Structure learning of antiferromagnetic Ising models

Authored by:

Devavrat Shah
Guy Bresler
David Gamarnik

Abstract

In this paper we investigate the computational complexity of learning the graph structure underlying a discrete undirected graphical model from i.i.d. samples. Our first result is an unconditional computational lower bound of $Omega (p{d/2})$ for learning general graphical models on p nodes of maximum degree d, for the class of statistical algorithms recently introduced by Feldman et al. The construction is related to the notoriously difficult learning parities with noise problem in computational learning theory. Our lower bound shows that the $widetilde O(p{d+2})$ runtime required by Bresler, Mossel, and Sly's exhaustive-search algorithm cannot be significantly improved without restricting the class of models. Aside from structural assumptions on the graph such as it being a tree, hypertree, tree-like, etc., most recent papers on structure learning assume that the model has the correlation decay property. Indeed, focusing on ferromagnetic Ising models, Bento and Montanari showed that all known low-complexity algorithms fail to learn simple graphs when the interaction strength exceeds a number related to the correlation decay threshold. Our second set of results gives a class of repelling (antiferromagnetic) models that have the emph{opposite} behavior: very strong repelling allows efficient learning in time $widetilde O(p2)$. We provide an algorithm whose performance interpolates between $widetilde O(p2)$ and $widetilde O(p{d+2})$ depending on the strength of the repulsion.

1 Paper Body

Graphical models have had tremendous impact in a variety of application domains. For unstructured high-dimensional distributions, such as in social networks, biology, and finance, an important first step is to determine which graphical model to use. In this paper we focus on the problem of structure learning: Given access to n independent and identically distributed samples ? (1) , . . . ? (n) from an undirected graphical model representing a discrete random vector

? = (?1 , . . . , ?p), the goal is to find the graph G underlying the model. Two basic questions are 1) How many samples are required? and 2) What is the computational complexity? In this paper we are mostly interested in the computational complexity of structure learning. We first consider the problem of learning a general discrete undirected graphical model of bounded degree. 1

1.1

Learning general graphical models

Several algorithms based on exhaustively searching over possible node neighborhoods have appeared in the last decade [4, 2, 5]. Abbeel, Koller, and Ng [4] gave algorithms for learning general graphical models close to the true distribution in Kullback-Leibler distance. Bresler, Mossel, and Sly [2] presented algorithms guaranteed to learn the true underlying graph. The algorithms in both [4] and [2] perform a search over candidate neighborhoods, and for a graph of maximum degree d, the computational complexity for recovering a graph on p ? d+2) (where the O ? notation hides logarithmic factors). nodes scales as O(p While the algorithms in [2] are guaranteed to reconstruct general models under basic nondegeneracy conditions using an optimal number of samples n = O(d log p) (sample complexity lower bounds were proved by Santhanam and Wainwright [6] as well as [2]), the ? d+2) run-time is impractically high even for constant but large graph exponent d in the O(p degrees. This has motivated a great deal of work on structure learning for special classes of graphical models. But before giving up on general models, we ask the following question: Question 1: Is it possible to learn the structure of general graphical models on p nodes with maximum degree d using substantially less computation than pd ? Our first result suggests that the answer to Question 1 is negative. We show an uncond ditional computational lower bound of p 2 for the class of statistical algorithms introduced by Feldman et al. [1]. This class of algorithms was introduced in order to understand the apparent difficulty of the Planted Clique problem, and is based on Kearns? statistical query model [7]. Kearns showed in his landmark paper that statistical query algorithms require exponential computation to learn parity functions subject to classification noise, and our hardness construction is related to this problem. Most known algorithmic approaches (including Markov chain Monte Carlo, semidefinite programming, and many others) can be implemented as statistical algorithms, so the lower bound is fairly convincing. We give background and prove the following theorem in Section 4. Theorem 1.1. Statistical algorithms require at least (p 2) computation steps in order to learn the structure of a general graphical models of degree d. d

