Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms

Guy Bresler[1,⋆], Elchanan Mossel[2,⋆⋆], and Allan Sly[3,⋆⋆⋆]

[1] Dept. of Electrical Engineering and Computer Sciences, U.C. Berkeley
gbresler@eecs.berkeley.edu
[2] Dept. of Statistics and Dept. of Electrical Engineering and Computer Sciences, U.C. Berkeley
mossel@stat.berkeley.edu
[3] Dept. of Statistics, U.C. Berkeley
sly@stat.berkeley.edu

Abstract. Markov random fields are used to model high dimensional distributions in a number of applied areas. Much recent interest has been devoted to the reconstruction of the dependency structure from independent samples from the Markov random fields. We analyze a simple algorithm for reconstructing the underlying graph defining a Markov random field on n nodes and maximum degree d given observations. We show that under mild non-degeneracy conditions it reconstructs the generating graph with high probability using $\Theta(d \log n)$ samples which is optimal up to a multiplicative constant. Our results seem to be the first results for general models that guarantee that *the* generating model is reconstructed. Furthermore, we provide an explicit $O(dn^{d+2} \log n)$ running time bound. In cases where the measure on the graph has correlation decay, the running time is $O(n^2 \log n)$ for all fixed d. In the full-length version we also discuss the effect of observing noisy samples. There we show that as long as the noise level is low, our algorithm is effective. On the other hand, we construct an example where large noise implies non-identifiability even for generic noise and interactions. Finally, we briefly show that in some cases, models with hidden nodes can also be recovered.

1 Introduction

In this paper we consider the problem of reconstructing the graph structure of a Markov random field from independent and identically distributed samples. Markov random fields (MRF) provide a very general framework for defining high dimensional distributions and the reconstruction of the MRF from observations

⋆ Supported by a Vodafone US-Foundation fellowship and NSF Graduate Research Fellowship.
⋆⋆ Supported by a Sloan fellowship in Mathematics, by NSF Career award DMS-0548249, NSF grant DMS-0528488 and ONR grant N0014-07-1-05-06.
⋆⋆⋆ Supported by NSF grants DMS-0528488 and DMS-0548249.

A. Goel et al. (Eds.): APPROX and RANDOM 2008, LNCS 5171, pp. 343–356, 2008.

has attracted much recent interest, in particular in biology, see e.g. [9] and a list of related references [10].

1.1 Our Results

We give sharp, up to a multiplicative constant, estimates for the number of independent samples needed to infer the underlying graph of a Markov random field. In Theorem 2 we use a simple information-theoretic argument to show that $\Omega(d \log n)$ samples are required to reconstruct a randomly selected graph on n vertices with maximum degree at most d. Then in Theorems 4 and 5 we propose two algorithms for reconstruction that use only $O(d \log n)$ samples assuming mild non-degeneracy conditions on the probability distribution. The two theorems differ in their running time and the required non-degeneracy conditions. It is clear that non-degeneracy conditions are needed to insure that there is a unique graph associated with the observed probability distribution.

Chickering [2] showed that maximum-likelihood estimation of the underlying graph of a Markov random field is NP-complete. This does not contradict our results which assume that the data is generated from a model (or a model with a small amount of noise). Although the algorithm we propose runs in time polynomial in the size of the graph, the dependence on degree (the run-time is $O(dn^{d+2} \log n)$) may impose too high a computational cost for some applications. Indeed, for some Markov random fields exhibiting a decay of correlation a vast improvement can be realized: a modified version of the algorithm runs in time $O(dn^2 \log n)$. This is proven in Theorem 8.

In addition to the fully-observed setting in which samples of all variables are available, we extend our algorithm in several directions. These sections are omitted due to space constraints; we refer the reader to the full version [14] for the discussion on these topics. In Section 5 of [14] we consider the problem of noisy observations. We first show by way of an example that if some of the random variables are perturbed by noise then it is in general impossible to reconstruct the graph structure with probability approaching 1. Conversely, when the noise is relatively weak as compared to the coupling strengths between random variables, we show that the algorithms used in Theorems 4 and 5 reconstruct the graph with high probability. Furthermore, we study the problem of reconstruction with partial observations, i.e. samples from only a subset of the nodes are available, and provide sufficient conditions on the probability distribution for correct reconstruction.

1.2 Related Work

Chow and Liu [1] considered the problem of estimating Markov random fields whose underlying graphs are trees, and provided an efficient (polynomial-time) algorithm based on the fact that in the tree case maximum-likelihood estimation amounts to the computation of a maximum-weight spanning tree with edge weights equal to pairwise empirical mutual information. Unfortunately, their approach does not generalize to the estimation of Markov random fields whose

graphs have cycles or hidden nodes. Much work in mathematical biology is devoted to reconstructing tree Markov fields when there are hidden nodes. For trees, given data that is generated from the model, the tree can be reconstructed efficiently from samples at a subset of the nodes given mild non-degeneracy conditions. See [12,13,11] for some of the most recent and tightest results in this setup.

Abbeel, *et al* [3] considered the problem of reconstructing graphical models based on factor graphs, and proposed a polynomial time and sample complexity algorithm. However, the goal of their algorithm was not to reconstruct the true structure, but rather to produce a distribution that is close in Kullback-Leibler divergence to the true distribution. In applications it is often of interest to reconstruct the true structure which gives some insight into the underlying structure of the inferred model.

