Learning the Sherrington-Kirkpatrick Model Even at Low Temperature

Gautam Chandrasekaran[*]
UT Austin

Adam R. Klivans[†]
UT Austin

Abstract

We consider the fundamental problem of learning the parameters of an undirected graphical model or Markov Random Field (MRF) in the setting where the edge weights are chosen at random. For Ising models, we show that a multiplicative-weight update algorithm due to Klivans and Meka learns the parameters in polynomial time for any inverse temperature $\beta \leq \sqrt{\log n}$.

This immediately yields an algorithm for learning the Sherrington-Kirkpatrick (SK) model beyond the high-temperature regime of $\beta < 1$. Prior work breaks down at $\beta = 1$ and requires heavy machinery from statistical physics or functional inequalities. In contrast, our analysis is relatively simple and uses only subgaussian concentration.

Our results extend to MRFs of higher order (such as pure p-spin models), where even results in the high-temperature regime were not known.

arXiv:2411.11174v1 [cs.LG] 17 Nov 2024

1 Introduction

In this work, we revisit the problem of structure learning and parameter recovery in undirected graphical models or Markov Random Fields (MRFs) over a binary alphabet. This class of distributions plays a significant role in various fields, including statistical physics, mathematics, and theoretical computer science. Perhaps the simplest and most well-studied class of graphical models is the Ising model. An Ising model $D_{A,h}$ is a distribution over $\{\pm 1\}^n$ given by the factorization

$$\Pr_{X \sim D_{A,h}}[X = x] \propto \exp\left(\sum_{i<j} A_{ij} x_i x_j + \sum_{i=1}^{n} h_i x_i \right).$$

Here A is called the interaction interaction matrix and h is the external field. The dependency graph of $D_{A,h}$ is the set of edges (i, j) for which A_{ij} is non zero.

The *structure learning* problem is to recover the underlying dependency graph G given only iid samples from $D_{A,h}$. We refer to the more difficult problem of recovering the edges of G and their associated weights as *parameter recovery*. The most commonly studied complexity measure for structure learning is the *width* of the Ising model denoted $\lambda(A, h) = \max_i(\sum_j |A_{ij}| + |h_i|)$. Klivans and Meka [KM17] obtained the first algorithm with nearly optimal sample complexity and running time for learning Ising models (see also [Bre15, VMLC16, HKM17]). Their multiplicative-weight update algorithm *Sparsitron* uses $N = \exp(O(\lambda)) \log n$ samples, runs in time $O(n^2 N)$, and additionally recovers the parameters of the graph.

1.1 Beyond Worst-Case Bounds and the SK Model

Although the above bounds are best possible for all bounded-width Ising models [SW12, KM17], it is natural to ask whether the exponential dependence on λ can be avoided for natural special cases of distributions. A prominent example is the extensively studied *Sherrington-Kirkpatrick* (SK) model. The SK model is the Ising model with factorization $\frac{\beta}{\sqrt{n}} \sum_{i<j} A_{ij} x_i x_j + \sum_{i=1}^{n} h_i x_i$ where the entries of A and h are sampled iid from $\mathcal{N}(0, 1)$. We refer to this distribution as $\mathsf{SK}(\beta)$.

When $\beta < 1$, the SK model is said to be in the high-temperature regime, where various desirable properties such as replica symmetry hold [Tal10]. The SK model experiences a phase transition at $\beta = 1$, and $\beta \geq 1$ is known as the low temperature regime (see [Tal03, Pan13] for a comprehensive discussion). Note that the underlying graph here is almost surely the complete graph, and so the structure learning problem is trivial. The parameter recovery problem, however, is still interesting. The problem had been first studied under the name *spin glass inversion*, and heuristic methods were proposed in [BRO17, MM09b]. Clearly, the width of this model is $O(\beta\sqrt{n})$ and thus the analysis from [KM17] only guarantees a sample complexity and running time that is superpolynomial in n.

For the easier problem (implied by parameter recovery) of recovering a distribution that is close in TV distance, [AJK$^+$24] obtained a polynomial time algorithm for the limited portion of the high-temperature regime ($\beta < \frac{1}{4}$). They rely on the *approximate tensorization of entropy* of these models at high temperatures, which have many implications including the polynomial-time mixing of the Glauber dynamics of these distributions. Furthermore, their work relies on powerful results in functional inequalities [BB19, EKZ22, ABXY24] that provably do not hold for $\beta > \frac{1}{4}$.

To overcome the above barrier at $\beta = \frac{1}{4}$ and to attain parameter recovery, a very nice recent work of Gaitonde and Mossel [GM24] takes a different approach. They first prove that bounding the operator norm of the covariance matrix of the SK model suffices for parameter recovery. They then apply a recent deep theorem on the boundedness of the operator norm of the covariance matrix of the SK model [EAG24, BXY23, BSXY24] for the entire high-temperature regime $\beta < 1$. The boundedness of the operator norm

provably does not hold for $\beta \geq 1$, and hence their techniques cannot work in the low-temperature regime. This suggests the existence of a barrier for efficient learnability beyond the high-temperature regime.

1.2 Our Results

We show that, surprisingly, there is no such barrier. We prove that Sparsitron run with $\exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot \text{poly}(n)$ samples recovers the parameters of the SK model. In particular, for $\beta \leq \sqrt{\log n}$, the algorithm runs in polynomial time. In doing so, we subsume and improve the corresponding result from [GM24].

Theorem 1.1 ([Theorem 3.7](), Informal). *With probability $1 - O(1/n)$ over $D_{A,h} \sim \mathsf{SK}(\beta)$, there exists an algorithm that takes $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot \text{poly}(n, 1/\epsilon)$ iid samples from $D_{A,h}$ and recovers a matrix $\widehat{A}$ and vector $\hat{h}$ such that $\left\| A - \widehat{A} \right\|_\infty \leq \epsilon$ and $\left\| h - \hat{h} \right\|_\infty \leq \epsilon$. The running time is $O(n^2 \cdot N)$.*

As a corollary of the above theorem, we can output an Ising model that is close in TV distance to the ground truth.

Corollary 1.2 ([Corollary 3.8]()). *With probability $1 - O(1/n)$ over $D_{A,h} \sim \mathsf{SK}(\beta)$, there exists an algorithm that takes $\exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot \text{poly}(n, 1/\epsilon)$ iid samples from $D_{A,h}$ and outputs an Ising model $D_{\widehat{A},\hat{h}}$ such that $\mathsf{d}_{\mathsf{TV}}(D_{\widehat{A},\hat{h}}, D_{A,h}) \leq \epsilon$. The running time is $O(n^2 \cdot N)$.*

We note that for $\beta < 1$, the running time of [GM24] is $\exp(O(1/(1 - \beta))) \cdot \text{poly}(n)$. For $\beta > 1 - o(1/\log n)$, this running time becomes super polynomial. Thus, our running time is better even in the high temperature regime for $\beta > 1 - o(1/\log n)$. A feature of our proof is that the Gaussian external field case does not require any additional work, whereas the corresponding proof of [GM24] relies on a highly technical argument that does not apply to the full high-temperature regime. Finally, we note that the analysis of [GM24] only implies learnability with probability $1 - O(1/\log n)$ over the choice $D_{A,h} \sim \mathsf{SK}(\beta)$. On the other hand, our algorithm works with probability $1 - O(1/n)$.

In fact, our techniques solve a more general problem: given samples from an Ising model with arbitrary dependency graph G and random weights, recover the graph and the parameters of the model. It is not clear if the techniques from prior work [AJK$^+$24, GM24] solve this problem, even in the high-temperature regime. Formally, we work in the following model:

Definition 1.3 (Random Ising Model). *Let G be a graph on n vertices with maximum degree d and let $\beta > 0$. We define the distribution $\mathcal{D}_{G,\beta}$ over Ising models as follows: a distribution $D \sim \mathcal{D}_{G,\beta}$ has weight matrix A such that*

$$A_{ij} = \begin{cases} \frac{\beta}{\sqrt{d+1}} \cdot \mathcal{N}(0,1) \text{ (or } \frac{\beta}{\sqrt{d+1}} \cdot Rad(1/2)) & \text{if } (i,j) \text{ is an edge in } G \\ 0 & \text{otherwise} \end{cases}$$

where each entry is sampled independently. Each entry of the external field h is either 0 or sampled from $\mathcal{N}(0,1)$ (or Rad(1/2)). The parameter β is called the inverse temperature.

Note that in the above definition, the graph G is arbitrary and only the weights of the interaction are random. Some variants of the above definition have been studied under the names *diluted spin glasses* [MPV87, MM09a], or the *Edwards-Anderson* (EA) model [EA75]. When G is instantiated to be the complete graph, the above definition corresponds to $\mathsf{SK}(\beta)$. We obtain the following theorem for parameter recovery.

Theorem 1.4 ([Theorem 3.5](), Informal). *With probability $1 - O(1/n)$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$ (with Gaussian weights), there exists an algorithm that takes $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot \text{poly}(d, \log n, 1/\epsilon)$ iid samples from $D_{A,h}$ and recovers a matrix $\widehat{A}$ and vector $\hat{h}$ such that $\left\| A - \widehat{A} \right\|_\infty \leq \epsilon$ and $\left\| h - \hat{h} \right\|_\infty \leq n\epsilon$. The running time is $O(n^2 \cdot N)$.*

Note that if one is only interested in parameter recovery for the interaction matrix A, then our sample complexity is sub-polynomial in n for bounded degree graphs. In fact, when both the random interaction matrix and external field have rademacher entries, we can recover the parameters exactly (by rounding) with sample complexity sub-polynomial in n and polynomial in $1/\beta$.

Theorem 1.5 (Theorem 3.9). *With probability $1 - O(1/n)$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$ (with rademacher weights), there exists an algorithm that takes $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot O(d \log(n/\beta)/\beta^2)$ samples and recovers matrix A and external field h exactly. The running time is $O(n^2 \cdot N)$.*

The structure learning algorithm is immediate: as long as each non-zero edge has value at least η, run algorithm from Theorem 1.4 with $\epsilon = \eta/2$ and add all edges (i, j) with $\widehat{A}_{ij} \geq \eta/2$ to the graph. In the Gaussian case, via standard Gaussian anticoncentration, with high probability, we have $\eta \geq \Omega(\beta/n^3)$ and thus we need polynomial sample complexity to identify these edges. In the rademacher case, since we exactly recover A, we just look at the non zero entries of A to recover the graph.

Corollary 1.6. *With probability $1 - O(1/n)$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$ (with rademacher weights), there exists an algorithm that takes $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot O(d \log(n/\beta)/\beta^2)$ iid samples from $D_{A,h}$ and recovers the graph G.*

Corollary 1.7. *With probability $1 - O(1/n)$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$ (with Gaussian weights), there exists an algorithm that takes $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot \mathrm{poly}(n, \log n, 1/\beta)$ iid samples from $D_{A,h}$ and recovers the graph G.*

We also obtain closeness in TV distance as a direct corollary.

Corollary 1.8. *With probability $1 - O(1/n)$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$ (with Gaussian weights), there exists an algorithm that takes $\exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot \mathrm{poly}(n, 1/\epsilon)$ iid samples from $D_{A,h}$ and outputs an Ising model $D_{\widehat{A},\widehat{h}}$ such that $\mathsf{d}_{\mathsf{TV}}(D_{\widehat{A},\widehat{h}}, D_{A,h}) \leq \epsilon$. The running time is $O(n^2 \cdot N)$.*

1.3 Our Results: Markov Random Fields

We now move on to the more general problem of parameter recovery and structure learning in higher order Markov Random Fields (MRFs) with random weights. We first give a formal definition of an MRF.

Definition 1.9 (Markov Random Field). *A distribution D_ψ is a t-wise Markov Random Field (MRF) with dependency graph G if*

$$\Pr_{X \sim D_\psi}[X = x] \propto \exp\big(\psi(x)\big) = \exp\left(\sum_{S \in C_t(G)} f_S(x)\right) \tag{1}$$

where $C_t(G)$ are the cliques of G of size at most t and f_S are arbitrary functions on the variables of S. The function $\psi(\mathbf{x})$ is a degree t polynomial called the factorization polynomial of D_ψ.

Note that Ising models are the special case when $t = 2$. The dependency graph can also be defined as the following: the set of neighbors of vertex v is the minimal set S such that X_v and $X_{[n]\setminus S \cup \{v\}}$ are independent conditioned on the variables in S. Furthermore, every distribution with such a dependency structure can be expressed as an MRF [CH71]. Now, the width of the model $\lambda(\psi) = \max_i \|\partial_i \psi\|_1$ where $\|p\|_1$ for a polynomial p is the sum of the absolute values of its coefficients. $\partial_i \psi$ is the partial derivative of ψ with respect to the index i. Similar to the case of the Ising model, Klivans and Meka [KM17] give an algorithm that uses $N = \exp(O(\lambda)) \log n$ samples, runs in time $O(n^t \cdot N)$ and recovers the graph G, see also ([HKM17]).

The first MRF we study is a generalization of the SK model called a pure t-spin model that is also widely studied (see [Tal10, Pan13] for a discussion). The problem of parameter recovery in these models for the full high-temperature region was stated as an open problem in [GM24] (they state a more general problem for mixed t-spin models, but even the special case of pure t-spin models is open and suffers the same technical barriers). This is the MRF with factorization $\psi(x) = \frac{\beta}{n^{(t-1)/2}} \sum_{\alpha \in [n]^t} \widehat{\psi}_\alpha \prod_{j=1}^t x_{\alpha_j}$ where $\widehat{\psi}_\alpha$ are sample iid from $\mathcal{N}(0,1)$ for every multi-index $\alpha \in [n]^t$. Note that $\lambda(\psi) = \Omega(n^{(t-1)/2})$ and hence the running time of Sparsitron is again exponentially large.

For a limited range of the high-temperature regime, the results of [AJK+24] do solve the weaker problem of recovering a distribution that is close in total variation distance. However, their techniques will provably not work for the entire high-temperature range, as these models experience the property of *shattering* in the high-temperature regime, and this provably rules out fast mixing and entropy factorization [AJ24, AMS23, GJK23].

Observing this barrier, [GM24] conjectured that boundedness of certain moment matrices are necessary for learning in the entire high-temperature regime. We prove that this is not the case. We show that it is possible to recover parameters even in the low temperature regime where the aforementioned moment matrices are provably unbounded [EAG24]. We now state our theorem on parameter recovery in pure t-spin models.