If complexity pd is to be considered intractable, what shall we consider as tractable? Writing algorithm complexity in the form c(d)pf (d) , for high-dimensional (large p) problems the exponent f (d) is of primary importance, and we will think of tractable algorithms as having an f (d) that is bounded by a constant independent of d. The factor c(d) is also important, and we will use it to compare algorithms with the same exponent f (d). In light of Theorem 1.1, reducing computation below p (d) requires restricting the class of models. One can either restrict the graph structure or the nature of the interactions between variables. The seminal paper of Chow and Liu [8] makes a model restriction

of the first type, assuming that the graph is a tree; generalizations include to polytrees [9], hypertrees [10], and others. Among the many possible assumptions of the second type, the correlation decay property is distinguished: to the best of our knowledge all existing low-complexity algorithms require the correlation decay property [3]. 1.2

Correlation decay property

Informally, a graphical model is said to have the correlation decay property (CDP) if any two variables ?s and ?t are asymptotically independent as the graph distance between s and t increases. Exponential decay of correlations holds when the distance from independence decreases exponentially fast in graph distance, and we will mean this stronger form when referring to correlation decay. Correlation decay is known to hold for a number of pairwise graphical models in the so-called high-temperature regime, including Ising, hard-core lattice gas, Potts (multinomial) model, and others (see, e.g., [11, 12, 13, 14, 15, 16]). 2

Bresler, Mossel, and Sly [2] observed that it is possible to efficiently learn models with (exponential) decay of correlations, under the additional assumption that neighboring variables have correlation bounded away from zero (as is true, e.g., for the ferromagnetic Ising model in the high temperature regime). The algorithm they proposed for this setting pruned the candidate set of neighbors for each node to roughly size O(d) by retaining only those variables with sufficiently high correlations, and then within this set performed the exhaustive search ? 2). The over neighborhoods mentioned before, resulting in a computational cost of dO(d) O(p greedy algorithms of Netrapali et al. [17] and Ray et al. [18] also require the correlation decay property and perform a similar pruning step by retaining only nodes with high pairwise correlation; they then use a different method to select the true neighborhood. A number of papers consider the problem of reconstructing Ising models on graphs with few short cycles, beginning with Anandkumar et al. [19]. Their results apply to the case of Ising models on sparsely connected graphs such as the Erd?os-Renyi random graph G(p, dp). They additionally require the interaction parameters to be either generic or ferromagnetic. Ferromagnetic models have the benefit that neighbors always have a non-negligible correlation because the dependencies cannot cancel, but in either case the results still require the CDP to hold. Wu et al. [20] remove the assumption of generic parameters in [19], but again require the CDP. Other algorithms for structure learning are based on convex optimization, such as Ravikumar et al.?s [21] approach using regularized node-wise logistic regression. While this algorithm does not explicitly require the CDP, Bento and Montanari [3] found that the logistic regression algorithm of [21] provably fails to learn certain ferromagnetic Ising model on simple graphs without correlation decay. Other convex optimization-based algorithms such as [22, 23, 24] require similar incoherence or restricted isometry-type conditions that are difficult to verify, but likely also require correlation decay. Since all known algorithms for structure learning require the CDP, we ask the following question (paraphrasing Bento and Montanari): Question 2: Is low-complexity structure learning possible for models which do not exhibit the CDP, on general bounded degree

graphs? Our second main result answers this question affirmatively by showing that a broad class of repelling models on general graphs can be learned using simple algorithms, even when the underlying model does not exhibit the CDP. 1.3

Repelling models

The antiferromagnetic Ising model has a negative interaction parameter, whereby neighboring nodes prefer to be in opposite states. Other popular antiferromagnetic models include the Potts or coloring model, and the hard-core model. Antiferromagnetic models have the interesting property that correlations between neighbors can be zero due to cancellations. Thus algorithms based on pruning neighborhoods using pairwise correlations, such as the algorithm in [2] for models with correlation decay, does not work for anti-ferromagnetic models. To our knowledge there are no previous results that improve on the pd computational complexity for structure learning of antiferromagnetic models on general graphs of maximum degree d. Our first learning algorithm, described in Section 2, is for the hard-core model. Theorem 1.2 (Informal). It is possible to learn strongly repelling models, such as the hard? 2). core model, with run-time O(p

