Comparing the Value of Labeled and Unlabeled Data in Method-of-Moments Latent Variable Estimation

Anonymous Author
Anonymous Institution

Abstract

Labeling data for modern machine learning is expensive and time-consuming. Latent variable models can be used to infer labels from weaker, easier-to-acquire sources operating on unlabeled data. Such models can also be trained using labeled data, presenting a key question: should a user invest in few labeled or many unlabeled points? We answer this via a framework centered on model misspecification in method-of-moments latent variable estimation. Our core result is an exact bias-variance decomposition of the generalization error, which shows that the unlabeled-only approach incurs additional bias under misspecification. We introduce a correction that provably removes this bias in certain cases. We apply our decomposition to three scenarios—well-specified, misspecified, and corrected models—to 1) choose between labeled and unlabeled data and 2) learn from their combination. We observe theoretically and with synthetic experiments that for well-specified models, labeled points are worth a constant factor more than unlabeled points. With misspecification, their relative value is higher due to the additional bias but is reduced with correction. We also apply our approach to study real-world weak supervision techniques for dataset construction.

1 Introduction

A key challenge in data-driven fields is the quality of training data. A fixed data collection budget can provide a large amount of incomplete training data, or a smaller but cleaner dataset. Given a choice between these two options, which should we select and which factors should determine this decision? This question, while fundamental, is especially relevant to modern machine learning, where vast amounts of unlabeled data is available. To exploit this without extensive hand-labeling, powerful techniques relying on *latent variable models*—in particular, *method-of-moments*—have been developed to generate labels.

Latent variable method-of-moments has been used to learn topic models (Anandkumar et al., 2014) and parse trees (Hsu et al., 2012), to evaluate crowdworkers (Joglekar et al., 2013), and to generate training datasets (Ratner et al., 2019; Fu et al., 2020). In these models, the outputs of sources—variables with some relation to the label—are observed and used to infer the latent variable. The core challenge is to learn the correlations (i.e., *accuracies*) between the sources and the unobserved label variable, which parametrize the model used to generate labels. Here, method-of-moments relies on decomposing multiple observable statistics based on independence among sources. When some labeled data is available, this setup also allows for the accuracy parameters to be directly estimated (Figure 1). Therefore, given a limited budget, a principle for choosing between labeled and unlabeled data is crucial, motivating a theoretical framework to understand the relative value between them.

Unmodeled dependencies among sources—a form of model misspecification—are common and yield inconsistent accuracy estimates, which in turn yield poor inferred labels. This affects the value of data produced with latent variable methods, so misspecification must play a role in our framework. While the question of how to analyze misspecification has been studied in classical statistics, the focus is on estimator asymptotics (Kleijn and van der Vaart, 2006, 2012). Our main challenge, however, is to analyze and understand misspecification for both parameter estimation and label inference in the finite—and often small—sample setting.

We theoretically analyze the two alternatives in latent variable methods. In both cases, the output is a

Preliminary work. Under review by AISTATS 2021. Do not distribute.

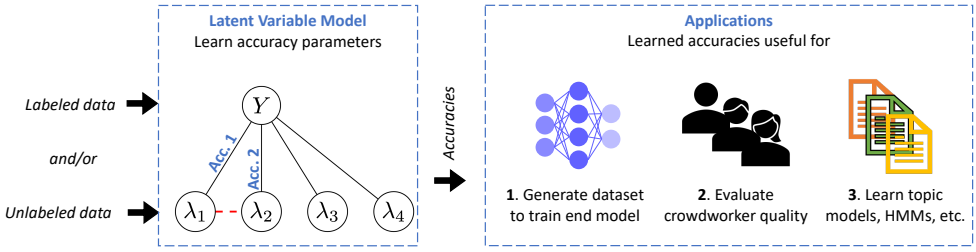


Figure 1: Latent variable methods (e.g., method-of-moments) can infer an unobserved variable (Y) by learning the accuracies of correlated sources ($\lambda_1, \ldots, \lambda_4$). This is done either from unlabeled data or directly from a small amount of labels; we seek a framework to explain the relative value of these choices. A major challenge are unmodeled dependencies between sources (red). Latent variable models have numerous applications.

joint distribution for the latent variable and observable sources. For the inputs, the choices are either n_L labeled or n_U unlabeled points (and the outputs of m sources). We examine misspecification in the form of unmodeled pairwise dependencies, giving a generalization error analysis for method-of-moments latent variable model performance of the two alternatives. We present a bias-variance decomposition of the generalization error, which for both the labeled and unlabeled data cases consists of (i) irreducible error, (ii) variance, and (iii) bias due to model misspecification at inference time. An important consequence is that for unlabeled data, we incur an additional (iv) standing bias due to incorrectly estimating accuracies that scales with the extent of misspecification, $\mathcal{O}(d/m)$ for m sources and d unmodeled dependencies among them.

Next, we turn to correcting this standing misspecification bias. In particular, a simple median approach is able to produce consistent estimators given that $d = o(m^2)$ and sufficient amounts of unlabeled data. Therefore, in certain cases, the bias $\mathcal{O}(d/m)$ from misspecification can be completely eliminated. This creates three scenarios to consider for our framework: well-specified (i.e. no unmodeled dependencies), misspecified, and corrected settings, depicted in Figure 2.

We give two applications of our theoretical framework for the three scenarios. First, we develop a criterion, *the data value ratio*, for choosing between labeled and unlabeled data., which is based on the relative minimum amount of labeled points needed to perform as well as a fixed amount of unlabeled points in terms of generalization error. For well-specified models, labeled data is a constant factor more valuable than unlabeled, but for misspecified models the value grows linearly in d and n_U. Furthermore, corrected models are able to *improve the value of unlabeled data*. Second, we combine the estimated parameters from the unlabeled approach, which are biased, with ones from the labeled approach—in certain cases outperforming either individually. We validate our framework with synthetic experiments, verify the scaling of our generalization error and data value ratio, and the performance of the combined estimator across the three settings.

An important real-world application of our results on latent variable methods are weak supervision (WS) frameworks, in particular data programming (Ratner et al., 2016), used in a huge range of products and systems across industry and academia. WS frameworks construct datasets without ground-truth annotations by using unlabeled points and distant or weak sources, such as heuristics (Gupta and Manning, 2014), external knowledge bases (Mintz et al., 2009; Craven and Kumlien, 1999; Takamatsu et al., 2012), or noisy crowd-sourced labels (Karger et al., 2011; Dawid and Skene, 1979). Data programming encompasses many such prior approaches, and has shown excellent results with the method-of-moments approach (Fu et al., 2020). We perform a real-world WS case study, where ground-truth source dependencies are not known, but sources are likely to be correlated to some extent. We observe that the relative value of labeled data is large, but the value of unlabeled data can be increased via our median approach. With equal amounts of data, the F1-score of a baseline unlabeled approach is 64.81 and the score of a labeled approach is 71.79, but the score of an unlabeled approach with correction is 68.12. This suggests that our theoretical explanation of the effects of misspecification can account for some of the behavior of models on real data.

2 Related Work

Misspecification in Graphical Models The asymptotic effect of misspecification on parameter estimation is studied by Kleijn and van der Vaart (2012), extending the Bernstein-Von Mises theorem to cases where observed samples are not of the parametric distribution being estimated. However, their main results do not fully extend to method-of-moments estimators. Other analyses of model misspecification directly examine families of models, such as Jog and Loh (2015)'s

lower bound on KL-separation of Gaussian graphical models. While this bound is important for modeling errors in inference, it does not illustrate our additional error in parameter estimation. More generally, works on misspecification either study a particular class of techniques (De Blasi and Walker, 2013) or a particular model and propose repairs (Grünwald et al., 2017)—while we compare effects on alternative datasets.

Structure Learning One way to reduce misspecification is to produce a more refined model. Graphical model structure learning aims to do so in both the supervised (Ravikumar et al., 2011; Loh and Wainwright, 2013) and unsupervised cases (Chandrasekaran et al., 2012; Meng et al., 2014; Bach et al., 2017; Varma et al., 2019). However, these works present computational challenges, require (often strong) conditions to hold, and do not analyze the downstream impact of errors. Our approach instead focuses on understanding the impact of errors and is applicable to partial recovery that often results from structure learning.

3 Background and Problem Setting

We start with background on latent variable models and introduce the model we analyze. We explain the stages—learning accuracies and inferring labels—for both the labeled and unlabeled cases, and conclude with model evaluation and key challenges.

Setup In latent variable models, a number of sources are observed and used to infer the latent variable. The input is usually n_U unlabeled data points, but in our setting we also consider a small *labeled* dataset of n_L samples. The output is a large, labeled dataset.

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y} = \{-1, 1\}$. We consider an unlabeled dataset $\boldsymbol{X}_U = \{x_i^U\}_{i=1}^{n_U}$ and a labeled dataset $(\boldsymbol{X}_L, \boldsymbol{Y}_L) = \{(x_i^L, y_i^L)\}_{i=1}^{n_L}$ drawn from the distribution of (X, Y). There are m sources, each outputting a value in $\{-1, +1\}$ via a deterministic function $\lambda_j : \mathcal{X} \to \mathcal{Y}$ for all $j \in [m]$. Our goal is to use the outputs of $\boldsymbol{\lambda}$, the vector of sources, to construct a model to infer Y.

To infer Y, we learn the model $\Pr(Y, \boldsymbol{\lambda})$ and then marginalize to produce soft labels $\widetilde{y}_i := 2\Pr(Y = 1|\boldsymbol{\lambda} = \boldsymbol{\lambda}(x_i)) - 1 \in [-1, 1]$ for each x_i by applying the m sources to $\boldsymbol{X}_U$ and $(\boldsymbol{X}_L, \boldsymbol{Y}_L)$. The overall approach has two steps: (i) learn the latent variable model (using labeled or unlabeled data), and (ii) infer labels $\tilde{y}_i$.

Theoretical model We pick a simple model that captures many latent variable model settings and still presents all of the challenges for comparing between the types of data. We assume an Ising model for $\Pr(Y, \boldsymbol{\lambda})$; the only difference between the labeled and unlabeled setting is that Y is latent in the latter. The set of canonical parameters is Θ, and the dependency graph is $G = (V, E)$, where $V = Y \cup \boldsymbol{\lambda}$ and E consists of edges from Y to the sources as well as the d edges among the sources, E_λ. The lack of an edge in G between a pair of variables indicates independence conditioned on a separator set (Lauritzen, 1996), so the true distribution can be modeled as

$$\Pr(Y, \boldsymbol{\lambda}) = \frac{1}{Z} \exp\Big(\theta_Y + \sum_{i=1}^{m} \theta_i \lambda_i Y + \sum_{(i,j) \in E_\lambda} \theta_{ij} \lambda_i \lambda_j\Big),$$

with cumulant function Z. For cleaner presentation, we assume $\Theta \geq 0$ (no sources that disagree with others or Y on average) and E_λ is sparse enough such that $\deg(\lambda_i) \leq 2$ for all λ_i (each source is conditionally dependent on at most one other source). This leads to model misspecification when edges are unknown.

Inference The label is computed using a naive Bayes approach that assumes all sources are conditionally independent with $E_\lambda = \emptyset$:

$$\widetilde{\Pr}(Y = 1|\lambda = \lambda(X)) = \frac{\prod_{i=1}^{m} \widetilde{\Pr}(\lambda_i = \lambda_i(X)|Y = 1) \Pr(Y = 1)}{\hat{\Pr}(\lambda = \lambda(X))} \quad (1)$$

where the class balance $\Pr(Y = 1)$ is assumed to be known, $\hat{\Pr}$ is an empirical probability , and $\widetilde{\Pr}$ indicates an estimated probability resulting from the parameter estimation step described below. In practice, the conditional independence assumptions required for (1) may not hold, but dependencies among sources are often unknown. Therefore, conditional independence is assumed, and we may suffer from *misspecification* in inferring our probabilistic labels.

Learning parameters with method-of-moments For the labeled dataset, we estimate $\widetilde{\Pr}(\lambda_i = \lambda_i(X)|Y = 1)$ in (1) directly from samples, as Y is observed.

For the unlabeled dataset, we use the method-of-moments estimator from Fu et al. (2020), which relies on the property that if $\lambda_i \perp\!\!\!\perp \lambda_j | Y$, then $\lambda_i Y \perp\!\!\!\perp \lambda_j Y$. This implies that $\mathbb{E}[\lambda_i Y] \cdot \mathbb{E}[\lambda_j Y] = \mathbb{E}\left[\lambda_i \lambda_j Y^2\right] = \mathbb{E}[\lambda_i \lambda_j]$, which is directly estimable. Define $a_i := \mathbb{E}[\lambda_i Y]$ as the unknown *accuracy* of λ_i. If we can introduce a third λ_k that is conditionally independent of λ_i and λ_j, we have a system of equations that can be solved using observable statistics. We use this *triplet method* to recover these accuracies: we choose two λ_j, λ_k at random for each λ_i and solve up to sign:

$$|\widetilde{a}_i^{(j,k)}| := \sqrt{\left|\frac{\hat{\mathbb{E}}[\lambda_i \lambda_j]\, \hat{\mathbb{E}}[\lambda_i \lambda_k]}{\hat{\mathbb{E}}[\lambda_j \lambda_k]}\right|}, \quad (2)$$

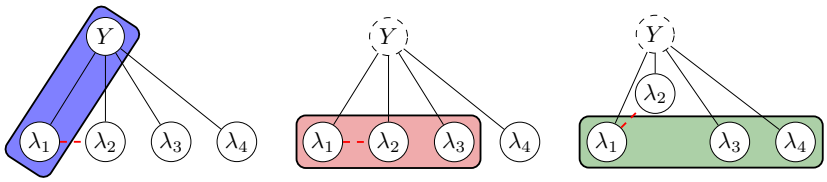


Figure 2: Latent variable models with unmodeled dependency (red edge), leading to misspecification. Boxes indicate observable variables used for accuracy estimation. Left: model with access to label Y. Pairs (λ_i, Y) directly estimate source accuracy. Center: latent model with unobserved Y. Triplet (red) includes unmodeled dependency, leading to inconsistent estimation. Right: Corrected model using medians. Triplet excludes the dependency, returning to consistency.

where $\hat{\mathbb{E}}$ is an empirical estimate of the expectation. We use the estimated $\widetilde{a}_i^U := \widetilde{a}_i^{(j,k)}$ to directly compute $\widetilde{\Pr}(\lambda_i = \pm 1 | Y = 1)$ for (1). However, random λ_j and λ_k may not satisfy conditional independence, and thus we incur error in estimating accuracies due to misspecification in a way unique to the unlabeled setting. We aim to capture this error in our evaluation.

Evaluating the model We define the model's generalization error as $R = \mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}[l(\widetilde{Y}, Y)]$ where expectation is taken over the distribution of $(Y, \boldsymbol{\lambda})$, $\mathcal{N}$ (the random dataset used), and τ (the algorithmic randomness, i.e. the triplets used in method-of-moments). $l(\cdot,\cdot)$ here is the cross entropy loss, $l(\widetilde{y}_i, y_i) = -\frac{1+y_i}{2}\log\widetilde{\Pr}(Y = 1|\lambda = \lambda(x_i)) - \frac{1-y_i}{2}\log\widetilde{\Pr}(Y = -1|\lambda = \lambda(x_i))$. Let R_U denote the error for the unlabeled dataset and R_L for labeled.

4 Theoretical Results

We theoretically analyze the quality of the latent variable model, taking into account the impact of misspecification when using unlabeled versus labeled data. In 4.1 we give an exact decomposition of the generalization error of the latent variable model, which demonstrates how misspecification is present in both the parameter learning and inference steps of the model when data is unlabeled and only present in the latter when data is labeled. In 4.2, we bound the generalization error using this framework to show how the unlabeled case has an additional standing bias of $\mathcal{O}(d/m)$. Given this standing bias, in 4.3, we introduce a simple method that under certain conditions can correct for dependency-based misspecification. We analyze this correction's impact on generalization error.

4.1 Decomposition Framework

Our first result is a decomposition of the generalization error into four components. The last two components, the inference bias and parameter estimation error, reflect the role of misspecification.

Theorem 1. *The generalization error has the following decomposition:*

$$\mathbb{E}\left[l(\widetilde{Y}, Y)\right] = \underbrace{H(Y|\boldsymbol{\lambda})}_{\text{Irreducible error}} - \underbrace{\mathbb{E}_{\mathcal{N}}\left[D_{\mathrm{KL}}(\Pr(\boldsymbol{\lambda})||\hat{\Pr}(\boldsymbol{\lambda}))\right]}_{\text{Observable sampling noise}} +$$

$$\sum_{(i,j)\in E_\lambda} \underbrace{I(\lambda_i;\lambda_j|Y)}_{\text{Inference bias}} + \sum_{i=1}^{m} \underbrace{\mathbb{E}_{Y,\mathcal{N},\tau}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right]}_{\text{Parameter estimation error}},$$

where $I(\lambda_i;\lambda_j|Y)$ is the conditional mutual information between sources and $H(Y|\boldsymbol{\lambda})$ is conditional entropy. Pr *refers to the true data distribution, while* $\hat{\Pr}$ *and* $\widetilde{\Pr}$ *refer to the estimated probabilities in* (1).

We now discuss each term above. The first two terms are independent of misspecification and are present in both the unlabeled and labeled cases:

- Irreducible error: an intrinsic property of the distribution of $(Y, \boldsymbol{\lambda})$ always present in bias-variance decomposition.
- Observable sampling noise: the expected KL divergence between the true marginal distribution of the observable sources and the empirical distribution. Particular to our inference approach, it is a common notion of sampling noise (Domingos, 2000; Yang et al., 2020) and approaches 0 asymptotically.