Note furthermore that two networks that differ only in the neighborhood of one node will have $O(1)$ KL distance. Therefore, even in cases where it is promised that the KL distance between the generating distribution and any other distribution defined by another graph is as large as possible, the lower bounds on the KL distance is $\Omega(1)$. Plugging this into the bounds in [3] yields a polynomial sampling complexity in order to find the generating network compared to our logarithmic sampling complexity. For other work based on minimizing the KL divergence see the references in [3].

Essentially the same problem as in the present work (but restricted to the Ising model) was studied by Wainwright, *et al* [5], where an algorithm based on ℓ_1-regularization was introduced. In that work, sufficient conditions—different than ours—for correct reconstruction were given. They require a condition (called A2) where the neighborhood of every vertex is only weakly affected by their neighbors. Verifying when the condition holds seems hard and no example is given in the paper where the condition holds. The simulation studies in the paper are conducted for graphs consisting of small disconnected components. In this setting the running time of their algorithm is $O(n^5)$. The result [5] is best compared to our result showing that under standard decay of correlation (e.g., for models satisfying the Dobrushin condition, which is satisfied for the models simulated in their work), the running time of our algorithm is $O(n^2 \log n)$ as given in Theorem 8. The algorithm of [5] has suboptimal sample complexity, requiring $\Theta(d^5 \log n)$ samples for reconstruction.

Subsequent to our work being posted on the Arxiv, Santhanam and Wainwright [4] again considered essentially the same problem for the Ising model, producing nearly matching lower and upper bounds on the asymptotic sampling complexity. A key difference from our work is that they restrict attention to the Ising model, i.e. Markov random fields with pairwise potentials and where each variable takes two values. Also, they consider models with a fixed number of total edges, and arbitrary node degree, in contrast to our study of models with bounded node degrees and an arbitrary number of edges. We note that their results are limited to determining the information theoretic sampling complexity for reconstruction, and provide no efficient algorithm.

2 Preliminaries

We begin with the definition of Markov random field.

Definition 1. *On a graph $G = (V, E)$, a* Markov random field *is a distribution X taking values in $\mathcal{A}^V$, for some finite set $\mathcal{A}$ with $\mid \mathcal{A} \mid = A$, which satisfies the Markov property*

$$P(X(W), X(U) \mid X(S)) = P(X(W) \mid X(S))P(X(U) \mid X(S)) \tag{1}$$

when W, U, and S are disjoint subsets of V such that every path in G from W to U passes through S and where $X(U)$ denotes the restriction of X from $\mathcal{A}^V$ to $\mathcal{A}^U$ for $U \subset V$.

Famously, by the Hammersley-Clifford Theorem, such distributions can be written in a factorized form as

$$P(\sigma) = \frac{1}{Z} \exp\left[\sum_a \Psi_a(\sigma_a)\right] \quad , \tag{2}$$

where Z is a normalizing constant, a ranges over the cliques in G, and $\Psi_a : \mathcal{A}^{|a|} \to \mathbb{R} \cup \{-\infty\}$ are functions called *potentials*.

The problem we consider is that of reconstructing the graph G, given k independent samples $\underline{X} = \{X^1, \ldots, X^k\}$ from the model. Denote by $\mathcal{G}_d$ the set of labeled graphs with maximum degree at most d. We assume that the graph $G \in \mathcal{G}_d$ is from this class. A structure estimator (or reconstruction algorithm) $\widehat{G} : \mathcal{A}^{kn} \to \mathcal{G}_d$ is a map from the space of possible sample sequences to the set of graphs under consideration. We are interested in the asymptotic relationship between the number of nodes in the graph, n, the maximum degree d, and the number of samples k that are required. An algorithm using number of samples $k(n)$ is deemed successful if in the limit of large n the probability of reconstruction error approaches zero.

3 Lower Bound on Sample Complexity

Suppose G is selected uniformly at random from $\mathcal{G}_d$. The following theorem gives a lower bound of $\Omega(d \log n)$ on the number of samples necessary to reconstruct the graph G. The argument is information theoretic, and follows by comparing the number of possible graphs with the amount of information available from the samples.

Theorem 2. *Let the graph G be drawn according to the uniform distribution on $\mathcal{G}_d$. Then there exists a constant $c = c(A) > 0$ such that if $k \leq cd \log n$ then for any estimator $\widehat{G} : \underline{X} \to \mathcal{G}_d$, the probability of correct reconstruction is $P(\widehat{G} = G) = o(1)$.*

Remark 1. Note that the theorem above doesn't need to assume anything about the potentials. The theorem applies for any potentials that are consistent with

the generating graph. In particular, it is valid both in cases where the graph is "identifiable" given many samples and in cases where it isn't.