Theorem 1.10 (Theorem C.3, Informal)**.** *Let $\mathcal{D}$ be a pure t-spin model with inverse temperature β. Then, we have that with probability at least $1 - O(1/n^t)$ over $D_\psi \sim \mathcal{D}$, there exists an algorithm that draws $N = \exp(O(\beta^2 t^2 + \beta t \sqrt{t \log n})) \cdot \operatorname{poly}(n^t, 1/\epsilon)$ samples and runs in time $O(N \cdot n^t)$ that outputs a t-MRF $D_{\tilde\psi}$ such that (1) $\left\| \psi - \tilde\psi \right\|_1 \le \epsilon$, (2) $d_{\mathsf{TV}}(D_\psi, D_{\tilde\psi}) \le \epsilon$.*

Again, our techniques solve the following more general problem: given iid samples from an MRF with dependency graph G and random coefficients, find G and recover the coefficients. Formally, we work in the following model:

Definition 1.11 (Random MRF)**.** *Let G be a graph on n vertices with maximum degree d. Let $\beta > 0$ be the inverse temperature and $t > 0$ be the degree of the MRF. The t-MRF ψ sampled from the distribution over MRFs $\mathcal{D}_{G,\beta,t}$ is the MRF with factorization polynomial ψ defined as*

$$\psi(x) := \frac{\beta}{(d+1)^{(t-1)/2}} \sum_{S \in C_t(G)} \widehat{\psi}(S) \cdot \prod_{i \in S} x_i$$

where $\{\widehat{\psi}(S)\}_{S \in C_t(G)}$ are sampled iid from $\mathcal{N}(0,1)$ (or $Rad(1/2)$).

We obtain the following theorems on parameter recovery and structure learning in these models.

Theorem 1.12 (Theorem 4.12, Informal)**.** *With probability $1 - O(1/n^t)$ over $D_\psi \sim D_{G,\beta,t}$ (with Gaussian weights), there exists an algorithm that draws $N = \exp(O(\beta^2 t + \beta t \sqrt{\log n})) \cdot \operatorname{poly}(n^t, 1/\epsilon)$ iid samples from D_ψ and runs in time $O(N \cdot n^t)$ that outputs a t-MRF $D_{\tilde\psi}$ such that (1) $\left\| \psi - \tilde\psi \right\|_1 \le \epsilon$, (2) $d_{\mathsf{TV}}(D_\psi, D_{\tilde\psi}) \le \epsilon$.*

Corollary 1.13 (Theorem 4.11, Informal)**.** *With probability $1 - O(1/n^t)$ over $D_\psi \sim D_{G,\beta,t}$ (with Gaussian weights), there exists an algorithm that draws $N = \exp(O(\beta^2 + \beta \sqrt{\log n})) \cdot \operatorname{poly}(n^t, 1/\beta))$ iid samples from D_ψ and runs in time $O(N \cdot n^t)$ that recovers the graph G.*

Similar to our results for the Ising model, we obtain improved bounds when the weights are rademacher. We give an algorithm that draws only a sub-polynomial number of samples and recovers ψ exactly.

Theorem 1.14 (Theorem 4.13, Informal)**.** *With probability $1 - O(1/n^t)$ over $D_\psi \sim D_{G,\beta,t}$ (with rademacher weights), there exists an algorithm that draws $N = \exp(O(\beta^2 + \beta \sqrt{\log n})) \cdot O(t d^t \log(n/\beta)/\beta^2)$ iid samples from D_ψ and runs in time $O(N \cdot n^t)$ that recovers the polynomial ψ exactly.*

1.4 Related Work

The problem of designing efficient algorithms for learning undirected graphical models has a long history of research [CL68, WLR06, AKN06, BMS08, NBSS10, TR14]. Current work on structure learning and parameter recovery for MRFs primarily focuses on the setting where λ, the sum of the absolute values of each node's edge weights, is bounded. In an important result, Bresler [Bre15] gave the first polynomial-time algorithm for constant λ that did not take any further assumptions on the underlying graph (such as correlation decay). His algorithm, however, has sample complexity that is doubly exponential in λ. Bresler's result was extended to general MRFs by Hamilton et al. [HKM17]. Vuffray et al. [VMLC16] gave the first efficient algorithm for learning Ising models with nearly optimal sample complexity (singly exponential in λ) but with suboptimal running time.

Using a different approach, Klivans and Meka [KM17] obtained the first algorithm for learning Ising models with both near-optimal sample complexity and near-optimal running time. Their algorithm also extends to MRFs and non-binary alphabets. In a follow-up work, Wu et al. [WSD19] showed how to replace the algorithm in [KM17] with an algorithm for sparse logistic regression. Their analysis follows the same framework as [KM17] and results in a slightly improved sample-complexity bound for large alphabets. The Wu et al. algorithm was extended to the case of MRFs in [ZKKW20] by closely following the analysis of the MRF case in [KM17]. Vuffray et al. [VML22] extended these results to more general models. Other recent works that extend Ising and MRF learning to different settings include [GKK19, PSBR20, MMS21, DKSS21, DDDK21, BGPV21, GMM24].

2 Preliminaries

Given a set $S \subseteq [n]$ and a vector $x \in \{\pm 1\}^n$, we use x_S to denote the vector obtained from x by restricting to the indices in the set S. Similarly we use $x_{\overline{S}}$ to denote the x restricted to coordinates outside S. We use $\chi_S(x)$ to denote the monomial $\prod_{i \in S} x_i$. Given a polynomial p, $\widehat{p}(S)$ refers to the coefficient of χ_S in p. $\mathcal{N}(0,1)$ refers to the univariate standard normal distribution. $\mathrm{Rad}(1/2)$ refers to the Rademacher random variable that takes ± 1 with probability $1/2$. For $x, y \in \{\pm 1\}^n$, the Hamming distance between x and y is defined as $d_H(x,y) := \sum_{i \in [n]} \mathbb{1}\{x_i \neq y_i\}$. Given a polynomial $p = \sum_{S \subseteq [n]} \widehat{p}(S)\chi_S(x)$, the partial derivative with respect to set T is defined as $\partial_T p(x) := \sum_{S \supseteq T} \widehat{p}(S)\chi_S(x)$. When $S = \{i\}$, we use the notation $\partial_i p$ for the polynomial $\partial_{\{i\}} p$. A maximal monomial of p is any set $S \subseteq [n]$ such that $\widehat{p}(T) = 0$ for all $T \supset S$. The sigmoid function $\sigma : \mathbb{R} \to \mathbb{R}$ is defined as $\sigma(x) := \frac{1}{1+e^{-x}}$.

Fact 2.1 (Properties of subgaussian distributions, [Ver18]). *For $\lambda > 0$, a random variable X is λ-subgaussian if there exists an absolute constant C such that $\Pr[|X| \geq t] \leq 2 \cdot \exp(-(t/\lambda)^2)$; $\mathbb{E}[\exp((X/(\lambda C))^2)] \leq 2$; and $\mathbb{E}[\exp(rX)] \leq \exp(C^2 r^2 \lambda^2)$, for all $r \in \mathbb{R}$.*

Fact 2.2 (Anti-Lipschitzness of the Sigmoid). *For $a, b \in \mathbb{R}$, we have that $|\sigma(a) - \sigma(b)| \geq \exp(-|a| - 3) \cdot \min(1, |a - b|)$.*

Fact 2.3. *Let D_ψ be a t-MRF with factorization polynomial ψ. Then, for $x \in \{\pm 1\}^n$ and $i \in [n]$, we have that $\Pr_{X \sim D}[X_i = 1 \mid X_{[n]\setminus\{i\}} = x_{[n]\setminus\{i\}}] = \sigma(2\partial_i \psi(x))$*

3 Learning Random Ising Models

In this section we present the argument for learning random Ising models with Gaussian external fields. We first review the Sparsitron algorithm and analysis from [KM17]. We then overview the techniques used by Gaitonde and Mossel [GM24] to solve the high-temperature case and show how we sidestep their technical barriers to achieve learnability in the low-temperature regime.

3.1 The Analysis of Sparsitron

Klivans and Meka [KM17] proved that the problem of parameter recovery in Ising models can be reduced to learning sparse generalized linear models with noise. We now sketch their argument (we consider the zero external field case for simplicity). From the definition of the Ising measure, it holds that for any $i \in [n]$, $\Pr_{X \sim D_A}[X_i = 1 \mid X_{[n] \setminus \{i\}} = x] = \sigma(2A_i \cdot x)$. Recall $\sup_{i \in [n]} \|A_i\|_1 \leq \lambda$. First, a multiplicative-weight update algorithm Sparsitron is run to obtain a vector w to minimize $\mathbb{E}_{X \sim D_A}[(\sigma(w \cdot X) - \sigma(A_i \cdot X))^2]$. Formally, their multiplicative-weight update algorithm has the following guarantee:

Theorem 3.1 ([KM17]). *Let $\lambda, \epsilon, \delta > 0$. Let D be a distribution on $\{\pm 1\}^n \times \{\pm 1\}$ where $\Pr[Y = +1|X] = \sigma(w \cdot X)$ for $(X, Y) \sim D$ and vector $w \in \mathbb{R}^n$ with $\|w\|_1 \leq \lambda$. There exists an algorithm that takes $N = O(\lambda^2 (\ln(n/\delta \epsilon))/\epsilon^2)$ independent samples from D, runs in time $O(nN)$, and outputs a vector $\widehat{w}$ such that*

$$\mathbb{E}_{(X,Y) \sim D}\left[(\sigma(w \cdot X) - \sigma(\widehat{w} \cdot X))^2\right] \leq \epsilon$$

with probability at least $1 - \delta$.

To recover the parameters of D_A, they run the above algorithm for each $i \in [n]$ with ϵ set to $\exp(-\Omega(\lambda))\epsilon^2$ and output the recovered vectors. The final sample complexity is $N = \exp(O(\lambda)) \log(n/\epsilon)/\epsilon^4$ and the running time is $O(n^2 \cdot N)$. The source of their (unavoidable in the worst case) exponential dependence on λ lies in their proof reducing parameter recovery to obtaining small squared loss. The proof of this reduction goes through two steps.

- (*Step 1*) Observe that the sigmoid function satisfies a weak anti-lipschitz property. Formally, we have for any $a, b \in \mathbb{R}$, $|\sigma(a) - \sigma(b)| \geq \exp(-|a| - 3) \cdot \min(1, |a - b|)$. Now, since $|A_i \cdot X| \leq \lambda$ for any $x \in \{\pm 1\}^n$, any w that attains squared loss less than ϵ^2 must satisfy the following property:

$$\mathbb{E}_{X \sim D_A}[|A_i \cdot X - w \cdot X|^2] \leq \exp(O(\lambda))\epsilon^2. \tag{2}$$

- (*Step 2*) Note that for any $j \in [n]$ and $y \in \{\pm 1\}^{n-1}$, it holds that $\min_{b \in \{\pm 1\}} \Pr_{X \sim D_A}[X_j = b \mid X_{[n] \setminus \{j\}} = y] \geq \exp(-4|A_j \cdot y|)/2 \geq \exp(-O(\lambda))$. A distribution satisfying this property is said to be $\exp(-O(\lambda))$-unbiased. Now, say a vector w has $|w_j - A_{ij}| = \alpha$. Then, for any y in $\{\pm 1\}^n$, there exists a $b \in \{\pm 1\}$ such that $|b(w_j - A_{ij}) + (A_i - w)_{[n] \setminus \{j\}} \cdot y)| \geq \alpha$. Using the fact that D_A is $\exp(-O(\lambda))$-unbiased, it now easily follows that $\mathbb{E}_{X \sim D_A}[|A_i \cdot X - w \cdot X|^2] \geq \exp(-O(\lambda)) \cdot \alpha^2$. From Equation (2), we have that that $|A_{ij} - w_j| = \alpha \leq \exp(O(\lambda)) \cdot \epsilon$. Setting the squared error to $\exp(-\Omega(\lambda)) \cdot \epsilon^2$ implies that $\|A_i - w\|_\infty \leq \epsilon$.

3.2 Learning in the High-Temperature Regime: the Analysis of Gaitonde and Mossel

Gaitonde and Mossel [GM24] showed that the exponential dependence on λ in the reduction from the previous subsection can be avoided when A has iid Gaussian entries. First, they showed that as long as there is some constant C such that $\Pr_{X \sim D_A}[|A_i \cdot X| \leq C] \geq 3/4$, for any w with small squared loss the following holds: $\mathbb{E}_{X \sim D_A}[|A_i \cdot X - w \cdot X|^2] \leq \exp(O(C)) \cdot \epsilon^2$. The intuition is that in a region of constant mass, the anti-lipschitzness of the sigmoid scales much better than $\exp(-O(\lambda))$, namely it scales as $\exp(-O(C))$. More precisely, to improve step one from the previous subsection, they consider the conditional expectation over the event $\{A_i \cdot X \leq C\}$. Note that the expected squared error (after conditioning) is at most a constant factor off the true expectation. To avoid the exponential dependence on λ from step two and recover the weight A_{ij}, they observe the following: if D_A satisfied the additional property that $\Pr_{X \sim D_A}[|A_j \cdot X| \leq C] \geq 3/4$, then, the unbiasedness of the distribution conditioned on the event $\{\max(|A_i \cdot X|, |A_j \cdot X|) \leq C\}$ is effectively at least $\exp(-O(C))$. Thus, it suffices to set the squared error of Sparsitron to $\exp(-\Omega(C)) \cdot \epsilon^2$. Formally, [GM24] proved the following:

Theorem 3.2. *(Theorem 6.1 from [GM24]) Let A be an interaction matrix and $h \in \mathbb{R}^n$ be the external field of an Ising model. Suppose there exists a bound $C \geq 1$ such that the following hold: (1) For each $i \in [j]$, we have $\|A_i\|_\infty \leq C$, and (2) for each $i \in [n]$, it holds with probability at least 3/4 over $X \sim D_{A,h}$ that $|A_i \cdot X + h_i| \leq C$. Let $\widehat{A}$ be a matrix from $\mathbb{R}^{n \times n}$ and $\widehat{h}$ a vector from $\mathbb{R}^n$. Then, we have that for all $i \in [n]$, $j \neq i$ and $\Delta(X) := \big(\sigma(2(A_i \cdot X + h_i) - \sigma(2(\widehat{A}_i \cdot X + \widehat{h}_i)))\big)^2$,*

$$\mathbb{E}_{X \sim D_{A,h}}[\Delta(X)] \geq \exp(-O(C)) \cdot \min\{1, 8(A_{ij} - \widehat{A}_{ij})^2\}. \tag{3}$$

Furthermore, if $\left\|A_i - \widehat{A}_i\right\|_\infty \leq |h_i - \widehat{h}_i|/2n$, then it holds that $\mathbb{E}_{X \sim D_{A,h}}[\Delta(X)] \geq \exp(-O(C)) \cdot \min\{1, 8(h_i - \widehat{h}_i)^2\}$.

They proceed by showing that the conditions of the above lemma hold for the SK model for $\beta < 1$. To do so, they use a recent result on the boundedness of the covariance matrix of the SK model at high temperatures [EAG24, BXY23, BSXY24]. These works show that $\left\|\mathbb{E}_{X \sim D_A}[XX^T]\right\|_{op} \leq \frac{2}{(1-\beta)^2}$ with probability at least $1 - o(1)$ over the randomness in A. From the definition of the operator norm and an application of Chebyshev's inequality, this implies that for any unit vector v, $\Pr_{X \sim D_A}[|v \cdot X| \geq 4/(1 - \beta)] \leq 1/4$. Now, applying this argument along the direction A_i for any $i \in [n]$ implies the conditions required to apply Theorem 3.2 with $C = 4/(1 - \beta)$.

This gives a learning algorithm that has sample complexity and running time $\exp(O(1/(1 - \beta))) \cdot \text{poly}(n)$. It is known that boundedness of the covariance matrix experiences a phase transition at $\beta = 1$ and becomes unbounded for $\beta > 1$ [GM24]. Thus, this approach will provably not work in the low-temperature regime. Furthermore, it is unclear if these covariance bounds hold for random Ising models on arbitrary graphs. We also note that the above argument does not hold for the case of non-zero external fields and requires significantly more technical work [GM24].

3.3 Learning Random Ising Models Beyond the High-Temperature Regime

In this section we prove our main result for learning random Ising models and obtain our claimed results for $\mathsf{SK}(\beta)$ as a corollary. We begin with the following observation: for the task of parameter recovery, we do not need $\mathbb{E}_{X \sim D_{A,h}}[v \cdot X]$ to be small for all directions v, as is implied by a distribution with bounded covariance. Rather, we only care about the directions $\{A_i\}_{i \in [n]}$ corresponding to the rows of the interaction matrix as required by Theorem 3.2. We directly analyze the correlation of $D_{A,h}$ with these directions. We use the additional property that the rows of A are subgaussian to show that with high probability over the choice of A, the random variable $A_i \cdot X + h_i X_i$ is subgaussian with subgaussianity constant $O(\beta^2 + \beta\sqrt{\log n})$.

To do this, we give a simple proof that there exists a universal constant B such that $\mathbb{E}_{X \sim D_{A,h}}[\exp(|A_i \cdot X + h_i X_i|^2/B\beta^2)] \leq \exp(\beta^2) \cdot O(n^2)$ with probability at least $1 - O(1/n)$ over the random choice of A and h (Lemma 3.3). Having shown this, a tail bound on $|A_i \cdot X + h_i X_i|$ follows immediately by using the standard trick of exponentiating and applying Markov's inequality (Lemma 3.4). Our main lemma is as follows:

Lemma 3.3. *With probability $1 - \frac{1}{n}$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$, there exists a universal constant B for all $i \in [n]$, we have*

$$\mathbb{E}_{X \sim D_{A,h}}\big[\exp\big(|A_i \cdot X + h_i X_i|^2/(B\beta^2)\big)\big] \leq n^2 \exp(O(\beta^2)). \tag{4}$$

Proof. We compute $\mathbb{E}_A \mathbb{E}_{X \sim D_A}\big[\exp\big(|A_i \cdot X + h_i X_i|^2/B\beta^2\big)\big]$ for all $i \in [n]$. We choose the universal constant B later in the proof. Let $(A, h)_{-i}$ denote the entries of (A, h) that do not involve the variable i.

These are independent from (A_i, h_i).

$$\mathbb{E}_{A,h} \mathbb{E}_{X \sim D_{A,h}} \big[\exp\big(|A_i \cdot X + h_i X_i|^2 / B\beta^2\big)\big] = \mathbb{E}_{(A,h)_{-i}} \mathbb{E}_{A_i,h_i} \mathbb{E}_{X \sim D_{A,h}} \big[\exp\big(|A_i \cdot X + h_i X_i|^2 / B\beta^2\big)\big]$$

$$= \mathbb{E}_{(A,h)_{-i}} \mathbb{E}_{A_i,h_i} \Big[\sum_{x \in \{\pm 1\}^n} \Pr_{X \sim D_{A,h}}[X = x] \cdot \exp\big(|A_i \cdot x + h_i x_i|^2 / B\beta^2\big)\Big]$$

$$= \mathbb{E}_{(A,h)_{-i}} \mathbb{E}_{A_i,h_i} \Big[\sum_{x \in \{\pm 1\}^n} \frac{\exp(\sum_{j<k} A_{ij} x_i x_j + \sum_{j=1}^n h_j x_j)}{Z(A,h)} \cdot \exp\big(|A_i \cdot x + h_i x_i|^2 / B\beta^2\big)\Big]$$

$$= \mathbb{E}_{(A,h)_{-i}} \mathbb{E}_{A_i,h_i} \Big[\sum_{x \in \{\pm 1\}^n} \frac{\exp(x_i \sum_{j \neq i} A_{ij} x_j + h_i x_i + g_{(A,h)_{-i}}(x))}{Z(A,h)} \cdot \exp\big(|A_i \cdot x + h_i x_i|^2 / B\beta^2\big)\Big]$$

$$= \mathbb{E}_{(A,h)_{-i}} \Big[\sum_{x \in \{\pm 1\}^n} \exp(g_{(A,h)_{-i}}(x)) \mathbb{E}_{A_i,h_i} \Big[\frac{\exp(x_i(A_i \cdot x) + h_i x_i)}{Z(A,h)} \cdot \exp\big(|A_i \cdot x + h_i x_i|^2 / B\beta^2\big)\Big]\Big]$$

$$(5)$$

where $g_{(A,h)_{-i}}(x) = \exp(\sum_{j<k, j \neq i \neq k} A_{jk} x_j x_k) + \sum_{j \neq i} h_j x_j$ only depends on $(A,h)_{-i}$ and is independent of A_i, h_i. To further bound the right hand side above, we need a lower bound on $Z(A,h)$, the partition function. We do so by marginalizing out the dependence of $Z(A,h)$ on A_i, h_i so that we can remove this term from the expectation over A_i, h_i. We have that

$$Z(A,h) = \sum_{x \in \{\pm 1\}^n} \exp(\sum_{j<k} A_{jk} x_j x_k + \sum_{i=1}^n h_i x_i) = \sum_{x \in \{\pm 1\}^n} \exp(x_i(A_i \cdot x) + h_i x_i) \exp(g_{(A,h)_{-i}}(x))$$

$$= \frac{1}{2} \sum_{x \in \{\pm 1\}^n} \big(\exp(A_i \cdot x + h_i) + \exp(-A_i \cdot x - h_i) \big) \exp(g_{(A,h)_{-i}}(x))$$

$$\geq \frac{1}{2} \sum_{x \in \{\pm 1\}^n} \exp(g_{(A,h)_{-i}}(x)) \qquad (6)$$

where the third equality follows from the fact that $A_i \cdot x + h_i$ and $g_{(A,h)_{-i}}(x)$ do not depend on x_i. The final inequality follows from the fact that $e^t + e^{-t} \geq 1$ for any $t \in \mathbb{R}$. Combining the Equations (5) and (6), we obtain that

$$\mathbb{E}_{A,h} \mathbb{E}_{X \sim D_{A,h}} \big[\exp\big(|A_i \cdot X + h_i X_i|^2 / 4\beta^2\big)\big]$$

$$= \mathbb{E}_{(A,h)_{-i}} \Big[\sum_{x \in \{\pm 1\}^n} \exp(g_{(A,h)_{-i}}(x)) \mathbb{E}_{A_i,h_i} \Big[\frac{\exp(x_i(A_i \cdot x) + h_i x_i)}{Z(A,h)} \cdot \exp\big(|A_i \cdot x + h_i x_i|^2 / B\beta^2\big)\Big]\Big]$$

$$\leq 2 \mathbb{E}_{(A,h)_{-i}} \Big[\sum_{x \in \{\pm 1\}^n} \exp(g_{(A,h)_{-i}}(x)) \mathbb{E}_{A_i,h_i} \Big[\frac{\exp(x_i(A_i \cdot x + h_i) + (A_i \cdot x + h_i x_i)^2 / B\beta^2)}{\sum_{x \in \{\pm 1\}^n} \exp(g_{(A,h)_{-i}}(x))}\Big]\Big]$$

$$= 2 \cdot \mathbb{E}_{(A,h)_{-i}} \Big[\sum_{x \in \{\pm 1\}^n} \frac{\exp(g_{(A,h)_{-i}}(x)) \mathbb{E}_{A_i,h_i}\big[\exp(x_i(A_i \cdot x + h_i) + |A_i \cdot x + h_i x_i|^2 / B\beta^2)\big]}{\sum_{x \in \{\pm 1\}^n} \exp(g_{(A,h)_{-i}}(x))}\Big] \quad (7)$$

To bound the above quantity, it suffices to bound $\mathbb{E}_{A_i,h_i}\big[\exp(x_i(A_i \cdot x + h_i)) \cdot \exp\big(|A_i \cdot x + h_i x_i|^2 / B\beta^2\big)\big]$ for arbitrary x. Since A_i is a vector of d independent $\frac{\beta}{\sqrt{d}}$ subgaussian random variables and h_i is $O(1)$-subgaussian, the distribution of $A_i \cdot x + h_i$ is $O(\beta)$-subgaussian for fixed x. This is because the sum of d

λ-subgaussian random variables is $\lambda\sqrt{d}$-subgaussian. Thus, we have

$$E_{A_i,h_i}\big[\exp(x_i(A_i \cdot x) + h_i x_i) \cdot \exp\big(|A_i \cdot x + h_i x_i|^2/B\beta^2\big)\big]$$
$$\leq \sqrt{\mathop{\mathbb{E}}_{A_i,h_i}\big[\exp(2x_i(A_i \cdot x + h_i))\big] \cdot \mathop{\mathbb{E}}_{A_i,h_i}\big[\exp(2(A_i \cdot x + h_i x_i)^2/B\beta^2)\big]} \leq \exp(O(\beta^2)) \quad (8)$$

where the first inequality follows from Cauchy Schwarz inequality and the last inequality follows from Fact 2.1 when B is an appropriately chosen universal constant. Combining Equations (7) and (8), we obtain that $\mathbb{E}_{A,h}\mathbb{E}_{X\sim D_{A,h}}\big[\exp\big(|A_i \cdot X + h_i X_i|^2/B\beta^2\big)\big] \leq \exp(O(\beta^2))$.

Now, applying Markov's inequality implies that with probability at least $1 - \frac{1}{n^2}$ over A, for a fixed $i \in [n]$, we have that $\mathbb{E}_{X\sim D_{A,h}}\big[\exp\big(|A_i \cdot X + h_i X_i|^2/B\beta^2\big)\big] \leq n^2 \exp(O(\beta^2))$. A union bound over $i \in [n]$ completes the proof. $\qquad\square$

Having proved the above lemma, the following tail bound on $A_i \cdot X + h_i X_i$ immediately follows.

Lemma 3.4. *Let G be a graph of degree d and let $\beta > 0$. With probability $1 - \frac{1}{n}$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$ with interaction matrix A and gaussian external field, for all $i \in [n]$, we have*

$$\Pr_{X\sim D_{A,h}}[|A_i \cdot X + h_i X_i| \leq O(\beta^2 + \beta\sqrt{\log n})] \geq 1 - \frac{1}{n} \quad (9)$$

Proof. The proof follows almost immediately from Lemma 3.3. Observe that with probability $1 - \frac{1}{n}$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$, for any $i \in [n]$, we have that

$$\Pr_{X\sim D_{A,h}}[|A_i \cdot X + h_i X_i| \geq t] = \Pr_{X\sim D_{A,h}}\left[\exp\big(\frac{|A_i \cdot X + h_i X_i|^2}{B\beta^2}\big) \geq \exp\big(\frac{t^2}{B\beta^2}\big)\right]$$
$$\leq \frac{\mathbb{E}_{X\sim D_{A,h}}\left[\exp\big(|A_i \cdot X + h_i X_i|^2/B\beta^2\big)\right]}{\exp\big(\frac{t^2}{B\beta^2}\big)} \leq n^2 \exp(O(\beta^2))\exp\big(-(t/B\beta)^2\big)$$

where B is the universal constant from Lemma 3.3. The first inequality follows from Markov's and the final inequality follows from Lemma 3.3. Setting $t = O(\beta^2 + \beta\sqrt{\log n})$ makes the above probability less than $\frac{1}{n}$. $\qquad\square$

Our theorem on parameter recovery now follows.

Theorem 3.5. *Let G be a graph of degree d and $0 < \epsilon, \delta \leq 1$. With probability at least $1 - O(1/n)$ over $D_{A,h} \sim \mathcal{D}_{G,\beta}$, there exists an algorithm that draws $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot O\big(\frac{\beta^2 d \log n \log(n/\delta\epsilon)}{\epsilon^4}\big)$ samples and runs in time $O(n^2 \cdot N)$ that outputs a matrix $\widehat{A}$ such that $\big\|A - \widehat{A}\big\|_\infty \leq \epsilon$ and $\big\|h - \widehat{h}\big\|_\infty \leq n\epsilon$. The algorithm succeeds with probability $1 - \delta$.*

Proof. Observe that the interaction matrix A and external field h for a random Ising model satisfies $\|A\|_\infty \leq O\big(\frac{\beta}{\sqrt{d}}\sqrt{\log n}\big)$ and $\|h\|_\infty \leq O(\sqrt{\log n})$ with high probability by applying standard subgaussian concentration. Now, Applying Theorem 3.2 with $C = O(\beta^2 + \beta\sqrt{\log n})$ and Theorem 3.1 with error $\epsilon = \exp(-O(\beta^2 + \beta\sqrt{\log n})) \cdot \epsilon^2$ and $\lambda = O(\beta\sqrt{d\log n})$, we obtain the theorem. $\qquad\square$

Our structure learning result immediately follows from the above theorem. From a standard argument that parameter recovery implies closeness in TV distance (Lemma B.3), we immediately obtain the following corollary by setting the error to ϵ/n^2.

Corollary 3.6. *Let G be a graph of degree d and $0 < \epsilon, \delta \leq 1$. With probability at least $1 - O(1/n)$ over $D_{A,h} \sim D_{G,\beta}$, there exists an algorithm that draws $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot O\left(\frac{\beta^2 n^8 d \log n \log(n/\delta\epsilon)}{\epsilon^8}\right)$ samples and runs in time $O(n^2 \cdot N)$ that outputs a matrix $\widehat{A}$ and vector $\widehat{h}$ such that (1) $\mathsf{d}_{\mathsf{KL}}(D_{A,h}, D_{\widehat{A},\widehat{h}}) \leq 2\epsilon^2$ and (2) $\mathsf{d}_{\mathsf{TV}}(D_{A,h}, D_{\widehat{A},\widehat{h}}) \leq \epsilon$. The algorithm succeeds with probability $1 - \delta$.*

We instantiate the above two statements to the SK model (complete graph with Gaussian weights) to obtain the following corollaries.

Theorem 3.7. *Let $\beta > 0$ and $0 < \epsilon, \delta \leq 1$. With probability at least $1 - O(1/n)$ over $D_{A,h} \sim \mathsf{SK}(\beta)$, there exists an algorithm that draws $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot O\left(\frac{\beta^2 n \log n \log(n/\delta\epsilon)}{\epsilon^4}\right)$ samples and runs in time $O(n^2 \cdot N)$ that outputs a matrix $\widehat{A}$ such that $\left\|A - \widehat{A}\right\|_\infty \leq \epsilon$ and $\left\|h - \widehat{h}\right\|_\infty \leq n\epsilon$.*

Corollary 3.8. *Let $\beta > 0$ and $0 < \epsilon, \delta \leq 1$. With probability at least $1 - O(1/n)$ over $D_{A,h} \sim \mathsf{SK}(\beta)$, there exists an algorithm that draws $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot O\left(\frac{\beta^2 n^9 \log n \log(n/\delta\epsilon)}{\epsilon^8}\right)$ samples and runs in time $O(n^2 \cdot N)$ that outputs a matrix $\widehat{A}$ and vector $\widehat{h}$ such that (1) $\mathsf{d}_{\mathsf{KL}}(D_{A,h}, D_{\widehat{A},\widehat{h}}) \leq 2\epsilon^2$ and (2) $\mathsf{d}_{\mathsf{TV}}(D_{A,h}, D_{\widehat{A},\widehat{h}}) \leq \epsilon$.*

In the case where the matrix A and external field are rademacher random variables, we show that we can exactly recover the model using only a sub-polynomial number of samples. This is in contrast to the Gaussian case where our sample complexity was polynomial in n, and we only recovered the model approximately.

Theorem 3.9. *Let G be a graph of degree d and $0 < \epsilon, \delta \leq 1$. With probability at least $1 - O(1/n)$ over $D_{A,h} \sim D_{G,\beta}$ (with rademacher weights), there exists an algorithm that draws $N = \exp(O(\beta^2 + \beta\sqrt{\log n})) \cdot O\left(\frac{d \log(n/\delta\beta)}{\beta^2}\right)$ samples and runs in time $O(n^2 \cdot N)$ that recovers the distribution $D_{A,h}$ exactly.*

This is a special case of a more general theorem (Theorem 4.13). We refer to Appendix D for the proof.

Remark 3.10. *We note that the analysis of Lemma 3.4 and hence all the learning results straightforwardly extend to the case where $h_i \sim \mathcal{N}(\mu, \sigma)$. In this case, for $\gamma = \sqrt{\beta^t + \sigma^2}$, we obtain that $|A_i \cdot X + h_i X_i|$ is at most $O(\mu + \gamma^2 + \gamma\sqrt{\log n})$ with probability at least $1 - O(1/n)$. This implies sample complexity and running time that scale with $\exp(O(\mu + \gamma^2 + \gamma\sqrt{\log n})) \cdot \mathrm{poly}(n)$.*

Remark 3.11. *Note that most of the analysis above would hold even if we had used a simpler algorithm for learning a sigmoid with respect to square loss such as GLMTron [KKSK11]. GLMtron, however, has polynomial sample complexity (in n) when the input norm is $\sqrt{n}$. On the other hand, Sparsitron has sample complexity $O(\lambda^2 \cdot \log n)$. This improved dependence on n is crucial to obtain our sub-polynomial sample complexity bound in Theorem 3.5 and Theorem 3.9.*

4 Learning Random MRFs

We now present the proofs of our result on learning random t-MRFs. First we sketch the analysis. Consider an MRF D_ψ. Note that $\Pr_{X \sim D_\psi}[X_i = 1 \mid X_{[n]\setminus\{i\}} = x] = \sigma(2\partial_i\psi(x))$. We run Sparsitron nodewise to obtain a polynomial p that attains squared error (for learning the sigmoid) of at most ϵ^2 with respect to $\partial_i\psi$. In [KM17], they argue that an MRF with width at most λ is $\exp(-O(\lambda))$-unbiased. They use this to show that squared loss of at most ϵ^2 implies that the maximal monomials of $\partial_i\psi$ and p are $\exp(O(\lambda t))\epsilon$ close. Recovering the maximal monomials is sufficient for structure recovery. As in Section 3, we want to avoid

worst-case bounds (i.e., naively setting $\lambda = \Omega(n^{(t-1)/2})$) by conditioning on some constant probability regions.

We condition on the event $\mathcal{E} := \{x \mid |\psi(x)-\psi(y)| \leq C$ for all y with $d_H(x,y) \leq t\}$ (see Definition 4.2). We argue that D_ψ conditioned on $\mathcal{E}$ behaves similarly to an $\exp(-O(C))$-unbiased distribution and hence obtain parameter recovery without an exponential dependence on λ (Lemmas A.2 and A.4). These results generalize Theorem 3.2 to the case of MRFs. Finally, we argue via a generalization of Lemmas 3.3 and 3.4 that for any random MRF, the event $\mathcal{E}$ occurs with constant probability (Theorem 4.4). In doing so, we bypass the need for boundedness of moment matrices as conjectured by [GM24].

4.1 Subgaussian Derivatives, C-smooth MRFs and their properties

We now define the notion of an MRF with subgaussian derivatives. This is a property that holds true for all the random MRF families that we study in this work. We argue that for any distribution over MRFs $\mathcal{D}$ with subgaussian derivative, $D_\psi \sim \mathcal{D}$ satisfies a deterministic condition (Definition 4.2) with high probability that implies efficient structure learning and parameter recovery by running Sparsitron.

Definition 4.1 (Subgaussian Derivatives). *Let $\mathcal{D}$ be a distribution over factorization polynomials ψ of degree t such that each coefficient of $\psi \sim \mathcal{D}$ is independently picked. Let $\lambda > 0$. We say that $\mathcal{D}$ has λ-subgaussian derivatives if it holds that*

1. *For all $x, y \in \{\pm 1\}^n$ such that $d_H(x,y) \leq t$, we have that the random variable $(\psi(x) - \psi(y))$ is λ-subgaussian where the randomness is over $\psi \sim \mathcal{D}$.*

For any vector $x \in \{\pm 1\}^n$ and set $S \subseteq [n]$, we denote the vector obtained by flipping the coordinates of x in the set S by x^S. Let the function ψ^S be defined as

$$\psi^S(x) := \psi(x) - \psi(x^S) = 2 \sum_{\substack{|T \cap S| \text{ is odd}}} \widehat{\psi}(T) \cdot \chi_T(x).$$

In particular, we have that $\psi^{\{i\}}(x) = 2x_i \cdot \partial_i \psi(x)$. We also refer to $\mathcal{D}$ as a distribution over MRFs because the factorization polynomial uniquely determines the MRF. We now define the deterministic conditions on $D_\psi \sim \mathcal{D}$ that imply efficient learning of D_ψ.

Definition 4.2 (C-smooth MRF). *Let D_ψ be a t-wise MRF with factorization polynomial ψ. Let $\mathcal{E} \subseteq \{\pm 1\}^n$ be the set defined as*

$$\mathcal{E} := \{x \mid |\psi(x) - \psi(y)| \leq C, \text{ for all } y \text{ such that } d_H(x,y) \leq t\}$$

We say that D_ψ is C-smooth if $\Pr_{X \sim D}[X \in \mathcal{E}] \geq \frac{7}{8}$.

To motivate the above definition and why it implies efficient learning, first consider the case of the Ising model $D_{A,h}$. In this case, the above definition corresponds to the property that with constant probability over $X \sim D_{A,h}$, we have that $|A_i \cdot x + h_i| \leq C$ for all $i \in [n]$. This was exactly the property used in the case of learning random Ising models in Section 3.

In the case of t-MRFs, the analysis of [KM17] used the worst case property of δ-unbiasedness of these distributions. They proved that for any distribution D that is δ-unbiased, given a polynomial p of degree at most t and maximal monomial S, it holds that $\Pr_{X \sim D}[|p(X)| \geq |\widehat{p}(S)|] \geq \delta^t$. Using this property and the fact that the width of the model is bounded by λ, they showed how to learn the coefficients of p from samples. Note that $\delta \geq \exp(-\Omega(\lambda))$. Since δ scales with the width, we cannot directly use their analysis directly for learning the polynomial without paying $\exp(O(\lambda))$. Here is where the above condition of C-smoothness helps us. It is easy to argue that for any polynomial p and point x, there exists a y at Hamming distance at most $t - 1$ from x such that $|p(y)| \geq |\widehat{p}(S)|$ (Lemma A.1). Now, using a simple argument, we see that any C-smooth MRF also has a similar anticoncentration property as the one required by [KM17].

12

Lemma 4.3 (Anticoncentration of C-smooth D_ψ)**.** *Let D_ψ be a C-smooth t-MRF. Then for any polynomial p of degree t with maximal monomial S, it holds that $\Pr_{X \sim D_\psi}[|p(X)| \geq |\widehat{p}(S)|] \geq 2^{-(t+1)} \cdot \exp(-2C)$.*

Proof. For any $x \in \mathcal{E}$ and y, z that differ from x only in the coordinates in S, it holds that $|\psi(y) - \psi(z)| \leq 2C$. Let z maximize the quantity $\Pr_{X \sim D_\psi}[X = z \mid X_{[n] \setminus S} = x]$. Thus, for any y with $y_{[n] \setminus S} = x_{[n] \setminus S}$, it holds that

$$\Pr_{X \sim D_\psi}[X = y \mid X_{[n] \setminus s} = x_{[n] \setminus S}] \geq 2^{-t} \cdot \frac{\Pr_{X \sim D_\psi}[X = y]}{\Pr_{X \sim D_\psi}[X = z]} \geq 2^{-t} \exp(-2C)$$

where the first inequality holds from the definition of z and the fact that at least one element has conditional probability greater than 2^{-t}. Let $\mathcal{E}_S$ be the set $\{y \mid x_{[n] \setminus S} = y, x \in \mathcal{E}\}$. Thus, we have

$$\Pr_{X \sim D_\psi}[|p(X)| \geq |\widehat{p}(S)|] \geq \sum_{w \in \mathcal{E}_S} \Pr_{X \sim D_\psi}[|p(X)| \geq |\widehat{p}(S)| \mid X_{[n] \setminus S} = w] \cdot \Pr_{X \sim D_\psi}[X_{[n] \setminus S} = w]$$

$$\geq 2^{-t} \exp(-2C) \Pr_{X \sim D_\psi}[X \in \mathcal{E}_S] \geq 2^{-(t+1)} \cdot \exp(-2C)$$

where the penultimate inequality follows from the fact that there exists some string z that differs from x only in S such that $|p(z)| \geq |\widehat{p}(S)|$. The second inequality follows from the previous argument that the probability of $X = z$ conditioned on $X_{[n] \setminus S} = w$ is at least $2^{-t} \exp(-2C)$. The last inequality follows from the fact that $\Pr_{X \sim D_\psi}[X \sim \mathcal{E}_S] \geq \Pr_{X \sim D_\psi}[X \in \mathcal{E}] \geq \frac{7}{8}$. $\qquad\square$

Although we do not use the above result directly in our proofs, we use similar ideas to extend the analysis of [KM17] to our setting.

We now prove a structural result that says that any MRF with subgaussian derivatives is also *C-smooth* for appropriate choice of C (a generalization of Lemma 3.4). This result contains most of the technical novelty of our work, and we believe it could be useful in other contexts as well.

Theorem 4.4. *Let $\mathcal{D}$ be a distribution over t-wise MRFs such that $\mathcal{D}$ has λ-subgaussian derivatives. Then, with probability at least $1 - \frac{1}{(n+1)^t}$ over $D_\psi \sim \mathcal{D}$, we have that D_ψ is $O(\lambda^2 + \lambda\sqrt{t \log n})$-smooth.*

Proof. Let the function ψ^{-S} be the polynomial containing the coefficients of ψ not in ψ^S. That is, $\psi^{-S}(x) := \sum_{|T \cap S| \text{ is even}} \widehat{\psi}_T \cdot \chi_T(x)$. Clearly, we have that $\psi(x) = \psi^{-S}(x) + \frac{\psi^S(x)}{2}$. We also have that $\psi(x^S) = \psi^{-S}(x) - \frac{\psi^S(x)}{2}$.

We are now ready to start the proof. The result follows by applying Markov's inequality to the following claim.

Claim 4.5. *With probability $1 - \frac{1}{(n+1)^t}$ over $D_\psi \sim \mathcal{D}$, for all $S \subseteq [n]$ with $|S| \leq t$, we have*

$$\mathbb{E}_{X \sim D_\psi}\left[\exp\left((\psi^S(X))^2/B\right)\right] \leq 4(n+1)^{2t} \exp(B) \tag{10}$$

where $B = 2(C\lambda)^2$ for large universal constant C.

Proof. We compute the quantity $\mathbb{E}_{D_\psi \sim \mathcal{D}} \mathbb{E}_{X \sim D_\psi}\left[\exp\left((\psi^S(X))^2/B\right)\right]$ for all sets S with size at most t. We have that

$$\mathbb{E}_{D_\psi \sim \mathcal{D}} \mathbb{E}_{X \sim D_\psi}\left[\exp\left((\psi^S(X))^2/B\right)\right] = \mathbb{E}_{\psi^{-S}} \mathbb{E}_{\psi^S} \mathbb{E}_{X \sim D_\psi}\left[\exp\left((\psi^S(X))^2/B\right)\right]$$

$$= \mathbb{E}_{\psi^{-S}} \mathbb{E}_{\psi^S}\left[\sum_{x \in \{\pm 1\}^n} \Pr_{X \sim D_\psi}[X = x] \cdot \exp\left((\psi^S(x))^2/B\right)\right]$$

$$= \mathbb{E}_{\psi^{-S}} \mathbb{E}_{\psi^S}\left[\sum_{x \in \{\pm 1\}^n} \frac{\exp(\psi(x))}{Z_\psi} \cdot \exp\left((\psi^S(x))^2/B\right)\right] \tag{11}$$

where Z_ψ is the partition function of the MRF defined as $Z_\psi = \sum_{x \in \{\pm 1\}^n} \exp(\psi(x))$. To proceed in bounding the quantity in Equation (11), we first need to decouple Z_ψ from ψ^S as this quantity is in the denominator and hence hard to analyze. To do this, we lower bound Z_ψ using only the polynomial ψ^{-S}. We have that

$$
\begin{aligned}
Z_\psi &= \sum_{x \in \{\pm 1\}^n} \exp(\psi(x)) = \frac{1}{2} \sum_{x \in \{\pm 1\}^n} \left(\exp(\psi(x)) + \exp(\psi(x^S)) \right) \\
&= \frac{1}{2} \sum_{x \in \{\pm 1\}^n} \left(\exp\left(\psi^{-S}(x) + \frac{\psi^S(x)}{2}\right) + \exp\left(\psi^{-S}(x) - \frac{\psi^S(x)}{2}\right) \right) \\
&\geq \frac{1}{2} \sum_{x \in \{\pm 1\}^n} \exp\left(\psi^{-S}(x)\right)
\end{aligned}
\tag{12}
$$

where we obtain the second equality by pairing terms that are equal to each other outside S and the complement of each other in S. The third equality follows from the definition of the polynomials ψ^S and ψ^{-S} and the final inequality follows from the fact that $e^t + e^{-t} \geq 1$ for all $t \in \mathbb{R}$. Combining Equations (11) and (12), we obtain that

$$
\mathop{\mathbb{E}}_{D_\psi \sim \mathcal{D}} \mathop{\mathbb{E}}_{X \sim D_\psi} \left[\exp\left((\psi^S(X))^2/B\right) \right]
$$

$$
\leq 2 \cdot \mathop{\mathbb{E}}_{\psi^{-S}} \mathop{\mathbb{E}}_{\psi^S} \left[\sum_{x \in \{\pm 1\}^n} \frac{\exp(\psi(x))}{\sum_{x \in \{\pm 1\}^n} \exp\left(\psi^{-S}(x)\right)} \cdot \exp\left((\psi^S(x))^2/B\right) \right]
$$

$$
= 2 \cdot \mathop{\mathbb{E}}_{\psi^{-S}} \mathop{\mathbb{E}}_{\psi^S} \left[\sum_{x \in \{\pm 1\}^n} \frac{\exp\left(\psi^{-S}(x) + \frac{\psi^S(x)}{2}\right)}{\sum_{x \in \{\pm 1\}^n} \exp\left(\psi^{-S}(x)\right)} \cdot \exp\left((\psi^S(x))^2/B\right) \right]
$$

$$
= 2 \cdot \mathop{\mathbb{E}}_{\psi^{-S}} \left[\sum_{x \in \{\pm 1\}^n} \frac{\exp(\psi^{-S}(x)) \, \mathbb{E}_{\psi^S} \left[\exp\left(\frac{\psi^S(x)}{2}\right) \cdot \exp\left((\psi^S(x))^2/B\right) \right]}{\sum_{x \in \{\pm 1\}^n} \exp(\psi^{-S}(x))} \right]
$$

where the penultimate equality follows from the fact that $\psi(x) = \psi^{-S}(x) + \frac{\psi^S(x)}{2}$ and the last equality follows linearity of expectation. Clearly, to bound the RHS in the above argument, it suffices to bound $\mathbb{E}_{\psi^S} \left[\exp\left(\frac{\psi^S(x)}{2}\right) \cdot \exp\left((\psi^S(x))^2/B\right) \right]$ pointwise for any $x \in \{\pm 1\}^n$. Note that from the assumption of λ-subgaussian derivatives, we have that the random variable $\psi^S(x)$ is λ-subgaussian. Let Y denote the random variable $\psi^S(x)$. We want to bound $\mathbb{E}_Y[\exp(Y/2) \cdot \exp(Y^2/(2 \cdot (C\lambda)^2))]$ given that Y is λ-subgaussian. We have that

$$
\mathop{\mathbb{E}}_Y \left[\exp(Y/2) \cdot \exp(Y^2/(2 \cdot (C\lambda)^2)) \right] \leq \sqrt{\mathop{\mathbb{E}}_Y \left[\exp(Y)\right] \cdot \mathop{\mathbb{E}}_Y \left[\exp(Y^2/(C\lambda)^2)\right]} \leq 2 \exp(C^2 \lambda^2)
$$

where the first inequality is Cauchy-Schwarz and the last inequality follows from Fact 2.1. Now, combining the above arguments, above gives us $\mathbb{E}_{D_\psi \sim \mathcal{D}} \mathbb{E}_{X \sim D_\psi} \left[\exp\left((\psi^S(X))^2/B\right) \right] \leq 4 \exp(C^2 \lambda^2) = 4 \exp(B)$. Thus, for any fixed S of size at most t, Markov's inequality implies that

$$
\mathop{\Pr}_{D_\psi \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{X \sim D_\psi} \left[\exp\left((\psi^S(X))^2/B\right) \right] \geq 4(n+1)^{2t} \exp(B) \right] \leq \frac{1}{(n+1)^{2t}}.
$$

Since the number of sets of size at most t is bounded above by $(n+1)^t$, a union bound implies the claim. $\qquad\square$

We are now ready to complete the proof of the theorem. Let $D_\psi \sim \mathcal{D}$ be an MRF for which the event from Claim 4.5 holds true. This happens with probability at least $1 - \frac{1}{(n+1)^t}$. For any set $|S| \leq t$, we have that

$$\Pr_{X \sim D_\psi}\left[|\psi^S(X)| \geq r\right] = \Pr_{X \sim D_\psi}\left[(\psi^S(X))^2/B \geq r^2/B\right] \leq \mathbb{E}_{X \sim D_\psi}\left[\exp\left((\psi^S(X))^2/B\right)\right] \cdot \exp(-r^2/B)$$

$$\leq 4(n+1)^{2t}\exp(B)\exp(-r^2/B) \leq \frac{1}{8(n+1)^t}$$

where $B = 2(C\lambda)^2$. The second inequality follows by taking the exponent on both sides and applying Markov's inequality. The third inequality follows from Claim 4.5. The final inequality follows by settings $r = O(B + \sqrt{Bt\log n})$. Now, a union bound over all sets of size at most t completes the proof. $\qquad\square$

4.2 Results on Structure Learning and Parameter Recovery

We now state our results on structure learning and parameter recovery for *C-smooth* MRFs. We need the following standard non-degeneracy condition (introduced in [KM17]) that is required to ensure identifiability of the underlying graph given samples from the MRF.

Definition 4.6 (η-identifiability). *A t-MRF with factorization polynomial ψ with dependency graph G is said to be η-identifiable for $\eta > 0$ if $|\widehat{\psi}(S)| \geq \eta$ for all maximal monomials S in ψ and all edges in G are contained in a monomial of ψ.*

We are now ready to state and prove the theorem on structure learning in *C-smooth* MRFs which are η-identifiable.

Theorem 4.7. *Let $C, \lambda, \eta > 0$. Let D_ψ be a C-smooth t-MRF with dependency graph G that is η-identifiable and suppose $\|\partial_i \psi\|_1 \leq \lambda$ for all $i \in [n]$. Then, there exists an algorithm that draws $N = O\left(\frac{\lambda^2 2^{O(t)}\exp(O(C))}{\eta^4}\log(n/\delta\eta)\right)$ samples from D_ψ and runs in time equal $O\left(N \cdot n^t\right)$ such that it finds the graph G with probability at least $1 - \delta$.*

Proof. The algorithm is exactly the same as Algorithm 3 in [KM17] with an appropriate choice of parameters. We only sketch a proof here since our analysis is almost identical to that of [KM17] except that we use Lemma A.2 instead of Lemma 6.2 from [KM17]. First, for each $i \in [n]$, we obtain polynomials p_i such that

$$\mathbb{E}_{X \sim D_\psi}\left[(\sigma(p_i(X)) - \sigma(2\partial_i\psi(X)))^2\right] \leq \frac{\eta^2}{2^{t+4}\exp(10C + 6)} \tag{13}$$

To do this, we use the property that $\Pr[X_i = 1 \mid X_{[n]\setminus\{i\}}] = \sigma(2\partial_i\psi(X))$ and run the algorithm from Theorem 3.1 after doing a feature expansion of X to the monomial basis containing all monomials of degree less than $t - 1$. We run Sparsitron such for each i the success probability is $O(\delta/n)$. Thus, with probability $O(\delta)$, we have that Equation (13) holds for all $i \in [n]$. The sample complexity so far was $\frac{\lambda^2 \cdot 2^{O(t)}\exp(O(C))}{\eta^4}\log(n/\delta\eta)$.

Now, since D_ψ is C-smooth, we have from Lemma A.2 that

$$\Pr_{X \sim D_\psi}\left[|\widehat{2\partial_i\psi}(S) - \partial_S p_i(X)| > \eta/2\right] \leq \frac{1}{4}.$$

The rest of the analysis is exactly the same as [KM17]. We construct an output graph H iteratively. Draw $K = O(\log(n^t/\delta))$ independent samples from D_ψ. For each $i \in [n]$, and $S \subseteq [t-1]$, evaluate $\frac{\partial_S p_i(X)}{2}$ on each of these samples. If the median of these K evaluations is greater than $\eta/2$, then add the clique on the

vertices in $S \cup \{i\}$ to the graph H. Using the concentration of median, η-identifiability of G and a union bound over all monomials of degree less than t and vertices $i \in [n]$, we have the graph H obtained at the end being equal to G with probability at least $1 - \delta$. The sample complexity is dominated by the number of samples N required for Sparsitron and the running time is at most $O(N \cdot n^t)$ where the n^t dependence comes from the feature expansion and the evaluation of the median on each of the monomials. □

We now state and prove the theorem on learning a distribution $D_{\tilde{\psi}}$ that is close to C-smooth D_ψ in KL divergence and TV distance.

Theorem 4.8. *Let $C, \lambda > 0$ and $0 < \epsilon, \delta < 1$. Let D_ψ be a C-smooth t-MRF and suppose $\|\partial_i \psi\|_1 \leq \lambda$ for all $i \in [n]$. Then, there exists an algorithm that draws $N = \frac{\lambda^2 ((nt)^{O(t)} \exp(O(C)) \log(n/\delta\epsilon)}{\epsilon^8}$ samples from D_ψ and runs in time equal to $O(N \cdot n^t)$ such that it outputs a t-MRF $D_{\tilde{\psi}}$ with factorization polynomial $\tilde{\psi}$ such that (1) $\left\| \psi - \tilde{\psi} \right\|_1 \leq \epsilon^2$, (2) $d_{KL}(D_\psi, D_{\tilde{\psi}}) \leq 2\epsilon^2$, and (3) $d_{TV}(D_\psi, D_{\tilde{\psi}}) \leq \epsilon$. The algorithm succeeds with probability at least $1 - \delta$.*

Proof. After doing a feature expansion, we run the algorithm from Theorem 3.1 nodewise to recover polynomials $\{p_i\}_{i \in [n]}$ such that $\mathbb{E}_{X \sim D_\psi}[(\sigma(p_i(X)) - \sigma(2\partial_i \psi(x)))^2] \leq \epsilon'$ for all $i \in [n]$ with $\epsilon' \leq O(\epsilon^4 \cdot \exp(-10C) \cdot (1/nt)^{2t+2})$. The sample complexity is $N = \frac{\lambda^2 ((nt)^{O(t)} \exp(O(C)) \log(n/\delta\epsilon)}{\epsilon^8}$. From Lemma A.4, we obtain polynomials $\{p_i\}_{i \in [n]}$ such that $\|p_i - \partial_i \psi\|_1 \leq \frac{\epsilon}{n}$ for all $i \in [n]$. Construct a polynomial $\tilde{\psi}$ such that $\widehat{\tilde{\psi}}(S) = \widehat{p}_i(S \setminus \{i\})$ where i is an arbitrary index in S. Observe that $\left\| \psi - \tilde{\psi} \right\|_1 \leq \epsilon^2$. Now, the theorem follows from Lemma B.3. □

We now argue that a random MRF drawn from the distribution over MRFs defined in Definition 1.11 satisfies the properties required for structure learning and parameter recovery in Theorems 4.7 and 4.8 with appropriate choice of parameters.

Lemma 4.9. *For a graph G of degree d and $t > 0$, let $\mathcal{D}_{G,\beta,t}$ be as defined in Definition 1.11 with Gaussian (or Rademacher) coefficients. Then, we have that with probability at least $1 - O(1/n^t)$ over $D_\psi \sim \mathcal{D}_{G,\beta,t}$,*

1. *D_ψ is (a) $\frac{\beta}{n^{5t/2}}$-identifiable in the Gaussian case, (b) $\frac{\beta}{d^{(t-1)/2}}$-identifiable in the Rademacher case,*

2. *for all $i \in [n]$, $\|\partial_i \psi\|_1$ is at most (1) $\beta \cdot d^{(t+1)/2} \sqrt{t \log n}$ in the Gaussian case (2) $\beta \cdot d^{(t+1)/2}$ in the Rademacher case,*

3. *D_ψ is $O(\beta^2 t + \beta t \sqrt{\log n})$-smooth.*

Proof. We begin with the proof of (1). From standard Gaussian anticoncentration, we have that $|\widehat{\psi}(S)| \geq \frac{\beta}{d^{(t-1)/2}} \cdot \eta$ with probability at least $1 - O(\eta)$ for any set S and $D_\psi \sim \mathcal{D}_{G,\beta,t}$. Choosing $\eta = O(1/n^{2t})$ and taking a union bound over all monomials, we obtain that with probability at least $1 - O(1/n^t)$, it holds that $|\widehat{\psi}(S)| \geq \frac{\beta}{n^{5t/2}}$. The bound for the Rademacher case is direct as each variable is $\frac{\beta}{d^{(t-1)/2}} \pm 1$. We now prove (2). From standard Gaussian tail bound, we have that $\|X\|_\infty \leq O(\sqrt{\log k})$ with probability at least $1 - 1/k$ for $X \sim \mathcal{N}(0,1)^k$. Applying this to all monomials, we have that $\|\partial_i \psi\|_1 \leq \beta \cdot d^{(t+1)/2} \sqrt{t \log n}$ for all $i \in [n]$ with probability at least $1 - O(1/n^t)$. Finally for (3), we prove the following lemma.

Lemma 4.10. *For a graph G of degree d and $t > 0$, let $\mathcal{D}_{G,\beta,t}$ be as defined in Definition 1.11 with Gaussian (or Rademacher) coefficients. Then, we have that with probability at least $1 - \frac{1}{(n+1)^t}$, $D_\psi \sim \mathcal{D}_{G,\beta,t}$ is $O(\beta^2 t + \beta t \sqrt{\log n})$-smooth.*

Proof. We prove that $\mathcal{D}_{G,\beta,t}$ has subgaussian derivatives and then use Theorem 4.4. Let ψ be the random function corresponding to the factorization polynomial of an MRF sampled from Theorem 4.4. Consider any $x, y \in \{\pm 1\}^n$ with x and y differing in set S and $|S| \leq t$. From Definition 4.1, it suffices to prove that the random variable $\psi(x) - \psi(y)$ is λ-subgaussian for appropriate λ. By the definition of the set S, we have that

$$\psi(x) - \psi(y) = 2 \sum_{\substack{T \in C_t(G) \\ |T \cap S| \text{ is odd}}} \widehat{\psi}(T) \cdot \chi_T(x).$$

Recall that each coefficient of ψ are independent and identically distributed to $\frac{\beta}{\sqrt{d^t}} Z$ where $Z \sim \mathcal{N}(0,1)$ or $Z \sim \text{Rad}(1/2)$. Note that the number of $T \in C_t(G)$ with which S has non zero intersection is at most $t \cdot \sum_{i=0}^{t-1} \binom{d}{i} \leq O(t \cdot d^{t-1})$. Since both $\mathcal{N}(0,1)$ and $\text{Rad}(1/2)$ are $O(1)$-subgaussian, we have that $\psi(x) - \psi(y)$ is the sum of $O(t \cdot d^{t-1})$ iid random variables which are $\frac{2\beta}{d^{(t-1)/2}}$-subgaussian. Since the sum of k λ-subgaussian variables is $O(\sqrt{k}\lambda)$-subgaussian (Proposition 2.6.1 from [Ver18]), we have that $\psi(x) - \psi(y)$ is $O(\beta\sqrt{t})$-subgaussian. Now, applying Theorem 4.4, we obtain that with probability at least $1 - \frac{1}{(n+1)^t}$, we have that $D_\psi \sim \mathcal{D}_{G,\beta,t}$ is $O(\beta^2 t + \beta t \sqrt{\log n})$-smooth. $\quad\square$

Thus, we have proved that $\mathcal{D}_{G,\beta,t}$ satisfies all the conditions that we need for efficient learnability. $\quad\square$

The following theorems are now immediate from Theorem 4.7 and Lemma B.3 and Lemma 4.9. The first theorem states the sample complexity and running time for structure learning.

Theorem 4.11 (Structure Learning in Random MRFs). *For a graph G of degree d and $t > 0$, let $\mathcal{D}_{G,\beta,t}$ be as defined in Definition 1.11. Then, we have that with probability at least $1 - O(1/n^t)$ over $D_\psi \sim \mathcal{D}_{G,\beta,t}$, there exists an algorithm that draws $N = \frac{\exp(O(\beta^2 t + \beta t \sqrt{\log n})) \cdot n^{O(t)} \cdot \log(1/\delta\beta)}{\beta^2}$ samples and recovers the graph G. The running time is $O(N \cdot n^t)$.*

Note that for $\beta \leq O(\sqrt{\log n})$ the above running time is polynomial in n^t. We now state the theorem of learning in TV distance.

Theorem 4.12. *For a graph G of degree d and $t > 0$, let $\mathcal{D}_{G,\beta,t}$ be as defined in Definition 1.11. Then, we have that with probability at least $1 - O(1/n^t)$ over $D_\psi \sim \mathcal{D}_{G,\beta,t}$, there exists an algorithm that draws $N = \frac{\exp(O(\beta^2 t + \beta t \sqrt{\log n})) \cdot n^{O(t)} \cdot \log(1/\delta\epsilon)}{\epsilon^8}$ samples and runs in time $O(N \cdot n^t)$ that outputs a t-MRF $D_{\tilde{\psi}}$ such that (1) $\left\| \psi - \tilde{\psi} \right\|_1 \leq \epsilon^2$, (2) $\mathsf{d}_{\mathsf{KL}}(D_\psi, D_{\tilde{\psi}}) \leq 2\epsilon^2$, and (3) $\mathsf{d}_{\mathsf{TV}}(D_\psi, D_{\tilde{\psi}}) \leq \epsilon$.*

For random MRFs with rademacher weights, we obtain improved bounds. We give an algorithm that draws a sub-polynomial number of samples and recovers the random MRF exactly.

Theorem 4.13. *For a graph G of degree d and $t > 0$, let $\mathcal{D}_{G,\beta,t}$ be as defined in Definition 1.11 with rademacher coefficients. Then, we have that with probability at least $1 - O(1/n^t)$ over $D_\psi \sim \mathcal{D}_{G,\beta,t}$, there exists an algorithm that draws $N = \exp(O(\beta^2 t + \beta t \sqrt{\log n})) \cdot O(td^t \log(n/\delta\beta)/\beta^2)$ samples and recovers ψ exactly. The running time is $O(N \cdot n^t)$.*

The proof is in Appendix D.

Remark 4.14. *We remark that when the graph G is known to the learner, the task of parameter recovery becomes less expensive. When running Sparsitron for the index i, it suffices to only search over the polynomials whose variables are neighbours of i in G. This is valid as the optimal polynomial $\partial_i\psi$ only contains the neighbouring variables. Thus, the cost of feature expansion is now d^t instead of n^t. Similarly, the error parameter in Lemma A.4 now scales with d^t instead of n^t. Thus, the sample complexity in Theorem 4.12 is now improved to $N = \frac{\exp(O(\beta^2 t + \beta t \sqrt{\log n})) \cdot d^{O(t)} \cdot n^8 \log(n/\delta\epsilon)}{\epsilon^8}$ and the running time is now $O(N \cdot nd^t)$. Thus, for bounded d, the running time is now polynomial in n.*

17

References

[ABXY24] Arka Adhikari, Christian Brennecke, Changji Xu, and Horng-Tzer Yau. Spectral gap estimates for mixed p-spin models at high temperature. *Probability Theory and Related Fields*, pages 1–29, 2024.

[AJ24] Gérard Ben Arous and Aukosh Jagannath. Shattering versus metastability in spin glasses. *Communications on Pure and Applied Mathematics*, 77(1):139–176, 2024.

[AJK+24] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Universality of spectral independence with applications to fast mixing in spin glasses. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5029–5056. SIAM, 2024.

[AKN06] Pieter Abbeel, Daphne Koller, and Andrew Y Ng. Learning factor graphs in polynomial time and sample complexity. *The Journal of Machine Learning Research*, 7:1743–1788, 2006.

[AMS23] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Shattering in pure spherical spin glasses. *arXiv preprint arXiv:2307.04659*, 2023.

[BB19] Roland Bauerschmidt and Thierry Bodineau. A very simple proof of the lsi for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, 2019.

[BGPV21] Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and NV Vinodchandran. Near-optimal learning of tree-structured distributions by chow-liu. In *Proceedings of the 53rd annual acm SIGACT symposium on theory of computing*, pages 147–160, 2021.

[BMS08] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 343–356. Springer, 2008.

[Bre15] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 771–782. ACM, 2015.

[BRO17] Ludovica Bachschmid-Romano and Manfred Opper. A statistical physics approach to learning curves for the inverse ising problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(6):063406, 2017.

[BSXY24] Christian Brennecke, Adrien Schertzer, Changji Xu, and Horng-Tzer Yau. The two point function of the sk model without external field at high temperature. *Probab. Math. Phys*, 5:131–175, 2024.

[BXY23] Christian Brennecke, Changji Xu, and Horng-Tzer Yau. Operator norm bounds on the correlation matrix of the sk model at high temperature. *arXiv preprint arXiv:2307.12535*, 2023.

[CH71] P Clifford and JM Hammersley. Markov fields on finite graphs and lattices. 1971.

[CL68] CKCN Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[DDDK21] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Anthimos Vardis Kandiros. Learning ising models from one or multiple samples. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 161–168, 2021.

[DKSS21] Ilias Diakonikolas, Daniel M Kane, Alistair Stewart, and Yuxin Sun. Outlier-robust learning of ising models under dobrushin's condition. In *Conference on Learning Theory*, pages 1645–1682. PMLR, 2021.

[EA75] Samuel Frederick Edwards and Phil W Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

[EAG24] Ahmed El Alaoui and Jason Gaitonde. Bounds on the covariance matrix of the sherrington–kirkpatrick model. *Electronic Communications in Probability*, 29:1–13, 2024.

[EKZ22] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: fast mixing in high-temperature ising models. *Probability theory and related fields*, 182(3):1035–1051, 2022.

[GJK23] David Gamarnik, Aukosh Jagannath, and Eren C Kızıldağ. Shattering in the ising pure p-spin model. *arXiv preprint arXiv:2307.07461*, 2023.

[GKK19] Surbhi Goel, Daniel M Kane, and Adam R Klivans. Learning ising models with independent failures. In *Conference on Learning Theory*, pages 1449–1469. PMLR, 2019.

[GM24] Jason Gaitonde and Elchanan Mossel. A unified approach to learning ising models: Beyond independence and bounded width. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 503–514, 2024.

[GMM24] Jason Gaitonde, Ankur Moitra, and Elchanan Mossel. Efficiently learning markov random fields from dynamics. *arXiv preprint arXiv:2409.05284*, 2024.

[HKM17] Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. *Advances in Neural Information Processing Systems*, 30, 2017.

[KKSK11] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

[KM17] Adam R. Klivans and Raghu Meka. Learning graphical models using multiplicative weights. *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354, 2017.

[MM09a] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[MM09b] Marc Mézard and Thierry Mora. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris*, 103(1-2):107–113, 2009.

[MMS21] Ankur Moitra, Elchanan Mossel, and Colin P Sandon. Learning to sample from censored markov random fields. In *Conference on Learning Theory*, pages 3419–3451. PMLR, 2021.

[MPV87] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[NBSS10] Praneeth Netrapalli, Siddhartha Banerjee, Sujay Sanghavi, and Sanjay Shakkottai. Greedy learning of markov network structure. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1295–1302. IEEE, 2010.

[Pan13] Dmitry Panchenko. *The sherrington-kirkpatrick model*. Springer Science & Business Media, 2013.

[PSBR20] Adarsh Prasad, Vishwak Srinivasan, Sivaraman Balakrishnan, and Pradeep Ravikumar. On learning ising models under huber's contamination model. *Advances in neural information processing systems*, 33:16327–16338, 2020.

[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.

[Tal03] Michel Talagrand. *Spin glasses: a challenge for mathematicians: cavity and mean field models*, volume 46. Springer Science & Business Media, 2003.

[Tal10] Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

[TR14] Rashish Tandon and Pradeep Ravikumar. Learning graphs with a few hubs. In *International conference on machine learning*, pages 602–610. PMLR, 2014.

[Ver18] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[VML22] Marc Vuffray, Sidhant Misra, and Andrey Y Lokhov. Efficient learning of discrete graphical models. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124017, 2022.

[VMLC16] Marc Vuffray, Sidhant Misra, Andrey Y. Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2595–2603, 2016.

[WLR06] Martin J Wainwright, John Lafferty, and Pradeep Ravikumar. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. *Advances in neural information processing systems*, 19, 2006.

[WSD19] Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. *Advances in Neural Information Processing Systems*, 32, 2019.

[ZKKW20] Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Steven Wu. Privately learning markov random fields. In *International conference on machine learning*, pages 11129–11140. PMLR, 2020.

A Lemmas on Polynomial Recovery

This section has some useful results on polynomial recovery in C-*smooth* MRFs. The proofs of these claims follow the steps of the analogous claims in Section 6 of [KM17], with appropriate changes to handle C-*smooth* distributions as opposed to the unbiased ones considered in [KM17]. The property of unbiasedness is a worst case property, but we extend their proofs to work under the average case properties of C-*smooth* MRFs.

First, we prove that for any polynomial p with maximal monomial S and any x, there exists a vector y at a Hamming distance at $|S|$ from x such that $|p(x)| \geq |\widehat{p}(S)|$.

Lemma A.1. *Let p be a polynomial on $\{\pm 1\}^n$ and let S be a maximal monomial of p. For any $x \in \{\pm 1\}^n$, there exists a vector $y \in \{\pm 1\}^n$ with $d_H(x, y) \leq |S|$ such that $|p(y)| \geq |\widehat{p}(S)|$.*

Proof. We prove the lemma by induction on $|S|$. We first consider the base case $|S| = 1$. Let $S = \{i\}$. We have that

$$p(x) = \widehat{p}(\{i\})\, x_i + p^{-i}(x)$$

where $p^{-i}(x)$ does not depend on x_i. Choosing y such that $y_i = \text{sign}\left(\widehat{p}(\{1\}) \cdot p^{-1}(x)\right)$ and $y_j = x_j$ for $j \neq i$, we have that $|p(y)| \geq |\widehat{p}(\{1\})|$. Clearly, $d_H(x, y) \leq 1$.

Say the claim is true for $|S| - 1$. Consider a maximal monomial S with $i \in S$ for some index i. We have that $p(x) = x_i \cdot \partial_i p(x) + p^{-i}(x)$ where $p^{-i}(x)$ and $\partial_i p$ do not depend on x_i. Observe that $S \setminus \{i\}$ is a maximal monomial for $\partial_i p$ with coefficient equal to $\widehat{p}(S)$. Thus, we have that there exists a vector z with $d_H(x, z) \leq |S| - 1$ and $z_i = x_i$ such that $|\partial_i p(z)| \geq |\widehat{p}(S)|$. Construct y such that $y_i = \text{sign}(\partial_i p(z) \cdot p^{-1}(z))$ and $y_j = z_j$ for $j \neq i$. Clearly, we have that $|p(y)| \geq |\partial_i p(z)| \geq |\widehat{p}(S)|$ we also have that $d_H(x, y) \leq |S|$ as y and z differ in at most one coordinate. This completes the proof. $\qquad\square$

The following lemma proves that for for any p such that $\mathbb{E}_{X \sim D_\psi}[(\sigma(p(X)) - \sigma(2\partial_i \psi(X)))^2]$ is small, it is possible to estimate the coefficient $\widehat{\partial_i \psi}(S)$ using p for any maximal monomial S.

Lemma A.2. *For $C > 0$, let D_ψ be a C-smooth t-MRF. Let p be a polynomial and $i \in [n]$ such that $\mathbb{E}_{X \sim D_\psi}[(\sigma(p(X)) - \sigma(2\partial_i \psi(X)))^2] \leq \epsilon$ for $\epsilon > 0$. Then, for any maximal monomial S of $\partial_i \psi$ of size at most $t - 1$, we have that*

$$\Pr_{X \sim D_\psi}\left[|\widehat{2\partial_i \psi}(S) - \partial_S p(X)| > \delta|\right] \leq \frac{2^{t-1} \exp(10C + 6) \cdot \epsilon}{\delta^2} + \frac{1}{8}$$

Proof. The proof follows the same structure as the proof of Lemma 6.2 in [KM17]. We make appropriate changes to account for the fact that we do not have unbiasedness anymore. Recall the definition of $\mathcal{E}$ from Definition 4.2. We define the set

$$\mathcal{E}_S := \{y \mid \text{there exists } x \in \mathcal{E} \text{ with } x_{[n] \setminus S} = y\}.$$

We first prove the following claim that $|\partial_i \psi(x)| \leq 2C$ for $x \in \{\pm 1\}^n$ with $x_{[n] \setminus S} \in \mathcal{E}_S$.

Claim A.3. *For $y \in \{\pm 1\}^n$ with $y_{[n] \setminus S} \in \mathcal{E}_S$, it holds that $|\partial_i \psi(y)| \leq 2C$.*

Proof. Since $y_{[n] \setminus S} \in \mathcal{E}_S$, we have that $y_{[n] \setminus S} = x_{[n] \setminus S}$ for $x \in \mathcal{E}$. For $x \in \mathcal{E}$, we have that

$$|\partial_i \psi(x)| = \frac{1}{2}|\psi^{\{i\}}(x)| \leq \frac{C}{2}$$

21

from the fact that $\psi^{\{i\}}(x) = 2x_i \cdot \partial_i\psi(x)$. Thus, we have that

$$|\partial_i\psi(x) - \partial_i\psi(y)| = \frac{1}{2}|x_i\psi^{\{i\}}(x) - y_i\psi^{\{i\}}(y)| \leq \frac{1}{2}|\psi(x) - \psi(x^{\{i\}})| + \frac{1}{2}|\psi(y) - \psi(y^{\{i\}})|$$

$$\leq \frac{1}{2}\left(|\psi(x) - \psi(x^{\{i\}})| + |\psi(y) - \psi(x)| + |\psi(y^{\{i\}}) - \psi(x)|\right) \leq \frac{3C}{2}.$$

The first inequality follows from the definition of $\psi^{\{i\}}$ and a triangle inequality . The second inequality again follows from a triangle inequality. The last inequality follows from the facts that $x \in \mathcal{E}$ and $\max\left(d_H(x, x^{\{i\}}), d_H(x, y), d_H(x, y^{\{i\}})\right) \leq t$. Thus, we obtain that $|\partial_i\psi(y)| \leq 2C$. $\qquad\square$

From Fact 2.2, we have that for any y with $y_{[n]\setminus S} \in \mathcal{E}_S$,

$$|\sigma(p(y)) - \sigma(2\partial_i\psi(y))| \geq \exp(-4C - 3) \cdot \min(1, |p(X) - 2\partial_i\psi(X)|).$$

since $|\partial_i\psi(y)| \leq 2C$ for all y with $y_{[n]\setminus S} \in \mathcal{E}_S$. Squaring and taking expectation, we obtain

$$\mathbb{E}_{X \sim D_\psi}\left[\min(1, |p(X) - 2\partial_i\psi(X)|^2) \cdot \mathbb{1}\{X_{[n]\setminus S} \in \mathcal{E}_S\}\right]$$

$$\leq \exp(8C + 6) \mathbb{E}_{X \sim D_\psi}\left[(\sigma(p(X)) - \sigma(2\partial_i\psi(X)))^2 \cdot \mathbb{1}\{X_{[n]\setminus S} \in \mathcal{E}_S\}\right] \leq \exp(8C + 6) \cdot \epsilon$$

From Markov's inequality, for any $0 < \delta < 1$, we have that

$$\Pr_{X \sim D_\psi}\left[|p(X) - 2\partial_i\psi(X)| \geq \delta, X_{[n]\setminus S} \in \mathcal{E}_S\right] \leq \frac{\exp(8C + 6) \cdot \epsilon}{\delta^2}.$$

Similar to the proof of Lemma 6.2 in [KM17], for any fixing z of variables not in S, let $r_z(x_S)$ be the polynomial obtained from $p(x) - 2\partial_i\psi(x)$ by fixing the variables outside S to z. Note that $\widehat{r_z}(S) = \widehat{2\partial_i\psi}(S) - \partial_S p(X)$ where $X_{[n]\setminus S} = z$ as S is maximal in $\partial_i\psi$.

For any $x_{[n]\setminus S} \in \mathcal{E}_S$ with $x_{[n]\setminus S} = z$, Lemma A.1 implies that there exists y with $d_H(x, y) \leq t - 1$ and $y_{[n]\setminus S} = x_{[n]\setminus S}$ such that $|r_z(y_S)| \geq |\widehat{r_z}(S)|$. Let $w \in \{\pm 1\}^n$ such that $\Pr_{X \sim D_\psi}\left[X = w \mid X_{[n]\setminus s} = z\right] \geq 2^{-t+1}$. There always exists one such vector as it is fixed in all but $t - 1$ indices. We have that for $z \in \mathcal{E}_S$,

$$\Pr_{X \sim D_\psi}\left[|r_z(X_S)| \geq |\widehat{r_z}(S)|, \mid X_{[n]\setminus S} = z\right] = \Pr_{X \sim D_\psi}\left[X = y \mid X_{[n]\setminus S} = z\right]$$

$$= \frac{\Pr_{X \sim D_\psi}[X = y]}{\Pr_{X \sim D_\psi}[X = w]} \cdot \frac{\Pr_{X \sim D_\psi}[X = w]}{\Pr_{X \sim D_\psi}\left[X_{[n]\setminus s} = z\right]}$$

$$\geq \exp(\psi(y) - \psi(w)) \cdot 2^{-t+1} \geq \exp(-2C) \cdot 2^{-t+1} \qquad (14)$$

where the first equality follows from the definition of conditional probability. The first inequality follows from the definition of w and the last inequality follows from the definition of $\mathcal{E}_S$.

We now have that

$$\Pr_{X \sim D_\psi}\left[|p(X) - 2\partial_i\psi(X)| \geq \delta, X_{[n]\setminus S} \in \mathcal{E}_S\right]$$

$$\geq \sum_{z \in \mathcal{E}_S} \Pr_{X \sim D_\psi}\left[|p(X) - 2\partial_i\psi(X)| \geq \delta \mid |\widehat{2\partial_i\psi}(S) - \partial_S p(X)| \geq \delta, X_{[n]\setminus S} = z\right]$$

$$\cdot \Pr_{X \sim D_\psi}\left[|\widehat{2\partial_i\psi}(S) - \partial_S p(X)| \geq \delta, X_{[n]\setminus S} = z\right]$$

$$\geq \sum_{z \in \mathcal{E}_S} \exp(-2C) \cdot 2^{-t+1} \cdot \Pr_{X \sim D_\psi}\left[|\widehat{2\partial_i\psi}(S) - \partial_S p(X)| \geq \delta, X_{[n]\setminus S} = z\right]$$

$$\geq \exp(-2C) \cdot 2^{-t+1} \Pr_{X \sim D_\psi}\left[|\widehat{2\partial_i\psi}(S) - \partial_S p(X)| \geq \delta, X_{[n]\setminus S} \in \mathcal{E}_S\right]$$

22

Putting things together, we obtain that

$$\Pr_{X \sim D_\psi} \left[|\widehat{2\partial_i \psi}(S) - \partial_S p(X)| \geq \delta, X_{[n]\setminus S} \in \mathcal{E}_S \right] \leq \frac{2^{t-1} \exp(10C+6) \cdot \epsilon}{\delta^2}$$

Now, we have that

$$\Pr_{X \sim D_\psi} \left[|\widehat{2\partial_i \psi}(S) - \partial_S p(X)| \geq \delta \right] \leq \Pr_{X \sim D_\psi} \left[|\widehat{2\partial_i \psi}(S) - \partial_S p(X)| \geq \delta, X_{[n]\setminus S} \in \mathcal{E}_S \right] + \frac{1}{8}$$

where we used the fact that $\Pr_{X \sim D_\psi}[X_{[n]\setminus S} \in \mathcal{E}_S] \geq \Pr_{X \sim D_\psi}[X \in \mathcal{E}] \geq \frac{7}{8}$. Combining the two equations above completes the proof. $\qquad \square$

The following lemma asserts that for a C-*smooth* MRF, the error $\|p - 2\partial_i \psi\|_1 \leq O\left(e^{O(C)} \binom{n}{t} \epsilon\right)$ whenever $\mathbb{E}_{X \sim D_\psi}\left[(\sigma(p(X)) - \sigma(2\partial_i \psi(X)))^2\right] \leq \epsilon$. This lemma generalizes Theorem 3.2 from [GM24]. The proof closely follows the proof of Lemma 6.4 in [KM17]. Their lemma also recovers the external fields, contrary to a remark in [GM24].

Lemma A.4. *For $C > 0$, let D_ψ be a C-smooth t-MRF. Let p be a polynomial and $i \in [n]$ such that $\mathbb{E}_{X \sim D_\psi}[(\sigma(p(X)) - \sigma(2\partial_i \psi(X)))^2] \leq \epsilon$ for $0 < \epsilon \leq \exp(-2C) \cdot 2^{-t}$. Then, we have that*

$$\|p - 2\partial_i \psi\|_1 \leq O(1)(4t)^t \binom{n}{t} \exp(5C)\sqrt{\epsilon}.$$

Proof. This proof follows closely the proof of Lemma 6.4 in [KM17]. Most of the details are the same, except that we make appropriate changes to handle the fact that we do not have unbiasedness, similar to Lemma A.2. We borrow notation from the proof of the aforementioned lemma in [KM17]. We highlight the parts where we make changes and do not re-derive the steps that are exactly the same. Let $r = p - 2\partial_i \psi$ be the difference polynomial. For $\ell \leq t - 1$, let $r_{=\ell}$ be the polynomial obtained from r by only considering monomials of size exactly equal to ℓ. For $\ell \leq t-1$, let $\rho_\ell = \|r_{=\ell}\|_1$. Clearly, the quantity we want to bound is $\|r\|_1 = \sum_{i=0}^{t-1} \rho_i$. We bound $\rho_0, \ldots \rho_{t-1}$ inductively, starting with ρ_{t-1}. We first, bound ρ_{t-1}.

Claim A.5. *Consider any maximal monomial S of $r = p - 2\partial_i \psi$ for p satisfying the assumptions of Lemma A.4. Then, it holds that $|\widehat{r}(S)| \leq \exp(5C+3) \cdot 2^{t/2}\sqrt{\epsilon}$.*

Proof. Recall the definition of $\mathcal{E}_S$. From the fact that $\Pr_{X \sim D_\psi}[X_{[n]\setminus S} \in \mathcal{E}_S] \geq \frac{7}{8}$ and the assumption of the lemma, we obtain (using an averaging argument) that there exists a vector $z \in \mathcal{E}_S$ such that

$$\mathbb{E}_{X \sim D_\psi} \left[(\sigma(p(X)) - \sigma(2\partial_i \psi(X)))^2 \mid X_{[n]\setminus S} = z \right] \leq 2\epsilon.$$

Similar to the proof of Lemma A.2, let r_z be the polynomial obtained from r by fixing the variables in $[n] \setminus S$ to z. Note that $\widehat{r_z}(S) = \widehat{r}(S)$. From the argument preceding Equation (14), we have that

$$\Pr_{X \sim D_\psi} \left[|r_z(X) \geq |\widehat{r_z}(S)| \mid X_{[n]\setminus S} = z \right] \geq \exp(-2C) \cdot 2^{-t+1}.$$

Recall from Claim A.3 that $|\partial_i \psi(y)| \leq 4C$ for any y with $y_{[n]\setminus S} \in \mathcal{E}_S$. This fact, together with Fact 2.2, implies that

$$2\epsilon \geq \mathbb{E}_{X \sim D_\psi} \left[(\sigma(p(X)) - \sigma(2\partial_i \psi(X)))^2 \mid X_{[n]\setminus S} = z \right] \geq \exp(-10C-6) \cdot 2^{-t+1} \cdot \min\left(1, \widehat{r}(S)\right)^2$$

From the bound $\epsilon \leq \frac{1}{2}\exp(-5C-3) \cdot 2^{-t+1}$, we have that $\widehat{r}(S) \leq 1$ which implies $|\widehat{r}(S)| \leq \exp(5C+3) \cdot 2^{t/2}\sqrt{\epsilon}$. $\qquad \square$

Since all degree $t-1$ terms are maximal, the above claim implies that $\rho_{t-1} \leq \binom{n}{t-1} \exp(5C+3) \cdot 2^{t/2} \sqrt{\epsilon}$. We now do the inductive step. For $|I| = \ell < t-1$, again by an averaging argument, we have z such that $\mathbb{E}_{X \sim D_\psi} \left[(\sigma(p(X)) - \sigma(2\partial_i \psi(X)))^2 \mid X_{[n] \setminus I} = z \right] \leq 2\epsilon$. Let r_z be the polynomial obtained from r by setting these variables to z. Again, repeating the same steps as before, we obtain that $\widehat{r_z}(I) \leq \exp(5C + 3) 2^{t/2} \sqrt{\epsilon}$. From here, the proof is exactly identical to the proof in [KM17] except we set $\epsilon_0 \equiv \exp(5C + 3) 2^{t/2} \sqrt{\epsilon} \cdot \binom{n}{t-1}$. Their analysis yields that

$$\|r\|_1 \leq 2^t t^t \cdot \epsilon_0 \leq O(1)(4t)^t \binom{n}{t} \exp(5C) \sqrt{\epsilon}.$$

$\square$

B Parameter Recovery vs TV Distance

We now argue that an MRF $\widehat{D}$ that has parameters close to D_ψ has low TV distance (in fact KL divergence) with respect to D_ψ. First we define KL divergence and TV distance.

Definition B.1 (KL Divergence). *Let P and Q be two distributions. The KL-divergence between P and Q is defined as*

$$d_{\mathsf{KL}}(P, Q) := \underset{X \sim P}{\mathbb{E}}[\log(P(X)/Q(X))]$$

We now define the Total Variation distance.

Definition B.2 (TV Distance). *Let P and Q be two distributions. The KL-divergence between P and Q is defined as*

$$d_{\mathsf{TV}}(P, Q) := \frac{1}{2} \sum_{x \in \{\pm 1\}^n} |P(x) - Q(x)|.$$

We are now ready to argue that parameter recovery in t-MRFs implies closeness in KL divergence/TV distance. The proof is almost the same as Lemma 3.6 in [GM24].

Lemma B.3. *Let D_ψ and $D_{\tilde{\psi}}$ be t-MRFs with factorization polynomials ψ and $\tilde{\psi}$ respectively such that $\left\|\psi - \tilde{\psi}\right\|_1 \leq \epsilon$. Then, (1) $d_{\mathsf{KL}}(D_\psi, D_{\tilde{\psi}}) \leq 2\epsilon$, and (2) $d_{\mathsf{TV}}(D_\psi, D_{\tilde{\psi}}) \leq \sqrt{\epsilon}$.*

Proof. We prove (1). (2) follows from Pinsker's inequality. Let Z_ψ and $Z_{\tilde{\psi}}$ are the partition functions of $D_\psi, D_{\tilde{\psi}}$ respectively. For any $x \in \{\pm 1\}^n$, we have that $\exp(\psi(x)) \leq \exp(\epsilon) \cdot \exp(\tilde{\psi}(x))$ and $\exp(\tilde{\psi}(x)) \leq \exp(\epsilon) \cdot \exp(\psi(x))$. From the definition of the partition function, we have that $Z_\psi \leq \exp(\epsilon) \cdot Z_{\tilde{\psi}}$ and $Z_{\tilde{\psi}} \leq \exp(\epsilon) \cdot Z_\psi$. For any $x \in \{\pm 1\}^n$, we thus have that $\frac{D_\psi(x)}{D_{\tilde{\psi}(x)}} \leq \exp(2\epsilon)$. Now, from the definition of KL divergence, we get that $d_{\mathsf{KL}}(D_\psi, D_{\tilde{\psi}}) \leq 2\epsilon$. $\square$

C Parameter Recovery in the Pure t-spin Model

Definition C.1 (Pure t-spin model). *A pure t-spin model is the distribution of t-MRFs such that that factorization polynomials ψ is a random variable of the form*

$$\psi(x) = \frac{\beta}{n^{(t-1)/2}} \sum_{(i_1, \ldots, i_t) \in [n]^t} \mathcal{N}(0, 1) \cdot \prod_{j=1}^{t} x_{i_j}.$$

We now prove that pure t-spin models satisfy the properties required for parameter recovery in Theorem 4.8.

Lemma C.2. *Let $\mathcal{D}$ be the pure t-spin model with inverse temperature β. Then, it holds with probability at least $1 - O(1/n^t)$ over $D_\psi \sim \mathcal{D}$ that*

1. *$\|\partial_i\psi\|_1 \leq \beta\sqrt{tn^{t+1}\log n}$ for all $i \in [n]$,*

2. *D_ψ is $O(\beta^2 t^2 + \beta t\sqrt{t\log n})$-smooth.*

Proof. We first prove (1). The number of tuples $(i_1, i_2, \ldots i_t)$ is at most n^t. By standard gaussian tails, the maximum absolute values of all these coefficients is at most $\frac{\beta}{n^{(t-1)/2}}\sqrt{t\log n}$ with probability at least $1 - O(1/n^t)$. Thus, we obtain $\|\partial_i\psi\|_1 \leq \beta\sqrt{tn^{t+1}\log n}$. To prove (2), we argue that $\mathcal{D}$ has subgaussian derivatives and then use Theorem 4.4. For any multiset $\alpha \in [n]^t$ and set $S \subseteq [n]$, we say that $|S \cap \alpha|$ is odd if the number of common elements between S and t (counting repetitions) is odd. Observe that for any set S with $|S| \leq t$ and any $x, y \in \{\pm 1\}^n$ with x and y differing in the set S, we have that

$$\psi(x) - \psi(y) = \frac{2\beta}{n^{(t-1)/2}} \sum_{\substack{\alpha \in [n]^t \\ |S \cap \alpha| \text{ is odd}}} \mathcal{N}(0,1) \cdot \prod_i^t x_{\alpha_i}.$$

We now count the number of terms in the above expression as that determines the subgaussianity of $\psi(x) - \psi(y)$. Since S has at most t terms and the intersection is at least 1, the number of terms is upper bounded by $|\bigcup_{i=1}^t \{\alpha \mid \alpha_i \in S, \alpha \in [n]^t\}| \leq t^2 \cdot n^{t-1}$. Thus, we have that $\psi(x) - \psi(y)$ is $O(\beta t)$-subgaussian. Now, from Theorem 4.4, we have that D_ψ is $O(\beta^2 t^2 + \beta t\sqrt{t\log n})$-smooth. $\square$

Note that the smoothness is a factor of t worse than what was obtained in Lemma 4.9. This is because in the definition of the pure t-spin model, the same set can be counted multiple times whereas this was not allowed in the definition of the random MRF (Definition 1.11). The following theorem on parameter recovery of these models is now immediate from Theorem 4.8.

Theorem C.3. *Let $\mathcal{D}$ be a pure t-spin model with inverse temperature β. With probability at least $1 - O(1/n^t)$ over $D_\psi \sim \mathcal{D}$, there exists an algorithm that draws $N = \frac{\exp(O(\beta^2 t^2 + \beta t\sqrt{t\log n}))\cdot n^{O(t)}\cdot\log(1/\delta\epsilon)}{\epsilon^8}$ samples and runs in time $O(N \cdot n^t)$ that outputs a t-MRF $D_{\tilde{\psi}}$ such that (1) $\left\|\psi - \tilde{\psi}\right\|_1 \leq \epsilon^2$, (2) $\mathsf{d}_{\mathsf{KL}}(D_\psi, D_{\tilde{\psi}}) \leq 2\epsilon^2$, and (3) $\mathsf{d}_{\mathsf{TV}}(D_\psi, D_{\tilde{\psi}}) \leq \epsilon$.*

D Improved bounds for Rademacher Random MRFs

For our algorithm with improved sample complexity to exactly learn the random MRF with Rademacher weights, we require a slightly modified version of Theorem 3.1.

Theorem D.1. *Let $\lambda, \epsilon, \delta > 0$. Let D be a distribution on $\{\pm 1\}^n \times \{\pm 1\}$ where $\Pr[Y = +1|X] = \sigma(w \cdot X + g(X))$ for $(X, Y) \sim D$ where $w \in \mathbb{R}^n$ is an unknown vector with $\|w\|_1 \leq \lambda$ and $g : \{\pm 1\}^n \to \mathbb{R}$ is a known function. There exists an algorithm that takes $N = O\left(\lambda^2(\ln(n/\delta\epsilon))/\epsilon^2\right)$ independent samples from D, runs in time $O(nN)$, and outputs a vector $\widehat{w}$ such that*

$$\mathop{\mathbb{E}}_{(X,Y)\sim D} \left[(\sigma(w \cdot X + g(X)) - \sigma(\widehat{w} \cdot X + g(X))^2\right] \leq \epsilon$$

with probability at least $1 - \delta$.

Note that the only difference between the above theorem and Theorem 3.1 is the addition of the known function $g(X)$ to the conditional probability. The proof of the above theorem is almost identical to that of Theorem 3.1 (also Theorem 3.1 in [KM17]) with very few additional changes (we change one line in their algorithm). We descrive the change below. We borrow the notation from the proof of [KM17] and only highlight key changes.

Proof of Theorem D.1. We now use u instead of σ to refer to the link function. We apply the transformation $(x, y) \to (x, \frac{y+1}{2})$ so that the $+1$ labels are mapped to $+1$ and the -1 labels are mapped to 0. From now on, we assume that D is the distribution of inputs after this transformation. Thus, we have that $\mathbb{E}_{(X,Y)\sim D}[Y \mid X = x] = u(w \cdot X + g(X))$.

The only changes we make to their algorithm are the following: (1) in line 4 of Algorithm 2 of [KM17], we redefine the loss vector ℓ^t to now be $\ell^t \coloneqq (1/2)(\mathbf{1} + (u(\lambda p^t \cdot x^t + g(x^t)) - y^t)x^t)$, and (2) in line 7 we compute the empirical risk as $\hat{\varepsilon}(\lambda p^t) = (1/M)\sum_{j=1}^{M}(u(\lambda p^t \cdot a^t + g(a^t)) - b^j)^2$. Note the addition of the term $g(x^t)$ in both the steps. We can do this as we know the function g.

Now, we highlight the changes in the analysis. The steps of the argument until Equation 3.3 of the proof of [KM17] are identical as the new loss vector ℓ^t is still a vector in $[0,1]^n$. The only change is in how we bound $\mathbb{E}_{(x^t,y^t)}[Q^t \mid (x^1, y^1), \dots, (x^{t-1}, y^{t-1})]$. We have that

$$
\begin{aligned}
\mathbb{E}_{(x^t,y^t)}[Q^t \mid (x^1, y^1), \dots, (x^{t-1}, y^{t-1})] &= \mathbb{E}_{(x^t,y^t)}[(p^t - (1/\lambda)w) \cdot \ell^t] \\
&= (1/2) \mathbb{E}_{(x^t,y^t)}[(p^t - (1/\lambda)w) \cdot (u(\lambda p^t \cdot x^t + g(x^t)) - y^t) \cdot x^t] \\
&= (1/2\lambda) \mathbb{E}_{x^t}[(\lambda p^t \cdot x^t + g(x^t) - w \cdot x^t - g(x^t))(u(\lambda p^t \cdot x^t + g(x^t)) - u(w \cdot x^t + g(x^t)))] \\
&\geq (1/2\lambda) \mathbb{E}_{x^t}[(u(\lambda p^t \cdot x^t + g(x^t)) - u(w \cdot x^t + g(x^t)))^2] = (1/2\lambda)\varepsilon(\lambda p^t)
\end{aligned}
$$

where $\varepsilon(v) \coloneqq \mathbb{E}_{(X,Y)\sim D}[(u(v \cdot X + g(X)) - u(w \cdot X + g(X)))^2]$ is the risk. The main difference from the proof of [KM17] is the third equation where we add and subtract $g(x^t)$ and then use the lipschitzness property. The rest of the proof is exactly identical. $\qquad\square$

We are now ready to prove Theorem 4.13.

Proof. Recall that $D_\psi \sim \mathcal{D}_{G,\beta,t}$ (with rademacher weights) is C-smooth with $C = O(\beta^2 t + \beta t \sqrt{\log n})$ with probability at least $1 - O(1/n^t)$. We henceforth assume D_ψ is C-smooth. We recover the coefficients of ψ iteratively, starting with the degree t terms and proceeding downwards. We use fresh samples per iteration. We first show the base case of recovering degree t terms. For each $i \in [n]$, we use Theorem 3.1 to find polynomials $\{p_i^t\}_{i\in[n]}$ such that

$$
\mathbb{E}_{X\sim D_\psi}[(\sigma(p_i^t(X)) - \sigma(2\partial_i\psi(X)))^2] \leq \exp(-10C - 6) \cdot (\beta^2/(16(2d)^t))
$$

Now, from Claim A.5, we have that $|\widehat{p_i^t}(S) - \widehat{2\partial_i\psi}(S)| < \frac{\beta}{4d^{(t-1)/2}}$ for any maximal monomial S of $p_i^t - 2\partial_i\psi$. To recover the coefficient $\hat{\psi}(S)$ for $|S| = t$, we consider any $i \in S$. Now, we have that $\widehat{\partial_i\psi(X)}(S \setminus \{i\}) = \widehat{\psi}(S)$. Note that $S \setminus \{i\}$ is a maximal monomial of $p_i^t - 2\partial_i\psi$ as it has degree $t - 1$ which is the degree of the polynomial. Thus, it holds that $|\widehat{p_i^t}(S \setminus \{i\})/2 - \widehat{\psi}(S)| \leq \frac{\beta}{2d^{(t-1)/2}}$. To obtain $\widehat{\psi}(S)$ exactly, we round $\widehat{p_i^t}(S \setminus \{i\})/2$ to the nearest multiple of $\beta/d^{(t-1)/2}$. Thus, we have obtained all the degree t coefficients of ψ exactly. The sample complexity of this step is $N^t = \exp(O(\beta^2 t + \beta t \sqrt{\log n})) \cdot O(d^t \log(n/\delta\beta)/\beta^2)$ and follows from Theorem 3.1.

We now describe how to obtain coefficients of degree j if we know $\widehat{\psi}(S)$ exactly for all $|S| > j$. Construct a function $g^j : \{\pm 1\}^n \to \{\pm 1\}$ such that $g^j(x) = \sum_{|T|>j} \widehat{\psi}(T)\chi_T(x)$. Let the polynomial g_i^j be

26

defined as $g_i^j = 2\partial_i g^j$. We can construct these polynomials as we know all coefficients of size greater than j. Now, for each $i \in [n]$, we find polynomials $\{p_i^j\}_{i \in [n]}$ of degree at most j such that

$$\mathop{\mathbb{E}}_{X \sim D_\psi}[(\sigma(p_i^j(X) + g_i^j(X)) - \sigma(2\partial_i \psi(X)))^2] \leq \exp(-10C - 6) \cdot (\beta^2/(16(2d)^t)).$$

We note that $2\partial_i \psi(X) = 2\sum_{|S| \leq j} \widehat{\partial_i \psi}(S)\chi_S(X) + g_i^j(X)$. Thus, we can find the above polynomials by $\{p_i^j\}_{i \in [n]}$ by running the modified Sparsitron algorithm from Theorem D.1 (with known function g_i^j) after expanding the features to contain all monomials of degree at most j. Observe that the degree j monomials in the polynomial $r = p_i^j + g_i^j - 2\partial_i \psi$ are maximal as all higher degree monomials are 0. Thus, again, we use Claim A.5 and repeat the argument from the base case to obtain that $|\widehat{p_i^j}(S \setminus \{i\})/2 - \widehat{\psi}(S)| \leq \frac{\beta}{2d^{(t-1)/2}}$ for all $|S| = j$. To obtain $\widehat{\psi}(S)$ exactly, we again round $\widehat{p_i^j}(S \setminus \{i\})/2$ to the nearest multiple of $\beta/d^{(t-1)/2}$. In this way iterating $j = t, t-1, \ldots 1$, we obtain, all the coefficients of ψ exactly. Since we use fresh samples in each iteration, we pay a multiplicative factor of t in the final sample complexity $\qquad\square$