For the last two terms, misspecification plays a different role depending on the data type.

- Inference bias: the conditional mutual information among dependent sources. Particular to our inference approach, it is the approximation error of using marginal singleton probabilities rather than their product distributions. Therefore, it represents the role of misspecification at the inference step (1) and is present for both data types. It is independent of parameter estimation method.
- Parameter estimation error: the difference between the true and estimated distribution of $\lambda_i|Y$. For the labeled approach, this error corresponds to sampling noise and asymptotically approaches

0. For the unlabeled approach, it directly depends on the estimation error of accuracies in (2). However, these estimators are biased, as are many method-of-moments approaches. Furthermore, misspecification makes the estimators inconsistent when λ_i, λ_j, and λ_k used to produce $\widetilde{a}_i^{(j,k)}$ are not pairwise conditionally independent.

We now discuss in detail the scaling of these last two terms, which highlights the tradeoff between labeled and unlabeled data under misspecification.

4.2 Scaling of the Generalization Error

We bound the terms in Theorem 1 to understand the scaling of error due to misspecification in both the unlabeled and labeled cases. Since the irreducible error is always present, we bound *excess generalization error*, defined as $R_L^e = R_L - H(Y|\boldsymbol{\lambda})$ for labeled data and similarly R_U^e for unlabeled data. We use $\mathcal{B}_I = \sum I(\lambda_i; \lambda_j|Y)$ for the inference bias in these bounds since it is independent of our two cases, and while it scales in d, it is simply a measurement over the true data distribution. We present upper bounds here and a lower asymptotic bound in the Appendix. We first bound R_L^e.

Theorem 2. *Suppose that there are $|E_\lambda| = d$ unmodeled dependencies. When we use the latent variable model described in section 3 with n_L labeled samples,*

$$R_L^e \leq \frac{m}{2n_L} + \mathcal{B}_I + o(1/n_L). \tag{3}$$

In this bound, $\frac{m}{2n_L}$ is an upper bound on parameter estimation error. It represents the sampling noise of $\widetilde{a}_i^L = \hat{\mathbb{E}}[\lambda_i Y]$, which asymptotically approaches 0. Therefore, the only standing bias is $\mathcal{B}_I$ due to inference approach. When there is no model misspecification, the excess error is $\mathcal{O}(1/n_L)$, and thus for large n_L our generated labels would eventually follow the true $\Pr(Y|\boldsymbol{\lambda})$.

We next present an upper bound on the excess generalization error in the unlabeled case. Define $\varepsilon_{ij} = \mathbb{E}[\lambda_i\lambda_j] - \mathbb{E}[\lambda_i Y]\,\mathbb{E}[\lambda_j Y]$ as the extent of misspecification on a single pair of sources, and let $0 \leq \varepsilon_{\min} \leq \varepsilon_{ij} \leq \varepsilon_{\max}$ for all pairs (i,j) under our model assumptions in section 3. The exact value of ε_{ij} in terms of canonical parameters is in the Appendix.

Theorem 3. *Suppose that there are $|E_\lambda| = d$ dependencies. When we use the latent variable model described in section 3 using n_U unlabeled samples,*

$$\begin{aligned} R_U^e \leq & \varepsilon_{\max}\left(\frac{c_1 d}{m} + \frac{c_2}{\sqrt{n_U}} + \frac{c_3 d}{m n_U}\right) \\ & + \frac{c_4 m}{n_U} + \mathcal{B}_I + o(1/n_U), \end{aligned} \tag{4}$$

where c_1, c_2, c_3, and c_4 are constants depending on the intrinsic quality of the sources (Appendix).

In this bound, we again have an observable sampling noise $\frac{c_4 m}{n_U}$, where the difference in the constant term comes from estimating $\hat{\mathbb{E}}[\lambda_i\lambda_j]$ in (2) rather than $\hat{\mathbb{E}}[\lambda_i Y]$ in the labeled approach. However, here the parameter estimation error has an additional term $\mathcal{B}_{\text{est}} := \varepsilon_{\max}\left(\frac{c_1 d}{m} + \frac{c_2}{\sqrt{n_U}} + \frac{c_3 d}{m n_U}\right)$ which depends on misspecification. Therefore, asymptotically the unlabeled approach has a standing bias bounded by $\frac{c_1 d \varepsilon_{\max}}{m} + \mathcal{B}_I$ in comparison to the labeled case's $\mathcal{B}_I$, and the finite sample regime contributes additional sampling noise for the unlabeled approach that scales in $\varepsilon_{\max}$. In the case of no misspecification ($d = 0, \varepsilon_{\max} = 0$), the only term present is $\frac{c_4 m}{n_U}$, so our latent variable model would also approach the true distribution of $\Pr(Y|\boldsymbol{\lambda})$ but at a different rate.

Partial Recovery Our result holds almost exactly for the partial recovery case, where d' out of d dependencies are recovered via structure learning or some other approach, and our method in (2) avoids choosing dependent sources. In particular, the additional estimation error now scales at rate $\frac{(d-d')\varepsilon_{\max}}{m-2d'}$.

4.3 Correcting for misspecification

How can we reduce the penalty for dealing with such unrecovered dependencies? We examine how to reduce misspecification for our estimator described in (2), but our correction can be applied to other method-of-moments approaches (Anandkumar et al., 2012; Chaganty and Liang, 2014), discussed in Appendix.

In our estimation approach, if there exists an λ_i such that there are no λ_j, λ_k where all three sources are pairwise conditionally independent given Y, then it is not possible to learn a_i. In less demanding cases, we suggest an alternative approach based on *medians*. Recall that misspecification impacts accuracy estimation error because random triplets that violate pairwise conditional independence are selected to compute our $\widetilde{a}_i^U$. To reduce this impact, we estimate each a_i by computing the median accuracy over all pairs λ_j, λ_k using (2) a total of $\binom{m-1}{2}$ times.

Proposition 1. *Let $\widetilde{a}_i^M = \text{median}(\{\widetilde{a}_i^{(j,k)}\ \forall\ j,k \neq i\})$. Then $\widetilde{a}_i^M$ is not affected by misspecification and is thus a consistent estimator if $m > 5$, $d < \frac{(m-1)(m-2)}{4}$, and $n_U \geq n_0$, where n_0 is $\omega(1/\varepsilon_{\min}^2)$.*

Refer to $\rho_{n_U} = \max_i \mathbb{E}\left[(\widetilde{a}_i^M - a_i)^2\right]$ as the rate of convergence for $\widetilde{a}^M$. Under these conditions, the excess generalization error R_M^e from using n_U unlabeled samples and a corrected model is, for constant c_ρ,

$$R_M^e \leq c_\rho m \rho_{n_U} + \mathcal{B}_I + o(1/n_U) \tag{5}$$

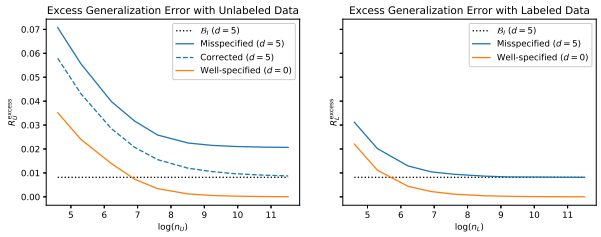


Figure 3: Excess generalization error vs. $\log(n)$ with different estimators for synthetic data. Left: comparison of unlabeled data performance under the three discussed settings. Right: comparison of labeled data performance for well-specified and misspecified models. A dashed line repesenting an empirical "$\mathcal{B}_I$" suggests how inference bias is present in both data cases.

While ρ can be analyzed in detail as a variant of the medians-of-means estimator, we stress that $\lim_{n_U \to \infty} \rho_{n_U} = 0$. Thus the standing bias of order $\mathcal{O}(d/m)$ due to misspecification can be eliminated. This reduction has significant implications for the value of labeled vs. unlabeled data in corrected settings.

4.4 Synthetic Experiments

We validate the fundamental principles of our theoretical framework using synthetic data. We measure the excess generalization error vs. $\log(n)$ in the well-specified, misspecified and corrected settings on synthetic data with $m = 10$ sources, accuracies drawn uniformly from $[.55, .75]$ and extents of misspecification fixed at $\varepsilon = 0.1$. To approximate expected excess generalization error for each n, we average results over 1000 samples. A more detailed protocol for synthetic experiments is available in the Appendix. Our results are in Figure 3. With no misspecification ($d = 0$) the labeled and unlabeled estimators both tend towards zero. Under misspecification ($d = 5$), we see that learning from unlabeled data results in an additional standing bias that parallels $\mathcal{B}_{\text{est}}$. Median aggregation reduces this bias and results in error converging to roughly similar values, paralleling $\mathcal{B}_I$, in both the unlabeled and labeled cases. These observations are consistent with our theoretical findings.

5 Applications

Based on our generalization error framework, we now have a rigorous way to analyze misspecification in latent variable models. We examine two practical applications of our theoretical results in three settings—well-specified, misspecified, and corrected.

- **Understanding the value of labeled data:** we address our motivating question about the value of labeled data–is a few labeled samples or many unlabeled samples better? This decision depends on the misspecification parameters (d, $\varepsilon_{\max}$), and n_U versus n_L.
- **Combining labeled and unlabeled data:** we show how simple linear combinations of the estimators can improve generalization error bounds over using one or the other. Then, we suggest a James-Stein type estimator from Green et al. (2005), which combines an unbiased estimator with biased information, to easily determine the weights of the linear combination.

5.1 Understanding the value of labeled data

We use our analysis from section 4.2 to develop a criterion for deciding between labeled and unlabeled points. Compute

$$\alpha(n_U) = \min_{n_L \in \mathbb{N}} \text{ s. t. } R_L^e(n_L) \leq R_U^e(n_U),$$

and define $V(n_U) = n_U/\alpha(n_U)$ to be the *data value ratio*. The intuitive idea here is to compare, for each amount of unlabeled data n_U, what factor less labeled data we would require to produce an equivalent error bound. We consider an approximation of the data value ratio $\widetilde{V}(n_U)$ based on our upper bounds for excess generalization error in 4.2. We examine the differences in $\widetilde{V}(n_U)$ for our three aforementioned settings:

- Well-specified setting: comparing excess risk when $d = 0$ and $\varepsilon_{\max} = 0$ reduces to examining $\frac{m}{2n_L}$ and $\frac{c_4 m}{n_U}$. Thus $\widetilde{V}(n_U) = 2c_4$ and our framework suggests that labeled data is only a constant factor more beneficial than unlabeled data.
- Misspecified setting: $\widetilde{V}(n_U)$ will capture the tradeoff between $\frac{m}{2n_L}$ and $\mathcal{B}_{\text{est}} + \frac{c_4 m}{n_U}$. We find that $\widetilde{V}(n_U) = 2\varepsilon_{\max}\left(\frac{c_1 d n_U}{m} + \frac{c_2\sqrt{n_U}}{m} + + \frac{c_3 d}{m^2}\right) + 2c_4$. That is, the value of labeled data *increases linearly in the amount of unlabeled data and misspecification* due to the standing bias in the generalization error for the unlabeled approach.
- Corrected setting: under our conditions from Proposition 1, we examine the difference between $\frac{m}{2n_L}$ and $c_\rho m \rho_{n_U}$, and thus $\widetilde{V}(n_U) = 2n_U c_\rho \rho_{n_U}$. Since ρ_{n_U} converges to 0, $\widetilde{V}(n_U)$ is sublinear in n_U, showing that the *corrected model increases the relative value of unlabeled data.*

Synthetic Experiments We measure $V(n)$ in well-specified, misspecified and corrected settings on synthetic data with the same setup as previously discussed. Our detailed protocol for approximating $V(n)$ is in the Appendix. We present the results in Figure 4. In the well-specified case ($d = 0$), $V(n)$ is small (less than 5) and roughly constant across n. Under misspecification however, the data value ratio grows with

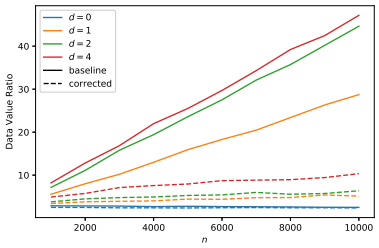


Figure 4: Data value ratio vs. n, using both the standard method-of-moments approach and the corrected approach, which aggregates results over triplets using medians. Note that $d = 0$ represents the well-specified setting.

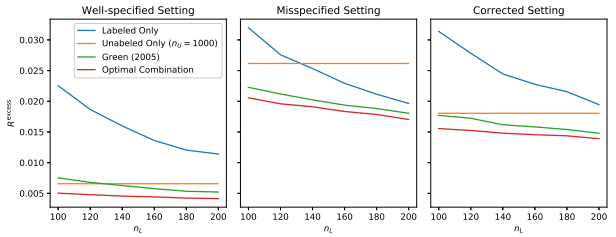


Figure 5: Excess generalization error for an optimally weighted combination of labeled and unlabeled estimators, and a combination weighted according to Green et al. (2005) across the well-specified (left), misspecified (center), and corrected (right) settings. The number of unlabeled points is fixed at $n_U = 1000$.

both d and n albeit much more slowly for the corrected setting, aligning with our theoretical findings.

5.2 Combining labeled and unlabeled data

While we now have a criterion to choose between datasets, how do we combine information from both? We examine ways to combine the accuracy parameters, namely $\widetilde{a}^U$ as defined in (2) for unlabeled data and an equivalent $\widetilde{a}^L := \hat{\mathbb{E}}[\boldsymbol{\lambda} Y]$ for labeled data. Recall that $\widetilde{a}^L$ is unbiased, while $\widetilde{a}^U$ is both biased and inconsistent.

First, we consider a simple linear combination, $a^{\text{lin}}(\alpha) = \alpha\widetilde{a}^U + (1-\alpha)\widetilde{a}^L$ for some $\alpha \in [0,1]$. Using our framework in 1, we can derive similar upper bounds on excess generalization error when the estimator is $a^{\text{lin}}(\alpha)$. We summarize our findings across the three settings below, where for the corrected setting we consider $\alpha\widetilde{a}^M + (1-\alpha)\widetilde{a}^L$.

- Well-specified setting: the upper bound on excess generalization error using a^{lin}, ignoring $\mathcal{B}_I$ and lower order terms, is $\alpha^2 \frac{c_4 m}{n_U} + (1-\alpha)^2 \frac{m}{2n_L}$. One can easily verify that there exists an $\alpha \in (0,1)$ that minimizes this upper bound. Since n_U is usually much larger than n_L, plugging in this optimal α shows that this new upper bound is roughly of the same order as the unlabeled case.
- Misspecified setting: the upper bound is a cubic polynomial in α. We find that the standing bias results in a generally lower optimal α and suggests that a combined estimator can yield an upper bound much smaller than that for the unlabeled case.
- Corrected setting: the upper bound now consists of $\alpha^2 c_\rho m \rho_{n_U} + (1-\alpha)^2 \frac{m}{2n_L}$. As a function of α, this differs from the well-specified setting's expression only in constant coefficients, so this again suggests an optimal $\alpha \in (0,1)$ and performance roughly similar to the unlabeled case.

In practice, we do not know the exact α that optimizes generalization error. However, there is vast literature on combined estimators that dominate the MLE estimator $\widetilde{a}^L$. In particular, we suggest using an approach from Green et al. (2005), who propose a way of setting α given knowledge of an unbiased estimator with biased information.

Synthetic Experiments We investigate the empirical performance of estimators which combine labeled and unlabeled data in well-specified, misspecified and corrected settings. We measure both the error when using the optimal α and the more practical approach of Green et al. (2005). We fix $n_U = 1000$ and vary n_L across a range of smaller values, aligning with the assumption that many more unlabeled than labeled points are typically available. Our results are in Figure 5. In the well-specified setting, the combined estimators perform roughly the same as just $\widetilde{a}^U$, matching up with our theoretical observations for large n_U. In the misspecified setting, both combined estimators result in much lower excess risk than either estimator individually, and as n_L increases, the labeled estimator curve approaches those of the combined estimators, suggesting that the weight on $\widetilde{a}^L$ increases as more labeled data becomes available. Lastly, in the corrected setting both combined estimators perform better than $\widetilde{a}^U$, but not by much. The weights α are reported in the Appendix. The optimal weights for the well-specified and corrected settings are higher (i.e. more weight on the unlabeled estimator) than the misspecified setting, and these weights decrease with n_L.

6 Real-World Case Study: Weak Supervision

We validate our findings on real-world weak supervision dataset. We expect that some amount of misspecification is inevitable, and that this causes additional bias when learning from only unlabeled data. Unlike our theoretical setting where we limit the number of

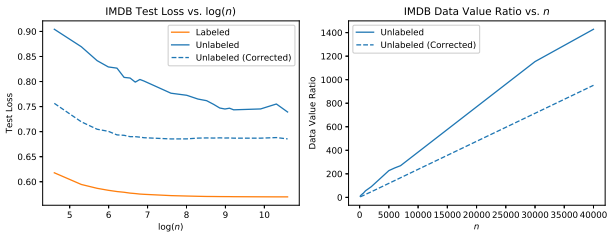
dependencies d for simplicity, with real-world data we anticipate many small dependencies which cannot be completely corrected by the medians approach. We seek to answer the following key questions.

- What is the standing parameter estimation bias due to misspecification? To what extent does the corrected estimator mitigate this bias?
- What is the data value ratio for the corrected estimator?
- Can a combined estimator with access to a small amount of labeled data provide substantial benefits over using only unlabeled data?

Protocol Our real-world task is the sentiment analysis task of determining whether IMDB movie reviews are positive or negative (Maas et al., 2011). The dataset contains 50K movie reviews, which we split into a training set of 40K reviews and a test set of 10K reviews. Our weak supervision sources are simple heuristics that vote "yes" when positive words appear and "no" when negative words appear. We provide further details in the Appendix.