Proof. To begin, we note that the probability of error is minimized by letting $\widehat{G}$ be the maximum a posteriori (MAP) decision rule,

$$\widehat{G}_{\mathrm{MAP}}(\underline{X}) = \mathrm{argmax}_{g \in G} P(G = g \mid \underline{X}).$$

By the optimality of the MAP rule, this bounds the probability of error using any estimator. Now, the MAP estimator $\widehat{G}_{\mathrm{MAP}}(\underline{X})$ is a deterministic function of $\underline{X}$. Clearly, if a graph g is not in the range of $\widehat{G}$ then the algorithm always makes an error when $G = g$. Let S be the set of graphs in the range of $\widehat{G}_{\mathrm{MAP}}$, so $P(\mathrm{error} \mid g \in S^c) = 1$. We have

$$\begin{aligned} P(\mathrm{error}) &= \sum_{g \in \mathcal{G}} P(\mathrm{error} \mid G = g) P(G = g) \\ &= \sum_{g \in S} P(\mathrm{error} \mid G = g) P(G = g) + \sum_{g \in S^c} P(\mathrm{error} \mid G = g) P(G = g) \\ &\geq \sum_{g \in S^c} P(G = g) = 1 - \sum_{g \in S} \mid \mathcal{G} \mid^{-1} \\ &\geq 1 - \frac{A^{nk}}{\mid \mathcal{G} \mid}, \end{aligned} \tag{3}$$

where the last step follows from the fact that $\mid S \mid \leq \mid \underline{\mathrm{X}} \mid \leq A^{nk}$. It remains only to express the number of graphs with max degree at most d, $\mid \mathcal{G}_d \mid$, in terms of the parameters n, d. The following lemma gives an adequate bound.

Lemma 3. *Suppose $d \leq n^\alpha$ with $\alpha < 1$. Then the number of graphs with max degree at most d, $\mid \mathcal{G}_d \mid$, satisfies*

$$\log \mid \mathcal{G}_d \mid = \Omega(nd \log n). \tag{4}$$

Proof. To make the dependence on n explicit, let $U_{n,d}$ be the number of graphs with n vertices with maximum degree at most d. We first bound $U_{n+2,d}$ in terms of $U_{n,d,}$. Given a graph G with n vertices and degree at most d, add two vertices a and b. Select d distinct neighbors $v_1, \ldots, v_d$ for vertex a, with d labeled edges; there are $\binom{n}{d} d!$ ways to do this. If v_i already has degree d in G, then v_i has at least one neighbor u that is not a neighbor of a, since there are only $d-1$ other neighbors of a. Remove the edge (v_i, u) and place an edge labeled i from vertex b to u. This is done for each vertex $v_1, \ldots, v_d$, so b has degree at most d. The graph G can be reconstructed from the resulting labeled graph on $n+2$ vertices as follows: remove vertex a, and return the neighbors of b to their correct original neighbors (this is possible because the edges are labeled).

Removing the labels on the edges from a and b sends at most $d!^2$ edge-labeled graphs of this type on $n+2$ vertices to the same unlabeled graph. Hence, the number of graphs with max degree d on $n+2$ vertices is lower bounded as

$$U_{n+2,d} \geq U_{n,d}\binom{n}{d} d! \frac{1}{d!^2} = U_{n,d}\binom{n}{d}\frac{1}{d!}.$$

It follows that for n even (and greater than $2d+4$)

$$U_{n,d} \geq \times_{i=1}^{n/2}\binom{n-2i}{d}\frac{1}{d!} \geq \left(\binom{n/2}{d}\frac{1}{d!}\right)^{n/4}. \tag{5}$$

If n is odd, it suffices to note that $U_{n+1,d} \geq U_{n,d}$. Taking the logarithm of equation (5) yields

$$\log U_{n,d} = \Omega(nd(\log n - \log d)) = \Omega(nd \log n), \tag{6}$$

assuming that $d \leq n^{\alpha}$ with $\alpha < 1$. □

Together with equation (3), Lemma 3 implies that for small enough c, if the number of samples $k \leq cd \log n$, then

$$P(\text{error}) \geq 1 - \frac{A^{nk}}{|\mathcal{G}|} = 1 - o(1).$$

This completes the proof of Theorem 2. □

4 Reconstruction

We now turn to the problem of reconstructing the graph structure of a Markov random field from samples. For a vertex v we let $N(v) = \{u \in V \setminus \{v\} : (u,v) \in E\}$ denote the set of neighbors of v. Determining the neighbors of v for every vertex in the graph is sufficient to determine all the edges of the graph and hence reconstruct the graph. Our algorithms reconstruct the graph by testing each candidate neighborhood of size at most d by using the Markov property, which states that for each $w \in V \setminus (N(v) \cup \{v\})$

$$P(X(v) \mid X(N(v)), X(w)) = P(X(v) \mid X(N(v))). \tag{7}$$

We give two algorithms for reconstructing networks; they differ in their non-degeneracy conditions and their running time. The first one, immediately below, has more stringent non-degeneracy conditions and faster running time.

4.1 Conditional Two Point Correlation Reconstruction

The first algorithm requires the following non-degeneracy condition:

Condition N1: There exist $\epsilon, \delta > 0$ such that for all $v \in V$, if $U \subset V \setminus \{v\}$ with $\mid U \mid \leq d$ and $N(v) \nsubseteq U$ then there exist values $x_v, x_w, x'_w, x_{u_1}, \ldots, x_{u_l}$ such that for some $w \in V \setminus (U \cup \{v\})$

$$\begin{aligned} \big|&P(X(v) = x_v \mid X(U) = x_U, X(w) = x_w) \\ &- P(X(v) = x_v \mid X(U) = x_U, X(w) = x'_w)\big| > \epsilon \end{aligned} \tag{8}$$

and

$$\begin{aligned} |P(X(U) = x_U, X(w) = x_w)| &> \delta, \\ |P(X(U) = x_U, X(w) = x'_w)| &> \delta. \end{aligned} \tag{9}$$

Remark 2. Condition (8) captures the notion that each edge should have sufficient strength. Condition (9) is required so that we can accurately calculate the empirical conditional probabilities.

