Privately Learning Markov Random Fields

Huanyu Zhang [1] **Gautam Kamath** [*2] **Janardhan Kulkarni** [*3] **Zhiwei Steven Wu** [*4]

Abstract

We consider the problem of learning Markov Random Fields (including the prototypical example, the Ising model) under the constraint of differential privacy. Our learning goals include both *structure learning*, where we try to estimate the underlying graph structure of the model, as well as the harder goal of *parameter learning*, in which we additionally estimate the parameter on each edge. We provide algorithms and lower bounds for both problems under a variety of privacy constraints – namely pure, concentrated, and approximate differential privacy. While non-privately, both learning goals enjoy roughly the same complexity, we show that this is not the case under differential privacy. In particular, only structure learning under approximate differential privacy maintains the non-private logarithmic dependence on the dimensionality of the data, while a change in either the learning goal or the privacy notion would necessitate a polynomial dependence. As a result, we show that the privacy constraint imposes a strong separation between these two learning problems in the high-dimensional data regime.

1. Introduction

Graphical models are a common structure used to model high-dimensional data, which find a myriad of applications in diverse research disciplines, including probability theory, Markov Chain Monte Carlo, computer vision, theoretical computer science, social network analysis, game theory, and computational biology (Levin et al., 2009; Chatterjee, 2005; Felsenstein, 2004; Daskalakis et al., 2011; Geman & Graffigne, 1986; Ellison, 1993; Montanari & Saberi, 2010). While statistical tasks involving general distributions over p variables often run into the curse of dimensionality (i.e., an exponential sample complexity in p), Markov Random Fields (MRFs) are a particular family of undirected graphical models which are parameterized by the "order" t of their interactions. Restricting the order of interactions allows us to capture most distributions which may naturally arise, and also avoids this severe dependence on the dimension (i.e., we often pay an exponential dependence on t instead of p). An MRF is defined as follows, see Section 2 for more precise definitions and notations we will use in this paper.

Definition 1.1. *Let $k, t, p \in \mathbb{N}$, $G = (V, E)$ be a graph on p nodes, and $C_t(G)$ be the set of cliques of size at most t in G. A* Markov Random Field *with alphabet size k and t-order interactions is a distribution $\mathcal{D}$ over $[k]^p$ such that*

$$\Pr_{Z \sim \mathcal{D}}[Z = z] \propto \exp\left(\sum_{I \in C_t(G)} \psi_I(z)\right),$$

where $\psi_I : [k]^p \to \mathbb{R}$ depends only on variables in I.

We note that each node corresponds to one coordinate of Z. Furthermore, the case when $k = t = 2$ corresponds to the prototypical example of an MRF, the Ising model (Ising, 1925) (Definition 2.1). More generally, if $t = 2$, we call the model *pairwise* (Definition 2.2), and if $k = 2$ but t is unrestricted, we call the model a *binary MRF* (Definition 2.4). In this paper, we mainly look at these two special cases of MRFs.

Given the wide applicability of these graphical models, there has been a great deal of work on the problem of graphical model estimation (Ravikumar et al., 2010; Santhanam & Wainwright, 2012; Bresler, 2015; Vuffray et al., 2016; Klivans & Meka, 2017; Hamilton et al., 2017; Rigollet & Hütter, 2017; Lokhov et al., 2018; Wu et al., 2018). That is, given a dataset generated from a graphical model, can we infer properties of the underlying distribution? Most of the attention has focused on two related learning goals.

1. *Structure learning* (Definition 2.5): Recover the set of non-zero edges in G.
2. *Parameter learning* (Definition 2.6): Recover the set of non-zero edges in G, as well as ψ_I for all cliques I of size at most t.

It is clear that structure learning is no harder than parameter learning. Nonetheless, the sample complexity of both

[*]These authors are in alphabetical order. [1]School of Electrical and Computer Engineering, Cornell University [2]Cheriton School of Computer Science, University of Waterloo [3]Microsoft Research Redmond [4]Computer Science & Engineering, University of Minnesota. Correspondence to: Huanyu Zhang <hz388@cornell.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

learning goals is known to be roughly equivalent. That is, both can be performed using a number of samples which is only *logarithmic* in the dimension p (assuming a model of bounded "width" λ[1]), thus facilitating estimation in very high-dimensional settings.

However, in modern settings of data analysis, we may be running our algorithms on datasets which are sensitive in nature. For instance, graphical models are often used to model medical and genetic data (Friedman et al., 2000; Lagor et al., 2001) – if our learning algorithm reveals too much information about individual datapoints used to train the model, this is tantamount to releasing medical records of individuals providing their data, thus violating their privacy. In order to assuage these concerns, we consider the problem of learning graphical models under the constraint of *differential privacy* (DP) (Dwork et al., 2006), considered by many to be the gold standard of data privacy. Informally, an algorithm is said to be differentially private if its distribution over outputs is insensitive to the addition or removal of a single datapoint from the dataset (a more formal definition is provided in Section 2). Differential privacy has enjoyed widespread adoption, including deployment in Apple (Differential Privacy Team, Apple, 2017), Google (Erlingsson et al., 2014), Microsoft (Ding et al., 2017), and the US Census Bureau for the 2020 Census (Dajani et al., 2017).

Our goal is to design algorithms which guarantee both:

- Accuracy: With high probability, the algorithm learns the underlying graphical model;
- Privacy: For *every* dataset, the algorithm guarantees differential privacy.

Thematically, we investigate the following question: how much additional data is needed to learn Markov Random Fields under the constraint of differential privacy? As mentioned before, absent privacy constraints, the sample complexity is logarithmic in p. Can we guarantee privacy with comparable amounts of data? Or if more data is needed, how much more?

1.1. Results and Techniques

We proceed to describe our results on privately learning Markov Random Fields. In this section, we will assume familiarity with some of the most common notions of differential privacy: pure ε-differential privacy, ρ-zero-concentrated differential privacy, and approximate (ε, δ)-differential privacy. In particular, one should know that these are in (strictly) decreasing order of strength (i.e., an algorithm which satisfies pure DP gives more privacy to the dataset than concentrated DP), formal definitions appear in Section 2. Furthermore, in order to be precise, some of our theorem statements will use notation which is defined later (Section 2) – these may be skipped on a first reading, as our prose will not require this knowledge.

[1]This is a common parameterization of the problem, which roughly corresponds to the graph having bounded-degree, see Section 2 for more details.

Upper Bounds. Our first upper bounds are for parameter learning. First, we have the following theorem, which gives an upper bound for parameter learning pairwise graphical models under concentrated differential privacy, showing that this learning goal can be achieved with $O(\sqrt{p})$ samples. In particular, this includes the special case of the Ising model, which corresponds to an alphabet size $k = 2$. Note that this implies the same result if one relaxes the learning goal to structure learning, or the privacy notion to approximate DP, as these modifications only make the problem easier. Further details are given in Section 3.3.

Theorem 1.2. *There exists an efficient ρ-zCDP algorithm which learns the parameters of a pairwise graphical model to accuracy α with probability at least $2/3$, which takes*

$$n = O\left(\frac{\lambda^2 k^5 \log(pk) e^{O(\lambda)}}{\alpha^4} + \frac{\sqrt{p}\lambda^2 k^{5.5} \log^2(pk) e^{O(\lambda)}}{\sqrt{\rho}\alpha^3}\right)$$

samples.

This result can be seen as a private adaptation of the elegant work of (Wu et al., 2018) (which in turn builds on the structural results of (Klivans & Meka, 2017)). Wu, Sanghavi, and Dimakis (Wu et al., 2018) show that ℓ_1-constrained logistic regression suffices to learn the parameters of all pairwise graphical models. We first develop a private analog of this method, based on the private Franke-Wolfe method of Talwar, Thakurta, and Zhang (Talwar et al., 2014; 2015), which is of independent interest. This method is studied in Section 3.1.