Unlike our theoretical model, where we assume that each source has a single accuracy parameter, we find that real-world sources have complex dependencies and can be better modeled with *class conditional* accuracies. The method-of-moments approach in this setting results in a quadratic version of the triplet method (Fu et al., 2020), the details of which we discuss in the Appendix. We use this version for our real-world case study, for which the same principles from our theoretical framework apply.

Standing bias and correction For our first real-world experiment, we measure the standing parameter estimation bias when learning from unlabeled data (paralleling $\mathcal{B}_{\text{est}}$), and measure the decrease in bias when using a corrected estimator. Recall that in our model a corrected estimator has an asymptotic estimation bias of 0 given certain conditions on d, m, and n. On real data, however, we anticipate many small but complex dependencies that are not necessarily corrected for via our medians approach. Hence, we anticipate that the corrected estimator reduces $\mathcal{B}_{\text{est}}$ by mitigating the effects of larger dependencies while still being biased by the many small dependencies. We compute the test cross entropy loss for a labeled model, a baseline unlabeled and an unlabeled model with correction while varying n and report results in Figure 6. Losses appear to converge, with a large gap between the labeled and unlabeled models and a smaller gap between the labeled model and the unlabeled model with correction. These gaps in loss are reflected by gaps in F1-scores, computed using a threshold of .5.



Model	Loss ($n = 40$K)	F1 ($n = 40$K)
Labeled	.570	71.79
Unlabeled	.740	64.81
Corrected	.686	68.12

Figure 6: We measure test losses and F1-scores for labeled, unlabeled and corrected models on the IMDB dataset. Top Left: losses vs. n; each model appears to flatten out by $n = 40,000$. Bottom: losses and F1-scores at $n = 40,000$, showing standing gaps in performance. Top Right: data value ratios for the two unlabeled models.

n_U	n_L	$\text{F1}_{\text{Unlabeled}}$	$\text{F1}_{\text{Labeled}}$	$\text{F1}_{\text{Combined}}$
40,000	40	68.12	64.70	67.06
40,000	80	68.12	67.65	68.81
40,000	120	68.12	68.92	69.64
40,000	200	68.12	69.97	70.41
40,000	400	68.12	70.81	71.04

Table 1: F1-scores for unlabeled, labeled and combined approaches on the IMDB dataset. We find that the combination generally outperforms either approach individually, and in particular both in cases where unlabeled only performs better and where labeled only performs better.

Measuring the value of labeled data Next, we measure the data value ratio in the real-world setting. Since both the unlabeled model and the unlabeled model with correction have a standing bias compared to the labeled model, we anticipate that the data value ratio for both unlabeled approaches grows linearly with n, with the data value ratio for the baseline unlabeled model having a greater constant factor due to its higher bias. We report these results in Figure 6.

Combining labeled and unlabeled data We finally measure the performance of the combined estimator from Green et al. (2005) in the setting where a small number of labeled points and many unlabeled points are available. We let $n_U = 40,000$ be the entire training set and vary n_L between 40 and 400. We use the corrected estimator for learning from unlabeled data. We report the F1-score using a threshold of .5. Results are in Table 1. We observe that the combined estimator outperforms either approach individually for $n_L > 40$.

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832.

Anandkumar, A., Hsu, D., and Kakade, S. M. (2012). A method of moments for mixture models and hidden markov models. volume 23 of *Proceedings of Machine Learning Research*, pages 33.1–33.34, Edinburgh, Scotland. JMLR Workshop and Conference Proceedings.

Bach, S. H., He, B., Ratner, A., and Ré, C. (2017). Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 273–282. JMLR. org.

Chaganty, A. T. and Liang, P. (2014). Estimating latent-variable graphical models using moments and likelihoods. In *International Conference on Machine Learning*, pages 1872–1880.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4):1935–1967.

Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *International Conference on Intelligent Systems for Molecular Biology (ISMB)*.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28.

De Blasi, P. and Walker, S. G. (2013). Bayesian asymptotics with misspecified models. *Statistica Sinica*, pages 169–187.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pages 231–238.

Fu, D. Y., Chen, M. F., Sala, F., Hooper, S. M., Fatahalian, K., and Ré, C. (2020). Fast and three-rious: Speeding up weak supervision with triplet methods. *arXiv preprint arXiv:2002.11955*.

Green, E. J., Strawderman, W. E., Amateis, R. L., and Reams, G. A. (2005). Improved Estimation for Multiple Means with Heterogeneous Variances. *Forest Science*, 51(1):1–6.

Grünwald, P., Van Ommen, T., et al. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Gupta, S. and Manning, C. D. (2014). Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 98–108.

Hsu, D., Kakade, S. M., and Liang, P. (2012). Identifiability and unmixing of latent parse trees. In *Advances in Neural Information Processing Systems, (NIPS 2012)*.

Jog, V. and Loh, P. (2015). On model misspecification and KL separation for gaussian graphical models. *CoRR*, abs/1501.02320.

Joglekar, M., Garcia-Molina, H., and Parameswaran, A. (2013). Evaluating the crowd with confidence. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 686–694.

Karger, D. R., Oh, S., and Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961.

Kleijn, B. and van der Vaart, A. (2012). The bernstein-von-mises theorem under misspecification. *Electron. J. Statist.*, 6:354–381.

Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional bayesian statistics. *Ann. Statist.*, 34(2):837–877.

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press.

Loh, P.-L. and Wainwright, M. J. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Annals of Statistics*, 41(6):3022–3049.

Maas, A., 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*, pages 142–150.

Meng, Z., Eriksson, B., and III, A. O. H. (2014). Learning latent variable gaussian graphical models. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, Beijing, China.

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: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

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., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. (2016). Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Takamatsu, S., Sato, I., and Nakagawa, H. (2012). Reducing wrong labels in distant supervision for relation extraction. In *Meeting of the Association for Computational Linguistics (ACL)*.

Varma, P., Sala, F., He, A., Ratner, A., and Ré, C. (2019). Learning dependency structures for weak supervision models. *arXiv preprint arXiv:1903.05844*.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks.

Supplementary Materials

A Glossary

The glossary is given in Table 2 below.

Symbol	Used for
X	An input vector $X \in \mathcal{X}$.
Y	A latent ground-truth label $Y \in \mathcal{Y} = \{-1, 1\}$.
m	Number of sources.
λ_j	jth source output $\lambda_j : \mathcal{X} \to \mathcal{Y}$; all m labels make up vector $\boldsymbol{\lambda}$
$\widetilde{Y}$	Probabilistic label in $[-1, 1]$ output by the latent variable model.
n_U	Number of unlabeled samples.
n_L	Number of labeled samples.
Θ	Canonical parameters of the Ising model for $\Pr(Y, \boldsymbol{\lambda})$.
G	Dependency graph $G = (V, E)$ over sources and the latent ground-truth label.
E_λ	Edges among sources in G.
d	Number of dependencies among sources $d = \lvert E_\lambda \rvert$.
a_i	True accuracy of the ith source $\mathbb{E}[\lambda_i Y]$.
$\widetilde{a}_i^U$	Estimated accuracy of the ith source using unlabeled data via the triplet method.
$\widetilde{a}_i^L$	Estimated accuracy of the ith source using labeled data, i.e. $\hat{\mathbb{E}}[\lambda_i Y]$.
$\widetilde{a}_i^M$	Estimated accuracy of the ith source using unlabeled data via the triplet method and median aggregation.
$\mathcal{N}$	Random variable representing dataset used.
τ	Algorithmic randomness for estimating accuracies via triplet method.
R, R_U, R_L, R_M	Generalization error $R = \mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}[l(\widetilde{Y}, Y)]$. R_U, R_L, R_M are for $\widetilde{a}_i^U, \widetilde{a}_i^L, \widetilde{a}_i^M$, respectively, and $l(\cdot, \cdot)$ is the cross-entropy loss.
R^e, R_U^e, R_L^e, R_M^e	Excess generalization error $R^e = R - H(Y \lvert \boldsymbol{\lambda})$.
$\mathcal{B}_I$	Inference bias $\mathcal{B}_I = \sum_{(i,j)\in E_\lambda} I(\lambda_i; \lambda_j \lvert Y)$.
$\mathcal{B}_{\text{est}}$	Parameter estimation error.
ε_{ij}	Extent of misspecification on a single pair of sources $\varepsilon_{ij} = \mathbb{E}[\lambda_i \lambda_j] - \mathbb{E}[\lambda_i Y]\,\mathbb{E}[\lambda_j Y]$.
$\varepsilon_{\min}, \varepsilon_{\max}$	Smallest and largest ε_{ij} for $(i, j) \in E_\lambda$.
ρ_{n_U}	Rate of convergence for $\tilde{a}_i^M$, $\rho_{n_U} = \max_i \mathbb{E}\left[(\widetilde{a}_i^M - a_i)^2\right]$.
$\alpha(n_U)$	Minimum labeled points needed for lower generalization error than n_U unlabeled points.
$V(n_U)$	Data value ratio at n_U unlabeled points.
$\widetilde{V}(n_U)$	Approximation of data value ratio using upper bounds at n_U unlabeled points.
α	Weight for unlabeled estimator to combine unlabeled and labeled estimators.
$a^{\text{lin}}(\alpha)$	Linear combination of unlabeled and labeled estimators using weight α.

Table 2: Glossary of variables and symbols used in this paper.

B Additional Theoretical Results

In this section, we discuss how our generalization error bounds, namely the standing $\mathcal{O}(d/m)$ bias for unlabeled data, and our results for the corrected medians estimator can still apply to other method-of-moments estimators that exploit conditionally independent views of hidden variables. Next, we present a lower asymptotic bound on the generalization error for labeled versus unlabeled data. Finally, we give more details about the combined estimators and the generalization bounds from using them.

B.1 Other Method-of-Moments Estimators

We present two other method-of-moments estimators and sketch out arguments for how using them (under misspecification) results in the same scaling of generalization error, and for how the median approach is able to help correct standing bias. We then provide an abstracted argument.

"Quadratic" Triplets This alternative latent variable model relies on class-conditional probability terms instead of mean parameters (Fu et al., 2020), which assume some symmetries in the distribution (see Lemma 2). For the ith source, we can write the parameters to be estimated as

$$\mu_i = \begin{bmatrix} \Pr(\lambda_i = 1|Y = 1) & \Pr(\lambda_i = 1|Y = -1) \\ \Pr(\lambda_i = -1|Y = 1) & \Pr(\lambda_i = -1|Y = -1) \end{bmatrix}.$$

Let

$$O_{ij} = \begin{bmatrix} \Pr(\lambda_i = 1, \lambda_j = 1) & \Pr(\lambda_i = 1, \lambda_j = -1) \\ \Pr(\lambda_i = -1, \lambda_j = 1) & \Pr(\lambda_i = -1, \lambda_j = -1) \end{bmatrix} \text{ and } P = \begin{bmatrix} \Pr(Y = 1) & 0 \\ 0 & \Pr(Y = -1) \end{bmatrix}.$$

Then, we obtain that

$$O_{ij} = \mu_i P \mu_j^\top. \tag{6}$$

The left-hand side is observable, and we can form triplets again to solve for each μ_i. Set $\alpha = P(\lambda_i = 1|Y = 1)$, $c_i = \frac{P(\lambda_i=1)}{P(Y=-1)}$ and $d_i = \frac{P(Y=1)}{P(Y=-1)}$. The top row of μ_i is then $[\alpha \quad c_i - d_i\alpha]$ with c_i and d_i known. For a triplet i, j, k, and the appropriate μ's, using the α, β, γ notation above and corresponding c_i, c_j, c_k and d_i, d_j, d_k terms, we obtain the system (see Fu et al. (2020) for more details)

$$\begin{aligned}
(1 + d_i d_j)\alpha\beta + c_i c_j - c_i d_j \beta - c_j d_i \alpha &= O_{ij}/\Pr(Y = 1), \\
(1 + d_i d_k)\alpha\gamma + c_i c_k - c_i d_k \gamma - c_k d_i \alpha &= O_{ik}/\Pr(Y = 1), \\
(1 + d_j d_k)\beta\gamma + c_j c_k - c_j d_k \gamma - c_k d_j \beta &= O_{jk}/\Pr(Y = 1).
\end{aligned}$$

To solve, α and γ are expressed with β for the first and third equations and this is plugged into the second—yielding a quadratic equation to be solved.

This approach incurs standing bias under misspecification. Quadratic triplets rely on conditional independence by assuming that $\Pr(\lambda_i = 1, \lambda_j = 1)$ and $\Pr(\lambda_i = 1|Y = 1)\Pr(\lambda_j = 1|Y = 1)\Pr(Y = 1) + \Pr(\lambda_i = 1|Y = -1)\Pr(\lambda_j = 1|Y = -1)\Pr(Y = -1)$ are equal. Suppose, however, that $(i, j) \in E_\lambda$. Then, $\mu_i P \mu_j^\top$ is no longer equal to O_{ij}, but $O_{ij} + \delta_{ij}$, where $\delta_{ij} = \Pr(Y = 1)[\Pr(\lambda_i|Y = 1)\Pr(\lambda_j|Y = 1) - \Pr(\lambda_i, \lambda_j|Y = 1)] + \Pr(Y = -1)[\Pr(\lambda_i|Y = -1)\Pr(\lambda_j|Y = -1) - \Pr(\lambda_i, \lambda_j|Y = -1)]$. This δ_{ij} can be written exactly in terms of the canonical parameters Θ and results in an inconsistent estimator of $\Pr(\lambda_i|Y)$. We note that the probability of selecting a bad triplet that leads to this is the same for this method and our main triplet method, so the standing bias still scales $\mathcal{O}(\frac{d\delta}{m})$.

This approach can also be corrected using medians. Out of $\binom{m-1}{2}$ triplets used in estimating $\Pr(\lambda_i|Y)$, there are $\binom{m-1}{2} - m - d - 3$ triplets that result in a consistent estimate. So as long as $\binom{m-1}{2} - m - d - 3 > \frac{1}{2} \cdot \binom{m-1}{2}$ and n_U is sufficiently large, using medians will result in a corrected estimator. See the proof of Proposition 1 in section C.4 for more details.

Method-of-moments for topic exchange Anandkumar et al. (2014) describes tensor method-of-moments estimators for a variety of applications, including topic models. In the topic model case, h is the topic latent variable, $x_1, \ldots, x_\ell$ are the words in the document, all assumed to be conditionally independent given h and drawn from an unknown conditional probability distribution μ_h parametrized by the latent topic variable. Here, $x_t = e_i$, the standard basis vector if the tth word is i. Anandkumar et al. (2014) uses the fact that

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i,$$

where w_i is the probability of h being topic i, to perform a tensor decomposition of the observable $\mathbb{E}[x_1 \otimes x_2 \otimes x_3]$ and learn μ_h. Note the similarity to our setting, where Y is used in place of h and where there are two (i.e., a matrix) instead of three views (giving a tensor). Conditional independence (of words given the topic) is required to for this expression to hold. Therefore, when conditional independence is violated, $\sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i$ is equal to $\mathbb{E}\left[x_1 \otimes x_2 \otimes x_3\right]$ plus some additional perturbation that is a function of the probability distribution. This error is propagated into the estimate of μ_h, and we assume Lipschitzness of this estimator. Furthermore, assuming random triples are selected to learn the accuracy of each word, using this approach to estimate accuracy parameters will again yield a standing bias.

Furthermore, the medians approach can again correct for this standing bias—there are $\binom{m-1}{2} - m - d - 3$ good triplets out of $\binom{m-1}{2}$, so we require the same conditions to yield consistent estimators as those for the quadratic triplets case.

Abstraction Consider in general some observable quantities $o_1, \ldots, o_v$, some unobservable quantities $u_1, \ldots, u_v$ that depend on the value of some latent variable h, and a relationship that holds when some set of dependencies Ω is taken into account,

$$f(o_1, \ldots, o_v) = g_\Omega(u_1, \ldots, u_v),$$

Next, we call $s(f(o_1, \ldots, o_v))$ an estimator that produces estimates of $u_1, \ldots, u_v$.

Our approach is simply to account for errors due to accessing an incorrect Ω', where $|\Omega \setminus \Omega'| = d$. Then,

$$f(o_1, \ldots, o_v) = g_{\Omega'}(u_1, \ldots, u_v) + d \times \Delta(u_1, \ldots, u_v),$$

where Δ is some error term. Given this setup, we then propagate the error term Δ in the estimator s, computing $s(f(o_1, \ldots, o_v)) - s(f(o_1, \ldots, o_v) - d\Delta(u_1, \ldots, u_v)$. This can be done either via perturbation analysis or Taylor approximation or other methods—the only requirement we place is Lipschitzness on the estimator s. Then, by randomly selecting subsets of $(o_1, \ldots, o_v)$ to estimate $u_1, \ldots, u_v$, the probability of picking a subset with error scales in d, showing that there exists a standing bias that is a function of the number of unmodeled dependencies. Moreover, there are some subsets of $(o_1, \ldots, o_v)$ that yield consistent estimators s; if this quantity is greater than half of all the subsets, then a medians approach can be beneficial when there is enough data.

B.2 Asymptotic lower bounds on generalization error

While Theorems 2 and 3 provide upper bounds on the excess generalization error, it is also important to consider the asymptotic lower bound—is the standing bias from misspecification in the unlabeled approach inevitable?