We extend this result to weakly repelling models (equivalent to the antiferromagnetic Ising model parameterized in a nonstandard way, see Section 3). Here ? is a repelling strength and h is an external field. Theorem 1.3 (Informal). Suppose ? ? (d ? ?)(h + ln 2) for an integer 0 ? ? ¡ d. Then ? 2+?). it is possible to learn a repelling model with interaction ?, with run-time O(p 3

? 2), achievable for The computational complexity of the algorithm interpolates between O(p d+2 ? strongly repelling models, and O(p), achievable for general models using exhaustive search. The complexity depends on the repelling strength of the model, rather than structural assumptions on the graph as in [19, 20]. We remark that the strongly repelling models exhibit long-range correlations, yet the algorithmic task of graph structure learning is possible using a local procedure. The focus of this paper is on structure learning, but the problem of parameter estimation is equally important. It turns out that the structure learning problem is strictly more challenging for the models we consider: once the graph is known, it is not difficult to estimate the parameters with low computational complexity (see, e.g., [4]).

2

Learning the graph of a hard-core model

We warm up by considering the hard-core model. The analysis in this section is straightforward, but serves as an example to highlight the fact that correlation decay is not a necessary condition for structure learning. Given a graph G = (V, E) on —V — = p nodes, denote by I(G) ? {0, 1}p the set of independent set indicator vectors ?, for which at least one of ?i or ?j is zero for each edge {i, j} ? E(G). The hardcore model with fugacity ? ¿ 0 assigns nonzero probability only to vectors in I(G), with ?—?— P(?) = , ? ? I(G) . (2.1) Z q —?— Here —?— is the number of entries of ? equal to one and Z = ??I(G) ? is the normalizing constant called the partition function. If ? ¿ 1 then more mass is assigned to larger independent sets. (We use indicator vectors to define the model in order to be consistent with the antiferromagnetic Ising model in the next section.)

Our goal is to learn the graph G = (V, E) underlying the model (2.1) given access to independent samples ? (1) , . . . , ? (n) . The following simple algorithm reconstructs G efficiently. Algorithm 1 simpleHC(? (1) , . . . , ? (n)) 1: FOR each i, j, k: (k) (k) 2: IF ?i = ?j = 1, THEN S = S ? {i, j} ? = Sc 3: OUTPUT E The idea behind the algorithm is very simple. If {i, j} belongs to the edge set E(G), then (k) (k) for every sample ? (k) either ?i = 0 or ?j = 0 (or both). Thus for every i, j and k such (k)

(k)

that ?i = ?j = 1 we can safely declare {i, j} not to be an edge. To show correctness of the algorithm it is therefore sufficient to argue that for every non-edge {i, j} there is a high likelihood that such an independent set ? (k) will be sampled.

Before doing this, we observe that simpleHC actually computes the maximum-likelihood (k) (k) estimate for the graph G. To see this, note that an edge e = {i, j} for which ?i = ?j = 1 ? since P(? (k) —G+e) ? ? Thus the ML estimate contains for some k cannot be in G, = 0 for any G. a subset of those edges e which have not been ruled out by ? (1) , . . . , ? (n) . But adding any such edge e to the graph decreases the value of the partition function in (2.1) (the sum is over fewer independent sets), thereby increasing the likelihood of each of the samples. The sample complexity and computational complexity of simpleHC is as follows, with proof in the Supplement. Theorem 2.1. Consider the hard-core model (2.1) on a graph G = (V, E) on —V — = p nodes and with maximum degree d. The sample complexity of simpleHC is n = O((2?)2d?2 log p) , 4

(2.2)

i.e. with this many samples the algorithm simpleHC correctly reconstructs the graph with probability 1 ? o(1). The computational complexity is O(np2) = O((2?)2d?2 p2 log p) .

(2.3)

We next show that the sample complexity bound in Theorem 2.1 is basically tight: Theorem 2.2 (Sample complexity lower bound). Consider the hard-core model (2.1). There is a family of graphs on p nodes with maximum degree d such that for the probability of successful reconstruction to approach one, the number of samples must scale as 1 p2 n= (2?)2d log . d Lemma 2.3. Suppose edge e = (i, j) ? / G, and let I be an independent set chosen according to the Gibbs distribution (2.1). Then P({i, j} ? I) ? (9 ? max{1, (2?)2d?2 })?1 , ? . The Supplementary Material contains proofs for Theorem 2.2 and Lemma 2.3.

