Probability on Graphs: Techniques and Applications to Data Science

5 - Correlation decay

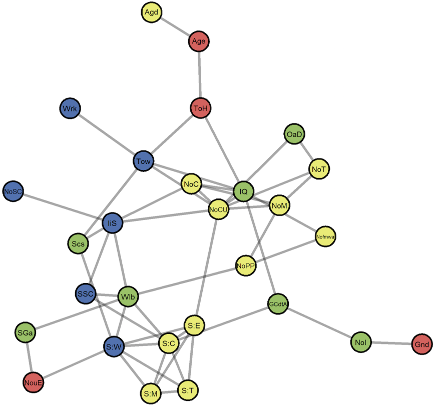
Sébastien Roch
UW–Madison
Mathematics

July 26, 2018

1. Correlation decay

2. Application: Reconstructing Markov random fields

An undirected graphical model



Recall: Ising model

Let $G = (V, E)$ be a finite, connected graph with maximum degree $\bar{\delta} = d$. Define $\mathcal{X} := A^V$ where $A = \{-1, +1\}$. Recall that the (ferromagnetic) Ising model on V with *inverse temperature* β is the probability distribution over *spin configurations* $\sigma \in \mathcal{X}$ given by

$$\mu_\beta(\sigma) := \frac{1}{\mathcal{Z}(\beta)} e^{-\beta \mathcal{H}(\sigma)},$$

where

$$\mathcal{H}(\sigma) := -\sum_{i \sim j} \sigma_i \sigma_j,$$

is the *Hamiltonian* and $\mathcal{Z}(\beta) := \sum_{\sigma \in \mathcal{X}} e^{-\beta \mathcal{H}(\sigma)}$.

Correlation decay

Spatial mixing: How much does the state at one vertex "influence" the state at a vertex far away?

There are many ways to measure this. Let $X \sim \mu_\beta$ on $G = (V, E)$. For $u, v \in V$, define

$$d_C(u, v) = \sum_{x_u, x_v \in S} |\mathbb{P}[X_u = x_u, X_v = x_v] - \mathbb{P}[X_u = x_u]\mathbb{P}[X_v = x_v]|$$

It can be shown in some cases that, when β is large enough, the measure above decays exponentially with the graph distance. Such a statement can be useful to analyze the behavior of Ising models. This is easier seen on an example.

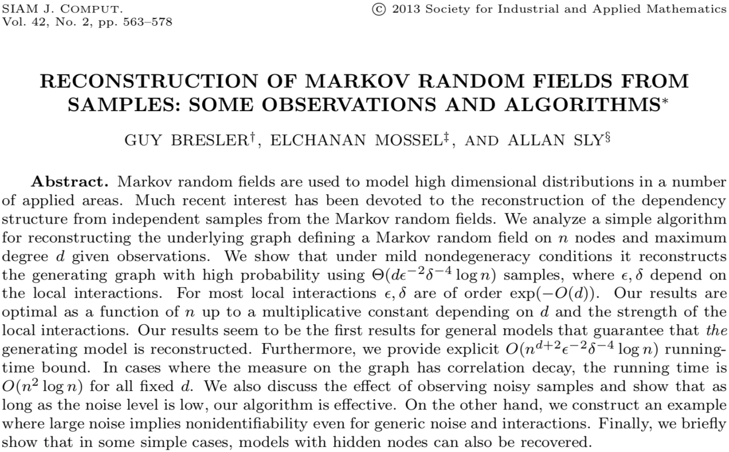
1 Correlation decay

2 Application: Reconstructing Markov random fields

Structure learning

Problem: Let $X^{(1)}, \ldots, X^{(k)}$ be i.i.d. $\sim \mu_\beta$ on an unknown graph $G = (V, E)$ with maximal degree $\bar{\delta} = d$. How to recover G from the samples $X^{(1)}, \ldots, X^{(k)}$?

