Goodness-of-fit Testing and Structure Learning of the Ising Model

Fred Zhang*

December 17, 2018

Abstract

In this note, we consider two algorithmic problems concerning the Ising model, *structure learning* and *testing goodness-of-fit*, and survey some recent developments.

- In the structure learning problem, we are interested in recovering its underlying graph structure given i.i.d. samples from a Ising model. A recent work by Klivans & Meka (2017) provides a nearly optimal algorithm for this problem in terms of both sample complexity and computational efficiency. The algorithm also extends to t-wise Markov random fields.
- The related problem of testing goodness-of-fit is that of deciding whether a given set of samples from an unknown Ising model is drawn from a specific hypothesis model. A recent work by Daskalakis et al. (2018) provides the first sample and time efficient testing algorithm.

1 Introduction

The Ising model is a fundamental model of ferromagnetism in statistical physics. The model consists of n discrete random variables over $\{-1, 1\}$ that represent magnetic spins. It‘ specifies a joint distribution over the variables, defined in terms of the external fields $h \in \mathbb{R}^n$ and a graph $G = (V, E)$ with edge weights A_{ij}, where the vertices represent the n variables. The distribution is then given by

$$\Pr(Z = z) \propto \exp\left(\sum_{i \neq j \in V} A_{ij} z_i z_j + \sum_i h_i z_i\right). \tag{1.1}$$

The Markov random fields (MRFs), also known as undirected graphical model, generalizes the Ising model, which allows t-wise interactions among the variables. For binary MRFs, it is specified through a graph $G = (V, E)$ on n vertices corresponding to the variables in $\{-1, 1\}^n$ such that each variable conditioned on its neighbors is independent of the remaining variables. The Hammersley-Clifford characterization of t-wise Markov random fields allows one to directly see that the model is a generalization of the Ising model:

$$\Pr(Z = z) \propto \exp\left(\sum_{I \in C_t(G)} \psi_I(z)\right), \tag{1.2}$$

where $C_t(G)$ denotes all cliques of size at most t in G, and ψ_I is a function that depends on the variables in I.

*John A. Paulson School of Engineering and Applied Sciences, Harvard University. Email: hzhang@g.harvard.edu

Structure Learning. A fundamental problem in statistics and machine leanring is that of recovering the underlying graph strucuture given samples from an undirected graphical model. The first provable algorithm for the problem considers the case when the graph is a tree, due to Chow & Liu (2006), and its analysis was later given by Chow & Wagner (1973). Subsequent works focus on generalizations of trees (Anandkumar et al. (2012), Chechetka & Guestrin (2008)) and graphs under certain restrictions (Ray et al. (2012), Netrapalli et al. (2010)). The recent work by Klivans & Meka (2017) provides the first algorithm for learning the graph structure of binary t-wise MRFs with nearly optimal sample and time complexity $n^{O(t)}$.

For simplicity, we focus on the degree-bounded zero-field Ising model of n variables in this article. For this special case, Klivans & Meka (2017) shows that one can recover each parameter A_{ij} up to an ϵ factor using roughly $O(\exp(d))$ samples in time $O\left(n^2\right)$ per sample. This achieves nearly optimal sample complexity, improving upon the doubly exponential bound given by Bresler (2015) and $\widetilde{O}\left(n^4\right)$ run time of Wainwright et al. (2007). The algorithm follows a simple multiplicative weight update paradigm. Its analysis hinges upon an interesting connection with the Hedge algorithm for online learning from the seminal work due to Freund & Schapire (1997).

Goodness-of-fit Testing. Consider the problem of goodness-of-fit testing for the Ising model. Suppose we have sample access to an unknown Ising model p (with unknown parameters). For some fixed Ising model q, can we distinguish between $p = q$ and $d_{\mathrm{TV}}(p, q) > \epsilon$ with high probability using a small number of samples? The problem was addressed by a recent work of Daskalakis et al. (2018). In particular, they show that it can be solved in polynomial time and sample complexity. In particular, $O\left(\frac{n^4\beta^2}{\epsilon^2}\right)$ samples suffice in the zero-field case, where $\beta = \max |A_{ij}|$.