Looking at the decomposition in Theorem 1, $\mathbb{E}_{\mathcal{N}}\left[D_{\mathrm{KL}}(\Pr(\boldsymbol{\lambda})||\hat{\Pr}(\boldsymbol{\lambda}))\right]$ approaches 0 asymptotically. We thus seek to asymptotically lower bound $\sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right]$. Note that in the labeled data case, parameter estimation error approaches 0 as n grows large since the observable estimated accuracy is both unbiased and consistent. In the unlabeled data case, we show that standing bias persists.

Theorem 4. *Suppose that there are $|E_\lambda| = d$ unmodeled dependencies. When we use the latent variable model described in section 3, the lower bound of the excess generalization error is asymptotically bounded by*

$$\lim_{n_U \to \infty} R_u^e \geq \frac{(m-2d)d^2\varepsilon_{\min}^2 b_{\min}^4}{2(m-1)^2(m-2)^2} + \mathcal{B}_I. \tag{7}$$

When d is $o(m)$, the asymptotic parameter estimation error is $\Omega\left(\frac{d^2\varepsilon_{\min}^2}{m^3}\right)$.

Proof. We compute an asymptotic lower bound for $\sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right]$. Applying Lemma 3, we see that

$$\lim_{n_U \to \infty} \sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right] = \sum_{i=1}^{m} \frac{1+a_i}{2}\log\left(1 + \frac{a_i - \bar{a}_i}{1+\bar{a}_i}\right) + \frac{1-a_i}{2}\log\left(1 + \frac{\bar{a}_i - a_i}{1-\bar{a}_i}\right) \tag{8}$$

We focus on the lower bound of any one element of this sum. For ease of notation, let $a := a_i$ and $x = a_i - \bar{a}_i$. Then this expression for an arbitrary i becomes

$$\frac{1+a_i}{2}\log\left(1+\frac{a_i-\bar{a}_i}{1+\bar{a}_i}\right)+\frac{1-a_i}{2}\log\left(1+\frac{\bar{a}_i-a_i}{1-\bar{a}_i}\right)=-\frac{1+a}{2}\log\left(1-\frac{x}{1+a}\right)-\frac{1-a}{2}\log\left(1+\frac{x}{1-a}\right) \quad (9)$$

Take the negative of this expression and define it as a function $f(x)$ to upper bound:

$$f(x)=\frac{1+a}{2}\log\left(1-\frac{x}{1+a}\right)+\frac{1-a}{2}\log\left(1+\frac{x}{1-a}\right) \quad (10)$$

We show that $f(x) \leq -\frac{1}{2}x^2$. Note that for $x = 0$, $f(x) = 0$ and $\frac{1}{2}x^2 = 0$. Then, we must show that for $x \geq 0$, $f'(x) \leq -x$ and for $x < 0$, $f'(x) > -x$. Taking the derivative of $f(x)$ gives us $f'(x) = \frac{-x}{1-(a-x)^2}$, and it is clear that the previous inequalities are satisfied.

Using this fact in (8), we have that $\lim_{n_U\to\infty}\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] = \sum_{i=1}^m \frac{1}{2}(a_i-\bar{a}_i)^2$. For $i \in E_\lambda$, note that by Lemma 1 it is possible to construct a graphical model such that $a_i - \bar{a}_i = 0$. For $i \in E_\lambda$, we know that $|a_i - \bar{a}_i|$ is at least $\frac{d\varepsilon_{\min}b_{\min}^2}{(m-1)(m-2)}$. Therefore,

$$\frac{1}{2}\sum_{i=1}^m(a_i-\bar{a}_i)^2 \geq \frac{1}{2}\sum_{i\notin E_\lambda}(a_i-\bar{a}_i)^2 \geq \frac{(m-2d)d^2\varepsilon_{\min}^2 b_{\min}^4}{2(m-1)^2(m-2)^2}. \quad (11)$$

□

B.3 Combined estimator analysis

The general form of the combined estimator we consider is $a^{\mathrm{lin}}(\alpha) = \alpha\widetilde{a}^U + (1-\alpha)\widetilde{a}^L$ for some weight $\alpha \in [0,1]$. The James-Stein type estimator from Green et al. (2005), which we evaluate empirically, uses the following:

$$\bar{a} := \widetilde{a}^U + \left(1-\frac{r}{\|\widetilde{a}^L-\widetilde{a}^U\|_{\Sigma^{-1}}}\right)_+(\widetilde{a}^L-\widetilde{a}^U), \quad (12)$$

where $\Sigma = \mathbf{Cov}\left[\widetilde{a}^L\right]$ and $r \in [0, 2(m-2)]$. Note that this is almost equivalent to $a^{\mathrm{lin}}\left(\frac{r}{\|\widetilde{a}^L-\widetilde{a}^U\|_{\Sigma^{-1}}}\right)$. Green et al. (2005) show that this estimator dominates $\widetilde{a}^L$ when the unbiased estimator is Gaussian and its covariance is known. However, since we can only estimate the covariance matrix, we replace Σ with an empirical estimate $\hat{\Sigma}$ in practice. Moreover, since $\widetilde{a}^L$ is only Gaussian asymptotically, we do not provide theoretical guarantees on $\bar{a}$. We instead focus on analyzing the performance of the general combined estimator $a^{\mathrm{lin}}(\alpha)$.

The change in estimator only impacts the generalization bound via the parameter estimation error, $\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right]$. We simplify this using Lemma 3, doing a Taylor approximation on a combined asymptotic estimate $\bar{a}_i^C := \alpha\bar{a}_i + (1-\alpha)a_i$ rather than $\bar{a}_i$. This gives us

$$\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] = \sum_{i=1}^m \frac{1+a_i}{2}\log\left(1+\frac{\alpha(a_i-\bar{a}_i)}{1+\bar{a}_i^C}\right)+\frac{1-a_i}{2}\log\left(1+\frac{\alpha(\bar{a}_i-a_i)}{1-\bar{a}_i^C}\right) \quad (13)$$

$$+\sum_{i=1}^m \frac{a_i-\bar{a}_i}{1-(\bar{a}_i^C)^2}\alpha^2\mathbb{E}\left[\bar{a}_i-\widetilde{a}_i^U\right]+\sum_{i=1}^m\frac{1}{2}\left(\frac{1}{1-(\bar{a}_i^C)^2}+\frac{2\alpha(\bar{a}_i-a_i)}{(1-(\bar{a}_i^C)^2)^2}\right)\left(\alpha^2\mathbb{E}\left[(\widetilde{a}_i^U-\bar{a}_i)^2\right]+(1-\alpha)^2\mathbb{E}\left[(\widetilde{a}_i^L-a_i)^2\right]\right) \quad (14)$$

We present bounds for the three settings discussed in the paper.

Well-specified setting In the well-specified setting, the unlabeled data accuracy estimator is consistent, so $\bar{a}_i = a_i$, and therefore

$$\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] = \sum_{i=1}^m\frac{1}{2}\left(\frac{1}{1-a_i^2}\right)\left(\alpha^2\mathbb{E}\left[(\widetilde{a}_i^U-\bar{a}_i)^2\right]+(1-\alpha)^2\mathbb{E}\left[(\widetilde{a}_i^L-a_i)^2\right]\right) \quad (15)$$

Using the results of the proof of Theorem 2 and the bound on $\mathbb{E}\left[(\widetilde{a}_i^U - \bar{a}_i)^2\right]$ in Lemma 6, we get that this is at most $\alpha^2 \frac{c_4 m}{n_U} + (1-\alpha)^2 \frac{m}{2n_L}$.

Misspecified Setting The constant terms for the bound on accuracy parameter estimation error will change due to $\bar{a}_i^C$ in the denominator rather than $\bar{a}_i$, but the derivation follows our proof for Theorem 3. Therefore, for some c',

$$\sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] \leq \varepsilon_{\max}\Big(\frac{c_1'\alpha d}{m} + \frac{c_2'\alpha^2}{\sqrt{n_U}} + \frac{c_3'\alpha^3 d}{m n_U} + \frac{\alpha(1-\alpha)^2 c_5' d}{m n_L}\Big) + \frac{c_4'\alpha^2 m}{n_U} + \frac{(1-\alpha)^2 m}{2n_L}.$$

Corrected Setting Here we consider the combined estimator $\alpha\widetilde{a}^M + (1-\alpha)\widetilde{a}^L$. Under certain conditions, we know that $\widetilde{a}^M$ asymptotically converges to a. Therefore, the accuracy parameter estimation error is

$$\sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] = \sum_{i=1}^{m} \frac{1}{2}\Big(\frac{1}{1-a_i^2}\Big)\Big(\alpha^2\mathbb{E}\left[(\widetilde{a}_i^M - \bar{a}_i)^2\right] + (1-\alpha)^2\mathbb{E}\left[(\widetilde{a}_i^L - a_i)^2\right]\Big) \tag{16}$$

$\mathbb{E}\left[(\widetilde{a}_i^M - \bar{a}_i)^2\right]$ is just the variance of the median estimator. Therefore, this summation is bounded by $\alpha^2 c_\rho m \rho_{n_U} + (1-\alpha)^2 \frac{m}{2n_L}$ under the conditions in Proposition 1.

C Proofs

First, we formally state our assumptions on the graphical model that are needed for our results.

Assumption 1. *Suppose that the distribution of* $\mathrm{Pr}(Y, \boldsymbol{\lambda})$ *takes on the form*

$$\mathrm{Pr}(Y, \boldsymbol{\lambda}) = \frac{1}{Z}\exp\Big(\theta_Y + \sum_{i=1}^{m}\theta_i\lambda_i Y + \sum_{(i,j)\in E_\lambda}\theta_{ij}\lambda_i\lambda_j\Big), \tag{17}$$

where Z *is the cumulant function, and all canonical parameters* Θ *are positive. This assumption also means that* $\mathbb{E}\left[\lambda_i\lambda_j\right], \mathbb{E}\left[\lambda_i Y\right] > 0$ *for all* i *and* j*. Define* $a_{\min} = \min_i a_i$ *as the minimum true accuracy. Define* $b_{\min} = \min_{i,j}\{\mathbb{E}\left[\lambda_i\lambda_j\right], \hat{\mathbb{E}}\left[\lambda_i\lambda_j\right]\}$*. Lastly, define* $\bar{a}_{\max} = \max_i \bar{a}_i = \max_{i,j,k}\mathbb{E}_\tau\left[\sqrt{\frac{\mathbb{E}[\lambda_i\lambda_j]\mathbb{E}[\lambda_i\lambda_k]}{\mathbb{E}[\lambda_j\lambda_k]}}\right]$.

C.1 Proof of Theorem 1

Our goal is to evaluate $\mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[l(\widetilde{Y}, Y)\right]$, where $\mathcal{N}$ is the randomness over a sample of n points (either n_U or n_L). This expected cross entropy loss can be written as

$$\mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[l(\widetilde{Y}, Y)\right] = -\mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[\log\frac{\widetilde{\mathrm{Pr}}(Y' = Y|\boldsymbol{\lambda}' = \boldsymbol{\lambda})}{\mathrm{Pr}(Y' = Y|\boldsymbol{\lambda}' = \boldsymbol{\lambda})}\right] + H(Y|\boldsymbol{\lambda}) \tag{18}$$

where Y', Y and $\boldsymbol{\lambda}', \boldsymbol{\lambda}$ are independent copies, and the conditional entropy $H(Y|\boldsymbol{\lambda})$ is by definition

$$H(Y|\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\lambda}}\left[-\mathrm{Pr}(Y = 1|\boldsymbol{\lambda}' = \boldsymbol{\lambda})\log\mathrm{Pr}(Y = 1|\boldsymbol{\lambda}' = \boldsymbol{\lambda}) - \mathrm{Pr}(Y = -1|\boldsymbol{\lambda}' = \boldsymbol{\lambda})\log\mathrm{Pr}(Y = 1|\boldsymbol{\lambda}' = \boldsymbol{\lambda})\right]. \tag{19}$$

Next, we evaluate $\log\frac{\widetilde{\mathrm{Pr}}(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\mathrm{Pr}(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}$. Define $\overline{\mathrm{Pr}}$ to be the conditionally independent label model parametrized by the true accuracies $a = \mathbb{E}\left[\boldsymbol{\lambda}Y\right]$ in the asymptotic regime; similar to $\widetilde{\mathrm{Pr}}$'s definition in (1),

$$\overline{\mathrm{Pr}}(Y' = Y|\boldsymbol{\lambda} = \boldsymbol{\lambda}(X)) = \frac{\overline{\mathrm{Pr}}(\boldsymbol{\lambda} = \boldsymbol{\lambda}(X)|Y' = Y)\,\mathrm{Pr}(Y' = Y)}{\mathrm{Pr}(\lambda = \lambda(X))} = \frac{\prod_{i=1}^{m}\mathrm{Pr}(\lambda_i = \lambda_i(X)|Y' = Y)\,\mathrm{Pr}(Y = 1)}{\mathrm{Pr}(\lambda = \lambda(X))} \tag{20}$$

Then,

$$\log \frac{\widetilde{\Pr}(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\Pr(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})} = \log \frac{\widetilde{\Pr}(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\overline{\Pr}(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})} + \log \frac{\overline{\Pr}(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\Pr(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}$$
$$= \sum_{i=1}^{m} \log \frac{\widetilde{\Pr}(\lambda_i'=\lambda_i|Y'=Y)}{\Pr(\lambda_i'=\lambda_i|Y'=Y)} + \log \frac{\Pr(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\hat{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda})} + \log \frac{\overline{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda}|Y'=Y)}{\Pr(\boldsymbol{\lambda}'=\boldsymbol{\lambda}|Y'=Y)}.$$

We have used the fact that the class balance $\Pr(Y'=Y)$ is the same value across the true distribution, $\widetilde{\Pr}$, and $\overline{\Pr}$. Plugging back into (18), we get

$$-\sum_{i=1}^{m} \mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[\log \frac{\widetilde{\Pr}(\lambda_i'=\lambda_i|Y'=Y)}{\Pr(\lambda_i'=\lambda_i|Y'=Y)}\right] - \mathbb{E}_{(Y,\boldsymbol{\lambda})}\left[\log \frac{\overline{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda}|Y'=Y)}{\Pr(\boldsymbol{\lambda}'=\boldsymbol{\lambda}|Y'=Y)}\right] - \mathbb{E}_{\boldsymbol{\lambda},\mathcal{N}}\left[\log \frac{\Pr(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\hat{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}\right] + H(Y|\boldsymbol{\lambda}) \tag{21}$$

We simplify each expectation now.

1. $-\sum_{i=1}^{m} \mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[\log \frac{\widetilde{\Pr}(\lambda_i'=\lambda_i|Y'=Y)}{\Pr(\lambda_i'=\lambda_i|Y'=Y)}\right]$:
   By definition of conditional KL divergence,

$$-\sum_{i=1}^{m} \mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[\log \frac{\widetilde{\Pr}(\lambda_i'=\lambda_i|Y'=Y)}{\Pr(\lambda_i'=\lambda_i|Y'=Y)}\right] = \sum_{i=1}^{m} \mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[\log \frac{\Pr(\lambda_i'=\lambda_i|Y'=Y)}{\widetilde{\Pr}(\lambda_i'=\lambda_i|Y'=Y)}\right] \tag{22}$$
$$= \sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau}\left[\mathbb{E}_Y\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i'|Y})\right]\right]. \tag{23}$$

2. $-\mathbb{E}_{(Y,\boldsymbol{\lambda})}\left[\log \frac{\overline{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda}|Y'=Y)}{\overline{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda}|Y'=Y)}\right]$:
   The key difference between $\overline{\Pr}$ and Pr is how the distributions factorize. The above expression can be written as

$$-\sum_{(i,j)\in E_\lambda} \mathbb{E}_{\lambda_i\lambda_j,Y}\left[\log \frac{\Pr(\lambda_i'=\lambda_i|Y'=Y)\Pr(\lambda_j'=\lambda_j|Y'=Y)}{\Pr(\lambda_i',\lambda_j'=\lambda_i,\lambda_j|Y'=Y)}\right]$$
$$= \sum_{(i,j)\in E_\lambda} \mathbb{E}_{\lambda_i,\lambda_j}\left[\log \frac{\Pr(\lambda_i',\lambda_j'=\lambda_i,\lambda_j|Y=1)}{\Pr(\lambda_i'=\lambda_i|Y=1)\Pr(\lambda_j'=\lambda_j|Y=1)}\middle| Y=1\right]\Pr(Y=1)$$
$$+ \mathbb{E}_{\lambda_i,\lambda_j}\left[\log \frac{\Pr(\lambda_i',\lambda_j'=\lambda_i,\lambda_j|Y=-1)}{\Pr(\lambda_i'=\lambda_i|Y=-1)\Pr(\lambda_j'=\lambda_j|Y=-1)}\middle| Y=-1\right]\Pr(Y=-1).$$

   Note that these expectations are equal to the mutual information between λ_i and λ_j conditional on $Y=1$ or $Y=-1$. Then by definition, the expression is equal to

$$\sum_{(i,j)\in E_\lambda} I(\lambda_i;\lambda_j|Y=1)\Pr(Y=1) + I(\lambda_i;\lambda_j|Y=-1)\Pr(Y=-1) = \sum_{(i,j)\in E_\lambda} I(\lambda_i;\lambda_j|Y).$$

3. $-\mathbb{E}_{\boldsymbol{\lambda},\mathcal{N}}\left[\log \frac{\Pr(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\widetilde{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}\right]$:
   This term is the expected negative KL divergence between the true and estimated distributions of $\boldsymbol{\lambda}$, $\mathbb{E}_{\mathcal{N}}\left[D_{\mathrm{KL}}(\Pr(\boldsymbol{\lambda})||\hat{\Pr}(\boldsymbol{\lambda}))\right]$. While there are many ways to estimate this distribution, we stick with simply the MLE estimate so that this expression will converge to 0 asymptotically.

