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 $\boldsymbol{\lambda}$.
$\widetilde{Y}$	Soft 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}	Mean squared error for $\tilde{a}_i^M$, $\rho_{n_U} = \max_i \mathbb{E}\left[(\widetilde{a}_i^M - a_i)^2\right]$.
$f(n_U)$	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.

Algorithm 1 Method-of-Moments Latent Variable Estimation (Fu et al., 2020)

Input: Empirical expectation estimates $\hat{\mathbb{E}}[\lambda_i\lambda_j]$
for $i = 1$ **to** m **do**
  $A = \emptyset$
  for $j, k \in \{1, \ldots, m\} \backslash \{i\}$ **do**
    $\widetilde{a}_i^{(j,k)} \leftarrow \sqrt{|\hat{\mathbb{E}}[\lambda_i\lambda_j] \cdot \hat{\mathbb{E}}[\lambda_i\lambda_k] / \hat{\mathbb{E}}[\lambda_j\lambda_k]|}$
    $A \leftarrow A \cup \widetilde{a}_i^{(j,k)}$
  end for
  $\widetilde{a}_i \leftarrow \textsc{Aggregate}(A)$
end for
return $\widetilde{a}$, estimates of $\mathbb{E}[\lambda_i Y]$ for all λ_i

B Additional Algorithmic Details

We provide more details on our algorithm for latent variable estimation. The input is either a labeled dataset $(\boldsymbol{X}_L, \boldsymbol{Y}_L)$ or unlabeled dataset $\boldsymbol{X}_U$ with m sources $\boldsymbol{\lambda}$. The output is an estimate of the distribution $\Pr(Y|\boldsymbol{\lambda}(X))$, which we construct using the factorization in (1). For both data types, this requires plugging in the values of $\Pr(Y = 1)$ and the empirical distribution of the sources, $\hat{\Pr}(\boldsymbol{\lambda} = \boldsymbol{\lambda}(X))$.

The approach to estimating $\widetilde{\Pr}(\lambda_i = \lambda_i(X)|Y = 1)$ is the only part of the method that differs between the labeled and unlabeled settings. For both, we can focus on estimating $\mathbb{E}[\lambda_i Y]$ since $\widetilde{\Pr}(\lambda_i = \pm 1|Y = 1) = \frac{1 \pm \widetilde{a}_i}{2}$ by Lemma 2. In the labeled setting, the expectation can be estimated directly, i.e. $\hat{\mathbb{E}}[\lambda_i Y] = \frac{1}{n_L}\sum_{j=1}^{n} \lambda_i(x_j)y_j$. On the other hand, for unlabeled data we use the triplet method from Fu et al. (2020), described in Algorithm 1, to estimate $\mathbb{E}[\lambda_i Y]$. This algorithm takes as input the pairwise rates of agreement between sources $\hat{\mathbb{E}}[\lambda_i\lambda_j]$ for all i, j, and returns an estimate of each $\mathbb{E}[\lambda_i Y]$.

The Aggregate subroutine in Algorithm 1 distinguishes between the unlabeled case with and without correction. For unlabeled data, we theoretically analyze the approach where we choose $\widetilde{a}_i \sim \text{Unif}(A)$; that is, we randomly select two λ_j, λ_k to compute $\widetilde{a}_i^U$, which is similarly done in other method-of-moments approaches. An alternate to this approach is to take the *mean* over all possible pairs λ_j, λ_k; note that this reduces the estimation error compared to the population-level estimate by a factor of $\binom{m-1}{2}$, but does not mitigate bias from misspecification. We use this approach in our synthetic and real-world experiments for the baseline unlabeled case without correction. Lastly, having $\textsc{Aggregate}(A)$ be the median of the set A is our proposed method of correcting for misspecification.

C Additional Theoretical Results

In Section C.1, 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, in Section C.2 we give more details about the combined estimators and the generalization bounds from using them. Finally, in Section C.3 we present a lower asymptotic bound on the generalization error for labeled versus unlabeled data and combining both.

C.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 and the same conditions from Proposition 1, which we prove in Section D.4, hold for the estimates to be consistent.

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 $\mathbb{E}[x_1 \otimes x_2 \otimes x_3]$ plus some additional perturbation that is a function of the probability distribution. This error is propagated into the estimate of μ_h, assuming 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.

C.2 Combined estimator analysis

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

$$\widetilde{a}^G := \widetilde{a}^U + \left(1 - \frac{r}{\|\widetilde{a}^L - \widetilde{a}^U\|_{\Sigma^{-1}}}\right)_+ (\widetilde{a}^L - \widetilde{a}^U), \tag{7}$$

where $\Sigma = \mathbf{Cov}\left[\widetilde{a}^L\right]$ and $r \in [0, 2(m-2)]$. Green et al. (2005) show that this estimator dominates $\widetilde{a}^L$ when the unbiased estimator is Gaussian and its covariance is known, but since we can only estimate the covariance matrix, we replace Σ with an empirical estimate $\hat{\Sigma}$ in practice. This estimator is equivalent to $a^{\text{lin}}\Big(\min\Big\{\frac{r}{\|\widetilde{a}^L - \widetilde{a}^U\|_{\hat{\Sigma}^{-1}}}, 1\Big\}\Big)$. We thus focus on analyzing the performance of the general combined estimator $a^{\text{lin}}(\alpha)$.

The change in estimator only impacts the generalization bound via the parameter estimation error, $\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\text{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right]$. 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

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

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

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

$$\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\text{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right] = \sum_{i=1}^m \frac{1}{2}\Big(\frac{1}{1-a_i^2}\Big)\Big(\alpha^2\mathbb{E}\left[(\widetilde{a}_i^U - \bar{a}_i)^2\right] + (1-\alpha)^2\mathbb{E}\left[(\widetilde{a}_i^L - a_i)^2\right]\Big) \tag{9}$$

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',

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

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

$$\sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] = \sum_{i=1}^{m} \frac{1}{2}\Big(\frac{1}{1-a_i^2}\Big)\Big(\alpha^2\mathbb{E}\left[(\widetilde{a}_i^M - \bar{a}_i)^2\right] + (1-\alpha)^2\mathbb{E}\left[(\widetilde{a}_i^L - a_i)^2\right]\Big) \tag{10}$$

$\mathbb{E}\left[(\widetilde{a}_i^M - \bar{a}_i)^2\right]$ 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.3 Lower bounds on generalization error

While Theorems 2 and 3 provide upper bounds on the excess generalization error, it is also important to consider lower bounds—is the standing bias from misspecification in the unlabeled approach inevitable? We analyze the asymptotic excess risk in the case of labeled data, unlabeled data, and both, and discuss how a lower bound approach to the data value ratio and analyzing combined estimators is possible.

Unlabeled data lower bound Looking at the decomposition in Theorem 1, $\mathbb{E}_{\mathcal{N}}\left[D_{\mathrm{KL}}(\mathrm{Pr}(\boldsymbol{\lambda})||\hat{\mathrm{Pr}}(\boldsymbol{\lambda}))\right]$ approaches 0 asymptotically and the inference bias $\sum_{(i,j)\in E_\lambda} I(\lambda_i;\lambda_j|Y)$ is independent of the amount of data. We thus seek to asymptotically lower bound $\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right]$. Note that in the labeled data case and when using the medians estimator $\widetilde{a}^M$ with unlabeled data, parameter estimation error approaches 0 as n grows large since the estimated accuracy parameters are 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*

$$\lim_{n_U\to\infty} R_u^e \geq \frac{(m-2d)d^2\varepsilon_{\min}^2 b_{\min}^4}{2(m-1)^2(m-2)^2} + \mathcal{B}_I. \tag{11}$$

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

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

$$\lim_{n_U\to\infty}\sum_{i=1}^{m} \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\mathrm{Pr}_{\lambda_i|Y}||\widetilde{\mathrm{Pr}}_{\lambda_i|Y})\right] = \sum_{i=1}^{m}\frac{1+a_i}{2}\log\left(1+\frac{a_i-\bar{a}_i}{1+\bar{a}_i}\right) + \frac{1-a_i}{2}\log\left(1+\frac{\bar{a}_i-a_i}{1-\bar{a}_i}\right). \tag{12}$$

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

$$\frac{1+a_i}{2}\log\left(1+\frac{a_i-\bar{a}_i}{1+\bar{a}_i}\right) + \frac{1-a_i}{2}\log\left(1+\frac{\bar{a}_i-a_i}{1-\bar{a}_i}\right) = -\frac{1+a}{2}\log\left(1-\frac{x}{1+a}\right) - \frac{1-a}{2}\log\left(1+\frac{x}{1-a}\right). \tag{13}$$

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

$$f(x) = \frac{1+a}{2}\log\left(1-\frac{x}{1+a}\right) + \frac{1-a}{2}\log\left(1+\frac{x}{1-a}\right). \tag{14}$$

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 (12), we have that $\lim_{n_U\to\infty}\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right] = \sum_{i=1}^m \frac{1}{2}(a_i-\bar{a}_i)^2$. 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 $i\notin E_\lambda$, we know that $|a_i-\bar{a}_i|$ is at least $\frac{d\varepsilon_{\min}b_{\min}^2}{(m-1)(m-2)}$. Therefore,

$$\frac{1}{2}\sum_{i=1}^m (a_i-\bar{a}_i)^2 \geq \frac{1}{2}\sum_{i\notin E_\lambda}(a_i-\bar{a}_i)^2 \geq \frac{(m-2d)d^2\varepsilon_{\min}^2 b_{\min}^4}{2(m-1)^2(m-2)^2}. \tag{15}$$

□

Combined estimator lower bound Next, we analyze the excess risk when we use the combined estimator $a^{\mathrm{lin}}(\alpha)$. Note that when we are in the well-specified and corrected settings, the asymptotic excess risk is 0. Therefore, we only consider the misspecified setting.

Corollary 1. *Denote $R^e_{\mathrm{lin}}(\alpha)$ as the excess risk of our latent variable model when we use accuracy parameter $a^{\mathrm{lin}}(\alpha)$. The lower bound of the excess generalization error when we combine labeled and unlabeled data (without correction) using weight α is asymptotically bounded by*

$$\lim_{n_U,n_L\to\infty} R^e_{\mathrm{lin}}(\alpha) \geq \frac{\alpha^2(m-2d)d^2\varepsilon_{\min}^2 b_{\min}^4}{2(m-1)^2(m-2)^2} + \mathcal{B}_I. \tag{16}$$

Proof. Based on (8), the asymptotic parameter estimation error is

$$\lim_{n_U,n_L\to\infty}\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right] = \sum_{i=1}^m \frac{1+a_i}{2}\log\Big(1+\frac{\alpha(a_i-\bar{a}_i)}{1+\bar{a}_i^C}\Big) + \frac{1-a_i}{2}\log\Big(1+\frac{\alpha(\bar{a}_i-a_i)}{1-\bar{a}_i^C}\Big), \tag{17}$$

where $\bar{a}_i^C = \alpha\bar{a}_i + (1-\alpha)a_i$. If we define $a := a_i$ and $x := \alpha(a_i-\bar{a}_i)$, then the ith element of this sum has the form in (13) and is thus at least $\frac{1}{2}x^2$. Therefore, using results from Lemma 1 the parameter estimation error is

$$\lim_{n_U,n_L\to\infty}\sum_{i=1}^m \mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right] \geq \sum_{i=1}^m \frac{\alpha^2}{2}(a_i-\bar{a}_i)^2 \geq \frac{\alpha^2(m-2d)d^2\varepsilon_{\min}^2 b_{\min}^4}{2(m-1)^2(m-2)^2}. \tag{18}$$

□

Applications to data value ratio and combined estimator analysis Finally, it is possible to define the data value ratio and analyze combined estimators based on lower bounds on the excess risk of labeled vs unlabeled data. To do this, we would use the expressions from Theorem 4 and Corollary 1 with standard finite-sample lower bounds on the estimates from observable data. For bounding the variance of accuracy parameters estimated via the triplet method on unlabeled data, we can use the lower bound from Theorem 2 of Fu et al. (2020).