A key challenge in obtaining this result is to bypass the barrier in the testing via learning approach. It is a fairly common technique in the distribution testing that one first computes an estimate of the model $\widehat{p}$ such that p and $\widehat{p}$ are close and compares the estimate with the hypothesis q. However, as we discussed earlier, fully recovering p requires expoentially many samples.

To bypass the sample complexity barrier in structural estimation, the testing algorithm of Daskalakis et al. (2018) leverages a localization argument—if p and q are sufficiently large, we can always blame on a node or an edge. Concretely, there must exist a pair of nodes u, v with significantly different covariance under p and q. It is easy to argue that the empirical covariance using polynomial samples suffices to give an accurate estimate. Thus, the test scheme simply computes the empirical covariances of all pairs from samples of p and compares them with the their expectations under q.

The general purpose testing algorithm via localization provides a polynomial baseline performance. This can be improved in various restricted regimes. See Table 1 of Daskalakis et al. (2018) for a summary. The authors also show a lower bound on the sample complexity that is linear in β, via the Le Cam's two-point method (LeCam et al. (1973)). The construction is fairly simple, involving merely one or two nodes. Thus, the correct dependence on n is still unkown.

2 Structure Learning

In this section we focus on the main result in Klivans & Meka (2017).

Theorem 2.1 (Theorem 5.2, Klivans & Meka (2017))**.** *Let $\mathcal{D}(A)$ be an n-variable zero-field Ising model with degree bounded by λ. There exists an algorithm that given λ, $\epsilon, \rho \in (0, 1)$, and $N = O(\lambda^2 \exp(O(\lambda))/\epsilon^4) \cdot (\log(n/\rho\epsilon))$ independent samples $Z^1, \ldots, Z^N \leftarrow \mathcal{D}(A)$ produces $\widehat{A}$ such that with probability at least $1 - \rho$,*

$$\left\|A - \widehat{A}\right\|_\infty \leq \epsilon.$$

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

The sample complexity nearly matches the information-theoretic lower bound due to Santhanam & Wainwright (2012). Further, the authors show that the quadratic runtime per sample may as well be optimal, by providing a reduction from the learning sparse parity with noise problem, a hard problem in computational learning theory.

The starting observation of the algorithm is a reduction to the problem of learning sparse generalized linear model. For a particular variable Z_i, we consider its conditional expectation fixing all other variables. On a high level, we view these $n-1$ conditioned variables as features and the value of Z_i as label. This enables us to reduce our unsupervised problem to a supervised one. It is natural to ask how the features and label are related so that a prediction model is possible. This is precisely given by the definition of the zero-field Ising model.

$$\Pr(Z = z) \propto \exp\left(\sum_{i \neq j \in V} A_{ij} z_i z_j\right). \tag{2.1}$$

Let $\sigma(z) = 1/(1-e^{-z})$ be the sigmoid function and Z_{-i} denote all variables except Z_i. It is easy to calculate that the conditional distribution of Z_i given $Z_{-i} = x$ is

$$\Pr\left(Z_i = -1 \mid Z_{-i} = x\right) = \sigma(w \cdot x),$$
$$\Pr\left(Z_i = 1 \mid Z_{-i} = x\right) = \sigma(-w \cdot x),$$

where $w_i = -2A_{ij}$. Now we can calculate the expected label Z_i given the features Z_{-1}. For simplicity, let $Y_i = (1 - Z_i)/2$, transforming -1 to 1 and 1 to 0. Thus,

$$\mathbb{E}\left[Y_i \mid Z_{-i} = x\right] = \sigma(w \cdot x). \tag{2.2}$$

This observation turns the problem into a supervised one. We are given a labeled dataset, where the label Y_i satisfies the property above, given features $Z_{-i} = x$. Can we recover the coefficients w from data? We assume that $\|w\|_1 \leq \lambda$, since the maximum degree of G is at most λ. This is precisely the problem called learning sparse generalized linear model.

Theorem 2.2 (Theorem 3.1, Klivans & Meka (2017)). *Let $\mathcal{D}$ be a distribution on $[-1,1]^n \times \{0,1\}$ where for $(X,Y) \sim \mathcal{D}$,*

$$\mathbb{E}[Y|X = x] = \sigma(w \cdot x). \tag{2.3}$$

Suppose that $\|w\|_1 \leq \lambda$ for a known $\lambda \geq 0$. Then, there exists an algorithm that for all $\epsilon, \delta \in [0,1]$ given $T = O\left(\lambda^2(\ln(n/\delta\epsilon))/\epsilon^2\right)$ independent examples from $\mathcal{D}$, produces a vector $v \in \mathbb{R}^n$ such that with probability at least $1-\delta$,

$$\mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}\left[\left(\sigma(v \cdot X) - \sigma(w \cdot X)\right)^2\right] \leq \epsilon. \tag{2.4}$$

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

This is the first time and sample efficient learning algorithm for sparse generalized linear model. For learning Ising model, we would simply run this algorithm viewing each variable as label Y and others as X, one at a time. The formal description will come up later.

To prove this theorem, first observe that we may assume without loss of generality that $w_i \geq 0$ for all i and that $\|w\|_1 = \lambda$. Otherwise, we can always transform the samples to $\left((x, -x, 0), y\right)$. Now the non-negativity assumption allows one to interpret the condition Equation 2.3 as a weighted voting procedure. More concretely, since the expected value of the label is monotonically increasing in every feature (*i.e.*, every coordinate of x), we view the positive features as votes for Y being close to 1 and the negative ones as votes for Y being close to 0. The expected value of Y is determined by a positive linear combination of the votes.

This leads to an connection with the expert advice problem in online learning. In this problem, a learner is assisted by n experts $\mathcal{A}$ in making a binary decision. In the tth round, each expert i offer their

one bit advice $a_{i,t}$, and the agent receives a binary feedback x_t and a loss $\ell_t(x_t, a_t)$ upon making his own binary decision a_t. The goal is to be competitive with the best expert over time. Formally, this means minimizing the regret.

$$\text{Regret}_T = \sum_{t=1}^{T} \ell_t(x_t, a_t) - \min_{i \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(x_t, a_{i,t}) \tag{2.5}$$

The celebrated work of Freund & Schapire (1997) gives a simple multiplicative weights update algorithm that achieves $O\left(\sqrt{T \ln n} + \ln n\right)$ regret bound.

HEDGE (x):
  Set $\boldsymbol{p}_1 = \mathbf{1}/n$
  For $t = 1, 2, \cdots, T$
    Pick expert i with probability proportional to $\boldsymbol{p}_t(i)$.
    Incur loos $\boldsymbol{\ell_t}(i) = \ell_t(x_t, a_{i,t})$ for each expert.
    Update weights $\boldsymbol{p}_{t+1}(i) = \boldsymbol{p}_t(i) \cdot \beta^{\boldsymbol{\ell_t}(i)}$.
    Normalize $\boldsymbol{p}_{t+1}$.

The parameter $\beta \in (0, 1)$ is called learning rate and set to be $1/\left(1 + \sqrt{(\ln n)/T}\right)$.

Back to our problem of sparse generalized linear model (Equation 2.3), we interpret each coordinate as an expert, and the loss function is defined as

$$\ell_i^t = \frac{1}{2}\left(1 + (\sigma\left(w^t \cdot x^t\right) - y^t)x_i^t\right). \tag{2.6}$$

Note that $\ell_i^t \in [0, 1]$. Here, the term $p^t = \sigma\left(w^t \cdot x^t\right)$ is our prediction of the label based on current weights. Observe that if we over-predicted the label ($p^t > y^t$), the positive coordinates of x^t would suffer higher loss. On the other hand, if we under-predicted the label, the negative ones would suffer higher loss. Our algorithm penalizes the weights that incur high loss each round.

Now we are ready to state and analyze our algorithm formally.

Proof of Theorem 2.2. Assume $w_i \geq 0$ for all i and that $\|w\|_1 = \lambda$. We propose the following multiplicative weights update algorithm. The algorithm takes $T + M$ independent samples, where $M = O\left(\ln\left(T/\delta\right)/\epsilon^2\right)$ samples are used in the final testing step.

SPARSITRON($\{\boldsymbol{x}^i, y^i\}, \{\boldsymbol{a}^i, b^i\}$):
  Set $\boldsymbol{w}_1 = \boldsymbol{\lambda}/n$
  For $t = 1, 2, \cdots, T$
    Compute prediction $p^t = \sigma\left(\boldsymbol{w}^t \cdot \boldsymbol{x}^t\right)$.
    Incur loos $\boldsymbol{\ell}_i^t = \frac{1}{2}\left(1 + \left(p^t - y^t\right)\boldsymbol{x}_i^t\right)$ for each coordinate.
    Update weights $\boldsymbol{w}_i^{t+1} = \boldsymbol{w}_i^t \cdot \beta^{\boldsymbol{\ell}_i^t}$.
    Normalize $\boldsymbol{w}_{t+1}$ to have ℓ_1 norm λ.
  For $t = 1, 2, \cdots, T$
    Compute the empirical risk
      $\widehat{\epsilon}(\boldsymbol{w}^t) = (1/M)\sum_{j=1}^{M}\left(\sigma(\boldsymbol{w}^t \cdot \boldsymbol{a}^j) - b^j\right)^2$.
    Return a minimizer of the empirical risk.

The proof works in two steps. We first show that the average of the true risks $\epsilon(\boldsymbol{w}^1), \cdots, \epsilon(\boldsymbol{w}^T)$ is small with high probability, where

$$\epsilon(\boldsymbol{w}) = \mathop{\mathbb{E}}_{\boldsymbol{x},y}\left(\sigma\left(\boldsymbol{w} \cdot \boldsymbol{x}\right) - y\right)^2.$$

Next we show that the empirical risks $\widehat{\epsilon}$ are close to the true risks when M is sufficiently large. This would imply that our empirical risk minimizer nearly minimizes the true risk.

Using the fact that $y^t \in \{0,1\}$ and $\left(\sigma(a)-\sigma(b)\right)^2 \le (a-b)\left(\sigma(a)-\sigma(b)\right)$ for any $a, b \in \mathbb{R}$, and conditioned on $\left(\boldsymbol{x}^1, y^1\right), \cdots, \left(\boldsymbol{x}^t, y^t\right)$

$$\begin{aligned}
\epsilon\left(\boldsymbol{w}^t\right) &= \mathop{\mathbb{E}}_{\boldsymbol{x}^t}\left(\sigma\left(\boldsymbol{w}^t \cdot \boldsymbol{x}^t\right) - \sigma\left(\boldsymbol{w} \cdot \boldsymbol{x}^t\right)\right)^2 \\
&\le \mathop{\mathbb{E}}_{\boldsymbol{x}^t}\left(\boldsymbol{w}^t \cdot \boldsymbol{x}^t - \boldsymbol{w} \cdot \boldsymbol{x}^t\right)\left(\sigma\left(\boldsymbol{w}^t \cdot \boldsymbol{x}^t\right) - \sigma\left(\boldsymbol{w} \cdot \boldsymbol{x}^t\right)\right) \\
&= 2\lambda \mathop{\mathbb{E}}_{\left(\boldsymbol{x}^t, y^t\right)}\left[\left((1/\lambda)\boldsymbol{w}^t - (1/\lambda)\boldsymbol{w}\right) \cdot \boldsymbol{\ell}^t\right] \\
&= 2\lambda \mathop{\mathbb{E}}_{\left(\boldsymbol{x}^t, y^t\right)}\left[Q^t \,\middle|\, \left(\boldsymbol{x}^1, y^1\right), \cdots, \left(\boldsymbol{x}^{t-1}, y^{t-1}\right)\right], \qquad (2.7)
\end{aligned}$$

where $Q^t = \left((1/\lambda)\boldsymbol{w}^t - (1/\lambda)\boldsymbol{w}\right) \cdot \boldsymbol{\ell}^t$. Let

$$Z^t = Q^t - \mathop{\mathbb{E}}_{\left(\boldsymbol{x}^t, y^t\right)}\left[Q^t \,\middle|\, \left(\boldsymbol{x}^1, y^1\right), \cdots, \left(\boldsymbol{x}^{t-1}, y^{t-1}\right)\right]$$

Since Z^t forms a martingale difference sequence, by Azuma-Hoeffding inequality, with probability $1-\delta$

$$\left|\sum_{t=1}^{T} Z^T\right| \le O\left(\sqrt{T\ln(1/\delta)}\right).$$

Thus, with probability $1-\delta$

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\left(\boldsymbol{x}^t, y^t\right)}\left[Q^t \,\middle|\, \left(\boldsymbol{x}^1, y^1\right), \cdots, \left(\boldsymbol{x}^{t-1}, y^{t-1}\right)\right] \le \sum_{t=1}^{T} Q^t + O\left(\sqrt{T\ln(1/\delta)}\right). \qquad (2.8)$$

By Equation 2.8 and Equation 2.7, with probability $1-\delta$

$$\left(1/2\lambda\right)\sum_{t=1}^{T}\epsilon(\boldsymbol{w}^t) \le \sum_{t=1}^{T} Q^t + O\left(\sqrt{T\ln\left(1/\delta\right)}\right)$$

Now it remains to bound Q^t. Applying Theorem 5 of Freund & Schapire (1997), we immediately get that

$$\sum_{t=1}^{T}\left(\boldsymbol{w}^t/\lambda\right)\cdot\boldsymbol{\ell}^t \le \min_{i\in[n]}\sum_{t=1}^{T}\boldsymbol{\ell}_i^t + O\left(\sqrt{T\ln n} + \ln n\right). \qquad (2.9)$$

Thus, by the definition of Q^t, and since $\|w\| \le \lambda$

$$\sum_{t=1}^{T} Q^t \le \min_{i\in[n]}\sum_{t=1}^{T}\boldsymbol{\ell}_i^t - \sum_{t=1}^{T}\left(\boldsymbol{w}/\lambda\right)\boldsymbol{\ell}^t + O\left(\sqrt{T\ln n} + \ln n\right) \le O\left(\sqrt{T\ln n} + \ln n\right)$$

Hence, with probability $1-\delta$

$$\left(1/2\lambda\right)\sum_{t=1}^{T}\epsilon(\boldsymbol{w}^t) = O\left(\sqrt{T\ln(1/\delta)}\right) + O\left(\sqrt{T\ln n} + (\ln n)\right).$$

Now let $T = \Omega(\lambda^2\ln(n/\delta)/\epsilon^2)$, with probability $1-\delta$,

$$\min_{t\in[T]}\epsilon(\boldsymbol{w}^t) \le O\left(\lambda\right)\cdot\frac{\sqrt{T\ln(1/\delta)} + \sqrt{T\ln n} + \ln n}{T} \le \epsilon/2.$$

Finally, for $M = C\ln(T/\delta)/\epsilon^2$, Chernoff bound implies $\left|\epsilon\left(\boldsymbol{w}^t\right) - \widehat{\epsilon}\left(\boldsymbol{w}^t\right)\right| \le \epsilon/4$ with probability $1-\delta$ for every t. □

We will not give a full proof of Theorem 2.1. The algorithm for recovering the Ising model should be clear. We would run the SPARSITRON algorithm for sparse generalized linear model on every variable Z_i, treating everything else as features. One structural insight about the Ising model is required to establish the correctness of the algorithm. We say that a distribution $\mathcal{D}$ on $\{-1,1\}^n$ is δ-*unbiased* if conditioned on other variables, each variable Z_i takes value -1 or 1 with probability at least δ. Klivans & Meka (2017) shows that unbiasedness implies that the guarantee of Equation 2.4 yields an ϵ additive recovery of each parameter. Moreover, the Ising model indeed satisfies this technical condition by Bresler (2015). These observations suffice to establish Theorem 2.1.

3 Goodness-of-fit Testing

We provide a sketch of the general purpose localization testing scheme of Daskalakis et al. (2018). For simplicity, we focus on the zero-field case. The algorithm achieves polynomial sample complexity. Let $d_{\mathrm{SKL}}(p,q) = \mathrm{KL}(p||q) + \mathrm{KL}(q||p)$ denote the symmetric KL divergence.

Theorem 3.1. *Given $\tilde{O}\left(\frac{n^4\beta^2}{\epsilon^2}\right)$ samples from an Ising model p and a description of an Ising model q, there exists a polynomial-time algorithm which distinguishes between the cases $p = q$ and $d_{SKL}(p,q) \geq \epsilon$ with probability at least $2/3$ where $\beta = \max\{|A_{ij}|\}$.*

Before proving this theorem, we need a technical lemma.

Lemma 3.2 (Santhanam & Wainwright (2012))**.** *Let p and q be two zero-field Ising models with paratmers A^p and A^q. Let $\mu^p_{ij} = \mathbb{E}_p\, Z_i Z_j$. Then the symmetric KL divergence between p and q is given by*

$$d_{SKL}(p,q) = \sum_{i \neq j} \left(A^p_{ij} - A^q_{ij}\right)\left(\mu^p_{ij} - \mu^q_{ij}\right). \tag{3.1}$$

This immediately implies a useful characterization of the two Ising models p and q being far from each other.

Lemma 3.3 (Lemma 4 of Daskalakis et al. (2018))**.** *Given two Ising models p and q, if $d_{\mathrm{SKL}}(p,q) \geq \epsilon$, then there exists $e = (i,j)$ such that*

$$\left(A^p_{ij} - A^q_{ij}\right)\left(\mu^p_{ij} - \mu^q_{ij}\right) \geq \epsilon/m, \tag{3.2}$$

where $m = \binom{n}{2}$.

Now we are ready to prove Theorem 3.1. On a high level, the algorithm is simple. We take enough samples to computes accurate estimates $\widehat{\mu}^p_{uv}$ of μ_{uv} for all u, v under p. Then if

$$\left|\widehat{\mu}^p_{uv} - \mu^q_{uv}\right| \gg \epsilon/m\beta$$

for some u, v, it certifies that Equation 3.2 is satisfied, and p, q are far. On the other hand, if the condition above does not hold for all pairs u, v, it means that p, q are close.

LOCALIZATIONTEST(p):
Draw $k = O\left(\frac{n^4\beta^2 \log n}{\epsilon^2}\right)$ samples $\left\{x^{(1)}\right\}_{i=1}^{k}$ from p.
Compute empirical estimates $\widehat{\mu}^p_{uv} = \frac{1}{k}\sum_{i=1}^{k} x^{(i)}_u x^{(i)}_v$ of all pairs (u,v) .
If for any pair (u,v), $\left|\widehat{\mu}^p_{uv} - \mu^q_{uv}\right| \geq \frac{\epsilon}{8m\beta}$,
    return that $d_{\mathrm{SKL}}(p,q) \geq \epsilon$.
otherwise return that $p = q$.

The analyais of the algorithm is straighforward. First it follows from simple Chernoff-Hoeffding bound that the sample estimates $\widehat{\mu}_{uv}$ are close to the true covariances μ^p_{uv} with high probability. Applying traingle inequality and Lemma 3.3 immediately shows that the algorithm is correct.

4 Concluding Remarks

An intriguing problem is to generalize the approach of Klivans & Meka (2017) to learning *Gaussian graphical models*, where the variables are Gaussian (over the reals). The graph structure corresponds to the precision matrix of a multivariate Gaussian, and it turns out that any multivariate Gaussian can be modeled as a Gaussian graphical model. Thus, the problem is exactly that of recovering the precision matrix given i.i.d. multivariate Gaussian samples. See Fan et al. (2016) for a survey on this topic.

It is also known that hte HEDGE algorithm can be interpreted as a mirror descent scheme; see Chapter 5 of Hazan et al. (2016). It would be interesting to consider an interpretation of the SPARSITRON algorithm under a unified framework.

To the best of the my knowledge, it remains open whether bounded degree t-wise MRFs admit goodness-of-fit testing schemes with polynomial time and sample complexity, although such results are known for Bayesian networks, a class of directed graphical models (Canonne et al. (2017)).

References

Anandkumar, A., Tan, V. Y., Huang, F., Willsky, A. S. et al. (2012), 'High-dimensional structure estimation in ising models: Local separation criterion', *The Annals of Statistics* **40**(3), 1346–1375. 2

Bresler, G. (2015), Efficiently learning ising models on arbitrary graphs, *in* 'Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing (STOC)', pp. 771–782. 2, 6

Canonne, C. L., Diakonikolas, I., Kane, D. M. & Stewart, A. (2017), Testing bayesian networks, *in* 'Proceedings of the 2017 Conference on Learning Theory (COLT)', pp. 370–448. 7

Chechetka, A. & Guestrin, C. (2008), Efficient principled learning of thin junction trees, *in* 'Advances in Neural Information Processing Systems (NIPS)', pp. 273–280. 2

Chow, C. & Liu, C. (2006), 'Approximating discrete probability distributions with dependence trees', *IEEE Trans. Inf. Theor.* **14**(3), 462–467. 2

Chow, C. & Wagner, T. (1973), 'Consistency of an estimate of tree-dependent probability distributions (corresp.)', *IEEE Transactions on Information Theory* **19**(3), 369–371. 2

Daskalakis, C., Dikkala, N. & Kamath, G. (2018), Testing ising models, *in* 'Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)', pp. 1989–2007. 1, 2, 6

Fan, J., Liao, Y. & Liu, H. (2016), 'An overview of the estimation of large covariance and precision matrices', *The Econometrics Journal* **19**(1), C1–C32. 7

Freund, Y. & Schapire, R. E. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences* **55**(1), 119 – 139. 2, 4, 5

Hazan, E. et al. (2016), 'Introduction to online convex optimization', *Foundations and Trends® in Optimization* **2**(3-4), 157–325. 7

Klivans, A. & Meka, R. (2017), Learning graphical models using multiplicative weights, *in* 'Proceedings of IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)', pp. 343–354. 1, 2, 3, 6, 7

LeCam, L. et al. (1973), 'Convergence of estimates under dimensionality restrictions', *The Annals of Statistics* **1**(1), 38–53. 2

Netrapalli, P., Banerjee, S., Sanghavi, S. & Shakkottai, S. (2010), Greedy learning of markov network structure, *in* 'Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on', IEEE, pp. 1295–1302. 2

Ray, A., Sanghavi, S. & Shakkottai, S. (2012), Greedy learning of graphical models with small girth, *in* 'Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on', IEEE, pp. 2024–2031. 2

Santhanam, N. P. & Wainwright, M. J. (2012), 'Information-theoretic limits of selecting binary graphical models in high dimensions.', *IEEE Trans. Information Theory* **58**(7), 4117–4134. 3, 6

Wainwright, M. J., Lafferty, J. D. & Ravikumar, P. K. (2007), High-dimensional graphical model selection using $\ell 1$-regularized logistic regression, *in* 'Advances in Neural Information Processing Systems (NIPS)', pp. 1465–1472. 2