We now describe the reconstruction algorithm, with the proof of correctness given by Theorem 4. In the following, $\widehat{P}$ denotes the empirical probability measure from the k samples.

Algorithm SIMPLERECON(*Input:* k i.i.d. samples from MRF; *Output:* estimated graph G)

- Initialize $E = \varnothing$.
- For each vertex v do
  - For each $U \subseteq V \setminus \{v\}$ with $|U| \leq d$, $w \in V \setminus (U \cup \{v\})$, and $x_1, \ldots, x_l, x_w, x'_w, x_v \in \mathcal{A}$
    - If
$$|\widehat{P}(X(U) = x_U, X(w) = x_w)| > \delta/2$$
and
$$|\widehat{P}(X(U) = x_U, X(w) = x'_w)| > \delta/2 \ ,$$
then compute
$$\begin{aligned} r(U, x_U, w, x_w, x'_w) = &\big|\widehat{P}(X(v) = x_v | X(U) = x_U, X(w) = x_w) \\ &- \widehat{P}(X(v) = x_v | X(U) = x_U, X(w) = x'_w)\big| \ . \end{aligned}$$
  - Let $N(v)$ be the minimum cardinality U such that $\max_{x_U, w, x_w, x'_w} r(U, x_U, w, x_w, x'_w) < \epsilon/2$.
  - Add the edges incident to v: $E = E \cup \{(v, u) : u \in N(v)\}$.
- Return the graph $G = (V, E)$.

Run-time analysis. The analysis of the running time is straightforward. There are n nodes, and for each node we consider $O(n^d)$ neighborhoods U. For each candidate neighborhood, we check $O(n)$ nodes x_w and perform a correlation test of complexity $O(d \log n)$. The run-time of SIMPLERECON is thus $O(dn^{d+2} \log n)$ operations.

We now give the main theorem.

Theorem 4 (Correctness of SimpleRecon). *Suppose the MRF satisfies condition* **N1**. *Then with the constant* $C = \left(\frac{81(d+2)}{\epsilon^2 \delta^4 2d} + C_1\right)$, *when* $k > Cd \log n$, *the estimator* SIMPLERECON *correctly reconstructs with probability at least* $1 - O(n^{-C_1})$.

Proof. Azuma's inequality gives that if $Y \sim \mathrm{Bin}(k,p)$ then

$$P(\mid Y - kp \mid > \gamma k) \leq 2\exp(-2\gamma^2 k)$$

and so for any collection $U = \{u_1, \ldots, u_l\} \subseteq V$ and $x_1, \ldots, x_l \in \mathcal{A}$ we have

$$P\left(\left|\widehat{P}(X(U) = x_U) - P(X(U) = x_U)\right| \leq \gamma\right) \leq 2\exp(-2\gamma^2 k). \tag{10}$$

There are $A^l\binom{n}{l} \leq A^l n^l$ such choices of $u_1, \ldots, u_l$ and $x_1, \ldots, x_l$. An application of the union bound implies that with probability at least $1 - A^l n^l 2\exp(-2\gamma^2 k)$ it holds that

$$\left|\widehat{P}(X(U) = x_U) - P(X(U) = x_U)\right| \leq \gamma \tag{11}$$

for all $\{u_i\}_{i=1}^l$ and $\{x_i\}_{i=1}^l$. If we additionally have $l \leq d+2$ and $k \geq C(\gamma) d \log n$, then equation (11) holds with probability at least $1 - A^{d+2} n^{d+2} 2/n^{2\gamma^2 C(\gamma) d}$. Choosing $C(\gamma) = \frac{d+2}{\gamma^2 2d} + C_1$, equation (11) holds with probability at least $1 - 2A^{d+2}/n^{C_1}$.

For the remainder of the proof assume (11) holds. Taking

$$\gamma(\epsilon, \delta) = \epsilon\delta^2/9 \ , \tag{12}$$

we can bound the error in conditional probabilities as

$$\begin{aligned}
&\mid \widehat{P}(X(v) = x_v \mid X(U) = x_U) - P(X(v) = x_v \mid X(U) = x_U) \mid \\
&= \left| \frac{\widehat{P}(X(v) = x_v, X(U) = x_U)}{\widehat{P}(X(U) = x_U)} - \frac{P(X(v) = x_v, X(U) = x_U)}{P(X(U) = x_U)} \right| \\
&\leq \left| \frac{\widehat{P}(X(v) = x_v, X(U) = x_U)}{P(X(U) = x_U)} - \frac{P(X(v) = x_v, X(U) = x_U)}{P(X(U) = x_U)} \right| \\
&\quad + \left| \frac{1}{\widehat{P}(X(U) = x_U)} - \frac{1}{P(X(U) = x_U)} \right| \\
&\leq \frac{\gamma}{\delta} + \frac{\gamma}{(\delta - \gamma)\delta} \leq \frac{\epsilon\delta^2}{9\delta} + \frac{\epsilon\delta^2}{9(\delta - \frac{\epsilon\delta^2}{9})\delta} = \frac{\epsilon\delta}{9} + \frac{\epsilon}{(9 - \epsilon\delta)} < \frac{\epsilon}{4} \ .
\end{aligned} \tag{13}$$

