Georgia Institute of Technology

CSIP Center for Signal & Image Processing

ECE/ML/CS/ISYE 8803

Approximate Inference in Graphical Models

Module 7: Part B
Markov Chain Monte Carlo (MCMC)

Faramarz Fekri

Center for Signal and Information Processing

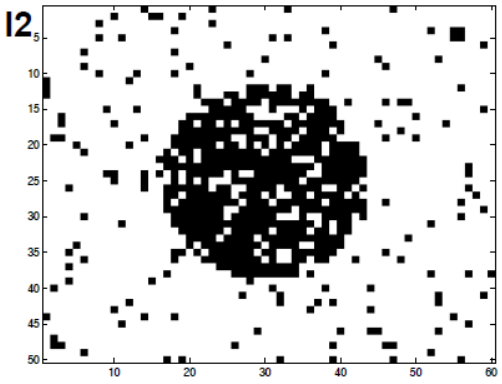
Overview

- Motivation to Markov Chain Monte Carlo (MCMC) sampling
- Introduction to Markov Chain
  - Time-invariant regular MC
    - Stationary distribution of MC
- MCMC Sampling
  - Gibbs sampling
    - Example
    - Mixing property
    - Practical aspects of Gibbs sampling
  - Rao-Blackwellized Particles

Read Chapter 12 of K&F

Saying Again: Why Sampling

- Exact and variational inference techniques focus on obtaining the entire posterior distribution $P(X_i|e)$.
- Often we only want to take expectations via posterior dist.; rather than full characterization of the posterior dist.
- Or we want to only compute the probability of an event.
- Sometime, we also want to see typical points from a distribution :



Thus far on Sampling

- Direct sampling
  - Hard to get rare events in high-dimensional spaces
  - Infeasible for MRFs, unless we know the normalizer Z
- Rejection sampling, Importance sampling
  - Do not work well if the proposal Q(x) is very different from P(x)
  - Constructing Q(x) similar to P(x) can prove to be difficult
    - Making a good proposal usually requires knowledge of the analytic form of P(x) – which is often not available.

- Intuition: instead of a fixed proposal Q(x), what if we could use an adaptive proposal?

Markov Chain Monte Carlo (MCMC)

- Limitations of importance sampling
  - An evidence node affects the sampling only for nodes that are its descendants.
  - The effect on nodes that are non-descendants is accounted for only by the weights w's.
  - What if much of the evidence is at the leaves of the network?
    - We are essentially sampling from the prior distribution P(**X**), which is often very far from the desired posterior P(**X**|**E**=**e**).

- An alternative approach to sampling
  - Define a sampling process that is guaranteed to converge to taking samples from the posterior distribution of interest P(X|E=e)
    - Generate samples from the sampling process
    - Estimate f(X) from the samples

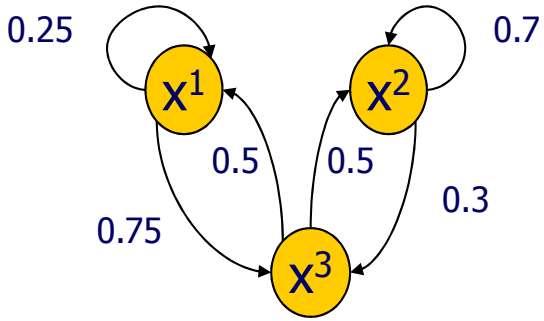
Introduction to Markov Chain

- A Markov chain consists of
  - A state space Val(**X**)
  - Transition probability T(x→x′) of going from state x to x′
  - Transition probabilities (transition kernels) are assumed independent of time.
- Distribution over subsequent states is defined as

Distribution over the "next" state

Distribution over the "current" state

$$P^{(t+1)}(X^{(t+1)} = x') = \sum_{x \in Val(X)} P^{(t)}(X^{(t)} = x) T(x \to x')$$



Example of a 3 states Markov Chain

- Note: the r.v.s $x^{(i)}$ can be vectors
  - We define $x^{(t)}$ to be the t-th sample of all variables in a graphical model
  - $X^{(t)}$ represents the entire state of the graphical model at time t

Markov Chain (MC)

- To understand Markov Chains, we define a few concepts:
  - Probability distributions over states: $P^{(t)}(x)$ is a distribution over the state of the system X=x, at time t.
  - When dealing with MC, we don't think of the system as being in one state, but as having a distribution over states.
  - For Graphical Models (GM), note that x represents all GM variables

