First, we provide an extended discussion of related work. Next, we provide a glossary of terms and notation that we use throughout this paper for easy summary. Next, we discuss additional algorithmic details, and we give the proofs of our main results (each theorem). Finally, we give additional experimental details.

A. Extended Related Work

The notion of the "triplet" of (conditionally) independent variables as the source of minimal signal in latent variable models was observed and exploited in two innovative works, both using moments to deal with the challenge of the latent variable. These are

- Joglekar et al. (2013), in the explicit context of crowdsourcing, and

- Chaganty & Liang (2014), for estimating the parameters of certain latent variable graphical models.

The "3-Differences Scheme" described in 3.1 of Joglekar et al. (2013) is equivalent to our approach in Algorithm 1 in the basic case where there are no abstains and the signs of the accuracies are non-negative. Joglekar et al. (2013) focuses on crowdsourcing, and thus offers two contributions for this setting: (i) computing confidence intervals for worker accuracies and (ii) a set of techniques for extending the three-voters case by collapsing multiple voters into a pair 'super-voters' in order to build a better triplet for a particular worker. Both of these are useful directions for extensions of our work. In contrast, our approach focuses on efficiently handling the non-binary abstains case critical for weak supervision and develops theoretical characterizations for the downstream model behavior when using our generated labels.

A more general approach to learning latent variable graphical models is described in Chaganty & Liang (2014). Here there is an explicit description of the "three-views" approach. It is shown how to estimate the canonical parameters of a remarkably wide class of graphical models (e.g., both directed and undirected) by applying the tensor decomposition idea (developed in Anandkumar et al. (2014)) to recover conditional parameters. By comparison, our work is more specialized, looking at undirected (in fact, specifically Ising) models in the context of weak supervision. The benefits of this specialization are that we can replace the use of the tensor power iteration technique with a non-iterative closed-form solution, even for non-binary variables. Nevertheless, the techniques in Chaganty & Liang (2014) can be useful for weak supervision as well, and their pseudolikelihood approach to recover canonical parameters suggests that forward methods of inference could be used in our label model. We also note that closed-form triplet methods can be used to estimate *part* of the parameters of a more complex exponential family model (where some variables are involved in pairwise interactions at most, others in more complex patterns), so that resorting to tensor power iterations can be minimized.

A further work that builds on the approach of Chaganty & Liang (2014) is Raghunathan et al. (2016), where moments are used in combination with a linear technique. However, the setting here is different from weak supervision. The authors of Raghunathan et al. (2016) study *indirect* supervision. Here, for any unlabeled data point x, the label y is not seen, but a variable o is observed. So far this framework resembles weak supervision, but in the indirect setting, the supervision distribution $S(o|y)$ is known—while for weak supervision, it is not. Instead, in Chaganty & Liang (2014), the S distribution is given for two particular applications: local privacy and a light-weight annotation scheme for POS tagging.

B. Glossary

The glossary is given in Table 1 below.

C. Further Algorithmic Details

In this section, we present more details on the main algorithm, extensions to more complex models, and the online variant.

C.1. Core Algorithm

We first present the general binary Ising model and the proof of Proposition 1 that follows from this construction. We also prove another independence property over this general class of Ising models that can be used to factorize expectations over arbitrarily large cliques. Next, we detail the exact setup of the graphical model when sources can abstain, as well as the special case when they never abstain, and define the mappings necessary to convert between values over $\boldsymbol{v}, G$ and $\boldsymbol{\lambda}, G_{dep}$.

Symbol	Used for
$\boldsymbol{X}$	Unlabeled data vector, $\boldsymbol{X} = [X_1, X_2, \dots, X_D] \in \mathcal{X}$
$\boldsymbol{X}^i$	ith unlabeled data vector
X_i	ith data element
D	Length of the unlabeled data vector
$\boldsymbol{Y}$	Latent, ground-truth label vector, $\boldsymbol{Y} = [Y_1, Y_2, \dots, Y_D] \in \mathcal{Y}$, also referred to as hidden variables
$\boldsymbol{Y}^i$	ith ground-truth label vector
Y_i	Ground-truth label for ith task, $Y_i \in \{-1, +1\}$
$\mathcal{D}$	Distribution from which we assume $(\boldsymbol{X}, \boldsymbol{Y})$ data points are sampled i.i.d.
S_i	ith weak supervision source
m	Number of weak supervision sources
λ_i	Label of S_i for $\boldsymbol{X}$ where $\lambda_i \in \{-1, 0, 1\}$; all m labels per $\boldsymbol{X}$ collectively denoted $\boldsymbol{\lambda}$
n	Number of data vectors
$\widetilde{\boldsymbol{Y}}$	Probabilistic training labels for a label vector
f_w	Discriminative classifier used as end model, parametrized by w
G_{dep}	Source dependency graph
G	Augmented graph $G = (V, E)$ used for binary Ising model, where $V = \{\boldsymbol{Y}, \boldsymbol{v}\}$
$\boldsymbol{v}$	Observed variables of the graphical model corresponding to $\boldsymbol{\lambda}$
L	Label matrix containing n samples of source labels $\lambda_1, \dots, \lambda_m$
$\mathcal{L}$	Augmented label matrix computed from L
$Y^{dep}(i)$	Task that λ_i labels
$Y(i)$	Hidden variable that the observed variable v_i acts on
$\mathcal{C}_{dep}$	Cliqueset (maximal and non-maximal) of G_{dep}
$\tilde{\mathcal{C}}_{dep}, \mathcal{S}_{dep}$	The maximal cliques and separator sets of the junction tree over G_{dep}
$\boldsymbol{\mu}$	The label model parameters collectively over all μ_C, μ_S, the marginal distributions of $C \in \tilde{\mathcal{C}}_{dep}$, $S \in \mathcal{S}_{dep}$
$P(\bar{\boldsymbol{Y}})$	Class prior for the $\boldsymbol{Y}$ label vector
a_i	$\mathbb{E}\left[v_i Y(i)\right]$, the unobservable mean parameters of binary Ising model G
Ω_G	Set of vertices in V to which the triplet method can be applied
$\mathcal{C}$	Cliqueset (maximal and non-maximal) of G
a_C	The expectation over the product of observed variables in clique $C \in \mathcal{C}$ and $Y(C)$
$a_{C_{dep}}$	The expectation over the product of sources in clique $C_{dep} \in \mathcal{C}_{dep}$ and $Y^{dep}(C_{dep})$

Table 1. Glossary of variables and symbols used in this paper.

We then formalize the linear transformation from $a_{C_{dep}}$ to $\mu_{C_{dep}}$, and finally we explain the RESOLVESIGNS function used in Algorithm 1.

First, we give the explicit form of the density for the Ising model we use. Given the graph $G = (V, E)$, we can write the corresponding joint distribution of $\boldsymbol{Y}, \boldsymbol{v}$ as

$$f_G(\boldsymbol{Y}, \boldsymbol{v}) = \frac{1}{Z} \exp\Big(\sum_{k=1}^{D} \theta_{Y_k} Y_k + \sum_{(Y_k, Y_l) \in E} \theta_{Y_k, Y_l} Y_k Y_l + \sum_{v_i \in \boldsymbol{v}} \theta_i v_i Y(i) + \sum_{(v_k, v_l) \in E} \theta_{k,l} v_k v_l \Big), \tag{1}$$

where Z is the partition function, and the θ terms collectively are the canonical parameters of the model. Note that this is the most general definition of the binary Ising model with multiple dependent hidden variables and observed variables that we use.

C.1.1. PROOF OF PROPOSITION 1

We present the proof of Proposition 1, which is the underlying independence property of (1) that enables us to use the triplet method. We aim to show that for any $a, b \in \{-1, +1\}^2$,

$$P\big(v_i Y(i) = a, v_j Y(i) = b\big) = P(v_i Y(i) = a) \cdot P(v_j Y(i) = b), \tag{2}$$

where $v_i \perp\!\!\!\perp v_j | Y(i)$. For now, assume that $Y(j) \neq Y(i)$.

Because v_i and v_j are conditionally independent given $Y(i)$, we have that $P(v_i = a, v_j = b | Y(i) = 1) = P(v_i = a | Y(i) =$

$1) \cdot P(v_j = b | Y(i) = 1)$, and similarly for $v_i = -a, v_j = -b$ conditional on $Y(i) = -1$. Then

$$\begin{aligned}
&P(v_i = a, v_j = b, Y(i) = 1) \cdot P(Y = 1) = P(v_i = a, Y(i) = 1) \cdot P(v_j = b, Y(i) = 1) \\
&P(v_i = -a, v_j = -b, Y(i) = -1) \cdot P(Y = -1) = P(v_i = -a, Y(i) = -1) \cdot P(v_j = -b, Y(i) = -1).
\end{aligned} \tag{3}$$

Note that terms in (2) can be split depending on if $Y(i)$ is 1 or -1, so proving independence of $v_i Y(i)$ and $v_j Y(i)$ is equivalent to

$$\begin{aligned}
&P(v_i = a, v_j = b, Y(i) = 1) + P(v_i = -a, v_j = -b, Y(i) = -1) \\
&= (P(v_i = a, Y(i) = 1) + P(v_i = -a, Y(i) = -1)) \cdot (P(v_j = b, Y(i) = 1) + P(v_j = -b, Y(i) = -1)).
\end{aligned}$$

We substitute (3) into the right hand side. After rearranging, our equation to prove is

$$\begin{aligned}
&P(v_i = a, v_j = b, Y(i) = 1) \cdot P(Y(i) = -1) + P(v_i = -a, v_j = -b, Y(i) = -1) \cdot P(Y(i) = 1) \\
&= P(v_i = -a, Y(i) = -1) \cdot P(v_j = b, Y(i) = 1) + P(v_i = a, Y(i) = 1) \cdot P(v_j = -b, Y(i) = -1).
\end{aligned}$$

Due to symmetry of the terms above, it is thus sufficient to prove

$$P(v_i = a, v_j = b, Y(i) = 1) \cdot P(Y(i) = -1) = P(v_i = -a, Y(i) = -1) \cdot P(v_j = b, Y(i) = 1). \tag{4}$$

Let $N(v_i)$ be the set of v_i's neighbors in $\boldsymbol{v}$, and $N(Y_i)$ be the set of Y_i's neighbors in $\boldsymbol{Y}$. Let $\mathcal{S}$ be the event space for the hidden and observed variables, such that each element of the set $\mathcal{S}$ is a sequence of $+1$s and -1s of length equal to $|V|$. Denote $\mathcal{S}(v_i, v_j, Y(i))$ to be the event space for V besides v_i, v_j, and $Y(i)$; we also have similar definitions used for $\mathcal{S}(Y(i)), \mathcal{S}(v_i, Y(i)), \mathcal{S}(v_j, Y(i))$.

Our approach is to write each probability in (4) as a summation of joint probabilities over $\mathcal{S}(v_i, Y(i)), \mathcal{S}(v_j, Y(i))$, and $\mathcal{S}(v_i, v_j, Y(i))$ using (1). To do this more efficiently, we can factor each joint probability defined according to (1) into a product over *isolated variables* and a product over *non-isolated variables*. Recall that our marginal variables are v_i, v_j and $Y(i)$. Define the set of non-isolated variables to be the marginal variables, plus all variables that interact directly with the marginal variables according to the potentials in the binary Ising model. Per this definition, the non-isolated variables are $V_{NI} = \{v_i, v_j, Y(i), Y(j), N(Y(i)), N(v_i), N(v_j), v_{Y(i)}\}$ where $v_{Y(i)} = \{v : Y(v) = Y(i)\}$ and the isolated variables are all other variables not in this set, $V_I = V \backslash V_{NI}$. We can thus factorize each probability into a term $\psi(\cdot)$ corresponding to factors of the binary Ising model that only have isolated variables and a term $\zeta(\cdot)$ coresponding to factors that have non-isolated variables.

$$\begin{aligned}
P(v_i = a, v_j = b, Y(i) = 1) &= \frac{1}{Z} \sum_{s^{(a,b)} \in \mathcal{S}(v_i, v_j, Y(i))} \psi(s^{(a,b)}) \cdot \zeta(v_i = a, v_j = b, Y(i) = 1, s^{(a,b)}) \\
P(Y(i) = -1) &= \frac{1}{Z} \sum_{s^{(Y)} \in \mathcal{S}(Y(i))} \psi(s^{(Y)}) \cdot \zeta(Y(i) = -1, s^{(Y)}) \\
P(v_i = -a, Y(i) = -1) &= \frac{1}{Z} \sum_{s^{(a)} \in \mathcal{S}(v_i, Y(i))} \psi(s^{(a)}) \cdot \zeta(v_i = -a, Y(i) = -1, s^{(a)}) \\
P(v_j = b, Y(i) = 1) &= \frac{1}{Z} \sum_{s^{(b)} \in \mathcal{S}(v_j, Y(i))} \psi(s^{(b)}) \cdot \zeta(v_j = b, Y(i) = 1, s^{(b)})
\end{aligned}$$

To be precise, $\psi(\cdot)$ is

$$\psi(s^{(a,b)}) = \exp\Big(\sum_{\substack{Y_k \notin N(Y(i)) \\ \cup Y(i) \cup Y(j)}} \theta_{Y_k} Y_k^{(a,b)} + \sum_{\substack{Y_k, Y_l \notin \\ N(Y(i)) \cup Y(i) \cup Y(j)}} \theta_{Y_k, Y_l} Y_k^{(a,b)} Y_l^{(a,b)} + \sum_{\substack{Y(k) \notin N(Y(i)) \cup Y(i) \cup Y(j), \\ k \notin N(v_j) \cup v_j}} \theta_k v_k^{(a,b)} Y(k)^{(a,b)} + \sum_{\substack{v_k, v_l \notin N(v_i) \cup v_i \\ \cup N(v_j) \cup v_j}} \theta_{l,k} v_k^{(a,b)} v_l^{(a,b)} \Big),$$

where $s^{(a,b)} = \{Y_1^{(a,b)}, \dots, Y_D^{(a,b)}, v_1^{(a,b)}, \dots\}$, and similar definitions hold for $s^{(a)}$, $s^{(b)}$, and $s^{(Y)}$. Then, (4) is equivalent to showing

$$\begin{aligned}
&\sum_{s^{(a,b)}, s^{(Y)}} \psi(s^{(a,b)}) \cdot \psi(s^{(Y)}) \cdot \zeta(v_i = a, v_j = b, Y(i) = 1, s^{(a,b)}) \cdot \zeta(Y(i) = -1, s^{(Y)}) \\
= &\sum_{s^{(a)}, s^{(b)}} \psi(s^{(a)}) \cdot \psi(s^{(b)}) \cdot \zeta(v_i = -a, Y(i) = -1, s^{(a)}) \cdot \zeta(v_j = b, Y(i) = 1, s^{(b)}).
\end{aligned}$$

We can show this by finding values of $s^{(a)}$ and $s^{(b)}$ that correspond to each $s^{(a,b)}$ and $s^{(Y)}$. Note that the ψ terms will cancel each other out if we directly set $s^{(a)}[V_I] = s^{(Y)}[V_I]$ and $s^{(b)}[V_I] = s^{(a,b)}[V_I]$. Therefore, we want to set $s^{(a)}[V_{NI}]$ and $s^{(b)}[V_{NI}]$ such that the products of ζs are equivalent. We write them out explicitly first:

$$\begin{aligned}
\zeta(v_i = a, v_j = b, Y(i) = 1, s^{(a,b)}) = \exp\Big(&\theta_{Y(i)} + \sum_{Y_k \in N(Y(i)) \cup Y(j)} \theta_{Y_k} Y_k^{(a,b)} + \sum_{Y_k \in N(Y(i))} \theta_{Y_k, Y(i)} Y_k^{(a,b)} + \sum_{\substack{Y_k \in N(Y(i)) \cup Y(j), \\ Y_l \notin N(Y(i)) \cup Y(i) \cup Y(j)}} \theta_{Y_k, Y_l} Y_k^{(a,b)} Y_l^{(a,b)} \\
&+ \theta_i a + \theta_j b Y(j)^{(a,b)} + \sum_{\substack{k \neq i,j, \\ Y(k) = Y(i)}} \theta_k v_k^{(a,b)} + \sum_{\substack{k \neq i,j, \\ Y(k) \in N(Y(i)) \cup Y(j) \\ | k \in N(v_j)}} \theta_k v_k^{(a,b)} Y(k)^{(a,b)} + \sum_{v_k \in N(v_i)} \theta_{i,k} a v_k^{(a,b)} \\
&+ \sum_{v_k \in N(v_j)} \theta_{j,k} b v_k^{(a,b)}\Big)
\end{aligned}$$

$$\begin{aligned}
\zeta(Y(i) = -1, s^{(Y)}) = \exp\Big(&-\theta_{Y(i)} + \sum_{Y_k \in N(Y(i)) \cup Y(j)} \theta_{Y_k} Y_k^{(Y)} - \sum_{Y_k \in N(Y(i))} \theta_{Y_k, Y(i)} Y_k^{(Y)} + \sum_{\substack{Y_k \in N(Y(i)) \cup Y(j), \\ Y_l \notin N(Y(i)) \cup Y(i) \cup Y(j)}} \theta_{Y_k, Y_l} Y_k^{(Y)} Y_l^{(Y)} \\
&- \theta_i v_i^{(Y)} + \theta_j v_j^{(Y)} Y(j)^{(Y)} - \sum_{\substack{k \neq i,j, \\ Y(k) = Y(i)}} \theta_k v_k^{(Y)} + \sum_{\substack{k \neq i,j, \\ Y(k) \in N(Y(i)) \cup Y(j) \\ | k \in N(v_j)}} \theta_k v_k^{(Y)} Y(k)^{(Y)} + \sum_{\substack{v_k \in N(v_i), \\ v_l \neq v_i}} \theta_{k,l} v_k^{(Y)} v_l^{(Y)} \\
&+ \sum_{\substack{v_k \in N(v_j), \\ v_l \neq v_j}} \theta_{k,l} v_k^{(Y)} v_l^{(Y)}\Big)
\end{aligned}$$

$$\begin{aligned}
\zeta(v_i = -a, Y(i) = -1, s^{(a)}) = \exp\Big(&-\theta_{Y(i)} + \sum_{Y_k \in N(Y(i)) \cup Y(j)} \theta_{Y_k} Y_k^{(a)} - \sum_{Y_k \in N(Y(i))} \theta_{Y_k, Y(i)} Y_k^{(a)} + \sum_{\substack{Y_k \in N(Y(i)) \cup Y(j), \\ Y_l \notin N(Y(i)) \cup Y(i) \cup Y(j)}} \theta_{Y_k, Y_l} Y_k^{(a)} Y_l^{(a)} \\
&+ \theta_i a + \theta_j v_j^{(a)} Y(j)^{(a)} - \sum_{\substack{k \neq i,j, \\ Y(k) = Y(i)}} \theta_k v_k^{(a)} + \sum_{\substack{k \neq i,j, \\ Y(k) \in N(Y(i)) \cup Y(j) \\ | k \in N(v_j)}} \theta v_k^{(a)} Y(k)^{(a)} + \sum_{\substack{v_k \in N(v_j), \\ v_l \neq v_j}} \theta_{k,l} v_k^{(a)} v_l^{(a)} \\
&- \sum_{v_k \in N(v_i)} \theta_{i,k} a v_k^{(a)}\Big)
\end{aligned}$$

$$\begin{aligned}
\zeta(v_j = b, Y(i) = 1, s^{(b)}) = \exp\Big(&\theta_{Y(i)} + \sum_{Y_k \in N(Y(i)) \cup Y(j)} \theta_{Y_k} Y_k^{(b)} + \sum_{Y_k \in N(Y(i))} \theta_{Y_k, Y(i)} Y_k^{(b)} + \sum_{\substack{Y_k \in N(Y(i)) \cup Y(j), \\ Y_l \notin N(Y(i)) \cup Y(i) \cup Y(j)}} \theta_{Y_k, Y_l} Y_k^{(b)} Y_l^{(b)} \\
&+ \theta_i v_i^{(b)} + \theta_j b Y(j)^{(b)} + \sum_{\substack{k \neq i,j, \\ Y(k) = Y(i)}} \theta_k v_k^{(b)} + \sum_{\substack{k \neq i,j, \\ Y(k) \in N(Y(i)) \cup Y(j) \\ | k \in N(v_j)}} \theta_k v_k^{(b)} Y(k)^{(b)} + \sum_{\substack{v_k \in N(v_i), \\ v_l \neq v_i}} \theta_{k,l} v_k^{(b)} v_l^{(b)} \\
&+ \sum_{v_k \in N(v_j)} \theta_{j,k} b v_k^{(b)}\Big)
\end{aligned}$$

We present a simple mapping from $s^{(a,b)}$ and $s^{(Y)}$ to $s^{(a)}$ and $s^{(b)}$ such that $\zeta(v_i = a, v_j = b, Y(i) = 1, s^{(a,b)}) \cdot \zeta(Y(i) = -1, s^{(Y)}) = \zeta(v_i = -a, Y(i) = -1, s^{(a)}) \cdot \zeta(v_j = b, Y(i) = 1, s^{(b)})$ holds:

	$s^{(a)}$	$s^{(b)}$
v_i	$-$	$-v_i^{(Y)}$
v_j	$v_j^{(Y)}$	$-$
$Y_k \in N(Y(i)) \cup Y(j)$	$Y_k^{(Y)}$	$Y_k^{(a,b)}$
$v_k \in N(v_i)$	$-v_k^{(a,b)}$	$-v_k^{(Y)}$
$v_k \in N(v_j)$	$v_k^{(Y)}$	$v_k^{(a,b)}$
$v_{Y(i)}$	$-v_k^{(a,b)}$	$-v_k^{(Y)}$

With this construction of $s^{(a)}$ and $s^{(b)}$, we have shown that $v_i Y(i)$ and $v_j Y(i)$ are independent. (In the case that $Y(j) = Y(i)$, the proof is almost exactly the same).