Bresler et al. (2013)



SIAM J. COMPUT.
Vol. 42, No. 2, pp. 563–578

RECONSTRUCTION OF MARKOV RANDOM FIELDS FROM SAMPLES: SOME OBSERVATIONS AND ALGORITHMS*

GUY BRESLER†, ELCHANAN MOSSEL‡, AND ALLAN SLY§

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 nondegeneracy conditions it reconstructs the generating graph with high probability using $\Theta(d\epsilon^{-2}\delta^{-4}\log n)$ samples, where ϵ, δ depend on the local interactions. For most local interactions ϵ, δ are of order $\exp(-O(d))$. Our results are optimal as a function of n up to a multiplicative constant depending on d and the strength of the local interactions. Our results seem to be the first results for general models that guarantee that *the* generating model is reconstructed. Furthermore, we provide explicit $O(n^{d+2}\epsilon^{-2}\delta^{-4}\log n)$ 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. We also discuss the effect of observing noisy samples and 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 nonidentifiability even for generic noise and interactions. Finally, we briefly show that in some simple cases, models with hidden nodes can also be recovered.

Recall: Gibbs sampling

The single-site Glauber dynamics of the Ising model is the Markov chain on $\mathcal{X}$ which, at each time, selects a site $i \in V$ uniformly at random and updates the spin σ_i according to $\mu_\beta(\sigma)$ conditioned on agreeing with σ at all sites in $V \backslash \{i\}$. Specifically, for $\gamma \in \{-1, +1\}$, $i \in V$, and $\sigma \in \mathcal{X}$, let $\sigma^{i,\gamma}$ be the configuration σ with the state at i being set to γ. Then the transition matrix is

$$Q_\beta(\sigma, \sigma^{i,\gamma}) := \frac{1}{n} \cdot \frac{e^{\gamma\beta S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}} = \frac{1}{n}\left\{\frac{1}{2} + \frac{1}{2}\tanh(\gamma\beta S_i(\sigma))\right\},$$

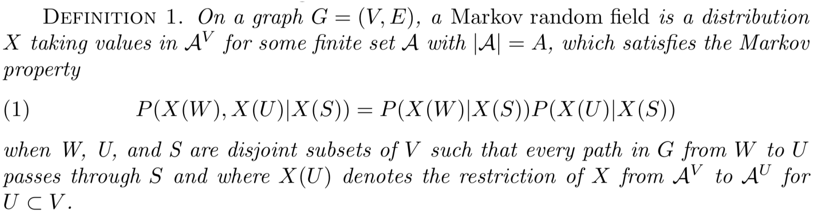
where

$$S_i(\sigma) := \sum_{j \sim i} \sigma_j.$$

All other transitions have probability 0. Recall that this chain is irreducible and reversible with respect to μ_β.

Markov property

Let $X \sim \mu_\beta$ on $G = (V, E)$. Then X satisfies the following.



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* $|\mathcal{A}| = A$, *which satisfies the Markov property*

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

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* $U \subset V$.

One of BMS's Main Results

THEOREM 3. *For an assignment $x_U = (x_{u_1}, \ldots, x_{u_l})$ and $y \in \mathcal{A}$, define*

$$x_U^i(y) = (x_{u_1}, \ldots, y, \ldots, x_{u_l})$$

to be the assignment obtained from x_U by replacing the ith element by y. Suppose there exist $\epsilon, \delta > 0$ such that the following condition holds: for all $v \in V$, if $N(v) = \{u_1, \ldots, u_l\}$, then for each $i, 1 \le i \le l$, and for any set $W \subset V - (\{v\} \cup N(v))$ with $|W| \le d$, there exist values $x_v, x_{u_1}, \ldots, x_{u_i}, \ldots, x_{u_l}, y \in \mathcal{A}$, and $x_W \in \mathcal{A}^{|W|}$ such that