For each vertex $v \in V$ we consider all candidate neighborhoods for v, subsets $U \subset V \setminus \{v\}$ with $\mid U \mid \leq d$. The estimate (13) and the triangle inequality imply that if $N(v) \subseteq U$, then by the Markov property,

$$\begin{aligned}
&\left|\widehat{P}(X(v) = x_v \mid X(U) = x_U, X(w) = x_w)\right. \\
&\quad \left. - \widehat{P}(X(v) = x_v \mid X(U) = x_U, X(w) = x'_w)\right| < \epsilon/2
\end{aligned} \tag{14}$$

for all $w \in V$ and $x_1, \ldots, x_l, x_w, x'_w, x_v \in \mathcal{A}$ such that

$$\begin{aligned}
&\left|\widehat{P}(X(U) = x_U, X(w) = x_w)\right| > \delta/2, \\
&\left|\widehat{P}(X(U) = x_U, X(w) = x'_w)\right| > \delta/2.
\end{aligned} \tag{15}$$

Conversely by condition **N1** and (9) and the estimate (13), we have that for any U with $N(v) \nsubseteq U$ there exists some $w \in V$ and $x_{u_1}, \ldots, x_{u_l}, x_w, x'_w, x_v \in \mathcal{A}$ such that equation (15) holds but equation (14) does not hold. Thus, choosing the smallest set U such that (14) holds gives the correct neighborhood.

To summarize, with number of samples

$$k = \left(\frac{81(d+2)}{\epsilon^2 \delta^4 2d} + C_1 \right) d \log n$$

the algorithm correctly determines the graph G with probability

$$P(\textsc{SimpleRecon}(X) = G) \geq 1 - 2A^{d+2}/n^{C_1} \quad .$$

□

4.2 General Reconstruction

While the algorithm SimpleRecon applies to a wide range of models, condition **N1** may occasionally be too restrictive. One setting in which condition **N1** does not apply is if the marginal spin at some vertex v is independent of the marginal spin at each of the other vertices, (i.e for all $u \in V \setminus \{v\}$ and all $x, y \in \mathcal{A}$ we have $P(X(v) = x, X(u) = y) = P(X(v) = x)P(X(u) = y)$. In this case the algorithm would incorrectly return the empty set for the neighborhood of v. The weaker condition for GeneralRecon holds on essentially all Markov random fields. In particular, (16) says that the potentials are non-degenerate, which is clearly a necessary condition in order to recover the graph. Equation (17) holds for many models, for example all models with soft constraints. This additional generality comes at a computational cost, with SimpleRecon having a faster running time, $O(dn^{d+2} \log n)$, versus $O(dn^{2d+1} \log n)$ for GeneralRecon.

We use the following notation in describing the non-degeneracy conditions. For an assignment $x_U = (x_{u_1}, \ldots, x_{u_l})$ and $x'_{u_i} \in \mathcal{A}$, define

$$x^i_U(x'_{u_i}) = (x_{u_1}, \ldots, x'_{u_i}, \ldots, x_{u_l})$$

to be the assignment obtained from x_U by replacing the ith element by x'_{u_i}.

Condition N2: There exist $\epsilon, \delta > 0$ such that the following holds: for all $v \in V$, if $N(v) = u_1, \ldots, u_l$, then for each $i, 1 \leq i \leq l$ and for any set $W \subset V \setminus (v \cup N(v))$ with $\mid W \mid \leq d$ there exist values $x_v, x_{u_1}, \ldots, x_{u_i}, \ldots, x_{u_l}, x'_{u_i} \in \mathcal{A}$ and $x_W \in \mathcal{A}^{|W|}$ such that

$$\begin{aligned} \big| & P(X(v) = x_v \mid X(N(v)) = x_{N(v)}) \\ & - P(X(v) = x_v \mid X(N(v)) = x^i_{N(v)}(x'_{u_i})) \big| > \epsilon \end{aligned} \tag{16}$$

and

$$\begin{aligned} & \mid P(X(N(v)) = x_{N(v)}, X(W) = x_W) \mid > \delta, \\ & \mid P(X(N(v)) = x^i_{N(v)}(x'_{u_i}), X(W) = x_W) \mid > \delta. \end{aligned} \tag{17}$$

We now give the algorithm GeneralRecon.

