Rapidly Mixing Gibbs Sampling for a Class of Factor Graphs Using Hierarchy Width

Christopher De Sa, Ce Zhang, Kunle Olukotun, and Chris Ré
cdesa@stanford.edu, czhang@cs.stanford.edu, kunle@stanford.edu, chrismre@cs.stanford.edu
Departments of Electrical Engineering and Computer Science, Stanford University

Overview

Everyone uses Gibbs sampling!

- ▷ De facto Markov Chain Monte Carlo method for inference.
- ▷ Works very well in practice.
- ▷ Used by many systems such as Factorie, OpenBugs, PGibbs, and DeepDive — including competition-winners.

DeepDive

But it's hard to tell when Gibbs sampling will work!

- ▷ Standard metric is ***mixing time***, the amount of time needed to produce samples that are "close" to the true distribution.
- ▷ ***Finding the mixing time is hard*** — there's little theory.

Our contribution: fast mixing with hierarchy width

- ▷ Introduce a new factor graph width: the ***hierarchy width***.
- ▷ Hierarchy width is a structural property of the factor graph.
- ▷ Bounding the hierarchy width is a sufficient condition to ensure that Gibbs sampling will mix in polynomial time.
- ▷ This gives us new understanding of a class of factor graphs for which ***Gibbs sampling is guaranteed to be feasible***.

Problem Setup

Gibbs sampling: Sample from distribution π over variables V

Require: Initial state X_i for $i \in V$, number of samples T.
for $t = 0$ **to** $T-1$ **do**
  Select i_t uniformly from V.
  Resample X_{i_t} conditionally from π given $X_{V \setminus \{i_t\}}$.
  Output sample $z_t \leftarrow X$.
end for

We study Gibbs sampling on discrete-valued ***factor graphs***. A factor graph is a graphical model over a set of variables V and factors Φ that has distribution

$$\pi(I) = \frac{1}{Z} \exp\left(\sum_{\phi \in \Phi} \phi(I) \right)$$

where I is a world — an assignment of a value to each variable in V — and Z is the constant required to make π a distribution.

We focus on bounding the *mixing time*, the first time t at which the estimated distribution μ_t is close to the true distribution π.

$$t_{\text{mix}} = \min\left\{ t : \max_{A \subset \Omega} |\mu_t(A) - \pi(A)| \le \frac{1}{4} \right\}.$$

Hierarchy Width and Rapid Mixing

The ***hierarchy width*** $\mathsf{hw}(G)$ of a factor graph G is defined such that, for any *connected* factor graph $G = \langle V, \Phi \rangle$,

$$\mathsf{hw}(G) = 1 + \min_{\phi^* \in \Phi} \mathsf{hw}(\langle V, \Phi - \{\phi^*\} \rangle),$$

and for any *disconnected* factor graph G with connected components $G_1, G_2, \ldots$,

$$\mathsf{hw}(G) = \max_i \mathsf{hw}(G_i).$$

All factor graphs G with no factors have

$$\mathsf{hw}(\langle V, \emptyset \rangle) = 0.$$

Main Theorem: Bounding the mixing time.

Let $G = \langle V, \Phi \rangle$ be a factor graph with n variables, at most s states per variable, e factors, and hierarchy width h. If we let

$$M = \max_{\phi \in \Phi} \left(\max_I \phi(I) - \min_I \phi(I) \right),$$

then we can bound the mixing time of Gibbs sampling on G with

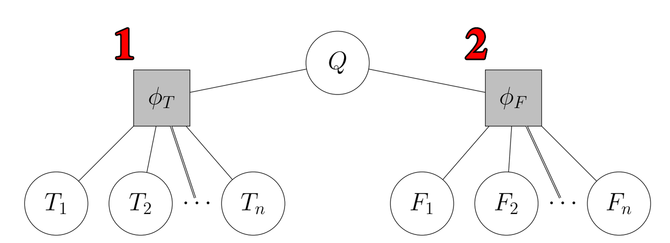
$$t_{\text{mix}} \le (\log(4) + n\log(s) + eM)\, n \exp(3hM).$$

In particular, if $hM = O(\log n)$, then Gibbs sampling mixes in polynomial time.