Therefore, (21) becomes

$$\mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[l(\widetilde{Y},Y)\right] = H(Y|\boldsymbol{\lambda}) - \mathbb{E}_{\mathcal{N}}\left[D_{\mathrm{KL}}(\Pr(\boldsymbol{\lambda})||\hat{\Pr}(\boldsymbol{\lambda}))\right] + \sum_{(i,j)\in E_\lambda} I(\lambda_i;\lambda_j|Y) + \sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right],$$

C.2 Proof of Theorem 2

Our goal is to evaluate $\sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right]$ on a labeled dataset. Using Lemma 3, note that $\mathbb{E}\left[\widetilde{a}_i^L\right] = \bar{a}_i = a_i$. Therefore,

$$\begin{aligned}\mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] &= \frac{1+a_i}{2}\cdot\frac{1}{2(1+a_i)^2}\mathbb{E}\left[(\widetilde{a}_i^L - a_i)^2\right] + \frac{1-a_i}{2}\cdot\frac{1}{2(1-a_i)^2}\mathbb{E}\left[(\widetilde{a}_i^L - a_i)^2\right] + o(1/n)\\ &= \frac{1}{2(1-a_i^2)}\mathrm{Var}\left(\widetilde{a}_i^L\right) + o(1/n)\end{aligned}$$

It can be shown that this is exactly $\frac{1}{2n_L}$. To see this, formally define $\widetilde{a}_i^L = \frac{1}{n_L}\sum_{j=1}^{n_L}\lambda_i^j Y^j$, where λ_i^j, Y^j belong the jth sample of the dataset. Then $\mathrm{Var}\left(\widetilde{a}_i^L\right) = \frac{1}{n_L^2}\sum_{j=1}^{n_L}\mathrm{Var}\left(\lambda_i^j Y^j\right) = \frac{1}{n_L^2}\sum_{j=1}^{n_L}\mathbb{E}\left[\lambda_i^{j2}Y^{j2}\right] - \mathbb{E}\left[\lambda_i Y\right]^2 = \frac{1-a_i^2}{n_L}$. Therefore, $\sum_{i=1}^{m}\mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] = \frac{m}{2n_L} + o(1/n_L)$, and our proof is complete.

C.3 Proof of Theorem 3

We restate the full theorem with the value of the constants. Under assumption 1, using n_U weakly labeled samples and a misspecified model yields excess generalization error

$$R_U^e \leq \varepsilon_{\max}\left(\frac{c_1 d}{m} + \frac{c_2}{\sqrt{n_U}} + \frac{c_3 d}{m n_U}\right) + \frac{c_4 m}{n_U} + \sum_{(i,j)\in E_\lambda} I(\lambda_i;\lambda_j|Y) + o(1/n_U),$$

where

$$\begin{aligned}c_1 &= \frac{2}{b_{\min}^2 a_{\min}^2}\left(1 + \frac{1}{(1-\bar{a}_{\max}^2)b_{\min}^2 a_{\min}^2}\right)\\ c_2 &= \frac{1}{(1-\bar{a}_{\max}^2)b_{\min}^2 a_{\min}^2}\sqrt{\frac{3(1-b_{\min}^2)}{b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)}\\ c_3 &= \frac{3(1-b_{\min}^2)}{(1-\bar{a}_{\max}^2)^2 b_{\min}^4 a_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)\\ c_4 &= \frac{3(1-b_{\min}^2)}{8b_{\min}^2(1-\bar{a}_{\max}^2)}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right),\end{aligned}$$

and $\varepsilon_{\max}$ is an upper bound on ε_{ij} defined in Lemma 5.

Define $\bar{a}_i = \mathbb{E}_\tau\left[\sqrt{\frac{\mathbb{E}[\lambda_i\lambda_j]\mathbb{E}[\lambda_i\lambda_k]}{\mathbb{E}[\lambda_j\lambda_k]}}\right]$ to be the asymptotic estimator with expectation over triplets. We apply Lemma 3 and simplify it to get

$$\begin{aligned}\sum_{i=1}^{m}\mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] &= \sum_{i=1}^{m}\left(\frac{1+a_i}{2}\log\left(1 + \frac{a_i - \bar{a}_i}{1+\bar{a}_i}\right) + \frac{1-a_i}{2}\log\left(1 + \frac{\bar{a}_i - a_i}{1-\bar{a}_i}\right)\right)\\ &+ \sum_{i=1}^{m}\frac{a_i - \bar{a}_i}{1-\bar{a}_i^2}\mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i - \widetilde{a}_i\right] + \sum_{i=1}^{m}\frac{1}{2}\Big(\frac{1}{1-\bar{a}_i^2} + \frac{2\bar{a}_i(\bar{a}_i - a_i)}{(1-\bar{a}_i^2)^2}\Big)\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i - \bar{a}_i)^2\right]\\ &+ o(1/n).\end{aligned} \tag{24}$$

This shows that there are three quantities to bound: $a_i - \bar{a}_i$, $\mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i - \widetilde{a}_i\right]$, and $\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i - \bar{a}_i)^2\right]$. Recall that for the unlabeled data case, $\widetilde{a}_i = \sqrt{\frac{\hat{\mathbb{E}}[\lambda_i\lambda_j]\hat{\mathbb{E}}[\lambda_i\lambda_k]}{\hat{\mathbb{E}}[\lambda_j\lambda_k]}}$ for random λ_j, λ_k, and $\bar{a}_i = \mathbb{E}_\tau\left[\sqrt{\frac{\mathbb{E}[\lambda_i\lambda_j]\mathbb{E}[\lambda_i\lambda_k]}{\mathbb{E}[\lambda_j\lambda_k]}}\right]$. The bounds for $\mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i - \widetilde{a}_i\right]$, and $\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i - \bar{a}_i)^2\right]$ are stated in Lemma 6; we focus on bounding the expected asymptotic gap $a_i - \bar{a}_i$ here.

Lemma 1. *For $i \in E_\lambda$, we have that*

$$\bar{a}_i - a_i \in \left[\frac{\varepsilon_{\min} b_{\min}}{m-1} - \frac{(d-1)\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2 a_{\min}^2},\ \frac{\varepsilon_{\max}}{(m-1)b_{\min}a_{\min}}\right] \tag{25}$$

For $i \notin E_\lambda$, we have that

$$\bar{a}_i - a_i \in \left[\frac{-d\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2 a_{\min}^2},\ \frac{-d\varepsilon_{\min} b_{\min}^2}{(m-1)(m-2)}\right] \tag{26}$$

And for all i, it is thus true that

$$|\bar{a}_i - a_i| \leq \frac{\varepsilon_{\max}}{(m-1)b_{\min}^2 a_{\min}^2}. \tag{27}$$

Proof. We define $\varepsilon_{ij} = \mathbb{E}[\lambda_i\lambda_j] - \mathbb{E}[\lambda_i Y]\,\mathbb{E}[\lambda_j Y]$ for $(i,j) \in E_\lambda$, i.e. the error we get from assuming conditional independence between λ_i and λ_j. We define the exact value of ε_{ij} in Lemma 5, and since all canonical parameters are assumed to be positive, we know that there exist $\varepsilon_{\min}, \varepsilon_{\max}$ that satisfy $0 < \varepsilon_{\min} \leq \varepsilon_{ij} \leq \varepsilon_{\max}$ over the entire edgeset E_λ. We now propagate this error to $\bar{a}_i$. Define $\bar{a}_i^{(j,k)}$ before we take the expectation over triplets as

$$\bar{a}_i^{(j,k)} := \sqrt{\frac{\mathbb{E}[\lambda_i\lambda_j]\,\mathbb{E}[\lambda_i\lambda_k]}{\mathbb{E}[\lambda_j\lambda_k]}}$$

Note that this means $\bar{a}_i \geq b_{\min}$. When each $\mathbb{E}[\lambda_i\lambda_j]$ can be written as $\mathbb{E}[\lambda_i Y]\,\mathbb{E}[\lambda_j Y]$, we get that $\bar{a}_i^{(j,k)} = a_i$. However, by our assumptions on the edgeset, at most one of the above pairwise expectations has nonzero ε_{ij}, in which case the true a_i is computed using $\mathbb{E}[\lambda_i\lambda_j] - \varepsilon_{ij}$, which is equal to $\mathbb{E}[\lambda_i Y]\,\mathbb{E}[\lambda_j Y]$, rather than $\mathbb{E}[\lambda_i\lambda_j]$.

If $(i,j) \in E_\lambda$ (but not (j,k) or (i,k)) then

$$a_i = \sqrt{\frac{(\mathbb{E}[\lambda_i\lambda_j] - \varepsilon_{ij})\mathbb{E}[\lambda_i\lambda_k]}{\mathbb{E}[\lambda_j\lambda_k]}}$$

This means that $\bar{a}_i \geq a_i$ and we asymptotically overestimate the accuracy. Then the difference between $\bar{a}_i^{(j,k)2}$ and a_i^2 is $\bar{a}_i^{(j,k)2} - a_i^2 = \frac{\varepsilon_{ij}\mathbb{E}[\lambda_i\lambda_k]}{\mathbb{E}[\lambda_j\lambda_k]} \in \left[\varepsilon_{\min} b_{\min}, \frac{\varepsilon_{\max}}{b_{\min}}\right]$. Moreover, $\bar{a}_i^{(j,k)} - a_i = \frac{\bar{a}_i^{(j,k)2} - a_i^2}{\bar{a}_i^{(j,k)} + a_i}$. Since $\bar{a}_i \geq a_i$ in this case, we have that $\bar{a}_i^{(j,k)} + a_i \in [2a_{\min}, 2]$; as a result,

$$\bar{a}_i^{(j,k)} - a_i \in \left[\frac{\varepsilon_{\min} b_{\min}}{2}, \frac{\varepsilon_{\max}}{2b_{\min}a_{\min}}\right] \tag{28}$$

Similarly, if $(i,k) \in E_\lambda$, we have the same bounds: $\bar{a}_i^{(j,k)2} - a_i^2 = \frac{\varepsilon_{ik}\mathbb{E}[\lambda_i\lambda_j]}{\mathbb{E}[\lambda_j\lambda_k]} \in \left[\varepsilon_{\min} b_{\min}, \frac{\varepsilon_{\max}}{b_{\min}}\right]$, and thus $\bar{a}_i^{(j,k)} - a_i \in \left[\frac{\varepsilon_{\min} b_{\min}}{2}, \frac{\varepsilon_{\max}}{2b_{\min}a_{\min}}\right]$. On the other hand, if $(j,k) \in E_\lambda$, the true accuracy is written as

$$a_i = \sqrt{\frac{\mathbb{E}[\lambda_i\lambda_j]\,\mathbb{E}[\lambda_i\lambda_k]}{(\mathbb{E}[\lambda_j\lambda_k] - \varepsilon_{jk})}}$$

This means that $\bar{a}_i^{(j,k)} \leq a_i$ and we asymptotically underestimate the accuracy. The difference between $\bar{a}_i^{(j,k)2}$ and a_i^2 is $a_i^2 - \bar{a}_i^{(j,k)2} = \frac{\varepsilon_{jk}\mathbb{E}[\lambda_i\lambda_j]\mathbb{E}[\lambda_i\lambda_k]}{\mathbb{E}[\lambda_j\lambda_k](\mathbb{E}[\lambda_j\lambda_k] - \varepsilon_{jk})} \in \left[\varepsilon_{\min} b_{\min}^2, \frac{\varepsilon_{\max}}{b_{\min}a_{\min}^2}\right]$. In this case, $a_i + \bar{a}_i^{(j,k)} \in [2b_{\min}, 2]$, so

$$a_i - \bar{a}_i^{(j,k)} \in \left[\frac{\varepsilon_{\min} b_{\min}^2}{2}, \frac{\varepsilon_{\max}}{2b_{\min}^2 a_{\min}^2}\right] \tag{29}$$

Lastly, if none of i, j, k share edges, $\bar{a}_i = a_i$. In our algorithm, we estimate each a_i using λ_j and λ_k chosen uniformly at random from the other $m-1$ sources. We thus need to compute the probabilities that $(i,j),(i,k)$ and (j,k) are in E_λ. Note that these probabilities depend on if $i \in E_\lambda$, which is true for $2d$ sources.

$$\Pr((i,j) \cup (i,k) \in E_\lambda \mid i \notin E_\lambda) = 0 \qquad \Pr((i,j) \cup (i,k) \in E_\lambda \mid i \in E_\lambda) = \frac{1(m-2)}{\binom{m-1}{2}} = \frac{2}{m-1}$$

$$\Pr((j,k) \in E_\lambda \mid i \notin E_\lambda) = \frac{2d}{(m-1)(m-2)} \quad \Pr((j,k) \in E_\lambda \mid i \in E_\lambda) = \frac{2(d-1)}{(m-1)(m-2)}$$

Therefore, if $i \in E_\lambda$, we use (28) and (29) to bound the expected error as

$$\bar{a}_i - a_i \leq \frac{2}{m-1} \cdot \frac{\varepsilon_{\max}}{2b_{\min}a_{\min}} + \frac{2(d-1)}{(m-1)(m-2)} \cdot \frac{-\varepsilon_{\min}b_{\min}^2}{2} \leq \frac{\varepsilon_{\max}}{(m-1)b_{\min}a_{\min}}$$

$$\bar{a}_i - a_i \geq \frac{2}{m-1} \cdot \frac{\varepsilon_{\min}b_{\min}}{2} + \frac{2(d-1)}{(m-1)(m-2)} \cdot \frac{-\varepsilon_{\max}}{2b_{\min}^2a_{\min}^2} = \frac{\varepsilon_{\min}b_{\min}}{m-1} - \frac{(d-1)\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2a_{\min}^2}$$

Note that this lower bound can be negative in this case, so it is not clear if $\bar{a}_i$ or a_i is bigger in expectation.

If $i \notin E_\lambda$, using (29) then the expected error is bounded as

$$\bar{a}_i - a_i \leq \frac{2d}{(m-1)(m-2)} \cdot \frac{-\varepsilon_{\min}b_{\min}^2}{2} = \frac{-d\varepsilon_{\min}b_{\min}^2}{(m-1)(m-2)}$$

$$\bar{a}_i - a_i \geq \frac{2d}{(m-1)(m-2)} \cdot \frac{-\varepsilon_{\max}}{2b_{\min}^2a_{\min}^2} = \frac{-d\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2a_{\min}^2}$$

In this case, $\bar{a}_i \leq a_i$. Finally, observe that regardless of if $i \in E_\lambda$ or not, the absolute value of the bias is bounded by

$$|\bar{a}_i - a_i| \leq \frac{\varepsilon_{\max}}{(m-1)b_{\min}^2a_{\min}^2}. \tag{30}$$

□

We return to (24). Since $a_i \geq \bar{a}_i$ when $i \notin E_\lambda$, we have that $\frac{1+a_i}{2}\log(1+\frac{a_i-\bar{a}_i}{1+\bar{a}_i}) + \frac{1-a_i}{2}\log(1+\frac{\bar{a}_i-a_i}{1-\bar{a}_i}) \leq \frac{1+a_i}{2}\log(1+\max\frac{a_i-\bar{a}_i}{1+\bar{a}_i})$ for $i \notin E_\lambda$. On the other hand when $i \in E_\lambda$, this expression can be upper bounded as $\frac{1+a_i}{2}\cdot\frac{a_i-\bar{a}_i}{1+\bar{a}_i} + \frac{1-a_i}{2}\frac{\bar{a}_i-a_i}{1-\bar{a}_i} = \frac{(\bar{a}_i-a_i)^2}{1-\bar{a}_i^2}$ using the inequality $\log(1+x) \leq x$ for $x > -1$ (it can be easily verified that $\frac{a_i-\bar{a}_i}{1+\bar{a}_i}$ and $\frac{\bar{a}_i-a_i}{1-\bar{a}_i}$ are at least -1). Since $|E_\lambda| = 2d$ and $\varepsilon_{\max} \leq 1$, the first summation of (24) is bounded by

$$(m-2d)\log\left(1 + \frac{d\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2a_{\min}^2(1+b_{\min})}\right) + 2d\frac{\varepsilon_{\max}^2}{(1-\bar{a}_{\max}^2)(m-1)^2b_{\min}^4a_{\min}^4} \tag{31}$$

$$\leq \frac{(m-2d)d\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2a_{\min}^2(1+b_{\min})} + \frac{2d\varepsilon_{\max}}{(1-\bar{a}_{\max}^2)(m-1)^2b_{\min}^4a_{\min}^4} \tag{32}$$

$$= \frac{d\varepsilon_{\max}}{(m-1)b_{\min}^2a_{\min}^2}\left(\frac{m-2d}{(m-2)(1+b_{\min})} + \frac{2}{(1-\bar{a}_{\max}^2)(m-1)b_{\min}^2a_{\min}^2}\right) \tag{33}$$

$$\leq \frac{d\varepsilon_{\max}}{(m-1)b_{\min}^2a_{\min}^2}\left(1 + \frac{1}{(1-\bar{a}_{\max}^2)b_{\min}^2a_{\min}^2}\right) \leq \frac{c_1d\varepsilon_{\max}}{m}, \tag{34}$$

where $c_1 = \frac{2}{b_{\min}^2a_{\min}^2}\left(1 + \frac{1}{(1-\bar{a}_{\max}^2)b_{\min}^2a_{\min}^2}\right)$. Next, we bound $\sum_{i=1}^m \frac{a_i-\bar{a}_i}{1-\bar{a}_i^2}\mathbb{E}_{\mathcal{N},\tau}[\bar{a}_i - \widetilde{a}_i]$.

