Improving Mixing Time in Gibbs Sampling

Samy

MLD Journal Club

February 20, 2014

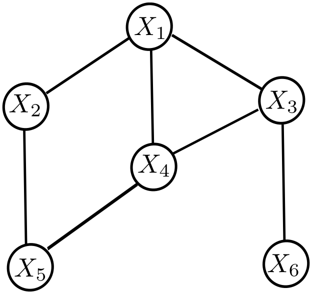
Papers

From Fields to Trees
Hamze, de Freitas 2004

Covariance Structure of the Gibbs Sampler ...
Liu, Wong, Kong 1994

Markov Chain Strategies in Scientific Computing
Jun Liu, 2001

The Problem



Given: Graphical Model with known Parameters.

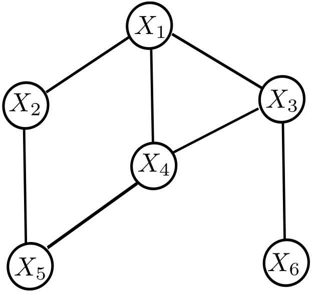
Goal: Estimate $\mathbb{E}[h(X)]$
E.g.: $\mathbb{1}_A(x)$, $\sum_i X_i$

Generate Samples: $X^{(1)}, X^{(2)}, \ldots X^{(N)}$

Empirical Estimator

$$\mu_0 = \frac{1}{N} \sum_{i=1}^{N} h(X^{(i)})$$

Gibbs Sampling



Current Sample: $X^{(t)}$

$X_1^{(t+1)} \mid X_{2,3,4,5,6}^{(t)}$

$X_2^{(t+1)} \mid X_1^{(t+1)}, X_{3,4,5,6}^{(t)}$

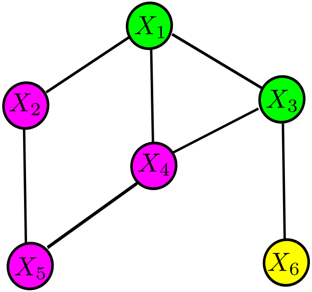
$\dots$

$X_6^{(t+1)} \mid X_{1,2,3,4,5}^{(t+1)}$

Next Sample: $X^{(t+1)}$

$X^{(1)} \to X^{(2)} \to \dots$ is a Markov Chain with eqlb$^{\text{m}}$ distribution P.

Blocked Gibbs Sampling



Current Sample: $X^{(t)}$

$X_{1,3}^{(t+1)} \mid X_{2,4,5,6}^{(t)}$

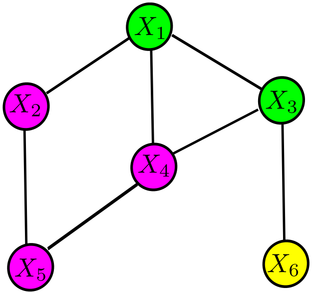
$X_{2,4,5}^{(t+1)} \mid X_{1,3}^{(t+1)}, X_6^{(t)}$

$X_6^{(t+1)} \mid X_{1,2,3,4,5}^{(t+1)}$

Next Sample: $X^{(t+1)}$

$X^{(1)} \rightarrow X^{(2)} \rightarrow \ldots$ is a Markov Chain with eqlb$^{\text{m}}$ distribution P.

Blocked Gibbs Sampling

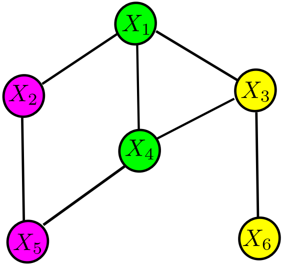


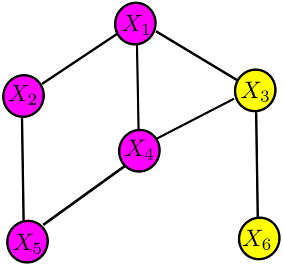
Why Blocked Gibbs ?

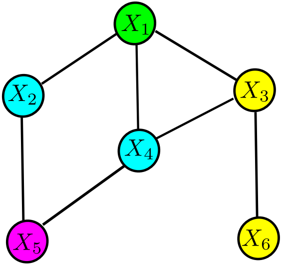
Sampling from conditional distribution is difficult.

BUT, Chain mixes faster $\implies$ better samples.

How to Block (Partition) the Variables?