Algorithm GeneralRecon(*Input:* k i.i.d. samples from MRF; *Output:* estimated graph G)

- Initialize $E = \varnothing$.
- For each vertex v do
  - Initialize $N(v) = \varnothing$.
  - For each $U \subseteq V \setminus \{v\}$ with $l = |U| \leq d$, $W \in V \setminus (U \cup \{v\})$ with $|W| \leq d$, each i, $1 \leq i \leq l$, and $x_v, x_W, x_U, x'_{u_i} \in \mathcal{A}$
    * If

$$\widehat{P}(X(W) = x_W, X(U) = x_U) > \delta/2$$
$$\widehat{P}(X(W) = x_W, X(U) = x^i_U(x'_{u_i})) > \delta/2$$

then compute

$$\begin{aligned} &r(U, W, i, x_v, x_W, x_U, x'_{u_i}) \\ &= \big|\widehat{P}(X(v) = x_v | X(W) = x_W, X(U) = x_U) \\ &\qquad - \widehat{P}(X(v) = x_v | X(W) = x_W, X(U) = x^i_U(x'_{u_i}))\big|. \end{aligned}$$

  - Let $N(v)$ be the maximum cardinality set U such that $\min_{W,i} \max_{x_v, x_W, x_U, x'_{u_i}} r(U, W, i, x_v, x_W, x_U, x'_{u_i}) > \epsilon/2$.
  - Add the edges incident to v: $E = E \cup \{(v, u) : u \in N(v)\}$.
- Return the graph $G = (V, E)$.

Run-time analysis. The analysis of the running time is similar to the previous algorithm. The run-time of GeneralRecon is $O(dn^{2d+1} \log n)$.

Theorem 5 (Correctness of GeneralRecon). *Suppose condition* **N2** *holds with ϵ and δ. Then for the constant $C = \frac{81(2d+1)}{\epsilon^2 \delta^4 2d} + C_1$, if $k > Cd \log n$ then the estimator* GeneralRecon *correctly reconstructs with probability at least $1 - O(n^{-C_1})$.*

Proof. As in Theorem 4 we have that with high probability

$$\left|\widehat{P}(X(U) = x_U) - P(X(U) = x_U)\right| \leq \gamma \tag{18}$$

for all $\{u_i\}_{i=1}^l$ and $\{x_i\}_{i=1}^l$ when $l \leq 2d+1$ and $k \geq C(\gamma) d \log n$; we henceforth assume that (18) holds. For each vertex $v \in V$ we consider all candidate neighborhoods for v, subsets $U = \{u_1, \ldots, u_l\} \subset V \setminus \{v\}$ with $0 \leq l \leq d$. For each candidate neighborhood U, the algorithm computes a score

$$\begin{aligned} &f(v;\ U) = \\ &min_{W,i}\, max_{x_v, x_W, x_U, x'_{u_i}} \big|\widehat{P}(X(v) = x_v \mid X(W) = x_W, X(U) = x_U) \\ &\qquad - \widehat{P}(X(v) = x_v \mid X(W) = x_W, X(U) = x^i_U(x'_{u_i}))\big|, \end{aligned}$$

where for each W, i, the maximum is taken over all x_v, x_W, x_U, x'_{u_i}, such that

$$\widehat{P}(X(W) = x_W, X(U) = x_U) > \delta/2 \tag{19}$$
$$\widehat{P}(X(W) = x_W, X(U) = x_U^i(x'_{u_i})) > \delta/2$$

and $W \subset V \setminus (U \cup \{v\})$ is an arbitrary set of nodes of size d, $x_W \in \mathcal{A}^d$ is an arbitrary assignment of values to the nodes in W, and $1 \leq i \leq l$.

The algorithm selects as the neighborhood of v the largest set $U \subset V \setminus \{v\}$ with $f(v;\ U) > \epsilon/2$. It is necessary to check that if U is the true neighborhood of v, then the algorithm accepts U, and otherwise the algorithm rejects U.

Taking $\gamma(\epsilon, \delta) = \epsilon\delta^2/9$, it follows exactly as in Theorem 4 that the error in each of the relevant empirical conditional probabilities satisfies

$$\begin{aligned} &| \widehat{P}(X(v) = x_v \mid X(W) = x_W, X(U) = x_U) \\ &\quad - P(X(v) = x_v \mid X(W) = x_W, X(U) = x_U) | < \frac{\epsilon}{4} \ . \end{aligned} \tag{20}$$

If $U \nsubseteq N(v)$, choosing $u_i \in U - N(v)$, we have when $N(v) \subset W \cup U$ that

$$\begin{aligned} &\big|P(X(v) = x_v \mid X(W) = x_W, X(U) = x_U) \\ &\quad - P(X(v) = x_v \mid X(W) = x_W, X(U) = x_U^i(x'_{u_i}))\big| \\ &= \big|P(X(v) = x_v \mid X(N(v)) = x_{N(v)}) - P(X(v) = x_v \mid X(N(v)) = x_{N(v)})\big| \\ &= 0 \ , \end{aligned}$$

by the Markov property (7). Assuming that equation (18) holds with γ chosen as in (12), the estimation error in $f(v;\ U)$ is at most $\epsilon/2$ by equation (20) and the triangle inequality, and it holds that $f(v;\ U) < \epsilon/2$ for each $U \nsubseteq N(v)$. Thus all $U \nsubseteq N(v)$ are rejected. If $U = N(v)$, then by the Markov property (7) and the conditions (16) and (17), for any i and $W \subset V$,

$$\begin{aligned} &\big|P(X(v) = x_v \mid X(W) = x_W, X(U) = x_U) \\ &\quad - P(X(v) = x_v \mid X(W) = x_W, X(U) = x_U^i(x'_{u_i}))\big| \\ &= \big|P(X(v) = x_v \mid X(N(v)) = x_{N(v)}) - P(X(v) = x_v \mid X(N(v)) = x^i_{N(v)}(x'_{u_i}))\big| \\ &> \epsilon \end{aligned}$$

for some x_v, x_W, x_U, x'_{u_i}. The error in $f(v;\ U)$ is less than $\epsilon/2$ as before, hence $f(v;\ U) > \epsilon/2$ for $U = N(v)$. Since $U = N(v)$ is the largest set that is not rejected, the algorithm correctly determines the neighborhood of v for every $v \in V$ when (18) holds.

