Privately Learning Markov Random Fields

Gautam Kamath[*] **Janardhan Kulkarni**[†] **Zhiwei Steven Wu**[‡] **Huanyu Zhang**[§]

Abstract

We consider the problem of learning Markov Random Fields, particularly the special case of the Ising model, under the constraint of differential privacy. This includes 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 clique. We provide algorithms and lower bounds for both problems under a variety of privacy models. While the non-private sample complexity bounds for these two problems are both logarithmic in the dimension, we show that this is not the case under differential privacy. In particular, we investigate the sample complexity bounds for both problems under the constraints of pure, approximate, and concentrated differential privacy. We show that only structure learning under approximate differential privacy has logarithmic sample complexity in the dimension, while a change in either the learning goal or the privacy notion would necessitate a polynomial dependence. This suggests that if we are operating on very high-dimensional data, the aforementioned setting is the only one which is tractable.

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 [LPW09, Cha05, Fel04, DMR11, GG86, Ell93, MS10]. While problems involving general distributions over d variables often run into the curse of dimensionality (i.e., an exponential sample complexity in d), 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 d). In this work, we focus on the fundamental case of *pairwise* graphical models (corresponding to $t = 2$), and binary t-wise MRFs (i.e., with alphabet size 2). Note that, as a special case, this includes the Ising model [Isi25]. More formally, we study the following objects.

Definition 1. *The p-variable pairwise graphical model is a distribution $\mathcal{D}(\mathcal{W}, \Theta)$ on $[k]^p$ that satisfies*

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

where $\mathcal{W} = \{W_{i,j} \in \mathbb{R}^{k \times k} : i \neq j \in [p]\}$ is a set of symmetric weight matrices, 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

[*]University of Waterloo. g@csail.mit.edu.

[†]Microsoft Research Redmond. jakul@microsoft.com.

[‡]University of Minnesota. zsw@umn.edu.

[§]Cornell University. hz388@cornell.edu.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

$G = (V, E)$, 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

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

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

We need the following definition of multilinear polynomial and its partial derivative in order to define binary t-wise MRFs.

Definition 2. *A 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.*

Now we can define binary t-wise MRFs. We use the same characterization of MRFs as [KM17].

Definition 3. *For a graph $G = (V, E)$ on p vertices, let $C_t(G)$ denote 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 $\varphi(x) = \sum_{I \in C_t(G)} \varphi_I(x)$ the factorization polynomial of the MRF. The width of $\mathcal{D}$ is defined as $\lambda = \max_i \|\partial_i h\|_1$, where $\|h\|_1 = \sum_I \left|\bar{h}(I)\right|$.

Given the wide applicability of these graphical models, there has been a great deal of work on the problem of graphical model estimation (see, e.g., [RWL10, SW12, Bre15, VMLC16, KM17, HKM17, RH17, LVMC18, WSD18]). 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. First, the easier goal is *structure learning*, which involves recovering the set of non-zero edges.

Definition 4. *An algorithm learns the* structure *of a graphical model if, given samples $X_1, \ldots, X_n$ drawn i.i.d. from $\mathcal{D}$, 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 5. *An algorithm learns the* parameters *of a pairwise graphical model if, given samples $X_1, \ldots, X_n \sim \mathcal{D}$, it outputs a set of matrices $\hat{\mathcal{W}}$ such that $|W_{i,j}(a, b) - \hat{W}_{i,j}(a, b)| \leq \alpha$, $\forall i \neq j \in [p], \forall a, b \in [k]$. An algorithm learns the* parameters *of a binary t-wise MRF with associated polynomial h if, given samples $X_1, \ldots, X_n \sim \mathcal{D}$, it finds another multilinear polynomial u such that that for all $I \subset [p]$ with $|I| = t$, $\left|\bar{h}(I) - \bar{u}(I)\right| \leq \alpha$.*

We note that, for both learning goals, the sample complexity is known to be only *logarithmic* in the dimension p (assuming a bound on the width of the model), 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, e.g., medical records, or emails. Therefore, it may be necessary to respect the privacy of individuals providing their data. In this work, we consider the problem of learning graphical models under the constraint of *differential privacy*. (We assume the reader is familiar with the notions of pure, concentrated, and approximate differential privacy). Given a set of points $X_1, \ldots, X_n$ in $[k]^p$, we say that two datasets are neighboring if they differ in exactly one point X_i. We wish for our algorithms to guarantee both:

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

A running theme of our inquiry focuses on when estimation in very high-dimensional settings, i.e., when the sample complexity is logarithmic in d.

2 Results and Techniques

We first describe our results for the harder problem of *parameter learning*. On the positive side, we provide the following upper bound for parameter learning of graphical models under concentrated differential privacy:

Theorem 6. *There exists an efficient ρ-zCDP algorithm which learns the parameters of a pairwise graphical model 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.

Theorem 7. *There exists an efficient ρ-zCDP algorithm which finds a multilinear polynomial u such that that with probability greater than $2/3$, for all $I \subset [p]$ with $|I| = t$, $\left|\bar{h}(I) - \bar{u}(I)\right| \leq \alpha$, given n i.i.d. samples $Z^1, \cdots, Z^n \sim \mathcal{D}$, where*

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

Moreover, we have $\|h - u\|_1 \leq O(\alpha \cdot p^t)$, where $\|h\|_1 = \sum_I \left|\bar{h}(I)\right|$.

These results can be seen as a private adaptation of the elegant work of [WSD18] (which in turn builds on the structural results of [KM17]) – in particular, the first term in the above expression is the sample complexity of their non-private algorithm. Wu, Sanghavi, and Dimakis [WSD18] show that ℓ_1-constrained logistic regression suffices to learn the parameters of all pairwise graphical models. By similar techniques, we can show that it also suffices to learn the parameters of binary t-wise MRFs. We first develop a private analog of this method, based on the private Franke-Wolfe method of Talwar, Thakurta, and Zhang [TTZ14, TTZ15], which is of independent interest:

Theorem 8. *If we consider the problem of sparse logistic regression, i.e., $\mathcal{L}(w; D) = \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-y_i\langle w, x_i\rangle})$, with the constraint that $\mathcal{C} = \{w : \|w\|_1 \leq \lambda\}$, and we further assume that $\forall i, \|x_i\|_\infty \leq 1, y_i \in \{\pm 1\}$, there exists an efficient ρ-zCDP algorithm that produces a parameter vector w^{priv}, such that with probability at least $1 - \beta$,*

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

We note that Theorem 8 avoids a polynomial dependence on the dimension p in favor of a polynomial dependence on the "sparsity" parameter λ. On this other hand, the sample complexity of Theorem 6 and Theorem 7 is $O(\sqrt{p})$, a large increase from the non-private complexity of $O(\log p)$. This arises due to Theorem 8 being applied to all p nodes, and composition of zCDP. It is natural to wonder whether this dependence on the dimension can be removed – unfortunately, we show that it can not be, even under the weaker notion of approximate differential privacy.

Theorem 9 (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 [BUV14, DSS$^+$15, SU17].

With this $\mathrm{poly}(p)$ barrier in mind, the question remains: when is graphical model estimation tractable in very high-dimensional settings? We relax our learning goal to *structure learning*, in which we only have to recover the sparsity pattern of the underlying graph. Under approximate differential privacy, we can circumvent this barrier and achieve a $O(\log d)$ sample complexity:

Theorem 10. *There exists an efficient (ε, δ)-differentially private algorithm which, with probability at least $2/3$, learns the structure of a pairwise graphical model. It requires*

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

samples.

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 [DL09] or stability-based histograms [KKMN09, BNSV15]. With this argument in mind, note that the above result can be immediately extended to higher-order Markov Random Fields, given a non-private algorithm with the appropriate guarantees, though we omit the statement of this result due to space restrictions.

Unfortunately, intractability returns if we try to strengthen the privacy notion beyond approximate differential privacy.

Theorem 11 (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 [HT10, BBKN14], 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.

3 Conclusions and Next Steps

Our results provide upper and lower bounds for private estimation of MRFs, under a variety of privacy notions, and both structure and parameter learning. Non-privately, both notions of learning require $O(\log p)$ samples. However, privately, we can only maintain a logarithmic dependence on the dimension if we consider structure learning under approximate differential privacy.

At this point, our results for binary t-wise MRFs only recover the coefficients of the polynomial corresponding to the highest degree terms. We hope to improve our results to estimate all maximal monomials of the polynomial.

Another direction we plan to pursue is better algorithms for privately learning the structure of a graphical model. Our current algorithm incurs a multiplicative cost of $O(\log(1/\delta)/\varepsilon)$ over the non-private algorithm in order to guarantee privacy. We would like to derive algorithms which incur only an additive cost over the non-private algorithms, thus guaranteeing "privacy for free" in various parameter regimes.

Acknowledgments

The authors would like to thank Kunal Talwar for suggesting the study of this problem. GK would like to thank Chengdu Style Restaurant (古月飘香) in Berkeley for inspiration in the conception of this project.

References

[BBKN14] Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3):401–437, 2014.

[BNSV15] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 634–649, Washington, DC, USA, 2015. IEEE Computer Society.

[Bre15] Guy Bresler. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 771–782, New York, NY, USA, 2015. ACM.

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 1–10, New York, NY, USA, 2014. ACM.

[Cha05] Sourav Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University, June 2005.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, STOC '09, pages 371–380, New York, NY, USA, 2009. ACM.

[DMR11] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. 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.

[DSS+15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 650–669, Washington, DC, USA, 2015. IEEE Computer Society.

[Ell93] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

[Fel04] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

[GG86] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, pages 1496–1517. American Mathematical Society, 1986.

[HKM17] Linus Hamilton, Frederic Koehler, and Ankur Moitra. 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.

[HT10] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing*, STOC '10, pages 705–714, New York, NY, USA, 2010. ACM.

[Isi25] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[KKMN09] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In *Proceedings of the 18th International World Wide Web Conference*, WWW '09, pages 171–180, New York, NY, USA, 2009. ACM.

[KM17] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 343–354, Washington, DC, USA, 2017. IEEE Computer Society.

[LPW09] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[LVMC18] Andrey Y. Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of Ising models. *Science Advances*, 4(3):e1700791, 2018.

[MS10] Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

[RH17] Philippe Rigollet and Jan-Christian Hütter. High dimensional statistics. `http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf`, 2017. Lecture notes.

[RWL10] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[SU17] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *The Journal of Privacy and Confidentiality*, 7(2):3–22, 2017.

[SW12] Narayana P. Santhanam and Martin J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

[TTZ14] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

[TTZ15] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Nearly-optimal private LASSO. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3025–3033. Curran Associates, Inc., 2015.

[VMLC16] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems 29*, NIPS '16, pages 2595–2603. Curran Associates, Inc., 2016.

[WSD18] Shanshan Wu, Sujay Sanghavi, and Alexandros G. Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. *arXiv preprint arXiv:1810.11905*, 2018.