$$\sum_{i=1}^m \frac{a_i-\bar{a}_i}{1-\bar{a}_i^2}\mathbb{E}_{\mathcal{N},\tau}[\bar{a}_i - \widetilde{a}_i] \leq \sum_{i=1}^m \frac{|\bar{a}_i-a_i|}{1-\bar{a}_i^2}\mathbb{E}_{\mathcal{N},\tau}[|\bar{a}_i - \widetilde{a}_i|] \tag{35}$$

$$\leq \frac{\sqrt{3}}{2\sqrt{n_U}} \cdot \sqrt{\frac{1-b_{\min}^2}{b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)}\frac{1}{1-\bar{a}_{\max}^2}\left(\frac{m\varepsilon_{\max}}{(m-1)b_{\min}^2a_{\min}^2}\right) \leq \frac{c_2\varepsilon_{\max}}{\sqrt{n_U}}, \tag{36}$$

where $c_2 = \frac{1}{(1-\bar{a}_{\max}^2)b_{\min}^2 a_{\min}^2}\sqrt{\frac{3(1-b_{\min}^2)}{b_{\min}^2}\left(\frac{1}{b_{\min}^4}+\frac{2}{b_{\min}^2}\right)}$. We bound $\sum_{i=1}^m \frac{1}{2}\Big(\frac{1}{1-\bar{a}_i^2}+\frac{2\bar{a}_i(\bar{a}_i-a_i)}{(1-\bar{a}_i^2)^2}\Big)\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right]$, which can be split into an expression independent of misspecification and one dependent on it:

$$\sum_{i=1}^m \frac{1}{2}\Big(\frac{1}{1-\bar{a}_i^2}+\frac{2\bar{a}_i(\bar{a}_i-a_i)}{(1-\bar{a}_i^2)^2}\Big)\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right] \le \frac{c_4 m}{n_U}+\sum_{i=1}^m \frac{\bar{a}_i-a_i}{(1-\bar{a}_i^2)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right], \tag{37}$$

where $c_4 = \frac{3(1-b_{\min}^2)}{8b_{\min}^2(1-\bar{a}_{\max}^2)}\left(\frac{1}{b_{\min}^4}+\frac{2}{b_{\min}^2}\right)$. The summation in (37) is bounded as follows, using the fact that $\bar{a}_i \le a_i$ for $i \notin E_\lambda$:

$$\sum_{i=1}^m \frac{\bar{a}_i-a_i}{(1-\bar{a}_i^2)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right] \le \frac{3}{4n_U}\cdot\frac{1-b_{\min}^2}{b_{\min}^2(1-\bar{a}_{\max}^2)^2}\left(\frac{1}{b_{\min}^4}+\frac{2}{b_{\min}^2}\right)\sum_{i\in E_\lambda}|\bar{a}_i-a_i| \tag{38}$$

$$\le \frac{3}{4n_U}\cdot\frac{1-b_{\min}^2}{b_{\min}^2(1-\bar{a}_{\max}^2)^2}\left(\frac{1}{b_{\min}^4}+\frac{2}{b_{\min}^2}\right)\left(\frac{2d\varepsilon_{\max}}{(m-1)b_{\min}^2 a_{\min}^2}\right) \le \frac{c_3 d\varepsilon_{\max}}{mn_U}, \tag{39}$$

where $c_3 = \frac{3(1-b_{\min}^2)}{(1-\bar{a}_{\max}^2)^2 b_{\min}^4 a_{\min}^2}\left(\frac{1}{b_{\min}^4}+\frac{2}{b_{\min}^2}\right)$. This concludes our proof.

C.4 Proof of Proposition 1

To prove the ability of using the median of the accuracies to correct for misspecification, we first examine the asymptotic case. For $i \in E_\lambda$, note that out of a total of $\binom{m-1}{2}$ triplets, $m-2$ of them will involve the edge $(i,j) \in E_\lambda$, resulting in a higher inconsistent estimate of the accuracy. $d-1$ of them will involve an edge $(j,k) \in E_\lambda$, resulting in a lower estimate of the accuracy. Therefore, $\frac{(m-1)(m-2)}{2} - m - d - 3$ triplets are consistent. As long as the $\binom{m-1}{2} - (m-2)$th largest triplet is greater than half of all the triplets, and the $d-1$th largest triplet is less than the half of all the triplets, then the median will be a consistent triplet. This gives us the conditions $m > 5$ and $d < \frac{(m-1)(m-2)}{4}$.

Next, for $i \notin E_\lambda$, d triplets will involve an edge $(j,k) \in E_\lambda$, resulting in lower estimated accuracy, while the other $\binom{m-1}{2} - d$ triplets are consistent. Therefore, as long as $d < \frac{(m-1)(m-2)}{4}$, the median triplet is consistent.

Lastly, we must consider the finite sample regime when the ordering of the accuracy estimates are perturbed by sampling noise. When each accuracy's expected sampling noise is less than half of the minimum standing bias of a triplet, the order of the accuracies will not change on average. This translates into the inequality $\mathbb{E}\left[|\widetilde{a}_i-\bar{a}_i|\right] \le \frac{1}{2}\min_{(j,k)}|a_i-\bar{a}_i^{(j,k)}|$. The minimum standing bias is $\frac{\varepsilon_{\min}b_{\min}^2}{2}$, and $\mathbb{E}\left[|\widetilde{a}_i-\bar{a}_i|\right] \sim \mathcal{O}(1/\sqrt{n})$ so this means that $n_U \ge n_0 \sim \Omega(1/\varepsilon_{\min}^2)$.

D Auxiliary Lemmas

Lemma 2. *For any source λ_i with accuracy $a_i = \mathbb{E}\left[\lambda_i Y\right]$,*

$$\Pr(\lambda_i = 1|Y=1) = \Pr(\lambda_i = -1|Y=-1) = \frac{1+a_i}{2}$$
$$\Pr(\lambda_i = -1|Y=1) = \Pr(\lambda_i = 1|Y=-1) = \frac{1-a_i}{2}.$$

Proof. By Proposition 2 of Fu et al. (2020), we know that $\lambda_i Y \perp\!\!\!\perp Y$ for the binary Ising model we use, defined in section 3. Intuitively, this means that the accuracy of a source is independent of the value of Y, and therefore $\Pr(\lambda_i Y = 1|Y=1) = \Pr(\lambda_i Y = 1) = \frac{1+a_i}{2}$, since $\mathbb{E}\left[\lambda_i Y\right] = 2\Pr(\lambda_i Y = 1) - 1$. Repeating this calculation with remaining configurations of $\Pr(\lambda_i Y = \pm 1|Y=\pm 1)$ concludes our proof. □

Lemma 3. *Define $a_i = \mathbb{E}\left[\lambda_i Y\right]$, and let $\widetilde{a}_i$ be our estimated accuracy on n points. Furthermore, let $\bar{a}_i$ be the*

expected asymptotic value of $\widetilde{a}_i$ over τ. Then, the estimation error is

$$\begin{aligned}\mathbb{E}_{Y,\mathcal{N},\tau}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] =& \frac{1+a_i}{2}\Big(\log\Big(1+\frac{a_i-\bar{a}_i}{1+\bar{a}_i}\Big)+\frac{\mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i-\widetilde{a}_i\right]}{1+\bar{a}_i}+\frac{1}{2(1+\bar{a}_i)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right]\Big)\\ &+\frac{1-a_i}{2}\Big(\log\Big(1+\frac{\bar{a}_i-a_i}{1-\bar{a}_i}\Big)+\frac{\mathbb{E}_{\mathcal{N},\tau}\left[\widetilde{a}_i-\bar{a}_i\right]}{1-\bar{a}_i}+\frac{1}{2(1-\bar{a}_i)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right]\Big)\\ &+o(1/n).\end{aligned}$$

Proof. As discussed previously, this term is equal to $-\mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[\log\frac{\widetilde{\mathrm{Pr}}(\lambda_i'=\lambda_i|Y'=Y)}{\mathrm{Pr}(\lambda_i'=\lambda_i|Y'=Y)}\right]$. By the law of total expectation, we now have

$$-\mathbb{E}_{\boldsymbol{\lambda},\mathcal{N},\tau}\left[\mathrm{Pr}(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\log\frac{\widetilde{\mathrm{Pr}}(\lambda_i'=\lambda_i|Y=1)}{\mathrm{Pr}(\lambda_i'=\lambda_i|Y=1)}+\mathrm{Pr}(Y=-1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\log\frac{\widetilde{\mathrm{Pr}}(\lambda_i'=\lambda_i|Y=-1)}{\mathrm{Pr}(\lambda_i'=\lambda_i|Y=-1)}\right]. \tag{40}$$

Suppose $\lambda_i\notin E_\lambda$. Conditioning on the value of λ_i and using Lemma 2, (40) becomes

$$\begin{aligned}&-\mathbb{E}_{\boldsymbol{\lambda}_{-i},\mathcal{N},\tau}\left[\mathbb{E}_{\lambda_i}\left[\mathrm{Pr}(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\log\frac{\widetilde{\mathrm{Pr}}(\lambda_i'=\lambda_i|Y=1)}{\mathrm{Pr}(\lambda_i'=\lambda_i|Y=1)}+\mathrm{Pr}(Y=-1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\log\frac{\widetilde{\mathrm{Pr}}(\lambda_i'=\lambda_i|Y=-1)}{\mathrm{Pr}(\lambda_i'=\lambda_i|Y=-1)}\Big|\boldsymbol{\lambda}_{-i}\right]\right]\\ =&-\mathbb{E}_{\boldsymbol{\lambda}_{-i},\mathcal{N},\tau}\Big[\big(\mathrm{Pr}(Y=1|\boldsymbol{\lambda}_{-i},\lambda_i=1)\,\mathrm{Pr}(\lambda_i=1|\boldsymbol{\lambda}_{-i})+\mathrm{Pr}(Y=-1|\boldsymbol{\lambda}_{-i},\lambda_i=-1)\,\mathrm{Pr}(\lambda_i=-1|\boldsymbol{\lambda}_{-i})\big)\log\frac{1+\widetilde{a}_i}{1+a_i}\\ &+\big(\mathrm{Pr}(Y=1|\boldsymbol{\lambda}_{-i},\lambda_i=-1)\,\mathrm{Pr}(\lambda_i=-1|\boldsymbol{\lambda}_{-i})+\mathrm{Pr}(Y=-1|\boldsymbol{\lambda}_{-i},\lambda_i=1)\,\mathrm{Pr}(\lambda_i=1|\boldsymbol{\lambda}_{-i})\big)\log\frac{1-\widetilde{a}_i}{1-a_i}\Big]\\ =&-\mathbb{E}_{\boldsymbol{\lambda}_{-i},\mathcal{N},\tau}\Big[\mathrm{Pr}(\lambda_iY=1|\boldsymbol{\lambda}_{-i})\log\frac{1+\widetilde{a}_i}{1-a_i}+\mathrm{Pr}(\lambda_iY=-1|\boldsymbol{\lambda}_{-i})\log\frac{1-\widetilde{a}_i}{1-a_i}\Big].\end{aligned}$$

Note that $\mathrm{Pr}(\lambda_i=1,Y=1|\boldsymbol{\lambda}_{-i})=\mathrm{Pr}(\lambda_i=1|Y=1)\frac{\mathrm{Pr}(\boldsymbol{\lambda}_{-i},Y=1)}{\mathrm{Pr}(\boldsymbol{\lambda}_{-i})}$ and $\mathrm{Pr}(\lambda_i=-1,Y=-1|\boldsymbol{\lambda}_{-i})=\mathrm{Pr}(\lambda_i=-1|Y=-1)\frac{\mathrm{Pr}(\boldsymbol{\lambda}_{-i},Y=-1)}{\mathrm{Pr}(\boldsymbol{\lambda}_{-i})}$ since λ_i and λ_{-i} are conditionally independent given Y, so $\mathrm{Pr}(\lambda_iY=1|\boldsymbol{\lambda}_{-i})=\mathrm{Pr}(\lambda_i=1|Y=1)=\frac{1+a_i}{2}$. Similarly, $\mathrm{Pr}(\lambda_iY=-1|\boldsymbol{\lambda}-i)=\mathrm{Pr}(\lambda_i=-1|Y=1)=\frac{1-a_i}{2}$, so the conditional KL divergence is equal to

$$\mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right]=-\mathbb{E}_{\mathcal{N},\tau}\left[\frac{1+a_i}{2}\log\frac{1+\widetilde{a}_i}{1+a_i}+\frac{1-a_i}{2}\log\frac{1-\widetilde{a}_i}{1-a_i}\right]. \tag{41}$$

Now suppose that $\lambda_i\in E_\lambda$ and has an edge to some λ_j. When we simplify (40) by conditioning on λ_i,λ_j, we find that $\sum_{l\in\{\pm1\}}\mathrm{Pr}(Y=1|\boldsymbol{\lambda}_{-i,j},\lambda_i=1,\lambda_j=l)\,\mathrm{Pr}(\lambda_i=1,\lambda_j=l|\boldsymbol{\lambda}_{-i,j})+\mathrm{Pr}(Y=-1|\boldsymbol{\lambda}_{-i,j},\lambda_i=-1,\lambda_j=l)\,\mathrm{Pr}(\lambda_i=-1,\lambda_j=l|\boldsymbol{\lambda}_{-i,j})$ (i.e, the coefficient for $\log\frac{1+\widetilde{a}_i}{1+a_i}$) is equal to $\mathrm{Pr}(\lambda_iY=1|\boldsymbol{\lambda}_{-i,j})$, and this is still equal to $\frac{1+a_i}{2}$. The same holds for the coefficient of $\log\frac{1-\widetilde{a}_i}{1-a_i}$. Therefore, (41) holds for all λ_i.

Next, we evaluate $-\mathbb{E}\left[\log\frac{1+\widetilde{a}_i}{1+a_i}\right]$ and $-\mathbb{E}\left[\log\frac{1-\widetilde{a}_i}{1-a_i}\right]$, where expectation is over $\mathcal{N}$ and τ. We apply a second-order Taylor approximation of $f(x)=\log\frac{1+x}{1+a_i}$ at $x=\bar{a}_i$:

$$\log\frac{1+\widetilde{a}_i}{1+a_i}\approx\log\frac{1+\bar{a}_i}{1+a_i}+\frac{1+a_i}{1+\bar{a}_i}\cdot\frac{1}{1+a_i}(\widetilde{a}_i-\bar{a}_i)-\frac{1}{2(1+\bar{a}_i)^2}(\widetilde{a}_i-\bar{a}_i)^2+o(1/n).$$

Taking the expectation on both sides, we get

$$\begin{aligned}-\mathbb{E}_{\mathcal{N},\tau}\left[\log\frac{1+\widetilde{a}_i}{1+a_i}\right]&\approx-\Big(\log\frac{1+\bar{a}_i}{1+a_i}+\frac{\mathbb{E}_{\mathcal{N},\tau}\left[\widetilde{a}_i\right]-\bar{a}_i}{1+\bar{a}_i}-\frac{1}{2(1+\bar{a}_i)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right]\Big)+o(1/n)\\ &=\log\Big(1+\frac{a_i-\bar{a}_i}{1+\bar{a}_i}\Big)+\frac{\mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i-\widetilde{a}_i\right]}{1+\bar{a}_i}+\frac{1}{2(1+\bar{a}_i)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right]+o(1/n),\end{aligned}$$

where we have used Lemma 4.

Similarly, we apply a second-order Taylor approximation of $f(x) = \log \frac{1-x}{1-a_i}$ at $x = \bar{a}_i$:

$$\log \frac{1-\widetilde{a}_i}{1-a_i} \approx \log \frac{1-\bar{a}_i}{1-a_i} + \frac{1-a_i}{1-\bar{a}_i} \cdot \frac{-1}{1-a_i}(\widetilde{a}_i - \bar{a}_i) - \frac{1}{2(1-\bar{a}_i)^2}(\widetilde{a}_i - \bar{a}_i)^2 + o(1/n).$$

Taking the expectation of both sides,

$$\begin{aligned} -\mathbb{E}\left[\log \frac{1-\widetilde{a}_i}{1-a_i}\right] &= -\Big(\log \frac{1-\bar{a}_i}{1-a_i} + \frac{\mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i - \widetilde{a}_i\right]}{1-\bar{a}_i} - \frac{1}{2(1-\bar{a}_i)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i - \bar{a}_i)^2\right]\Big) + o(1/n) \\ &= \log\Big(1 + \frac{\bar{a}_i - a_i}{1-\bar{a}_i}\Big) + \frac{\mathbb{E}_{\mathcal{N},\tau}\left[\widetilde{a}_i - \bar{a}_i\right]}{1-\bar{a}_i} + \frac{1}{2(1-\bar{a}_i)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i - \bar{a}_i)^2\right] + o(1/n). \end{aligned}$$

Substituting these expressions into (41), we get our desired equation.

□

Lemma 4. *The remainder of the Taylor approximation done in Lemma 3 is $o(1/n)$ for estimation done on n samples in both the labeled and unlabeled cases.*

Proof. The remainder for $-\mathbb{E}_{\mathcal{N},\tau}\left[\log \frac{1+\widetilde{a}_i}{1+a_i}\right]$ is bounded by $\frac{1}{3(1+\bar{a}_i)^3}\mathbb{E}_{\mathcal{N},\tau}\left[(\bar{a}_i - \widetilde{a}_i)^3\right]$, and the remainder for $-\mathbb{E}_{\mathcal{N},\tau}\left[\log \frac{1-\widetilde{a}_i}{1-a_i}\right]$ is bounded by $\frac{1}{(1-\bar{a}_i)^3}\mathbb{E}_{\mathcal{N},\tau}\left[(\bar{a}_i - \widetilde{a}_i)^3\right]$.