To summarize, with number of samples

$$k = \left(\frac{81(2d+1)}{\epsilon^2\delta^4 2d} + C_1\right) d \log n$$

the algorithm correctly determines the graph G with probability

$$P(\textsc{GeneralRecon}(X) = G) \geq 1 - 2A^{2d+1}/n^{C_1} \ .$$

□

4.3 Non-degeneracy of Models

We can expect condition **N2** to hold in essentially all models of interest. The following proposition shows that the condition holds for any model with soft constraints.

Proposition 6 (Models with soft constraints). *In a graphical model with maximum degree d given by equation* (2) *suppose that all the potentials Ψ_{uv} satisfy $\|\Psi_{uv}\|_\infty \le K$ and*

$$max_{x_1,x_2,x_3,x_4 \in \mathcal{A}} |\Psi_{uv}(x_1,x_2) - \Psi_{uv}(x_3,x_2) - \Psi_{uv}(x_1,x_4) + \Psi_{uv}(x_3,x_4)| > \gamma, \tag{21}$$

for some $\gamma > 0$. Then there exist $\epsilon, \delta > 0$ depending only on d, K and γ such that condition **N2** *holds.*

Proof. It is clear that for some sufficiently small $\delta = \delta(d,m,K) > 0$ we have that for all $u_1, \ldots, u_{2d+1} \in V$ and $x_{u_1}, \ldots, x_{u_{2d+1}} \in \mathcal{A}$ that

$$P(X(u_1) = x_{u_1}, \ldots, X(u_{2d+1}) = x_{u_{2d+1}}) > \delta. \tag{22}$$

Now suppose that $u_1, \ldots, u_l$ is the neighborhood of v. Then for any $1 \le i \le l$ it follows from equation (21) that there exists $x_v, x'_v, x_{u_i}, x'_{u_i} \in \mathcal{A}$ such that for any $x_{u_1} \ldots, x_{u_{i-1}}, x_{u_{i+1}}, \ldots, x_{u_l} \in \mathcal{A}$,

$$\begin{aligned} &\frac{P(X(v) = x_v \mid X(u_1) = x_{u_1}, \ldots, X(u_i) = x'_{u_i}, \ldots, X(u_l) = x_{u_l})}{P(X(v) = x'_v \mid X(u_1) = x_{u_1}, \ldots, X(u_i) = x'_{u_i}, \ldots, X(u_l) = x_{u_l})} \\ \ge\ & e^{\gamma} \frac{P(X(v) = x_v \mid X(u_1) = x_{u_1}, \ldots, X(u_i) = x_{u_i}, \ldots, X(u_l) = x_{u_l})}{P(X(v) = x'_v \mid X(u_1) = x_{u_1}, \ldots, X(u_i) = x_{u_i}, \ldots, X(u_l) = x_{u_l})}. \end{aligned}$$

Combining with equation (22), equation (16) follows, showing that condition **N2** holds. □

Although the results to follow hold more generally, for ease of exposition we will keep in mind the example of the Ising model with no external magnetic field,

$$P(x) = \frac{1}{Z} \exp\left(\sum_{(u,v) \in E} \beta_{uv} x_u x_v \right) , \tag{23}$$

where $\beta_{uv} \in \mathbb{R}$ are coupling constants and Z is a normalizing constant.

The following lemma gives explicit bounds on ϵ and δ in terms of bounds on the coupling constants in the Ising model, showing that condition **N2** can be expected to hold quite generally.

Proposition 7. *Consider the Ising model with all parameters satisfying*

$$0 < c < | \beta_{ij} | < C$$

on a graph G with max degree at most d . Then condition **N2** *is satisfied with*

$$\epsilon \ge \frac{\tanh(2c)}{2C^2 + 2C^{-2}} \quad \textit{and} \quad \delta \ge \frac{e^{-4dC}}{2^{2d}}.$$

Proof. We refer the reader to the full version [14] for the proof. □

4.4 $O(n^2 \log n)$ Algorithm for Models with Correlation Decay

The reconstruction algorithms SIMPLERECON and GENERALRECON run in polynomial time $O(dn^{d+2} \log n)$ and $O(dn^{2d+1} \log n)$, respectively. It would be desirable for the degree of the polynomial to be independent of d, and this can be achieved for Markov random fields with exponential decay of correlations. For two vertices $u, v \in V$, let $d(u, v)$ denote the graph distance and let $d_C(u, v)$ denote the correlation between the spins at u and v defined as

$$d_C(u,v) = \sum_{x_u, x_v \in \mathcal{A}} \left| P(X(u) = x_u, X(v) = x_v) - P(X(u) = x_u) P(X(v) = x_v) \right|.$$

If the interactions are sufficiently weak, the graph will satisfy the Dobrushin-Shlosman condition (see e.g. [8]) and there will be exponential decay of correlations between vertices, i.e. $d_C(u, v) \leq \exp(-\alpha d(u, v))$ for some $\alpha > 0$.