3

Learning anti-ferromagnetic Ising models

In this section we consider the anti-ferromagnetic Ising model on a graph G = (V, E). We parametrize the model in such a way that each configuration has probability) * 1 P(?) = exp H(?) , ? ? {0, 1}p , (3.1) Z where ? ? H(?) = ?? ?i ?j + hi ? i . (3.2) (i,j)?E

i?V

Here ? ¿ 0 and {hi }i?V are real-valued parameters, and we assume that —hi — ? h for all i. Working with configurations in {0, 1}p rather than the more typical {?1, +1}p amounts to a reparametrization (which is without loss

of generality as shown for example in Appendix 1 of [25]). Setting hi = h = ln ? for all i, we recover the hard-core model with fugacity ? in the limit ? ? ?, so we think of (3.2) as a ?soft? independent set model. 3.1

Strongly antiferromagnetic models

We start by considering the situation in which the repelling strength ? is sufficiently large that we can modify the approach used for the hard-core model. We require some notation to work with conditional probabilities: for each vertex b ? V , let (i)

and

Bb = {? (i) : ?b = 1} ,

? a = 1—?b = 1) := 1 —{i ? B : ? (i) = 1}— . P(? a —B— ! " ? a = 1—?b = 1) = P(?a = 1—?b = 1). The algorithm, described next, Of course, E P(? ? to a threshold. determines whether each edge {a, b} is present based on comparing P Algorithm 2 StrongRepelling

? Input: ?, h, d, and n samples ? (1) , . . . , ? (n) ? {0, 1}p . Output: edge set E. d h(d?1) ?2 1: Let ? = (1 + 2 e) ! " 2: FOR each possible edge {a, b} ? V2 : ? 3: IF P? (?a = 1—?b = 1) ? (1 + e??h)?1 + ? THEN add edge (a, b) to E ? 4: OUTPUT E Algorithm StrongRepelling obtains the following performance. The proof of Proposition 3.1 is similar to that of Theorem 2.1, replacing Lemma 2.3 by Lemma 3.2 below. 5

Proposition 3.1. Consider the antiferromagnetic Ising model (3.2) on a graph G = (V, E) on p nodes and with maximum degree d. If ? ? d(h + ln 2) ,

then algorithm StrongRepelling has sample complexity 1 2 n = O 22d e2h(d+1) log p ,

i.e. this many samples are sufficient to reconstruct the graph with probability 1 ? o(1). The computational complexity of StrongRepelling is 1 2 O(np2) = O 22d e2h(d+1) p2 log p . When the interaction parameter ? ? d(h+ln 2) it is possible to identify edges using pairwise statistics. The next lemma, proved in the Supplement, shows the desired separation. Lemma 3.2. We have the following estimates: (i) If (a, b) ? / E(G), then P(?a = 1—?b = 1) ?

1 1+2deg(a) eh(deg(a)+1)

(ii) Conversely, if (a, b) ? E(G), then P(?a = 1—?b = 1) ? (ii) For any b ? V , P(?b = 1) ? 3.2

1 1+2deg(b) eh(deg(b)+1)

.

1 1+e??h

.

.

Weakly antiferromagnetic models

In this section we focus on learning weakly repelling models and show a trade-off between computational complexity and strength of the repulsion. Recall that for strongly repelling models our algorithm has run-time O(p2 log p), the same as for the hard-core model (infinite repulsion). For a subset of nodes U ? V , let GU denote the graph obtained from G by removing nodes in U (as well as any edges incident to nodes in U). The following corollary is immediate from Lemma 3.2. Corollary 3.3. We have the conditional probability estimates for

deleting subsets of nodes: (i) If (a, b) ? / E(G), then for any subset of nodes U ? V {a, b}, PGU (?a = 1—?b = 1) ?

1

1+2

degGU (a) h(degGU (a)+1)

e

.

(ii) Conversely, if (a, b) ? E(G), then for any subset of nodes U ? V {a, b} PGU (?a = 1—?b = 1) ?