For the labeled data case, it is easy to check that $\mathbb{E}_{\mathcal{N}}\left[(\bar{a}_i - a_i)^3\right] \sim \mathcal{O}(1/n_L^2)$. Therefore, we focus on analyzing the unlabeled data case's estimator by bounding $\mathbb{E}_{\mathcal{N}}\left[|\bar{a}_i - \widetilde{a}_i|^3 \mid \lambda_j, \lambda_k\right]$ independent of choice of j and k. For ease of notation, define $X = \lambda_i\lambda_j$ and $Y = \lambda_i\lambda_k$, such that $XY = \lambda_j\lambda_k$, and let

$$a := \bar{a}_i^{(j,k)} = \sqrt{\frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}}, \qquad \hat{a} := \widetilde{a}_i = \sqrt{\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]}}. \tag{42}$$

Note $a \in [-1, 1]$, so clip $\hat{a} \in [-1, 1]$. Because $X \in \{-1, 1\}$ and $\hat{\mathbb{E}}[X]$ is an i.i.d. sum of $n = n_U$ samples from X, we can apply Hoeffding's inequality to get:

$$\Pr\left(|\hat{\mathbb{E}}[X] - \mathbb{E}[X]| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{n2^2}\right) = 2\exp\left(-\frac{n\epsilon^2}{2}\right) \tag{43}$$

The same is true for $\hat{\mathbb{E}}[Y]$ and $\hat{\mathbb{E}}[XY]$. Thus, by union bound,

$$\Pr\left(|\hat{\mathbb{E}}[X] - \mathbb{E}[X]| \geq \epsilon \vee |\hat{\mathbb{E}}[Y] - \mathbb{E}[Y]| \geq \epsilon \vee |\hat{\mathbb{E}}[XY] - \mathbb{E}[XY]| \geq \epsilon\right) \leq 6\exp\left(-\frac{n\epsilon^2}{2}\right) \tag{44}$$

Refer to the event $\left(|\hat{\mathbb{E}}[X] - \mathbb{E}[X]| \geq \epsilon \vee |\hat{\mathbb{E}}[Y] - \mathbb{E}[Y]| \geq \epsilon \vee |\hat{\mathbb{E}}[XY] - \mathbb{E}[XY]| \geq \epsilon\right)$ as B. If $\neg B$ and $\epsilon < \frac{1}{2}\min(\mathbb{E}[X], \mathbb{E}[Y], \mathbb{E}[XY]) < 1$, then

$$|\hat{\mathbb{E}}[X] - \mathbb{E}[X]| < \epsilon, \quad |\hat{\mathbb{E}}[Y] - \mathbb{E}[Y]| < \epsilon, \quad |\hat{\mathbb{E}}[XY] - \mathbb{E}[XY]| < \epsilon \tag{45}$$

By the mean value theorem with $f(x) = \sqrt{x}$, there exists a u between $\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]}$ and $\frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}$ such that

$$|\hat{a} - a| = \left|\frac{1}{2\sqrt{u}}\left(\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right)\right| \tag{46}$$

Note that

$$u \geq \min\left(\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]}, \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right) \geq \min\left(\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)}{\mathbb{E}[XY]+\epsilon}, \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right) \tag{47}$$

$$\geq \min\left(\frac{(\mathbb{E}[X]/2)(\mathbb{E}[Y]/2)}{1+\epsilon}, \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right) \geq \min\left(\frac{\mathbb{E}[X]\mathbb{E}[Y]}{8}, \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right) \geq \frac{\mathbb{E}[X]\mathbb{E}[Y]}{8} \tag{48}$$

Thus,

$$|\hat{a}-a| \leq \frac{\sqrt{2}}{\sqrt{\mathbb{E}[X]\mathbb{E}[Y]}}\left|\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right|. \tag{49}$$

For the term on the right inside the absolute value:

$$\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)}{\mathbb{E}[XY]+\epsilon} \leq \frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]} \leq \frac{(\mathbb{E}[X]+\epsilon)(\mathbb{E}[Y]+\epsilon)}{\mathbb{E}[XY]-\epsilon} \tag{50}$$

$$\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)}{\mathbb{E}[XY]+\epsilon} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]} \leq \frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]} \leq \frac{(\mathbb{E}[X]+\epsilon)(\mathbb{E}[Y]+\epsilon)}{\mathbb{E}[XY]-\epsilon} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]} \tag{51}$$

$$\left|\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right| \leq \max\left(\left|\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)}{\mathbb{E}[XY]+\epsilon} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right|, \tag{52}$$

$$\left|\frac{(\mathbb{E}[X]+\epsilon)(\mathbb{E}[Y]+\epsilon)}{\mathbb{E}[XY]-\epsilon} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right|\right). \tag{53}$$

Examining the left term in the max,

$$\left|\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)}{\mathbb{E}[XY]+\epsilon} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right| = \left|\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y](\mathbb{E}[XY]+\epsilon)}{\mathbb{E}[XY](\mathbb{E}[XY]+\epsilon)}\right| \tag{54}$$

$$= \left|\frac{-\epsilon(\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[XY] + \mathbb{E}[Y]\mathbb{E}[XY] - \epsilon\mathbb{E}[XY])}{\mathbb{E}[XY](\mathbb{E}[XY]+\epsilon)}\right| \tag{55}$$

$$\leq \epsilon\left|\frac{\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[XY] + \mathbb{E}[Y]\mathbb{E}[XY]}{\mathbb{E}[XY]^2}\right| \tag{56}$$

$$= \epsilon C_1 > 0 \tag{57}$$

Examining the right term in the max,

$$\left|\frac{(\mathbb{E}[X]+\epsilon)(\mathbb{E}[Y]+\epsilon)}{\mathbb{E}[XY]-\epsilon} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right| = \left|\frac{(\mathbb{E}[X]+\epsilon)(\mathbb{E}[Y]+\epsilon)\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y](\mathbb{E}[XY]-\epsilon)}{\mathbb{E}[XY](\mathbb{E}[XY]-\epsilon)}\right| \tag{58}$$

$$= \left|\frac{\epsilon(\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[XY] + \mathbb{E}[Y]\mathbb{E}[XY] + \epsilon\mathbb{E}[XY])}{\mathbb{E}[XY](\mathbb{E}[XY]-\epsilon)}\right| \tag{59}$$

$$\leq \epsilon\left|\frac{\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[XY] + \mathbb{E}[Y]\mathbb{E}[XY] + \mathbb{E}[XY]}{\mathbb{E}[XY]^2/2}\right| \tag{60}$$

$$= \epsilon C_2 > 0 \tag{61}$$

Combining the max argument bounds, we have that $\left|\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right| \leq \epsilon\max(C_1, C_2) \leq \epsilon C_2$. Therefore,

$$|\hat{a}-a| \leq \epsilon\frac{\sqrt{2}C_2}{\sqrt{\mathbb{E}[X]\mathbb{E}[X]}} = \epsilon C_3 \tag{62}$$

where C_3 is a positive function of $\mathbb{E}[X]$, $\mathbb{E}[Y]$, and $\mathbb{E}[XY]$. To recap, this is satisfied if $\neg B$ and ϵ is small. Let $\epsilon = n^{-3/8}$, thus for large enough n, ϵ is smaller than any constant. Recall, $\Pr(B) \leq 6\exp(-n\epsilon^2/2)$. With this definition of ϵ, $\Pr(B) \leq 6\exp(-n^{1/4}/2)$.

Now, we are finally ready to evaluate the limit:

$$\lim_{n\to\infty} n\mathbb{E}[|\hat{a}-a|^3] = \lim_{n\to\infty} n\left(\mathbb{E}[|\hat{a}-a|^3|B]\Pr(B) + \mathbb{E}[|\hat{a}-a|^3|\neg B]P(\neg B)\right) \tag{63}$$

$$\leq \lim_{n\to\infty} n\left(C_3^3\epsilon^3\cdot 1 + 2^3\cdot 6\exp(-n^{1/4}/2)\right) \tag{64}$$

$$= C_3^3 \lim_{n\to\infty} n(n^{-3/8})^3 + 48\lim_{n\to\infty} n\exp(-n^{1/4}/2) \tag{65}$$

$$= C_3^3 \lim_{n\to\infty} n^{-1/8} + 48\lim_{m\to\infty} m^4\exp(-m/2) = 0 \tag{66}$$

Trivially, $\lim_{n\to\infty} n\mathbb{E}[|\hat{a}-a|^3] \geq 0$. Thus, $\lim_{n\to\infty} n\mathbb{E}[|\hat{a}-a|^3] = 0$. □

Lemma 5. *If $(i,j)\in E_\lambda$, then*

$$\varepsilon_{ij} = \Delta_{ij} - \Delta_i a'_j - \Delta_j a'_i - \Delta_i\Delta_j, \tag{67}$$

where

$$\Delta_i = \frac{2}{z_{ij}z'_{ij}}(\exp(\theta_{ij}) - \exp(-\theta_{ij}))(\exp(2\theta_j) - \exp(-2\theta_j)) \tag{68}$$

$$\Delta_j = \frac{2}{z_{ij}z'_{ij}}(\exp(\theta_{ij}) - \exp(-\theta_{ij}))(\exp(2\theta_i) - \exp(-2\theta_i)) \tag{69}$$

$$\Delta_{ij} = \frac{2}{z_{ij}z'_{ij}}(\exp(\theta_{ij}) - \exp(-\theta_{ij}))(\exp(2\theta_i) + \exp(-2\theta_i) + \exp(2\theta_j) + \exp(-2\theta_j)) \tag{70}$$

$$a'_i = \frac{2}{z'_{ij}}\exp(\theta_i)(\exp(\theta_j) + \exp(-\theta_j)) - 1 \tag{71}$$

$$a'_j = \frac{2}{z'_{ij}}\exp(\theta_j)(\exp(\theta_i) + \exp(-\theta_i)) - 1 \tag{72}$$

$$z_{ij} = \sum_{s_i,s_j}\exp(s_i\theta_i + s_j\theta_j + s_is_j\theta_{ij}) \tag{73}$$

$$z'_{ij} = \sum_{s_i,s_j}\exp(s_i\theta_i + s_j\theta_j) \tag{74}$$

Using these values, it is also possible to verify that $\varepsilon_{ij}\in(0,1)$ if $\theta_i,\theta_j,\theta_{ij}>0$.

Proof. We define a new distribution, which we denote by $\Pr'$ and $\mathbb{E}'$:

$$\Pr'(Y,\boldsymbol{\lambda}) = \frac{1}{Z'}\exp\left(\theta_Y + \sum_{i=1}^m \theta_i\lambda_i Y + \sum_{(k,l)\neq(i,j)}\theta_{kl}\lambda_k\lambda_l\right). \tag{75}$$

This distribution uses all the same canonical parameters as (**??**) except $\theta_{ij}\lambda_i\lambda_j$. We know that for this distribution, $\mathbb{E}'[\lambda_i\lambda_j] = \mathbb{E}'[\lambda_i Y]\,\mathbb{E}'[\lambda_j Y]$. Our approach to compute $\varepsilon_{ij} = \mathbb{E}[\lambda_i\lambda_j] - \mathbb{E}[\lambda_i Y]\,\mathbb{E}[\lambda_j Y]$ is to bound the differences between $\mathbb{E}$ and $\mathbb{E}'$.

First, we evaluate $\mathbb{E}[\lambda_i Y] - \mathbb{E}[\lambda_j Y]$. We write $\mathbb{E}[\lambda_i Y]$ as $2\Pr(\lambda_i Y = 1) - 1 = \frac{2}{p}\Pr(\lambda_i = 1, Y = 1) - 1$ and $\mathbb{E}'[\lambda_i Y]$ as $\frac{2}{p}\Pr'(\lambda_i = 1, Y = 1) - 1$ by Lemma 2, where $p = \Pr(Y = 1)$. Then, letting s_{-i} represent all combinations of labels on all $\boldsymbol{\lambda}$ besides λ_i,

$$\Delta_i = \mathbb{E}[\lambda_i Y] - \mathbb{E}'[\lambda_i Y] = \frac{2}{p}\sum_{s_{-i}}\exp\left(\theta_Y + \theta_i + \sum_{k\neq i}\theta_k l_k + \sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\right)\left(\frac{\exp(\theta_{ij}l_j)}{Z} - \frac{1}{Z'}\right) \tag{76}$$

Next, note that $p = \frac{z_Y}{Z}$ and $p = \frac{z'_Y}{Z'}$, where $z_Y = \sum_s \exp(\theta_Y + \sum_{k=1}^m \theta_k s_k + \sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l + \theta_{ij}s_i s_j)$ and $z'_Y = \sum_s \exp(\theta_Y + \sum_{k=1}^m \theta_i s_i + \sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l)$ (we can check these expressions for p are equal, since the

edgewise potentials are canceled out). Δ_i is now

$$2\exp(\theta_i)\sum_{s_{-i}}\exp\left(\theta_Y+\sum_{k\neq i}\theta_k l_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\right)\left(\frac{\exp(\theta_{ij}l_j)}{z_Y}-\frac{1}{z'_Y}\right) \tag{77}$$

$$=2\exp(\theta_i+\theta_j)\sum_{s_{-i,j}}\exp\left(\theta_Y+\sum_{k\neq i,j}\theta_k l_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\right)\left(\frac{\exp(\theta_{ij})}{z_Y}-\frac{1}{z'_Y}\right) \tag{78}$$

$$+2\exp(\theta_i-\theta_j)\sum_{s_{-i,j}}\exp\left(\theta_Y+\sum_{k\neq i,j}\theta_k l_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\right)\left(\frac{\exp(-\theta_{ij})}{z_Y}-\frac{1}{z'_Y}\right) \tag{79}$$

$\frac{\exp(\pm\theta_{ij})}{z_Y}-\frac{1}{z'_Y}$ can be written as $\frac{1}{z_Y z'_Y}\sum_{s'}\exp(\theta_Y+\sum_k\theta_k s'_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s'_k s'_l)(\exp(\pm\theta_{ij})-\exp(\theta_{ij}s'_i s'_j))$. Then for positive θ_{ij}, this becomes $\frac{1}{z_Y z'_Y}(\exp(\theta_i-\theta_j)+\exp(-\theta_i+\theta_j))\sum_{s'}\exp(\theta_Y+\sum_{k\neq i,j}\theta_k s'_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s'_k s'_l)(\exp(\theta_{ij})-\exp(-\theta_{ij}))$, and for negative $-\theta_{ij}$, this becomes $\frac{1}{z_Y z'_Y}(\exp(\theta_i+\theta_j)+\exp(-\theta_i-\theta_j))\sum_{s'}\exp(\theta_Y+\sum_{k\neq i,j}\theta_k s'_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s'_k s'_l)(\exp(-\theta_{ij})-\exp(\theta_{ij}))$. Then, our expression becomes

$$\frac{2}{z_Y z'_Y}\Bigg(\sum_{s_{-i,j}}\exp\Big(\theta_Y+\sum_{k\neq i,j}\theta_k l_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\Big)\Bigg)^2(\exp(\theta_{ij})-\exp(-\theta_{ij})) \tag{80}$$

$$\times\big(\exp(\theta_i+\theta_j)(\exp(\theta_i-\theta_j)+\exp(-\theta_i+\theta_j))-\exp(\theta_i-\theta_j)(\exp(\theta_i+\theta_j)+\exp(-\theta_i-\theta_j))\big) \tag{81}$$

The second line simplifies $\exp(2\theta_i)+\exp(2\theta_j)-\exp(2\theta_i)-\exp(-2\theta_j)=\exp(2\theta_j)-\exp(-2\theta_j)$. Lastly, note that $z_Y=\sum_{s_{-i,j}}\exp\Big(\theta_Y+\sum_{k\neq i,j}\theta_k l_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\Big)\cdot\sum_{s_i,s_j}\exp(s_i\theta_i+s_j\theta_j+s_i s_j\theta_{ij})$, and $z'_Y=\sum_{s_{-i,j}}\exp\Big(\theta_Y+\sum_{k\neq i,j}\theta_k l_k+\sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\Big)\cdot\sum_{s_i,s_j}\exp(s_i\theta_i+s_j\theta_j)$. Canceling out the summations over the other sources, we have our desired expression for Δ_i. We can do the same to get our result for Δ_j.