The following theorem shows that by restricting the candidate neighborhoods of the GENERALRECON algorithm to those nodes with sufficiently high correlation, one can achieve a run-time of $O(dn^2 \log n)$.

Theorem 8 (Reconstruction with correlation decay). *Suppose that G and X satisfy the hypothesis of Theorem 5 and that for all $u, v \in V$, $d_C(u, v) \leq \exp(-\alpha d(u, v))$ and there exists some $\kappa > 0$ such that for all $(u, v) \in E$, $d_C(u, v) > \kappa$. Then for some constant $C = C(\alpha, \kappa, \epsilon, \delta) > 0$, if $k > Cd \log n$ then there exists an estimator $\widehat{G}(\underline{X})$ such that the probability of correct reconstruction is $P(G = \widehat{G}(\underline{X})) = 1 - o(1)$ and the algorithm runtime is $O(nd^{\frac{d \ln(4/\kappa)}{\alpha}} + dn^2 \ln n)$ with high probability.*

Proof. Denote the correlation neighborhood of a vertex v as $N_C(v) = \{u \in V : \widehat{d_C}(u, v) > \kappa/2\}$ where $\widehat{d_C}(u, v)$ is the empirical correlation of u and v. For large enough C, with high probability for all $v \in V$, we have that $N(v) \subseteq N_C(v) \subseteq \{u \in V : d(u, v) \leq \frac{\ln(4/\kappa)}{\alpha}\}$. Now, we have the estimate $| \{u \in V : d(u, v) \leq \frac{\ln(4/\kappa)}{\alpha}\} | \leq d^{\frac{\ln(4/\kappa)}{\alpha}}$, which is independent of n.

When reconstructing the neighborhood of a vertex v we modify GENERALRECON to only test candidate neighborhoods U and sets W which are subsets of $N_C(v)$. The algorithm restricted to the smaller range of possible neighborhoods correctly reconstructs the graph since the true neighborhood of a vertex is always in its correlation neighborhood. For each vertex v the total number of choices of candidate neighborhoods U and sets W the algorithm has to check is $O(d^{\frac{d \ln(4/\kappa)}{\alpha}})$, so the reconstruction algorithm takes $O(nd^{\frac{d \ln(4/\kappa)}{\alpha}})$ operations. It takes $O(dn^2 \ln n)$ operations to calculate all the correlations, which for large n dominates the run time. □

Acknowledgment. E.M. thanks Marek Biskup for helpful discussions on models with hidden variables.

References

1. Chow, C.K., Liu, C.N.: Approximating discrete probability distributions with dependence trees. IEEE Trans. Info. Theory IT-14, 462–467 (1968)
2. Chickering, D.: Learning Bayesian networks is NP-complete. In: Proceedings of AI and Statistics (1995)
3. Abbeel, P., Koller, D., Ng, A.: Learning factor graphs in polynomial time and sample complexity. Journal of Machine Learning Research 7, 1743–1788 (2006)
4. Santhanam, N., Wainwright, M.J.: Information-theoretic limits of graphical model selection in high dimensions (submitted, January 2008)
5. Wainwright, M.J., Ravikumar, P., Lafferty, J.D.: High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In: NIPS 2006, Vancouver, BC, Canada (2006)
6. Baldassi, C., Braunstein, A., Brunel, N., Zecchina, R.: Efficient supervised learning in networks with binary synapses; arXiv:0707.1295v1
7. Mahmoudi, H., Pagnani, A., Weigt, M., Zecchina, R.: Propagation of external and asynchronous dynamics in random Boolean networks; arXiv:0704.3406v1
8. Dobrushin, R.L., Shlosman, S.B.: Completely analytical Gibbs fields. In: Fritz, J., Jaffe, A., Szasz, D. (eds.) Statistical mechanics and dynamical systems, pp. 371–403. Birkhauser, Boston (1985)
9. Friedman, N.: Infering cellular networks using probalistic graphical models. In: Science (February 2004)
10. Kasif, S.: Bayes networks and graphical models in computational molecular biology and bioinformatics, survey of recent research (2007), `http://genomics10.bu.edu/bioinformatics/kasif/bayes-net.html`
11. Daskalakis, C., Mossel, E., Roch, S.: Optimal phylogenetic reconstruction. In: STOC 2006: Proceedings of the 38th Annual ACM Symposium on Theory of Computing, pp. 159–168. ACM, New York (2006)
12. Erdös, P.L., Steel, M.A., Székely, L.A., Warnow, T.A.: A few logs suffice to build (almost) all trees (part 1). Random Struct. Algor. 14(2), 153–184 (1999)
13. Mossel, E.: Distorted metrics on trees and phylogenetic forests. IEEE/ACM Trans. Comput. Bio. Bioinform. 4(1), 108–116 (2007)
14. Bresler, G., Mossel, E., Sly, A.: Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms; arXiv:0712.1402v1