1 . 1 + e??h

We can effectively remove nodes from the graph by conditioning: The family of models (3.2) has the property that conditioning on ?i = 0 amounts to removing node i from the graph. Fact 3.4 (Self-reducibility). Let G = (V, E), and consider the model 3.2. Then for any subset of nodes U ? V , the probability law PG (? ? ? —?U = 0) is equal to PGU (?V U ? ?). The final ingredient is to show that we can condition by restricting attention to a subset of the observed data, ? (1) , . . . , ? (n) , without throwing away too many samples. Lemma 3.5. Let U ? V be a subset of nodes and denote the subset of samples with variables (i) ?U equal to zero by AU = {? (i) : ?u = 0 for all u ? U }. Then with probability at least h 2—U — 1 ? exp(n/2(1 + e)) the number —AU — of such samples is at least n2 ? (1 + eh)?—U — .

We now present the algorithm. Effectively, it reduces node degree by removing nodes (which can be done by conditioning on value zero), and then applies the strong repelling algorithm to the residual graph. 6

Algorithm 3 WeakRepelling

? Input: ?, h, d, and n samples ? (1) , . . . , ? (n) ? {0, 1}p . Output: edge set E. 1: Let ? = (1 + 2d eh(d?1))?2 ! ” 2: FOR each possible edge (a, b) ? V2 : 3: FOR each U ? V {a, b} of size —U — ? ?d ? ?/(h + ln 2)? 4: Compute P?GU (?a = 1—?b = 1) ? 5: IF minU :—U —= P?GU (?a = 1—?b = 1) ? (1 + e??h) + ? THEN add edge (a, b) to E ? 6: OUTPUT E Theorem 3.6. Let ? be a nonnegative integer strictly smaller than d, and consider the antiferromagnetic Ising model 3.2 with ? ? (d ? ?)(h + ln 2) on a graph G. Algorithm WeakRepelling reconstructs the graph with probability 1 ? o(1) as p ? ? using 1 2 n = O (1 + eh)? 22d e2h(d+1) log p i.i.d. samples, with run-time

4

! ” ?h,d (p2+?) . O np2+? = O

Statistical algorithms and proof of Theorem 1.1

We start by describing the statistical algorithm framework introduced by [1]. In this section it is convenient to work with variables taking values in {?1, +1} rather than {0, 1}. 4.1

Background on statistical algorithms

Let X = {?1, +1}p denote the space of configurations and let D be a set of distributions over X . Let F be a set of solutions (in our case, graphs) and Z : D ? 2F be a map taking each distribution D ? D to a subset of solutions Z(D) ? F that are defined to be valid solutions for D. In our setting, since each

graphical model is identifiable, there is a single graph Z(D) corresponding to each distribution D. For n ¿ 0, the distributional search problem Z over D and F using n samples is to find a valid solution f ? Z(D) given access to n random samples from an unknown D ? D.

The class of algorithms we are interested in are called unbiased statistical algorithms, defined by access to an unbiased oracle. Other related classes of algorithms are defined in [1], and similar lower bounds can be derived for those as well. Definition 4.1 (Unbiased Oracle). Let D be the true distribution. The algorithm is given access to an oracle, which when given any function h : X ? {0, 1}, takes an independent random sample x from D and returns h(x). These algorithms access the sampled data only through the oracle: unbiased statistical algorithms outsource the computation. Because the data is accessed through the oracle, it is possible to prove unconditional lower bounds using information-theoretic methods. As noted in the introduction, many algorithmic approaches can be implemented as statistical algorithms. We now define a key quantity called average correlation. The average correlation of a subset of distributions D? ? D relative to a distribution D is denoted ?(D? , D), ¿ ? -= D1 1 D2 ? ?(D , D) := ? 2 (4.1) - D ? 1, D ? 1 - , —D — D1 ,D2 ?D ?

D

where ?f, g?D := Ex?D [f (x)g(x)] and the ratio D1 /D represents the ratio of probability mass functions, so (D1 /D)(x) = D1 (x)/D(x).