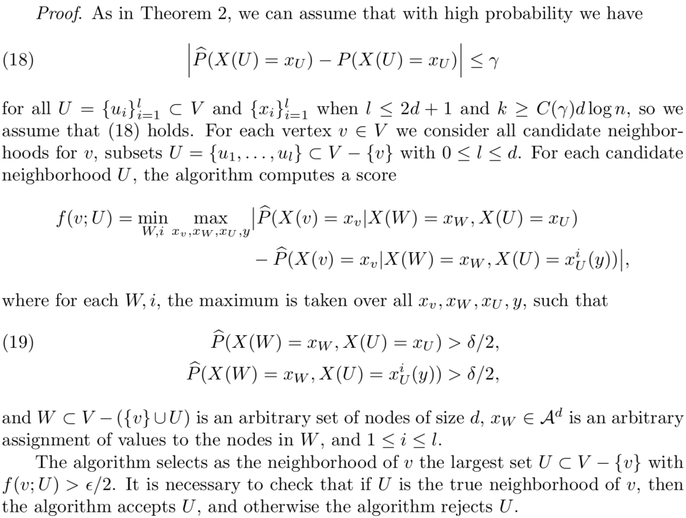
$$\begin{aligned}(16)\qquad & \big|P(X(v) = x_v | X(N(v)) = x_{N(v)}) \\ & \quad - P(X(v) = x_v | X(N(v)) = x^i_{N(v)}(y))\big| > \epsilon\end{aligned}$$

and

$$\begin{aligned}(17)\qquad & P(X(N(v)) = x_{N(v)}, X(W) = x_W) > \delta, \\ & P(X(N(v)) = x^i_{N(v)}(y), X(W) = x_W) > \delta.\end{aligned}$$

Then for some constant $C = C(\epsilon, \delta) > 0$, if $k > Cd \log n$, then there exists an estimator $\widehat{G}(\underline{X})$ such that the probability of correct reconstruction is $P(G = \widehat{G}(\underline{X})) = 1 - o(1)$. The estimator $\widehat{G}$ is computable in time $O(n^{2d+1} \log n)$.

Reconstructing neighborhoods



Proof. As in Theorem 2, we can assume that with high probability we have

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

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

$$\begin{aligned}f(v;U)=\min_{W,i}\max_{x_v,x_W,x_U,y}\big|&\widehat{P}(X(v)=x_v|X(W)=x_W,X(U)=x_U)\\&-\widehat{P}(X(v)=x_v|X(W)=x_W,X(U)=x_U^i(y))\big|,\end{aligned}$$

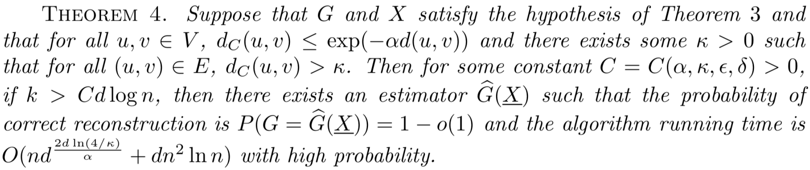
where for each W,i, the maximum is taken over all x_v,x_W,x_U,y, such that

$$\begin{aligned}\widehat{P}(X(W)=x_W,X(U)=x_U)&>\delta/2,\\\widehat{P}(X(W)=x_W,X(U)=x_U^i(y))&>\delta/2,\end{aligned}\tag{19}$$

and $W\subset V-(\{v\}\cup U)$ 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 $U\subset V-\{v\}$ 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.

A faster method under correlation decay



THEOREM 4. *Suppose that G and X satisfy the hypothesis of Theorem 3 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 running time is $O(nd^{\frac{2d\ln(4/\kappa)}{\alpha}} + dn^2 \ln n)$ with high probability.*

Cutting down the number of potential neighborhoods

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 the size of $|\{u \in V : d(u,v) \leq \frac{\ln(4/\kappa)}{\alpha}\}|$ is at most $d^{\frac{\ln(4/\kappa)}{\alpha}}$, which is independent of n.

When reconstructing the neighborhood of a vertex v we modify the algorithm in Theorem 3 to test only 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 with high probability since the true neighborhood of a vertex is 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 $O(d^{\frac{2d\ln(4/\kappa)}{\alpha}})$, so running the reconstruction algorithm takes $O(nd^{\frac{2d\ln(4/\kappa)}{\alpha}})$ operations. It takes $O(dn^2 \ln n)$ operations to calculate all the correlations, which for large n dominates the running time. □

Go deeper

More details and results in:

- Bresler, Mossel, Sly, *Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms*, SIAM J. Comput., 42(2):563–578.
- Bresler, *Efficiently Learning Ising Models on Arbitrary Graphs*, STOC 2015.