Outline

MCMC Theory (Jun Liu, 2001)

$$X^{(1)} \to X^{(2)} \to \cdots \to X^{(t)} \to X^{(t+1)} \to \ldots$$

$$\pi \leftarrow \text{Equilibrium Distribution}$$

$$L_0(\pi) = \{h : \Omega \to \mathbb{R} : \quad \mathbb{E}_\pi h(X) = 0, \quad \mathbb{V}_\pi h(X) < \infty\}$$
$$\langle h, g \rangle = \mathrm{Covar}_\pi(h(X), g(X))$$
$\big(L_0(\pi), \langle \cdot, \cdot \rangle\big)$ is a Hilbert Space.

$$\|h\|^2 = \langle h, h \rangle = \mathrm{Covar}_\pi(h(X), h(X)) = \mathbb{V}_\pi(h(X))$$

Forward Operator

Define $F : L_0(\pi) \to L_0(\pi)$, $[Fh](z) = \mathbb{E}[h(X^{(1)})|X^{(0)} = z]$

Norm of operator: $\|F\| = \sup_{\|h\| \leq 1} \|Fh\|$.

$$\begin{aligned}\|Fh\|^2 &= \mathbb{V}_{X^{(0)} \sim \pi}(\mathbb{E}[t(X^{(1)})|X^{(0)}]) \\ &\leq \mathbb{V}_{X^{(1)} \sim \pi}(h(X^{(1)})) = \|h\|^2\end{aligned}$$

Hence, $\|F\| \leq 1$.

If MC is reversible, F is self-adjoint $\implies \|F^n\| = \|F\|^n$.

Convergence of a MC

Rate of Convergence: $\|F\|$

$\|F\| = \sup_{f,g} \text{Corr}(f(X^{(t+1)}), g(X^{(t)}))$

Theorem: Let $X^{(0)} \sim P^{(0)}$. $P^{(n)}, \mathbb{E}^{(n)}$: n$^{\text{th}}$ step evolution.

$$|\mathbb{E}^{(n)}h(X) - \mathbb{E}_{\pi}h(X)| \leq C \, \|F\|^n \, \|h\|.$$

Finite State space: $\|F\| \equiv 2^{\text{nd}}$ eigval of Transition Matrix.

Rao-Blackwellisation

$(X, Y) \sim P$. Samples:
$\{(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots (X^{(N)}, Y^{(N)})\}$

Goal: Estimate $\mathbb{E}_P[h(X)]$

Empirical Estimator $\quad \mu_0 = \frac{1}{N} \sum_{i=1}^{N} h(X^{(i)})$

Rao-Blackwellised Estimator $\quad \mu_{rb} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[h(X) | Y^{(i)}\right]$

Rao-Blackwellisation

$(X, Y) \sim P$. Samples:
$\{(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots (X^{(N)}, Y^{(N)})\}$

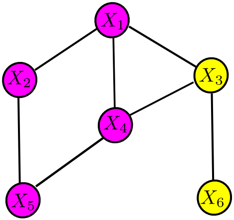
Independent samples: RB Estimator is better.

2-Stage Gibbs: RB Estimator is better (LKW 1994).

$$\mathbb{V}(\mu_0) - \mathbb{V}(\mu_{rb}) \leq \frac{\mathbb{V}(h(X))}{n^2} \Big(n(1 - \|F\|) + 2(n-1)(\|F\| - \|F\|^2) \\ \ldots + (\|F\|^{n-1} - \|F\|^n) \Big)$$

Outline

Primary Idea



Blocked Gibbs Sampler:

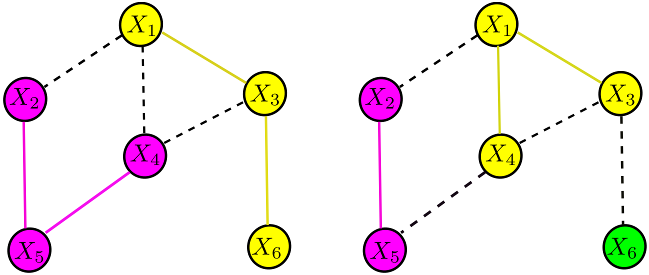
Correlated Partitions $\Longrightarrow$
Successive Samples Correlated $\Longrightarrow$
Large $\|F\|$

Blocked Gibbs : $\|F\|$ is determined by the partitions.

Partition the Graph so that $\|F\|$ is small !

Is this the best strategy ?