Hierarchy Width Examples

Intuitively, we can think of labeling each factor with a positive integer, its *level in the hierarchy*. For two factors F and G to have the same level, they must only interact through their superiors: every path from F to G must pass through a factor with a smaller label. The hierarchy width is the minimum value, across all labellings, of the largest label. Here are some examples (labels in red).

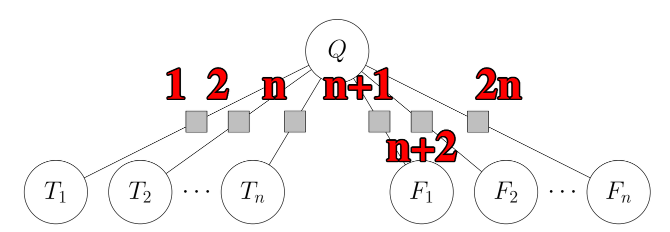
Example: Voting model (logical).



This model has only two (large) factors, which can't have the same label because they are adjacent. Therefore, its hierarchy width is $\mathsf{hw}(G) = 2$.

- ▷ Actually mixes in $O(n \log n)$ time.

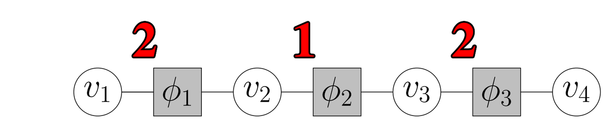
Example: Voting model (linear).



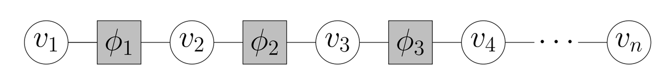
This model has $2n$ factors, all of which are adjacent. Therefore, its hierarchy width is $\mathsf{hw}(G) = 2n$.

- ▷ Actually mixes in $\exp(\Omega(n))$ time.
- ▷ This means Gibbs is *infeasible*.

Example: Path graph.



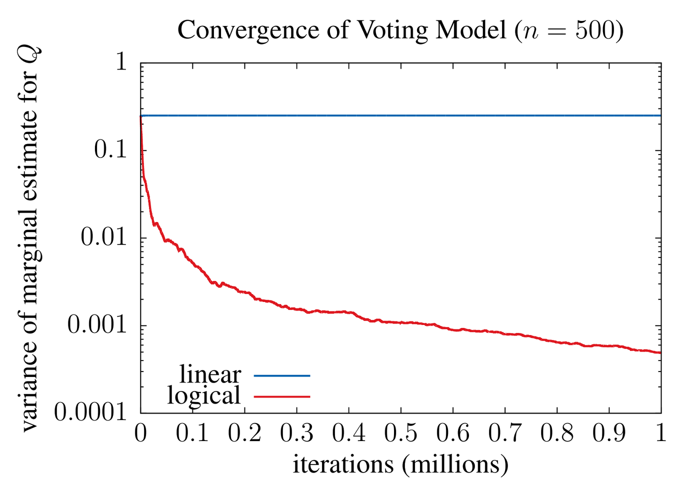
Removing factor ϕ_2 disconnects the graph, so we can label both ϕ_1 and ϕ_3 as 2. So, this graph has $\mathsf{hw}(G) = 2$.