Theorem 1.3. *If we consider the problem of private sparse logistic regression, there exists an efficient ρ-zCDP algorithm that produces a parameter vector w^{priv}, such that with probability at least $1 - \beta$, the empirical risk satisfies*

$$\mathcal{L}(w^{priv}; D) - \mathcal{L}(w^{erm}; D) = O\left(\frac{\lambda^{\frac{4}{3}} \log(\frac{np}{\beta})}{(n\sqrt{\rho})^{\frac{2}{3}}}\right).$$

We note that Theorem 1.3 avoids a polynomial dependence on the dimension p in favor of a polynomial dependence on the "sparsity" parameter λ. The greater dependence on p which arises in Theorem 1.2 is from applying Theorem 1.3 and then using composition properties of concentrated DP.

We go on to generalize the results of (Wu et al., 2018), showing that ℓ_1-constrained logistic regression can also learn the parameters of binary t-wise MRFs. This result is novel even in the non-private setting. Due to the page limit, we defer coverage of binary MRFs to the supplement.

The following theorem shows that we can learn the parameters of binary t-wise MRFs with $\tilde{O}(\sqrt{p})$ samples.

Theorem 1.4. *Let $\mathcal{D}$ be an unknown binary t-wise MRF with associated polynomial h. Then there exists an ρ-zCDP algorithm which learns the maximal monomials of h to accuracy α, given n i.i.d. samples $Z^1, \cdots, Z^n \sim \mathcal{D}$, where*

$$n = O\left(\frac{e^{5\lambda t}\sqrt{p}\log^2(p)}{\sqrt{\rho}\alpha^{\frac{9}{2}}} + \frac{t\lambda^2\sqrt{p}\log p}{\rho\alpha^2} + \frac{e^{6\lambda t}\log(p)}{\alpha^6}\right).$$

To obtain the rate above, our algorithm uses the Private Multiplicative Weights (PMW) method by (Hardt & Rothblum, 2010) to estimate all parity queries of all orders no more than t. The PMW method runs in time exponential in p, since it maintains a distribution over the data domain. We can also obtain an *oracle-efficient* algorithm that runs in polynomial time when given access to an empirical risk minimization oracle over the class of parities. By replacing PMW with such an oracle-efficient algorithm FEM in (Vietri et al., 2019), we obtain a slightly worse sample complexity

$$n = O\left(\frac{e^{5\lambda t}\sqrt{p}\log^2(p)}{\sqrt{\rho}\alpha^{\frac{9}{2}}} + \frac{t\lambda^2\sqrt{p^3}\log p}{\rho\alpha^2} + \frac{e^{6\lambda t}\log(p)}{\alpha^6}\right).$$

For the special case of structure learning under approximate differential privacy, we provide a significantly better algorithm. In particular, we can achieve an $O(\log p)$ sample complexity, which improves exponentially on the above algorithm's sample complexity of $O(\sqrt{p})$. The following is a representative theorem statement for pairwise graphical models, though we derive similar statements for binary MRFs of higher order.

Theorem 1.5. *There exists an efficient (ε, δ)-differentially private algorithm which, with probability at least $2/3$, learns the structure of a pairwise graphical model, which requires a sample complexity of*

$$n = O\left(\frac{\lambda^2 k^4 \exp(14\lambda)\log(pk)\log(1/\delta)}{\varepsilon\eta^4}\right),$$

where η is the minimum parameter weight in absolute value. The detailed definition is in Section 2.

This result can be derived using stability properties of non-private algorithms. In particular, in the non-private setting, the guarantees of algorithms for this problem recover the entire graph *exactly* with high probability. This allows us to derive private algorithms at a multiplicative cost of $O(\log(1/\delta)/\varepsilon)$ samples, using either the propose-test-release framework (Dwork & Lei, 2009) or stability-based histograms (Korolova et al., 2009; Bun et al., 2015). Further details are given in Section 5.

Lower Bounds. We note the significant gap between the aforementioned upper bounds: in particular, our more generally applicable upper bound (Theorem 1.2) has a $O(\sqrt{p})$ dependence on the dimension, whereas the best known lower bound is $\Omega(\log p)$ (Santhanam & Wainwright, 2012). However, we show that our upper bound is tight. That is, even if we relax the privacy notion to approximate differential privacy, *or* relax the learning goal to structure learning, the sample complexity is still $\Omega(\sqrt{p})$. Perhaps surprisingly, if we perform both relaxations simultaneously, this falls into the purview of Theorem 1.5, and the sample complexity drops to $O(\log p)$.

First, we show that even under approximate differential privacy, learning the parameters of a graphical model requires $\Omega(\sqrt{p})$ samples. The formal statement is given in Section 4.

Theorem 1.6 (Informal)**.** *Any algorithm which satisfies approximate differential privacy and learns the parameters of a pairwise graphical model with probability at least $2/3$ requires $\mathrm{poly}(p)$ samples.*

This result is proved by constructing a family of instances of binary pairwise graphical models (i.e., Ising models) which encode product distributions. Specifically, we consider the set of graphs formed by a perfect matching with edges $(2i, 2i+1)$ for $i \in [p/2]$. In order to estimate the parameter on every edge, one must estimate the correlation between each such pair of nodes, which can be shown to correspond to learning the mean of a particular product distribution in ℓ_∞-distance. This problem is well-known to have a gap between the non-private and private sample complexities, due to methods derived from fingerprinting codes (Bun et al., 2014; Dwork et al., 2015; Steinke & Ullman, 2017), or differentially private Fano's inequality (Acharya et al., 2020).

Second, we show that learning the structure of a graphical model, under either pure or concentrated differential privacy, requires $\mathrm{poly}(p)$ samples. The formal theorem appears in Section 6.

Theorem 1.7 (Informal)**.** *Any algorithm which satisfies pure or concentrated differential privacy and learns the structure of a pairwise graphical model with probability at least $2/3$ requires $\mathrm{poly}(p)$ samples.*

We derive this result via packing arguments (Hardt & Talwar, 2010; Beimel et al., 2014; Acharya et al., 2020), by showing that there exists a large number (exponential in p) of different binary pairwise graphical models which must be distinguished. The construction of a packing of size m implies lower bounds of $\Omega(\log m)$ and $\Omega(\sqrt{\log m})$ for learning under pure and concentrated differential privacy, respectively.

1.1.1. Summary and Discussion

We summarize our findings on privately learning Markov Random Fields in Table 1, focusing on the specific case of the Ising model. We note that qualitatively similar relation-

ships between problems also hold for general pairwise models as well as higher-order binary Markov Random Fields. Each cell denotes the sample complexity of a learning task, which is a combination of an objective and a privacy constraint. Problems become harder as we go down (as the privacy requirement is tightened) and to the right (structure learning is easier than parameter learning).

The top row shows that both learning goals require only $\Theta(\log p)$ samples to perform absent privacy constraints, and are thus tractable even in very high-dimensional settings or when data is limited. However, if we additionally wish to guarantee privacy, our results show that this logarithmic sample complexity is only achievable when one considers structure learning under approximate differential privacy. If one changes the learning goal to parameter learning, *or* tightens the privacy notion to concentrated differential privacy, then the sample complexity jumps to become polynomial in the dimension, in particular $\Omega(\sqrt{p})$. Nonetheless, we provide algorithms which match this dependence, giving a tight $\Theta(\sqrt{p})$ bound on the sample complexity.

Due to space restrictions, details of our results on t-wise MRFs and several proofs appear in the supplement.

1.2. Related Work

As mentioned before, there has been significant work in learning the structure and parameters of graphical models, see, e.g., (Chow & Liu, 1968; Csiszár & Talata, 2006; Abbeel et al., 2006; Ravikumar et al., 2010; Jalali et al., 2011a;b; Santhanam & Wainwright, 2012; Bresler et al., 2014; Bresler, 2015; Vuffray et al., 2016; Klivans & Meka, 2017; Hamilton et al., 2017; Rigollet & Hütter, 2017; Lokhov et al., 2018; Wu et al., 2018). Perhaps a turning point in this literature is the work of Bresler (Bresler, 2015), who showed for the first time that general Ising models of bounded degree can be learned in polynomial time. Since this result, following works have focused on both generalizing these results to broader settings (including MRFs with higher-order interactions and non-binary alphabets) as well as simplifying existing arguments. There has also been work on learning, testing, and inferring other statistical properties of graphical models (Bhattacharya & Mukherjee, 2016; Martín del Campo et al., 2016; Daskalakis et al., 2017; Mukherjee et al., 2018; Bhattacharya, 2019). In particular, learning and testing Ising models in statistical distance have also been explored (Daskalakis et al., 2018; Gheissari et al., 2018; Devroye et al., 2018; Daskalakis et al., 2019; Bezakova et al., 2019), and are interesting questions under the constraint of privacy.

Recent investigations at the intersection of graphical models and differential privacy include (Bernstein et al., 2017; Chowdhury et al., 2019; McKenna et al., 2019). Bernstein et al. (Bernstein et al., 2017) privately learn graphical models by adding noise to the sufficient statistics and use an expectation-maximization based approach to recover the parameters. However, the focus is somewhat different, as they do not provide finite sample guarantees for the accuracy when performing parameter recovery, nor consider structure learning at all. Chowdhury, Rekatsinas, and Jha (Chowdhury et al., 2019) study differentially private learning of Bayesian Networks, another popular type of graphical model which is incomparable with Markov Random Fields. McKenna, Sheldon, and Miklau (McKenna et al., 2019) apply graphical models in place of full contingency tables to privately perform inference.

Graphical models can be seen as a natural extension of product distributions, which correspond to the case when the order of the MRF t is 1. There has been significant work in differentially private estimation of product distributions (Blum et al., 2005; Bun et al., 2014; Dwork et al., 2006; Steinke & Ullman, 2017; Kamath et al., 2019; Cai et al., 2019; Bun et al., 2019). Recently, this investigation has been broadened into differentially private distribution estimation, including sample-based estimation of properties and parameters, see, e.g., (Nissim et al., 2007; Smith, 2011; Bun et al., 2015; Diakonikolas et al., 2015; Karwa & Vadhan, 2018; Acharya et al., 2018; Kamath et al., 2019; Bun et al., 2019). For further coverage of differentially private statistics, see (Kamath & Ullman, 2020).

2. Preliminaries

Given a set of points $Z^1, \cdots, Z^n$, we use superscripts, i.e., Z^i to denote the i-th datapoint. Given a vector $Z \in \mathbb{R}^p$, we use subscripts, i.e., Z_i to denote its i-th coordinate. We also use Z_{-i} to denote the vector after deleting the i-th coordinate, i.e. $Z_{-i} = [Z_1, \cdots, Z_{i-1}, Z_{i+1}, \cdots, Z_p]$.

2.1. Markov Random Field Preliminaries

We first introduce the definition of the Ising model, which is a special case of general MRFs when $k = t = 2$.

Definition 2.1. *The p-variable Ising model is a distribution $\mathcal{D}(A, \theta)$ on $\{-1, 1\}^p$ that satisfies*

$$\Pr(Z = z) \propto \exp\left(\sum_{1 \le i \le j \le p} A_{i,j} z_i z_j + \sum_{i \in [p]} \theta_i z_i\right),$$

where $A \in \mathbb{R}^{p \times p}$ is a symmetric weight matrix with $A_{ii} = 0, \forall i \in [p]$ and $\theta \in \mathbb{R}^p$ is a mean-field vector. The dependency graph of $\mathcal{D}(A, \theta)$ is an undirected graph $G = (V, E)$, with vertices $V = [p]$ and edges $E = \{(i, j) : A_{i,j} \neq 0\}$. The width of $\mathcal{D}(A, \theta)$ is defined as

$$\lambda(A, \theta) = \max_{i \in [p]} \left(\sum_{j \in [p]} |A_{i,j}| + |\theta_i|\right).$$

	Structure Learning	Parameter Learning
Non-private	$\Theta(\log p)$ (folklore)	$\Theta(\log p)$ (folklore)
Approximate DP	$\Theta(\log p)$ (Theorems 5.3)	$\Theta(\sqrt{p})$ (Theorems 3.3 and 4.1)
Zero-concentrated DP	$\Theta(\sqrt{p})$ (Theorems 3.3 and 6.1)	$\Theta(\sqrt{p})$ (Theorems 3.3 and 4.1)
Pure DP	$\Omega(p)$ (Theorem 6.1)	$\Omega(p)$ (Theorem 6.1)

Table 1. Sample complexity (dependence on p) of privately learning an Ising model.

Let $\eta(A,\theta)$ be the minimum edge weight in absolute value, i.e., $\eta(A,\theta) = \min_{i,j\in[p]:A_{i,j}\neq 0} |A_{i,j}|$.

We note that the Ising model is supported on $\{-1,1\}^p$. A natural generalization is to generalize its support to $[k]^p$, and maintain pairwise correlations.

Definition 2.2. *The p-variable pairwise graphical model is a distribution $\mathcal{D}(\mathcal{W},\Theta)$ on $[k]^p$ that satisfies*

$$\Pr(Z=z) \propto \exp\left(\sum_{1\le i\le j\le p} W_{i,j}(z_i,z_j) + \sum_{i\in[p]} \theta_i(z_i)\right),$$

where $\mathcal{W} = \{W_{i,j} \in \mathbb{R}^{k\times k} : i \neq j \in [p]\}$ is a set of weight matrices satisfying $W_{i,j} = W_{j,i}^T$, and $\Theta = \{\theta_i \in \mathbb{R}^k : i \in [p]\}$ is a set of mean-field vectors. The dependency graph of $\mathcal{D}(\mathcal{W},\Theta)$ is an undirected graph $G = (V,E)$, with vertices $V = [p]$ and edges $E = \{(i,j) : W_{i,j} \neq 0\}$. The width of $\mathcal{D}(\mathcal{W},\Theta)$ is defined as

$$\lambda(\mathcal{W},\Theta) = \max_{i\in[p],a\in[k]} \left(\sum_{j\in[p]\setminus i} \max_{b\in[k]} |W_{i,j}(a,b)| + |\theta_i(a)|\right).$$

Define $\eta(\mathcal{W},\Theta) = \min_{(i,j)\in E} \max_{a,b} |W_{i,j}(a,b)|$.

Both the above models only consider pairwise interactions between nodes. In order to capture higher-order interactions, we examine the more-general model of Markov Random Fields (MRFs). In this paper, we will restrict our attention to MRFs over a binary alphabet (i.e., distributions over $\{\pm 1\}^p$). In order to define binary t-wise MRFs, we first need the following definition of multilinear polynomials, partial derivatives and maximal monomials.

Definition 2.3. *Multilinear polynomial is defined as $h : \mathbb{R}^p \to \mathbb{R}$ such that $h(x) = \sum_I \bar{h}(I) \prod_{i\in I} x_i$ where $\bar{h}(I)$ denotes the coefficient of the monomial $\prod_{i\in I} x_i$ with respect to the variables $(x_i : i \in I)$. Let $\partial_i h(x) = \sum_{J:i\notin J} \bar{h}(J\cup\{i\}) \prod_{j\in J} x_j$ denote the partial derivative of h with respect to x_i. We say $I \subseteq [p]$ is a maximal monomial of h if $\bar{h}(J) = 0$ for all $J \supset I$.*

Now we are able to formally define binary t-wise MRFs.

Definition 2.4. *For a graph $G = (V,E)$ on p vertices, let $C_t(G)$ denotes all cliques of size at most t in G. A binary t-wise Markov random field on G is a distribution $\mathcal{D}$ on $\{-1,1\}^p$ which satisfies*

$$\Pr_{Z\sim\mathcal{D}}(Z=z) \propto \exp\left(\sum_{I\in C_t(G)} \varphi_I(z)\right),$$

and each $\varphi_I : \mathbb{R}^p \to \mathbb{R}$ is a multilinear polynomial that depends only on the variables in I. We call G the dependency graph of the MRF and $h(x) = \sum_{I\in C_t(G)} \varphi_I(x)$ the factorization polynomial of the MRF. The width of $\mathcal{D}$ is defined as $\lambda = \max_{i\in[p]} \|\partial_i h\|_1$, where $\|h\|_1 := \sum_I |\bar{h}(I)|$.

Finally, we define two possible goals for learning graphical models. First, the easier goal is *structure learning*, which involves recovering the set of non-zero edges.

Definition 2.5. *An algorithm learns the* structure *of a graphical model if, given samples $Z_1,\dots,Z_n \sim \mathcal{D}$, it outputs a graph $\hat{G} = (V,\hat{E})$ over $V = [p]$ such that $\hat{E} = E$, the set of edges in the dependency graph of $\mathcal{D}$.*

The more difficult goal is *parameter learning*, which requires the algorithm to learn not only the location of the edges, but also their parameter values.

Definition 2.6. *An algorithm learns the* parameters *of an Ising model (resp. pairwise graphical model) if, given samples $Z_1,\dots,Z_n \sim \mathcal{D}$, it outputs a matrix $\hat{A}$ (resp. set of matrices $\hat{\mathcal{W}}$) such that $\max_{i,j\in[p]} |A_{i,j} - \hat{A}_{i,j}| \le \alpha$ (resp. $|W_{i,j}(a,b) - \widehat{W}_{i,j}(a,b)| \le \alpha$, $\forall i \neq j \in [p], \forall a,b \in [k]$).*

Definition 2.7. *An algorithm learns the* parameters *of a binary t-wise MRF with associated polynomial h if, given samples $X^1,\dots,X^n \sim \mathcal{D}$, it outputs another multilinear polynomial u such that that for all maximal monomial $I \subseteq [p]$, $\left|\bar{h}(I) - \bar{u}(I)\right| \le \alpha$.*

2.2. Privacy Preliminaries

A *dataset* $X = (X^1,\dots,X^n) \in \mathcal{X}^n$ is a collection of points from some universe $\mathcal{X}$. We say that two datasets X and X' are neighboring, which are denoted as $X \sim X'$ if they differ in exactly one single point. In our work we consider a few different variants of differential privacy. The first is the standard notion of differential privacy.

Definition 2.8 (Differential Privacy (DP) (Dwork et al., 2006))**.** *A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{S}$ satisfies*

(ε, δ)-differential privacy ((ε, δ)-DP) *if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$, and any event $S \subseteq \mathcal{S}$,*

$$\Pr\left(\mathcal{A}(X) \in S\right) \leq e^{\varepsilon} \Pr\left(\mathcal{A}(X') \in S\right) + \delta.$$

The second is *concentrated differential privacy* (Dwork & Rothblum, 2016). In this work, we specifically consider its refinement *zero-mean concentrated differential privacy* (Bun & Steinke, 2016).

Definition 2.9 (Concentrated Differential Privacy (zCDP) (Bun & Steinke, 2016))**.** *A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{S}$ satisfies ρ-zCDP if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$,*

$$\forall \alpha \in (1, \infty) \quad D_\alpha\left(M(X) || M(X')\right) \leq \rho\alpha,$$

where $D_\alpha\left(M(X) || M(X')\right)$ is the α-Rényi divergence between $M(X)$ and $M(X')$.

The following lemma quantifies the relationships between $(\varepsilon, 0)$-DP, ρ-zCDP and (ε, δ)-DP.

Lemma 2.10 (Relationships Between Variants of DP (Bun & Steinke, 2016))**.** *For every $\varepsilon \geq 0$,*

1. *If $\mathcal{A}$ satisfies $(\varepsilon, 0)$-DP, then $\mathcal{A}$ is $\frac{\varepsilon^2}{2}$-zCDP.*
2. *If $\mathcal{A}$ satisfies $\frac{\varepsilon^2}{2}$-zCDP, then $\mathcal{A}$ satisfies $(\frac{\varepsilon^2}{2} + \varepsilon\sqrt{2\log(\frac{1}{\delta})}, \delta)$-DP for every $\delta > 0$.*

Roughly speaking, $(\varepsilon, 0)$-DP is stronger than zCDP, which is stronger than (ε, δ)-DP with $\delta > 0$.

A crucial property of all the variants of differential privacy is that they can be composed adaptively. By adaptive composition, we mean a sequence of algorithms $\mathcal{A}_1(X), \ldots, \mathcal{A}_T(X)$ where the algorithm $\mathcal{A}_t(X)$ may also depend on the outcomes of the algorithms $\mathcal{A}_1(X), \ldots, \mathcal{A}_{t-1}(X)$.

Lemma 2.11 (Composition of DP (Dwork et al., 2010; Bun & Steinke, 2016))**.** *If $\mathcal{A}$ is an adaptive composition of differentially private algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_T$, then the following two properties hold:*

1. *If $\mathcal{A}_1, \ldots, \mathcal{A}_T$ are $(\varepsilon_0, \delta_1), \ldots, (\varepsilon_0, \delta_T)$-DP for some $\varepsilon_0 \leq 1$, then for every $\delta_0 > 0$, $\mathcal{A}$ is (ε, δ)-DP for $\varepsilon = \varepsilon_0\sqrt{6T\log(1/\delta_0)}$ and $\delta = \delta_0 + \sum_t \delta_t$.*
2. *If $\mathcal{A}_1, \ldots, \mathcal{A}_T$ are $\rho_1, \ldots, \rho_T$-zCDP then $\mathcal{A}$ is ρ-zCDP for $\rho = \sum_t \rho_t$.*

3. Parameter Learning of Pairwise Graphical Models

3.1. Private Sparse Logistic Regression

As a subroutine of our parameter learning algorithm, we will solve the following problem of private sparse logistic regression: given a training data set D consisting of n examples $d^j = (x^j, y^j)$ drawn from a distribution P, where $x^j \in \mathbb{R}^p$ with $\left\|x^j\right\|_\infty \leq 1$ and $y^j \in \{\pm 1\}$, a constraint set $\mathcal{C} = \{w \in \mathbb{R}^p : \|w\|_1 \leq \lambda\}$, we want to minimize population logistic loss $\mathbb{E}_P\left[\log(1 + e^{-Y\langle w, X\rangle})\right]$ subject to privacy constraint. To do so, we will leverage the private Frank-Wolfe (FW) algorithm by (Talwar et al., 2014), which minimizes the empirical risk $\mathcal{L}(w; D) = \frac{1}{n}\sum_{j=1}^n \ell(w; d^j) = \frac{1}{n}\sum_{j=1}^n \log(1 + e^{-y^j\langle w, x^j\rangle})$. We show that their algorithm also satisfies zCDP; meanwhile establishes the empirical loss guarantee and population loss guarantee in sparse logistic regression. These results are stated in Theorem 1.3 and Theorem 3.1, respectively, and we defer the proof to the supplement.

Algorithm 1 $\mathcal{A}_{PFW}(D, \mathcal{L}, \rho, \mathcal{C})$: Private FW Algorithm

Input: Data set: $D = \{d^1, \cdots, d^n\}$, loss function: $\mathcal{L}(w; D) = \frac{1}{n}\sum_{j=1}^n \log(1 + e^{-y^j\langle w, x^j\rangle})$, convex set: $\mathcal{C} = \{w \in \mathbb{R}^p : \|w\|_1 \leq \lambda\}$, iteration times: T, and privacy parameters: ρ
Initialize w from an arbitrary point in $\mathcal{C}$.
for $t = 1$ to $T - 1$ **do**
  $\forall s \in S$, $\alpha_s \leftarrow \langle s, \nabla\mathcal{L}(w; D)\rangle + \text{Lap}\left(0, \frac{L_1\|C\|_1\sqrt{T}}{n\sqrt{\rho}}\right)$.
  $\tilde{w}_t \leftarrow \arg\min_{s \in S} \alpha_s$.
  $w_{t+1} \leftarrow (1 - \mu_t)w_t + \mu_t\tilde{w}_t$, where $\mu_t = \frac{2}{t+2}$.
end for
Output: $w^{priv} = w_T$

Theorem 3.1 (Private sparse logistic regression)**.** *Algorithm 1 satisfies ρ-zCDP. Given a data set D drawn i.i.d. from an unknown distribution P, with probability at least $1 - \beta$ over the randomness of the algorithm and D,*

$$\mathbb{E}_P\left[\ell(w^{priv}; (X, Y))\right] - \min_{w \in C} \mathbb{E}_P\left[\ell(w; (X, Y))\right]$$
$$\leq O\left(\frac{\lambda^{\frac{4}{3}}\log(\frac{np}{\beta})}{(n\sqrt{\rho})^{\frac{2}{3}}} + \frac{\lambda\log\left(\frac{1}{\beta}\right)}{\sqrt{n}}\right).$$

3.2. Privately Learning Ising Models

We first consider the problem of estimating the weight matrix of the Ising model. To be precise, given n i.i.d. samples $\{z^1, \cdots, z^n\}$ generated from an unknown distribution $\mathcal{D}(A, \theta)$, our goal is to design an ρ-zCDP estimator $\hat{A}$ such that with probability at least $\frac{2}{3}$, $\max_{i,j \in [p]}\left|A_{i,j} - \hat{A}_{i,j}\right| \leq \alpha$.

An observation of the Ising model is that for any node Z_i, the probability of $Z_i = 1$ conditioned on the values of the remaining nodes Z_{-i} follows from a sigmoid function. The next lemma comes from (Klivans & Meka, 2017), which formalizes this observation.

Algorithm 2 Privately Learning Ising Models

Input: n samples $\{z^1, \cdots, z^n\}$, where $z^m \in \{\pm 1\}^p$ for $m \in [n]$, an upper bound on $\lambda(A, \theta) \leq \lambda$, privacy parameter ρ

for $i = 1$ to p **do**

$\forall m \in [n]$, $x^m \leftarrow [z^m_{-i}, 1]$, $y^m \leftarrow z^m_i$.

$w^{priv} \leftarrow \mathcal{A}_{PFW}(D, \mathcal{L}, \rho', \mathcal{C})$, where $\rho' = \frac{\rho}{p}$, $D = \{(x^m, y^m)\}_{m=1}^n$, $\mathcal{C} = \{\|w\|_1 \leq 2\lambda\}$, and $\mathcal{L}(w; D) = \frac{1}{n}\sum_{m=1}^n \log\left(1 + e^{-y^m\langle w, x^m\rangle}\right)$.

$\forall j \in p$, $\hat{A}_{i,j} \leftarrow \frac{1}{2} w^{priv}_{\tilde{j}}$, where $\tilde{j} = j$ when $j < i$ and $\tilde{j} = j - 1$ if $j > i$.

end for

Output: $\hat{A} \in \mathbb{R}^{p \times p}$

Lemma 3.2. *Let $Z \sim \mathcal{D}(A, \theta)$ and $Z \in \{-1, 1\}^p$, then $\forall i \in [p]$, $\forall x \in \{-1, 1\}^{[p]\setminus\{i\}}$,*

$$\Pr(Z_i = 1 | Z_{-i} = x) = \sigma\left(\sum_{j \neq i} 2A_{i,j}x_j + 2\theta_i\right) = \sigma(\langle w, x'\rangle).$$

where $w = 2[A_{i,1}, \cdots, A_{i,i-1}, A_{i,i+1}, \cdots, A_{i,p}, \theta_i] \in \mathbb{R}^p$, and $x' = [x, 1] \in \mathbb{R}^p$.

By Lemma 3.2, we can estimate the weight matrix by solving a logistic regression for each node, which is utilized in (Wu et al., 2018) to design non-private estimators. Our algorithm uses the private FW method to solve the per-node logistic regression problem and achieves the the following theoretical guarantee.

Theorem 3.3. *Let $\mathcal{D}(A, \theta)$ be an unknown p-variable Ising model with $\lambda(A, \theta) \leq \lambda$. There exists an efficient ρ-zCDP algorithm which outputs a weight matrix $\hat{A} \in \mathbb{R}^{p \times p}$ such that with probability greater than $2/3$, $\max_{i,j \in [p]} \left|A_{i,j} - \hat{A}_{i,j}\right| \leq \alpha$ if the number of i.i.d. samples satisfies*

$$n = \Omega\left(\frac{\lambda^2 \log(p) e^{12\lambda}}{\alpha^4} + \frac{\sqrt{p}\lambda^2 \log^2(p) e^{9\lambda}}{\sqrt{\rho}\alpha^3}\right).$$

Proof. We first prove that Algorithm 2 satisfies ρ-zCDP. Notice that in each iteration, the algorithm solves a private sparse logistic regression under $\frac{\rho}{p}$-zCDP. Therefore, Algorithm 2 satisfies ρ-zCDP by composition (Lemma 2.11).

For the accuracy analysis, we start by looking at the first iteration ($i = 1$) and showing that $\left|A_{1,j} - \hat{A}_{1,j}\right| \leq \alpha$, $\forall j \in [p]$, with probability greater than $1 - \frac{1}{10p}$.

Given a random sample $Z \sim \mathcal{D}(A, \theta)$, we let $X = [Z_{-1}, 1]$, $Y = Z_1$. From Lemma 3.2, $\Pr(Y = 1 | X = x) = \sigma(\langle w^*, x\rangle)$, where $w^* = 2[A_{1,2}, \cdots, A_{1,p}, \theta_1]$. We also note that $\|w^*\|_1 \leq 2\lambda$, as a consequence of the width constraint of the Ising model.

For any n i.i.d. samples $\{z^m\}_{m=1}^n$ drawn from the Ising model, let $x^m = [z^m_{-1}, 1]$ and $y^m = z^m_1$, it is easy to check that each (x^m, y^m) is the realization of (X, Y). Let w^{priv} be the output of $\mathcal{A}\Big(D, \mathcal{L}, \frac{\rho}{p}, \{w : \|w\|_1 \leq 2\lambda\}\Big)$, where $D = \{(x^m, y^m)\}_{m=1}^n$. By Lemma 3.1, when $n = O\left(\frac{\sqrt{p}\lambda^2 \log^2(p)}{\sqrt{\rho}\gamma^{\frac{3}{2}}} + \frac{\lambda^2 \log(p)}{\gamma^2}\right)$, with probability greater than $1 - \frac{1}{10p}$, $\mathbb{E}_{Z \sim \mathcal{D}(A,\theta)}\left[\mathcal{L}(w^{priv}; (X, Y))\right] - \mathbb{E}_{Z \sim \mathcal{D}(A,\theta)}\left[\mathcal{L}(w^*; (X, Y))\right] \leq \gamma$.

We will use the following lemma from (Wu et al., 2018). Roughly speaking, with the assumption that the samples are generated from an Ising model, any estimator w^{priv} which achieves a small error in the loss $\mathcal{L}$ guarantees an accurate parameter recovery in ℓ_∞ distance.

Lemma 3.4. *Let X, Y be defined above. We suppose the joint distribution of (X, Y) is P, and $\Pr(Y = 1 | X = x) = \sigma(\langle u_1, x\rangle + \theta_1)$ for some $u_1 \in \mathbb{R}^{p-1}$ and $\theta_1 \in \mathbb{R}$. If $\mathbb{E}_{(X,Y)\sim P}\left[\log\left(1 + e^{-Y(\langle u_1, X\rangle + \theta_1)}\right)\right] - \mathbb{E}_{(X,Y)\sim P}\left[\log\left(1 + e^{-Y(\langle u_2, X\rangle + \theta_2)}\right)\right] \leq \gamma$ for some $u_2 \in \mathbb{R}^{p-1}, \theta_2 \in \mathbb{R}$, and $\gamma \leq \frac{1}{2}e^{-4\lambda - 6}$, then $\|u_1 - u_2\|_\infty = O(e^{2\lambda} \cdot \sqrt{\gamma})$.*

By Lemma 3.4, if $\mathbb{E}_{Z \sim \mathcal{D}(A,\theta)}\left[\mathcal{L}(w^{priv}; (X, Y))\right] - \mathbb{E}_{Z \sim \mathcal{D}(A,\theta)}\left[\mathcal{L}(w^*; (X, Y))\right] \leq O\left(\alpha^2 e^{-6\lambda}\right)$, we have $\left\|w^{priv} - w^*\right\|_\infty \leq \alpha$. By replacing $\gamma = \alpha^2 e^{-6\lambda}$, we prove that $\left\|A_{1,j} - \hat{A}_{1,j}\right\|_\infty \leq \alpha$ with probability greater than $1 - \frac{1}{10p}$. Noting that similar argument works for the other iterations and non-overlapping part of the matrix is recovered in different iterations. By union bound over p iterations, we prove that with probability at least $\frac{2}{3}$, $\max_{i,j \in [p]} \left|A_{i,j} - \hat{A}_{i,j}\right| \leq \alpha$. □

3.3. Privately Learning Pairwise Graphical Model over General Alphabet

Next, we study parameter learning for pairwise graphical models over general alphabet. Given n i.i.d. samples $\{z^1, \cdots, z^n\}$ drawn from an unknown distribution $\mathcal{D}(\mathcal{W}, \Theta)$, we want to design an ρ-zCDP estimator $\hat{\mathcal{W}}$ such that with probability at least $\frac{2}{3}$, $\forall i \neq j \in [p], \forall u, v \in [k]$, $\left|W_{i,j}(u, v) - \widehat{W}_{i,j}(u, v)\right| \leq \alpha$. To facilitate our presentation, we assume that $\forall i \neq j \in [p]$, every row (and column)

vector of $W_{i,j}$ has zero mean.[2]

Analogous to Lemma 3.2 for the Ising model, a pairwise graphical model has the following property, which can be utilized to recover its parameters. The proof is similar and we omit it for simplicity.

Lemma 3.5. *Let $Z \sim \mathcal{D}(\mathcal{W}, \Theta)$ and $Z \in [k]^p$. For any $i \in [p]$, any $u \neq v \in [k]$, and any $x \in [k]^{p-1}$,*

$$\begin{aligned}&\Pr\left(Z_i = u | Z_i \in \{u, v\}, Z_{-i} = x\right)\\ =\ &\sigma\left(\sum_{j \neq i} \left(W_{i,j}(u, x_j) - W_{i,j}(v, x_j)\right) + \theta_i(u) - \theta_i(v)\right).\end{aligned}$$

Now we introduce our algorithm. Without loss of generality, we consider estimating $W_{1,j}$ for all $j \in [p]$ as a running example. We fix a pair of values (u, v), where $u, v \in [k]$ and $u \neq v$. Let $S_{u,v}$ be the samples where $Z_1 \in \{u, v\}$. In order to utilize Lemma 3.5, we do the following transformation on the samples in $S_{u,v}$: for the m-th sample z^m, let $y^m = 1$ if $z_1^m = u$, else $y^m = -1$. And x^m is the one hot encoding of the vector $[z_{-1}^m, 1]$, where OneHotEncode(s) is a mapping from $[k]^p$ to $\mathbb{R}^{p \times k}$, and the i-th row is the t-th standard basis vector given $s_i = t$. Then we define $w^* \in \mathbb{R}^{p \times k}$ as follows:

$$\begin{aligned}&w^*(j, \cdot) = W_{1,j+1}(u, \cdot) - W_{1,j+1}(v, \cdot), \forall j \in [p-1];\\ &w^*(p, \cdot) = [\theta_1(u) - \theta_1(v), 0, \cdots, 0].\end{aligned}$$

Lemma 3.5 implies that $\forall t$, $\Pr\left(Y^t = 1\right) = \sigma(\langle w^*, X^t \rangle)$, where $\langle \cdot, \cdot \rangle$ is the element-wise multiplication of matrices. According to the definition of the width of $\mathcal{D}(\mathcal{W}, \Theta)$, $\|w^*\|_1 \leq \lambda k$. Now we can apply the sparse logistic regression in Algorithm 3 to the samples in $S_{u,v}$.

Suppose $w_{u,v}^{priv}$ is the output of the private Frank-Wolfe algorithm, we define $U_{u,v} \in \mathbb{R}^{p \times k}$ as follows: $\forall b \in [k]$,

$$\begin{aligned}U_{u,v}(j, b) &= w_{u,v}^{priv}(j, b) - \frac{1}{k} \sum_{a \in [k]} w_{u,v}^{priv}(j, a), \forall j \in [p-1];\\ U_{u,v}(p, b) &= w_{u,v}^{priv}(p, b) + \frac{1}{k} \sum_{j \in [p-1]} \sum_{a \in [k]} w_{u,v}^{priv}(j, a).\end{aligned} \tag{1}$$

$U_{u,v}$ can be seen as a "centered" version of $w_{u,v}^{priv}$ (for the first $p-1$ rows). It is not hard to see that $\langle U_{u,v}, x \rangle = \langle w_{u,v}^{priv}, x \rangle$, so $U_{u,v}$ is also a minimizer of the sparse logistic regression.

[2]The assumption that $W_{i,j}$ is centered is without loss of generality and widely used in the literature (Klivans & Meka, 2017; Wu et al., 2018). We present the argument here for completeness. Suppose the a-th row of $W_{i,j}$ is not centered, i.e., $\sum_b W_{i,j}(a, b) \neq 0$, we can define $W'_{i,j}(a, b) = W_{i,j}(a, b) - \frac{1}{k} \sum_b W_{i,j}(a, b)$ and $\theta'_i(a) = \theta_i(a) + \frac{1}{k} \sum_b W_{i,j}(a, b)$, and the probability distribution remains unchanged.

Algorithm 3 Privately Learning Pairwise Graphical Model

Input: alphabet size k, n i.i.d. samples $\{z^1, \cdots, z^n\}$, where $z^m \in [k]^p$ for $m \in [n]$, an upper bound on $\lambda(\mathcal{W}, \Theta) \leq \lambda$, privacy parameter ρ

for $i = 1$ to p **do**

  for each pair $u \neq v \in [k]$ **do**

    $S_{u,v} \leftarrow \{z^m, m \in [n] : z_i^m \in \{u, v\}\}$.

    $\forall z^m \in S_{u,v}$, $x^m \leftarrow$ OneHotEncode$([z_{-i}^m, 1])$, $y^m \leftarrow 1$ if $z_i^m = u$; $y^t \leftarrow -1$ if $z_i^m = v$.

    $w_{u,v}^{priv} \leftarrow \mathcal{A}_{PFW}(D, \mathcal{L}, \rho', \mathcal{C})$, where $\rho' = \frac{\rho}{k^2 p}$, $D = \{(x^m, y^m) : z^m \in S_{u,v}\}$, $\mathcal{L}(w; D) = \frac{1}{|S_{u,v}|} \sum_{m=1}^{|S_{u,v}|} \log\left(1 + e^{-y^m \langle w, x^m \rangle}\right)$, $\mathcal{C} = \{\|w\|_1 \leq 2\lambda k\}$.

    Define $U_{u,v} \in \mathbb{R}^{p \times k}$ by centering the first $p-1$ rows of $w_{u,v}^{priv}$, as in Equation 1.

  end for

  for $j \in [p] \backslash i$ and $u \in [k]$ **do**

    $\widehat{W}_{i,j}(u, :) \leftarrow \frac{1}{k} \sum_{v \in [k]} U_{u,v}(\tilde{j}, :)$, where $\tilde{j} = j$ when $j < i$ and $\tilde{j} = j - 1$ when $j > i$.

  end for

end for

Output: $\widehat{W}_{i,j} \in \mathbb{R}^{k \times k}$ for all $i \neq j \in [p]$

For now, let we suppose $\forall j \in [p-1], b \in [k]$, $U_{u,v}(j, b)$ is a "good" approximate of $(W_{1,j+1}(u, b) - W_{1,j+1}(v, b))$, which is proved in the supplement. If we sum over $v \in [k]$, it can be shown that $\frac{1}{k} \sum_{v \in [k]} U_{u,v}(j, b)$ is also a "good" approximate of $W_{1,j+1}(u, b)$, for all $j \in [p-1]$, and $u, b \in [k]$, because of the centering assumption of $\mathcal{W}$, i.e., $\forall j \in [p-1], b \in [k], \sum_{v \in [k]} W_{1,j+1}(v, b) = 0$.

The following theorem is the main result of this section, where its proof is structurally similar to that of Theorem 3.3 and we leave it to the supplement.

Theorem 3.6. *Let $\mathcal{D}(\mathcal{W}, \Theta)$ be an unknown p-variable pairwise graphical model distribution, and we suppose that $\mathcal{D}(\mathcal{W}, \Theta)$ has width $\lambda(\mathcal{W}, \Theta) \leq \lambda$. There exists an efficient ρ-zCDP algorithm which outputs $\widehat{W}$ such that with probability greater than $2/3$, $\left|W_{i,j}(u, v) - \widehat{W}_{i,j}(u, v)\right| \leq \alpha$, $\forall i \neq j \in [p], \forall u, v \in [k]$ if the number of i.i.d. samples satisfy*

$$n = \Omega\left(\frac{\lambda^2 k^5 \log(pk) e^{O(\lambda)}}{\alpha^4} + \frac{\sqrt{p} \lambda^2 k^{5.5} \log^2(pk) e^{O(\lambda)}}{\sqrt{\rho} \alpha^3}\right).$$

4. Lower Bounds for Parameter Learning

The lower bound for parameter estimation is based on mean estimation in ℓ_∞ distance. For details on the construction, refer to Section 1.1, and the proof appears in the supplement.

Theorem 4.1. *Suppose $\mathcal{A}$ is an (ε, δ)-differentially private algorithm that takes n i.i.d. samples $Z^1, \ldots, Z^n$ drawn from any unknown p-variable Ising model $\mathcal{D}(A, \theta)$ and outputs $\hat{A}$ such that $\mathbb{E}\left[\max_{i,j\in[p]} |A_{i,j} - \hat{A}_{i,j}|\right] \leq \alpha \leq 1/50$. Then $n = \Omega\left(\frac{\sqrt{p}}{\alpha\varepsilon}\right)$.*

5. Structure Learning of Graphical Models

In this section, we will give an (ε, δ)-differentially private algorithm for learning the *structure* of a Markov Random Field. The dependence on the dimension d will be only *logarithmic*, in comparison to the complexity of privately learning the parameters. The following lemma is immediate from stability-based mode arguments (see, e.g., Proposition 3.4 of (Vadhan, 2017)).

Lemma 5.1. *Suppose there exists a (non-private) algorithm which takes $X = (X^1, \ldots, X^n)$ sampled i.i.d. from some distribution $\mathcal{D}$, and outputs some fixed value Y (which may depend on $\mathcal{D}$) with probability at least $2/3$. Then there exists an (ε, δ)-differentially private algorithm which takes $O\left(\frac{n\log(1/\delta)}{\varepsilon}\right)$ samples and outputs Y with probability at least $1 - \delta$.*

We can now directly import the following theorem from (Wu et al., 2018).

Theorem 5.2 ((Wu et al., 2018))**.** *There exists an algorithm which, with probability at least $2/3$, learns the structure of a pairwise graphical model. It requires $n = O\left(\frac{\lambda^2 k^4 e^{14\lambda}\log(pk)}{\eta^4}\right)$ samples.*

This gives us the following private learning result as a corollary. Similar results for binary MRFs appear in the supplement.

Corollary 5.3. *There exists an (ε, δ)-differentially private algorithm which, with probability at least $2/3$, learns the structure of a pairwise graphical model. It requires $n = O\left(\frac{\lambda^2 k^4 e^{14\lambda}\log(pk)\log(1/\delta)}{\varepsilon\eta^4}\right)$ samples.*

6. Lower Bounds for Structure Learning

In this section we state our structure learning lower bounds under pure DP or zCDP, for learning either Ising models or pairwise graphical models. We show that under both ε-DP and ρ-zCDP, a polynomial dependence on the dimension is unavoidable. Due to the page limit, we defer the proofs to the supplement.

Theorem 6.1. *Any $(\varepsilon, 0)$-DP algorithm which learns the structure of an Ising model with minimum edge weight η requires $n = \Omega\left(\frac{\sqrt{p}}{\eta\varepsilon} + \frac{p}{\varepsilon}\right)$ samples. Furthermore, $n = \Omega\left(\sqrt{\frac{p}{\rho}}\right)$ samples are required for the same task under ρ-zCDP.*

Theorem 6.2. *Any $(\varepsilon, 0)$-DP algorithm which learns the structure of a p-variable pairwise graphical model with minimum edge weight η requires $n = \Omega\left(\frac{\sqrt{p}}{\eta\varepsilon} + \frac{k^2 p}{\varepsilon}\right)$ samples. Furthermore, $n = \Omega\left(\sqrt{\frac{k^2 p}{\rho}}\right)$ samples are required for the same task under ρ-zCDP.*

Acknowledgments

The authors would like to thank Kunal Talwar for suggesting the study of this problem, and Adam Klivans, Frederic Koehler, Ankur Moitra, and Shanshan Wu for helpful and inspiring conversations. GK would like to thank Chengdu Style Restaurant (古月飘香) in Berkeley for inspiration in the conception of this project.

Huanyu Zhang is supported by NSF #1815893 and by NSF #1704443. This work was partially done while the author was an intern at Microsoft Research Redmond.

Gautam Kamath is supported by a University of Waterloo startup grant. Part of this work was done while supported as a Microsoft Research Fellow, as part of the Simons-Berkeley Research Fellowship program, and while visiting Microsoft Research Redmond.

Zhiwei Steven Wu is supported in part by the NSF FAI Award #1939606, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, and a Mozilla Research Grant.

References

Abbeel, P., Koller, D., and Ng, A. Y. Learning factor graphs in polynomial time and sample complexity. *Journal of Machine Learning Research*, 7(Aug):1743–1788, 2006.

Acharya, J., Kamath, G., Sun, Z., and Zhang, H. Inspectre: Privately estimating the unseen. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pp. 30–39. JMLR, Inc., 2018.

Acharya, J., Sun, Z., and Zhang, H. Differentially private assouad, fano, and le cam. *arXiv preprint arXiv:2004.06830*, 2020.

Beimel, A., Brenner, H., Kasiviswanathan, S. P., and Nissim, K. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3):401–437, 2014.

Bernstein, G., McKenna, R., Sun, T., Sheldon, D., Hay, M., and Miklau, G. Differentially private learning of undirected graphical models using collective graphical models. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 478–487. JMLR, Inc., 2017.

Bezakova, I., Blanca, A., Chen, Z., Štefankovič, D., and Vigoda, E. Lower bounds for testing graphical models: Colorings and antiferromagnetic Ising models. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pp. 283–298, 2019.

Bhattacharya, B. B. A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):575–602, 2019.

Bhattacharya, B. B. and Mukherjee, S. Inference in Ising models. *Bernoulli*, 2016.

Blum, A., Dwork, C., McSherry, F., and Nissim, K. Practical privacy: The SuLQ framework. In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pp. 128–138, New York, NY, USA, 2005. ACM.

Bresler, G. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pp. 771–782, New York, NY, USA, 2015. ACM.

Bresler, G., Gamarnik, D., and Shah, D. Structure learning of antiferromagnetic Ising models. In *Advances in Neural Information Processing Systems 27*, NIPS '14, pp. 2852–2860. Curran Associates, Inc., 2014.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pp. 635–658, Berlin, Heidelberg, 2016. Springer.

Bun, M., Ullman, J., and Vadhan, S. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pp. 1–10, New York, NY, USA, 2014. ACM.

Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pp. 634–649, Washington, DC, USA, 2015. IEEE Computer Society.

Bun, M., Kamath, G., Steinke, T., and Wu, Z. S. Private hypothesis selection. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19. Curran Associates, Inc., 2019.

Cai, T. T., Wang, Y., and Zhang, L. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

Chatterjee, S. *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University, June 2005.

Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

Chowdhury, A. R., Rekatsinas, T., and Jha, S. Data-dependent differentially private parameter learning for directed graphical models. *arXiv preprint arXiv:1905.12813*, 2019.

Csiszár, I. and Talata, Z. Consistent estimation of the basic neighborhood of Markov random fields. *The Annals of Statistics*, 34(1):123–145, 2006.

Dajani, A. N., Lauger, A. D., Singer, P. E., Kifer, D., Reiter, J. P., Machanavajjhala, A., Garfinkel, S. L., Dahl, S. A., Graham, M., Karwa, V., Kim, H., Lelerc, P., Schmutte, I. M., Sexton, W. N., Vilhuber, L., and Abowd, J. M. The modernization of statistical disclosure limitation at the U.S. census bureau, 2017. Presented at the September 2017 meeting of the Census Scientific Advisory Committee.

Daskalakis, C., Mossel, E., and Roch, S. Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture. *Probability Theory and Related Fields*, 149(1):149–189, 2011.

Daskalakis, C., Dikkala, N., and Kamath, G. Concentration of multilinear functions of the Ising model with applications to network data. In *Advances in Neural Information Processing Systems 30*, NIPS '17. Curran Associates, Inc., 2017.

Daskalakis, C., Dikkala, N., and Kamath, G. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pp. 1989–2007, Philadelphia, PA, USA, 2018. SIAM.

Daskalakis, C., Dikkala, N., and Kamath, G. Testing Ising models. *IEEE Transactions on Information Theory*, 65 (11):6829–6852, 2019.

Devroye, L., Mehrabian, A., and Reddad, T. The minimax learning rate of normal and Ising undirected graphical models. *arXiv preprint arXiv:1806.06887*, 2018.

Diakonikolas, I., Hardt, M., and Schmidt, L. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 2566–2574. Curran Associates, Inc., 2015.

Differential Privacy Team, Apple. Learning with privacy at scale. https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/

appledifferentialprivacysystem.pdf, December 2017.

Ding, B., Kulkarni, J., and Yekhanin, S. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems 30*, NIPS '17, pp. 3571–3580. Curran Associates, Inc., 2017.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, STOC '09, pp. 371–380, New York, NY, USA, 2009. ACM.

Dwork, C. and Rothblum, G. N. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pp. 265–284, Berlin, Heidelberg, 2006. Springer.

Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pp. 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

Dwork, C., Smith, A., Steinke, T., Ullman, J., and Vadhan, S. Robust traceability from trace amounts. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pp. 650–669, Washington, DC, USA, 2015. IEEE Computer Society.

Ellison, G. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

Erlingsson, Ú., Pihur, V., and Korolova, A. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS '14, pp. 1054–1067, New York, NY, USA, 2014. ACM.

Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

Geman, S. and Graffigne, C. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, pp. 1496–1517. American Mathematical Society, 1986.

Gheissari, R., Lubetzky, E., and Peres, Y. Concentration inequalities for polynomials of contracting Ising models. *Electronic Communications in Probability*, 23(76):1–12, 2018.

Hamilton, L., Koehler, F., and Moitra, A. Information theoretic properties of Markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems 30*, NIPS '17. Curran Associates, Inc., 2017.

Hardt, M. and Rothblum, G. N. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pp. 61–70, Washington, DC, USA, 2010. IEEE Computer Society.

Hardt, M. and Talwar, K. On the geometry of differential privacy. In *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing*, STOC '10, pp. 705–714, New York, NY, USA, 2010. ACM.

Ising, E. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

Jalali, A., Johnson, C. C., and Ravikumar, P. K. On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems 24*, NIPS '11, pp. 1935–1943. Curran Associates, Inc., 2011a.

Jalali, A., Ravikumar, P. K., Vasuki, V., and Sanghavi, S. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, AISTATS '11, pp. 378–387. JMLR, Inc., 2011b.

Kamath, G. and Ullman, J. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

Kamath, G., Li, J., Singhal, V., and Ullman, J. Privately learning high-dimensional distributions. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pp. 1853–1902, 2019.

Karwa, V. and Vadhan, S. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pp. 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Klivans, A. and Meka, R. Learning graphical models using multiplicative weights. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pp. 343–354, Washington, DC, USA, 2017. IEEE Computer Society.

Korolova, A., Kenthapadi, K., Mishra, N., and Ntoulas, A. Releasing search queries and clicks privately. In *Proceedings of the 18th International World Wide Web*

Conference, WWW '09, pp. 171–180, New York, NY, USA, 2009. ACM.

Lagor, C., Aronsky, D., Fiszman, M., and Haug, P. J. Automatic identification of patients eligible for a pneumonia guideline: comparing the diagnostic accuracy of two decision support models. *Studies in Health Technology and Informatics*, 84(1):493–497, 2001.

Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

Lokhov, A. Y., Vuffray, M., Misra, S., and Chertkov, M. Optimal structure and parameter learning of Ising models. *Science Advances*, 4(3):e1700791, 2018.

Martín del Campo, A., Cepeda, S., and Uhler, C. Exact goodness-of-fit testing for the Ising model. *Scandinavian Journal of Statistics*, 2016.

McKenna, R., Sheldon, D., and Miklau, G. Graphical-model based estimation and inference for differential privacy. *arXiv preprint arXiv:1901.09136*, 2019.

Montanari, A. and Saberi, A. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

Mukherjee, R., Mukherjee, S., and Yuan, M. Global testing against sparse alternatives under Ising models. *The Annals of Statistics*, 46(5):2062–2093, 2018.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on the Theory of Computing*, STOC '07, pp. 75–84, New York, NY, USA, 2007. ACM.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Rigollet, P. and Hütter, J.-C. High dimensional statistics. http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf, 2017. Lecture notes.

Santhanam, N. P. and Wainwright, M. J. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

Smith, A. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pp. 813–822, New York, NY, USA, 2011. ACM.

Steinke, T. and Ullman, J. Between pure and approximate differential privacy. *The Journal of Privacy and Confidentiality*, 7(2):3–22, 2017.

Talwar, K., Thakurta, A., and Zhang, L. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

Talwar, K., Thakurta, A., and Zhang, L. Nearly-optimal private LASSO. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 3025–3033. Curran Associates, Inc., 2015.

Vadhan, S. The complexity of differential privacy. In Lindell, Y. (ed.), *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, chapter 7, pp. 347–450. Springer International Publishing AG, Cham, Switzerland, 2017.

Vietri, G., Tian, G., Bun, M., Steinke, T., and Wu, Z. S. New oracle efficient algorithms for private synthetic data release. *NeurIPS PriML workshop*, 2019.

Vuffray, M., Misra, S., Lokhov, A., and Chertkov, M. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems 29*, NIPS '16, pp. 2595–2603. Curran Associates, Inc., 2016.

Wu, S., Sanghavi, S., and Dimakis, A. G. Sparse logistic regression learns all discrete pairwise graphical models. *arXiv preprint arXiv:1810.11905*, 2018.