D Proofs

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

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

$$\Pr(Y,\boldsymbol{\lambda};\theta) = \frac{1}{Z}\exp\Big(\theta_Y + \sum_{i=1}^m \theta_i\lambda_i Y + \sum_{(i,j)\in E_\lambda}\theta_{ij}\lambda_i\lambda_j\Big), \tag{19}$$

where Z is the cumulant function, and the set of 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 $\bar{a}_{\max} = \max_i \bar{a}_i = \max_{i,j,k}\mathbb{E}_\tau\left[\sqrt{\frac{\mathbb{E}[\lambda_i\lambda_j]\mathbb{E}[\lambda_i\lambda_k]}{\mathbb{E}[\lambda_j\lambda_k]}}\right]$.

D.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

$$\mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[l(\widetilde{Y},Y)\right] = -\mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[\log\frac{\widetilde{\Pr}(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\Pr(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}\right] + H(Y|\boldsymbol{\lambda}), \tag{20}$$

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

$$H(Y|\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\lambda}}\left[-\Pr(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\log\Pr(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda}) - \Pr(Y=-1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\log\Pr(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\right]. \tag{21}$$

Next, we evaluate $\log\frac{\widetilde{\Pr}(Y'=Y|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\Pr(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})}$. Define $\overline{\Pr}$ 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 $\widetilde{\Pr}$'s definition in (1),

$$\overline{\Pr}(Y'=Y|\boldsymbol{\lambda}=\boldsymbol{\lambda}(X)) = \frac{\overline{\Pr}(\boldsymbol{\lambda}=\boldsymbol{\lambda}(X)|Y'=Y)\Pr(Y'=Y)}{\Pr(\lambda=\lambda(X))} = \frac{\prod_{i=1}^m \Pr(\lambda_i=\lambda_i(X)|Y'=Y)\Pr(Y'=Y)}{\Pr(\lambda=\lambda(X))}. \tag{22}$$

Then,

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

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 (20), we get

$$-\sum_{i=1}^m \mathbb{E}_{(Y,\boldsymbol{\lambda}),\mathcal{N},\tau}\left[\log\frac{\widetilde{\Pr}(\lambda_i'=\lambda_i|Y'=Y)}{\Pr(\lambda_i'=\lambda_i|Y'=Y)}\right] - \mathbb{E}_{(Y,\boldsymbol{\lambda})}\left[\log\frac{\overline{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda}|Y'=Y)}{\Pr(\boldsymbol{\lambda}'=\boldsymbol{\lambda}|Y'=Y)}\right] - \mathbb{E}_{\boldsymbol{\lambda},\mathcal{N}}\left[\log\frac{\Pr(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\hat{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}\right] + H(Y|\boldsymbol{\lambda}). \tag{23}$$

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,

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

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

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

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. $-\mathbb{E}_{\boldsymbol{\lambda},\mathcal{N}}\left[\log\frac{\Pr(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}{\hat{\Pr}(\boldsymbol{\lambda}'=\boldsymbol{\lambda})}\right]$:
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, (23) becomes

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

D.2 Proof of Theorem 2

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

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

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, $\sum_{i=1}^{m}\mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right] = \frac{m}{2n_L} + o(1/n_L)$, and our proof is complete.

D.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

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

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*

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

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

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

And for all i, it is thus true that

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

Proof. We define $\varepsilon_{ij}=\mathbb{E}\left[\lambda_i\lambda_j\right]-\mathbb{E}\left[\lambda_iY\right]\mathbb{E}\left[\lambda_jY\right]$ 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 $0<\varepsilon_{\min}\le\varepsilon_{ij}\le\varepsilon_{\max}$ 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

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

Note that this means $\bar{a}_i\ge b_{\min}$. When each $\mathbb{E}\left[\lambda_i\lambda_j\right]$ can be written as $\mathbb{E}\left[\lambda_iY\right]\mathbb{E}\left[\lambda_jY\right]$, 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 $\mathbb{E}\left[\lambda_i\lambda_j\right]-\varepsilon_{ij}$, which is equal to $\mathbb{E}\left[\lambda_iY\right]\mathbb{E}\left[\lambda_jY\right]$, rather than $\mathbb{E}\left[\lambda_i\lambda_j\right]$.

