Appendix

A Further discussion of related work

Computational Results for Markov Random Fields There is a long line of work on computational aspects of inference (e.g. MLE, MAP) in Markov Random Field models similar to Model 1 (Veksler, 1999; Boykov and Veksler, 2006; Komodakis and Tziritas, 2007; Schraudolph and Kamenetsky, 2009; Chandrasekaran et al., 2012). To our knowledge none of these results shed light on the statistical recovery rates that are attainable for this setting — computationally efficiently or not.

Censored Block Model A recent line of research has studied recovery under the so-called *censored block model* (CBM). In CBM, vertices are labeled by ±1 and for every edge uv, the number Y_uY_v is observed independently with probability $1-q$ (where Y_u, Y_v are the labels of the vertices). The goal is to find the true label Y_uY_v of each edge uv correctly with high probability (based on the noisy observations). For partial recovery in the censored block model we ask for a prediction whose correlation with the ground truth (up to sign) is constant strictly greater than $1/2$ as $n \to \infty$. For the Erdös-Rényi random graph model, $G(n, \alpha/n)$ both the threshold (how large α needs to be in terms of p) for partial Saade et al. (2015) and exact Abbe et al. (2014) recovery have been determined Exact recovery is obtained through maximum likelihood estimation which is generally intractable. The authors provide a polynomial time algorithm based on semidefinite programming that matches this threshold up to constant factors.

We observe that in our setting, due to the presence of side information, there is a simple and efficient algorithm that achieves exact recovery with high probability when the minimal degree is $\Omega(\log n)$: Theorem 6. Such exact recovery algorithms are known for CBM model only under additional spectral expansion conditions Abbe et al. (2014).

Recovery from Pairwise Measurements Chen and Goldsmith (2014) provide conditions on exact recovery in a censored block model-like setting which, like our own, considers structured classes of graphs. Motivated by applications in computational biology and social networks analysis, Chen et al. (2016) have recently considered *exact recovery* for edges in this setting. Like the present work, they consider sparse graphs with local structure such as grids and rings. Because their focus is *exact* recover and their model does not have side information, their results mainly apply to graphs of logarithmic degree and our incomparable to our own results. For example, on the ring lattice $R_{n,k}$ in Example 7 their exact recovery result requires $k = \Omega(\log(n))$, whereas our partial recover result concerns constant k.

Correlation Clustering Correlation clustering focuses on a combinatorial optimization problem closely related to the maximum likelihood estimation problem for our setting when we are only given edge labels. The main difference from our work is that the number of clusters is not predetermined. Most work on this setting has focused on obtaining approximation algorithms and has not considered any particular generative model for the weights (as in our case). An exception is Joachims and Hopcroft (2005), which gives partial recovery results in a model similar to the one we consider, in which a ground truth partition is fixed and the observed edge labels correspond to some noisy notion of similarity. However, these authors focus on the case where G is the complete graph.

Makarychev et al. (2015) consider correlation clustering where the model is a semi-random variant of the one we consider for the edge inference problem: Fix a graph $G = (V, E)$ and a vertex label Y. For each $uv \in E$, we observe X_{uv} where $X_{uv} = Y_uY_v$ with probability $1-p$ and has its value in selected by an adversary otherwise. They do not consider side information, nor are they interested in concrete structured classes of graphs like grids.

B Omitted proofs

B.1 Proofs from Section 2

Proof of Theorem 1. By the Bernstein inequality it holds that with probability at least $1-\delta/2$,

$$\sum_{(u,v)\in E} \mathbb{1}\{Y_u \neq X_{u,v}Y_v\} \leq 2pn + 2\log(2/\delta).$$

Thus, if we take $\mathcal{F} = \left\{\widehat{Y} : \sum_{(u,v)\in E} \mathbb{1}\{\widehat{Y}_u \neq X_{u,v}\widehat{Y}_v\} \leq 2pn + 2\log(2/\delta)\right\}$, then $Y \in \mathcal{F}$ with probability at least $1-\delta/2$.

Fix $\widehat{Y} \in \{\pm 1\}^V$. We can verify by substitution that for each $v \in V$,

$$\mathbb{1}\{\widehat{Y}_v \neq Y_v\} = \frac{1}{1-2q}\Big[\mathbb{P}_Z(\widehat{Y}_v \neq Z_v) - \mathbb{P}_Z(Y_v \neq Z_v)\Big].$$

This implies that when $Y \in \mathcal{F}$ we have the following relation for Hamming error:

$$\sum_{v\in V} \mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\} = \frac{1}{1-2q}\left[\sum_{v\in V}\mathbb{P}(\widehat{Y}_v \neq Z_v) - \min_{Y'\in\mathcal{F}}\sum_{v\in V}\mathbb{P}\left(Y'_v \neq Z_v\right)\right].$$

Corollary 2 now implies that if we take $\widehat{Y} = \arg\min_{Y'\in\mathcal{F}}\sum_{v\in V}\mathbb{1}\{Y'_v \neq Z_v\}$, which is precisely the solution to (2), then with probability at least $1-\delta/2$,

$$\sum_{v\in V}\mathbb{P}\left(\widehat{Y}_v \neq Z_v\right) - \min_{Y'\in\mathcal{F}}\sum_{v\in V}\mathbb{P}\left(Y'_v \neq Z_v\right) \leq \left(\frac{4}{3}+\frac{1}{\epsilon}\right)\log\left(\frac{2|\mathcal{F}|}{\delta}\right).$$

Using that $|\mathcal{F}| \leq \sum_{k=0}^{2pn+2\log(2/\delta)}\binom{n}{k} \leq (e/p)^{2pn+2\log(2/\delta)}$ and $\epsilon \leq 1/2$ we further have that the RHS is bounded as $\frac{2}{\epsilon}\log(2e/p\delta)(2pn+2\log(2/\delta)+1)$. Putting everything together (and recalling $1-2q=2\epsilon$), it holds that with probability at least $1-\delta$

$$\sum_{v\in V}\mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\} \leq \frac{1}{\epsilon^2}(2pn+2\log(2/\delta)+1)\log(2e/p\delta).$$

□

B.2 Proofs from Section 3

Proof of Theorem 3. The minimax value of the estimation problem is given by

$$\min_{\widehat{Y}}\max_{Y}\mathop{\mathbb{E}}_{X,Z|Y}\sum_{v\in V}\mathbb{1}\left\{\widehat{Y}_v(X,Z) \neq Y_v\right\}.$$

We can move to the following lower bound by considering a game where each vertex predictor $\widehat{Y}_v$ is given access to the true labels Y of all other vertices in G:

$$\min_{\left\{\widehat{Y}_v\right\}_{v\in V}}\max_{Y}\mathop{\mathbb{E}}_{X,Z|Y}\sum_{v\in V}\mathbb{1}\left\{\widehat{Y}_v(X,Z,Y_{V\setminus\{v\}}) \neq Y_v\right\}.$$

Under the new model, the minimax optimal predictor for a given node v is given by the MAP predictor:

$$\widehat{Y}_v = \mathop{\arg\min}_{\widehat{Y}\in\{\pm1\}}\log\left(\frac{1-q}{q}\right)\mathbb{1}\left\{\widehat{Y} \neq Z_v\right\} + \log\left(\frac{1-p}{p}\right)\sum_{u\in N_v}\mathbb{1}\left\{\widehat{Y} \neq Y_uX_{uv}\right\}.$$

When $p < q$, the minimax optimal estimator for v takes the majority of the predictions suggested by its edges (that is, $Y_u \cdot X_{uv}$ for each neighbor u) and uses the vertex observation Z_v to break ties.

When $\texttt{deg}(v)$ is odd, the majority will be wrong if at least $\lceil\texttt{deg}(v)\rceil$ of the edges in the neighbor of v are flipped, and will be correct otherwise. When $\texttt{deg}(v)$ is even there are two cases: 1) Strictly more than $\lceil\texttt{deg}(v)\rceil$ of the edges in $N(v)$ have been flipped, in which case the majority will be wrong. 2) Exactly half the edges are wrong, in which the optimal estimator will take the label Z_v as its prediction, which will be wrong with probability q.

We thus have

$$\begin{aligned}\mathbb{P}(\widehat{Y}_v \neq Y_v) &= \sum_{k=\lceil \mathsf{deg}(v)/2\rceil}^{\mathsf{deg}(v)} \binom{\mathsf{deg}(v)}{k} p^k (1-p)^{\mathsf{deg}(v)-k} \\ &\geq \binom{\mathsf{deg}(v)}{\lceil \mathsf{deg}(v)/2\rceil} p^{\lceil \mathsf{deg}(v)/2\rceil}(1-p)^{\mathsf{deg}(v)-k} \\ &\geq \left(\frac{\mathsf{deg}(v)}{\lceil \mathsf{deg}(v)/2\rceil}\right)^{\lceil \mathsf{deg}(v)/2\rceil} p^{\lceil \mathsf{deg}(v)/2\rceil}(1/2)^{\lceil \mathsf{deg}(v)/2\rceil} \\ &\geq \Omega(p^{\lceil \mathsf{deg}(v)/2\rceil}).\end{aligned}$$

In the last line we have used that we treat $\mathsf{deg}(v)$ as constant to suppress a weak dependence on it that arises when $\mathsf{deg}(v)$ is odd. Putting everything together, we see that in expectation we have the bound

$$\mathbb{E}\left[\sum_{v\in V} \mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\}\right] \geq \Omega\left(q \sum_{v\in V} p^{\lceil \mathsf{deg}(v)/2\rceil}\right).$$

□

Proof of Theorem 4. Recall that the minimax value of the estimation problem is given by

$$\min_{\widehat{Y}} \max_{Y} \mathop{\mathbb{E}}_{X,Z|Y} \sum_{v\in V} \mathbb{1}\left\{\widehat{Y}_v(X,Z) \neq Y_v\right\}.$$

As in the proof of Theorem 3, we will move to a lower bound where predictors are given access to extra data. In this case, we consider a set of disjoint predictors $\left\{\widehat{Y}^W\right\}$, one for each component $W \in \mathcal{W}$. We assume that $\widehat{Y}^W$ see the ground truth Y_v for each vertex $v \notin W$, and further sees the product $Y_{uv} \triangleq Y_u Y_v$ for each edge $e \in E(W)$. Assuming $G(W)$ is connected (this clearly can only make the problem easier), the learner now only needs to infer one bit of information per component. The minimax value of the new game can be written as:

$$\geq \min_{\left\{\widehat{Y}^W\right\}_{W\in\mathcal{W}}} \max_{Y} \mathop{\mathbb{E}}_{X,Z|Y} \sum_{W\in\mathcal{W}} \sum_{v\in W} \mathbb{1}\left\{\widehat{Y}^W_v(X, Z, Y_{V\setminus W}, \{Y_{uv} \mid uv \in E(W)\}) \neq Y_v\right\}.$$

Because the learner only needs to infer a single bit per component, we have reduced to the setting of Theorem 3, components in our setting as vertices in that setting (so $\mathsf{deg}(v)$ is replaced by $\delta_G(W)$). The only substantive difference is the following: In that lower bound, we required that $p < q$. For the new setting, we have that "q" is actually (pessimistically) $q^{|W|}$, and so we require that $p < q^{\max_{W\in\mathcal{W}}|W|}$ for the bound to apply across all components. Using the final bound from Theorem 3, we have

$$\mathbb{E}\left[\sum_{v\in V} \mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\}\right] \geq \Omega\left(q^{\max_{W\in\mathcal{W}}|W|} \sum_{W\in\mathcal{W}} p^{\lceil \delta_G(W)/2\rceil}\right).$$

□

B.3 Proofs from Section 4

Proof of Example 1. We will show that $\Omega(pn)$ Hamming error is optimal for all trees by establishing that all trees have constant fraction of vertices whose degree is at most two, then appealing to Theorem 3.

Let T be the tree under consideration. T is bipartite. Let (A, B) be the bipartition of T into two disjoint independent sets. Suppose without loss of generality that $|A| \geq n/2$. If a is the number of vertices in A of degree at least 3 and $a' = |A| - a$, we have that $3a \leq n-1$, hence $a \leq (n-1)/3$. Therefore $a' \geq n/2 - a \geq (n-1)/6$. Letting A' be the set of vertices in A with at most 2 neighbors, we see that A' is an independent set of size at least $(n-1)/6$, and so we appeal to Theorem 3 for the result. □

Proof of Example 2. Fix $d \geq 3$. We will construct a graph G of size $(d+1)n$. By building up from components as follows:

- For each $k \in [n]$ let G_k be the complete graph on $d+1$ vertices. Remove an edge from an arbitrary pair of vertices (u_k, v_k).
- Form G by taking the collection of all G_k, then adding an edge connecting v_k to u_{k+1} for each k, with the convention $u_{n+1} = u_1$.

This construction for $d = 3$ is illustrated in Figure 1.

Observe that G is d-regular. We obtain the desired result by applying Theorem 4 with the collection $\{G_k\}$ as the set system and observing that the each component G_k has only two edges leaving.

□

Proof of Example 3. We first examine the case where $c = 3$. Here we take the tree decomposition illustrated in Figure 2a, where we cover the graph with overlapping 3×2 components, and take $W^{\star} = \bigcup_{v \in W} N_v$. This yields $\mathsf{mincut}^{\star}(W) = 3$ for all components except those at the graph's endpoints. We now connect the components as a path graph and appeal to Theorem 2, which implies a rate of $\widetilde{O}(p^2 n)$.

When $c = \omega(1)$ we can build a decomposition as follows (informally): Produce E' as in Figure 2b by performing the zig-zag cut with every third row of edges, leaving only 3 edges on the left or right side (alternating). We can now produce T (a path graph) by tiling G' with overlapping 3×3 components. Again, take $W^{\star} = \bigcup_{v \in W} N_v$.

We can verify that if we perform extended inference we have $\mathsf{mincut}^{\star}(W) = 3$ for the $O(n)$ components in the interior of the graph and $\mathsf{mincut}^{\star}(W) = 2$ for the $O(\sqrt{n})$ components at the boundary.

The tree decomposition is illustrated in Figure 3. We have $\mathsf{wid}^{\star}(T) = O(1)$ and $\mathsf{deg}_E(T) = O(1)$. Applying Theorem 2 thus gives an upper bound of $\widetilde{O}(p^2 n + p\sqrt{n})$ with probability at least $1 - \delta$.

Since T is a path graph, we pay $O(n\lceil p^2 n \rceil)$ in computation as per Appendix D.

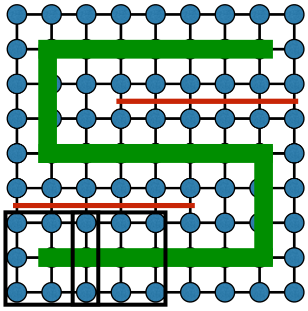

Figure 3: Tree decomposition for 2D grid.

□

Proof of Example 4. We will prove this result for the three-dimensional case. We first show the lower bound.

Suppose $c \geq 3$ is constant, so that we are in the "hypertube" regime. Note that vertices on the outermost "edges" of the hypertube, examples of which are circled in Figure 4, have degree at most 4. There are $\Omega(n)$ such vertices, so appealing to Theorem 3 yields a lower bound on Hamming error of $\Omega(p^2 n)$. In fact for the $n/c^2 \times c \times c$ hyper-tube one can achieve the $O(p^2 n)$ rate using our method. Simply take each components of size $2 \times c \times c$ connected in a path as in the example for the 2D grid. Since the minimum cut for each component is already at least 3, we don't need to consider extended components and simply use brute-force on the components themselves.

We now sketch the upper bound for the $n^{1/3} \times n^{1/3} \times n^{1/3}$ hypergrid. We use a technique similar to that used for the 2D grid in Example 3: We take T to be a path graph obtained by covering the hypergrid in overlapping

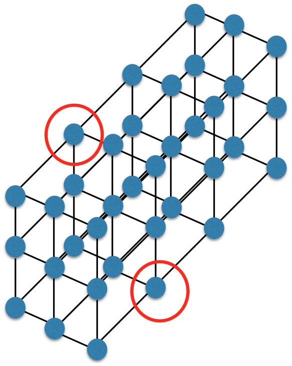

Figure 4: Lower bound argument for $n/c^2 \times c \times c$ hypergrid.

$3 \times 3 \times 3$ components in a zig-zagging pattern. Note that each $3 \times 3 \times 3$ component will contain nodes similar to those highlighted in Figure 4 with degree at most 4. This means $\mathsf{mincut}^\star(W) = 4$, so to obtain the $O(p^3 n)$ Hamming error we must consider extended components. Take $W^\star = \bigcup_{v \in W} N_v$. Then $\mathsf{mincut}^\star(W) = 6$ for all components except those at the boundary of the hypergrid, which have $\mathsf{mincut}^\star(W) \in \{4, 5\}$. There are only $o(n)$ such components, so we achieve the $O(p^3 n)$ upper bound by appealing to Theorem 2.

For higher-dimensional hypergrids, the strategy of taking components to be constant-sized hypergrids and T to be a zig-zagging path graph readily extends. The lower bound stated follows from a simple counting argument.

In general, we can associated vertices of a $c_1 \times c_2 \times \ldots \times c_d$ hypergrid with the elements of $\mathbb{Z}_{c_1} \times \mathbb{Z}_{c_2} \times \ldots \times \mathbb{Z}_{c_d}$. For a vertex $v = (v_1, \ldots, v_d)$, the degree is given by $\mathsf{deg}(v) = |\{k \in [d] \mid v_k \in \{0, c_k\}\}|$.

Consider the case where $c_1, \ldots c_{d-1} = c$, $c_d = n/c^{d-1}$. In this case, the degree argument above implies

$$|\{v \mid \mathsf{deg}(v) = d + 1\}| \geq \sum_{v_k \in \{0,c\} : k \neq d} (n - 2) = \Omega(n).$$

Thus, a constant fraction of vertices have degree $d + 1$, and so Theorem 3 implies a lower bound of $\Omega(p^{\lceil \frac{d+1}{2} \rceil} n)$.

□

Proof of Example 7.
Upper bound: Tree decomposition We first formally define the tree decomposition $T = (\mathcal{W}, F)$ that we will use with Algorithm 1. Assume for simplicity what $n = n' \cdot (2k + 1)$. We will define a vertex set $\{v_1, \ldots, v_{n'}\}$ as follows: $v_1 = 1$, $v_{i+1} = v_i + k + 1$. We will now define a component for each of these vertices:

$$W(v_i) = N_G(v_i).$$

Let $\mathcal{W}$ will be the union of these components. Since we assumed n to be divisible by $(2k + 1)$, the components a partition of V. We now define the Extend function for this decomposition:

$$\textsc{Extend}(W) = \bigcup_{v \in W} N_G(v).$$

That is, the extended component $W^\star(v_i)$ is the set of all vertices removed from v_i by paths of length 2.

Finally, we construct the edge set F by adding edges of the form $(W(v_i), W(v_{i+1}))$ for $i \in \{1, \ldots, n' - 1\}$. This means that the decomposition is a path graph. The decomposition is clearly admissible in the sense of Definition 3.

We can observe that $\mathsf{mincut}^\star(W) = 2k$ just as the minimum cut of $R_{n,k}$ is itself $2k$. Theorem 2 thus implies a recovery rate of $\widetilde{O}(p^k n)$. Since T is a path graph, the algorithm runs in time $O(\lceil p^k n \rceil n)$.

Lower bound That $O(p^k n)$ is optimal can be seen by appealing to Theorem 3 with the fact that $R_{n,k}$ is $2k$-regular.

□

Proof of Example 8. The average number of vertices added is αn. By the Chernoff bound, with high probability the number of vertices added is bounded as $\alpha n + c\sqrt{\alpha n \log n}$ for some constant c. This means that for any $\epsilon > 0$, there is some minimum n for which an $(1 - \alpha + \epsilon)$ fraction of vertices have no edges added. This means that there are at least $(1 - \alpha + \epsilon)n$ edges with degree $2k$, so Theorem 3 yields the result. □

B.4 Analysis of TreeDecompositionDecoder

Properties of Tree Decompositions We begin by recalling a few properties of tree decompositions that are critical for proving the performance bounds for Algorithm 1.

Proposition 2. For any tree decomposition $T = (\mathcal{W}, F)$, the following properties hold:

1. For each $v \in V$ there exists W with $v \in W$.
   This guarantees that we produce a prediction for each vertex.

2. If $(W_1, W_2) \in F$, there is some $v \in V$ with $v \in W_1, W_2$.
   This guarantees that the class $\mathcal{F}$ (see (19)) is well-defined.

3. T is connected
   This implies that $|\mathcal{F}| \lesssim 2^K$.

4. $|\mathcal{W}| \leq n$.
   This implies that a mistake bound for components of the tree decomposition translates to a mistake bound for vertices of G.

Proof of Proposition 2. 1. Definition 1.

2. Suppose there is some edge $(W_1, W_2) \in F$ with no common vertices. Consider the subtrees T_1 and T_2 created by removing $(W_1, W_2) \in F$. By the coherence property (Definition 1), the subgraphs of G' associated with these decompositions (call them G'_{T_1} and G'_{T_2}) must have no common nodes. Yet, G' is connected, so there must be $(u, v) \in E'$ with $u \in G'_{T_1}$, $v \in G'_{T_2}$. Our hypothesis now implies that there is no $W \in \mathcal{W}$ containing u and v, so T violates the edge inclusion property of the tree decomposition.

3. Definition 1

4. This follows directly from the non-redundancy assumption of Definition 1. See, e.g., (Kleinberg and Tardos, 2006, 10.16).

□

Estimation in Tree Decomposition Components We now formally define and analyze the component-wise estimators computed in line 5 of Algorithm 1.

Definition 6 (Extended Component Estimator)**.** *Consider the (edge) maximum likelihood estimator over $W^\star$:*

$$\widetilde{Y}^{W^\star} \triangleq \operatorname*{arg\,min}_{\widetilde{Y}\in\{\pm1\}^{W^\star}} \sum_{uv\in E'(W^\star)} \mathbb{1}\{\widetilde{Y}_u\widetilde{Y}_v \neq X_{uv}\}. \tag{10}$$

We define the ***extended component estimator*** *$\widehat{Y}^{W^\star} \in \{\pm1\}^W$ as restriction of $\widetilde{Y}^{W^\star}$ to W.*

For $\widehat{Y}^{W^\star}$ estimation performance is governed by $\mathsf{mincut}^\star(W)$ rather than $\mathsf{mincut}(W)$, as the next lemma shows:

Lemma 2 (Error Probability for Extended Component Estimator)**.**

$$\mathbb{P}\left(\min_{s\{\pm1\}} \mathbb{1}\{s\widehat{Y}^{W^\star} \neq Y^W\} > 0\right) \leq 2^{|W^\star|}p^{\lceil \mathsf{mincut}^\star(W)/2\rceil}.$$

Proof of Lemma 2. Suppose $\widehat{Y}^{W^\star} \neq Y^W$ and consider $D = \{v \in W^\star : \widetilde{Y}_v^{W^\star} \neq Y_v\}$. Then there is some maximal connected component S of D containing at least one vertex of W. It must then be the case that at least half the edge samples in $\delta(S)$ are flipped with respect to the ground truth. Consequently it holds that

$$\begin{aligned}\mathbb{P}\left(\min_{s\{\pm 1\}} \mathbb{1}\{s\widehat{Y}^{W^\star} \neq Y^W\} > 0\right) &\leq \sum_{S\subseteq W^\star : S\cap W\neq\emptyset, \bar{S}\cap W\neq\emptyset} p^{\lceil|\delta(S)|/2\rceil} \\ &\leq \sum_{S\subseteq W^\star} p^{\lceil \mathsf{mincut}^\star(W)/2\rceil} \\ &\leq 2^{|W^\star|} p^{\lceil \mathsf{mincut}^\star(W)/2\rceil}.\end{aligned}$$

□

Lemma 2 shows that considering mincut* offers improved failure probability over mincut because it allows us to take advantage of all of the information in $W^\star$, yet only pay (in terms of errors) for cuts that involve nodes in the core component W. In Figure 2a, all components of the tree decomposition except the endpoints have $\mathsf{mincut}^\star(W) = 3$, and so their extended component estimators achieve $O(p^2)$ failure probability.

Concentration We begin by stating a concentration result for functions of independent random variables, which we will use to establish a bound on the total number of components that fail in the first stage of our algorithm. Let $X_1, \ldots, X_n$ be independent random variables each taking values in a probability space $\mathcal{X}$, and let $F : \mathcal{X}^n \to \mathbb{R}$. We will be interested in the concentration of the random variable $S = F(X_1, \ldots, X_n)$. Letting $X_1', \ldots, X_n'$ be independent copies of $X_1, \ldots, X_n$, we define $S^{(i)} = F(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$. Finally, we define a new random variable

$$V_+ = \sum_{i=1}^n \mathbb{E}\left[(S - S^{(i)})_+^2 \mid X_1, \ldots, X_n\right].$$

Theorem 5 (Entropy Method with Efron-Stein Variance (Boucheron et al., 2003))**.** *If there exists a constant $a > 0$ such that $V_+ \leq aS$ then*

$$\mathbb{P}\{S \geq \mathbb{E}[S] + t\} \leq \exp\left(\frac{-t^2}{4a\,\mathbb{E}[S] + 2at}\right).$$

Subsequently, with probability at least $1 - \delta$,

$$S \leq \mathbb{E}[S] + \max\left\{4a\log(1/\delta), 2\sqrt{2a\,\mathbb{E}[S]\log(1/\delta)}\right\} \leq 2\,\mathbb{E}[S] + 6a\log(1/\delta).$$

With Theorem 5 in mind, we may proceed to a bound on the number of components with mistakes when the basic component estimator (8) is used.

Lemma 3 (Formal Version of Lemma 1)**.** For all $\delta > 0$, with probability at least $1 - \delta$ over the draw of X,

$$\min_{s\in\{\pm 1\}^{\mathcal{W}}} \sum_{W\in\mathcal{W}} \mathbb{1}\{s_W\widehat{Y}^W \neq Y^W\} \leq 2\sum_{W\in\mathcal{W}} 2^{|W|} p^{\lceil \mathsf{mincut}(W)/2\rceil} + 6\max_{e\in E}|\mathcal{W}(e)| \max_{W\in\mathcal{W}}|E'(W)|\log(1/\delta). \tag{11}$$

$$\tag{12}$$

Proof of Lemma 3. Define a random variable

$$S(X) = \sum_{W\in\mathcal{W}} \min_{s\in\{\pm 1\}} \mathbb{1}\left\{s\widehat{Y}^W(X) \neq Y^W\right\},$$

where $\widehat{Y}^W$ are the component-wise estimators produced by Algorithm 1 and X are the edge observations. To prove the lemma we will apply Theorem 5 by showing that there is a constant a such that the necessary variance bound $V_+ \leq aS$ holds.

To this end, consider

$$S(X) - S(X^{(e)}) = \sum_{W\in\mathcal{W}} \left(\min_{s\in\{\pm 1\}} \mathbb{1}\left\{s\widehat{Y}^W(X) \neq Y^W\right\} - \min_{s\in\{\pm 1\}} \mathbb{1}\left\{s\widehat{Y}^W(X^{(e)}) \neq Y^W\right\}\right),$$

where $X^{(e)}$ is defined as in Theorem 5. To be more precise, we draw $(X'_e)_{e\in E}$ from the same distribution as X, then let $X^{(e)}$ be the result of replacing X_e with X'_e.

We have

$$S(X) - S(X^{(e)}) = \sum_{W\in\mathcal{W}(e)}\left(\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\} - \min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X^{(e)})\neq Y^W\right\}\right),$$

since changing X_e can only change $\widehat{Y}^W$ if $e\in W$. Now, since $S(X^{(e)})$ is nonnegative we have

$$\begin{aligned}
(S(X) - S(X^{(e)})^2_+ &= \left(\sum_{W\in\mathcal{W}(e)}\left(\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\} - \min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X^{(e)})\neq Y^W\right\}\right)\right)^2_+\\
&\leq \left(\sum_{W\in\mathcal{W}(e)}\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\}\right)^2\\
&\leq |\mathcal{W}(e)|\sum_{W\in\mathcal{W}(e)}\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\}.
\end{aligned}$$

We now sum over all edges to arrive at an upper bound on V_+:

$$\begin{aligned}
V_+ &= \sum_{e\in E}\mathbb{E}\left[(S(X) - S(X^{(e)})^2_+ \mid X\right]\\
&\leq \max_{e\in E}|\mathcal{W}(e)|\sum_{e\in E}\sum_{W\in\mathcal{W}(e)}\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\}\\
&= \max_{e\in E}|\mathcal{W}(e)|\sum_{W\in\mathcal{W}}\sum_{e\in E(W)}\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\}\\
&\leq \max_{e\in E}|\mathcal{W}(e)|\max_{W\in\mathcal{W}}|E(W)|\sum_{W\in\mathcal{W}}\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\}\\
&\leq \max_{e\in E}|\mathcal{W}(e)|\max_{W\in\mathcal{W}}|E(W)|\sum_{W\in\mathcal{W}}\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\}\\
&= \max_{e\in E}|\mathcal{W}(e)|\max_{W\in\mathcal{W}}|E(W)|S(X).
\end{aligned}$$

We now appeal to Theorem 5 with $a = \max_{e\in E}|\mathcal{W}(e)|\max_{W\in\mathcal{W}}|E(W)|$, which yields that with probability at least $1-\delta$,

$$S \leq 2\,\mathbb{E}[S] + 6\max_{e\in E}|\mathcal{W}(e)|\max_{W\in\mathcal{W}}|E(W)|\log(1/\delta).$$

Finally, the bound on $\mathbb{E}[S]$ follows from Proposition 1:

$$\mathbb{E}[S] = \sum_{W\in\mathcal{W}}\mathbb{P}\left(\min_{s\in\{\pm1\}}\mathbb{1}\left\{s\widehat{Y}^W(X)\neq Y^W\right\}\right) \leq \sum_{W\in\mathcal{W}}2^{|W|}p^{\lceil\mathsf{mincut}(W)/2\rceil}.$$

□

An analogous concentration result to Lemma 3 holds to bounds the number of components that fail over the whole graph when the extended component estimator is used:

Lemma 4. For all $\delta > 0$, with probability at least $1-\delta$ over the draw of X,

$$\min_{s\in\{\pm1\}^{\mathcal{W}}}\sum_{W\in\mathcal{W}}\mathbb{1}\{s_W\widehat{Y}^{W^\star}\neq Y^{W^\star}\} \leq 2\sum_{W\in\mathcal{W}}2^{|W^\star|}p^{\lceil\mathsf{mincut}^\star(W)/2\rceil} + 6\max_{e\in E}|\mathcal{W}^\star(e)|\max_{W\in\mathcal{W}}|E'(W^\star)|\log(1/\delta). \tag{13}$$

Proof of Lemma 4. This proof proceeds exactly as in the proof of Lemma 3 using

$$S(X) = \sum_{W \in \mathcal{W}} \min_{s \in \{\pm 1\}} \mathbb{1}\left\{s\widehat{Y}^{W^\star}(X) \neq Y^W\right\}.$$

The only difference is that edges are more influential than in that lemma because each extended component estimator $\widehat{Y}^{W^\star}$ may depend on more edges than the simpler component estimator $\widehat{Y}^W$. To this end, define $\mathcal{W}^\star(e) = \{W \mid e \in E'(W^\star)\}$. One can verify that if we replace every instance of $\mathcal{W}(e)$ in the proof of Lemma 3 with $\mathcal{W}^\star(e)$ it holds that $V_+ \leq aS$ with $a = \max_{e \in E}|\mathcal{W}^\star(e)| \max_{W \in \mathcal{W}}|E(W^\star)|$. Theorem 5 then implies that with probability at least $1 - \delta$,

$$\begin{aligned} S &\leq 2\,\mathbb{E}[S] + 6 \max_{e \in E}|\mathcal{W}^\star(e)| \max_{W \in \mathcal{W}}|E(W^\star)| \log(1/\delta) \\ &= 2\,\mathbb{E}[S] + 6\mathsf{deg}^\star_E(T) \max_{W \in \mathcal{W}}|E(W^\star)| \log(1/\delta). \end{aligned}$$

□

Proof of Theorem 2.
Full theorem statement We will prove the following error bound: If $T = (\mathcal{W}, F)$ is admissible, with probability at least $1 - \delta$ over the draw of X and Z, $\widehat{Y}$ satisfies:

$$\sum_{v \in V} \mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\} \tag{14}$$

$$\leq O\left(\frac{1}{\epsilon^2}\left(2^{\mathsf{wid}^\star(T)} \sum_{W \in \mathcal{W}} p^{\lceil \mathsf{mincut}^\star(W)/2 \rceil} + \mathsf{deg}^\star_E(T) \max_{W \in \mathcal{W}}|E(W^\star)| \log(1/\delta)\right) \cdot (\mathsf{wid}(T) + \mathsf{deg}(T) \log n)\right) \tag{15}$$

This statement specializes to (6) when all of the tree decomposition quantities are constant and $\delta = 1/n$.

Error bound for individual components Lemma 2 implies that for a fixed component $W \in \mathcal{W}$, the probability that the estimator produced by the brute-force enumeration routine fails to exactly recover the labels in W (up to sign) is bounded as

$$\mathbb{P}\left(\min_{s\{\pm 1\}} \mathbb{1}\{s\widehat{Y}^{W^\star} \neq Y^W\} > 0\right) \leq 2^{|W^\star|} p^{\lceil \mathsf{mincut}^\star(W)/2 \rceil}.$$

Error bound across all components Consider the following random variable, which is the total number components

$$S(X) = \sum_{W \in \mathcal{W}} \min_{s \in \{\pm 1\}} \mathbb{1}\left\{s\widehat{Y}^{W^\star}(X) \neq Y^W\right\}.$$

The bound on component failure probability immediately implies in in-expectation bound on S:

$$\mathbb{E}[S] \leq \sum_{W \in \mathcal{W}} 2^{|W^\star|} p^{\lceil \mathsf{mincut}^\star(W)/2 \rceil}.$$

Lemma 4 shows that S concentrates tightly around its expectation. More precisely, let $A \triangleq 6\mathsf{deg}^\star_E(T) \max_{W \in \mathcal{W}}|E(W^\star)|$ and

$$K_n \triangleq 2^{\mathsf{wid}^\star(T)+2} \sum_{W \in \mathcal{W}} p^{\lceil \mathsf{mincut}^\star(W)/2 \rceil} + A \log(2/\delta). \tag{16}$$

Then Lemma 4 implies that with probability at least $1 - \delta/2$,

$$\begin{aligned} \min_{s \in \{\pm 1\}^{\mathcal{W}}} \sum_{W \in \mathcal{W}} \mathbb{1}\{s_W \widehat{Y}^{W^\star} \neq Y^W\} &\leq 2 \sum_{W \in \mathcal{W}} 2^{|W^\star|} p^{\lceil \mathsf{mincut}^\star(W)/2 \rceil} + A \log(2/\delta) \\ &\leq K_n \end{aligned} \tag{17}$$

Inference with side information: Hypothesis class Consider the following binary signing of the components in T:

$$s^\star = \arg\min_{s\in\{\pm1\}^{\mathcal{W}}} \sum_{W\in\mathcal{W}} \mathbb{1}\{s_W \widehat{Y}^{W^\star} \neq Y^W\}.$$

$s^\star$ is signing of the component-wise predictions $(\widehat{Y}^{W^\star})$ that best matches the ground truth. If we knew the value of $s^\star$ we could use it to produce a vertex prediction with at most K_n mistakes. Computing the $s^\star$ is information-theoretically impossible because we do not have access to Y, but we will show that the signing we produce using the side information Z is close.

Define

$$L_n = \mathsf{deg}(T)\cdot K_n. \tag{18}$$

We will argue that (17) implies that $s^\star$ lies in the class

$$\mathcal{F}(X) \triangleq \left\{ s\in\{\pm1\}^{\mathcal{W}} \mid \sum_{(W_1,W_2)\in F} \mathbb{1}\{s_{W_1} \neq s_{W_2}\cdot S(W_1,W_2)\} \leq L_n \right\}. \tag{19}$$

First, consider the `for` loop on Algorithm 1, line 11. Proposition 2 implies that $S(W_1,W_2)$ as defined in this loop is well-defined, because there always exists some $v\in W_1\cap W_2$.

Second, consider the value of

$$\sum_{(W_1,W_2)\in F} \mathbb{1}\{s^\star_{W_1} \neq s^\star_{W_2}\cdot S(W_1,W_2)\} = \sum_{(W_1,W_2)\in F} \mathbb{1}\{s^\star_{W_1} \neq s^\star_{W_2}\cdot \widehat{Y}_v^{W_1^\star}\cdot \widehat{Y}_v^{W_2^\star}\}.$$

We can bound this quantity in terms of the number of components W for which

$$\min_{s\in\{\pm1\}} \mathbb{1}\left\{s\widehat{Y}^{W^\star} \neq Y^W\right\} = 1.$$

Observe that if $\min_{s\in\{\pm1\}} \mathbb{1}\left\{s\widehat{Y}^{W^\star} \neq Y^W\right\} = 0$ then there is some $\bar{s}_W\in\{\pm1\}$ such that $\widehat{Y}^{W^\star} = \bar{s}_W Y^W$. If we take $s^\star_W = \bar{s}_W$ in all the components with no errors, and choose the sign arbitrarily for others, we will have $\mathbb{1}\{s^\star_{W_1} \neq s^\star_{W_2}\cdot \widehat{Y}_v^{W_1^\star}\cdot \widehat{Y}_v^{W_2^\star}\} = 0$ whenever both W_1 and W_2 have no errors. Pessimistically, there are at most $L_n = \mathsf{deg}(T)\cdot K_n$ edges (W_1,W_2) where at least one of W_1 or W_2 has an error, and therefore (17) implies that with probability at least $1-\delta/2$, $s^\star\in\mathcal{F}$.

We conclude this discussion by showing that $|\mathcal{F}(X)|$ small. Since by Proposition 2 T is connected, labelings of the edges of T are in one to one correspondence with labelings of the components. Consequently,

$$|\mathcal{F}(X)| \leq \sum_{k=0}^{L_n}\binom{|\mathcal{W}|}{k} \leq \left(\frac{e|\mathcal{W}|}{L_n}\right)^{L_n} \leq \left(\frac{en}{L_n}\right)^{L_n}. \tag{20}$$

The last inequality uses that, from Proposition 2, $|\mathcal{W}|\leq n$.

Final error bound for inference with side information We now use the properties of $\mathcal{F}(X)$ to derive an error bound for the prediction $\widehat{Y}$. Recall from Algorithm 1 that $\widehat{Y}$ is defined in terms of

$$\hat{s} = \min_{s\in\mathcal{F}(X)} \sum_{W\in\mathcal{W}}\sum_{v\in W} \mathbb{1}\{s_W \widehat{Y}_v^{W^\star} \neq Z_v\}. \tag{21}$$

We reduce the analysis of the error rate of $\hat{s}$ to analysis of excess risk in a manner that parallels the proof of Theorem 1, but is slightly more involved because the best predictor in $\mathcal{F}$ does not perfectly match the ground truth. Fix $\hat{s} \in \{\pm 1\}^{\mathcal{W}}$. For each component $W \in \mathcal{W}$ we have

$$\begin{aligned}
\sum_{v \in W} \mathbb{1}\{\hat{s}_W \widehat{Y}_v^{W^\star} \neq Y_v\} &\leq \sum_{v \in W} \mathbb{1}\{\hat{s}_W \widehat{Y}_v^{W^\star} \neq s^*_W \widehat{Y}_v^{W^\star}\} + \sum_{v \in W} \mathbb{1}\{s^*_W \widehat{Y}_v^{W^\star} \neq Y_v\} \\
&\leq \sum_{v \in W} \mathbb{1}\{\hat{s}_W \widehat{Y}_v^{W^\star} \neq s^*_W \widehat{Y}_v^{W^\star}\} + |W| \; \mathbb{1}\{s^*_W \widehat{Y}^{W^\star} \neq Y^W\} \\
&= \frac{1}{1-2q} \sum_{v \in W : s^*_W \widehat{Y}_v^{W^\star} = Y_v} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right) \\
&\quad - \frac{1}{1-2q} \sum_{v \in W : s^*_W \widehat{Y}_v^{W^\star} \neq Y_v} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right) \\
&\quad + |W| \; \mathbb{1}\{s^*_W \widehat{Y}_W \neq Y_W\}.
\end{aligned}$$

Now note that given that Z_v is drawn as a noisy version of Y_v, $\left|\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right| = 1 - 2q$ and so

$$\begin{aligned}
&- \frac{1}{1-2q} \sum_{v \in W : s^*_W \widehat{Y}_v^{W^\star} \neq Y_v} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right) \\
&\qquad \leq 2 \sum_{v \in W} \mathbb{1}\{s^*_W \widehat{Y}_v^{W^\star} \neq Y_v\} + \frac{1}{1-2q} \sum_{v \in W : s^*_W \widehat{Y}_v^{W^\star} \neq Y_v} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right) \\
&\qquad \leq 2|W| \mathbb{1}\{s^*_W \widehat{Y}^{W^\star} \neq Y^W\} + \frac{1}{1-2q} \sum_{v \in W : s^*_W \widehat{Y}_v^{W^\star} \neq Y_v} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right).
\end{aligned}$$

We conclude that

$$\begin{aligned}
&\sum_{v \in W} \mathbb{1}\{\hat{s}_W \widehat{Y}_v^{W^\star} \neq Y_v\} \\
&\leq 3|W| \mathbb{1}\{s^*_W \widehat{Y}^{W^\star} \neq Y^W\} + \frac{1}{1-2q} \sum_{v \in W} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right).
\end{aligned}$$

Summing over all the components $W \in \mathcal{W}$ we arrive at the bound

$$\begin{aligned}
&\sum_{W \in \mathcal{W}} \sum_{v \in W} \mathbb{1}\{\hat{s}_W \widehat{Y}_v^{W^\star} \neq Y_v\} \\
&\quad \leq 3 \left(\max_{W \in \mathcal{W}} |W| \right) \sum_{w \in \mathcal{W}} \mathbb{1}\{s^*_W \widehat{Y}^{W^\star} \neq Y^W\} + \frac{1}{1-2q} \sum_{W \in \mathcal{W}} \sum_{v \in W} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right) \\
&\quad \leq 3 \left(\max_{W \in \mathcal{W}} |W| \right) K_n + \frac{1}{1-2q} \sum_{W \in \mathcal{W}} \sum_{v \in W} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right)
\end{aligned}$$

We can now appeal to the statistical learning bounds from Appendix C to handle the right-hand side of this expression. Lemma 5 implies that if we take $\hat{s} = \arg\min_{s \in \mathcal{F}} \sum_{W \in \mathcal{W}} \sum_{v \in W} \mathbb{1}\left\{\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right\}$, which is precisely the solution to (21), we obtain the excess risk bound,

$$\begin{aligned}
&\sum_{W \in \mathcal{W}} \sum_{v \in W} \left(\mathbb{P}_Z\left(\hat{s}_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right) - \mathbb{P}_Z\left(s^*_W \widehat{Y}_v^{W^\star} \cdot Z_v < 0\right)\right) \\
&\leq \left(\frac{2}{3} + \frac{c}{2} \right) \log(2|\mathcal{F}|/\delta) + \frac{1}{c} \sum_{w \in \mathcal{W}} \sum_{v \in W} \mathbb{1}\{\hat{s}_W \widehat{Y}_v^{W^\star} \neq Y_v\},
\end{aligned}$$

with probability at least $1 - \delta/2$ over Z for all $c > 0$. If we choose $c = 1/\epsilon$, rearrange, and apply the union bound, this implies that with probability at least $1 - \delta$ over the draw of X and Z we have

$$\sum_{W \in \mathcal{W}} \sum_{v \in W} \mathbb{1}\{\hat{s}_W \widehat{Y}_v^{W^\star} \neq Y_v\} \leq 6 \left(\max_{W \in \mathcal{W}} |W| \right) K_n + \frac{2}{\epsilon^2} \log(2|\mathcal{F}|/\delta).$$

Recall that $|\mathcal{F}| \leq (e|\mathcal{W}|/L_n)^{L_n}$, which implies a bound of

$$
\begin{aligned}
&\sum_{W\in\mathcal{W}}\sum_{v\in W}\mathbb{1}\{\hat{s}_W\widehat{Y}_v^{W^\star}\neq Y_v\}\\
&\leq O\left(\frac{1}{\epsilon^2}[\mathsf{wid}(T)\cdot K_n + L_n\cdot\log(en/L_n)+\log(1/\delta)]\right)\\
&\leq O\left(\frac{1}{\epsilon^2}[K_n\cdot(\mathsf{wid}(T)+\mathsf{deg}(T)\cdot\log(en/K_n))+\log(1/\delta)]\right)\\
&\leq O\left(\frac{1}{\epsilon^2}\left(2^{\mathsf{wid}^\star(T)}\sum_{W\in\mathcal{W}}p^{\lceil\mathsf{mincut}^\star(W)/2\rceil}+\mathsf{deg}^\star_E(T)\max_{W\in\mathcal{W}}|E(W^\star)|\log(1/\delta)\right)\cdot(\mathsf{wid}(T)+\mathsf{deg}(T)\log n)\right)
\end{aligned}
$$

Our choice of $\widehat{Y}$ in Algorithm 1 ensures that the Hamming error $\sum_{v\in V}\mathbb{1}\left\{\widehat{Y}_v\neq Y_v\right\}$ inherits this bound. Proposition 2 implies that every $v\in V$ is in some component, so this choice is indeed well-defined. □

C Statistical learning

Here we consider a fixed design variant of the statistical learning setting. Fix an input space $\mathcal{X}$ and output space $\mathcal{Z}$. We are given a fixed set $X_1,\ldots,X_n\in\mathcal{X}$ and samples $Z_1,\ldots,Z_n\in\mathcal{Z}$ with Z_i drawn from $P(Z_i\mid X_i)$ for some distribution P. We fix a hypothesis class $\mathcal{F}$ which is some subset of mappings from $\mathcal{X}$ to $\mathcal{Z}$, and we would like to use Z to find $\widehat{Y}\in\mathcal{F}$ that will predict future observations of Z on X. To evaluate prediction we define a loss function $\ell:\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}_+$, and define $L_i(Y)=\mathbb{E}_{Z|X_i}[\ell(Y,Z)]$. Our goal is to use Z to select $\widehat{Y}\in\mathcal{F}$ to guarantee low *excess risk*:

$$\sum_{i\in[n]}L_i(\widehat{Y}(X_i))-\min_{Y\in\mathcal{F}}\sum_{i\in[n]}L_i(Y(X_i)). \tag{22}$$

Typically this is accomplished using the *empirical risk minimizer* (ERM):

$$\widehat{Y}=\arg\min_{Y\in\mathcal{F}}\sum_{i\in[n]}\ell(Y(X_i),Z_i)^{5}.$$

In this paper we consider a specific instantiation of the above framework in which

- $\mathcal{X}=V$, the vertex set for some graph (possibly a tree decomposition), and $X_1,\ldots,X_n$ are an arbitrary ordering of V (so $n=|V|$). In light of this we index all variables using V going forward.
- $\mathcal{Z}=\{\pm1\}$. We fix $Y\in\{\pm1\}^V$ and let $Z_v=Y_v$ with probability $1-q$ and $Z_v=-Y_v$ otherwise (as in Model 1).
- $\ell(Y,Z)=\mathbb{1}\{Y\neq V\}$, so $L_i(Y)=\mathbb{P}_Z(Y\neq Z_v)$.
- $\mathcal{F}\subseteq\{\pm1\}^V$ is arbitrary.

For this setting the excess risk for a predictor $\widehat{Y}\in\{\pm1\}^V$ can be written as

$$\sum_{v\in V}\mathbb{P}(\widehat{Y}_v\neq Z_v)-\min_{Y'\in\mathcal{F}}\sum_{v\in V}\mathbb{P}\left(Y'_v\neq Z_v\right), \tag{23}$$

and the empirical risk minimizer is given by $\widehat{Y}=\arg\min_{Y'\in\mathcal{F}}\sum_{v\in V}\mathbb{1}\{Y'_v\neq Z_v\}$.

We assume this setting exclusively for the remainder of the section.

[5] There are a many standard bounds quantifying the performance of ERM in settings beyond the one we consider. See Bousquet et al. (2004) for a survey.

Lemma 5 (Excess risk bound for ERM)**.** Let $\widehat{Y}$ be the ERM and let $Y^{\star} = \arg\min_{Y' \in \mathcal{F}} \sum_{v \in V} \mathbb{P}(Y' \neq Z)$. Then with probability at least $1-\delta$ over the draw of Z,

$$\sum_{v \in V} \mathbb{P}\left(\widehat{Y}_v \neq Z_v\right) - \min_{Y' \in \mathcal{F}} \sum_{v \in V} \mathbb{P}\left(Y'_v \neq Z_v\right) \leq \left(\frac{2}{3} + \frac{c}{2}\right) \log\left(\frac{|\mathcal{F}|}{\delta}\right) + \frac{1}{c} \sum_{v \in V} \mathbb{1}\left\{\widehat{Y}_v \neq Y^{\star}_v\right\} \tag{24}$$

for all $c > 0$.

Corollary 2 (ERM excess risk: Well-specified case)**.** When $Y \in \mathcal{F}$ we have that with probability at least $1-\delta$,

$$\sum_{v \in V} \mathbb{P}\left(\widehat{Y}_v \neq Z_v\right) - \min_{Y \in \mathcal{F}} \sum_{v \in V} \mathbb{P}\left(Y_v \neq Z_v\right) \leq \left(\frac{4}{3} + \frac{1}{\epsilon}\right) \log\left(\frac{|\mathcal{F}|}{\delta}\right), \tag{25}$$

recalling $q = 1/2 - \epsilon$.

Proof of Corollary 2. When $Y \in \mathcal{F}$, $Y^{\star} = Y$, and we have

$$\sum_{v \in V} \mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\} = \frac{1}{1-2q} \sum_{v \in V} \left(\mathbb{P}\left(\widehat{Y}_v \neq Z_v\right) - \mathbb{P}(Y_v \neq Z_v)\right).$$

Applying this inequality to the right hand side of (24) and rearranging yields

$$\left(1 - \frac{1}{c(1-2q)}\right) \sum_{v \in V} \left(\mathbb{P}\left(\widehat{Y}_v \neq Z_v\right) - \mathbb{P}(Y_v \neq Z_v)\right) \leq \left(\frac{2}{3} + \frac{c}{2}\right) \log(|\mathcal{F}|/\delta).$$

To complete the proof we take $c = \frac{2}{1-2q}$, which gives

$$\frac{1}{2} \sum_{v \in V} \left(\mathbb{P}\left(\widehat{Y}_v \neq Z_v\right) - \mathbb{P}(Y_v \neq Z_v)\right) \leq \left(\frac{2}{3} + \frac{1}{1-2q}\right) \log(|\mathcal{F}|/\delta).$$

□

Proof of Lemma 5. We will use Lemma 6 with $\mathcal{F}$ as the index set so that every $i \in [N]$ corresponds to one $Y' \in \mathcal{F}$. We define our collection of random variables as

$$T_v^{Y'} = \mathbb{1}\{Y'_v \neq Z_v\} - \mathbb{1}\{Y^{\star}_v \neq Z_v\}$$

where Y is the ground truth and Y' is any element of $\mathcal{F}$. Now using Lemma 6 and recalling $\sigma^2_{Y'} = \sum_{v \in V} \mathrm{Var}(T_v^{Y'})$, we have that with probability at least $1-\delta$, simultaneously for all Y',

$$\begin{aligned}
\sum_{v \in V} (\mathbb{E}[T_v^{Y'}] - T_v^{Y'}) &\leq \frac{2}{3} \log(|\mathcal{F}|/\delta) + \sqrt{2\sigma^2_{Y'} \log(|\mathcal{F}|/\delta)} \\
&\leq \inf_{c>0} \left[\left(\frac{2}{3} + \frac{c}{2}\right) \log(|\mathcal{F}|/\delta) + \sigma^2_{Y'}/c\right] \\
&\leq \inf_{c>0} \left[\left(\frac{2}{3} + \frac{c}{2}\right) \log(|\mathcal{F}|/\delta) + \frac{1}{c} \sum_{v \in V} \mathbb{E}[(T_v^{Y'})^2]\right].
\end{aligned}$$

In particular this implies that for $\widehat{Y} = \arg\min_{Y' \in \mathcal{F}} \sum_{v \in V} \mathbb{1}\{Y'_v \neq Z_v\}$ we have that for all $c > 0$,

$$\begin{aligned}
\sum_{v \in V} \left(\mathbb{P}\left(\widehat{Y}_v \neq Z_v\right) - \mathbb{P}\left(Y^{\star}_v \neq Z_v\right)\right) \leq& \sum_{v \in V} \left(\mathbb{1}\left\{\widehat{Y}_v \neq Z_v\right\} - \mathbb{1}\left\{Y^{\star}_v \neq Z_v\right\}\right) + \left(\frac{2}{3} + \frac{c}{2}\right) \log(|\mathcal{F}|/\delta) \\
&+ \frac{1}{c} \sum_{v \in V} \mathbb{E}\left[\left(\mathbb{1}\{\widehat{Y}_v \neq Z_v\} - \mathbb{1}\{Y^{\star}_v \neq Z_v\}\right)^2\right].
\end{aligned}$$

Now since $Y^\star \in \mathcal{F}$ and $\widehat{Y}$ is the ERM, we get that $\sum_{v\in V}\left(\mathbb{1}\left\{\widehat{Y}_v \neq Z_v\right\} - \mathbb{1}\left\{Y^\star_v \neq Z_v\right\}\right) \leq 0$ and so,

$$\sum_{v\in V}\left(\mathbb{P}\left(\widehat{Y}_v \neq Z_v\right) - \mathbb{P}\left(Y^\star_v \neq Z_v\right)\right) \leq \left(\frac{2}{3}+\frac{c}{2}\right)\log(|\mathcal{F}|/\delta) + \frac{1}{c}\sum_{v\in V}\mathbb{E}\left[\left(\mathbb{1}\{\widehat{Y}_v \neq Z_v\} - \mathbb{1}\{Y^\star_v \neq Z_v\}\right)^2\right]$$
$$= \left(\frac{2}{3}+\frac{c}{2}\right)\log(|\mathcal{F}|/\delta) + \frac{1}{c}\sum_{v\in V}\mathbb{1}\left\{\widehat{Y}_v \neq Y^\star_v\right\}.$$

□

Lemma 6 (Maximal Inequality)**.** For each $i \in [N]$, let $\{T^i_v\}_{v\in V}$ be a random process with each variable T^i_v bounded in absolute value by 1. Define $\sigma_i^2 = \sum_{v\in V}\mathrm{Var}(T^i_v)$. With probability at least $1-\delta$,

$$\sum_{v\in V}(\mathbb{E}[T^i_v] - T^i_v) \leq \frac{2}{3}\log(N/\delta) + \sqrt{2\sigma_i^2\log(N/\delta)} \quad \forall i \in [N]. \tag{26}$$

Proof of Lemma 6. Let us start by writing out the Bernstein bound for the random variable $\sum_{t=1}^n Z^i_t$:

$$\mathbb{P}\left(\sum_{v\in V}(\mathbb{E}[T^i_v] - T^i_v) > \theta\right) \leq \exp\left(-\frac{\theta_i^2}{2\sigma_i^2 + \frac{2}{3}\theta_i}\right).$$

We now consider the family of processes $\{T^i_v\}_{v\in V}$ and see that by union bound we have

$$\mathbb{P}\left(\max_{i\in[N]}\sum_{v\in V}(\mathbb{E}[T^i_v] - T^i_v) - \theta_i > 0\right) \leq \sum_{i\in[N]}\exp\left(-\frac{\theta_i^2}{2\sigma_i^2 + \frac{2}{3}\theta_i}\right).$$

Solving the quadratic formula, it holds that if we take

$$\theta_i \geq \frac{1}{3}\log(N/\delta) + \sqrt{\log^2(N/\delta)/9 + 2\sigma_i^2\log(N/\delta)},$$

then we have

$$\sum_{i\in[N]}\exp\left(-\frac{\theta_i^2}{2\sigma_i^2+\frac{4}{3}}\right) \leq \delta.$$

We can conclude that

$$\mathbb{P}\left(\forall i \in [N], \quad \sum_{v\in V}(\mathbb{E}[T^i_v] - T^i_v) > \frac{1}{3}\log(N/\delta) + \sqrt{\log^2(N/\delta)/9 + 2\sigma_i^2\log(N/\delta)}\right) \leq \delta.$$

□

D Algorithms

The tree inference algorithm from Section 2 and the full tree decomposition inference algorithm, Algorithm 1, rely on the solution of a constrained minimization problem over the edges and vertices of a tree T. This minimization problem is stated in its most general form as Algorithm 2. This problem can be solved efficiently using the following tree-structured graphical model:

- Fix an arbitrary order on T, and let $p(v)$ denote the parent of a vertex v under this order.
- Define variables $s \in \{\pm 1\}^V$ and $C \in \{1,\ldots,K_n\}^V$.
- For each variable $v \in V$ define factor:

$$\psi_v(s_v, s_{p(v),C_v,C_{\delta_+(v)}}) = e^{-\mathbb{1}\{\mathrm{Cost}_v[s_v]\}}\cdot\mathbb{1}\left\{\sum_{u\in\delta_+(v)}C_u \leq C_v - \mathbb{1}\left\{s_v \neq s_{p(v)}\cdot S(v,p(v))\right\}\right\}.$$