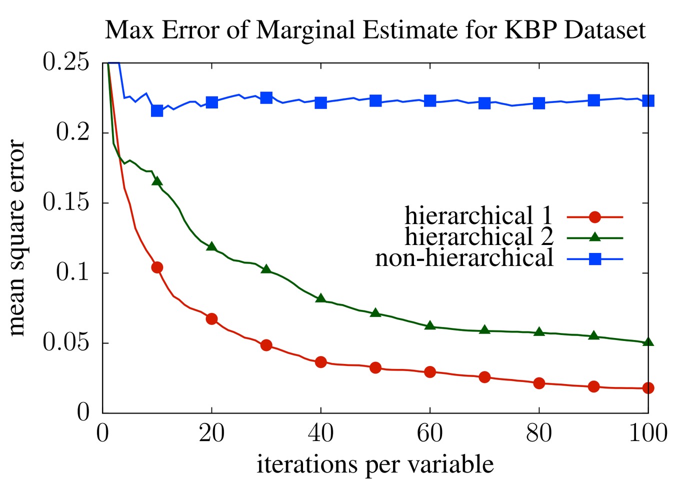


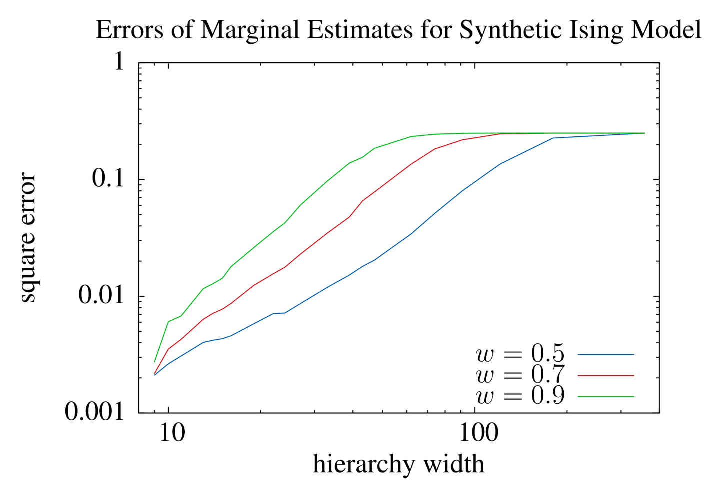
In general, the path graph has hierarchy width $\mathsf{hw}(G) = \lceil \log_2 n \rceil$.

- ▷ Guaranteed to mix in polynomial time.

Experiments







The first plot shows that, of the two voting models, the ***bounded-hierarchy-width model has lower error***. The second plot shows the same thing for templates on a real dataset — in particular, the model in Hierarchical 2 was used as part of a ***competition-winning system*** (TAC KBP '14). The third plot shows, for an ensemble of synthetic Ising models, how ***error varies with hierarchy width***.

Facts about Hierarchy Width

One of the useful properties of the hierarchy width is that, for any fixed k, computing whether a graph G has ***hierarchy width*** $\mathsf{hw}(G) \le k$ ***can be done in time polynomial in the size of G***.

- ▷ This is similar to many other useful graph widths.

Hierarchy width is an ***upper bound*** on the commonly-used graph metric, ***hypertree width***. Hierarchy width is also an upper bound on the maximum degree of a variable in the graph.

Hierarchical Templates

A factor graph template is an abstract model that can be *instantiated* on a dataset to produce a factor graph. They are commonly used to construct models, including in state-of-the-art systems.

Our contribution: we introduce ***hierarchical templates***, which when instantiated on any dataset produce models that are guaranteed to ***mix in polynomial time***.

A template consists of template factors like

$$\phi\left(\mathsf{TweetedAbout}(\hat{x}, y), \mathsf{IsPopular}(\hat{x})\right).$$

We call $\hat{x}$ a *head symbol*, and y a *body symbol*. (Details of template instantiation appear in the paper.)

A template factor is ***hierarchical*** if all its head symbols appear in the same order in each of its terms. (In particular, our example above is hierarchical.) A template is hierarchical if all its factors are hierarchical.

Hierarchical templates always mix fast.

The hierarchy width of a template instance is no greater than the number of template factors in the template. Combining this with our other result, **hierarchical templates produce models that always mix in polynomial time!**

Here is an outline of our results: