58th Annual IEEE Symposium on Foundations of Computer Science

Learning Graphical Models Using Multiplicative Weights

Adam R. Klivans
Department of Computer Science
UT-Austin
Austin, USA
Email: klivans@cs.utexas.edu

Raghu Meka
Department of Computer Science
UCLA
Los Angeles, USA
Email: raghu@cs.ucla.edu

***Abstract*—We give a simple, multiplicative-weight update algorithm for learning undirected graphical models or Markov random fields (MRFs). The approach is new, and for the well-studied case of Ising models or Boltzmann machines, we obtain an algorithm that uses a nearly optimal number of samples and has running time $\tilde{O}(n^2)$ (where n is the dimension), subsuming and improving on all prior work. Additionally, we give the first efficient algorithm for learning Ising models over *non-binary* alphabets.**

Our main application is an algorithm for learning the structure of t-wise MRFs with nearly-optimal sample complexity (up to polynomial losses in necessary terms that depend on the weights) and running time that is $n^{O(t)}$. In addition, given $n^{O(t)}$ samples, we can also learn the parameters of the model and generate a hypothesis that is close in statistical distance to the true MRF. All prior work runs in time $n^{\Omega(d)}$ for graphs of bounded degree d and does not generate a hypothesis close in statistical distance even for $t = 3$. We observe that our runtime has the correct dependence on n and t assuming the hardness of learning sparse parities with noise.

Our algorithm– the *Sparsitron*– is easy to implement (has only one parameter) and holds in the on-line setting. Its analysis applies a regret bound from Freund and Schapire's classic Hedge algorithm. It also gives the first solution to the problem of learning sparse Generalized Linear Models (GLMs).

I. Introduction

Undirected graphical models or *Markov random fields* (MRFs) are one of the most well-studied and influential probabilsitic models with applications to a wide range of scientific disciplines [1]–[8]. Here we focus on binary undirected graphical models which are distributions $(Z_1, \ldots, Z_n)$ on $\{1, -1\}^n$ with an associated undirected graph G - known as the *dependency graph* - on n vertices where each Z_i conditioned on the values of $(Z_j : j$ adjacent to i in $G)$ is independent of the remaining variables.

Developing efficient algorithms for inferring the structure of the underlying graph G from random samples from $\mathcal{D}$ is a central problem in machine learning, statistics, physics, and computer science and has attracted considerable attention from researchers in these fields. Most works have placed a strong assumption on the structure of the graphical model (e.g., restricted strong convexity [9], [10] or correlation decay [11], [12]).

The current frontier of MRF learning has focused on the *Ising model* (also known as *Boltzmann machines*) on *bounded-degree graphs*, a special class of graphical models with only *pairwise interactions* and each vertex having degree at most d in the underlying dependency graph. We refer to [13] for an extensive historical overview of the problem. Two important works of note are due to Bresler [13] and [14] who learn Ising models on bounded degree graphs.

Bresler's algorithm is a combinatorial (greedy) approach that runs in time $\tilde{O}(n^2)$ but requires doubly exponential in d many samples from the distribution (only singly exponential is necessary). [14] use machinery from convex programming to achieve nearly optimal sample complexity for learning Ising models with zero external field and with running time $\tilde{O}(n^4)$. Neither of these results are proved to hold over non-binary alphabets or for general MRFs.

A. Our Results

The main contribution of this paper is a simple, multiplicative-weight update algorithm for learning MRFs. Using our algorithm we obtain the following new results:

0272-5428/17 $31.00 © 2017 IEEE
DOI 10.1109/FOCS.2017.39

- An efficient online algorithm for learning Ising models on arbitrary graphs with nearly optimal sample complexity and running time $\tilde{O}(n^2)$ per example (precise statements can be found in Section V). In particular, for bounded degree graphs we achieve a run-time of $\tilde{O}(n^2)$ with nearly optimal sample complexity. This subsumes and improves all prior work including the above mentioned results of Bresler [13] and [14]. Our algorithm is the first that works even for *unbounded-degree* graphs as long as the ℓ_1 norm of the weight vector of each neighborhood is bounded, a condition necessary for efficiency (see discussion following Corollary V.4).
- An algorithm for learning the dependency graph of binary t-wise Markov random fields with nearly optimal sample complexity and run-time $n^{O(t)}$ (precise statements can be found in Section VII). Moreover, given access to roughly $n^{O(t)}$ samples (suppressing necessary terms depending on the weights), we can also reconstruct the parameters of the model and output a t-wise MRF that gives a point-wise approximation to the original distribution.

As far as we are aware, these are the *first efficient algorithms* for learning higher-order MRFs. All previous work on learning general t-wise MRFs runs in time $n^{\Omega(d)}$ (where d is the underlying degree of the graph) and does not output a function f that can generate an approximation to the distribution in statistical distance, *even for the special case of* $t = 3$. We give evidence that the $n^{O(t)}$ dependence in our running time is nearly optimal by applying a simple reduction from the problem of learning sparse parities with noise on t variables to learning t-wise MRFs due to Bresler, Gamarnik, and Shah [15] (learning sparse parities with noise is a notoriously difficult challenge in theoretical computer science). Bresler [13] observed that even for the simplest possible Ising model where the graph has a single edge, beating $O(n^2)$ run-time corresponds to fast algorithms for the well-studied *light bulb* problem [16], for which the best known algorithm runs in time $O(n^{1.62})$ [17].

Moreover, our algorithm is easy to implement, has only one tunable parameter, and works in an on-line fashion. The algorithm– the *Sparsitron*– solves the problem of learning a sparse Generalized Linear Model. That is, given examples $(X, Y) \in [-1, 1]^n \times [0, 1]$ drawn from a distribution $\mathcal{D}$ with the property that $\mathbb{E}[Y|X = x] = \sigma(w \cdot x)$ for some monotonic, Lipschitz σ and unknown w with $\|w\|_1 \leq \lambda$, the *Sparsitron* efficiently outputs a w' such that $\sigma(w' \cdot x)$ is close to $\sigma(w \cdot x)$ in *squared-loss* and has sample complexity $O(\lambda^2 \log n)$.

In an independent and concurrent work, Hamilton, Koehler, and Moitra [18] generalized Bresler's approach to hold for both higher-order MRFs as well as MRFs over general (non-binary) alphabets. For learning binary MRFs on bounded-degree—degree at most d—graphs, under the same non-degeneracy assumption taken by Hamilton et al.,[1] we obtain sample complexity that is singly exponential in d^t, whereas theirs is doubly exponential in d^t (both of our papers obtain sample complexity that depends only logarithmically on n, the number of vertices).

B. Our Approach

For a graph $G = (V, E)$ on n vertices, let $C_t(G)$ denote all cliques of size at most t in G. We use the Hammersley-Clifford characterization of Markov random fields and define a binary t-wise Markov random field on G to be a distribution $\mathcal{D}$ on $\{1, -1\}^n$ where

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

and each $\psi_I : \mathbb{R}^n \to \mathbb{R}$ is a function that depends only on the variables in I.

For ease of exposition, we will continue with the case of $t = 2$, the Ising model, and subsequently describe the extension to larger values of t. Let $\sigma(z)$ denote the *sigmoid* function. That is $\sigma(z) = 1/1 + e^{-z}$. Since $t = 2$, we have

$$\Pr[Z = z] \propto \exp\left(\sum_{i \neq j \in [n]} A_{ij} z_i z_j + \sum_i \theta_i z_i\right)$$

for a weight matrix $A \in \mathbb{R}^{n \times n}$ and $\theta \in \mathbb{R}^n$; here, a weight $A_{ij} \neq 0$ if and only if $\{i, j\}$ is an edge in the underlying dependency graph. For a node Z_i, it is easy to see that the probability $Z_i = -1$ conditioned on any setting of the remaining nodes to some value $x \in \{-1, 1\}^{[n]\setminus\{i\}}$ is equal to $\sigma(w \cdot x + \theta)$ where $w \in \mathbb{R}^{[n]\setminus\{i\}}$, $w_j = -2A_{ij}$, $\theta = -\theta_i$.

As such, if we set $X \equiv (Z_j : j \neq i)$ and $Y = (1 - Z_i)/2$, then the conditional expectation of Y given X is *equal* to a sigmoid with an unknown weight vector w and threshold θ_i. We can now rephrase our original *unsupervised* learning task as the following *supervised* learning problem: Given random examples (X, Y) with conditional mean function $\mathbb{E}[Y|X = x] = \sigma(w \cdot x + \theta)$, recover w and θ.

Learning a conditional mean function of the form $u(w \cdot x)$ with a fixed, known *transfer function* $u : \mathbb{R} \to \mathbb{R}$ is *precisely* the problem of learning a *Generalized Linear Model* or GLM and has been studied extensively in machine learning. The first provably efficient algorithm for learning GLMs where u is both monotone and Lipschitz was given by Kalai and Sastry [19], who called their algorithm the "Isotron". Their result was simplified, improved, and extended by Kakade, Kalai, Kanade, and Shamir [20] who introduced the "GLMtron" algorithm.

[1] A previous version of this paper needed a slightly stronger non-degeneracy assumption.

Notice that $\sigma(z)$ is both monotone and 1-Lipschitz. Therefore, directly applying the GLMtron in our setting will result in a w' and θ' such that

$$\mathbb{E}[(\sigma(w' \cdot x + \theta') - \sigma(w \cdot x + \theta))^2] \leq \varepsilon. \qquad \text{(I.1)}$$

Unfortunately, the sample complexity of the GLMtron depends on $\|w\|_2$, which results in sub-optimal bounds on sample complexity for our setting[2]. We desire sample complexity dependent on $\|w\|_1$, essentially the *sparsity* of w. In addition, we need an *exact recovery* algorithm. That is, we need to ensure that w' itself is close to w and not just that the ℓ_2*-error* as in Equation I.1 is small. We address these two challenges next.

Our algorithm, the *Sparsitron*, uses a multiplicative-weight update rule for learning w, as opposed to the GLMtron or Isotron, both of which use additive update rules. This enables us to achieve essentially optimal sample complexity. The Sparsitron is simple to describe (see Algorithm 2) and depends on only one parameter λ, the upper bound on the ℓ_1-norm. Its analysis only uses a regret bound from the classic Hedge algorithm due to Freund and Schapire [21].

Although the Sparsitron algorithm finds a vector $w' \in \mathbb{R}^n$ such that $\mathbb{E}_X[(\sigma(w' \cdot X + \theta') - \sigma(w \cdot X + \theta))^2]$ is small, we still must prove that w' is actually close to w. Achieving such strong recovery guarantees for arbitrary distributions is typically a much harder problem (and can be provably hard in some cases for related problems [22], [23]). In our case, we exploit the nature of MRFs by a clean property of such distributions: Call a distribution $\mathcal{D}$ on $\{1, -1\}^n$ δ*-unbiased* if each variable Z_i is 1 or -1 with probability at least δ conditioned on any setting of the other variables. It turns out that under conditions that are necessary for reconstruction, the distributions of MRFs are δ-unbiased for a non-negligible δ. We show that for such δ-unbiased distributions achieving reasonably small ℓ_2-error as in Equation I.1 implies that the recovered coefficient w' is in fact close to w.

To obtain our results for learning t-wise Markov random fields, we generalize the above approach to handle functions of the form $\sigma(p(x))$ where p is a degree t multilinear polynomial. Sparsitron can be straightforwardly extended to handle low-degree polynomials by *linearizing* such polynomials (i.e., working in the (n^t)-dimensional space of coefficients). We then have to show that achieving small ℓ_2-error - $\mathbb{E}_X[(\sigma(p(X)) - \sigma(q(X)))^2] \ll 1$ - implies that the polynomials p, q are close. This presents several additional technical challenges; still, in a self-contained proof, we show this holds whenever the underlying distribution is δ-unbiased as is the case for MRFs.

[2]GLMtron in our setting would require $\Omega(n)$ samples; we are aiming for an information-theoretically optimal logarithmic dependence in the dimension n.

C. Best-Experts Interpretation of Our Algorithm

Our algorithm can be viewed as a surprisingly simple weighted voting scheme (a.k.a. "Best-Experts" strategy) to uncover the underlying graph structure $G = (\{v_1, \ldots, v_n\}, E)$ of a Markov random field. Consider an Ising model where for a fixed vertex v_i, we want to determine v_i's neighborhood and edge weights. Let $Z = (Z_1, \ldots, Z_n)$ denote random draws from the Ising model.

- Initially, all vertices $v_j (j \neq i)$ could be neighbors. We create a vector of "candidate" neighbors of length $2n-2$ with entries $(j, +)$ and $(j, -)$ for all $j \neq i$. Intuitively, since we do not know if node v_j will be negatively or positively correlated with v_i, we include two candidate neighbors, $(j, +), (j, -)$ to cover the two cases.
- At the outset, every candidate is equally likely to be a neighbor of v_i and so receives an initial *weight* of $1/(2n-2)$. Now consider a random draw from the Ising model $Z = (Z_1, \ldots, Z_n)$. For each $j \neq i$ we view each Z_j (and its negation -Z_j) as the *vote* of $(j, +)$ for the value Z_i (respectively of $(j, -)$). The overall *prediction* p of our candidates is equal to a weighted sum of their votes (we always assume the weights are non-negative and normalized appropriately).
- For a candidate neighbor v_j, let the *penalty* of the prediction p (as motivated by the conditional mean function) be equal to $\ell_j = (\sigma(-2p) - (1 - Z_i)/2)Z_j$. Each candidate v_j's weight is simply multiplied by β^{ℓ_j} (for some suitably chosen *learning rate* β[3]). It is easy to see that candidates who predict Z_i *correctly* will be penalized *less* than neighbors whose predictions are incorrect.

Remarkably, the weights of this algorithm will converge to the weights of the underlying Ising model, and the rate of this convergence is optimal. Weights of vertices that are not neighbors of v_i will rapidly decay to zero.

For clarity, we present the updates for a single iteration of our Sparsitron algorithm applied to Ising model in Algorithm 1. The iterative nature of the algorithm is reminiscent of algorithms such as belief propagation and stochastic gradient descent that are commonly used in practice. Exploring connections with these algorithms (if any) is an intriguing question.

D. Organization

We begin by describing the Sparsitron algorithm for learning sparse generalized models and prove its correctness. We then show, given a hypothesis output by the Sparsitron, how to recover the underlying weight vector *exactly* under δ-unbiased distributions. For ease of exposition, we begin by assuming that we are learning an Ising model.

[3]For our analysis, the learning rate can be set using standard techniques, e.g., $\beta = 1 - \sqrt{\log n/T}$ when processing T examples.

Algorithm 1 Updates for SPARSITRON applied to learning Ising models

Initialize $W_{ij}^+ = W_{ij}^- = 1/2(n-1)$ and $\hat{A}_{ij} = 0$ for $i \neq j$. PARAMETERS: *Sparsity bound* λ.

1: **for** each new example $(Z_1, \ldots, Z_n)$ **do**:
2: Compute the current *predictions*: $p_i = \sum_{j \neq i} \hat{A}_{ij} Z_j$ for all i.
3: **for** each $i \neq j$ **do**
4: Compute the penalties: Set $\ell_{ij} = (\sigma(-2p_i) - (1 - Z_i)/2) \cdot Z_j$.
5: Update the weights: Set $W_{ij}^+ = W_{ij}^+ \cdot \beta^{\ell_{ij}}$; $W_{ij}^- = W_{ij}^- \cdot \beta^{-\ell_{ij}}$.
6: **for** each $i \neq j$ **do**
7: Compute edge weights: $\hat{A}_{ij} = \frac{\lambda}{\sum_{\ell \neq i}(W_{i\ell}^+ + W_{i\ell}^-)} \cdot \left(W_{ij}^+ - W_{ij}^-\right)$.

We then describe how to handle the more general case of learning t-wise MRFs. This requires working with multilinear polynomials, and studying their behavior (especially, how small they can be) under δ-unbiased distributions.

II. PRELIMINARIES

We will use the following notations and conventions.

- For a vector $x \in \mathbb{R}^n$, $x_{-i} \in \mathbb{R}^{[n]\setminus\{i\}}$ denotes $(x_j : j \neq i)$.
- We write multilinear polynomials $p : \mathbb{R}^n \to \mathbb{R}$ as $p(x) = \sum_I \hat{p}(I) \prod_{i \in I} x_i$; in particular, $\hat{p}(I)$ denotes the coefficient of the monomial $\prod_{i \in I} x_i$ in the polynomial. Let $\|p\|_1 = \sum_I |\hat{p}(I)|$.
- For a multilinear polynomial $p : \mathbb{R}^n \to \mathbb{R}$, we let $\partial_i p(x) = \sum_{J : J \not\ni i} \hat{p}(J \cup \{i\}) \prod_{j \in J} x_j$ denote the partial derivative of p with respect to x_i. Similarly, for $I \subseteq [n]$, let $\partial_I p(x) = \sum_{J : J \cap I = \emptyset} \widehat{p}(J \cup I) \prod_{j \in J} x_j$ denote the partial derivative of p with respect to the variables $(x_i : i \in I)$.
- For a multilinear polynomial $p : \mathbb{R}^n \to \mathbb{R}$, we say $I \subseteq [n]$ is a *maximal monomial* of p if $\hat{p}(J) = 0$ for all $J \supset I$ (i.e., there is no non-zero monomial that strictly contains I).

III. LEARNING SPARSE GENERALIZED LINEAR MODELS

We first describe our *Sparsitron* algorithm for learning sparse GLMs. In the next section we show how to learn MRFs using this algorithm. The main theorem of this section is the following:

Theorem III.1. *Let $\mathcal{D}$ be a distribution on $[-1,1]^n \times \{0,1\}$ where for $(X,Y) \sim \mathcal{D}$, $E[Y|X = x] = u(w \cdot x)$ for a non-decreasing 1-Lipschitz function $u : \mathbb{R} \to [0,1]$. Suppose that $\|w\|_1 \leq \lambda$ for a known $\lambda \geq 0$. Then, there exists an algorithm that for all $\varepsilon, \delta \in [0,1]$ given $T = O(\lambda^2(\ln(n/\delta\varepsilon))/\varepsilon^2)$ independent examples from $\mathcal{D}$, produces a vector $v \in \mathbb{R}^n$ such that with probability at least $1 - \delta$,*

$$\mathop{\mathbb{E}}_{(X,Y) \leftarrow \mathcal{D}}[(u(v \cdot X) - u(w \cdot X))^2] \leq \varepsilon. \tag{III.1}$$

The run-time of the algorithm is $O(nT)$. Moreover, the algorithm can be run in an online manner.

Proof: We assume without loss of generality that $w_i \geq 0$ for all i and that $\|w\|_1 = \lambda$; if not, we can map examples (x, y) to $((x, -x, 0), y)$ and work in the new space. For any vector $v \in \mathbb{R}^n$, define the *risk* of v $\varepsilon(v) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[(u(v \cdot X) - u(w \cdot X))^2]$. Let $\mathbf{1}$ denote the all 1's vector.

Our approach is to use the regret bound for the *Hedge* algorithm of Freund and Schapire [21]. Let $T \geq 1$, $\beta \in [0,1]$ be parameters to be chosen later and $M = C'''T\ln(1/\delta)/\varepsilon^2$ for a constant C''' to be chosen later. The algorithm is shown in Algorithm 2. The inputs to the algorithm are $T + M$ independent examples $(x^1, y^1,), \ldots, (x^T, y^T)$ and $(a^1, b^1), \ldots, (a^M, b^M)$ drawn from $\mathcal{D}$.

Algorithm 2 SPARSITRON

1: Initialize $w^0 = \mathbf{1}/n$.
2: **for** $t = 1, \ldots, T$ **do**
3: Let $p^t = w^{t-1}/\|w^{t-1}\|_1$.
4: Define $\ell^t \in \mathbb{R}^n$ by $\ell^t = (1/2)(\mathbf{1} + (u(\lambda p^t \cdot x^t) - y^t)x^t)$.
5: Update the weight vectors w^t: for each $i \in [n]$, set $w_i^t = w_i^{t-1} \cdot \beta^{\ell_i^t}$.
6: **for** $t = 1, \ldots, T$ **do**
7: Compute the *empirical risk*

$$\hat{\varepsilon}(\lambda p^t) = (1/M)\sum_{j=1}^{M} \left(u(\lambda p^t \cdot a^j) - b^j\right)^2.$$

8: RETURN $v = \lambda p^j$ for $j = \arg\min_{t \in [T]} \hat{\varepsilon}(\lambda p^t)$.

We add the $\mathbf{1}$ in Step 4 of Algorithm 2 to be consistent with [21] who work with loss vectors in $[0,1]^n$.

We next analyze our algorithm and show that for suitable parameters β, T, M, it achieves the guarantees of the theorem. We first show that the sum of the risks $\varepsilon(\lambda p^1), \ldots, \varepsilon(\lambda p^T)$ is small with high probability over the examples; the claim then follows by a simple Chernoff bound to argue that for M sufficiently big, the empirical estimates of the risk, $\hat{\varepsilon}(\lambda p^1), \ldots, \hat{\varepsilon}(\lambda p^T)$ are close to the true risks.

Observe that $\ell^t \in [0,1]^n$ and associate each $i = 1, \ldots, n$ with an expert and then apply the analysis of Freund and Schapire (c.f. [21], Theorem 5). In particular, setting $\beta = 1/(1 + \sqrt{(\ln n)/T})$, we get that

$$\sum_{t=1}^{T} p^t \cdot \ell^t \leq \min_{i \in [n]} \sum_{t=1}^{T} \ell_i^t + O(\sqrt{T \ln n} + (\ln n)). \tag{III.2}$$

Let random variable $Q^t = p^t \cdot \ell^t - (w/\lambda)\cdot\ell^t$. Note that $Q^t \in [-1,1]$. Let

$$Z^t = Q^t - \mathop{\mathbb{E}}_{(x^t,y^t)}[Q^t \mid (x^1,y^1),\ldots,(x^{t-1},y^{t-1})].$$

Then, $Z^1,\ldots,Z^T$ form a martingale difference sequence with respect to the sequence $(x^1,y^1),\ldots,(x^T,y^T)$ and are bounded between $[-2,2]$. Therefore, by Azuma-Hoeffding inequality for bounded martingale difference sequences, with probability at least $1-\delta$, we have $\left|\sum_{t=1}^T Z^t\right| \leq O(\sqrt{T\ln(1/\delta)})$. Thus, with probability at least $1-\delta$,

$$\sum_{t=1}^T \mathop{\mathbb{E}}_{(x^t,y^t)}[Q^t \mid (x^1,y^1),\ldots,(x^{t-1},y^{t-1})] \leq \sum_{t=1}^T Q^t + O(\sqrt{T\ln(1/\delta)}). \quad \text{(III.3)}$$

Now, for a fixed $(x^1,y^1),\ldots,(x^{t-1},y^{t-1})$, taking expectation with respect to (x^t,y^t), we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x^t,y^t)}[Q^t \mid (x^1,y^1),\ldots,(x^{t-1},y^{t-1})] = \\
&\qquad \mathop{\mathbb{E}}_{(x^t,y^t)}\left[(p^t - (1/\lambda)w)\cdot\ell^t\right] \\
&= (1/2)\mathop{\mathbb{E}}_{(x^t,y^t)}\left[(p^t - (1/\lambda)w)\cdot(u(\lambda p^t\cdot x^t) - y^t)x^t\right] \\
&= (1/2\lambda)\mathop{\mathbb{E}}_{x^t}\left[(\lambda p^t\cdot x^t - w\cdot x^t)(u(\lambda p^t\cdot x^t) - u(w\cdot x^t))\right] \\
&\geq (1/2\lambda)\mathop{\mathbb{E}}_{x^t}\left[(u(\lambda p^t\cdot x^t) - u(w\cdot x^t))^2\right] \\
&\text{(for all } a,b\in\mathbb{R},\ (a-b)(u(a)-u(b)) \geq (u(a)-u(b))^2\text{).} \\
&= (1/2\lambda)\cdot\varepsilon(\lambda p^t). \quad \text{(III.4)}
\end{aligned}$$

Therefore, for a fixed $(x^1,y^1),\ldots,(x^{t-1},y^{t-1})$, we have

$$(1/2\lambda)\varepsilon(\lambda p^t) \leq \mathop{\mathbb{E}}_{(x^t,y^t)}[Q^t \mid (x^1,y^1),\ldots,(x^{t-1},y^{t-1})].$$

Combining the above with Equations III.2, III.3, we get that with probability at least $1-\delta$,

$$\begin{aligned}
&(1/2\lambda)\sum_{t=1}^T \varepsilon(\lambda p^t) \leq \sum_{t=1}^T Q^t + O(\sqrt{T\ln(1/\delta)}) \\
&\leq \min_{i\in[n]}\sum_{t=1}^T \ell_i^t - \sum_{t=1}^T (w/\lambda)\cdot\ell^t + O(\sqrt{T\ln n} + (\ln n)) + \\
&\qquad O(\sqrt{T\ln(1/\delta)}). \quad \text{(III.5)}
\end{aligned}$$

Now, let $L = \sum_{t=1}^T \ell^t$. Then,

$$\min_{i\in[n]}\sum_{t=1}^T \ell_i^t - \sum_{t=1}^T (1/\lambda)w\cdot\ell^t = \min_{i\in[n]} L_i - (w/\lambda)\cdot L \leq 0,$$

where the last inequality follows as $\|w\|_1 = \lambda$. Therefore, with probability at least $1-\delta$,

$$(1/2\lambda)\sum_{t=1}^T \varepsilon(\lambda p^t) = O(\sqrt{T\ln(1/\delta)}) + O(\sqrt{T\ln n} + (\ln n)).$$

In particular, for $T > C''\lambda^2(\ln(n/\delta))/\varepsilon^2$ for a sufficiently big constant C'', with probability at least $1-\delta$,

$$\min_{t\in[T]} \varepsilon(\lambda p^t) \leq O(\lambda)\cdot\frac{\sqrt{T\ln(1/\delta)} + \sqrt{T\ln n} + \ln n}{T} \leq \varepsilon/2.$$

Now set $M = C'''\ln(T/\delta)/\varepsilon^2$ so that by a Chernoff-Hoeffding bound as in Fact III.2, with probability at least $1-\delta$, for every $t\in[T]$, $|\varepsilon(\lambda p^t) - \hat{\varepsilon}(\lambda p^t)| \leq \varepsilon/4$. Therefore, with probability at least $1-2\delta$, $\varepsilon(v) \leq \varepsilon/4 + \hat{\varepsilon}(v) \leq \varepsilon$. Note that the number of samples needed is $T + M = O(\lambda^2\ln(n/\varepsilon\delta)/\varepsilon^2)$. The theorem follows.

■

Fact III.2. *There exists a constant $C > 0$ such that the following holds. Let $v\in\mathbb{R}^n$ and let $(a^1,b^1),\ldots,(a^M,b^M)$ be independent examples from $\mathcal{D}$. Then, for all $\rho,\gamma \geq 0$, and $M \geq C\ln(1/\rho)/\gamma^2$,*

$$\mathsf{Pr}\left[\left|(1/M)\left(\sum_{j=1}^M (u(v\cdot a^j) - b^j)^2\right) - \varepsilon(v)\right| \geq \gamma\right] \leq \rho.$$

IV. Recovering affine functions from ℓ_2 minimization

In this section we show that running the Sparsitron algorithm with sufficiently low error parameter ε will result in an ℓ_∞ approximation to the unknown weight vector. We will use this strong approximation to reconstruct the dependency graphs of Ising models as well as the edge weights.

Our analysis relies on the following important definition:

Definition IV.1. *A distribution $\mathcal{D}$ on $\{1,-1\}^n$ is δ-unbiased if for $X\sim\mathcal{D}$, $i\in[n]$, and any partial assignment x to $(X_j : j\neq i)$,*

$$\min(\mathsf{Pr}[X_i = 1 | X_{-i} = x], \mathsf{Pr}[X_i = -1 | X_{-i} = x]) \geq \delta.$$

We will use the following elementary property of sigmoid.

Claim IV.2. *For $a,b\in R$,*

$$|\sigma(a) - \sigma(b)| \geq e^{-|a|-3}\cdot\min(1,|a-b|).$$

Proof: Fix $a\in R$ and let $\gamma = \min(1,|a-b|)$. Then, since σ is monotonic

$$|\sigma(a) - \sigma(b)| \geq \min(\sigma(a+\gamma) - \sigma(a), \sigma(a) - \sigma(a-\gamma)).$$

Now, it is easy to check by a case-analysis that for all $a,a'\in\mathbb{R}$,

$$|\sigma(a) - \sigma(a')| \geq \min(\sigma'(a),\sigma'(a'))\cdot|a-a'|.$$

Further, for any t, $\sigma'(t) = 1/(2+e^t+e^{-t}) \geq e^{-|t|}/4$. Combining the above two, we get that

$$\sigma(a+\gamma) - \sigma(a) \geq (1/4)\min(e^{-|a+\gamma|}, e^{-|a|})\cdot\gamma \geq (1/4)e^{(-|a|-\gamma)}\gamma.$$

Similarly, we get

$$\sigma(a) - \sigma(a-\gamma) \geq 4\min(e^{-|a-\gamma|}, e^{-|a|})\cdot\gamma \geq (1/4)e^{(-|a|-\gamma)}\gamma.$$

The claim now follows by substituting $\gamma = \min(1, |a-b|)$ (and noting that $1/4 \geq e^{-2}$). ∎

Lemma IV.3. *Let D be a δ-unbiased distribution on $\{1,-1\}^n$. Suppose that for two vectors $v, w \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$, $\mathbb{E}_{X\sim D}[(\sigma(w\cdot X+\alpha)-\sigma(v\cdot X+\beta))^2] \leq \varepsilon$ where $\varepsilon < \delta\cdot\exp(-2\|w\|_1 - 2|\alpha| - 6)$. Then,*

$$\|v-w\|_\infty \leq O(1)\cdot e^{\|w\|_1+|\alpha|}\cdot\sqrt{\varepsilon/\delta}.$$

Proof: For brevity, let $p(x) = w\cdot x+\alpha$, and $q(x) = v\cdot x+\beta$. Fix an index $i\in[n]$ and let $X\sim D$.

Now, for any $x\in\{1,-1\}^n$, by Claim IV.2,

$$|\sigma(p(x))-\sigma(q(x))| \geq e^{-\|w\|_1-|\alpha|-3}\cdot\min\left(1,|p(x)-q(x)|\right).$$

Let $x^{i,+}\in\{1,-1\}^n$ (respectively $x^{i,-}$) denote the vector obtained from x by setting $x_i = 1$ (respectively $x_i=-1$). Note that $p(x^{i,+})-p(x^{i,-}) = 2w_i$ and $q(x^{i,+})-q(x^{i,-}) = 2v_i$. Therefore,

$$p(x^{i,+})-q(x^{i,+})-(p(x^{i,-})-q(x^{i,-})) = 2(w_i-v_i).$$

Thus,

$$\max\left(\left|p(x^{i,+})-q(x^{i,+})\right|,\left|p(x^{i,-})-q(x^{i,-})\right|\right) \geq |w_i-v_i|.$$

Therefore, for any fixing of X_{-i}, as X is δ-unbiased,

$$\Pr_{X_i|X_{-i}}[|p(X)-q(X)|\geq|w_i-v_i|]\geq\delta.$$

Hence, combining the above inequalities, $\varepsilon \geq \mathbb{E}_X\left[(\sigma(p(X))-\sigma(q(X)))^2\right] \geq e^{-2\|w\|_1-2|\alpha|-6}\cdot\delta\cdot\min\left(1,|w_i-v_i|^2\right)$. As $\varepsilon < e^{-2\|w\|_1-2|\alpha|-6}\delta$, the above inequality can only hold if $|w_i-v_i|<1$ so that $|w_i-v_i| < e^{\|w\|_1+|\alpha|+3}\cdot\sqrt{\varepsilon/\delta}$. The claim now follows. ∎

V. Learning Ising Models

Definition V.1. *Let $A\in\mathbb{R}^{n\times n}$ be a* weight matrix *and $\theta\in\mathbb{R}^n$ be a* mean-field *vector. The associated n-variable* Ising model *is a distribution $\mathcal{D}(A,\theta)$ on $\{1,-1\}^n$ given by the condition*

$$\Pr_{Z\leftarrow\mathcal{D}(A,\theta)}[Z=z]\propto\exp\left(\sum_{i\neq j\in[n]}A_{ij}z_iz_j+\sum_i\theta_iz_i\right).$$

The dependency graph *of $\mathcal{D}(A,\theta)$ is the graph G formed by all pairs $\{i,j\}$ with $|A_{ij}|\neq 0$. We define $\lambda(A,\theta) = \max_i(\sum_j|A_{ij}|+|\theta_i|)$ to be the* width *of the model.*

We give a simple, sample-efficient, and online algorithm for recovering the parameters of an Ising model.

Theorem V.2. *Let $\mathcal{D}(A,\theta)$ be an n-variable Ising model with width $\lambda(A,\theta)\leq\lambda$. There exists an algorithm that given λ, $\varepsilon,\rho\in(0,1)$, and $N = O(\lambda^2\exp(O(\lambda))/\varepsilon^4)\cdot(\log(n/\rho\varepsilon))$ independent samples $Z^1,\ldots,Z^N\leftarrow\mathcal{D}(A,\theta)$ produces $\hat{A}$ such that with probability at least $1-\rho$,*

$$\|A-\hat{A}\|_\infty\leq\varepsilon.$$

The run-time of the algorithm is $O(n^2N)$. Moreover, the algorithm can be run in an online manner.

Proof: The starting point for our algorithm is the following observation. Let $Z\leftarrow\mathcal{D}(A,\theta)$. Then, for any $i\in[n]$ and any $x\in\{1,-1\}^{[n]\setminus\{i\}}$,

$$\Pr[Z_i=-1|Z_{-i}=x] = \frac{1}{1+\exp(2\sum_{j\neq i}A_{ij}x_j+\theta_i)} = \sigma(w(i)\cdot x+\theta_i), \quad \text{(V.1)}$$

where we define $w(i)\in R^{[n]\setminus\{i\}}$ with $w(i)_j = -2A_{ij}$ for $j\neq i$. This allows us to use our Sparsitron algorithm for learning GLMs.

For simplicity, we describe our algorithm to infer the coefficients A_{nj} for $j\neq n$; it extends straightforwardly to recover the weights $\{A_{ij}:j\neq i\}$ for each i. Let $Z\leftarrow\mathcal{D}(A,\theta)$ and let $X\equiv(Z_1,\ldots,Z_{n-1},1)$, and $Y=(1-Z_n)/2$. Then, from the above we have that

$$\mathbb{E}[Y|X] = \sigma(w(n)\cdot X),$$

where $w(n)\in\mathbb{R}^n$ with $w(n)_j = -2A_{nj}$ for $j<n$, and $w(n)_n=\theta_i$. Note that $\|w(n)\|_1\leq 2\lambda$. Further, σ is a monotone 1-Lipschitz function. Let $\gamma\in(0,1)$ be a parameter to be chosen later. We now apply the Sparsitron algorithm to compute a vector $v(n)\in\mathbb{R}^n$ so that with probability at least $1-\rho/n^2$,

$$\mathbb{E}[(\sigma(w(n)\cdot X)-\sigma(v(n)\cdot X))^2]\leq\gamma. \quad \text{(V.2)}$$

We set $\hat{A}_{nj} = -(v(n)_j)/2$ for $j<n$. We next argue that Equation V.2 in fact implies $\|w(n)-v(n)\|_\infty\ll 1$. To this end, we will use the following easy fact (see e.g. Bresler [13]):

Fact V.3. *For $Z\leftarrow\mathcal{D}(A,\theta)$, $i\in[n]$, and any partial assignment x to Z_{-i}, $\min(\Pr[Z_i=-1|Z_{-i}=x],\Pr[Z_i=1|Z_{-i}=x]) \geq (1/2)e^{-2\lambda(A,\theta)}\geq(1/2)e^{-2\lambda}$.*

That is, the distribution Z is δ-unbiased for $\delta = (1/2)e^{-2\lambda}$. Note that $w(n)\cdot X = \sum_{j<n}w(n)_jZ_j+w(n)_n$ and $v(n)\cdot X = \sum_{j<n}v(n)_jZ_j+v(n)_n$. Therefore, as $(Z_1,\ldots,Z_{n-1})$ is δ-unbiased, by Lemma IV.3 and Equation V.2, we get

$$\max_{j<n}|v(n)_j-w(n)_j|\leq O(1)\exp(2\lambda)\cdot\sqrt{\gamma/\delta},$$

if $\gamma\leq c\delta\cdot\exp(-4\lambda)\leq c\exp(-5\lambda)$ for a sufficiently small c. Thus, if we set $\gamma = c'\exp(-5\lambda)\varepsilon^2$ for a sufficiently small constant c', then we get

$$\max_{j<n}|A_{nj}-\hat{A}_{nj}| = (1/2)\|v(n)-w(n)\|_\infty\leq\varepsilon.$$

By a similar argument for $i=1,\ldots,n-1$ and taking a union bound, we get estimates $\hat{A}_{ij}$ for all $i\neq j$ so that with

probability at least $1-\rho$,

$$\max_{i\neq j} |A_{ij} - \hat{A}_{ij}| \leq \varepsilon.$$

Note that by Theorem III.1, the number of samples needed to satisfy Equation V.2 is

$$O((\lambda/\gamma)^2 \cdot (\log(n/\rho\gamma))) = O(\lambda^2 \exp(10\lambda)/\varepsilon^4) \cdot (\log(n/\rho\varepsilon)).$$

This proves the theorem. ∎

The above theorem immediately implies an algorithm for recovering the dependency graph of an Ising model with nearly optimal sample complexity.

Corollary V.4. *Let $\mathcal{D}(A,\theta)$ be an n-variable Ising model with width $\lambda(A,\theta) \leq \lambda$ and each non-zero entry of A at least $\eta > 0$ in absolute value. There exists an algorithm that given λ, $\eta, \rho \in (0,1)$, and $N = O(\exp(O(\lambda))/\eta^4) \cdot (\log(n/\rho\eta))$ independent samples $Z^1, \ldots, Z^N \leftarrow \mathcal{D}(A,\theta)$ recovers the underlying dependency graph of $\mathcal{D}(A,\theta)$ with probability at least $1-\rho$. The run-time of the algorithm is $O(n^2 N)$. Moreover, the algorithm can be run in an online manner.*

Proof: The claim follows immediately from Theorem V.2 by setting $\varepsilon = \eta/2$ to compute $\hat{A}$ and taking the edges E to be $\{\{i,j\} : |\hat{A}_{ij}| \geq \eta/2\}$. ∎

It is instructive to compare the upper bounds from Corollary V.4 with known unconditional lower bounds on the sample complexity of learning Ising models with n vertices due to Santhanam and Wainwright [24]. They prove that, even if the weights of the underlying graph are known, any algorithm for learning the graph structure must use $\Omega(\frac{2^{\lambda/4} \cdot \log n}{\eta \cdot 2^{3\eta}})$ samples. Hence, the sample complexity of our algorithm is near the best-known information-theoretic lower bound.

VI. Recovering polynomials from ℓ_2 minimization

In order to obtain results for learning general Markov Random Fields, we need to extend our learning results from previous sections to the case of sigmoids of low-degree polynomials. In this section, we prove that for any polynomial $p : \mathbb{R}^n \to \mathbb{R}$, minimizing the ℓ_2-loss with respect to a sigmoid under a δ-unbiased distribution $\mathcal{D}$ also implies closeness as a polynomial. That is, for two polynomials $p, q : \mathbb{R}^n \to \mathbb{R}$ if $\mathbb{E}_{X\sim\mathcal{D}}[(\sigma(p(X)) - \sigma(q(X)))^2]$ is sufficiently small, then $\|p-q\|_1 \ll 1$ (Lemma VI.4) and that the coefficients of maximal monomials of p can be inferred from q (Lemma VI.2). These results will allow us to recover the structure and parameters of MRFs when combined with Sparsitron.

The exact statements and arguments here are similar in spirit to Lemma IV.3 and its proof but are more subtle. To start with, we need the following property of δ-unbiased distributions which says that low-degree polynomials are not too small with non-trivial probability (aka *anti-concentration*) under δ-unbiased distributions.

Lemma VI.1. *There is a constant $c > 0$ such that the following holds. Let D be a δ-unbiased distribution on $\{1,-1\}^n$. Then, for any multilinear polynomial $s : \mathbb{R}^n \to \mathbb{R}$, and any maximal monomial $I \neq \emptyset \subseteq [n]$ in s,*

$$\Pr_{X\sim\mathcal{D}}[|s(X)| \geq |\widehat{s}(I)|] \geq \delta^{|I|}.$$

Proof: We prove the claim by induction on $|I|$. For an $i \in [n]$, let $x^{i,+} \in \{1,-1\}^n$ (respectively $x^{i,-}$) denote the vector obtained from x by setting $x_i = 1$ (respectively $x_i = -1$). Note that $x^{i,+}, x^{i,-}$ only depend on x_{-i}. Let $X \sim \mathcal{D}$.

Suppose $I = \{i\}$ so that $s(x) = \widehat{s}(\{i\})x_i + s'(x_{-i})$ for some polynomial s' that only depends on x_{-i}. Note that $\max(|s(x^{i,+})|, |s(x^{i,-})|) \geq |\widehat{s}(\{i\})|$. Therefore, for any fixing of X_{-i}, as X is δ-unbiased,

$$\Pr_{X_i|X_{-i}}[|s(X)| \geq |\widehat{s}(\{i\})|] \geq \delta.$$

Now, suppose $|I| = \ell \geq 2$ and that the claim is true for all polynomials and all monomials of size at most $\ell - 1$. Let $i \in I$. Then, $s(x) = x_i \cdot \partial_i(s(x_{-i})) + s'(x_{-i})$ for some polynomial s' that only depends on x_{-i}. Thus, $\max(|s(x^{i,+})|, |s(x^{i,-})|) \geq |\partial_i s(x_{-i})|$. Therefore, for any fixing of X_{-i}, as X is δ-unbiased,

$$\Pr_{X_i|X_{-i}}[|s(X)| \geq |\partial_i s(X_{-i})|] \geq \delta.$$

Now, let $J = I \setminus \{i\}$ and observe that J is a maximal monomial in $r(x_{-i}) \equiv \partial_i s(x_{-i})$ with $\widehat{r}(J) = \widehat{s}(I)$. Therefore, by the induction hypothesis,

$$\Pr_{X_{-i}}[|\partial_i s(X_{-i})| \geq |\widehat{s}(I)|] \geq \delta^{\ell-1}.$$

Combining the last two inequalities, we get that $\Pr[|s(X)| \geq \widehat{s}(I)] \geq \delta^{\ell}$. The claim now follows by induction. ∎

The next lemma shows that for unbiased distributions $\mathcal{D}$, and two low-degree polynomials $p, q : \mathbb{R}^n \to \mathbb{R}$, if $\mathbb{E}_{X\sim\mathcal{D}}[(\sigma(p(x)) - \sigma(q(x))^2]$ is small, then one can infer the coefficients of the maximal monomials of p from q[4].

Lemma VI.2. *Let $\mathcal{D}$ be a δ-unbiased distribution on $\{1,-1\}^n$. Let p, q be two multilinear polynomials $p, q : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}_{X\sim\mathcal{D}}[(\sigma(p(x)) - \sigma(q(x))^2] \leq \varepsilon$. Then, for every maximal monomial $I \subseteq [n]$ of p, and any $\rho > 0$,*

$$\Pr_{X\sim\mathcal{D}}[|\widehat{p}(I) - \partial_I q(X)| > \rho] \leq \frac{e^{2\|p\|_1+6}\varepsilon}{\rho^2 \delta^{|I|}}.$$

Proof: Let $X \sim \mathcal{D}$ and fix a maximal monomial $I \subseteq [n]$ in p. Now, for any $x \in \{1,-1\}^n$, by Claim IV.2,

$$|\sigma(p(x)) - \sigma(q(x))| \geq e^{-\|p\|_1-3} \cdot \min(1, |p(x) - q(x)|).$$

Therefore,

$$\mathbb{E}\left[\min\left(1, |p(X) - q(X)|^2\right)\right] \leq e^{2\|p\|_1+6}\varepsilon.$$

[4]Note that under the hypothesis of the lemma, the coefficients of p and q can nevertheless be far.

Hence, for every $\rho \in (0,1)$,

$$\Pr_X[|p(X) - q(X)| > \rho] \leq e^{2\|p\|_1+6}\varepsilon/\rho^2.$$

Now consider a fixing of all variables not in I to $z \in \{1,-1\}^{[n]\setminus I}$ and let $r_z(x_I)$ be the polynomial obtained by the resulting fixing. Now, $\Pr_X[|p(X) - q(X)| > \rho] = \sum_{z\in\{1,-1\}^{[n]\setminus I}} \Pr[X_{[n]\setminus I} = z] \cdot \Pr[|r_z(X_I)| > \rho \mid X_{[n]\setminus I} = z]$. Further, $\widehat{r}(I) = \widehat{p}(I) - \partial_I q(z)$ as I is maximal in p.

Conditioned on the event that $|\widehat{r}(I)| > \rho$, for a random choice of $X_{[n]\setminus I}$, we have from Lemma VI.1 that $\Pr_{X_I}[|r_z(X_I)| > \rho] \geq \delta^{|I|}$. Thus we have

$$\Pr_X[|p(X) - q(X)| > \rho] \geq \delta^{|I|} \cdot \Pr_{X_{[n]\setminus I}}\left[\left|\widehat{p}(I) - \partial_I q(X_{[n]\setminus I})\right| > \rho\right]$$

Combining the above equations we get that

$$\Pr_X[|\widehat{p}(I) - \partial_I q(X)| > \rho] \leq \frac{e^{2\|p\|_1+6}\varepsilon}{\rho^2\delta^{|I|}}.$$

■

The next claim shows that under the assumptions of Lemma VI.2, the highest degree monomials of p, q are close to each other.

Lemma VI.3. *Let $\mathcal{D}$ be a δ-unbiased distribution on $\{1,-1\}^n$. Let p, q be two multilinear polynomials $p, q : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}_{X\sim\mathcal{D}}[(\sigma(p(x)) - \sigma(q(x))^2] \leq \varepsilon$ where $\varepsilon < e^{-2\|p\|_1-6}\delta^{|I|}$. Then, for every maximal monomial $I \subseteq [n]$ of $(p-q)$,*

$$|\widehat{p}(I) - \widehat{q}(I)| \leq e^{\|p\|_1+3} \cdot \sqrt{\varepsilon/\delta^{|I|}}.$$

Proof: Fix a maximal monomial $I \subseteq [n]$ in $(p-q)$. Now, for any X, by Claim IV.2,

$$|\sigma(p(X)) - \sigma(q(X))| \geq e^{-\|p\|_1-3} \cdot \min(1, |p(X) - q(X)|).$$

On the other hand, as X is δ-unbiased, by Lemma VI.1, with probability at least $\delta^{|I|}$, $|p(X) - q(X)| \geq |\widehat{p}(I) - \widehat{q}(I)|$. Therefore, $\varepsilon \geq \mathbb{E}_X\left[(\sigma(p(X)) - \sigma(q(X)))^2\right] \geq e^{-2\|p\|_1-6} \cdot \delta^{|I|} \cdot \min\left(1, |\widehat{p}(I) - \widehat{q}(I)|^2\right)$.

As $\varepsilon < e^{-2\|p\|_1-6}\delta^{|I|}$, the above inequality can only hold if $|\widehat{p}(I) - \widehat{q}(I)| < 1$ so that

$$|\widehat{p}(I) - \widehat{q}(I)| < e^{\|p\|_1+3}\sqrt{\varepsilon/\delta^{|I|}}.$$

The claim follows. ■

We next show that if $\mathbb{E}_{X\sim\mathcal{D}}[(\sigma(p(x)) - \sigma(q(x))^2] \ll n^{-t}$ is sufficiently small, then $\|p - q\|_1 \ll 1$.

Lemma VI.4. *Let D be a δ-unbiased distribution on $\{1,-1\}^n$. Let p, q be two multilinear polynomials $p, q : \mathbb{R}^n \to \mathbb{R}$ of degree t such that $\mathbb{E}_{X\sim\mathcal{D}}[(\sigma(p(x)) - \sigma(q(x))^2] \leq \varepsilon$ where $\varepsilon < e^{-2\|p\|_1-6}\delta^t$. Then,*

$$\|p - q\|_1 = O(1) \cdot (2t)^t e^{\|p\|_1} \cdot \sqrt{\varepsilon/\delta^t} \cdot \binom{n}{t}.$$

Proof: For a polynomial $s : \mathbb{R}^n \to \mathbb{R}$ of degree at most t, and $\ell \leq t$, let $s_{\leq\ell}$ denote the polynomial obtained from s by only taking monomials of degree at most ℓ and let $s_{=\ell}$ denote the polynomial obtained from s by only taking monomials of degree exactly ℓ.

For brevity, let $r = p - q$, and for $\ell \leq t$, let $\rho_\ell = \|r_{=\ell}\|_1 = \|p_{=\ell} - q_{=\ell}\|_1$. We will inductively bound $\rho_t, \rho_{t-1}, \ldots, \rho_1$.

From Lemma VI.3 applied to the polynomials p, q, we immediately get that

$$\rho_t = \|r_{=t}\|_1 \leq e^{\|p\|_1+3} \cdot \sqrt{\varepsilon/\delta^t} \cdot \binom{n}{t} \equiv \varepsilon_0. \tag{VI.1}$$

Now consider $I \subseteq [n]$ with $|I| = \ell$. Then, by an averaging argument, there is some fixing of the variables not in X_I so that for the polynomials p_I, q_I obtained by this fixing, and for the resulting distribution $\mathcal{D}_I$ on $\{1,-1\}^I$,

$$\mathbb{E}_{Y\sim\mathcal{D}_I}[(\sigma(p_I(Y)) - \sigma(q_I(Y)))^2] \leq \varepsilon.$$

Note that $\mathcal{D}_I$ is also δ-unbiased. Therefore, by Lemma VI.3 applied to the polynomials p, q, letting $r_I = p_I - q_I$, we get that

$$|\widehat{r}_I(I)| = |\widehat{p}_I(I) - \widehat{q}_I(I)| \leq e^{\|p\|_1+3} \cdot \sqrt{\varepsilon/\delta^{|I|}}.$$

We next relate the coefficients of r_I to that of r. As the polynomial r_I is obtained from r by fixing the variables not in I to some values in $\{1,-1\}$,

$$|\widehat{r}_I(I)| \geq |\widehat{r}(I)| - \sum_{J:J\supset I} |\widehat{r}(J)|.$$

Combining the above two inequalities, we get that

$$|\widehat{r}(I)| \leq e^{\|p\|_1+3} \cdot \sqrt{\varepsilon/\delta^\ell} + \sum_{J\supset I} |\widehat{r}(J)|.$$

Summing the above equation over all I of size exactly ℓ, we get

$$\begin{aligned}\|r_{=\ell}\|_1 = \sum_{I:|I|=\ell} |\widehat{r}(I)| &\leq e^{\|p\|_1+3} \cdot \sqrt{\varepsilon/\delta^\ell} \cdot \binom{n}{\ell} + \\ &\sum_{I:|I|=\ell}\left(\sum_{J\supset I} |\widehat{r}(J)|\right) \\ &\leq \varepsilon_0 + \sum_{I:|I|=\ell}\left(\sum_{J\supset I} |\widehat{r}(J)|\right) \\ = \varepsilon_0 + \sum_{j=\ell+1}^{t}\binom{j}{\ell} \cdot \left(\sum_{J:|J|=j} |\widehat{r}(J)|\right) &= \varepsilon_0 + \sum_{j=\ell+1}^{t}\binom{j}{\ell}\|r_{=j}\|_1.\end{aligned} \tag{VI.2}$$

Therefore, we get the recurrence,

$$\rho_\ell \leq \varepsilon_0 + \sum_{j=\ell+1}^{t}\binom{j}{\ell}\rho_j. \tag{VI.3}$$

We can solve the above recurrence by induction on ℓ. Specifically, we claim that the above implies $\rho_j \leq (2t)^{t-j} \cdot \varepsilon_0$. For $j = t$, the claim follows from Equation VI.1. Now, suppose the inequality holds for all $j > \ell$. Then, by Equation VI.3, as $\binom{j}{\ell} \leq j^{j-\ell}$,

$$\rho_\ell \leq \varepsilon_0 + \sum_{j=\ell+1}^{t} j^{j-\ell}(2t)^{t-j}\varepsilon_0 \leq \varepsilon_0 + \sum_{j=\ell+1}^{t} t^{j-\ell}(2t)^{t-j}\varepsilon_0$$

$$\leq t^{t-\ell} \cdot \varepsilon_0 \cdot \left(1 + \sum_{j=\ell+1}^{t} 2^{t-j}\right) = t^{t-\ell} \cdot \varepsilon_0 \cdot 2^{t-\ell}.$$

Therefore,

$$\|r\|_1 = \sum_{\ell=0}^{t} \|r_{=\ell}\|_1 \leq \sum_{\ell=0}^{t} (2t)^{t-\ell}\varepsilon_0 \leq \varepsilon_0 \cdot 2^{t+1}t^t.$$

The lemma now follows by plugging in the value of ε_0. ■

VII. Learning Markov Random Fields

We now describe how to apply the Sparsitron algorithm to recover the structure as well as parameters of binary t-wise MRFs.

We will use the characterization of MRFs via the Hammersley-Clifford theorem. Given a graph $G = (V, E)$ on n vertices, let $C_t(G)$ denote all cliques of size at most t in G. A binary t-wise MRF with dependency graph G is a distribution $\mathcal{D}$ on $\{1, -1\}^n$ where the probability density function of $\mathcal{D}$ can be written as

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

where $\mathcal{S} \subseteq C_t(G)$ and each $\psi_I : \mathbb{R}^n \to \mathbb{R}$ is a function that depends only on the variables in I. Note that if $t = 2$, this corresponds exactly to the Ising model. We call $\psi(x) = \sum_{I\in\mathcal{S}} \psi_I(x)$ the *factorization polynomial* of the MRF and G the *dependency graph* of the MRF.

Note that the factorization polynomial is a polynomial of degree at most t. However, different graphs and factorizations (i.e., functions $\{\psi_I\}$) could potentially lead to the same polynomial. To get around this we enforce the following non-degeneracy condition:

Definition VII.1. *For a t-wise MRF $\mathcal{D}$ on $\{1, -1\}^n$ we say an associated dependency graph G and factorization*

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

for $\mathcal{S} \subseteq C_t(G)$ is η-identifiable if for every maximal monomial J *in $\psi(x) = \sum_{I\in\mathcal{S}} \psi_I(x)$, $\left|\widehat{\psi}(J)\right| \geq \eta$ and every edge in G is covered by a non-zero monomial of ψ.*

We now state our main theorems for learning MRFs. Our first result is about *structure learning*, i.e., recovering the underlying dependency graph of a MRF. Roughly speaking, using $N = 2^{O(\lambda t)} \log(n/\eta)/\eta^4$ samples we can recover the underlying dependency graph of a η-identifiable MRF where λ is the maximum ℓ_1-norm of the derivatives of the factorization polynomial. The run-time of the algorithm is $O(M \cdot n^t)$. Note that $\max_i \|\partial_i \psi\|_1$ is analogous to the notion of width for Ising models (as in Corollary V.4). Thus, exponential dependence on it is necessary as in the Ising model and our sample complexity is in fact nearly optimal in all parameters.

Theorem VII.2. *Let $\mathcal{D}$ be a t-wise MRF on $\{1, -1\}^n$ with underlying dependency graph G and factorization polynomial $p(x) = \sum_{I\in C_t(G)} p_I(x)$ with $\max_i \|\partial_i p\|_1 \leq \lambda$. Suppose that $\mathcal{D}$ is η-identifiable. Then, there exists an algorithm that given λ, $\eta, \rho \in (0, 1/2)$, and*

$$N = \frac{e^{O(t)}e^{O(\lambda t)}}{\eta^4} \cdot (\log(n/\rho\eta))$$

independent samples from $\mathcal{D}$, recovers the underlying dependency graph G with probability at least $1 - \rho$. The run-time of the algorithm is $O(N \cdot n^t)$. Moreover, the algorithm can be run in an online manner.

Along with learning the dependency graph, given more samples, we can also approximately learn the parameters of the MRF: i.e., compute a t-wise MRF whose distribution is close as a pointwise-approximation to the original probability density function.

Theorem VII.3. *Let $\mathcal{D}$ be a t-wise MRF on $\{1, -1\}^n$ with underlying dependency graph G and factorization polynomial $\psi(x) = \sum_{I\in C_t(G)} \psi_I(x)$ with $\max_i \|\partial_i \psi\|_1 \leq \lambda$. There exists an algorithm that given λ, and $\varepsilon, \rho \in (0, 1/2)$, and*

$$N = \frac{(2t)^{O(t)}e^{O(\lambda t)}}{\varepsilon^4} \cdot n^{4t} \cdot (\log(n/\rho\varepsilon))$$

independent samples $Z^1, \ldots, Z^N \leftarrow \mathcal{D}$ produces a t-wise MRF $\mathcal{D}'$ with dependency graph H and a factorization polynomial $\varphi(x) = \sum_{I\in C_t(H)} \varphi_I(x)$ such that with probability at least $1 - \rho$:

$$\forall x,\ \Pr_{Z\sim\mathcal{D}}[Z = x] = (1 \pm \varepsilon) \Pr_{Z\sim\mathcal{D}'}[Z = x].$$

The algorithm runs in time $O(Nn^t)$ and can be run in an online manner.

We in fact show how to recover the parameters of a *log-polynomial* density defined as follows:

Definition VII.4. *A distribution $\mathcal{D}$ on $\{1, -1\}^n$ is said to be a log-polynomial distribution of degree t if for some multilinear polynomial $p : \mathbb{R}^n \to \mathbb{R}$ of degree t,*

$$\Pr_{X\sim\mathcal{D}}[X = x] \propto \exp(p(x)).$$

Theorem VII.5. *Let $\mathcal{D}$ be a log-polynomial distribution of degree at most t on $\{1, -1\}^n$ with the associated polynomial*

$p : \mathbb{R}^n \to \mathbb{R}$ *such that* $\max_i \|\partial_i p\|_1 \le \lambda$. *There exists an algorithm that given* λ, *and* $\varepsilon, \rho \in (0,1)$ *and*

$$N = \frac{(2t)^{O(t)} \cdot e^{O(\lambda t)}}{\varepsilon^4} \cdot (\log(n/\rho\varepsilon)),$$

independent samples $Z^1, \ldots, Z^N \leftarrow \mathcal{D}$, *finds a multilinear polynomial* $q : \mathbb{R}^n \to \mathbb{R}$ *such that with probability at least* $1 - \rho$

$$\|p - q\|_1 \le \varepsilon \cdot \binom{n}{t}.$$

Moreover, we can also find coefficients $(\hat{s}(I) : I \subseteq [n], |I| \le t)$ *such that with probability at least* $1 - \rho$, *for every maximal monomial* I *of* p, *we have* $|\widehat{p}(I) - \widehat{s}(I)| < \varepsilon$. *The run-time of the algorithm is* $O(N \cdot n^t)$ *and the algorithm can be run in an online manner.*

A. Learning the structure of MRFs

The following elementary properties of MRFs play a critical role in our analysis.

Lemma VII.6. *Let* $\mathcal{D}$ *be a* t*-wise MRF on* $\{1,-1\}^n$ *with underlying dependency graph* G *and factorization polynomial* $p(x) = \sum_{I \in C_t(G)} p_I(x)$ *with* $\max_i \|\partial_i p\|_1 \le \lambda$. *Then, the following hold for* $Z \leftarrow \mathcal{D}$:

- *For any* i, *and a partial assignment* $x \in \{1,-1\}^{[n]\setminus\{i\}}$, $\mathsf{Pr}[Z_i = -1 | Z_{-i} = x] = \sigma(-2\partial_i p(x))$.
- $\mathcal{D}$ *is* $(e^{-2\lambda}/2)$*-unbiased.*

Proof: For any $x \in \{1,-1\}^{[n]\setminus\{i\}}$,

$$\frac{\mathsf{Pr}[Z_i = 1 | Z_{-i} = x]}{\mathsf{Pr}[Z_i = -1 | Z_{-i} = x]} = \exp(2\partial_i p(x)).$$

Thus,

$$\mathsf{Pr}[Z_i = -1 | Z_{-i} = x] = \sigma(-2\partial_i p(x)).$$

Next, for each i, and any partial assignment x to Z_{-i},

$$\begin{aligned}\min\left(\mathsf{Pr}[Z_i = -1 | Z_{-i} = x], \mathsf{Pr}[Z_i = 1 | Z_{-i} = x]\right) &= \\ \min\left(\sigma(-2\partial_i p(x)), 1 - \sigma(-2\partial_i p(x))\right) &\\ \ge (1/2)e^{-2\|\partial_i p\|_1} \ge (1/2)e^{-2\lambda}.&\end{aligned}$$

■

We also need the following elementary fact about median:

Claim VII.7. *Let* X *be a real-valued random variable such that for some* $\alpha, \gamma \in \mathbb{R}$, $\mathsf{Pr}[|X - \alpha| > \gamma] < 1/4$. *Then, for* K *independent copies of* X, $X_1, X_2, \ldots, X_K$,

$$\mathsf{Pr}[|\textsc{Median}(X_1, \ldots, X_K) - \alpha| > \gamma] \le 2\exp(-\Omega(K)).$$

Proof of Theorem VII.2: We will show how to recover neighbors of the vertex n (for ease of notation). By repeating the argument for all $i \in [n]$, we will get the graph G.

The starting point for our algorithm is Lemma VII.6 that allow us to use Sparsitron algorithm via *feature expansion* and the properties of δ-unbiased distributions developed in Section VI.

Concretely, let $p' = -2\partial_n p$ and $\boldsymbol{p'} = (\widehat{p'}(I) : I \subseteq [n-1], |I| \le t-1)$. Similarly, for $x \in \{1,-1\}^{n-1}$, let $\boldsymbol{v}(x) = (\prod_{i \in I} x_i : I \subseteq [n-1], |I| \le t-1)$. Let $Z \sim \mathcal{D}$ and X be the distribution of $\boldsymbol{v}(Z_{-n})$ and let $Y = (1 - Z_n)/2$. Then, by Lemma VII.6, we have

$$\mathbb{E}[Y|X] = \sigma(\boldsymbol{p'} \cdot X).$$

Let $\delta = e^{-2\lambda}/2$, and let $\varepsilon \in (0,1)$, $K \ge 1$ be parameters to be chosen later. Our algorithm is shown in Figure 3. The intuition is as follows: We first apply Sparsitron to recover a polynomial q that approximates $\partial_n p$ in the sense that

$$\mathop{\mathbb{E}}_{Z}[(\sigma(-2\partial_n p(Z)) - \sigma(-2q(Z)))^2] < \varepsilon.$$

However, the above does not guarantee that the coefficients of q are close to those of $\partial_n p$. To overcome this, we exploit Lemma VI.2 that guarantees that for any maximal monomial I in $\partial_n p$, $\partial_I q(Z)$ is close to $\widehat{\partial_n p}(I)$ with high probability for $Z \sim \mathcal{D}$; concretely, in steps (4), (5), (6), we draw fresh samples from $\mathcal{D}$ and use the median evaluation of $\partial_I q(\ )$ as our estimate for $\widehat{\partial_n p}(I)$.

Algorithm 3 MRF RECOVERY

1: Initialize $H = \emptyset$ to be the empty graph.
2: Apply the Sparsitron algorithm as in Theorem III.1 to compute a vector $\boldsymbol{q}$ such that with probability at least $1 - \rho/2n^2$,
$$\mathbb{E}[(\sigma(\boldsymbol{p'} \cdot X) - \sigma(\boldsymbol{q} \cdot X))^2] \le \varepsilon.$$
3: Define a polynomial $q : \mathbb{R}^{n-1} \to \mathbb{R}$ by setting $\widehat{q}(I) = (-1/2)\boldsymbol{q}_I$ for all $I \subseteq [n-1]$.
4: Let $Z^1, \ldots, Z^K$ be additional independent samples from $\mathcal{D}$.
5: **for** each $I \subseteq [n-1]$, $|I| \le t-1$ **do**
6: If $\left|\textsc{median}\left(\partial_I q(Z^1), \ldots, \partial_I q(Z^K)\right)\right| > \eta/2$, then add the complete graph on $\{n\} \cup I$ to H.

We next argue that for a suitable choice of ε, K, with probability at least $1 - \rho/n$, the graph H contains all edges of G adjacent to vertex n.

Observe that by our definitions of $\boldsymbol{p'}, \boldsymbol{q}, X$ $\mathbb{E}_Z\left[(\sigma(-2\partial_n p(Z)) - \sigma(-2q(Z))^2\right] = \mathbb{E}[(\sigma(\boldsymbol{p'} \cdot X) - \sigma(\boldsymbol{q} \cdot X))^2] \le \varepsilon$.

(Here, we abuse notation and write $q(Z) = q(Z_1, \ldots, Z_{n-1})$ as the latter does not depend on Z_n.)

Further, as Z is δ-unbiased by Lemma VII.6, by Lemma VI.2 for any maximal monomial $I \subseteq [n-1]$ of $\partial_n p$, we have

$$\mathsf{Pr}\left[\left|\widehat{\partial_n p}(I) - \partial_I q(Z)\right| > \eta/4\right] < \frac{16e^{2\|p\|_1 + 6}\varepsilon}{\eta^2 \delta^{|I|}}.$$

Let $\varepsilon = e^{-2\lambda - 6}\eta^2\delta^t/64$ so that

$$\mathsf{Pr}\left[\left|\widehat{\partial_n p}(I) - \partial_I q(Z)\right| > \eta/4\right] < 1/4.$$

Therefore, by Claim VII.7, $\mathsf{Pr}\left[\left|\textsc{Median}(\partial_I q(Z^1), \ldots, \partial_I q(Z^K)) - \widehat{\partial_n p}(I)\right| > \eta/4\right] < 2\exp(-\Omega(K))$.

Taking $K = C\log(n^t/\rho)$ for a sufficiently big constant C, we get that with probability at least $1 - \rho/n$, for all maximal monomials I of $\partial_n p$, $\left|\textsc{Median}(\partial_I q(Z^1), \partial_I q(Z^2), \ldots, \partial_I q(Z^K)) - \widehat{\partial_n p}(I)\right| < \eta/4$.

Now, whenever the above happens, as the coefficients of maximal monomials of p are at least η in magnitude (by η-identifiability), our algorithm will add the complete graph on the variables of all maximal monomials of p involving vertex n to H.

Thus, the algorithm recognizes the neighbors of vertex n exactly with probability at least $1 - \rho/n$. Repeating the argument for each vertex $i \in [n]$ and taking a union bound over all vertices gives us the recovery guarantee of the theorem. It remains to bound the sample-complexity.

Note that $\|\boldsymbol{p'}\|_1 = 2\|\partial_n p\|_1 \le 2\lambda$. Therefore, by Theorem III.1, the number of samples needed for the call to Sparsitron in Step (2) of Algorithm 3 is

$$O(\lambda^2 \cdot \ln(n^t/\rho\varepsilon)/\varepsilon^2) = e^{O(t)} \cdot e^{O(\lambda \cdot t)} \cdot \ln(n/\rho\eta) \cdot (1/\eta^4).$$

As $K = Ct\ln(n/\rho)$, the above bound dominates the number of samples proving the theorem. ∎

B. Learning log-polynomial densities and parameters of MRFs

We first observe that Theorem VII.5 implies Theorem VII.3

Proof of Theorem VII.3: We apply Theorem VII.5 with error $\varepsilon' = \varepsilon n^{-t}$ to samples from $\mathcal{D}$ to obtain a polynomial $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that $\|\psi - \varphi\|_1 \le \varepsilon$. We build a new graph H as follows: For each monomial $I \subseteq [n]$ with $\widehat{\varphi}(I) \neq 0$, add all the edges in I to H. Let $\mathcal{D}'$ denote the t-wise MRF with dependency graph H and factorization polynomial φ. Since, $\|\psi - \varphi\|_1 \le \varepsilon$, it follows that for all x, $|\psi(x) - \varphi(x)| < \varepsilon$. Therefore, for all x,

$$\exp(\psi(x)) = \exp(\varphi(x) \pm \varepsilon) = (1 \pm 2\varepsilon)\exp(\varphi(x)).$$

The theorem now follows. ∎

We next prove Theorem VII.5. The proof is similar to that of Theorem V.2 and Theorem VII.2.

Proof of Theorem VII.5: For each i, we will show how to recover a polynomial q_i such that $\|\partial_i p - q_i\|_1 < \varepsilon \cdot \binom{n}{t-1}$. We can then combine these polynomials to obtain a polynomial q. One way to do so is as follows: For each $I \subseteq [n]$, let $i = \arg\min(I)$, and define $\widehat{q}(I) = \widehat{q_i}(I \setminus \{i\})$. Then,

$$\|p - q\|_1 = \sum_I |\widehat{p}(I) - \widehat{q}(I)| = \sum_{i=1}^n \sum_{I : \arg\min(I) = i} |\widehat{p}(I) - \widehat{q}(I)|$$
$$\le \sum_{i=1}^n \|\partial_i p - q_i\| \le \varepsilon \cdot n \cdot \binom{n}{t-1}.$$

Here we show how to find a polynomial q_n such that with probability at least $1 - \rho/n$,

$$\|\partial_n p - q_n\|_1 < \varepsilon \cdot \binom{n}{t-1}. \tag{VII.1}$$

The other cases can be handled similarly and the theorem then follows from the above argument.

As in Theorem VII.2, we exploit Lemma VII.6 to employ our Sparsitron algorithm for learning GLMs via *feature expansion*. Concretely, let $p' = -2\partial_n p$ and $\boldsymbol{p'} = (\widehat{p'}(I) : I \subseteq [n-1], |I| \le t-1)$. Similarly, for $x \in \{1, -1\}^{n-1}$, let $\boldsymbol{v}(x) = (\prod_{i \in I} x_i : I \subseteq [n-1], |I| \le t-1)$. Let $Z \sim \mathcal{D}$ and X be the distribution of $\boldsymbol{v}(x)$ and let $Y = (1 - Z_n)/2$. Then, from the above arguments, we have

$$\mathbb{E}[Y|X] = \sigma(\boldsymbol{p'} \cdot X).$$

Note that $\|\boldsymbol{p'}\|_1 = 2\|\partial_n p\|_1 \le 2\lambda$. Let $\gamma \in (0, 1)$ be a parameter to be chosen later. We now apply the Sparsitron algorithm as in Theorem III.1 to compute a vector $\boldsymbol{q'} \in \mathbb{R}^n$ such that with probability at least $1 - \rho/n$,

$$\mathbb{E}[(\sigma(\boldsymbol{p'} \cdot X) - \sigma(\boldsymbol{q'} \cdot X))^2] \le \gamma.$$

We define polynomial q_n by setting $\widehat{q_n}(I) = (-1/2) \cdot \boldsymbol{q'}_I$ for all $I \subseteq [n-1]$. Then, the above implies that

$$\mathop{\mathbb{E}}_Z\left[(\sigma(-2\partial_n p(Z)) - \sigma(-2q_n(Z))^2\right] \le \gamma. \tag{VII.2}$$

Now, an argument similar to that of Lemma VII.6 shows that Z is δ-unbiased for $\delta = e^{-2\lambda}/2$. Therefore, by Equation VII.2, and Lemma VI.4, for $\gamma < c\exp(-4\lambda) \cdot \delta^{-t}$ for a sufficiently small constant c, we get

$$\|\partial_n p - q_n\|_1 \le O(1)(2t)^t \cdot e^{2\lambda} \cdot \sqrt{\gamma/\delta^t} \cdot \binom{n}{t-1} \le \varepsilon \cdot \binom{n}{t-1},$$

where $\gamma = \varepsilon^2 \cdot \exp(-C\lambda t)/C(2t)^{2t}$ for a sufficiently large constant $C > 0$. Note that by Theorem III.1, the number of samples needed to satisfy Equation VII.2 is

$$O((\lambda/\gamma)^2 \cdot (\log(n/\rho\gamma)) = \frac{(2t)^{O(t)} \cdot e^{O(\lambda t)}}{\varepsilon^4} \cdot (\log(n/\rho\varepsilon)).$$

This proves Equation VII.1 and hence the main part of the theorem. The moreover part of the statement follows from an argument nearly identical to that of Theorem VII.2 and is omitted here. ∎

VIII. Nonbinary Case

Due to space considerations we defer this to the full version (or see the version on arxiv.org).

REFERENCES

[1] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[2] S. L. Lauritzen, *Graphical Models*. Oxford University Press, 1998.

[3] E. Mossel, S. Roch, and A. Sly, "Robust estimation of latent tree graphical models: Inferring hidden states with inexact parameters," *IEEE Trans. Information Theory*, vol. 59, no. 7, pp. 4357–4373, 2013. [Online]. Available: http://dx.doi.org/10.1109/TIT.2013.2251927

[4] G. Hinton and T. Sejnowski, "Learning and relearning in boltzmann machines," in *Parallel Distributed Processing*, Rummelhart and McClelland, Eds., 1986, pp. 283–335.

[5] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Information Theory*, vol. 47, no. 2, pp. 498–519, 2001. [Online]. Available: http://dx.doi.org/10.1109/18.910572

[6] R. Salakhutdinov, "Learning in markov random fields using tempered transitions," in *NIPS*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc, 2009, pp. 1598–1606. [Online]. Available: http://papers.nips.cc/book/advances-in-neural-information-processing-systems-22-2009

[7] P. Clifford, "Markov random fields in statistics," in *Disorder in Physical Systems. A Volume in Honour of John M. Hammersley*, G. R. Grimmett and D. J. A. Welsh, Eds. Oxford: Clarendon Press, 1990, pp. 19–32.

[8] A. Jaimovich, G. Elidan, H. Margalit, and N. Friedman, "Towards an integrated protein-protein interaction network: A relational markov network approach," *Journal of Computational Biology*, vol. 13, no. 2, pp. 145–164, 2006. [Online]. Available: http://dx.doi.org/10.1089/cmb.2006.13.145

[9] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty, "High-dimensional graphical model selection using ℓ_1-regularized logistic regression," in *NIPS*, B. Schölkopf, J. C. Platt, and T. Hofmann, Eds. MIT Press, 2006, pp. 1465–1472. [Online]. Available: http://papers.nips.cc/book/advances-in-neural-information-processing-systems-19-2006

[10] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, no. 4, Mar. 12 2010, http://dx.doi.org/10.1214/12-STS400. [Online]. Available: http://arxiv.org/abs/1010.2731

[11] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of markov random fields from samples: Some observations and algorithms," *SIAM J. Comput*, vol. 42, no. 2, pp. 563–578, 2013. [Online]. Available: http://dx.doi.org/10.1137/100796029

[12] A. Ray, S. Sanghavi, and S. Shakkottai, "Greedy learning of graphical models with small girth," in *Allerton*. IEEE, 2012, pp. 2024–2031. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6475439

[13] G. Bresler, "Efficiently learning ising models on arbitrary graphs," in *STOC*. ACM, 2015, pp. 771–782. [Online]. Available: http://dl.acm.org/citation.cfm?id=2746539

[14] M. Vuffray, S. Misra, A. Y. Lokhov, and M. Chertkov, "Interaction screening: Efficient and sample-optimal learning of ising models," in *NIPS*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 2595–2603. [Online]. Available: http://papers.nips.cc/book/advances-in-neural-information-processing-systems-29-2016

[15] G. Bresler, D. Gamarnik, and D. Shah, "Structure learning of antiferromagnetic ising models," in *NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 2852–2860. [Online]. Available: http://papers.nips.cc/book/advances-in-neural-information-processing-systems-27-2014

[16] L. G. Valiant, "Functionality in neural nets," in *Proceedings of the First Annual Workshop on Computational Learning Theory*, ser. COLT '88. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988, pp. 28–39. [Online]. Available: http://dl.acm.org/citation.cfm?id=93025.93038

[17] G. Valiant, "Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem," *J. ACM*, vol. 62, no. 2, pp. 13:1–13:45, May 2015. [Online]. Available: http://doi.acm.org/10.1145/2728167

[18] L. Hamilton, F. Koehler, and A. Moitra, "Information theoretic properties of markov random fields, and their algorithmic applications," 2017, https://arxiv.org/pdf/1705.11107.pdf.

[19] A. T. Kalai and R. Sastry, "The isotron algorithm: High-dimensional isotonic regression," in *COLT*, 2009.

[20] S. M. Kakade, A. Kalai, V. Kanade, and O. Shamir, "Efficient learning of generalized linear and single index models with isotonic regression," in *NIPS*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 927–935. [Online]. Available: http://papers.nips.cc/book/advances-in-neural-information-processing-systems-24-2011

[21] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *JCSS: Journal of Computer and System Sciences*, vol. 55, 1997.

[22] V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami, "On agnostic learning of parities, monomials, and halfspaces," *SIAM J. Comput.*, vol. 39, no. 2, pp. 606–645, 2009.

[23] V. Guruswami and P. Raghavendra, "Hardness of learning halfspaces with noise," *SIAM J. Comput.*, vol. 39, no. 2, pp. 742–765, 2009.

[24] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. Information Theory*, vol. 58, no. 7, pp. 4117–4134, 2012. [Online]. Available: http://dx.doi.org/10.1109/TIT.2012.2191659