With this formulation it is clear that given (s, C) maximizing the potential

$$\psi(s, C) = \prod_{v \in V} \psi_v(s_v, s_{p(v), C_v, C_{\delta_+(v)}})$$

the node labels s are a valid solution for Algorithm 2. Since ψ is a tree-structured MRF the maximizer can be calculated exactly using max-sum message passing (see e.g. Cowell et al. (2006)). The only catch is that naively this procedure's running time will scale as $n^{\mathsf{deg}(T)}$, because each of the variables C_v has a range that scales with n. For example, the range of C_v is $\widetilde{O}(pn)$ for the setup in Section 2. We now show that the structure of the factors can be exploited to perform message passing in polynomial time in $\mathsf{deg}(T)$ and n. In particular, message passing can be performed in time time $\widetilde{O}(K_n n^2)$ for general trees and time $\widetilde{O}(K_n n)$ when T is a path graph.

```
Algorithm 2 TreeDecoder
Input: Tree $T = (V, E)$, $\{\mathrm{Cost}_v\}_{v \in V}$, $\{S(u, v)\}_{(u,v) \in E}$, $K_n \in \mathbb{N}$.
$\hat{s} = \arg\min_{s \in \{\pm 1\}^V} \sum_{v \in V} \mathrm{Cost}_v[s_v]$
 s.t. $\sum_{(u,v) \in E} \mathbb{1}\{s_u \neq s_v \cdot S(u, v)\} \leq K_n$
Return: $\hat{s} \in \{\pm 1\}^V$.
```

To solve TreeDecoder efficiently, we first turn T into a DAG by running a BFS from a given vertex r and directing edges according to the time of discovery. We denote this DAG by $\overrightarrow{T}$. We root this directed tree at r, and denote the parent of a vertex $u \neq r$ by $p(u)$. For $u \in V$, let $\overrightarrow{T}_u$ denote the (directed) subtree rooted at u. Given a labeling Y to the vertices of T, an edge uv for which $s_u \neq s_v \cdot S(u, v)$ is called a *violated* edge.

We now define a table OPT that will be used to store values for sub-problems of Algorithm 2. For $u \neq r$, and budget K, we define $OPT(u, K|1)$ to be the optimal value of the optimization problem in Algorithm 2 over the subtree $\overrightarrow{T}_u$ for budget K, where the label of $p(u)$ is constrained to have value 1. Importantly, the edge $(u, p(u))$ is also considered in the count of violated edges (in addition to the edges in $\overrightarrow{T}_u$). $OPT(u, K| - 1)$ is defined likewise, but for $p(u)$ constrained to label value -1.

$$OPT(u, K|1) = \min_{s \in \{-1,1\}} \min_{\sum_{v \in N_u} K_v = K - \mathbb{1}\{s \neq S_{p(u)} \cdot S(u, p(u))\}} \left(\sum_{v \in N(u)} OPT(v, K_v|s) + \mathrm{Cost}_v[s] \right).$$

Here s is simply the value assigned to u. We constrain the budgets K_v to satisfy $0 \leq K_v \leq |\overrightarrow{T}_v|$ (clearly no subtree $\overrightarrow{T}_v$ can violate more than $|\overrightarrow{T}_v|$ edges). For the sake of readability, we do not include this constraint in the recursive formula above. A similar recursion can be obtained for $OPT(u, K| - 1)$.

One can verify that if we can compute $OPT(u, K|s)$ for all nonroot nodes and all values of $K \leq K_n, s \in \{-1, 1\}$ then we can find the optimum of the problem of our whole tree. To achieve this, simply attach a degree one node r' to the root of the tree, add a directed edge (r', r) and set the label of the root to equal 1. Then we simply solve for $OPT(r, K|1)$, where $S(r, r') = 1$ as well as $OPT(r, K, 1)$, where $S(r', r)$ is -1 and return the minimum of the the values.

For a leaf node w, the value of $OPT(w, K'|s)$ can be calculated as follows: it is $\min(cost[s_w = -1], cost[s_w = 1])$, for $K' \geq 1$. If $K = 0$, it is $cost[s']$ where s' is the unique label not violating the constraint $s \neq s' \cdot S(w, p(w))$

We now show how to calculate $OPT(u, K_u|s)$ for any vertex in the tree, assuming OPT has already been calculated for its children. To do this, we try both values of s_u, and then condition on its value to optimize

$$\min_{\sum_{j \in [1,k]} K_j = K - \mathbb{1}\{s \neq s_{p(u)} \cdot S(u, p(u))\}} \sum_{u \in [1,k]} OPT(j, K_j|s).$$

The function $\sum_{v \in N_u} OPT(v, K_v|s)$ can be minimized using another layer of dynamic programming as follows: For $r \leq s$, let $[r, s]$ be the set of integers between r and s. Assuming we enumerate the vertices in $N(u)$ by

$1, ..., k := |N(u)|$ and setting K_j to be the budget for the jth node, we have the equality

$$\min_{\sum_{j\in[1,k]} K_j = K - \mathbb{1}\{s\neq s_{p(u)}\cdot S(u,p(u))\}} \sum_{u\in[1,k]} OPT(j, K_j|s)$$

$$= \min_{K_1\in[0,K-\mathbb{1}\{s\neq s_{p(u)}\cdot S(u,p(u))\}]} OPT(1, K_1|s) + \min_{\sum_{j\in[2,k]} K_j = K - K_1 - \mathbb{1}\{s\neq s_{p(u)}\cdot S(u,p(u))\}} \sum_{j\in[2,k]} OPT(j, K_j|s).$$

The minimization problem can be solved in time $O(|N(u)|K_n^2)$ time. We first calculate the minimum cost for the first two vertices where the number of constraints violated can range between 1 to K. This can be done in time $O(K^2)$. We then examine the minimum cost for the first three vertices (assuming of course u has at least three descendants) where the number of violated constraints ranges between 0 and K. Since we have the information for the first two vertices, these values can be calculated again in time $O(K^2)$. We repeat this iteration until all descendants of u are considered. It follows that the overall running time of this algorithm is $\sum_{u\in V} |N(u)|K_n^2 = O(nK_n^2)$, since T is a tree.

When T is a path graph each node has a single child, the recursion collapses to time $O(nK_n)$.

E Further techniques for general graphs

Here we give a simple proof that if the minimal degree of G is $\Omega(\log n)$, then there is an algorithm that achieves arbitrarily small error for each vertex as $n \to \infty$ as soon as $q = 1/2 - \epsilon$ is constant.

Theorem 6. *There is an efficient algorithm that guarantees*

$$\mathbb{E}\left[\sum_{v\in v} \mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\}\right] \leq \sum_{v\in V} \exp(-C\mathsf{deg}(v)\epsilon^2(1-2p)^2).$$

for some $C > 0$.

Observe that this rate quickly approaches 0 with n as soon as $\mathsf{deg}(G) = \Omega(\log n)$ (i.e., it has $o(n)$ Hamming error) . On the other hand, if degree is constant (say d), then even when $p = 0$ the rate of this algorithm is only $e^{-dO(\epsilon^2)}n$, so the algorithm does not have the desired property of having error approach 0 as $p \to 0$.

Proof of Theorem 6. Fix a vertex v and, for each vertex u in its neighborhood, define an estimate $S_u = Z_u \cdot X_{uv}$. We can observe that $\mathbb{P}(S_u = Y_v) = (1-p)(1-q) + pq = \frac{1}{2} + \epsilon(1-2p)$. Our algorithm will be to use the estimator $\widehat{Y}_v = \text{Majority}(\{S_u\}_{u\in N(v)})$. Since each S_u is independent, the Hoeffding bound gives that

$$\mathbb{P}(\widehat{Y}_v \neq Y_v) \leq \exp(-C\mathsf{deg}(v)\epsilon^2(1-2p)^2).$$

Taking this prediction for each vertex gives an expected Hamming error bound of

$$\mathbb{E}\left[\sum_{v\in v} \mathbb{1}\left\{\widehat{Y}_v \neq Y_v\right\}\right] \leq \sum_{v\in V} \exp(-C\mathsf{deg}(v)\epsilon^2(1-2p)^2).$$

□