C.1.2. Handling Larger Cliques

We discuss how arbitrarily large cliques can be factorized into mean parameters and observable statistics to compute values of a_C in Algorithm 2. This is due to the following general independence property that arises from construction of the Ising model in (1):

Proposition 1. *For a clique C of v_k's all connected to a single $Y(C)$, we have that $\prod_{k \in C} v_k \perp\!\!\!\perp Y(C)$ if $|C|$ is even, and $\prod_{k \in C} v_k Y(C) \perp\!\!\!\perp Y(C)$ if $|C|$ is odd.*

Therefore, if $|C|$ is even, then $a_C = \mathbb{E}\left[\prod_{k \in C} v_k\right] \cdot \mathbb{E}\left[Y(C)\right]$. If $|C|$ is odd, then $a_C = \mathbb{E}\left[\prod_{k \in C} v_k\right] / \mathbb{E}\left[Y(C)\right]$.

Proof. We assume that there is only one hidden variable Y, although generalizing to the case where $D > 1$ is straightforward because our proposed independence property only acts on the hidden variable associated with a clique of observed variables.

We first prove the case where $|C|$ is even. We aim to show that for any $a, b \in \{-1, +1\}^2$,

$$P\Big(\prod_{k \in C} v_k = a, Y = b\Big) = P\Big(\prod_{k \in C} v_k = a\Big) P(Y = b).$$

Using the concept of isolated variables and non-isolated variables earlier, the set of all observed variables V_I besides those in C and their neighbors can be ignored. Furthermore, suppose that $\mathcal{S}^{(C,a)}$ is the set of all $k \in C$ such that $\prod_{k \in C} v_k = a$. For example, if $C = \{i, j\}$ and $a = -1$, $\mathcal{S}^{(C,-1)} = \{(v_i, v_j) = (1, -1), (-1, 1)\}$. We write out each of the above probabilities as well as the partition function Z:

$$P\Big(\prod_{i \in C} v_i = a, Y = b\Big) = \frac{1}{Z} \sum_{s^{(a,b)} \in \mathcal{S}(C,Y)} \psi\big(s^{(a,b)}\big) \sum_{s^{(C_1,a)} \in \mathcal{S}^{(C)}} \exp\Big(\theta_Y b + \sum_{i \in C} \theta_i b s_{v_i}^{(C_1)} + \sum_{i \notin C} \theta_i b v_i^{(a,b)}$$
$$+ \sum_{(i,j) \in C} \theta_{i,j} s_{v_i}^{(C_1)} s_{v_j}^{(C_1)} + \sum_{i \in C} \sum_{j \in N(v_i) \setminus v_C} \theta_{i,j} s_{v_i}^{(C_1)} v_j^{(a,b)}\Big)$$

$$P\Big(\prod_{i \in C} v_i = a\Big) = \frac{1}{Z} \sum_{s^{(a)} \in \mathcal{S}(C)} \psi\big(s^{(a)}\big) \sum_{s^{(C_2,a)} \in \mathcal{S}^{(C)}} \exp\Big(\theta_Y Y^{(a)} + \sum_{i \in C} \theta_i s_{v_i}^{(C_1)} Y^{(a)} + \sum_{i \notin C} \theta_i v_i^{(a)} Y^{(a)}$$
$$+ \sum_{(i,j) \in C} \theta_{i,j} s_{v_i}^{(C_2)} s_{v_j}^{(C_2)} + \sum_{i \in C} \sum_{j \in N(v_i) \setminus v_C} \theta_{i,j} s_{v_i}^{(C_1)} v_j^{(a)}\Big)$$

$$
\begin{aligned}
P(Y=b) = \sum_{s^{(b)} \in \mathcal{S}(Y)} \psi\big(s^{(b)}\big) \exp\Big(&\theta_Y b + \sum_{i \in C} \theta_i b v_i^{(b)} + \sum_{i \notin C} \theta_i v_i^{(b)} Y^{(b)} \\
&+ \sum_{(i,j) \in C} \theta_{i,j} v_i^{(b)} v_j^{(b)} + \sum_{i \in C} \sum_{j \in N(v_i) \setminus v_C} \theta_{i,j} v_i^{(b)} v_j^{(b)} \Big) \\
Z = \sum_{s^{(z)} \in \mathcal{S}} \psi\big(s^{(z)}\big) \exp\Big(&\theta_Y Y^{(z)} + \sum_{i \in C} \theta_i v_i^{(z)} Y^{(z)} + \sum_{i \notin C} \theta_i v_i^{(z)} Y^{(z)} + \sum_{(i,j) \in C} \theta_{i,j} v_i^{(z)} v_j^{(z)} \\
&+ \sum_{i \in C} \sum_{j \in N(v_i) \setminus v_C} \theta_{i,j} v_i^{(z)} v_j^{(z)} \Big)
\end{aligned}
$$

We want to show that we can map from each $s^{(a,b)}$, $s^{(z)}$ and $s^{(C_1)}$ to a respective $s^{(a)}$, $s^{(b)}$, and $s^{(C_2)}$. The $\psi(\cdot)$ terms can be ignored since we can just directly set $s^{(a)}[V_I] = s^{(a,b)}[V_I]$ and $s^{(b)}[V_I] = s^{(z)}[V_I]$. Using the above expressions for probabilities and the cumulant function, our desired statement to prove for each $s^{(a,b)}$, $s^{(z)}$ and $s^{(C_1)}$ is

$$
\begin{aligned}
&\exp\Big(\theta_Y\big(b + Y^{(z)}\big) + \sum_{i \in C} \theta_i \big(b s_{v_i}^{(C_1)} + v_i^{(z)} Y^{(z)}\big) + \sum_{i \notin C} \theta_i \big(b v_i^{(a,b)} + v_i^{(z)} Y^{(z)}\big) \\
&\quad + \sum_{(i,j) \in C} \theta_{i,j} \big(s_{v_i}^{(C_1)} s_{v_j}^{(C_1)} + v_i^{(z)} v_j^{(z)}\big) + \sum_{i \in C} \sum_{j \in N(v_i) \setminus v_C} \theta_{i,j} \big(s_{v_i}^{(C_1)} v_k^{(a,b)} + v_i^{(z)} v_k^{(z)}\big)\Big) \\
= \; &\exp\Big(\theta_Y\big(b + Y^{(a)}\big) + \sum_{i \in C} \theta_i \big(s_{v_i}^{(C_2)} Y^{(a)} + b v_i^{(b)}\big) + \sum_{i \notin C} \theta_i \big(v_i^{(a)} Y^{(a)} + b v_i^{(b)}\big) \\
&\quad + \sum_{(i,j) \in C} \theta_{i,j} \big(s_{v_i}^{(C_2)} s_{v_j}^{(C_2)} + v_i^{(b)} v_j^{(b)}\big) + \sum_{i \in C} \sum_{j \in N(v_i) \setminus v_C} \theta_{i,j} \big(s_{v_i}^{(C_2)} v_j^{(a)} + v_i^{(b)} v_j^{(b)}\big)\Big)
\end{aligned} \tag{5}
$$

We can ensure that the above expression is satisfied with the following relationship between $s^{(a,b)}$, $s^{(z)}, s^{(C_1)}$ and $s^{(a)}, s^{(b)}, s^{(C_2)}$. If $Y^{(z)} = b$, then we set $Y^{(a)} = b$, $s_{v_i}^{(C_2)} = s_{v_i}^{(C_1)}$ for $i \in C$, and $v_i^{(b)} = v_i^{(z)}, v_i^{(a)} = v_i^{(a,b)}$ for all v_i. If $Y^{(z)} = -b$, then we set $Y^{(a)} = -b$, $s_{v_i}^{(C_2)} = -s_{v_i}^{(C_1)}$ for $i \in C$, and $v_i^{(b)} = -v_i^{(z)}, v_i^{(a)} = -v_i^{(a,b)}$ for all v_i. However, note that setting either all $s_{v_i}^{(C_2)}$ to be $s_{v_i}^{(C_1)}$ or $-s_{v_i}^{(C_1)}$ means that both $s^{(C_1)}$ and $-s^{(C_1)}$ are in $\mathcal{S}^{(C)}$. This is only true when $|C|$ is even because $\prod_{i \in C}(-v_i) = (-1)^{|C|} \prod_{i \in C} v_i = (-1)^{|C|} a$.

Our proof approach is similar when $|C|$ is odd. We aim to show that for any $a, b \in \{-1, +1\}^2$,

$$
P\Big(\prod_{k \in C} v_k Y = a, Y = b\Big) = P\Big(\prod_{k \in C} v_k Y = a\Big) P(Y = b).
$$

$P(\prod_{k \in C} v_k Y = a, Y = b)$ can be written as $P(\prod_{k \in C} v_k = \frac{a}{b}, Y = b)$, which follows the same format of the probability we used for the case where $|C|$ is even. We will end up with a desired equation to prove that is identical to (5), except that we must modify $s^{(C_1)}$ and $s^{(C_2)}$. $s^{(C_1)}$ is now from the set $\mathcal{S}^{(C, a/b)}$, and $s^{(C_2)}$ is from the set $\mathcal{S}^{(C, a/b)}$ when $Y^{(a)} = b$ and from the set $s^{(C, -a/b)}$ when $Y^{(a)} = -b$. We can set $s^{(a)}$, $s^{(b)}$, and $s^{(C_2)}$ the exact same way as before; in particular, $s_{v_i}^{(C_2)} = s_{v_i}^{(C_1)}$ when $Y^{(a)} = b$ and $s_{v_i}^{(C_2)} = -s_{v_i}^{(C_1)}$ when $Y^{(a)} = -b$. Both $s_{v_i}^{(C_1)}, Y^{(a)} = b$ and $-s_{v_i}^{(C_1)}, Y^{(a)} = -b$ satisfy $\prod_{i \in C} v_i Y = a$, since $\prod_{i \in C}(-v_i)(-Y) = (-1)^{|C|+1} \prod_{i \in C} v_i Y = a$ when $|C|$ is odd. □

C.1.3. Augmenting the dependency graph

We define the graphical model particular to how G_{dep} is augmented, which gives way to a concise mapping between each a_C and $a_{C_{dep}}$.

In the case where no sources can abstain at all, λ_i takes on values $\{\pm 1\}$ and thus the augmentation is not necessary. We have that $G = G_{dep}$, $\boldsymbol{v} = \boldsymbol{\lambda}$, and the graphical model's joint distribution (1) reduces to

$$
f_G(Y, \boldsymbol{\lambda}) = \frac{1}{Z} \exp\Big(\sum_{k=1}^{D} \theta_{Y_k} Y_k + \sum_{(Y_k, Y_l) \in E} \theta_{Y_k, Y_l} Y_k Y_l + \sum_{i=1}^{m} \theta_i \lambda_i Y(i) + \sum_{(\lambda_i, \lambda_j) \in E} \theta_{i,j} \lambda_i \lambda_j\Big). \tag{6}
$$

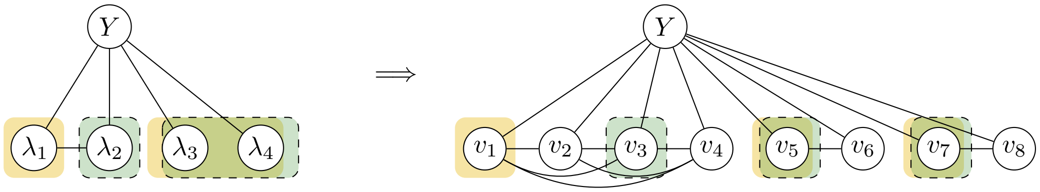

Figure 1. Example of mapping from G_{dep} to G. Left: G_{dep}, where boxes indicate valid triplet groupings of sources. Right: G, where boxes indicate the triplets of observed variables that are sufficient to recover all mean parameters.

All of Algorithm 2 will be done on $\{\boldsymbol{Y}, \boldsymbol{\lambda}\}$. While the triplet method is still used for recovering mean parameters, the mapping from a_C to $a_{C_{dep}}$ is trivial, and the linear transformation back to $\mu_{C_{dep}}$ will have terms containing $\lambda_i = 0$ reduced to 0.

In the case where sources abstain, we have discussed how to generate $\boldsymbol{v}$ from $\boldsymbol{\lambda}$ and G from G_{dep}, of which an example is shown in Figure 1. Most importantly, we suppose that when $\lambda_i = 0$, we set (v_{2i-1}, v_{2i}) to either $(1,1)$ or $(-1,-1)$ with equal probability such that

$$P\big((v_{2i-1}, v_{2i}) = (1,1), V \backslash \{v_{2i-1}, v_{2i}\}\big) = P\big((v_{2i-1}, v_{2i}) = (-1,-1), V \backslash \{v_{2i-1}, v_{2i}\}\big) = \frac{1}{2} P(\lambda_i = 0, V \backslash \{v_{2i-1}, v_{2i}\}). \tag{7}$$

The joint distribution over $\{\boldsymbol{Y}, \boldsymbol{v}\}$ follows from (1):

$$\begin{aligned} f_G(\boldsymbol{Y}, \boldsymbol{v}) = \frac{1}{Z} \exp\Bigg(& \sum_{k=1}^{D} \theta_{Y_k} Y_k + \sum_{(Y_k, Y_l) \in E} \theta_{Y_k, Y_l} Y_k Y_l + \sum_{i=1}^{m} \theta_i \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} v_{2i-1} \\ v_{2i} \end{bmatrix} Y^{dep}(i) \\ & + \sum_{i=1}^{m} \theta_{i,i} v_{2i-1} v_{2i} + \sum_{i,j : (\lambda_i, \lambda_j) \in E_{dep}} \theta_{i,j} \begin{bmatrix} v_{2i-1} & v_{2i} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} v_{2j-1} \\ v_{2j} \end{bmatrix} \Bigg), \end{aligned} \tag{8}$$

where E_{dep} is G_{dep}'s edge set. Note that this graphical model has the same absolute values of the canonical parameters for both $v_{2i-1} Y^{dep}(i)$ and for all four terms $(v_{2i-1}, v_{2i}) \times (v_{2j-1}, v_{2j})$ due to the balancing in (7). As a result, the mean parameters also exhibit the same symmetry, which we show in the following lemma.

Lemma 1. *For each λ_i, we have that $\mathbb{E}\left[\lambda_i Y^{dep}(i)\right] = \mathbb{E}\left[v_{2i-1} Y^{dep}(i)\right] = -\mathbb{E}\left[v_{2i} Y^{dep}(i)\right]$.*

Proof. First, we can write out $\mathbb{E}\left[\lambda_i Y^{dep}(i)\right]$ as

$$\begin{aligned} \mathbb{E}\left[\lambda_i Y^{dep}(i)\right] &= P(\lambda_i Y^{dep}(i) = 1) - P(\lambda_i Y^{dep}(i) = -1) = P(\lambda_i Y^{dep}(i) = 1) \\ &- (1 - P(\lambda_i Y^{dep}(i) = 1) - P(\lambda_i Y^{dep}(i) = 0)) \\ &= 2P(\lambda_i Y^{dep}(i) = 1) + P(\lambda_i = 0) - 1. \end{aligned}$$

We know that if we have $v_{2i-1} = 1$ or $v_{2i} = -1$, then λ_i is either 1 or 0, but never -1; similarly, $v_{2i-1} = -1$ and $v_{2i} = 1$ imply that $\lambda_i \neq 1$. We write out $\mathbb{E}\left[v_{2i-1} Y^{dep}(i)\right]$:

$$\begin{aligned} \mathbb{E}\left[v_{2i-1} Y^{dep}(i)\right] &= 2\left(P(v_{2i-1} = 1, Y^{dep}(i) = 1) + P(v_{2i-1} = -1, Y^{dep}(i) = -1)\right) - 1 \\ &= 2\big(P((v_{2i-1}, v_{2i}) = (1,1), Y^{dep}(i) = 1) + P(\lambda_i = 1, Y^{dep}(i) = 1) \\ &+ P(\lambda_i = -1, Y^{dep}(i) = -1) + P((v_{2i-1}, v_{2i}) = (-1,-1), Y^{dep}(i) = -1)\big) - 1 \\ &= 2\Big(P(\lambda_i Y^{dep}(i) = 1) + \frac{1}{2} P(\lambda_i = 0, Y^{dep}(i) = 1) + \frac{1}{2} P(\lambda_i = 0, Y^{dep}(i) = -1)\Big) - 1 \\ &= 2P(\lambda_i Y^{dep}(i) = 1) + P(\lambda_i = 0) - 1 = \mathbb{E}\left[\lambda_i Y^{dep}(i)\right]. \end{aligned}$$

Similarly, $\mathbb{E}\left[v_{2i}Y^{dep}(i)\right]$ is

$$\begin{aligned}\mathbb{E}\left[v_{2i}Y^{dep}(i)\right] &= 2\big(P((v_{2i-1}, v_{2i}) = (1,1), Y^{dep}(i) = 1) + P(\lambda_i = -1, Y^{dep}(i) = 1) \\ &\quad + P(\lambda_i = 1, Y^{dep}(i) = -1) + P((v_{2i-1}, v_{2i}) = (-1,-1), Y^{dep}(i) = -1)\big) - 1 \\ &= 2\Big(P(\lambda_i Y^{dep}(i) = -1) + \frac{1}{2}P(\lambda_i = 0, Y^{dep}(i) = 1) + \frac{1}{2}P(\lambda_i = 0, Y^{dep}(i) = -1)\Big) - 1 \\ &= 2P(\lambda_i Y^{dep}(i) = -1) + P(\lambda_i = 0) - 1 \\ &= P(\lambda_i Y^{dep}(i) = -1) - (1 - P(\lambda_i = 0) - P(\lambda_i Y^{dep}(i) = -1)) \\ &= P(\lambda_i Y^{dep}(i) = -1) - P(\lambda_i Y^{dep}(i) = 1) = -\mathbb{E}\left[\lambda_i Y^{dep}(i)\right].\end{aligned}$$

□

The triplets in Algorithm 1 thus only need to be computed over exactly half of $\boldsymbol{v}$, each corresponding to one source, as shown in Figure 1. Moreover, this augmentation method for $\boldsymbol{v}$ and G allows us to conclude for any clique of sources $C_{dep} \in \mathcal{C}_{dep}$,

$$\mathbb{E}\bigg[\prod_{k \in C_{dep}} v_{2k-1} Y^{dep}(C_{dep})\bigg] = \mathbb{E}\bigg[\prod_{k \in C_{dep}} \lambda_k Y^{dep}(C_{dep})\bigg].$$

In general, the expectation over a clique in G_{dep} containing $\{\lambda_i\}_{i \in C_{dep}}$ is equal to the expectation over the corresponding clique C in G containing $\{v_{2i-1}\}_{i \in C_{dep}}$ such that $a_C = a_{C_{dep}}$.

C.1.4. Linear Transformation to Label Model Parameters

To convert these $a_{C_{dep}}$ into $\mu_{C_{dep}}$, we present a way to linearly map from these product probabilities and expectations back to marginal distributions, focusing on the unobservable distributions over a clique of sources and a task that the sources vote on. We first restate our example stated in Section 3.2. Define $\mu_i(a, b) = P(Y^{dep}(i) = a, \lambda_i = b)$ for $a \in \{-1, 1\}$ and $b \in \{-1, 0, 1\}$. We can set up a series of linear equations and denote it as $A_1\mu_i = r_i$:

$$\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0\end{bmatrix} \begin{bmatrix}\mu_i(1,1) \\ \mu_i(-1,1) \\ \mu_i(1,0) \\ \mu_i(-1,0) \\ \mu_i(1,-1) \\ \mu_i(-1,-1)\end{bmatrix} = \begin{bmatrix}1 \\ P(Y^{dep}(i) = 1) \\ P(\lambda_i = 1) \\ P(\lambda_i Y^{dep}(i) = 1) \\ P(\lambda_i = 0) \\ P(\lambda_i = 0, Y^{dep}(i) = 1)\end{bmatrix}. \tag{9}$$

Note that four entries on the right of the equation are observable or known. $P(\lambda_i Y^{dep}(i) = 1)$ can be written in terms of a_i, and by construction of (v_{2i-1}, v_{2i}) and (7), we can factorize $P(\lambda_i = 0, Y^{dep}(i) = 1)$ into observable terms:

$$\begin{aligned}P(\lambda_i = 0, Y^{dep}(i) = 1) &= P((v_{2i-1}, v_{2i}) = (1,1), Y^{dep}(i) = 1) + P((v_{2i-1}, v_{2i}) = (-1,-1), Y^{dep}(i) = 1) \\ &= (P((v_{2i-1}, v_{2i}) = (1,1)) + P((v_{2i-1}, v_{2i}) = (-1,-1)))P(Y^{dep}(i) = 1) \\ &= P(\lambda_i = 0)P(Y^{dep}(i) = 1).\end{aligned}$$

Here we use the fact that $v_{2i-1}v_{2i}$ and $Y^{dep}(i)$ are independent by Proposition 1. We can verify that A_1 is invertible, so $\mu_i(a, b)$ can be obtained from this system.

There is a way to extend this system to the general case. We form a system of linear equations $A_s\mu_C = r_C$ for each clique of sources C in G_{dep}, where $s = |C|$ is the number of weak sources λ_i in the clique and μ_C is the marginal distribution over these s sources and 1 task. A_s is a $2(3^s) \times 2(3^s)$ matrix of 0s and 1s that will help map from r_C, a vector of probabilities known from prior steps of the algorithms or from direct estimation, to the desired label model parameter μ_C. Define

$$A_0 = \begin{bmatrix}1 & 1 \\ 1 & 0\end{bmatrix} \qquad B_0 = \begin{bmatrix}0 & 0 \\ 0 & 1\end{bmatrix}$$

$$D = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad E = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Then A_s and B_s can be recursively constructed with

$$\begin{aligned} A_s &= D \otimes A_{s-1} + E \otimes B_{s-1} \\ B_s &= E \otimes A_{s-1} + D \otimes B_{s-1}, \end{aligned}$$

where $\otimes$ is the Kronecker product. To define r_C, we first specify an ordering of elements of μ_C. Let the last λ_{C_s} in the joint probability μ_C take on value $\lambda_{C_s} = 1$ for the first $2 \times 3^{s-1}$ entries, $\lambda_{C_s} = 0$ for the next $2 \times 3^{s-1}$ entries, and $\lambda_{C_s} = -1$ for the last $2 \times 3^{s-1}$ entries. In general, the ith λ_{C_i} in μ_C will alternate among $1, 0, -1$ every $2 \times 3^{i-1}$ entries. Finally, the $Y(i)$ entry of μ_C alternates every other value between 1 and -1.

The ordering of r_C follows a similar structure. If we rename the Y and λ variables to $z_1, \ldots, z_{s+1}$ for generality, each entry $r_C(U, Z)$ is equal to $P(\prod_{z_i \in Z} z_i = 1, z_j = 0 \; \forall z_j \in U)$, where $U \cap Z = \emptyset$, and $U \subseteq C \backslash Y(i)$, $Z \subseteq C$. We also write $r_C(\emptyset, \emptyset) = 1$. The entries of r_C will alternate similarly to μ_C, for each λ_{C_i}, the first $2 \times 3^{i-1}$ terms will not contain λ_{C_i} in either U or Z, the second $2 \times 3^{i-1}$ terms will have $\lambda_{C_i} \in Z$, and the last $2 \times 3^{i-1}$ terms will have $\lambda_{C_i} \in U$. For $Y(i)$, elements of r_C will alternate every other value between not having $Y(i)$ in Z and having $Y(i)$ in Z. (9) illustrates an example of the orderings for μ_C and r_C.

Furthermore, we also have the system $B_s \mu_C = r_C^B$, where $r_C^B(U, Z) = P(\prod_{z_i \in Z} z_i = -1, z_j = 0 \; \forall z_j \in U)$ when $Z \neq \emptyset$, and $r_B^C(U, \emptyset) = 0$. The ordering of r_C^B is the same as that of r_C.

Lemma 2. *With the setup above, $A_s \mu_C = r_C$.*

Proof. We prove that $A_s \mu_C = r_C$ and $B_s \mu_C = r_C^B$ by induction on s. For the base case $s = 0$, we examine a clique over just a single Y:

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} P(Y = 1) \\ P(Y = -1) \end{bmatrix} = \begin{bmatrix} 1 \\ P(Y = 1) \end{bmatrix} \qquad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P(Y = 1) \\ P(Y = -1) \end{bmatrix} = \begin{bmatrix} 0 \\ P(Y = -1) \end{bmatrix},$$

which are both clearly true. Next, we assume that $A_k \mu_C = r_C$ and $B_k \mu_C = r_C^B$ for $s = k$. We want to show that $A_{k+1} \mu_{C'} = r_{C'}$ and $B_{k+1} \mu_{C'} = r_{C'}^B$ for a larger clique C' where $C \subset C'$ and $|C'| = s + 1$. By construction of A_{k+1} and B_{k+1},

$$A_{k+1} = \begin{bmatrix} A_k & A_k & A_k \\ A_k & 0 & B_k \\ 0 & A_k & 0 \end{bmatrix} \qquad B_{k+1} = \begin{bmatrix} B_k & B_k & B_k \\ B_k & 0 & A_k \\ 0 & B_k & 0 \end{bmatrix}.$$

$\mu_{C'}$, $r_{C'}$, and $r_{C'}^B$ can be written as

$$\mu_{C'} = \begin{bmatrix} \mu_C(\lambda_{C'_{k+1}} = 1)) \\ \mu_C(\lambda_{C'_{k+1}} = 0) \\ \mu_C(\lambda_{C'_{k+1}} = -1) \end{bmatrix} \qquad r_{C'} = \begin{bmatrix} r_C \\ r_C(\lambda_{C'_{k+1}} \in Z') \\ r_C(\lambda_{C'_{k+1}} \in U') \end{bmatrix} \qquad r_{C'}^B = \begin{bmatrix} r_C^B \\ r_C^B(\lambda_{C'_{k+1}} \in Z') \\ r_C^B(\lambda_{C'_{k+1}} \in U') \end{bmatrix},$$

where $\mu_C(\lambda_{C'_{k+1}} = 1) = P(Y, \lambda_{C_1}, \ldots, \lambda_{C_k}, \lambda_{C'_{k+1}} = 1)$, $r_C(\lambda_{C'_{k+1}} \in Z') = r_C(U, Z \cup \{\lambda_{C'_{k+1}}\})$, and so on. U', Z' for C' are constructed similarly to U, Z for C.

Then the three equations for A_k we want to show are

$$\begin{aligned} &A_k(\mu_C(\lambda_{C'_{k+1}} = 1) + \mu_C(\lambda_{C'_{k+1}} = 0) + \mu_C(\lambda_{C'_{k+1}} = -1)) = r_C \\ &A_k(\mu_C(\lambda_{C'_{k+1}} = 1)) + B_k(\mu_C(\lambda_{C'_{k+1}} = -1)) = r_C(\lambda_{C'_{k+1}} \in Z') \\ &A_k(\mu_C(\lambda_{C'_{k+1}} = 0)) = r_C(\lambda_{C'_{k+1}} \in U'). \end{aligned}$$

The first equation is true because $\lambda_{C'_{k+1}}$ is marginalized out to yield $A_k \mu_C = r_C$, which is true by our inductive hypothesis. In the third equation, the term $\lambda_{C'_{k+1}} = 0$ is added as a joint probability to all probabilities in μ_C and r_C, so this also holds by the inductive hypothesis. In the second equation, $A_k(\mu_C(\lambda_{C'_{k+1}} = 1))$ is equal to r_C with each probability having $\lambda_{C'_{k+1}} = 1$ as an additional joint probability, and similarly $B_k(\mu_C(\lambda_{C'_{k+1}} = -1))$ is equal to r_C^B with each nonzero probability having $\lambda_{C'_{k+1}} = -1$ as an additional joint probability. For entries where $Z \neq \emptyset$, summing these up yields

$$\begin{aligned} &P\Big(\prod_{z_i \in Z} z_i = 1, \lambda_{C'_{k+1}} = 1, z_j = 0 \; \forall z_j \in U\Big) + P\Big(\prod_{z_i \in Z} z_i = -1, \lambda_{C'_{k+1}} = -1, z_j = 0 \; \forall z_j \in U\Big) \\ =&P\Big(\prod_{z_i \in Z} z_i \lambda_{C'_{k+1}} = 1, z_j = 0 \; \forall z_j \in U\Big). \end{aligned}$$

And when $Z = \emptyset$, we have $P(\lambda_{C'_{k+1}} = 1, z_j = 0 \; \forall z_j \in U)$, so all together these probabilities make up $r_C(\lambda_{C'_{k+1}} \in Z')$.

The three equations for B_k are similar:

$$\begin{aligned} &B_k(\mu_C(\lambda_{C'_{k+1}} = 1) + \mu_C(\lambda_{C'_{k+1}} = 0) + \mu_C(\lambda_{C'_{k+1}} = -1)) = r_C^B \\ &B_k(\mu_C(\lambda_{C'_{k+1}} = 1)) + A_k(\mu_C(\lambda_{C'_{k+1}} = -1)) = r_C^B(\lambda_{C'_{k+1}} \in Z') \\ &B_k(\mu_C(\lambda_{C'_{k+1}} = 0)) = r_C^B(\lambda_{C'_{k+1}} \in U'). \end{aligned}$$

Again, the first and third equations are clearly true using the inductive hypothesis, and the second equation is also true when we decompose $\prod_{z_i \in Z'} z_i = -1$ into $\prod_{z_i \in Z} z_i = 1, \lambda_{C'_{k+1}} = -1$ and $\prod_{z_i \in Z} z_i = -1, \lambda_{C'_{k+1}} = 1$.

We complete this proof by induction to conclude that $A_s \mu_C = r_C$ and $B_s \mu_C = r_C^B$, showing a recursive approach for mapping from r_C to μ_C for any clique or separator set C. □

Finally, we note that each r_C is made up of computable terms. Entries of the form $r_C(\emptyset, Z) = P(\prod_{z_i \in Z} z_i = 1)$ are immediately calculated from a_c for cliqes $c \subseteq C$, and entries where $Y(i) \notin Z$ can be directly estimated. Entries where $Z = \{Y(i)\}, U \neq \emptyset$ can be factorized into known or directly estimated probabilities, and all other entries can be computed by calculating each a_c conditional on U.

As an example, to construct r_{ij} for a clique $\{\lambda_i, \lambda_j, Y^{dep}(i,j)\}$, the only entries of r_{ij} that are unobservable from the data are $P(\lambda_i Y^{dep}(i,j) = 1)$, $P(\lambda_j Y^{dep}(i,j) = 1)$, $P(\lambda_i \lambda_j Y^{dep}(i,j) = 1)$, $P(\lambda_i = 0, Y^{dep}(i,j) = 1)$, $P(\lambda_j = 0, Y^{dep}(i,j) = 1)$, $P(\lambda_i = 0, \lambda_j Y^{dep}(i,j) = 1)$, $P(\lambda_j = 0, \lambda_i Y^{dep}(i,j) = 1)$, and $P(\lambda_i = 0, \lambda_j = 0, Y^{dep}(i,j) = 1)$. We have discussed how to estimate all but the last three.

To estimate $P(\lambda_i = 0, \lambda_j Y^{dep}(i,j) = 1)$, we can write this as

$$\begin{aligned} P(\lambda_j Y^{dep}(i,j) = 1, \lambda_i = 0) &= P(\lambda_j Y^{dep}(i,j) = 1 | \lambda_i = 0) P(\lambda_i = 0) \\ &= \frac{1 + \mathbb{E}\left[\lambda_j Y^{dep}(i,j) | \lambda_i = 0\right] - P(\lambda_j = 0 | \lambda_i = 0)}{2} \cdot P(\lambda_i = 0) \\ &= \frac{1}{2} P(\lambda_i = 0) + \frac{1}{2} \mathbb{E}\left[\lambda_j Y^{dep}(i,j) | \lambda_i = 0\right] P(\lambda_i = 0) + \frac{1}{2} P(\lambda_j = 0, \lambda_i = 0). \end{aligned}$$

We can solve $\mathbb{E}\left[\lambda_j Y^{dep}(i,j) | \lambda_i = 0\right]$ using the triplet method conditional on samples where λ_i abstains. $P(\lambda_i = 0, \lambda_j = 0, Y^{dep}(i,j) = 1)$ can be written as $P(\lambda_i = 0, \lambda_j = 0) P(Y^{dep}(i,j) = 1)$, of which all probabilities are observable, by Proposition 1.

C.1.5. ResolveSigns

This function is used to determine the signs after we have recovered the magnitudes of accuracy terms such as $|\mathbb{E}[v_i Y(i)]|$. One way to implement this function is to use one known accuracy sign per Y. We observe that if we know the sign of $a_i = \mathbb{E}[v_i Y(i)]$, then we are able to obtain the sign of any other term $a_j = \mathbb{E}[v_j Y(j)]$ where $Y(j) = Y(i)$. If v_i and v_j are conditionally independent given $Y(i)$, we directly use $a_i a_j = \mathbb{E}\left[v_i v_j\right]$ and knowledge of a_i's sign to get the sign of a_j. If v_i and v_j are not conditionally independent given $Y(i)$, we need two steps to recover the sign: for some v_k that is conditionally independent of both v_i and v_j given $Y(i)$, we first use $a_i\mathbb{E}\left[v_kY(i)\right] = \mathbb{E}\left[v_iv_k\right]$ to get the sign of $\mathbb{E}\left[v_kY(i)\right]$. Then we use $a_j\mathbb{E}\left[v_kY(i)\right] = \mathbb{E}\left[v_jv_k\right]$ to get the sign of a_j. Therefore, knowing the sign of one accuracy per Y is sufficient to recover all signs.

The ResolveSigns used in Algorithm 1 uses another approach and follows from the assumption that on average per Y, the accuracies a_i are better than zero. We apply this procedure to the sets of accuracies corresponding to each hidden variable; for each set, we have two sign choices, and we check which of these two produces a non-negative sum for the accuracies. In the common case where there is just one task, there are only two choices to check overall.

C.2. Extensions to More Complex Graphical Models

Recall that our Ising model is constructed for binary task labels, with sufficient conditional independence on G and G_{dep} such that $\Omega_G = V$, and without singleton potentials. We address how to extend our method when each of these conditions do not hold.

Multiclass Case We have given an algorithm for binary classes for $\boldsymbol{Y}$ (and ternary for the sources, since these can also abstain). To extend this to higher-class cases, we can apply a one-versus-all reduction repeatedly to apply our core algorithm.

Extension to More Complex Graphs In Algorithm 1, we rely on the fact $\Omega_G = V$ to compute all accuracies. However, certain a_i's cannot be recovered when there are fewer than 3 conditionally independent subgraphs in G, where a subgraph V_a is defined as a set of vertices such that if $v_i \in V_a$ and $v_j \notin V_a$, $v_i \perp\!\!\!\perp v_j | Y(i)$. Instead, when there are only 1 or 2 subgraphs, we use another independence property, which states that $v_iY(i) \perp\!\!\!\perp Y(i)$ for all v_i. This means that $\mathbb{E}\left[v_iY(i)\right] \cdot \mathbb{E}\left[Y(i)\right] = \mathbb{E}\left[v_iY(i)^2\right] = \mathbb{E}\left[v_i\right]$, and thus $a_i = \frac{\mathbb{E}[v_i]}{\mathbb{E}[Y(i)]}$. This independence property does not require us to choose triplets of sources; instead we can directly divide to compute a_i. However, this approach fails in the presence of singleton potentials and can be very inaccurate when $\mathbb{E}\left[Y(i)\right]$ is close to 0. One can use this independence property in addition to Proposition 1 on G with 2 conditionally independent subgraphs, and when G only consists of 1 subgraph, we require that there are no singleton potentials on any of the sources.

Dealing with Singleton Potentials Our current Ising model does not include singleton potentials except on Y_i terms. However, we can handle cases where sources are modeled to have singleton potentials. Proposition 1 holds as long as either v_i or v_j belongs to a subgraph that has no potentials on individual observed variables. Therefore, the triplet method is able to recover mean parameters as long as we have at least two conditionally independent subgraph with no singleton potentials on observed variables. For example, just two sources conditionally independent of all the others with no singleton potential suffices to guarantee that this modified graphical model still allows for our algorithm to recover label model parameters.

In the case where we have singleton potentials on possibly every source, we have the following alternative approach. We use a slightly different parametrization and a quadratic version of the triplet method. Instead of tracking mean parameters (and thus accuracies like $\mathbb{E}\left[v_iY(i)\right]$, we shall instead directly compute parameters that involve *class-conditional* probabilities. These are, in particular, for v_i,

$$\mu_i = \begin{bmatrix} P(v_i = 1|Y(i) = 1) & P(v_i = 1|Y(i) = -1) \\ P(v_i = -1|Y(i) = 1) & P(v_i = -1|Y(i) = -1) \end{bmatrix}.$$

Note that these parameters are minimal (the terms $P(v_i = 0|Y(i) = \pm 1)$, indicating the conditional abstain rate, are determined by the columns above.

We set

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

For a pair of conditionally independent sources, we have that

$$\mu_i P \mu_j^T = O_{ij}. \tag{10}$$

Because we can observe terms like O_{ij}, we can again form triplets with i, j, k as before, and solve. Note that this alternative parametrization does not depend on the presence or absence of singleton potentials in the Ising model, only on the conditional independences directly defined by it.

Moreover, there is a closed form solution to the resulting system of non-linear equations. To see this, consider the following. Note that

$$P(v_i = 1|Y(i) = -1) = \frac{P(v_i = 1)}{P(Y(i) = -1)} - \frac{P(v_i = 1|Y(i) = 1)P(Y(i) = 1)}{P(Y(i) = -1)}.$$

Note that everything is observable (or known, for class balances), so that we can write the top row of μ_i as a function of a single variable. That is, we set $\alpha = P(v_i = 1|Y(i) = 1)$, $c_i = \frac{P(v_i=1)}{P(Y(i)=-1)}$ and $d_i = \frac{P(Y(i)=1)}{P(Y(i)=-1)}$. Then, the top row of μ_i becomes $[\alpha \quad c_i - d_i\alpha]$, and c_i and d_i are known.

Next, consider some triplets i, j, k, with corresponding μ's. Similarly, we set the top-left corner in the corresponding μ's to be α, β, γ, and the corresponding terms for the top-right corner are c_i, c_j, c_k and d_i, d_j, d_k. Then, by considering the upper-left position in (10), we get the system

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

To solve this system, we express α and γ in terms of β, using the first and third equations, and then we can plug these into the second and multiply (for example, when using α, by $((1 + d_i d_j)\beta - c_j d_i)^2$) to obtain a quadratic in terms of β. Solving this quadratic and selecting the correct root, then obtaining the remaining parameters (α, γ) and filling in the rest of the μ_i, μ_j, μ_k terms completes the procedure. Note that we have to carry out the triplet procedure here twice per μ_i, since there are two rows. Lastly, we can convert probabilities over $\boldsymbol{v}$ into equivalent probabilities over $\boldsymbol{\lambda}$ as discussed in Appendix C.1.3.

C.3. Online Algorithm

The online learning setting presents new challenges for weak supervision. In the offline setting, the weak supervision pipeline has two distinct components: first, computing all probabilistic labels for a dataset and then using them to train an end model. In the online setting however, samples are introduced one by one, so we see each $\boldsymbol{X}^i$ only once and are not able to store it.

Fortunately, Algorithm 1 and Algorithm 2 both rely on computing estimates of expected moments over the observable weak sources. Since these are just averages, we can efficiently produce an estimate of the label model parameters at each time step. For each new sample, we update the averages of the moments using a rolling window and use them to output its probabilistic label; then the end model is trained on this sample, and the data point itself is no longer needed for further computation. Our method is fast enough that we can "interleave" the two components of the weak supervision pipeline, in comparison to Ratner et al. (2019) and Sala et al. (2019), which require a full covariance matrix inversion and SGD.

The online learning environment is also subject to *distributional drift* over time, where old samples may come from very different distributions compared to more recent samples. Formally, define distributional drift as the following property: for $(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_t$, the KL-divergence between P_i and P_{i+1} is less than $KL(P_t, P_{t+1}) \leq \Delta$ for any t. If there were no distributional drift, i.e., $\Delta = 0$, we would invoke Algorithm 1 or 2 at each time step t for the new sample's output label, where the estimates of $\hat{\mathbb{E}}[v_i v_j]$ and other observable moments would be cumulatively over t rather than n. However, because of distributional drift, it is important to prioritize most recent samples. We propose a rolling window of size W, which can be optimized theoretically, to average over rather than all past t samples. Algorithm 1 describes the general meta-algorithm for the online setting.

C.3.1. Theoretical Analysis

Similar to the offline setting, we analyze our method for online label model parameter recovery and provide bounds on its performance. First, we derive a bound on the sampling error $||\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}_t||_2$ in terms of the window size W, concluding that there exists an optimal W^* to minimize this error. Then, we present an online generalization result that describes how well our end model can "track" new samples coming from a drifting distribution.

Controlling the Online Sampling Error with W The sampling error at each time step t $||\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}_t||_2$ is dependent on the window size W which we average samples over to produce estimates. On one hand, a small window will ensure that the

Algorithm 1 Online Weak Supervision

```
Input: dependency graph $G_{dep}$, window $W$ for rolling averages
for $t = 1, 2, \ldots$ : do
    Receive source output vector $l_t$ and distribution prior $P_t(\bar{\boldsymbol{Y}})$.
    Run Algorithm 1 and Algorithm 2 with estimates computed over $W$ samples $l_{t-W+1:t}$ and their augmented equivalents to output $\hat{\boldsymbol{\mu}}_t$.
    Use junction tree formula to produce probabilistic output $\widetilde{\boldsymbol{Y}}^t \sim P_{\hat{\boldsymbol{\mu}}_t}(\,\cdot\,|l_t)$.
    Use $\widetilde{\boldsymbol{Y}}^t$ to update $w_t$, the parametrization of the end model $f_w$.
end for
```

estimate will be computed using samples from distributions close to P_t, but using few samples results in a high empirical estimation error. On the other hand, a larger window will allow us to use many samples; however, samples farther in the past will be from distributions that may not be similar to P_t. Hence, W must be selected to minimize both the effect of using drifting distributions and the estimation error in the number of samples used.

Theorem 1. *Let $\hat{\boldsymbol{\mu}}_t$ be an estimate of $\boldsymbol{\mu}_t$, the label model parameters at time t, over W previous samples from the product distribution $\mathbf{Pr}_W = \prod_{i=t-W+1}^{t} P_i$, which suffers a Δ-distributional drift. Then, still assuming cliques in G_{dep} are limited to 3 vertices,*

$$\mathbb{E}_{\mathbf{Pr}_W}\left[||\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t||_2\right] = \frac{1}{a_{\min}^5}\left(3.19C_1\sqrt{\frac{m}{W}} + \frac{6.35C_2}{\sqrt{r}}\frac{m}{\sqrt{W}}\right) + \frac{2c(|\mathcal{C}_{dep}| + |\mathcal{S}_{dep}|)\Delta W^{3/2}}{\sqrt{6\alpha_{P_t}}}.$$

where α_{P_t} is the minimum non-zero probability that P_t takes. A global minimum for the sampling error as a function of W exists, so the window size can be set such that $W^* = argmin_W\ \mathbb{E}\left[||\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t||_2\right]$.

Proof. Denote $P_t^W = \underbrace{P_t \times \ldots P_t}_{W}$. We first bound the difference between $\mathbb{E}_{\mathbf{Pr}_W}\left[||\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t||_2\right]$ and $\mathbb{E}_{P_t^W}\left[||\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t||_2\right]$.

$$\begin{aligned}
\left|\mathbb{E}_{\mathbf{Pr}_W}\left[\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2\right] - \mathbb{E}_{P_t^W}\left[\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2\right]\right| &= \left|\sum_{\{x_i\}_{i=t-W+1}^{t}} \|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2 \cdot \left(\mathbf{Pr}_W(x_{t-w+1}, \ldots, x_t) - P_t^W(x_{t-w+1}, \ldots, x_t)\right)\right| \\
&\le \max \|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2 \cdot \sum_{\{x_i\}_{i=t-W+1}^{t}} \left|\mathbf{Pr}_W(x_{t-w+1}, \ldots, x_t) - P_t^W(x_{t-w+1}, \ldots, x_t)\right| \\
&= \max \|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2 \cdot 2TV(\mathbf{Pr}_W, P_t^W).
\end{aligned}$$

Since the label model parameters are all probabilities, $\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2$ is bounded by $c \cdot (|\mathcal{C}_{dep}| + |\mathcal{S}_{dep}|)$, where c is a constant. To compute $TV(\mathbf{Pr}_W, P_t^W)$, we use Pinsker's inequality and tensorization of the KL-divergence:

$$\begin{aligned}
TV(\mathbf{Pr}_W, P_t^W) &\le \sqrt{\frac{1}{2}KL(\mathbf{Pr}_W||P_t^W)} = \sqrt{\frac{1}{2}KL(P_{t-W+1} \times \cdots \times P_t||P_t \times \cdots \times P_t)} \\
&= \sqrt{\frac{1}{2}\sum_{i=t-W+1}^{t} KL(P_i||P_t)}.
\end{aligned}$$

Each $KL(P_i||P_t)$ can be bounded above by $\frac{2}{\alpha_{P_t}}TV(P_i, P_t)^2$ by the inverse of Pinsker's inequality, where $\alpha_{P_t} = \min_{x \in \mathcal{X}, P_t(x) > 0} P_t(x)$. Since the triangle inequality is satisfied for total variation distance, $TV(P_i, P_t) \le \Delta(t-i)$. Plugging this back in, we get

$$\begin{aligned}
TV(\mathbf{Pr}_W, P_t^W) &\le \sqrt{\frac{1}{2} \cdot \frac{2}{\alpha_{P_t}}\Delta^2 \sum_{i=t-W+1}^{t}(t-i)^2} = \sqrt{\frac{\Delta^2}{\alpha_{P_t}}\sum_{i=0}^{W-1} i^2} \\
&= \sqrt{\frac{\Delta^2}{\alpha_{P_t}} \cdot \frac{(W-1)W(2W-1)}{6}} \le \frac{\Delta W^{3/2}}{\sqrt{6\alpha_{P_t}}}.
\end{aligned}$$

Therefore,

$$\left|\mathbb{E}_{\mathbf{Pr}_W}\left[\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2\right] - \mathbb{E}_{P_t^W}\left[\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2\right]\right| \leq \frac{2c(|\mathcal{C}_{dep}| + |\mathcal{S}_{dep}|)\Delta W^{3/2}}{\sqrt{6\alpha_{P_t}}}.$$

Furthermore, the offline sampling error result applies over P_t^W, so $\mathbb{E}_{P_t^W}\left[\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2\right] \leq \frac{1}{a_{\min}^5}\left(3.19C_1\sqrt{\frac{m}{W}} + \frac{6.35C_2}{\sqrt{r}}\frac{m}{\sqrt{W}}\right)$. Hence,

$$\mathbb{E}_{\mathbf{Pr}_W}\left[\|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}_t\|_2\right] \leq \frac{1}{a_{\min}^5}\left(3.19C_1\sqrt{\frac{m}{W}} + \frac{6.35C_2}{\sqrt{r}}\frac{m}{\sqrt{W}}\right) + \frac{2c(|\mathcal{C}_{dep}| + |\mathcal{S}_{dep}|)\Delta W^{3/2}}{\sqrt{6\alpha_{P_t}}},$$

and we set a window size W^* to minimize this expression. □

Online Generalization Bound We provide a bound quantifying the gap in probability of incorrectly classifying an unseen $t+1$th sample between our learned end model parametrization and an optimal end model parametrization.

Because the online learning setting is subject to distributional drift over time, our methods must be able to predict the next time step's label with some guarantee despite the changing environment. The Δ drift is aggravated by (1) potential model misspecification for each P_t and (2) sample noise. However, we are able to take into account these additional conditions by modeling the overall drift Δ^μ to be a combination of intrinsic distributional drift Δ, model misspecification, and estimation error of parameters.

Recall that $\boldsymbol{X}^i \sim P_i$ is drawn from the true distribution at time i, while $\widetilde{\boldsymbol{Y}}_i \sim P_{\hat{\boldsymbol{\mu}}_i}(\cdot|\boldsymbol{\lambda}(\boldsymbol{X}^i))$ is the probabilistic output of our label model. Define the joint distribution of a sample to be $(\boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) \sim P_{i,\hat{\boldsymbol{\mu}}_i}$. At each time step t, our goal is train our end model $f_w \in \mathcal{F}$ and evaluate its performance against the true $(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_t$, given that we have $t-1$ previous samples drawn from $P_{i,\hat{\boldsymbol{\mu}}_i}$.

We define a binary loss function $L(w,x,y) = |f_w(x) - y|$ and choose $\hat{w}_t$ to minimize over the past s samples such that

$$\hat{w}_t = \text{argmin}_w \frac{1}{s}\sum_{i=t-s}^{t-1} L(w, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i).$$

We present a new generalization result that bounds the probability that $f_{\hat{w}_t}(\boldsymbol{X}^t)$ does not equal the true $\boldsymbol{Y}^t$ and also accounts for model misspecification and error from parameter estimation.

Theorem 2. *Define $\Delta^\mu := d_{TV}(P_{i,\hat{\boldsymbol{\mu}}_i}, P_{i+1,\hat{\boldsymbol{\mu}}_{i+1}})$ to be the distributional drift between the two samples and $D^\mu := \max_i d_{TV}(P_i, P_{i,\hat{\boldsymbol{\mu}}_i})$ to be an upper bound for the total variational distance between the true distribution and the noise aware misspecified distribution. If $\Delta^\mu \leq \frac{c(\epsilon - 8D^\mu)^3}{\text{VCdim}(\mathcal{F})}$ for some constant $c > 0$, there exists a $\hat{w}_t$ computed over the past $s = \left\lfloor \frac{\epsilon - 8D^\mu}{16\Delta^\mu} \right\rfloor$ samples such that, for any time $t > s$ and $\epsilon \in (8D^\mu, 1)$,*

$$\mathbf{Pr}_{\hat{\boldsymbol{\mu}},t}(L(\hat{w}_t, \boldsymbol{X}^t, \boldsymbol{Y}^t) = 1) \leq \epsilon + \min_{w^*} P_t(L(w^*, \boldsymbol{X}^t, \boldsymbol{Y}^t) = 1),$$

where $\mathbf{Pr}_{\hat{\boldsymbol{\mu}},t} = \prod_{i=t-s}^{t-1} P_{i,\hat{\boldsymbol{\mu}}_i} \cdot P_t$. Furthermore,

$$D^\mu \leq \sqrt{\frac{1}{2}\max_i KL(P_i(\boldsymbol{Y}|\boldsymbol{X}) \mid\mid P_{\boldsymbol{\mu}_i}(\boldsymbol{Y}|\boldsymbol{X}))} + m^{\frac{1}{4}}\sqrt{\frac{1}{e_{min}}\max_i ||\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i||_2}.$$

Proof. We adapt Theorem 2 from Long (1999). Choose $\epsilon \leq 1$. Let $s = \left\lfloor \frac{\epsilon - 8D^\mu}{16(\Delta + 2D^\mu)} \right\rfloor$ and $\Delta^\mu = \Delta + 2D^\mu \leq \frac{(\epsilon - 8D^\mu)^3}{5000000d}$, where d is the end model's VC dimension. Let $L(w,x,y) = |y - f_w(x)| \in \{0,1\}$, where $f_w(x)$ is the output of the end model parametrized by w when given input x.

At time t, the sequence of inputs to the end model so far is $(\boldsymbol{X}^1, \widetilde{\boldsymbol{Y}}^1), (\boldsymbol{X}^2, \widetilde{\boldsymbol{Y}}^2), \ldots (\boldsymbol{X}^{t-1}, \widetilde{\boldsymbol{Y}}^{t-1})$, where $(\boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) \sim P_{i,\hat{\boldsymbol{\mu}}_i}$. We evaluate the end model's performance by using a parametrization w_t that is a function of the $t-1$ inputs so far and computing $L(w_t, \boldsymbol{X}^t, \boldsymbol{Y}^t)$ where $(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_t$. In particular, let $w_t^* = \text{argmin}_w \mathbb{E}_{(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_t}[L(w, \boldsymbol{X}^t, \boldsymbol{Y}^t)]$, and $\hat{w}_t = \text{argmin}_w \frac{1}{s} \sum_{i=t-s}^{t-1} L(w, x_i, \tilde{y}_i)$ where $x_i, \tilde{y}_i$ are the values of the random variables $\boldsymbol{X}^i$ and $\widetilde{\boldsymbol{Y}}_i$.

Suppose that $TV(P_i, P_{i+1}) \leq \Delta$. Then $TV(P_{i,\hat{\boldsymbol{\mu}}_i}, P_{i+1,\hat{\boldsymbol{\mu}}_{i+1}})$ is

$$TV(P_{i,\hat{\boldsymbol{\mu}}_i}, P_{i+1,\hat{\boldsymbol{\mu}}_{i+1}}) \leq TV(P_{i,\hat{\boldsymbol{\mu}}_i}, P_i) + \Delta + TV(P_{i+1}, P_{i+1,\hat{\boldsymbol{\mu}}_{i+1}}) \leq \Delta + 2D^\mu = \Delta^\mu.$$

Let $\beta \geq 6\Delta^\mu s + 4D^\mu$, and $\alpha = \frac{\beta}{2} - 2D^\mu \geq 3\Delta^\mu s$. Note that $TV(P_{i,\hat{\boldsymbol{\mu}}_i}, P_{t,\hat{\boldsymbol{\mu}}_t}) \leq \Delta^\mu s = \kappa$ for any $i = t-s, \ldots, t-1$. Denote $\mathbf{Pr}_{\hat{\boldsymbol{\mu}}} = \prod_{i=t-s}^{t-1} P_{i,\hat{\boldsymbol{\mu}}_i}$. Then by Lemma 12 of Long (1999),

$$\mathbf{Pr}_{\hat{\boldsymbol{\mu}}}\Big\{\exists w : \Big|\frac{1}{s}\sum_{i=t-s}^{t-1} L(w, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) - \mathbb{E}_{(\boldsymbol{X}^t, \widetilde{\boldsymbol{Y}}^t) \sim P_{t,\hat{\boldsymbol{\mu}}_t}}\left[L(w, \boldsymbol{X}^t, \widetilde{\boldsymbol{Y}}^t)\right]\Big| > \alpha\Big\} \leq 8 \cdot 41^d \exp\left(-\frac{(\alpha - \kappa)^2 s}{1600}\right).$$

For any real numbers a, b, c, and $x > y$, if $|a - b| \geq x$ and $|b - c| \leq y$, then $|a - b| - |b - c| \geq x - y$ and thus $|a - c| = |a - b + b - c| \geq ||a - b| - |b - c|| \geq x - y$. Applying this,

$$\begin{aligned}
&\mathbf{Pr}_{\hat{\boldsymbol{\mu}}}\Big\{\exists w : \Big|\frac{1}{s}\sum_{i=t-s}^{t-1} L(w, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) - \mathbb{E}_{(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_t}\left[L(w, \boldsymbol{X}^t, \boldsymbol{Y}^t)\right]\Big| > \alpha + 2D^\mu, \\
&\Big|\mathbb{E}_{(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_t}[L(w, \boldsymbol{X}^t, \boldsymbol{Y}^t)] - \mathbb{E}_{(\boldsymbol{X}^t, \widetilde{\boldsymbol{Y}}^t) \sim P_{t,\hat{\boldsymbol{\mu}}_t}}\left[L(w, \boldsymbol{X}^t, \widetilde{\boldsymbol{Y}}^t)\right]\Big| < 2D^\mu\Big\} \\
\leq\ &\mathbf{Pr}_{\hat{\boldsymbol{\mu}}}\Big\{\exists w : \Big|\frac{1}{s}\sum_{i=t-s}^{t-1} L(w, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) - \mathbb{E}_{(\boldsymbol{X}^t, \widetilde{\boldsymbol{Y}}^t) \sim P_{t,\hat{\boldsymbol{\mu}}_t}}\left[L(w, \boldsymbol{X}^t, \widetilde{\boldsymbol{Y}}^t)\right]\Big| > \alpha\Big\} \leq 8 \cdot 41^d \exp\left(-\frac{(\alpha - \kappa)^2}{1600}\right).
\end{aligned}$$

By Lemma 3, the difference in the expected loss $\mathbb{E}[L(w, \boldsymbol{X}^t, \boldsymbol{Y}^t)]$ when $\boldsymbol{X}^t, \boldsymbol{Y}^t$ is from P_t versus $P_{t,\hat{\boldsymbol{\mu}}_t}$ is always less than $2D^\mu$, so the above becomes

$$\begin{aligned}
&\mathbf{Pr}_{\hat{\boldsymbol{\mu}}}\Big\{\exists w : \Big|\frac{1}{s}\sum_{i=t-s}^{t-1} L(w, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) - \mathbb{E}_{(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_t}\left[L(w, \boldsymbol{X}^t, \boldsymbol{Y}^t)\right]\Big| > \alpha + 2D^\mu\Big\} \\
&\leq 8 \cdot 41^d \exp\left(-\frac{(\alpha - \kappa)^2 s}{1600}\right).
\end{aligned}$$

We can write this in terms of β. Note that $\Delta^\mu s \leq \frac{\beta}{6} - \frac{2D^\mu}{3}$. The RHS is equivalent to

$$\begin{aligned}
8 \cdot 41^d \exp\left(-\frac{(\alpha - \kappa)^2 m}{1600}\right) &= 8 \cdot 41^d \exp\left(-\frac{s}{1600}\left(\frac{\beta}{2} - 2D^\mu - \Delta^\mu s\right)^2\right) \\
\leq 8 \cdot 41^d \exp\left(-\frac{s}{1600}\left(\frac{\beta}{2} - 2D^\mu - \frac{\beta}{6} + \frac{2D^\mu}{3}\right)^2\right) &= 8 \cdot 41^d \exp\left(-\frac{s}{14400}(\beta - 4D^\mu)^2\right).
\end{aligned}$$

So the probability becomes

$$\mathbf{Pr}_{\hat{\boldsymbol{\mu}}}\Big\{\exists w : \Big|\frac{1}{s}\sum_{i=t-s}^{t-1} L(w, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) - \mathbb{E}_{(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_t}\left[L(w, \boldsymbol{X}^t, \boldsymbol{Y}^t)\right]\Big| > \frac{\beta}{2}\Big\} \leq 8 \cdot 41^d \exp\left(-\frac{s}{14400}(\beta - 4D^\mu)^2\right).$$

Next, note that the probability that at least one of $\hat{w}_t$ or w_t^* satisfies $\Big|\frac{1}{s}\sum_{i=t-s}^{t-1} L(w, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) - \mathbb{E}_{(\boldsymbol{X}^t, \boldsymbol{Y}^t) \sim P_i}\left[L(w, \boldsymbol{X}^t, \boldsymbol{Y}^t)\right]\Big| > \frac{\beta}{2}$ is less than the probability that there exists a w that satisfies the above inequality. In general, if $|a-b| > \beta$, then $|a| > \frac{\beta}{2}$ or $|b| > \frac{\beta}{2}$ (or both). Then

$$\mathbf{Pr}_{\hat{\boldsymbol{\mu}}}\Big\{\Big|\frac{1}{s}\sum_{i=t-s}^{t-1} L(w_t^*, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) - \mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(w_t^*, \boldsymbol{X}^t, \boldsymbol{Y}^t)]$$
$$- \frac{1}{s}\sum_{i=t-s}^{t-1} L(\hat{w}_t, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) + \mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(\hat{w}_t, \boldsymbol{X}^t, \boldsymbol{Y}^t)]\Big| > \beta\Big\}$$
$$\le \mathbf{Pr}_{\hat{\boldsymbol{\mu}}}\Big\{\Big|\frac{1}{s}\sum_{i=t-s}^{t-1} L(w_t^*, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) - \mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(w_t^*, \boldsymbol{X}^t, \boldsymbol{Y}^t)]| > \frac{\beta}{2}, \cup$$
$$\Big| - \frac{1}{s}\sum_{i=t-s}^{t-1} L(\hat{w}_t, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) + \mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(\hat{w}_t, \boldsymbol{X}^t, \boldsymbol{Y}^t)]\Big| > \frac{\beta}{2}\Big\}$$
$$\le 8 \cdot 41^d \exp\left(-\frac{s}{14400}(\beta - 4D^\mu)^2\right).$$

By definition of w_t^* and $\hat{w}_t$, $\frac{1}{s}\sum_{i=t-s}^{t-1} L(w_t^*, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i) > \frac{1}{s}\sum_{i=t-s}^{t-1} L(\hat{w}_t, \boldsymbol{X}^i, \widetilde{\boldsymbol{Y}}^i)$ and $\mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(\hat{w}_t, \boldsymbol{X}^t, \boldsymbol{Y}^t)] > \mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(w_t^*, \boldsymbol{X}^t, \boldsymbol{Y}^t)]$. Therefore,

$$\mathbf{Pr}_{\hat{\boldsymbol{\mu}}}\Big\{\mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(\hat{w}_t, \boldsymbol{X}^t, \boldsymbol{Y}^t)] - \mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(w_t^*, \boldsymbol{X}^t, \boldsymbol{Y}^t)] > \beta\Big\}$$
$$\le 8 \cdot 41^d \exp\left(-\frac{s}{14400}(\beta - 4D^\mu)^2\right).$$

Now we apply Lemma 13 from Long (1999). Define

$$\phi(\beta) = \begin{cases} 8 \cdot 41^d \exp\left(-\frac{s}{14400}(\beta - 4D^\mu)^2\right) & \beta \geq 6\Delta^\mu s + 4D^\mu \\ 1 & o.w. \end{cases}.$$

Let $a_0 = 0$ and $a_1 = 6\Delta^\mu s + 4D^\mu$. For all other a_i where $i > 1$ until some a_n where $a_{n+1} > 1$, define $a_i = \sqrt{\frac{14400(\ln 8 + (\ln 41)d + i \ln 2)}{s}} + 4D^\mu$. This way, $\phi(a_{i>1}) = 2^{-i}$. Then Lemma 13 states

$$\mathbb{E}_{\{(\boldsymbol{X}^i,\widetilde{\boldsymbol{Y}}^i)\sim P_{i,\hat{\boldsymbol{\mu}}_i}\}_{i=t-s}^{t-1}}[P_t(L(\hat{w}_t, \boldsymbol{X}^t, \boldsymbol{Y}^t) = 1) - P_t(L(w_t^*, \boldsymbol{X}^t, \boldsymbol{Y}^t) = 1)]$$
$$\le 1 \cdot a_1 + \sum_{i=1}^{\infty}\left(\sqrt{\frac{14400(\ln 8 + (\ln 41)d + i \ln 2)}{s}} + 4D^\mu\right)2^{-i}$$
$$\le 6\Delta^\mu s + 4D^\mu + 341\sqrt{\frac{d}{s}} + 4D^\mu = 6\Delta^\mu s + 8D^\mu + 341\sqrt{\frac{d}{s}}.$$

Plugging in our values of s and Δ^μ, we get that $6\Delta^\mu s + 8D^\mu + 341\sqrt{\frac{d}{s}} \le \epsilon$. Therefore, if the drift between two consecutive samples is less than $TV(P_{i,\hat{\boldsymbol{\mu}}_i}, P_{i+1,\hat{\boldsymbol{\mu}}_{i+1}}) \le \Delta^\mu \le \frac{(\epsilon - 8D^\mu)^3}{5000000d}$, there exists an algorithm that computes a $\hat{w}_t$ over the past $s = \left\lfloor \frac{\epsilon - 8D^\mu}{16(\Delta + 2D^\mu)} \right\rfloor$ inputs to the end model, such that

$$\mathbf{Pr}_{\hat{\boldsymbol{\mu}},t}(L(\hat{w}_t, \boldsymbol{X}^t, \boldsymbol{Y}^t) = 1) \le \epsilon + \min_{w^*} P_t(L(w^*, \boldsymbol{X}^t, \boldsymbol{Y}^t) = 1),$$

where $D^\mu \le \sqrt{\frac{1}{2}\max_i \mathbb{E}_{\boldsymbol{X}\sim P_i}[KL(P_i(\boldsymbol{Y}|\boldsymbol{X}) \mid\mid P_{\boldsymbol{\mu}_i}(\boldsymbol{Y}|\boldsymbol{X}))]} + m^{1/4}\sqrt{\frac{1}{\sigma_{min}}\max_i ||\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i||_2}$ by Lemma 4. □

Lemma 3. *The difference in the expected value of $L(w, \boldsymbol{X}, \boldsymbol{Y})$ when samples are drawn from $P_{t,\hat{\boldsymbol{\mu}}_t}$ versus P_t is*

$$\Big|\mathbb{E}_{(\boldsymbol{X}^t,\widetilde{\boldsymbol{Y}}^t)\sim P_{t,\hat{\boldsymbol{\mu}}_t}}[L(w, \boldsymbol{X}^t, \widetilde{\boldsymbol{Y}}^t)] - \mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(w, \boldsymbol{X}^t, \boldsymbol{Y}^t)]\Big| \le 2D^\mu.$$

Proof. We use the definition of total variation distance:

$$\begin{aligned}
&\left|\mathbb{E}_{(\boldsymbol{X}^t,\widetilde{\boldsymbol{Y}}^t))\sim P_{t,\hat{\boldsymbol{\mu}}_t}}[L(w,\boldsymbol{X}^t,\widetilde{\boldsymbol{Y}}^t] - \mathbb{E}_{(\boldsymbol{X}^t,\boldsymbol{Y}^t)\sim P_t}[L(w,\boldsymbol{X}^t,\boldsymbol{Y}^t)]\right| \\
&= \left|\sum_{x,y} L(w,x,y)(P_{t,\hat{\boldsymbol{\mu}}_t}(x,y) - P_t(x,y))\right| \\
&\leq \sum_{x,y} L(w,x,y)|P_{t,\hat{\boldsymbol{\mu}}_t}(x,y) - P_t(x,y)| \\
&\leq \sum_{x,y} |P_{t,\hat{\boldsymbol{\mu}}_t}(x,y) - P_t(x,y)| = 2TV(P_{t,\hat{\boldsymbol{\mu}}_t}, P_t) \leq 2D^\mu.
\end{aligned}$$

$\square$

Lemma 4.

$$D^\mu \leq \sqrt{\frac{1}{2}\max_i KL(P_i(\boldsymbol{Y}|\boldsymbol{X}) \mid\mid P_{\boldsymbol{\mu}_i}(\boldsymbol{Y}|\boldsymbol{X}))} + m^{1/4}\sqrt{\frac{1}{\sigma_{min}}\max_i ||\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i||_2}.$$

Here, σ_{min} is the minimum singular value of the covariance matrix Σ of the variables $V = \{\boldsymbol{Y}, \boldsymbol{v}\}$ in the graphical model.

Proof. We first use the triangle inequality on TV distance to split D^μ into two KL-divergences.

$$\begin{aligned}
D^\mu &\leq \max_i TV(P_{i,\hat{\boldsymbol{\mu}}_i}, P_i) \leq \max_i TV(P_{i,\hat{\boldsymbol{\mu}}_i}, P_{i,\boldsymbol{\mu}_i}) + \max_i TV(P_{i,\boldsymbol{\mu}_i}, P_i) \\
&\leq \sqrt{\frac{1}{2}\max_i KL(P_{i,\boldsymbol{\mu}_i}||P_{i,\hat{\boldsymbol{\mu}}_i})} + \sqrt{\frac{1}{2}\max_i KL(P_i||P_{i,\boldsymbol{\mu}_i})}.
\end{aligned}$$

To simplify the first divergence, we use the binary Ising model definition in (1), which for simplicity we write as $f_G(\boldsymbol{Y}, \boldsymbol{v}) = \frac{1}{Z}\exp(\theta^T\phi(V))$, where $\phi(V)$ is the vector of all potentials.

$$\begin{aligned}
KL(P_{i,\boldsymbol{\mu}_i}||P_{i,\hat{\boldsymbol{\mu}}_i}) &= (\hat{\theta}_i - \theta_i)^T\mathbb{E}[\phi(V)] + \ln\frac{\hat{Z}}{Z} \leq |\hat{\theta}_i - \theta_i|_1 + \ln\frac{\hat{Z}}{Z} \leq \sqrt{m}||\hat{\theta}_i - \theta_i||_2 + \ln\frac{\sum_{s\in\mathcal{S}}\exp(\hat{\theta}_i^T\phi(s))}{\sum_{s\in\mathcal{S}}\exp(\theta_i^T\phi(s))} \\
&\leq \sqrt{m}||\hat{\theta}_i - \theta_i||_2 + \frac{1}{\hat{Z}}\sum_{s\in\mathcal{S}}\exp(\hat{\theta}_i^T\phi(s))\ln\frac{\exp(\hat{\theta}_i^T\phi(s))}{\exp(\theta_i^T\phi(s))} \\
&\leq \sqrt{m}||\hat{\theta}_i - \theta_i||_2 + \frac{1}{\hat{Z}}\sum_{s\in\mathcal{S}}\exp(\hat{\theta}_i^T\phi(s))((\hat{\theta}_i - \theta_i)^T\phi(s)) \\
&\leq \sqrt{m}||\hat{\theta}_i - \theta_i||_2 + \frac{1}{\hat{Z}}\sum_{s\in\mathcal{S}}\exp(\hat{\theta}_i^T\phi(s))\sqrt{m}||\hat{\theta}_i - \theta_i||_2 \leq 2\sqrt{m}||\hat{\theta}_i - \theta_i||_2 \\
&\leq \frac{2\sqrt{m}}{\sigma_{min}}||\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i||_2.
\end{aligned}$$

Here we used $\phi(s), \mathbb{E}[\phi(V)] \in [-1, +1]$, the log sum inequality, and Lemma 8. The second divergence can be simplified into a conditional KL-divergence.

$$\begin{aligned}
KL(P_i||P_{i,\boldsymbol{\mu}_i}) &= \sum_{x,y} P_i(x,y)\ln\frac{P_i(x,y)}{P_{i,\boldsymbol{\mu}_i}(x,y)} = \sum_{x,y} P_i(x,y)\ln\frac{P_i(y|x)P_i(x)}{P_{i,\boldsymbol{\mu}_i}(y|x)P_{i,\boldsymbol{\mu}_i}(x)} \\
&= \sum_{x,y} P_i(x,y)\ln\frac{P_i(y|x)P_i(x)}{P_{\boldsymbol{\mu}_i}(y|x)P_i(x)} = \sum_x P_i(x)\sum_y P_i(y|x)\ln\frac{P_i(y|x)}{P_{\boldsymbol{\mu}_i}(y|x)} \\
&= \sum_x P_i(x)KL(P_i(\boldsymbol{Y}|x)||P_{\boldsymbol{\mu}_i}(\boldsymbol{Y}|x)) = KL(P_i(\boldsymbol{Y}|\boldsymbol{X}) \mid\mid P_{\boldsymbol{\mu}_i}(\boldsymbol{Y}|\boldsymbol{X})),
\end{aligned}$$

where

$$KL(P_i(\boldsymbol{Y}|\boldsymbol{X}) \mid\mid P_{\boldsymbol{\mu}_i}(\boldsymbol{Y}|\boldsymbol{X})) = \mathbb{E}_{P_i}[KL(P_i(\boldsymbol{Y}|x) \mid\mid P_{\boldsymbol{\mu}_i}(\boldsymbol{Y}|x))].$$

□

This result suggests that, with a small enough Δ^{μ}, our parametrization of the end model using past data will perform only ϵ worse in probability than the best possible parametrization of the end model on the next data point. Furthermore, note that s is decreasing in D^{μ}; more model misspecification and sampling error intuitively suggests that we want to use fewer previous data points to compute $\hat{w}_t$, so again having a simple yet suitable graphical model allows the end model to train on more data for better prediction.

D. Proofs of Main Results

D.1. Proof of Theorem 1 (Sampling Error)

We first present three concentration inequalities - one on the accuracies estimated via the triplet method, and the other two on directly observable values. Afterwards, we discuss how to combine these inequalities into a sampling error result for $\boldsymbol{\mu}$ when G_{dep} has small cliques of size 3 or less.

Estimation error for a_i using Algorithm 1

Lemma 5. *Denote M as the second moment matrix over all observed variables, e.g. $M_{ij} = \mathbb{E}\left[v_i v_j\right]$. Let $\hat{a}$ be an estimate of the m desired accuracies a using $\hat{M}$ computed from n samples. Define $a_{\min} = \min\{\min_i |\hat{a}_i|, \min_i |a_i|\}$, and assume sign$(a_i)$ = sign$(\hat{a}_i)$ for all a_i. Furthermore, assume that the number of samples n is greater than some n_0 such that $a_{\min} > 0$, and $\hat{M}_{ij} \neq 0$. Then the estimation error of the accuracies is*

$$\Delta_a = \mathbb{E}[\|\hat{a} - a\|_2] \leq C_a \frac{1}{a_{\min}^5} \sqrt{\frac{m}{n}},$$

for some constant C_a.

Proof. We start with a few definitions. Denote a triplet as $T_i(1), T_i(2), T_i(3)$, and in total suppose we need τ number of triplets. Recall that our estimate of a can be obtained with

$$|\hat{a}_{T_i(1)}| = \left(\frac{|\hat{M}_{T_i(1)T_i(2)}||\hat{M}_{T_i(1)T_i(3)}|}{|\hat{M}_{T_i(2)T_i(3)}|}\right)^{\frac{1}{2}}.$$

Because we assume that signs are completely recoverable,

$$\|\hat{a} - a\|_2 = \||\hat{a}| - |a|\|_2 \leq \left(\sum_{i=1}^{\tau} (|\hat{a}_{T_i(1)}| - |a_{T_i(1)}|)^2 + (|\hat{a}_{T_i(2)}| - |a_{T_i(2)}|)^2 + (|\hat{a}_{T_i(3)}| - |a_{T_i(3)}|)^2\right)^{\frac{1}{2}}. \tag{11}$$

Note that $|\hat{a}_i^2 - a_i^2| = |\hat{a}_i - a_i||\hat{a}_i + a_i|$. By the reverse triangle inequality, $(|\hat{a}_i| - |a_i|)^2 = \||\hat{a}_i| - |a_i|\|^2 \leq |\hat{a}_i - a_i|^2 = \left(\frac{|\hat{a}_i^2 - a_i^2|}{|\hat{a}_i + a_i|}\right)^2 \leq \frac{1}{4a_{\min}^2}|\hat{a}_i^2 - a_i^2|^2$, because $|\hat{a}_i + a_i| = |\hat{a}_i| + |a_i| \geq 2a_{\min}$. For ease of notation, suppose we examine a particular $T_i = \{1, 2, 3\}$. Then

$$\begin{aligned}
(|\hat{a}_1| - |a_1|)^2 &\leq \frac{1}{4a_{\min}^2}|\hat{a}_1^2 - a_1^2|^2 = \frac{1}{c^2}\left|\frac{|\hat{M}_{12}||\hat{M}_{13}|}{|\hat{M}_{23}|} - \frac{|M_{12}||M_{13}|}{|M_{23}|}\right|^2 \\
&= \frac{1}{4a_{\min}^2}\left|\frac{|\hat{M}_{12}||\hat{M}_{13}|}{|\hat{M}_{23}|} - \frac{|\hat{M}_{12}||\hat{M}_{13}|}{|M_{23}|} + \frac{|\hat{M}_{12}||\hat{M}_{13}|}{|M_{23}|} - \frac{|\hat{M}_{12}||M_{13}|}{|M_{23}|} + \frac{|\hat{M}_{12}||M_{13}|}{|M_{23}|} - \frac{|M_{12}||M_{13}|}{|M_{23}|}\right|^2 \\
&\leq \frac{1}{4a_{\min}^2}\left(\left|\frac{\hat{M}_{12}\hat{M}_{13}}{\hat{M}_{23}M_{23}}\right|\left||\hat{M}_{23}| - |M_{23}|\right| + \left|\frac{\hat{M}_{12}}{M_{23}}\right|\left||\hat{M}_{13}| - |M_{13}|\right| + \left|\frac{M_{13}}{M_{23}}\right|\left||\hat{M}_{12}| - |M_{12}|\right|\right)^2 \\
&\leq \frac{1}{4a_{\min}^2}\left(\left|\frac{\hat{M}_{12}\hat{M}_{13}}{\hat{M}_{23}M_{23}}\right||\hat{M}_{23} - M_{23}| + \left|\frac{\hat{M}_{12}}{M_{23}}\right||\hat{M}_{13} - M_{13}| + \left|\frac{M_{13}}{M_{23}}\right||\hat{M}_{12} - M_{12}|\right)^2. \qquad (12)
\end{aligned}$$

Clearly, all elements of $\hat{M}$ and M must be less than 1. We further know that elements of $|M|$ are at least a_{min}^2, since $\mathbb{E}\left[v_i v_j\right] = \mathbb{E}\left[v_i Y\right]\mathbb{E}\left[v_j Y\right] \geq a_{\min}^2$. Furthermore, elements of $|\hat{M}|$ are also at least $a_{\min}^2$ because $|\hat{M}_{ij}| = \hat{a}_i\hat{a}_j \geq a_{\min}^2$ by construction of our algorithm. Define $\Delta ij = \hat{M}_{ij} - M_{ij}$. Then

$$\begin{aligned}
(|\hat{a}_1| - |a_1|)^2 &\leq \frac{1}{4a_{\min}^2}\left(\frac{1}{a_{\min}^4}|\Delta_{23}| + \frac{1}{a_{\min}^2}|\Delta_{13}| + \frac{1}{a_{\min}^2}|\Delta_{12}|\right)^2 \\
&\leq \frac{1}{4a_{\min}^2}(\Delta_{23}^2 + \Delta_{13}^2 + \Delta_{12}^2)\left(\frac{1}{a_{\min}^8} + \frac{2}{a_{\min}^4}\right).
\end{aligned}$$

(11) is now

$$\|\hat{a} - a\|_2 \leq \left(\frac{3}{4a_{\min}^2}\left(\frac{1}{a_{\min}^8} + \frac{2}{a_{\min}^4}\right)\sum_{i=1}^{\tau}\left(\Delta_{T_i(1)T_i(2)}^2 + \Delta_{T_i(1)T_i(3)}^2 + \Delta_{T_i(2)T_i(3)}^2\right)\right)^{\frac{1}{2}}.$$

To bound the maximum absolute value between elements of $\hat{M}$ and M, note that the Frobenius norm of the 3×3 submatrix defined over T_i is

$$\|\hat{M}_{T_i} - M_{T_i}\|_F = \left(2\left(\Delta_{T_i(1)T_i(2)}^2 + \Delta_{T_i(1)T_i(3)}^2 + \Delta_{T_i(2)T_i(3)}^2\right)\right)^{\frac{1}{2}}.$$

Moreover, $\|\hat{M}_{T_i} - M_{T_i}\|_F = \sqrt{\sum_{j=1}^3 \sigma_j^2(\hat{M}_{T_i} - M_{T_i})} \leq \sqrt{3}\|\hat{M}_{T_i} - M_{T_i}\|_2$. Putting everything together,

$$\begin{aligned}
\|\hat{a} - a\|_2 &\leq \left(\frac{3}{4a_{\min}^2}\left(\frac{1}{a_{\min}^8} + \frac{2}{a_{\min}^4}\right) \cdot \frac{1}{2}\sum_{i=1}^{\tau}\|\hat{M}_{T_i} - M_{T_i}\|_F^2\right)^{\frac{1}{2}} \\
&\leq \left(\frac{3}{4a_{\min}^2}\left(\frac{1}{a_{\min}^8} + \frac{2}{a_{\min}^4}\right) \cdot \frac{3}{2}\sum_{i=1}^{\tau}\|\hat{M}_{T_i} - M_{T_i}\|_2^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Lastly, to compute $\mathbb{E}[\|\hat{a} - a\|_2]$, we use Jensen's inequality and linearity of expectation:

$$\mathbb{E}\|\hat{a} - a\|_2] \leq \left(\frac{3}{4a_{\min}^2}\left(\frac{1}{a_{\min}^8} + \frac{2}{a_{\min}^4}\right) \cdot \frac{3}{2}\sum_{i=1}^{\tau}\mathbb{E}[\|\hat{M}_{T_i} - M_{T_i}\|_2^2]\right)^{\frac{1}{2}}.$$

We use the matrix Hoeffding inequality as described in Ratner et al. (2019), which says

$$P(\|\hat{M} - M\|_2 \geq \gamma) \leq 2m\exp\left(-\frac{n\gamma^2}{32m^2}\right).$$

To get the probability distribution over $\|\hat{M} - M\|_2^2$, we just note that $P(\|\hat{M} - M\|_2 \geq \gamma) = P(\|\hat{M} - M\|_2^2 \geq \gamma^2)$ to get

$$P(\|\hat{M} - M\|_2^2 \geq \gamma) \leq 2m \exp\left(-\frac{n\gamma}{32m^2}\right).$$

From which we can integrate to get

$$\mathbb{E}[\|\hat{M}_{T_i} - M_{T_i}\|_2^2] = \int_0^\infty P(\|\hat{M}_{T_i} - M_{T_i}\|_2^2 \geq \gamma) d\gamma \leq \frac{64(3)^3}{n}.$$

Substituting this back in, we get

$$\begin{aligned}\mathbb{E}[\|\hat{a} - a\|_2] &\leq \left(\frac{3}{4a_{\min}^2}\left(\frac{1}{a_{\min}^8} + \frac{2}{a_{\min}^4}\right) \cdot \frac{3\tau}{2}\frac{1728}{n}\right)^{\frac{1}{2}} \\ &\leq \left(\frac{1944}{a_{\min}^2} \cdot \left(\frac{1}{a_{\min}^8} + \frac{2}{a_{\min}^4}\right) \cdot \frac{\tau}{n}\right)^{\frac{1}{2}}.\end{aligned}$$

Finally, note that

$$\frac{1}{a_{\min}^2} \cdot \left(\frac{1}{a_{\min}^8} + \frac{2}{a_{\min}^4}\right) = \frac{1}{a_{\min}^2} \cdot \frac{1 + 2a_{\min}^4}{a_{\min}^8} \leq \frac{3}{a_{\min}^{10}}.$$

Therefore, the sampling error for the accuracy is bounded by

$$\mathbb{E}[\|\hat{a} - a\|_2] \leq \left(\frac{1944 \cdot 3}{a_{\min}^{10}} \cdot \frac{\tau}{n}\right)^{\frac{1}{2}} \leq C_a \frac{1}{a_{\min}^5}\sqrt{\frac{m}{n}}.$$

This is because at most we will use a triplet to compute each relevant a_i, meaning that $\tau \leq m$. The term C_a here is $18\sqrt{6}$.

□

Remark 1. *Although a lower bound on accuracy $a_{\min}$ invariably appears in this result, the dependence on a single low-accuracy source $\lambda_{\min}$ can be reduced. We improve our bound from having a $\frac{1}{a_{\min}^5}$ dependency to one additive term of order $\frac{1}{a_{\min}\sqrt{n}}$, while other terms are not dependent on $a_{\min}$ and are overall of order $\sqrt{\frac{m-1}{n}}$. In (12), the $4a_{\min}^2$ can be tightened to $4a_i^2$ for each λ_i, and M_{23} and $\hat{M}_{23}$ are not in terms of $a_{\min}$ if neither of the two labeling functions at hand are $\lambda_{\min}$. Therefore, for any $\lambda_i \neq \lambda_{\min}$, we do not have a dependency on $a_{\min}$ if we ensure that the triplet used to recover its accuracy in Algorithm 1 does not include $\lambda_{\min}$. Then only one term in our final bound will have a $\frac{1}{a_{\min}\sqrt{n}}$ dependency compared to the previous $\frac{1}{a_{\min}^5}\sqrt{\frac{m}{n}}$.*

Concentration inequalities on observable data

Lemma 6. *Define $p^{(i)}(x) = P(\lambda_i = x)$ and $\hat{p}^{(i)}(x) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}\left\{L_k^{(i)} = x\right\}$, and let $p(x), \hat{p}(x) \in \mathbb{R}^m$ denote the vectors over all i. Then*

$$\Delta_p := \mathbb{E}\left[\|\hat{p}(x) - p(x)\|_2\right] \leq \sqrt{\frac{m}{n}}.$$

Proof. Note that $\mathbb{E}\left[\mathbb{1}\left\{L_k^{(i)} = x\right\}\right] = P(\lambda_i = 1)$. Then using Hoeffding's inequality, we have that

$$P(|\hat{p}^{(i)}(x) - p^{(i)}(x)| \geq \epsilon) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{n(1)^2}\right) \leq 2\exp\left(-2n\epsilon^2\right).$$

This expression is equivalent to

$$P(|p^{(i)}(x) - p^{(i)}(x)|^2 \geq \epsilon) \leq 2\exp\left(-2n\epsilon\right).$$

We can now compute $\mathbb{E}\left[|\hat{p}^{(i)}(x) - p^{(i)}(x)|^2\right]$:

$$\mathbb{E}\left[|\hat{p}^{(i)}(x) - p^{(i)}(x)|^2\right] \leq \int_0^\infty 2\exp\left(-2n\epsilon\right) d\epsilon = -2 \cdot \frac{1}{2n}\exp\left(-2n\epsilon\right)\Big|_0^\infty = \frac{1}{n}.$$

The overall L2 error for $p(x)$ is then

$$\mathbb{E}\left[\|\hat{p}(x) - p(x)\|_2\right] = \mathbb{E}\left[\Big(\sum_{i=1}^m |\hat{p}^{(i)}(x) - p^{(i)}(x)|^2\Big)^{1/2}\right] \leq \sqrt{\sum_{i=1}^m \mathbb{E}\left[|\hat{p}^{(i)}(x) - p^{(i)}(x)|^2\right]} \leq \sqrt{\frac{m}{n}}.$$

□

Lemma 7. *Define $M(a,b)$ to be a second moment matrix where $M(a,b)_{ij} = \mathbb{E}\left[a_i b_j\right]$ for some random variables $a_i, b_j \in \{-1, 0, 1\}$ each corresponding to λ_i, λ_j. Let $\|\cdot\|_{ij}$ be the Frobenius norm over elements indexed at (i,j), where λ_i and λ_j share an edge in the dependency graph. If G_{dep} has d conditionally independent subgraphs, the estimation error of M is*

$$\Delta_M := \mathbb{E}[\|\hat{M}(a,b) - M(a,b)\|_{ij}] \leq C_m\sqrt{\frac{d-1+(m-d+1)^2}{n}} \leq C_m\frac{m}{\sqrt{n}}.$$

For some constant C_m.

Proof. Recall that the subgraphs are defined as sets $V_1, \ldots, V_d$, and let $E_1, \ldots, E_d$ be the corresponding sets of edges within the subgraphs. We can split up the norm $\|\hat{M}(a,b) - M(a,b)\|_{ij}$ into summations over sets of edges.

$$\begin{aligned}\|\hat{M}(a,b) - M(a,b)\|_{ij} &= \Big(\sum_{(i,j)\in E_{dep}} (\hat{M}(a,b)_{ij} - M(a,b)_{ij})^2\Big)^{\frac{1}{2}} = \Big(\sum_{k=1}^d \sum_{(i,j)\in E_k} (\hat{M}(a,b)_{ij} - M(a,b)_{ij})^2\Big)^{\frac{1}{2}} \\ &\leq \Big(\sum_{k=1}^d \sum_{i,j\in V_k} (\hat{M}(a,b)_{ij} - M(a,b)_{ij})^2\Big)^{\frac{1}{2}} = \Big(\sum_{k=1}^d \frac{1}{2}\|\hat{M}(a,b)_{V_k} - M(a,b)_{V_k}\|_F^2\Big)^{\frac{1}{2}}.\end{aligned}$$

We take the expectation of both sides by using linearity of expectation and Jensen's inequality:

$$\mathbb{E}[\|\hat{M}(a,b) - M(a,b)\|_{ij}] \leq \Big(\sum_{k=1}^d \frac{1}{2}\mathbb{E}[\|\hat{M}(a,b)_{V_k} - M(a,b)_{V_k}\|_F^2]\Big)^{\frac{1}{2}}.$$

We are able to modify Proposition A.3 of Bunea & Xiao (2015) into a concentration inequality for the second moment matrix rather than the covariance matrix, which states that $\mathbb{E}[\|\hat{M}(a,b)_{V_k} - M(a,b)_{V_k}\|_F^2] \leq (32e^{-4} + e + 64)\left(\frac{4c_1 tr(M_{V_k})}{\sqrt{n}}\right)^2$ for some constant c_1. We are able to use this result because our random variables are sub-Gaussian and have bounded higher order moments. Then our bound becomes

$$\mathbb{E}[\|\hat{M}(a,b) - M(a,b)\|_{ij}] \leq \Big(\sum_{k=1}^d \frac{1}{2}(32e^{-4} + e + 64)\frac{16c_1^2|V_k|^2}{n}\Big)^{\frac{1}{2}} \leq \Big(\frac{8c_1^2(32e^{-4} + e + 64)}{n}\sum_{k=1}^d |V_k|^2\Big)^{\frac{1}{2}}.$$

$\sum_{k=1}^d |V_k|^2$ is maximized when we have $d-1$ sugraphs of size 1 and 1 subgraph of size $m-d+1$, in which case the summation is $d - 1 + (m-d+1)^2$. Intuitively, when there are more subgraphs, this value will be smaller and closer to an order of m rather than m^2. Putting this together, our bound is

$$\mathbb{E}[\|\hat{M}(a,b) - M(a,b)\|_{ij}] \leq \Big(8c_1^2(32e^{-4} + e + 64)\frac{d-1+(m-d+1)^2}{n}\Big)^{\frac{1}{2}} \leq C_m\frac{m}{\sqrt{n}}.$$

Where $C_m = \sqrt{8c_1^2(32e^{-4} + e + 64)}$. □

Estimating $\mu_{\mathbf{i}}$ We first estimate $\mu_i = P(\lambda_i, Y^{dep}(i))$ for all relevant λ_i. For ease of notation, let Y refer to $Y^{dep}(i)$ in this section. Denote $\boldsymbol{\mu}_i$ to be the vector of all μ_i across all $\boldsymbol{\lambda}$. Note that

$$\|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|_2 \leq \|diag_m(A_1^{-1})\|_2 \|\hat{\rho} - \rho\|_2.$$

ρ is the vector of all r_i for $i = 1, \dots, m$, and $diag_m(A_1^{-1})$ is a block matrix containing m A_1^{-1} on its diagonal; note that the 2-norm of a block diagonal matrix is just the maximum 2-norm over all of the block matrices, which is $\|A_1^{-1}\|_2$. Recall that $r_i = [1 \;\; P(\lambda_i = 1) \;\; P(\lambda_i = 0) \;\; P(Y = 1) \;\; P(\lambda_i Y = 1) \;\; P(\lambda_i = 0, Y = 1)]^T$. For each term of r_i, we have a corresponding sampling error to compute over ρ:

- $P(\lambda_i = 1)$: We need to compute $\hat{P}(\lambda_i = 1) - P(\lambda_i = 1)$ for each λ_i. All together, the sampling error for this term is equivalent to $\|\hat{p}(1) - p(1)\|_2$.
- $P(\lambda_i = 0)$: The sampling error over all $\hat{P}(\lambda_i = 0) - P(\lambda_i = 0)$ is equivalent to $\|\hat{p}(0) - p(0)\|_2$.
- $P(\lambda_i Y = 1)$: Since $a_i = \mathbb{E}\left[v_{2i-1} Y\right] = \mathbb{E}\left[\lambda_i Y\right] = P(\lambda_i Y = 1) - P(\lambda_i Y = -1) = 2P(\lambda_i Y = 1) + P(\lambda_i = 0) - 1$ and the sampling error over all $\hat{P}(\lambda_i Y = 1) - P(\lambda_i Y = 1)$ is at most $\frac{1}{2}\|(\hat{a} - a) - (\hat{p}(0) - p(0))\|_2 \leq \frac{1}{2}\left(\|\hat{a} - a\|_2 + \|\hat{p}(0) - p(0)\|_2\right)$.
- $P(\lambda_i = 0, Y = 1)$: This expression is equal to $P(\lambda_i = 0)P(Y = 1)$, so the sampling error is $P(Y = 1)\|\hat{p}(0) - p(0)\|_2 \leq \|\hat{p}(0) - p(0)\|_2$.

Putting these error terms together, we have an expression for the sampling error for ρ:

$$\begin{aligned}
\|\hat{\rho} - \rho\|_2 &= \sqrt{\|\hat{p}(1) - p(1)\|_2^2 + 2\|\hat{p}(0) - p(0)\|_2^2 + \frac{1}{4}(\|\hat{a} - a\|_2 + \|\hat{p}(0) - p(0)\|_2)^2} \\
&\leq \|\hat{p}(1) - p(1)\|_2 + \sqrt{2}\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}(\|\hat{a} - a\| + \|\hat{p}(0) - p(0)\|) \\
&= \|\hat{p}(1) - p(1)\|_2 + \Big(\frac{1}{2} + \sqrt{2}\Big)\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}\|\hat{a} - a\|_2,
\end{aligned}$$

where we use concavity of the square root in the first step. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\rho} - \rho\|_2\right] &\leq \mathbb{E}\left[\|\hat{p}(1) - p(1)\|_2\right] + \Big(\frac{1}{2} + \sqrt{2}\Big)\mathbb{E}\left[\|\hat{p}(0) - p(0)\|_2\right] + \frac{1}{2}\mathbb{E}\left[\|\hat{a} - a\|_2\right] \\
&= \Big(\frac{3}{2} + \sqrt{2}\Big)\Delta_p + \frac{1}{2}\Delta_a.
\end{aligned}$$

Plugging this back into our error for $\boldsymbol{\mu}_i$ and using Lemmas 5 and 6,

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|_2\right] \leq \|A_1^{-1}\|_2 \left(\left(\frac{3}{2} + \sqrt{2}\right)\sqrt{\frac{m}{n}} + \frac{C_a}{2a_{|min|}^5}\sqrt{\frac{m}{n}}\right).$$

Therefore, if there are no cliques of size 3 or greater in G_{dep}, the sampling error is $\mathcal{O}(\sqrt{m/n})$.

Estimating all μ_{ij} Now we estimate $\mu_{ij} = P(\lambda_i, \lambda_j, Y^{dep}(i,j))$ for λ_i, λ_j sharing an edge in G_{dep}. For ease of notation, let Y refer to $Y^{dep}(i,j)$ in this section. Denote $\boldsymbol{\mu}_{ij}$ to be the vector of all μ_{ij}. Note that

$$\|\hat{\boldsymbol{\mu}}_{ij} - \boldsymbol{\mu}_{ij}\|_2 \leq \|diag_{|E|}(A_2)^{-1}\|_2 \|\hat{\psi} - \psi\|_2 = \|A_2^{-1}\|_2 \|\hat{\psi} - \psi\|_2.$$

ψ is the vector of all r_{ij} for all $(i,j) \in E$. Recall that $a_i = \mathbb{E}\left[v_i Y\right]$, $a_{ij} = \mathbb{E}\left[v_i v_j Y\right]$. We also define $X_i^{(a)} = \mathbb{1}\left\{\lambda_i = a\right\}$ and $M(X^{(a)}, X^{(b)})_{ij} = \mathbb{E}\left[X_i^{(a)} X_j^{(b)}\right] = P(\lambda_i = a, \lambda_j = b)$. For each term of r_i, we have a corresponding estimation error to compute.

- $P(\lambda_i = 1)$: We need to compute $\hat{P}(\lambda_i = 1) - P(\lambda_i = 1)$ over all $(i, j) \in E$, so the sampling error for this term is $\sqrt{\sum_{(i,j)\in E}(\hat{P}(\lambda_i = 1) - P(\lambda_i = 1))^2} \leq \sqrt{\sum_{i=1}^{m} m(\hat{P}(\lambda_i = 1) - P(\lambda_i = 1))^2} = \sqrt{m}\|\hat{p}(1) - p(1)\|_2$.

- $P(\lambda_i = 0)$: The sampling error is equivalent to $\sqrt{m}\|\hat{p}(0) - p(0)\|_2$.

- $P(\lambda_j = 1)$: The sampling error is equivalent to $\sqrt{m}\|\hat{p}(1) - p(1)\|_2$.

- $P(\lambda_j = 0)$: The sampling error is equivalent to $\sqrt{m}\|\hat{p}(0) - p(0)\|_2$.

- $P(\lambda_i\lambda_j = 1)$: This probability can be written as $P(\lambda_i = 1, \lambda_j = 1) + P(\lambda_i = -1, \lambda_j = -1)$, so we would need to compute $\hat{P}(\lambda_i = 1, \lambda_j = 1) - P(\lambda_i = 1, \lambda_j = 1) + \hat{P}(\lambda_i = -1, \lambda_j = -1) - P(\lambda_i = -1, \lambda_j = -1)$. Then the sampling error is equivalent to $\|\hat{M}(X^{(1)}, X^{(1)}) - M(X^{(1)}, X^{(1)}) + \hat{M}(X^{(-1)}, X^{(-1)}) - M(X^{(-1)}, X^{(-1)})\|_{ij}$.

- $P(\lambda_i = 0, \lambda_j = 1)$: Using the definition of M, the sampling error over all $(i, j) \in E$ for this is $\|\hat{M}(X^{(0)}, X^{(1)}) - M(X^{(0)}, X^{(1)})\|_{ij}$.

- $P(\lambda_i = 1, \lambda_j = 0)$: Similarly, the sampling error is $\|\hat{M}(X^{(1)}, X^{(0)}) - M(X^{(1)}, X^{(0)})\|_{ij}$.

- $P(\lambda_i = 0, \lambda_j = 0)$: Similarly, the sampling error is $\|\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)})\|_{ij}$.

- $P(\lambda_i Y = 1)$: Similar to before, the sampling error is $\frac{1}{2}\sqrt{m}\left(\|\hat{a} - a\|_2 + \|\hat{p}(0) - p(0)\|_2\right)$.

- $P(\lambda_i = 0, Y = 1)$: Similar to our estimate of $\boldsymbol{\mu_i}$, the sampling error is $\sqrt{m}\|\hat{p}(0) - p(0)\|_2$.

- $P(\lambda_j Y = 1)$: The sampling error is $\frac{1}{2}\sqrt{m}\left(\|\hat{a} - a\|_2 + \|\hat{p}(0) - p(0)\|_2\right)$.

- $P(\lambda_j = 0, Y = 1)$: The sampling error is $\sqrt{m}\|\hat{p}(0) - p(0)\|_2$.

- $P(\lambda_i\lambda_j Y = 1)$: Note that $\mathbb{E}\left[\lambda_i\lambda_j Y\right] = 2P(\lambda_i\lambda_j Y = 1) + P(\lambda_i\lambda_j = 0) - 1$. Moreover, $\mathbb{E}\left[\lambda_i\lambda_j Y\right]$ can be expressed as $\mathbb{E}\left[Y\right] \cdot \mathbb{E}\left[\lambda_i\lambda_j\right]$. Then the sampling error over all $\hat{P}(\lambda_i\lambda_j Y = 1) - P(\lambda_i\lambda_j Y = 1)$ is at least $\frac{1}{2}\|\mathbb{E}\left[Y\right](\hat{\mathbb{E}}\left[\lambda_i\lambda_j\right] - \mathbb{E}\left[\lambda_i\lambda_j\right]) - (\hat{P}(\lambda_i\lambda_j = 0) - P(\lambda_i\lambda_j = 0))\|_{ij}$. Furthermore, we can write $P(\lambda_i\lambda_j = 0)$ as $P(\lambda_i = 0) + P(\lambda_j = 0) - P(\lambda_i = 0, \lambda_j = 0)$, so our sampling error is now less than $\frac{1}{2}\|\hat{M}(\lambda, \lambda) - M(\lambda, \lambda)\|_{ij} + \frac{1}{2}\sqrt{m}\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}\sqrt{m}\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}\|\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)})\|_{ij}$.

- $P(\lambda_i = 0, \lambda_j Y = 1)$: Note that this can be written as $\frac{1}{2}\left(P(\lambda_i = 0) + \mathbb{E}\left[\lambda_j Y|\lambda_i = 0\right]P(\lambda_i = 0) - P(\lambda_i = 0, \lambda_j = 0)\right)$. Then the sampling error over all $\hat{P}(\lambda_i = 0, \lambda_j Y = 1) - P(\lambda_i = 0, \lambda_j Y = 1)$ is equivalent to

$$\begin{aligned}
&\frac{1}{2}\sqrt{m}\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}\|\hat{\mathbb{E}}\left[\lambda_j Y|\lambda_i = 0\right]\hat{P}(\lambda_i = 0) - \mathbb{E}\left[\lambda_j Y|\lambda_i = 0\right]P(\lambda_i = 0) \\
&\quad - (\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)}))\|_{ij} \\
&= \frac{1}{2}\sqrt{m}\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}\|\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)})\|_{ij} + \frac{1}{2}\|\hat{\mathbb{E}}\left[\lambda_j Y|\lambda_i = 0\right](\hat{P}(\lambda_i = 0) - P(\lambda_i = 0)) \\
&\quad - (\mathbb{E}\left[\lambda_j Y|\lambda_i = 0\right] - \hat{\mathbb{E}}\left[\lambda_j Y|\lambda_i = 0\right])P(\lambda_i = 0)\|_{ij} \\
&\leq \frac{\sqrt{m}}{2}\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}\|\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)})\|_{ij} + \frac{\sqrt{m}}{2}\|\hat{p}(0) - p(0)\|_2 \\
&\quad + \frac{1}{2}\|\mathbb{E}\left[\lambda_j Y|\lambda_i = 0\right] - \hat{\mathbb{E}}\left[\lambda_j Y|\lambda_i = 0\right]\|_{ij} \\
&= \sqrt{m}\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}\|\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)})\|_{ij} + \frac{1}{2}\|\mathbb{E}\left[\lambda_j Y|\lambda_i = 0\right] - \hat{\mathbb{E}}\left[\lambda_j Y|\lambda_i = 0\right]\|_{ij}
\end{aligned}$$

- $P(\lambda_j = 0, \lambda_i Y = 1)$: Symmetric to the previous case, the sampling error is $\sqrt{m}\|\hat{p}(0) - p(0)\|_2 + \frac{1}{2}\|\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)})\|_{ij} + \frac{1}{2}\|\mathbb{E}\left[\lambda_j Y|\lambda_i = 0\right] - \hat{\mathbb{E}}\left[\lambda_j Y|\lambda_i = 0\right]\|_{ij}$.

- $P(\lambda_i = 0, \lambda_j = 0, Y = 1)$: This expression is equal to $P(\lambda_i = 0, \lambda_j = 0)P(Y = 1)$, so the sampling error is $P(Y = 1)\|\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)})\|_{ij} \leq \|\hat{M}(X^{(0)}, X^{(0)}) - M(X^{(0)}, X^{(0)})\|_{ij}$.

After combining terms and taking the expectation, we have that

$$\begin{aligned}\mathbb{E}\left[\|\hat{\psi}-\psi\|_2\right] &\leq 2\sqrt{2m}\Delta_p + 2\Delta_M + 3\Delta_M + \frac{1}{\sqrt{2}}(\sqrt{m}\Delta_a + \sqrt{m}\Delta_p) + \sqrt{2m}\Delta_p + \frac{1}{2}(\Delta_M + 2\sqrt{m}\Delta_p + \Delta_M) \\ &+ \frac{1}{\sqrt{2}}(2\sqrt{m}\Delta_p + \|\hat{\mathbb{E}}\left[\lambda_i Y|\lambda_j = 0\right] - \mathbb{E}\left[\lambda_i Y|\lambda_j = 0\right]\|_{ij} + \Delta_M) + \Delta_M \\ &= \left(7 + \frac{1}{\sqrt{2}}\right)\Delta_M + \left(\frac{9}{2}\sqrt{2m} + \sqrt{m}\right)\Delta_p + \sqrt{\frac{m}{2}}\Delta_a + \frac{1}{\sqrt{2}}\|\hat{\mathbb{E}}\left[\lambda_i Y|\lambda_j = 0\right] - \mathbb{E}\left[\lambda_i Y|\lambda_j = 0\right]\|_{ij}.\end{aligned}$$

For $\mathbb{E}\left[\lambda_i Y|\lambda_j = 0\right]$, this term is equal to 0 when no sources can abstain. Otherwise, suppose that among the sources that do abstain, each label abstains with frequency at least r. Then $\|\hat{\mathbb{E}}\left[\lambda_i Y|\lambda_j = 0\right] - \mathbb{E}\left[\lambda_i Y|\lambda_j = 0\right]\|_{ij} \leq \sqrt{m} \cdot \frac{C_a}{a^5_{\text{min}}}\sqrt{\frac{m}{rn}}$ since there are rn samples used to produce the estimate. Using Lemma 5, 6, and 7, we now get that

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}_{ij} - \boldsymbol{\mu}_{ij}\|_2\right] \leq \|A_2^{-1}\|\left(\left(7 + \frac{1}{\sqrt{2}}\right)C_m\frac{m}{\sqrt{n}} + \left(\frac{9\sqrt{2}}{2} + 1\right)\frac{m}{\sqrt{n}} + \frac{C_a}{a^5_{|min|}} \cdot \frac{m}{\sqrt{n}}\left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2r}}\right)\right).$$

Finally, we can compute $\|A_1^{-1}\|$ and $\|A_2^{-1}\|$ since both matrices are constants, so the total estimation error is

$$\begin{aligned}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2\right] \leq& 3.19\left(\left(\frac{3}{2} + \sqrt{2}\right)\sqrt{\frac{m}{n}} + \frac{C_a}{2a^5_{|min|}}\sqrt{\frac{m}{n}}\right) + \\ & 6.35\left(\left(7 + \frac{1}{\sqrt{2}}\right)C_m\frac{m}{\sqrt{n}} + \left(\frac{9\sqrt{2}}{2} + 1\right)\frac{m}{\sqrt{n}} + \frac{C_a}{a^5_{|min|}} \cdot \frac{m}{\sqrt{n}}\left(\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2r}}\right)\right).\end{aligned}$$

D.2. Proof of Theorem 2 (Information Theoretical Lower Bound)

For Theorem 2 and Theorem 3, we will need the following lemma.

Lemma 8. *Let θ_1 and θ_2 be two sets of canonical parameters for an exponential family model, and let μ_1 and μ_2 be the respective mean parameters. If we define e_{min} to be the smallest eigenvalue of the covariance matrix Σ for the random variables in the graphical model,*

$$\|\theta_1 - \theta_2\| \leq \frac{1}{e_{min}}\|\mu_1 - \mu_2\|$$

.

Proof. Let $A(\theta)$ be the log partition function. Now, recall that the Hessian $\nabla^2 A(\theta)$ is equal to Σ above. Next, since e_{min} is the smallest eigenvalue, $\nabla^2 A(\theta) - e_{min}I = \Sigma - e_{min}I$ is positive semi-definite, so $A(\theta)$ is strongly convex with parameter e_{min}.

Note that since $A(\cdot)$ is strongly convex with parameter e_{min}, then $A^*(\cdot)$, its Fenchel dual, has Lipchitz continuous gradients with parameter $\frac{1}{e_{min}}$ (Zhou, 2018). This means that

$$\|\nabla A^*(\mu_1) - \nabla A^*(\mu_2)\| \leq \frac{1}{e_{min}}\|\mu_1 - \mu_2\|.$$

But $\nabla A^*(\mu)$ is the inverse mapping from mean parameters to canonical parameters, so this is just

$$\|\theta_1 - \theta_2\| \leq \frac{1}{e_{min}}\|\mu_1 - \mu_2\|$$

. □

Now, we provide the proof for Theorem 2. Consider the following family of distributions for a graphical model with one hidden variable Y, m observed variables that are all conditionally independent given Y, and no sources abstaining:

$$\mathcal{P} = \left\{P = \frac{1}{z}\exp(\theta_Y Y + \sum_{j=1}^{m}\theta_j\lambda_j Y) : \theta \in \mathbb{R}^{m+1}\right\}$$

We define a set of canonical parameters $\theta_v = \delta v$, where $\delta > 0$, $v \in \{-1, 1\}^m$ (θ_Y is fixed since it maps to a known mean parameter), and P_v is the corresponding distribution in $\mathcal{P}$. $\mathcal{P}$ induces a $\frac{\delta}{\sqrt{m}}$-Hamming separation for the L2 loss because

$$\begin{aligned}\|\theta - \theta_v\|_2 = \Big(\sum_{j=1}^{m} |\theta_j - [\theta_v]_j|^2\Big)^{1/2} &\geq \frac{\sum_{j=1}^{m} 1 \cdot |\theta_j - [\theta_v]|_j}{\big(\sum_{j=1}^{m} 1^2\big)^{1/2}} \\ &= \frac{1}{\sqrt{m}} \sum_{j=1}^{m} |\theta_j - [\theta_v]_j| \geq \frac{\delta}{\sqrt{m}} \sum_{j=1}^{m} \mathbf{1}\{\text{sign}(\theta_j) \neq v_j\}.\end{aligned}$$

We use Cauchy-Schwarz inequality in the first line and the fact that if the sign of θ_j is different from v_j, then θ_j and $[\theta_v]_j$ must be at least δ apart. Then applying Assouad's Lemma (Yu, 1997), the minimax risk is bounded by

$$\mathcal{M}_n(\theta(\mathcal{P}), L2) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\|\hat{\theta}(X_1, \ldots, X_n) - \theta(P)\|_2] \geq \frac{\delta}{2\sqrt{m}} \sum_{j=1}^{m} 1 - \|P^n_{+j} - P^n_{-j}\|_{TV}.$$

$\hat{\theta}(X_1, \ldots, X_n)$ is an estimate of θ based on the n observable data points, while $\theta(P)$ is the canonical parameters of a distribution P. $P^n_{\pm j} = \frac{1}{2^{m-1}} \sum_v P^n_{v,\pm j}$, where $P^n_{v,\pm j}$ is the product of n distributions parametrized by θ_v with $v_j = \pm 1$. We use the convexity of total variation distance, Pinsker's inequality, and decoupling of KL-divergence to get

$$\|P^n_{+j} - P^n_{-j}\|^2_{TV} \leq \max_{d_{ham}(v,v') \leq 1} \|P^n_v - P^n_{v'}\|^2_{TV} \leq \frac{1}{2} \max_{d_{ham}(v,v') \leq 1} KL(P^n_v \| P^n_{v'}) = \frac{n}{2} \max_{d_{ham}(v,v') \leq 1} KL(P_v \| P_{v'}).$$

v and v' above only differ in one term. Then our lower bound becomes

$$\mathcal{M}_n(\theta(\mathcal{P}), L2) \geq \frac{\delta}{2\sqrt{m}} \sum_{j=1}^{m} 1 - \sqrt{\frac{n}{2} \max_{d_{ham}(v,v') \leq 1} KL(P_v \| P_{v'})} = \frac{\delta\sqrt{m}}{2} \left(1 - \sqrt{\frac{n}{2} \max_{d_{ham}(v,v') \leq 1} KL(P_v \| P_{v'})}\right). \quad (13)$$

We must bound the KL-divergence between P_v and P'_v. Suppose WLOG that v and v' differ at the ith index with $v_i = 1, v'_i = -1$, and let z_v and $z_{v'}$ be the respective terms used to normalize the distributions. Then the KL divergence is

$$KL(P_v \| P_{v'}) = \mathbb{E}_v[\langle \theta_v - \theta_{v'}, \lambda Y \rangle] + \ln \frac{z_{v'}}{z_v} = 2\delta \mathbb{E}_v[\lambda_i Y] + \ln \frac{z_{v'}}{z_v}. \quad (14)$$

We can write an expression for $\mathbb{E}_v[\lambda_i Y]$:

$$\begin{aligned}\mathbb{E}_v[\lambda_i Y] &= 2(P_v(\lambda_i = 1, Y = 1) + P_v(\lambda_i = -1, Y = -1)) - 1 \\ &= \frac{2}{z_v}\Big(\sum_{\lambda_{\neg i}} \exp(\theta_Y + \delta + \sum_{j \neq i}^{m} (\delta v_j)\lambda_j) + \exp(-\theta_Y + \delta - \sum_{j \neq i}^{m} (\delta v_j)\lambda_j)\Big) - 1 \\ &= \frac{2}{z_v} \exp(\delta) \sum_{\lambda_{\neg i}} 2\cosh(\theta_Y + \sum_{j \neq i}^{m} (\delta v_j)\lambda_j) - 1. \quad (15)\end{aligned}$$

Similarly, z_v and $z_{v'}$ can be written as

$$\begin{aligned}z_v &= \exp(\delta) \sum_{\lambda_{\neg i}} 2\cosh(\theta_Y + \sum_{j \neq i}^{m} (\delta v_j)\lambda_j) + \sum_{\lambda_{\neg i}} \exp(\theta_Y - \delta + \sum_{j \neq i} (\delta v_j)\lambda_j) + \sum_{\lambda_{\neg i}} \exp(-\theta_Y - \delta - \sum_{j \neq i} (\delta v_j)) \\ &= (\exp(\delta) + \exp(-\delta)) \sum_{\lambda_{\neg i}} 2\cosh(\theta_Y + \sum_{j \neq i} (\delta v_j)\lambda_j) = 4\cosh(\delta) \sum_{\lambda_{\neg i}} \cosh(\theta_Y + \sum_{j \neq i} (\delta v_j)\lambda_j) \\ z_{v'} &= 4\cosh(\delta) \sum_{\lambda_{\neg i}} \cosh(\theta_Y + \sum_{j \neq i} (\delta v'_j)\lambda_j)\end{aligned}$$

Plugging z_v back into (15), we get:

$$\mathbb{E}_v\left[\lambda_i Y\right] = 4 \cdot \frac{\exp(\delta)\sum_{\lambda_{\neg i}}\cosh(\theta_Y + \sum_{j\neq i}^{m}(\delta v_j)\lambda_j)}{4\cosh(\delta)\sum_{\lambda_{\neg i}}\cosh(\theta_Y + \sum_{j\neq i}^{m}(\delta v_j)\lambda_j)} - 1 = \frac{\exp(\delta)}{\cosh(\delta)} - 1.$$

Also note that $\frac{z_{v'}}{z_v} = 1$ since $v'_j = v_j$ for all $j \neq i$. The KL-divergence expression (14) now becomes

$$KL(P_v \| P_{v'}) = 2\delta\left(\frac{\exp(\delta)}{\cosh(\delta)} - 1\right) + \ln(1) = 2\delta\left(\frac{\exp(\delta)}{\cosh(\delta)} - 1\right).$$

We finally show that this expression is less than $2\delta^2$. Note that for positive δ, $f(\delta) = \frac{\exp(\delta)}{\cosh(\delta)} - 1 < \delta$, because $f(\delta)$ is concave and $f'(0) = 1$. Then we clearly have that $KL(P_v \| P_{v'}) \leq 2\delta^2$. Putting this back into our expression for the minimax risk, (13) becomes

$$\mathcal{M}_n(\theta(\mathcal{P}), L2) \geq \frac{\delta\sqrt{m}}{2}(1 - \sqrt{n\delta^2}).$$

Then if we set $\delta = \frac{1}{2\sqrt{n}}$, we get that

$$\mathcal{M}_n(\theta(\mathcal{P}), L2) \geq \frac{\sqrt{m}}{8\sqrt{n}}.$$

Lastly, to convert to a bound over the mean parameters, we use Lemma 8 to conclude that

$$\inf_{\hat{\mu}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\|\hat{\mu}(X_1, \ldots, X_n) - \mu(P)\|_2\right] \geq \frac{e_{min}}{8}\sqrt{\frac{m}{n}}.$$

From this, we can conclude that the estimation error on the label model parameters $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2$ is also at least $\frac{e_{min}}{8}\sqrt{\frac{m}{n}}$.

D.3. Proof of Theorem 3 (Generalization Error)

We base our proof off of Theorem 1 of Ratner et al. (2019) with modifications to account for model misspecification. To learn the parametrization of our end model f_w, we want to minimize a loss function $L(w, \boldsymbol{X}, \boldsymbol{Y}) \in [0, 1]$. The expected loss we would normally minimize using some $w^* = \operatorname{argmin}_w L(w)$ is

$$L(w) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{D}}\left[L(w, \boldsymbol{X}, \boldsymbol{Y})\right].$$

However, since we do not have access to the true labels $\boldsymbol{Y}$, we instead minimize the expected noise-aware loss. Recall that $\boldsymbol{\mu}$ is the parametrization of the label model we would learn with population-level statistics, and $\hat{\boldsymbol{\mu}}$ is the parametrization we learn with the empirical estimates from our data. Denote $P_{\boldsymbol{\mu}}$ and $P_{\hat{\boldsymbol{\mu}}}$ as the respective distributions. If we were to have a population-level estimate of $\boldsymbol{\mu}$, the loss to minimize would be

$$L_{\boldsymbol{\mu}}(w) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{D}}\left[\mathbb{E}_{\widetilde{\boldsymbol{Y}}\sim P_{\boldsymbol{\mu}}(\cdot|\boldsymbol{\lambda}(\boldsymbol{X}))}\left[L(w, \boldsymbol{X}, \widetilde{\boldsymbol{Y}})\right]\right].$$

However, because we must estimate $\hat{\boldsymbol{\mu}}$ and further are minimizing loss over n samples, we want to estimate a $\hat{w}$ that minimizes the empirical loss,

$$\hat{L}_{\hat{\boldsymbol{\mu}}}(w) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{\boldsymbol{Y}}\sim P_{\hat{\boldsymbol{\mu}}}(\cdot|\boldsymbol{\lambda}(\boldsymbol{X_i}))}\left[L(w, \boldsymbol{X_i}, \widetilde{\boldsymbol{Y}})\right].$$

We first write $L(w)$ in terms of $L_{\boldsymbol{\mu}}(w)$.

$$\begin{aligned}
L(w) &= \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{D}}\left[L(w,\boldsymbol{X},\boldsymbol{Y})\right] = \mathbb{E}_{(\boldsymbol{X'},\boldsymbol{Y'})\sim D}\left[\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim D}\left[L(w,\boldsymbol{X'},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{X'}\right]\right] \\
&= \mathbb{E}_{(\boldsymbol{X'},\boldsymbol{Y'})\sim D}\Big[\mathbb{E}_{(\boldsymbol{X},\widetilde{\boldsymbol{Y}})\sim P_{\boldsymbol{\mu}}}\left[L(w,\boldsymbol{X'},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{X'}\right] + \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{D}}\left[L(w,\boldsymbol{X'},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{X'}\right] \\
&- \mathbb{E}_{(\boldsymbol{X},\widetilde{\boldsymbol{Y}})\sim P_{\boldsymbol{\mu}}}\left[L(w,\boldsymbol{X'},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{X'}\right]\Big] \\
&\leq \mathbb{E}_{(\boldsymbol{X'},\boldsymbol{Y'})\sim\mathcal{D}}\left[\mathbb{E}_{(\boldsymbol{\lambda},\widetilde{\boldsymbol{Y}})\sim P_{\boldsymbol{\mu}}}\left[L(w,\boldsymbol{X'},\boldsymbol{Y})|\boldsymbol{\lambda}=\boldsymbol{\lambda}')\right]\right] \\
&+ \mathbb{E}_{(\boldsymbol{X'},\boldsymbol{Y'})\sim\mathcal{D}}\left[\Big|\sum_{x,y} L(w,\boldsymbol{X'},y)(\mathcal{D}(\boldsymbol{X}=x,\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'}) - P_{\boldsymbol{\mu}}(\boldsymbol{X}=x,\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'}))\Big|\right] \\
&\leq L_{\boldsymbol{\mu}}(w) + \mathbb{E}_{(\boldsymbol{X'},\boldsymbol{Y'})\sim\mathcal{D}}\Big[\sum_{x,y} L(w,\boldsymbol{X'},y)\cdot\big|\mathcal{D}(\boldsymbol{X}=x,\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'}) - P_{\boldsymbol{\mu}}(\boldsymbol{X}=x,\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'})\big|\Big] \\
&\leq L_{\boldsymbol{\mu}}(w) + \mathbb{E}_{(\boldsymbol{X'},\boldsymbol{Y'})\sim\mathcal{D}}\Big[\sum_{x,y}\big|\mathcal{D}(\boldsymbol{X}=x,\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'}) - P_{\mu}(\boldsymbol{X}=x,\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'})\big|\Big]
\end{aligned}$$

Here we have used the fact that $L(w,\boldsymbol{X'},y)\leq 1$. Note that $\mathcal{D}(\boldsymbol{X}=x,\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'}) = \mathcal{D}(\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'})$ only when $\boldsymbol{X'}=x$, and is 0 otherwise. The same holds for $P_{\boldsymbol{\mu}}$, so

$$L(w) \leq L_{\boldsymbol{\mu}}(w) + \mathbb{E}_{(\boldsymbol{X'},\boldsymbol{Y'})\sim\mathcal{D}}\left[\sum_{y}\big|\mathcal{D}(\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'}) - P_{\boldsymbol{\mu}}(\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'})\big|\right].$$

Note that the expression $\sum_y \big|\mathcal{D}(\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'}) - P_{\boldsymbol{\mu}}(\boldsymbol{Y}=y|\boldsymbol{X}=\boldsymbol{X'})\big|$ is just half the total variation distance between $\mathcal{D}(\boldsymbol{Y}|\boldsymbol{X'})$ and $P_{\boldsymbol{\mu}}(\boldsymbol{Y}|\boldsymbol{X'})$. Then, using Pinsker's inequality, we bound $L(w)$ in terms of the conditional KL divergence between $\mathcal{D}$ and P_{μ}:

$$\begin{aligned}
L(w) &\leq L_{\boldsymbol{\mu}}(w) + \mathbb{E}_{\boldsymbol{X'}\sim\mathcal{D}}\left[2\cdot TV(\mathcal{D}(\boldsymbol{Y}|\boldsymbol{X'}), P_{\boldsymbol{\mu}}(\boldsymbol{Y}|\boldsymbol{X'}))\right] \\
&\leq L_{\boldsymbol{\mu}}(w) + 2\cdot\mathbb{E}_{\boldsymbol{X}\sim\mathcal{D}}\left[\sqrt{(1/2)KL(\mathcal{D}(\boldsymbol{Y}|\boldsymbol{X})\parallel P_{\boldsymbol{\mu}}(\boldsymbol{Y}|\boldsymbol{X}))}\right] \\
&\leq L_{\boldsymbol{\mu}}(w) + \sqrt{2\cdot KL(\mathcal{D}(\boldsymbol{Y}|\boldsymbol{X})\parallel P_{\boldsymbol{\mu}}(\boldsymbol{Y}|\boldsymbol{X}))}.
\end{aligned}$$

There is a similar lower bound on $L(w)$ if we perform the same steps as above on the inequality $L(w)\geq L_{\boldsymbol{\mu}}(w) - \mathbb{E}_{(\boldsymbol{X'},\boldsymbol{Y'})\sim\mathcal{D}}\left[\Big|\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{D}}\left[L(w,\boldsymbol{X'},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{X'}\right] - \mathbb{E}_{(\boldsymbol{X},\widetilde{\boldsymbol{Y}})\sim P_{\boldsymbol{\mu}}}\left[L(w,\boldsymbol{X'},\boldsymbol{Y})|\boldsymbol{X}=\boldsymbol{X'}\right]\Big|\right]$. This yields

$$L(w) \geq L_{\boldsymbol{\mu}}(w) - \sqrt{2\cdot KL(\mathcal{D}(\boldsymbol{Y}|\boldsymbol{X})\parallel P_{\boldsymbol{\mu}}(\boldsymbol{Y}|\boldsymbol{X}))}.$$

Therefore,

$$L(\hat{w}) - L(w^*) \leq L_{\boldsymbol{\mu}}(\hat{w}) - L_{\boldsymbol{\mu}}(w^*) + 2\sqrt{2\cdot KL(\mathcal{D}(\boldsymbol{Y}|\boldsymbol{X})\parallel P_{\boldsymbol{\mu}}(\boldsymbol{Y}|\boldsymbol{X}))}.$$

We finish the proof of the generalization bound with the procedure from Ratner et al. (2019) but also use the conversion from canonical parameters to mean parameters as stated in Lemma 8, and note that the estimation error of the mean parameters is always less than the estimation error of the label model parameters. Then our final generalization result is

$$L(\hat{w}) - L(w^*) \leq \gamma(n) + \frac{8|\mathcal{Y}|}{e_{min}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 + \delta(\mathcal{D}, P_{\boldsymbol{\mu}}),$$

where $\delta(\mathcal{D}, P_{\boldsymbol{\mu}}) = 2\sqrt{2\cdot KL(\mathcal{D}(\boldsymbol{Y}|\boldsymbol{X})\parallel P_{\boldsymbol{\mu}}(\boldsymbol{Y}|\boldsymbol{X}))}$, e_{min} is the minimum eigenvalue of **Cov** $[\boldsymbol{\lambda}, \boldsymbol{Y}]$ over the construction of the binary Ising model, and $\gamma(n)$ bounds the empirical risk minimization error.

E. Extended Experimental Details

We describe additional details about the tasks, including details about data sources, supervision sources, and end models. We also report details about our ablation studies. All timing measurements were taken on a machine with an Intel Xeon E5-2690 v4 CPU and Tesla P100-PCIE-16GB GPU. Details about the sizes of the train/dev/test splits and end models are shown in Table 2.

E.1. Dataset Details

Dataset	**End Model**	N_{train}	N_{dev}	N_{test}
Spouse	LSTM	22,254	2,811	2,701
Spam	Logistic Regression	1,586	120	250
Weather	Logistic Regression	187	50	50
Commercial	ResNet-50	64,130	9,479	7,496
Interview	ResNet-50	6,835	3,026	3,563
Tennis Rally	ResNet-50	6,959	746	1,098
Basketball	ResNet-18	3,594	212	244

Table 2. We report the train/dev/test split of each dataset. The dev and test set have ground truth labels, and we assign labels to the training set using our method or one of the baseline methods.

Spouse, Weather We use the datasets from Ratner et al. (2018) and the train/dev/test splits from that work (**Weather** is called **Crowd** in that work).

Spam We use the dataset as provided by Snorkel[1] and those train/dev/test splits.

Interview, Basketball We use the datasets from Sala et al. (2019) and the train/dev/test splits from that work.

Commercial We use the dataset from Fu et al. (2019) and the train/dev/test splits from that work.

Tennis Rally We obtained broadcast footage from four professional tennis matches, and annotated segments when the two players are in a rally. We temporally downsampled the images at 1 FPS. We split into dev/test by taking segments from each match (using contiguous segments for dev and test, respectively) to ensure that dev and test come from the same distribution.

E.2. Task-Specific End Models

For the datasets we draw from previous work (each dataset except for **Tennis Rally**), we use the previously published end model architectures (LSTM (Hochreiter & Schmidhuber, 1997) for **Spouse**, logistic regression over bag of n-grams for **Spam** and over Bert features for **Weather** (Devlin et al., 2018), ResNet pre-trained on ImageNet for the video tasks). For **Tennis Rally**, we use ResNet-50 pre-trained on ImageNet to classify individual frames. We do not claim that these end models achieve the best possible performance for each task; our goal is the compare the relative imporovements that our weak supervision models provide compare to other baselines through label quality, which is orthogonal to achieving state-of-the-art performance for these specific tasks.

For end models that come from previous works, we use the hyperparameters from those works. For the label model baselines, we use the hyperparameters from previous works as well. For our label model, we use class balance from the dev set, or tune the class balance ourselves with a grid search. We also tune which triplets we use for parameter recovery on the dev set. For our end model parameters, we either use the hyperparameters from previous works, or run a simple grid search over learning rate and momentum.

[1] https://www.snorkel.org/use-cases/01-spam-tutorial

	Spouse	Spam	Weather
Random abstains	20.9	64.1	69.1
FlyingSquid	49.6	92.3	88.9
Single Triplet Worst	4.5	67.0	0.0
Single Triplet Best	51.2	83.6	77.6
Single Triplet Average	37.9	73.4	31.0
FlyingSquidLabel Model	47.0	89.1	77.6

Table 3. End model performance in terms of F1 score with random votes replacing abstentions (first row), compared to FlyingSquid, for the benchmark applications.

E.3. Supervision Sources

Supervision sources are expressed as short Python functions. Each source relied on different information to assign noisy labels:

Spouse, Weather, Spam For these tasks, we used the same supervision sources as used in previous work (Ratner et al., 2018). These are all text classification tasks, so they rely on text-based heuristics such as the presence or absence of certain words, or particular regex patterns.

Interview, Basketball Again, we use sources from previous work (Sala et al., 2019). For **Interview**, these sources rely on the presence of certain faces in the frame, as determined by an identity classifier, or certain text in the transcript. For **Basketball**, these sources rely on an off-the-shelf object detector to detect balls or people, and use heuristics based on the average pixel of the detected ball or distance between the ball and person to determine whether the sport being played is basketball or not.

Commercial In this dataset, there is a strong signal for the presence or absence of commercials in pixel histograms and the text; in particular, commercials are book-ended on either side by sequences of black frames, and commercial segments tend to have mixed-case or missing transcripts (whereas news segments are in all caps). We use these signals to build the weak supervision sources.

Tennis Rally This dataset uses an off-the-shelf pose detector to provide primitives for the weak supervision sources. The supervision sources are heuristics based on the number of people on court and their positions. Additional supervision sources use color histograms of the frames (i.e., how green the frame is, or whether there are enough white pixels for the court markings to be shown).

E.4. Ablation Studies

We report the results of two ablation studies on the benchmark applications. In the first study, we examine the effect of randomly replacing abstains with votes, instead of augmenting G_{dep}. In the second study, we examine the effect of using a single random selection of triplets instead of taking the mean or median over all triplet assignments.

Table 3 (top) shows end model performance for the three benchmark tasks when replacing abstains with random votes (top row), compared to FlyingSquid end model performance. Replacing abstentions with random votes results in a major degradation in performance.

Table 3 (bottom) shows label model performance when using a single random assignment of triplets, compared to the FlyingSquid label model, which takes the median or mean of all possible triplets. There is large variance when taking a single random assignment of triplets, whereas using an aggregation is more stable. In particular, while selecting a good seed can result in performance that matches (**Weather**) or exceeds (**Spouse**) FlyingSquid label model performance, selecting a *bad* seed result in much worse performance (including catastrophically bad predictors). As a result, FlyingSquid outperforms random assignments on average.

As a final note, we comment on using means vs. medians for aggregating accuracy scores. For all tasks except for **Weather**, there is no difference in label model performance. For **Weather**, using medians is more accurate, since the supervision sources have a large abstention rate. As a result, many triplets result in accuracy scores of zero (hence the 0 F1 score in Table 3). This throws off the median aggregation, since the median accuracy score becomes zero for many sources. However, mean aggregation is more robust to these zero's, since the positive accuracy scores from the triplets can correct for the accuracy.

References

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

Bunea, F. and Xiao, L. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. *Bernoulli*, 21(5):1200–1230, 2015.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Fu, D. Y., Crichton, W., Hong, J., Yao, X., Zhang, H., Truong, A., Narayan, A., Agrawala, M., Ré, C., and Fatahalian, K. Rekall: Specifying video events using compositions of spatiotemporal labels. *arXiv preprint arXiv:1910.02993*, 2019.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

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

Long, P. M. The complexity of learning according to two models of a drifting environment. *Machine Learning*, 37(3): 337–354, Dec 1999. ISSN 1573-0565. doi: 10.1023/A:1007666507971. URL https://doi.org/10.1023/A:1007666507971.

Raghunathan, A., Frostig, R., Duchi, J., and Liang, P. Estimation from indirect supervision with linear moments. In *International conference on machine learning*, pp. 2568–2577, 2016.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the 44th International Conference on Very Large Data Bases (VLDB)*, Rio de Janeiro, Brazil, 2018.

Ratner, A. J., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., and Ré, C. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, 2019.

Sala, F., Varma, P., Fries, J., Fu, D. Y., Sagawa, S., Khattar, S., Ramamoorthy, A., Xiao, K., Fatahalian, K., Priest, J., and Ré, C. Multi-resolution weak supervision for sequential data. In *Advances in Neural Information Processing Systems 32*, pp. 192–203, 2019.

Yu, B. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.

Zhou, X. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.