.039

Table 8: Mean and canonical parameter estimation error of Firebolt when learning from a set of conditionally independent unipolar LFs.

Results We first present the results on predictive performance in Table 7. As can be seen, Firebolt outperforms alternatives on predictive performance measured by both test AUC and AP. Next, we present results on parameter learning performance and interpretability, summarized in Table 8. As can be seen, both the mean parameter estimation error and the canonical parameter estimation error decrease as the sample size increases. This demonstrates the capacity of Firebolt in recovering the parameters of the label model as well as its interpretability for the importance of each labeling function in the data generation process. It should also be noticed that the parameter estimation error associated with the class balance $\|\hat{\mu}_{00} - \mu^*_{00}\|_2$ and $\|\hat{\theta}_{00} - \theta^*_{00}\|_2$ decrease more slowly compared to that of the accuracy parameters of the labeling functions. This highlight the difficulty of imbalanced classification.

Method	Majority Vote	Production	CLL	Flyingsquid	Firebolt
AUC	0.776	0.844	0.770	0.844	**0.896**
AP	0.296	0.558	0.007	0.302	**0.632**

Table 9: Test performance of label models learned from an imbalanced dataset of 50,000 samples of conditionally independent unipolar and bipolar LFs.

D.2.2 Learning from a Mix of Unipolar and Bipolar LFs

We study the performance of Firebolt in learning from a mixed of conditionally independent unipolar and bipolar labeling functions. For this purpose, we consider a conditionally independent ground truth label model with two positive labeling functions and one bipolar labeling function. The ground truth positive rate of label is 15.114%. We want to understand the predictive performance, the parameter learning performance, and the interpretability of Firebolt in this setting.

We use the same metrics and follow the same protocol as described in Appendix D.2.1. We produce two datasets with $5,000$ and $50,000$ samples, respectively. We evaluate the predictive performance of Firebolt on the dataset with $50,000$ samples. We follow the inference procedure described in Appendix B.7.3.

Sample Size	$\|\hat{\mu}_{0+} - \mu^*_{0+}\|_2$	$\|\hat{\mu}_{00} - \mu^*_{00}\|_2$	$\|\hat{\theta}_{0+} - \theta^*_{0+}\|_2$	$\|\hat{\theta}_{00} - \theta^*_{00}\|_2$
$5,000$	0.076	0.056	0.171	0.197
$50,000$	0.019	0.013	0.035	0.044

Table 10: Mean and canonical parameter estimation error of Firebolt when learning from a mix of unipolar and bipolar LFs.

The predictive performance is shown in Table 10, where Firebolt outperforms alternatives in this setting. The parameter learning performance and interpretability is shown in Table 10, where both the mean and canonical parameter error decrease as we have more samples to learn from.

D.2.3 Learning with Complex Dependencies

In Section 5.4, we study the predictive performance of Firebolt compared to other methods when learning from labeling functions with complex dependency. Here, we extend this study by also understanding the parameter learning performance and interpretability of Firebolt learning from labeling functions with complex dependency. To this end, we consider the same ground truth label model used in Section 5.4.

We produce two datasets of $80,000$ and $800,000$ samples respectively and use Firebolt to learn label models from these two datasets. We then measure the mean and canonical parameter estimation error as metrics to understand the parameter learning performance and its interpretability.

Sample Size	$\|\hat{\mu}_{0+} - \mu^*_{0+}\|_2$	$\|\hat{\mu}_{00} - \mu^*_{00}\|_2$	$\|\hat{\theta}_{0+} - \theta^*_{0+}\|_2$	$\|\hat{\theta}_{00} - \theta^*_{00}\|_2$
$5,000$	0.378	0.205	0.293	0.825
$50,000$	0.034	0.008	0.080	0.032

Table 11: Mean and canonical parameter estimation error of Firebolt when learning from LFs with complex dependency.

The results are summarized in Table 11. As can be seen, both the mean and canonical parameter estimation errors decrease as Firebolt learns from an increased number of samples. This indicates the improved performance of Firebolt in parameter learning and reflecting the contribution of each labeling function in producing probabilistic labels under the ground truth data generation process.

D.2.4 Robustness

We empirically demonstrate the robustness of Firebolt against common violation of assumptions and conditions made by the algorithm such as using worse-than-random LFs and misspecified dependencies.

Method	Majority Vote	Production	CLL	Flyingsquid	Firebolt
AUC	0.554	0.894	0.640	0.955	**0.965**
AP	0.007	0.123	0.008	0.468	**0.544**

Table 12: Test performance of label models learned from an imbalanced dataset of 80,000 samples of conditionally independent positive labeling functions where one of the LFs is worse than random guessing.

Worse-Than-Random LFs We seek to understand the robustness of Firebolt when some of the LFs that Firebolt learns from are worse than random guessing. We are interested in understanding how the predictive performance may deteriorate with worse than random guessing LFs. For this purpose, we modify the ground truth label model in Appendix D.2.1 by changing one of the LFs from better than random guessing to worse than.

We use the same metrics and follow the same protocol as described in Appendix D.2.1 to evaluate the predictive performance of Firebolt. The results are summarized in Table 12. As we can observe, Firebolt outperforms alternatives under the scenario where not all the labeling functions are better than random guessing. Nonetheless, comparing Table 12 with Table 7, we notice that the predictive performance of all methods drop when not all LFs are better than random guessing, with Firebolt among the ones being influenced the least.

Misspecified Dependency We demonstrate the robustness of Firebolt against misspecified dependency. We are concerned with the deterioration of predictive performance with misspecified dependency. To this end, we run Firebolt on the dataset with $50,000$ samples described in Appendix D.2.2 but we remove the dependency between the two labeling functions that represent the bipolar labeling function to create a conditionally independent dependency graph $\hat{G}$. We feed this $\hat{G}$ to Firebolt instead of G^* to learn a label model and measure its test performance. We get a test AUC of 0.871 and a test AP of 0.588. This result is only second to Firebolt with the correctly specified dependency graph, comparing to the performance reported in Table 9. This result suggests the robustness of Firebolt against misspecified dependency.

E Discussion

In this section, we discuss the limitations and potential societal impact of our work.

Limitation The assumptions made by Firebolt as described in Appendix C.2.1 may induce potential limitation, especially when those assumptions are not met in practice. Furthermore, it is worth noting that Firebolt does not deal with dependency graph with more than one latent variable, such as the ones dealt with in Sala et al. (2019); Fu et al. (2020); Hooper et al. (2020). In addition, Firebolt also requires user specifying dependency graph among the labeling functions as an input, which can be challenging for the users to provide without structure learning (Bach et al., 2017; Varma et al., 2019). Lastly, based on our theoretical analysis, the convergence of Firebolt relies on the convergence of the covariance tensor, whose optimal rate is still an open statistical problem as we previously described in Section 4.

Societal Impact Firebolt focuses on the setting where we learn with limited to no label availability. Similar to it's supervised learning counterpart, the performance of our method is highly dependent on the input it receives which is the labeling functions in our case. In practice we encourage users to define good labeling functions and to ensure that the labeling functions are not biased unfairly in a way that will cause the model make biased predictions towards certain population or demographic. While fairness is beyond the scope of our paper, we encourage users to make adequate fairness considerations when using our model.