We quote the definition of statistical dimension with average correlation from [1], and then state a lower bound on the number of queries needed by any statistical algorithm. 7

Definition 4.2 (Statistical dimension). Fix ? ¿ 0, ? ¿ 0, and search problem Z over set of solutions F and class of distributions D over X. We consider pairs (D, DD) consisting of a ?reference distribution? D over X and a finite set of distributions DD ? D with the following property: for any solution f ? F, the set Df = DD Z ?1 (f) has size at least (1 ? ?) ? —DD —. Let ?(D, DD) be the largest integer ? so that for any subset D? ? Df with —D? — ? —Df —/?, the average correlation is —?(D? , D)— ¡ ? (if there is no such ? one can take ? = 0). The statistical dimension with average correlation ? and solution set bound ? is defined to be the largest ?(D, DD) for valid pairs (D, DD) as described, and is denoted by SDA(Z, ?, ?). Theorem 4.3 ([1]). Let X be a domain and Z a search problem over a set of solutions F and a class of distributions D over X . For ? ¿ 0 and ? ? (0, 1), let ? = SDA(Z, ?, ?). Any (possibly randomized) unbiased statistical algorithm that solves Z with probability ? requires at least m calls to the Unbiased Oracle for ; ¡ ?(? ? ?) (? ? ?)2 m = min , . 2(1 ? ?) 12? In particular, if ? ? 1/6, then any algorithm with success probability at least 2/3 requires at least min{?/4, 1/48?} samples from the Unbiased Oracle.

In order to show that a graphical model on p nodes of maximum degree d requires computation p (d) in this computational model, we therefore would like to show that SDA(Z, ?, ?) = p (d) with ? = p? (d) . 4.2

Soft parities

r For any subset S ? [p] of cardinality —S— = d, let ?S (x) = i?S xi be the parity of variables in S. Define a probability distribution by assigning mass to

x ? {?1, +1}p according to 1 pS (x) = exp(c ? ?S (x)) . (4.2) Z Here c is a constant, and the partition function is ? Z= exp(c ? ?S (x)) = 2p?1 (ec + e?c) . (4.3) x

Our !p" family of distributions D is given by these soft parities over subsets S ? [p], and —D— = d . The following lemma, proved in the supplementary material, computes correlations between distributions. Lemma 4.4. Let U denote the uniform distribution on {?1, +1}p . For S ?= T , the correlation ? pUS ? 1, pUT ? 1? is exactly equal to zero for any value of c. If S = T , the correlation ? pUS ? 1, pUS ? 1? = 1 ? (ec +e4?c)2 ? 1. Lemma 4.5. For any set D? ? D of size at least —D—/pd/2 , the average correlation satisfies ?(D? , U) ? dd p?d/2 .

Proof. By the preceding lemma, the only contributions to the sum (4.1) comes from choosing ? the same set S in the sum, of which there are a fraction 1/—D such correlation is at ! " —. Each ? d/2 d/2 p most one !by" Lemma 4.4, so ? ? 1/—D — ? p /—D— = p / d ? dd /pd/2 . Here we used the estimate nk ? (nk)k . Proof of Theorem 1.1. Let ? = 1/6 and ? = dd p?d/2 , and consider the set of distributions D given by soft parities as defined above. With reference distribution D = U , the uniform distribution, Lemma 4.5 implies that SDA(Z, ?, ?) of the structure learning problem over distribution (4.2) is at least ? = pd/2 /dd . The result follows from Theorem 4.3. Acknowledgments This work was supported in part by NSF grants CMMI-1335155 and CNS-1161964, and by Army Research Office MURI Award W911NF-11-1-0036. 8