Tree Partitions



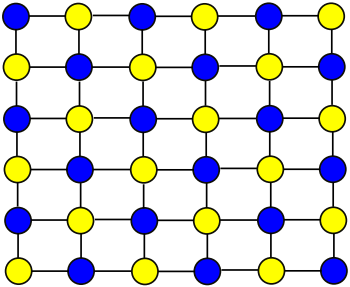
Lets focus on Tree Partitions.

Computing conditional distribution on trees is easy.

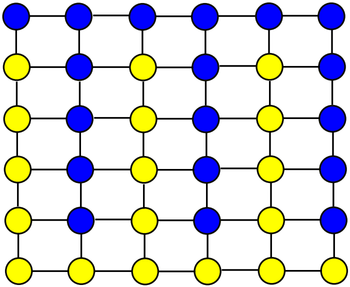
Best Tree Partition ? Not easy.

Outline

Grid Graph

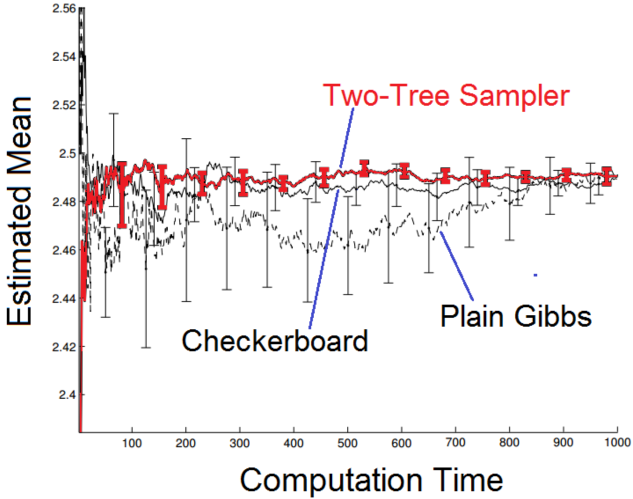


Checkerboard



Two-Trees

Results

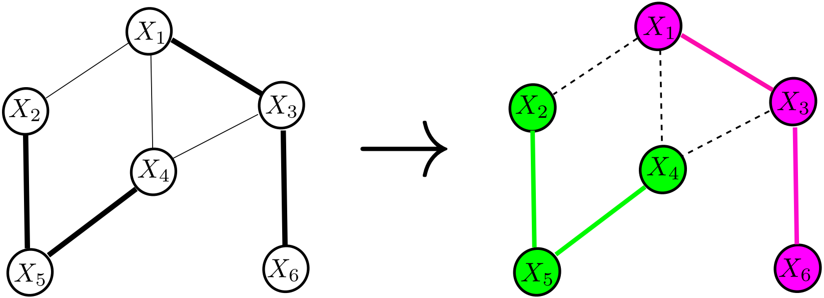


H/deF use a a Rao-Blackwellised Estimator.

Outline

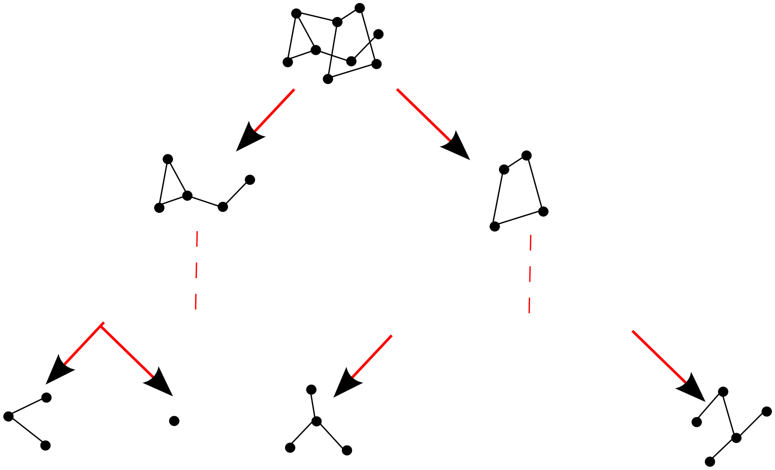
My 701 Course Project

Incorporate Correlations via Weights (thanks Geoff !)



Partitioning into K trees is NP-Hard ☹ (credit: Elara W)

Greedy Splitting & Recombination



My 701 Course Project

Questions

1. What measure of Correlation for Weights ?
2. How to Greedy Split ? (Min-Cut, N-Min-Cut, Max-Cut)
   How optimal is this ?
3. Guarantees about $\|F\|$?
4. On what kind of graphs can we expect to win ?