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

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

This means that $\bar{a}_i\ge a_i$ 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 $\bar{a}_i\ge a_i$ 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

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

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

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

$$\bar{a}_i - a_i \geq \frac{2}{m-1} \cdot \frac{\varepsilon_{\min} b_{\min}}{2} + \frac{2(d-1)}{(m-1)(m-2)} \cdot \frac{-\varepsilon_{\max}}{2b_{\min}^2 a_{\min}^2} = \frac{\varepsilon_{\min} b_{\min}}{m-1} - \frac{(d-1)\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2 a_{\min}^2}. \tag{31}$$

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

$$\bar{a}_i - a_i \leq \frac{2d}{(m-1)(m-2)} \cdot \frac{-\varepsilon_{\min} b_{\min}^2}{2} = \frac{-d\varepsilon_{\min} b_{\min}^2}{(m-1)(m-2)}, \tag{32}$$

$$\bar{a}_i - a_i \geq \frac{2d}{(m-1)(m-2)} \cdot \frac{-\varepsilon_{\max}}{2b_{\min}^2 a_{\min}^2} = \frac{-d\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2 a_{\min}^2}. \tag{33}$$

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

$$|\bar{a}_i - a_i| \leq \frac{\varepsilon_{\max}}{(m-1)b_{\min}^2 a_{\min}^2}. \tag{34}$$

□

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

$$\begin{aligned}
&(m-2d)\log\left(1+\frac{d\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2 a_{\min}^2(1+b_{\min})}\right)+2d\frac{\varepsilon_{\max}^2}{(1-\bar{a}_{\max}^2)(m-1)^2 b_{\min}^4 a_{\min}^4} \qquad (35)\\
\leq&\frac{(m-2d)d\varepsilon_{\max}}{(m-1)(m-2)b_{\min}^2 a_{\min}^2(1+b_{\min})}+\frac{2d\varepsilon_{\max}}{(1-\bar{a}_{\max}^2)(m-1)^2 b_{\min}^4 a_{\min}^4}\\
=&\frac{d\varepsilon_{\max}}{(m-1)b_{\min}^2 a_{\min}^2}\left(\frac{m-2d}{(m-2)(1+b_{\min})}+\frac{2}{(1-\bar{a}_{\max}^2)(m-1)b_{\min}^2 a_{\min}^2}\right)\\
\leq&\frac{d\varepsilon_{\max}}{(m-1)b_{\min}^2 a_{\min}^2}\left(1+\frac{1}{(1-\bar{a}_{\max}^2)b_{\min}^2 a_{\min}^2}\right)\leq\frac{c_1 d\varepsilon_{\max}}{m},
\end{aligned}$$

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

$$\begin{aligned}
\sum_{i=1}^m \frac{a_i-\bar{a}_i}{1-\bar{a}_i^2}\mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i-\widetilde{a}_i\right]\leq&\sum_{i=1}^m\frac{|\bar{a}_i-a_i|}{1-\bar{a}_i^2}\mathbb{E}_{\mathcal{N},\tau}\left[|\bar{a}_i-\widetilde{a}_i|\right] \qquad (36)\\
\leq&\frac{\sqrt{3}}{2\sqrt{n_U}}\cdot\sqrt{\frac{1-b_{\min}^2}{b_{\min}^2}\left(\frac{1}{b_{\min}^4}+\frac{2}{b_{\min}^2}\right)}\frac{1}{1-\bar{a}_{\max}^2}\left(\frac{m\varepsilon_{\max}}{(m-1)b_{\min}^2 a_{\min}^2}\right)\leq\frac{c_2\varepsilon_{\max}}{\sqrt{n_U}},
\end{aligned}$$

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 $\sum_{i=1}^m \frac{1}{2}\left(\frac{1}{1-\bar{a}_i^2}+\frac{2\bar{a}_i(\bar{a}_i-a_i)}{(1-\bar{a}_i^2)^2}\right)\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right]$, which can be split into an expression independent of misspecification and one dependent on it:

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

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 $\bar{a}_i \leq a_i$ for $i \notin E_\lambda$:

$$\begin{aligned}
\sum_{i=1}^m\frac{\bar{a}_i-a_i}{(1-\bar{a}_i^2)^2}\mathbb{E}_{\mathcal{N},\tau}\left[(\widetilde{a}_i-\bar{a}_i)^2\right]\leq&\frac{3}{4n_U}\cdot\frac{1-b_{\min}^2}{b_{\min}^2(1-\bar{a}_{\max}^2)^2}\left(\frac{1}{b_{\min}^4}+\frac{2}{b_{\min}^2}\right)\sum_{i\in E_\lambda}|\bar{a}_i-a_i| \qquad (38)\\
\leq&\frac{3}{4n_U}\cdot\frac{1-b_{\min}^2}{b_{\min}^2(1-\bar{a}_{\max}^2)^2}\left(\frac{1}{b_{\min}^4}+\frac{2}{b_{\min}^2}\right)\left(\frac{2d\varepsilon_{\max}}{(m-1)b_{\min}^2 a_{\min}^2}\right)\leq\frac{c_3 d\varepsilon_{\max}}{mn_U},
\end{aligned}$$

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.

D.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 $\mathbb{E}\left[|\widetilde{a}_i - \bar{a}_i|\right] \leq \frac{1}{2}\min_{(j,k)}|a_i - \bar{a}_i^{(j,k)}|$. 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 $n_U \geq n_0 \sim \Omega(1/\varepsilon_{\min}^2)$.

Lastly, we compute the excess risk when using the corrected estimator. From Lemma 1, since the asymptotic expectation $\bar{a}$ of the estimator is equal to the true accuracy a, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right] = \frac{1+a_i}{2}\left(\frac{\mathbb{E}[a_i - \widetilde{a}_i^M]}{1+a_i} + \frac{1}{2(1+a_i)^2}\mathbb{E}[(\widetilde{a}_i^M - a_i)^2]\right) \\ + \frac{1-a_i}{2}\left(\frac{\mathbb{E}[\widetilde{a}_i^M - a_i]}{1-a_i} + \frac{1}{2(1-a_i)^2}\mathbb{E}[(\widetilde{a}_i^M - a_i)^2]\right).\end{aligned} \tag{39}$$

Note that $\frac{1+a_i}{2}\cdot\frac{\mathbb{E}[a_i - \widetilde{a}_i^M]}{1+a_i} + \frac{1-a_i}{2}\cdot\frac{\mathbb{E}[\widetilde{a}_i^M - a_i]}{1-a_i} = 0$. Then the parameter estimation error is

$$\sum_{i=1}^m\left(\frac{1}{4(1+a_i)} + \frac{1}{4(1-a_i)}\right)\mathbb{E}\left[(\widetilde{a}_i^M - a_i)^2\right] = \sum_{i=1}^m \frac{1}{2(1-a_i^2)}\mathbb{E}\left[(\widetilde{a}_i^M - a_i)^2\right] \leq \frac{1}{2(1-\max_i a_i^2)}\cdot m\rho_{n_U}. \tag{40}$$

This completes our proof, where $c_\rho = \frac{1}{2(1-\max_i a_i^2)}$ in Proposition 1.

E Auxiliary Lemmas

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

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

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*

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

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

$$-\mathbb{E}_{\boldsymbol{\lambda},\mathcal{N},\tau}\left[\Pr(Y=1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\log\frac{\widetilde{\Pr}(\lambda_i' = \lambda_i|Y=1)}{\Pr(\lambda_i' = \lambda_i|Y=1)} + \Pr(Y=-1|\boldsymbol{\lambda}'=\boldsymbol{\lambda})\log\frac{\widetilde{\Pr}(\lambda_i' = \lambda_i|Y=-1)}{\Pr(\lambda_i' = \lambda_i|Y=-1)}\right]. \tag{41}$$

Suppose $\lambda_i \notin E_\lambda$. Conditioning on the value of λ_i and using Lemma 2, (41) becomes

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

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

$$
\mathbb{E}_{\mathcal{N},\tau,Y}\left[D_{\mathrm{KL}}(\Pr_{\lambda_i|Y}||\widetilde{\Pr}_{\lambda_i|Y})\right]=-\mathbb{E}_{\mathcal{N},\tau}\left[\frac{1+a_i}{2}\log\frac{1+\widetilde{a}_i}{1+a_i}+\frac{1-a_i}{2}\log\frac{1-\widetilde{a}_i}{1-a_i}\right]. \tag{42}
$$

Now suppose that $\lambda_i \in E_\lambda$ and has an edge to some λ_j. When we simplify (41) by conditioning on λ_i,λ_j, we find that $\sum_{l\in\{\pm1\}}\Pr(Y=1|\boldsymbol{\lambda}_{-i,j},\lambda_i=1,\lambda_j=l)\Pr(\lambda_i=1,\lambda_j=l|\boldsymbol{\lambda}_{-i,j})+\Pr(Y=-1|\boldsymbol{\lambda}_{-i,j},\lambda_i=-1,\lambda_j=l)\Pr(\lambda_i=-1,\lambda_j=l|\boldsymbol{\lambda}_{-i,j})$ (i.e, the coefficient for $\log\frac{1+\widetilde{a}_i}{1+a_i}$) is equal to $\Pr(\lambda_i Y=1|\boldsymbol{\lambda}_{-i,j})$, 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, (42) 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 (42), 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 $\frac{1}{3(1-\bar{a}_i)^3}\mathbb{E}_{\mathcal{N},\tau}\left[(\bar{a}_i-\widetilde{a}_i)^3\right]$.

For the labeled data case, it is easy to check that $\mathbb{E}_{\mathcal{N}}\left[(\bar{a}_i-\widetilde{a}_i)^3\right] \sim \mathcal{O}(1/n_L^2)$. 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

$$a := \bar{a}_i^{(j,k)} = \sqrt{\frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}}, \qquad \hat{a} := \widetilde{a}_i = \sqrt{\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]}}. \tag{43}$$

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:

$$\Pr\left(|\hat{\mathbb{E}}[X]-\mathbb{E}[X]| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{n\cdot 2^2}\right) = 2\exp\left(-\frac{n\epsilon^2}{2}\right). \tag{44}$$

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

$$\Pr\left(|\hat{\mathbb{E}}[X]-\mathbb{E}[X]| \geq \epsilon \vee |\hat{\mathbb{E}}[Y]-\mathbb{E}[Y]| \geq \epsilon \vee |\hat{\mathbb{E}}[XY]-\mathbb{E}[XY]| \geq \epsilon\right) \leq 6\exp\left(-\frac{n\epsilon^2}{2}\right). \tag{45}$$

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

$$|\hat{\mathbb{E}}[X]-\mathbb{E}[X]| < \epsilon, \quad |\hat{\mathbb{E}}[Y]-\mathbb{E}[Y]| < \epsilon, \quad |\hat{\mathbb{E}}[XY]-\mathbb{E}[XY]| < \epsilon. \tag{46}$$

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

$$|\hat{a}-a| = \left|\frac{1}{2\sqrt{u}}\left(\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right)\right|. \tag{47}$$

Note that

$$\begin{aligned} u &\geq \min\left(\frac{\hat{\mathbb{E}}[X]\hat{\mathbb{E}}[Y]}{\hat{\mathbb{E}}[XY]}, \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right) \geq \min\left(\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)}{\mathbb{E}[XY]+\epsilon}, \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right) \\ &\geq \min\left(\frac{(\mathbb{E}[X]/2)(\mathbb{E}[Y]/2)}{1+\epsilon}, \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right) \geq \min\left(\frac{\mathbb{E}[X]\mathbb{E}[Y]}{8}, \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right) \geq \frac{\mathbb{E}[X]\mathbb{E}[Y]}{8}. \end{aligned} \tag{48}$$

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:

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

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

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

Examining the left term in the max,

$$\begin{aligned}
\left|\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)}{\mathbb{E}[XY]+\epsilon} - \frac{\mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[XY]}\right| &= \left|\frac{(\mathbb{E}[X]-\epsilon)(\mathbb{E}[Y]-\epsilon)\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y](\mathbb{E}[XY]+\epsilon)}{\mathbb{E}[XY](\mathbb{E}[XY]+\epsilon)}\right| \\
&= \left|\frac{-\epsilon(\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[XY] + \mathbb{E}[Y]\mathbb{E}[XY] - \epsilon\mathbb{E}[XY])}{\mathbb{E}[XY](\mathbb{E}[XY]+\epsilon)}\right| \\
&\leq \epsilon\left|\frac{\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[XY] + \mathbb{E}[Y]\mathbb{E}[XY]}{\mathbb{E}[XY]^2}\right| \\
&= \epsilon C_1 > 0
\end{aligned} \tag{51}$$

Examining the right term in the max,

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

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

$$\leq \epsilon\left|\frac{\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[XY] + \mathbb{E}[Y]\mathbb{E}[XY] + \mathbb{E}[XY]}{\mathbb{E}[XY]^2/2}\right| \tag{54}$$

$$= \epsilon C_2 > 0 \tag{55}$$

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,

$$|\hat{a} - a| \leq \epsilon\frac{\sqrt{2}C_2}{\sqrt{\mathbb{E}[X]\mathbb{E}[X]}} = \epsilon C_3 \tag{56}$$

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:

$$\lim_{n\to\infty} n\mathbb{E}[|\hat{a}-a|^3] = \lim_{n\to\infty} n\left(\mathbb{E}[|\hat{a}-a|^3|B]\Pr(B) + \mathbb{E}[|\hat{a}-a|^3|\neg B]P(\neg B)\right) \tag{57}$$

$$\leq \lim_{n\to\infty} n\left(C_3^3\epsilon^3\cdot 1 + 2^3\cdot 6\exp(-n^{1/4}/2)\right) \tag{58}$$

$$= C_3^3\lim_{n\to\infty} n(n^{-3/8})^3 + 48\lim_{n\to\infty} n\exp(-n^{1/4}/2) \tag{59}$$

$$= C_3^3\lim_{n\to\infty} n^{-1/8} + 48\lim_{m\to\infty} m^4\exp(-m/2) = 0 \tag{60}$$

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. *(Quantifying per-edge misspecification.) If $(i,j) \in E_\lambda$, then*

$$\varepsilon_{ij} = \Delta_{ij} - \Delta_i a'_j - \Delta_j a'_i - \Delta_i \Delta_j, \tag{61}$$

where

$$\Delta_i = \frac{2}{z_{ij} z'_{ij}} (\exp(\theta_{ij}) - \exp(-\theta_{ij}))(\exp(2\theta_j) - \exp(-2\theta_j)) \tag{62}$$

$$\Delta_j = \frac{2}{z_{ij} z'_{ij}} (\exp(\theta_{ij}) - \exp(-\theta_{ij}))(\exp(2\theta_i) - \exp(-2\theta_i)) \tag{63}$$

$$\Delta_{ij} = \frac{2}{z_{ij} z'_{ij}} (\exp(\theta_{ij}) - \exp(-\theta_{ij}))(\exp(2\theta_i) + \exp(-2\theta_i) + \exp(2\theta_j) + \exp(-2\theta_j)) \tag{64}$$

$$a'_i = \frac{2}{z'_{ij}} \exp(\theta_i)(\exp(\theta_j) + \exp(-\theta_j)) - 1 \tag{65}$$

$$a'_j = \frac{2}{z'_{ij}} \exp(\theta_j)(\exp(\theta_i) + \exp(-\theta_i)) - 1 \tag{66}$$

$$z_{ij} = \sum_{s_i, s_j} \exp(s_i \theta_i + s_j \theta_j + s_i s_j \theta_{ij}) \tag{67}$$

$$z'_{ij} = \sum_{s_i, s_j} \exp(s_i \theta_i + s_j \theta_j) \tag{68}$$

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 $\mathbb{E}'$, that does not have an edge between λ_i and λ_j:

$$\Pr'(Y, \boldsymbol{\lambda}) = \frac{1}{Z'} \exp\Big(\theta_Y + \sum_{i=1}^{m} \theta_i \lambda_i Y + \sum_{(k,l) \neq (i,j)} \theta_{kl} \lambda_k \lambda_l\Big). \tag{69}$$

This distribution uses all the same canonical parameters as (19) 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,

$$\Delta_i = \mathbb{E}[\lambda_i Y] - \mathbb{E}'[\lambda_i Y] = \frac{2}{p} \sum_{s_{-i}} \exp\left(\theta_Y + \theta_i + \sum_{k \neq i} \theta_k l_k + \sum_{(k,l) \neq (i,j)} \theta_{kl} s_k s_l\right) \left(\frac{\exp(\theta_{ij} l_j)}{Z} - \frac{1}{Z'}\right) \tag{70}$$

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

$$\begin{aligned}
&2\exp(\theta_i) \sum_{s_{-i}} \exp\left(\theta_Y + \sum_{k \neq i} \theta_k l_k + \sum_{(k,l) \neq (i,j)} \theta_{kl} s_k s_l\right) \left(\frac{\exp(\theta_{ij} l_j)}{z_Y} - \frac{1}{z'_Y}\right) \\
=&2\exp(\theta_i + \theta_j) \sum_{s_{-i,j}} \exp\left(\theta_Y + \sum_{k \neq i,j} \theta_k l_k + \sum_{(k,l) \neq (i,j)} \theta_{kl} s_k s_l\right) \left(\frac{\exp(\theta_{ij})}{z_Y} - \frac{1}{z'_Y}\right) \\
&+2\exp(\theta_i - \theta_j) \sum_{s_{-i,j}} \exp\left(\theta_Y + \sum_{k \neq i,j} \theta_k l_k + \sum_{(k,l) \neq (i,j)} \theta_{kl} s_k s_l\right) \left(\frac{\exp(-\theta_{ij})}{z_Y} - \frac{1}{z'_Y}\right)
\end{aligned} \tag{71}$$

$\frac{\exp(\pm\theta_{ij})}{z_Y} - \frac{1}{z'_Y}$ 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

$$\frac{2}{z_Y z'_Y}\Bigg(\sum_{s_{-i,j}}\exp\Big(\theta_Y + \sum_{k\neq i,j}\theta_k l_k + \sum_{(k,l)\neq(i,j)}\theta_{kl}s_k s_l\Big)\Bigg)^2(\exp(\theta_{ij}) - \exp(-\theta_{ij})) \tag{72}$$
$$\times\big(\exp(\theta_i+\theta_j)(\exp(\theta_i-\theta_j)+\exp(-\theta_i+\theta_j)) - \exp(\theta_i-\theta_j)(\exp(\theta_i+\theta_j)+\exp(-\theta_i-\theta_j))\big)$$

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

$$\Pr(\lambda_i=1,\lambda_j=1) - \Pr'(\lambda_i=1,\lambda_j=1) \tag{73}$$
$$= \sum_{Y,s_{-i,j}}\exp\Big(\theta_Y Y + \theta_i Y + \theta_j Y + \sum_{k\neq i,j}\theta_k s_k Y + \sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\Big)\left(\frac{\exp(\theta_{ij})}{Z} - \frac{1}{Z'}\right),$$
$$\Pr(\lambda_i=-1,\lambda_j=-1) - \Pr'(\lambda_i=-1,\lambda_j=-1) \tag{74}$$
$$= \sum_{Y,s_{-i,j}}\exp\Big(\theta_Y Y - \theta_i Y - \theta_j Y + \sum_{k\neq i,j}\theta_k s_k Y + \sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\Big)\left(\frac{\exp(\theta_{ij})}{Z} - \frac{1}{Z'}\right).$$

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 two times (73) plus (74):

$$\Delta_{ij} = 2\left(\frac{\exp(\theta_{ij})}{Z} - \frac{1}{Z'}\right)\sum_{Y,s_{-i,j}}\exp\big(\theta_Y Y + \sum_{k\neq i,j}\theta_k s_k Y + \sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\big)\big(\exp(\theta_i Y + \theta_j Y) + \exp(-\theta_i Y - \theta_j Y)\big) \tag{75}$$
$$= \frac{2}{Z'Z}(\exp(\theta_{ij}) - \exp(-\theta_{ij}))(\exp(\theta_i-\theta_j) + \exp(-\theta_i+\theta_j))(\exp(\theta_i+\theta_j) + \exp(-\theta_i-\theta_j))$$
$$\times\Bigg(\sum_{Y,s_{-i,j}}\exp\big(\theta_Y Y + \sum_{k\neq i,j}\theta_k s_k Y + \sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\big)\Bigg)^2$$
$$= \frac{2}{Z'Z}(\exp(\theta_{ij}) - \exp(-\theta_{ij}))(\exp(2\theta_i) + \exp(-2\theta_i) + \exp(2\theta_j) + \exp(-2\theta_j))$$
$$\times\Bigg(\sum_{Y,s_{-i,j}}\exp\big(\theta_Y Y + \sum_{k\neq i,j}\theta_k s_k Y + \sum_{(k,l)\in(i,j)}\theta_{kl}s_k s_l\big)\Bigg)^2.$$

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}:

$$\varepsilon_{ij} = \mathbb{E}[\lambda_i\lambda_j] - \mathbb{E}[\lambda_i Y]\,\mathbb{E}[\lambda_j Y] = \mathbb{E}'[\lambda_i\lambda_j] + \Delta_{ij} - (\mathbb{E}'[\lambda_i Y] + \Delta_i)(\mathbb{E}'[\lambda_j Y] + \Delta_j) \tag{76}$$
$$= \Delta_{ij} - \Delta_i\mathbb{E}'[\lambda_j Y] - \Delta_j\mathbb{E}'[\lambda_i Y] - \Delta_i\Delta_j.$$

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

$$\begin{aligned}\mathbb{E}'[\lambda_i Y] =& 2\left(\Pr'(\lambda_i = 1, Y = 1) + \Pr'(\lambda_i = -1, Y = -1)\right) - 1 \\ =& \frac{2}{Z'}\exp(\theta_i)(\exp(\theta_j) + \exp(-\theta_j)) \\ & \times \sum_{s_{-i,j}} \exp\Big(\sum_{(k,l)\neq(i,j)} \theta_{kl} s_k s_l\Big)\Big(\exp\Big(\theta_Y + \sum_{k\neq i,j} \theta_k s_k\Big) + \exp\Big(-\theta_Y - \sum_{k\neq i,j} \theta_k s_k\Big)\Big) - 1.\end{aligned} \tag{77}$$

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 $\mathbb{E}'[\lambda_i Y]$ is equal to

$$\mathbb{E}'[\lambda_i Y] = \frac{2\exp(\theta_i)(\exp(\theta_j) + \exp(-\theta_j))}{\sum_{s_i,s_j} \exp(s_i\theta_i + s_j\theta_j)} - 1. \tag{78}$$

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. *(Estimation error of accuracies via triplet method.) In the case of unlabeled data, accuracies estimated using the triplet method in* (2) *satisfy*

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

Proof. First, note that $\mathbb{E}_{\mathcal{N},\tau}[\bar{a}_i - \widetilde{a}_i] = \mathbb{E}_{\mathcal{N},\tau}\left[\mathbb{E}_\tau\left[\bar{a}_i^{(j,k)}\right] - \widetilde{a}_i\right] = \mathbb{E}_{\mathcal{N},\tau}\left[\bar{a}_i^{(j,k)} - \widetilde{a}_i\right]$. 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 refer to this expectation as $\mathbb{E}[\bar{a}_i - \widetilde{a}_i]$. Then, $\mathbb{E}[\bar{a}_i - \widetilde{a}_i] = \mathbb{E}\left[\frac{\bar{a}_i^2 - \widetilde{a}_i^2}{\bar{a}_i + \widetilde{a}_i}\right] \leq \frac{1}{2b_{\min}}\mathbb{E}\left[|\bar{a}_i^2 - \widetilde{a}_i^2|\right]$. Denote $M_{ij} = \mathbb{E}[\lambda_i\lambda_j]$ and $\hat{M}_{ij} = \hat{\mathbb{E}}[\lambda_i\lambda_j]$. Then, by definition of our estimator in (2),

$$\begin{aligned}\mathbb{E}[\bar{a}_i - \widetilde{a}_i] &\leq \frac{1}{2b_{\min}}\mathbb{E}\left[\frac{\hat{M}_{ij}\hat{M}_{ik}}{\hat{M}_{jk}M_{jk}}|\hat{M}_{jk} - M_{jk}| + \frac{\hat{M}_{ij}}{M_{jk}}|\hat{M}_{ik} - M_{ik}| + \frac{M_{ik}}{M_{jk}}|\hat{M}_{ij} - M_{ij}|\right] \\ &\leq \frac{1}{2b_{\min}}\mathbb{E}\left[\frac{1}{b_{\min}^2}|\delta_{jk}| + \frac{1}{b_{\min}}|\delta_{ik}| + \frac{1}{b_{\min}}|\delta_{ij}|\right],\end{aligned} \tag{79}$$

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

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

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 $[.55, .75]$.

Formally, $\hat{M}_{ij} = \frac{1}{n_U}\sum_{l=1}^{n_U}\lambda_i^l\lambda_j^l$. Therefore, $\text{Var}\left(M_{ij}\right) = \frac{1}{n_U^2}\sum_{l=1}^{n_U}\mathbb{E}\left[(\lambda_i^l)^2(\lambda_j^l)^2\right] - M_{ij}^2 = \frac{1-M_{ij}^2}{n_U} \leq \frac{1-b_{\min}^2}{n_U}$, 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 (79),

$$\begin{aligned}
\mathbb{E}\left[(\widetilde{a}_i - \bar{a}_i)^2\right] &\leq \frac{1}{4b_{\min^2}}\mathbb{E}\left[\left(\frac{\hat{M}_{ij}\hat{M}_{ik}}{\hat{M}_{jk}M_{jk}}|\hat{M}_{jk} - M_{jk}| + \frac{\hat{M}_{ij}}{M_{jk}}|\hat{M}_{ik} - M_{ik}| + \frac{M_{ik}}{M_{jk}}|\hat{M}_{ij} - M_{ij}|\right)^2\right] \qquad (80)\\
&\leq \frac{1}{4b_{\min}^2}\mathbb{E}\left[\left(\frac{1}{b_{\min}^2}|\delta_{jk}| + \frac{1}{b_{\min}}|\delta_{ik}| + \frac{1}{b_{\min}}|\delta_{ij}|\right)^2\right]\\
&\leq \frac{1}{4b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)\left(\text{Var}\left(\hat{M}_{ij}\right) + \text{Var}\left(\hat{M}_{ik}\right) + \text{Var}\left(\hat{M}_{jk}\right)\right)\\
&\leq \frac{3}{4n_U} \cdot \frac{1-b_{\min}^2}{b_{\min}^2}\left(\frac{1}{b_{\min}^4} + \frac{2}{b_{\min}^2}\right)
\end{aligned}$$

□

F Additional Experimental Details

We provide additional details on experiments. Our code can be found at `https://github.com/bencw99/comparing-labeled-and-unlabeled-data`.

F.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 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

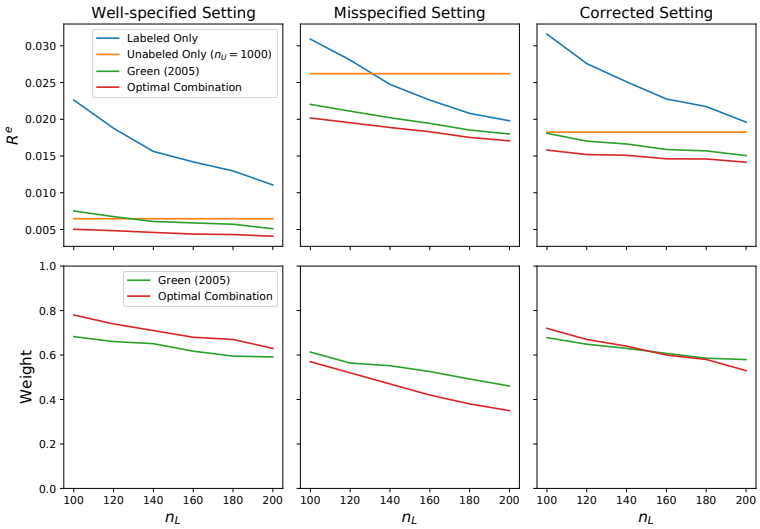

Figure 7: Excess generalization error and associated combination weight α for an optimally weighted combination of labeled and unlabeled estimators, and a combination weighted according to Green et al. (2005) across the well-specified (left), misspecified (center), and corrected (right) settings. The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

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. Therefore, 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 C.2, 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.

F.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 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

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.

and (2) we measure generalization error on the test set, since we cannot compute the expected generalization error directly.