2 References

[1] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao, ?Statistical algorithms and a lower bound for detecting planted cliques,? in STOC, pp. 655?664, ACM, 2013. [2] G. Bresler, E. Mossel, and A. Sly, ?Reconstruction of Markov random fields from samples: Some observations and algorithms,? Approximation, Randomization and Combinatorial Optimization, pp. 343?356, 2008. [3] J. Bento and A. Montanari, ?Which graphical models are difficult to learn?,? in NIPS, 2009. [4] P. Abbeel, D. Koller, and A. Y. Ng, ?Learning factor graphs in polynomial time and sample complexity,? The Journal of Machine Learning Research, vol. 7, pp. 1743?1788, 2006. [5] I. Csisz? ar and Z. Talata, ?Consistent estimation of the basic neighborhood of markov random fields,? The Annals of Statistics, pp. 123?145, 2006. [6] N. P. Santhanam and M. J. Wainwright, ?Information-theoretic limits of selecting binary graphical models in high dimensions,? Info. Theory, IEEE Trans. on, vol. 58, no. 7, pp. 4117?4134, 2012. [7] M. Kearns, ?Efficient noise-tolerant learning from statistical queries,? Journal of the ACM (JACM), vol. 45, no. 6, pp. 983?1006, 1998. [8] C. Chow and C. Liu, ?Approximating discrete probability distributions with dependence trees,? Information Theory, IEEE Transactions on, vol. 14, no. 3, pp. 462?467, 1968. [9] S. Dasgupta, ?Learning polytrees,? in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 134?141, Morgan Kaufmann Publishers Inc., 1999. [10] N. Srebro, ?Maximum likelihood bounded

tree-width markov networks,? in Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, pp. 504?511, Morgan Kaufmann Publishers Inc., 2001. [11] R. L. Dobrushin, ?Prescribing a system of random variables by conditional distributions,? Theory of Probability & Its Applications, vol. 15, no. 3, pp. 458?486, 1970. [12] R. L. Dobrushin and S. B. Shlosman, ?Constructive criterion for the uniqueness of gibbs field,? in Statistical physics and dynamical systems, pp. 347?370, Springer, 1985. [13] J. Salas and A. D. Sokal, ?Absence of phase transition for antiferromagnetic potts models via the dobrushin uniqueness theorem,? Journal of Statistical Physics, vol. 86, no. 3-4, pp. 551?579, 1997. [14] D. Gamarnik, D. A. Goldberg, and T. Weber, ?Correlation decay in random decision networks,? Mathematics of Operations Research, vol. 39, no. 2, pp. 229?261, 2013. [15] D. Gamarnik and D. Katz, ?Correlation decay and deterministic fptas for counting listcolorings of a graph,? in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1245?1254, Society for Industrial and Applied Mathematics, 2007. [16] D. Weitz, ?Counting independent sets up to the tree threshold,? in Proceedings of the thirtyeighth annual ACM symposium on Theory of computing, pp. 140?149, ACM, 2006. [17] P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai, ?Greedy learning of markov network structure,? in 48th Allerton Conference, pp. 1295?1302, 2010. [18] A. Ray, S. Sanghavi, and S. Shakkottai, ?Greedy learning of graphical models with small girth,? in 50th Allerton Conference, 2012. [19] A. Anandkumar, V. Tan, F. Huang, and A. Willsky, ?High-dimensional structure estimation in Ising models: Local separation criterion,? Ann. of Stat., vol. 40, no. 3, pp. 1346?1375, 2012. [20] R. Wu, R. Srikant, and J. Ni, ?Learning loosely connected Markov random fields,? Stochastic Systems, vol. 3, no. 2, pp. 362?404, 2013. [21] P. Ravikumar, M. Wainwright, and J. Lafferty, ?High-dimensional Ising model selection using ?1 -regularized logistic regression,? The Annals of Statistics, vol. 38, no. 3, pp. 1287?1319, 2010. [22] S.-I. Lee, V. Ganapathi, and D. Koller, ?Efficient structure learning of markov networks using l 1-regularization,? in Advances in neural Information processing systems, pp. 817?824, 2006. [23] A. Jalali, C. C. Johnson, and P. D. Ravikumar, ?On learning discrete graphical models using greedy methods.,? in NIPS, pp. 1935?1943, 2011. [24] A. Jalali, P. Ravikumar, V. Vasuki, S. Sanghavi, and U. ECE, ?On learning discrete graphical models using group-sparse regularization,? in Inter. Conf. on AI and Statistics (AISTATS), vol. 14, 2011. [25] A. Sinclair, P. Srivastava, and M. Thurley, ?Approximation algorithms for two-state antiferromagnetic spin systems on bounded degree graphs,? Journal of Statistical Physics, vol. 155, no. 4, pp. 666?686, 2014.

9