Next, we compute $\Delta_{ij}=\mathbb{E}\left[\lambda_i\lambda_j\right]-\mathbb{E}'\left[\lambda_i\lambda_j\right]$, which is equal to $2(\Pr(\lambda_i=1,\lambda_j=1)-\Pr'(\lambda_i=1,\lambda_j=1)+\Pr(\lambda_i=-1,\lambda_j=-1)-\Pr'(\lambda_i=-1,\lambda_j=-1))$:

$$\Pr(\lambda_i=1,\lambda_j=1)-\Pr'(\lambda_i=1,\lambda_j=1) \tag{82}$$

$$=\sum_{Y,s_{-i,j}}\exp\Big(\theta_Y Y+\theta_i Y+\theta_j Y+\sum_{k\neq i,j}\theta_k s_k Y+\sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\Big)\left(\frac{\exp(\theta_{ij})}{Z}-\frac{1}{Z'}\right) \tag{83}$$

$$\Pr(\lambda_i=-1,\lambda_j=-1)-\Pr'(\lambda_i=-1,\lambda_j=-1) \tag{84}$$

$$=\sum_{Y,s_{-i,j}}\exp\Big(\theta_Y Y-\theta_i Y-\theta_j Y+\sum_{k\neq i,j}\theta_k s_k Y+\sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\Big)\left(\frac{\exp(\theta_{ij})}{Z}-\frac{1}{Z'}\right) \tag{85}$$

We can write $\frac{\exp(\theta_{ij})}{Z}-\frac{1}{Z'}$ as $\frac{1}{Z'Z}\sum_{Y,s}\exp\big(\theta_Y Y+\sum_{k=1}^m\theta_k s_k Y+\sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\big)(\exp(\theta_{ij})-\exp(\theta_{ij}s_i s_j))$, which is equal to $\frac{1}{Z'Z}(\exp(\theta_{ij})-\exp(-\theta_{ij}))(\exp(\theta_i-\theta_j)+\exp(-\theta_i+\theta_j))\sum_{Y,s_{-i,j}}\exp\big(\theta_Y Y+\sum_{k\neq i,j}\theta_k s_k Y+\sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\big)$. Therefore, Δ_{ij} is equal to

$$\Delta_{ij}=2\left(\frac{\exp(\theta_{ij})}{Z}-\frac{1}{Z'}\right)\sum_{Y,s_{-i,j}}\exp\big(\theta_Y Y+\sum_{k\neq i,j}\theta_k s_k Y+\sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\big)\big(\exp(\theta_i Y+\theta_j Y)+\exp(-\theta_i Y-\theta_j Y)\big) \tag{86}$$

$$=\frac{2}{Z'Z}(\exp(\theta_{ij})-\exp(-\theta_{ij}))(\exp(\theta_i-\theta_j)+\exp(-\theta_i+\theta_j))(\exp(\theta_i+\theta_j)+\exp(-\theta_i-\theta_j)) \tag{87}$$

$$\times\Bigg(\sum_{Y,s_{-i,j}}\exp\big(\theta_Y Y+\sum_{k\neq i,j}\theta_k s_k Y+\sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\big)\Bigg)^2 \tag{88}$$

$$=\frac{2}{Z'Z}(\exp(\theta_{ij})-\exp(-\theta_{ij}))(\exp(2\theta_i)+\exp(-2\theta_i)+\exp(2\theta_j)+\exp(-2\theta_j)) \tag{89}$$

$$\times\Bigg(\sum_{Y,s_{-i,j}}\exp\big(\theta_Y Y+\sum_{k\neq i,j}\theta_k s_k Y+\sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\big)\Bigg)^2 \tag{90}$$

Note that $Z = \sum_{Y,s_{-i,j}} \exp(\theta_Y Y + \sum_{k\neq i,j} \theta_k s_k Y + \sum_{(k,l)\in(i,j)} \theta_{kl} s_k s_l) \sum_{s_i,s_j} \exp(s_i\theta_i + s_j\theta_j + s_i s_j \theta_{ij})$ and $Z' = \sum_{Y,s_{-i,j}} \exp(\theta_Y Y + \sum_{k\neq i,j} \theta_k s_i Y + \sum_{(k,l)\in(i,j)} \theta_{kl} s_k s_l) \sum_{s_i,s_j} \exp(s_i\theta_i + s_j\theta_j)$. Plugging this back in and canceling out summations, we obtain our desired result for Δ_{ij}.

We now can compute ε_{ij}:

$$\begin{aligned}\varepsilon_{ij} &= \mathbb{E}\left[\lambda_i\lambda_j\right] - \mathbb{E}\left[\lambda_i Y\right]\mathbb{E}\left[\lambda_j Y\right] = \mathbb{E}'\left[\lambda_i\lambda_j\right] + \Delta_{ij} - (\mathbb{E}'\left[\lambda_i Y\right] + \Delta_i)(\mathbb{E}'\left[\lambda_j Y\right] + \Delta_j) &(91)\\ &= \Delta_{ij} - \Delta_i\mathbb{E}'\left[\lambda_j Y\right] - \Delta_j\mathbb{E}'\left[\lambda_i Y\right] - \Delta_i\Delta_j &(92)\end{aligned}$$

Lastly, we compute $\mathbb{E}'\left[\lambda_i Y\right]$ and $\mathbb{E}'\left[\lambda_j Y\right]$:

$$\begin{aligned}\mathbb{E}'\left[\lambda_i Y\right] =& 2\left(\Pr'(\lambda_i = 1, Y = 1) + \Pr'(\lambda_i = -1, Y = -1)\right) - 1 &(93)\\ =& \frac{2}{Z'}\exp(\theta_i)(\exp(\theta_j) + \exp(-\theta_j)) &(94)\\ &\times \sum_{s_{-i,j}} \exp\Big(\sum_{(k,l)\neq(i,j)} \theta_{kl} s_k s_l\Big)\Big(\exp\Big(\theta_Y + \sum_{k\neq i,j}\theta_k s_k\Big) + \exp\Big(-\theta_Y - \sum_{k\neq i,j}\theta_k s_k\Big)\Big) - 1 &(95)\end{aligned}$$

Z' can be written as $\sum_{s_i,s_j}\exp(s_i\theta_i + s_j\theta_j)\sum_{s_{-i,j}}\exp\Big(\sum_{(k,l)\notin(i,j)}\theta_{kl}s_k s_l\Big)\Big(\exp(\theta_Y + \sum_{k\neq i,j}\theta_k s_k) + \exp(-\theta_Y - \sum_{k\neq i,j}\theta_k s_k)\Big)$, Therefore $\mathbb{E}'\left[\lambda_i Y\right]$ is equal to

$$\mathbb{E}'\left[\lambda_i Y\right] = \frac{2\exp(\theta_i)(\exp(\theta_j) + \exp(-\theta_j))}{\sum_{s_i,s_j}\exp(s_i\theta_i + s_j\theta_j)} - 1 \tag{96}$$

The key takeaways from this Lemma are:

1. Impact of misspecification in our computations exhibits some form of Lipschitzness, i.e. it is bounded in terms of the canonical parameters of our distribution.
2. One misspecified edge only contributes error defined in terms of the canonical parameters on the two vertices and the unmodeled edge between them.
3. Under our assumptions, $\varepsilon_{ij} > 0$.

□

Lemma 6. *In the case of unlabeled data, accuracies estimated using the triplet method in* (2) *satisfy*

$$\begin{aligned}\mathbb{E}_{\mathcal{N},\tau}\left[\widetilde{a}_i - \bar{a}_i\right] &\leq \frac{\sqrt{3}}{2\sqrt{n_U}} \cdot \sqrt{\frac{1-b_{\min}^2}{b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)}\\ \mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i - \bar{a}_i)^2\right] &\leq \frac{3}{4n_U} \cdot \frac{1-b_{\min}^2}{b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)\end{aligned}$$

Proof. First, note that $\mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i - \widetilde{a}_i\right] = \mathbb{E}_{\mathcal{N},\tau}\left[\mathbb{E}_\tau\left[\bar{a}_i^{(j,k)}\right] - \widetilde{a}_i\right] = \mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i^{(j,k)} - \widetilde{a}_i\right]$. Therefore, is it sufficient to produce an upper bound on $\mathbb{E}_{\mathcal{N}}\left[\bar{a}_i^{(j,k)} - \widetilde{a}_i | \lambda_j, \lambda_k\right]$ independent of j, k. For ease of notation, we efer to this expectation as $\mathbb{E}\left[\bar{a}_i - \widetilde{a}_i\right]$. Then, $\mathbb{E}\left[\bar{a}_i - \widetilde{a}_i\right] = \mathbb{E}\left[\frac{\bar{a}_i^2 - \widetilde{a}_i^2}{\bar{a}_i + \widetilde{a}_i}\right] \leq \frac{1}{2b_{\min}}\mathbb{E}\left[|\bar{a}_i^2 - \widetilde{a}_i^2|\right]$. Denote $M_{ij} = \mathbb{E}\left[\lambda_i\lambda_j\right]$ and $\hat{M}_{ij} = \hat{\mathbb{E}}\left[\lambda_i\lambda_j\right]$. Then, by definition of our estimator in (2),

$$\begin{aligned}\mathbb{E}\left[\bar{a}_i - \widetilde{a}_i\right] &\leq \frac{1}{2b_{\min}}\mathbb{E}\left[\frac{\hat{M}_{ij}\hat{M}_{ik}}{\hat{M}_{jk}M_{jk}}|\hat{M}_{jk} - M_{jk}| + \frac{\hat{M}_{ij}}{M_{jk}}|\hat{M}_{ik} - M_{ik}| + \frac{M_{ik}}{M_{jk}}|\hat{M}_{ij} - M_{ij}|\right] &(97)\\ &\leq \frac{1}{2b_{\min}}\mathbb{E}\left[\frac{1}{b_{\min}^2}|\delta_{jk}| + \frac{1}{b_{\min}}|\delta_{ik}| + \frac{1}{b_{\min}}|\delta_{ij}|\right],\end{aligned}$$

i	0	1	2	3	4	5	6	7	8	9
Accuracy	.6893	.6072	.5954	.6603	.6939	.6346	.7462	.6870	.6462	.6284

Table 3: The source accuracies used for synthetic experiments. They were each drawn uniformly from $[.55, .75]$.

where $\delta_{ij} = \hat{M}_{ij} - M_{ij}$ is the estimation error for the pairwise expectations. Using Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[\bar{a}_i - \widetilde{a}_i\right] &\leq \frac{1}{2b_{\min}}\sqrt{\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}}\mathbb{E}\left[\sqrt{\delta_{ij}^2 + \delta_{ik}^2 + \delta_{jk}^2}\right] \\ &\leq \frac{1}{2b_{\min}}\sqrt{\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}}\sqrt{\mathrm{Var}\left(\hat{M}_{ij}\right) + \mathrm{Var}\left(\hat{M}_{ik}\right) + \mathrm{Var}\left(\hat{M}_{jk}\right)}\end{aligned}$$

Formally, $\hat{M}_{ij} = \frac{1}{n_U}\sum_{l=1}^{n_U}\lambda_i^l\lambda_j^l$. Therefore, $\mathrm{Var}\left(M_{ij}\right) = \frac{1}{n_U^2}\sum_{l=1}^{n_U}\mathbb{E}\left[\lambda_i^{l2}\lambda_j^{l2}\right] - M_{ij}^2 = \frac{1-M_{ij}^2}{n_U} \leq \frac{1-b_{\min}^2}{n_U}$, and our bound becomes

$$\mathbb{E}\left[\bar{a}_i - \widetilde{a}_i\right] \leq \frac{\sqrt{3}}{2\sqrt{n_U}} \cdot \sqrt{\frac{1-b_{\min}^2}{b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)}$$

Next, to bound $\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i - \bar{a}_i)^2\right]$, it is sufficient to upper bound $\mathbb{E}_{\mathcal{N}}\left[(\widetilde{a}_i - \bar{a}_i^{(j,k)})^2 \mid \lambda_j, \lambda_k\right]$ independent of choice of j and k. Refer to this expectation as $\mathbb{E}\left[(\widetilde{a}_i - \bar{a}_i)^2\right]$. Then, $\mathbb{E}\left[(\widetilde{a}_i - \bar{a}_i)^2\right] = \mathbb{E}\left[\frac{(\widetilde{a}_i^2 - \bar{a}_i^2)^2}{(\widetilde{a}_i + \bar{a}_i)^2}\right] \leq \frac{1}{4b_{\min}^2}\mathbb{E}\left[(\widetilde{a}_i^2 - \bar{a}_i^2)^2\right]$. Similar to (97),

$$\begin{aligned}\mathbb{E}\left[(\widetilde{a}_i - \bar{a}_i)^2\right] &\leq \frac{1}{4b_{\min^2}}\mathbb{E}\left[\left(\frac{\hat{M}_{ij}\hat{M}_{ik}}{\hat{M}_{jk}M_{jk}}|\hat{M}_{jk} - M_{jk}| + \frac{\hat{M}_{ij}}{M_{jk}}|\hat{M}_{ik} - M_{ik}| + \frac{M_{ik}}{M_{jk}}|\hat{M}_{ij} - M_{ij}|\right)^2\right] && (98) \\ &\leq \frac{1}{4b_{\min}^2}\mathbb{E}\left[\left(\frac{1}{b_{\min}^2}|\delta_{jk}| + \frac{1}{b_{\min}}|\delta_{ik}| + \frac{1}{b_{\min}}|\delta_{ij}|\right)^2\right] && (99) \\ &\leq \frac{1}{4b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)\left(\mathrm{Var}\left(\hat{M}_{ij}\right) + \mathrm{Var}\left(\hat{M}_{ik}\right) + \mathrm{Var}\left(\hat{M}_{jk}\right)\right) && (100) \\ &\leq \frac{3}{4n_U} \cdot \frac{1-b_{\min}^2}{b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right) && (101)\end{aligned}$$

□

E Additional Experimental Details

E.1 Synthetic Experiments

In this section, we first provide our protocol for generating synthetic data, which is fixed across our synthetic experiments. We then discuss the details of the experiments performed for each of the plots in section 4 and section 5.

Generating synthetic data We use the same synthetic data distributions for all of our synthetic experiments. We set the number of sources to $m = 10$, and draw accuracies uniformly from $[.55, .75]$, both of which would be typical in relevant applications (ex., in weak supervision). We report these accuracies in Table 3. For experiments with dependencies, when $d = 1$ we add the edge $(0, 1)$, when $d = 2$ we add a second edge $(2, 3)$ and so on. Every dependency is fixed at $\varepsilon_{ij} = \mathbb{E}[\lambda_i\lambda_j] - \mathbb{E}[\lambda_i]\mathbb{E}[\lambda_j] = 0.1$.

Figure 3: Excess generalization error We measure the expected excess generalization error for several different estimators and values of n. For each value of n, we take 1000 samples and measure the generalization error of an estimator trained on this sample. We average the results over these 1000 samples.

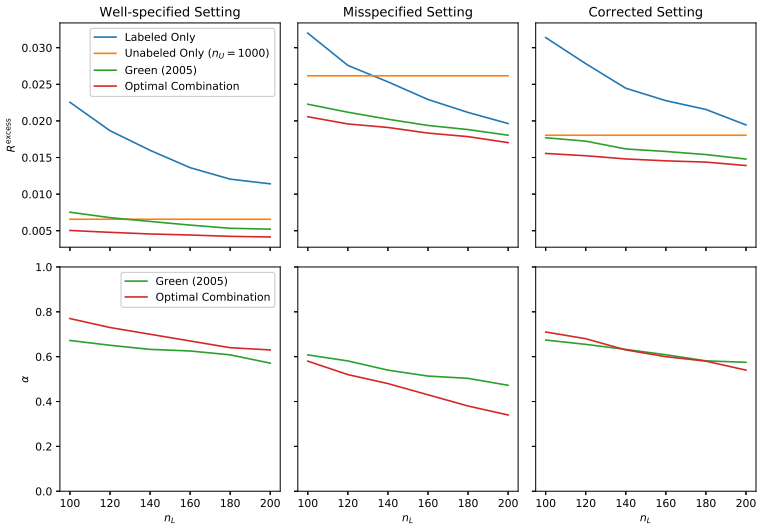


Figure 7: Excess generalization error and associated combination weight α for an optimally weighted combination of labeled and unlabeled estimators, and a combination weighted according to Green et al. (2005) across the well-specified (left), misspecified (center), and corrected (right) settings. The number of unlabeled points is fixed at $n_U = 1000$.

Figure 4: Computing the data value ratio We compute the data value ratio for unlabeled models with mean and median aggregation for different numbers of dependencies d. The definition of the data value ratio requires finding the smallest n_L with which learning from n_L labeled points achieves lower expected generalization error than learning from n_U unlabeled points. To measure the expected generalization error for some n, we average over 1000 samples, which would be intractable to do for every n_L. So, we measure the expected generalization error for every n_L between 10 and 100, every n_L divisible by 2 between 100 and 1000 and every n_L divisible by 10 between 1000 and 5000. Besides this shortcut, we compute the data value ratio according to its definition.

Figure 5: Combining labeled and unlabeled data We compare the practical approach of weighting the unlabeled and labeled estimators according to Green et al. (2005), formally defined in section B.3, with the optimal weight. We let the optimal weight vary with n_U and n_L, but not with the specific data points drawn. In other words, we compute the optimal weight to be that which minimizes the average generalization error over 1000 trials for each n_L. On the other hand, the weight from Green et al. (2005) is a function of the learned accuracies (and thus of the specific data points drawn). In Figure 7 we report the optimal α for each n_L (n_U is fixed at 1000) as well as the *average* weight from Green et al. (2005) over 1000 trials.

E.2 Real-World Case Study: Weak Supervision

We discuss the weak supervision dataset we create and clarify the details of our experimental protocol for the real-world case study.

Creating a weak supervision dataset In weak supervision, soft labels from latent variable estimation are used as an alternative to a hand-labeled dataset. The sources used are usually heuristics which incorporate domain-specific knowledge about a particular task and can be acquired relatively cheaply. For our real-world case study, we choose the simple sentiment analysis task of classifying IMDB reviews as positive or negative. Our sources are defined simply: for a collection of positive sentiment words, output "yes" if the word appears in the review and "no" otherwise; for a collection of negative sentiment words, similarly output "no" if the word appears and "yes" otherwise. The specific words used and their sentiments are reported in Table 4. We select these words because they are empirically predictive, appear relatively frequently in reviews and are intuitively associated with positive/negative reviews.

Figure 6 and Table 1: Experiments with real data We measure excess generalization error, the data value ratio and the performances of combined estimators for the real-world dataset. Our protocols for these

Word	love	like	good	great	best	excellent	terrible	worst	bad	better	could	would
Sentiment	+	+	+	+	+	+	-	-	-	-	-	-

Table 4: The words used as sources for the real-world weak supervision task of classifying IMDB reviews as positive or negative.

experiments mirror those we used for synthetic datasets, with two key differences: (1) for each trial, we sample points uniformly from the training set of 40,000 points, since we cannot sample directly from the distribution and (2) we measure generalization error on the test set, since we cannot compute the expected generalization error directly.