Markov Chain: Stationary Distribution

Distribution over the "next" state

Distribution over the "current" state

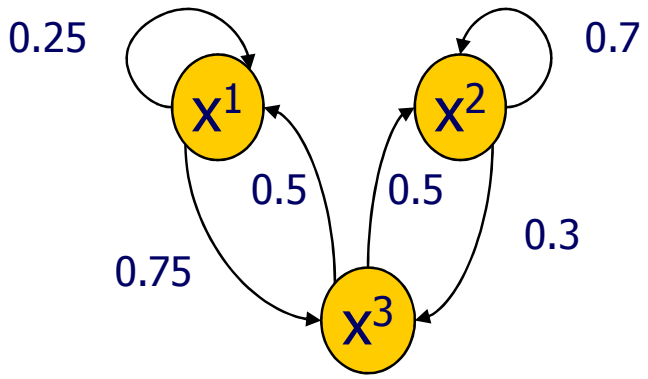
$$P^{(t+1)}(X^{(t+1)} = x') = \sum_{x \in Val(X)} P^{(t)}(X^{(t)} = x) T(x \to x')$$

- A distribution $\pi(X)$ is a stationary distribution for a Markov chain T if it satisfies

$$\pi(X = x') = \sum_{x \in Val(X)} \pi(X = x) T(x \to x')$$

$$\begin{aligned}
\pi(x^1) &= 0.25\pi(x^1) + 0.5\pi(x^3) \\
\pi(x^2) &= 0.7\pi(x^2) + 0.5\pi(x^3) \\
\pi(x^3) &= 0.75\pi(x^1) + 0.3\pi(x^2)
\end{aligned}$$

$$\begin{aligned}
\pi(x^1) &= 0.2 \\
\pi(x^2) &= 0.5 \\
\pi(x^3) &= 0.3
\end{aligned}$$



-Lee

Regular (Ergodic) Markov Chain

- A Markov chain is regular if there is k such that for every x, x′∈Val(X), the probability of traveling from x to x′ in exactly k steps in time is greater than zero.

> Theorem: A finite state Markov chain T has a unique stationary distribution if and only if it is regular

- A regular Markov Chain will forget its starting distribution, and will converge to its stationary distribution (after sufficient steps/walks) no matter which initial distribution was used to start the chain.

Goal: Design a Markov chain whose stationary distribution is the target distribution we wish to sample, e.g., P(**X**|**e**)

MCMC Sampling

- Stationary distributions in MC are of great importance in MCMC sampling.
- Key idea: Construct a Markov chain whose stationary distribution is the target distribution we wish to generate samples from.
- Sampling process: random walk in the Markov chain.

- Want to sample from $P(X)$, start with a random initial vector X.
  - X^t: X at time step t
  - X^t transition to X^{t+1} with probability
    - $Q(X^{t+1}|X^t, \dots, X^1) = T\ (X^t \rightarrow X^{t+1})$
  - The stationary distribution of $T\ (X^t \rightarrow X^{t+1})$ is our $P(X)$

- Run for an initial M samples (burn-in time) until the chain converges/ mixes to the stationary distribution. Then collect N sample as x_i

- Key issues: Designing the transition kernel, and diagnose convergence.

Gibbs Sampling (GS) of P(X)

- Let $\mathcal{X}$ be the collection of all variables in P(X)

- States of MC:
- we define the states of the Markov chain to be instantiations **x** to $\mathcal{X}$
- Transition probability $T(x \rightarrow x')$
  - $T = T_1 \cdot \ldots \cdot T_k$
  - For each variable X_i, let $\mathbf{X}_{-i}$ be $\mathbf{X}$-$\{X_i\}$. Say that $\mathbf{X}_{-i} = \mathbf{x}_{-i}$.
    $T_i = T((\mathbf{x}_{-i}, x_i) \rightarrow (\mathbf{x}_{-i}, x'_i)) = P(x'_i | \mathbf{x}_{-i})$
- Claim: P(X) is a stationary distribution to the MC formed by $T(x \rightarrow x')$

$$P(x_i' \mid \mathbf{x}_{-i}) = \frac{P(x_i', \mathbf{x}_{-i})}{\sum_{x_i'} P(x_i', \mathbf{x}_{-i})}$$

- Gibbs-sampling Markov chain is regular if:
  - Bayesian networks: all CPDs are strictly positive
  - Markov networks: all clique potentials are strictly positive

Gibbs Sampling of P(X)

Protocol of Gibbs Sampling of P(X)

- We have variables set $X = \{X_1, \dots, X_K\}$ variables in a GM.
- At each step, one variable X_i is selected (at random or some fixed sequence), denote the remaining variables as X_{-i}, and its current value as x_{-i}^t
- Compute the conditional distribution $P(X_i \mid x_{-i}^t)$
- A value x_i^t is sampled from this distribution
- This sample x_i^t replaces the previous sampled value of X_i in X

- Gibbs Sampling is an MCMC algorithm that samples each random variable of a graphical model, one at a time.
  - GS is a special case of the Metropolis-Hasting algorithm (will see)
- We will see that a very special transition kernel T(x→x′) used in GS that works nicely with Markov blanket in graphical models.

Gibbs Sampling of P(X|E=e)

- Wish to define a chain for which **P(X|e)** is the stationary distribution.
- Define the states of the Markov chain to be instantiations **x** to $\mathcal{X}$ – **E**.
- Assume that **P(X | e) = P_Φ** for some set of factors Φ that are defined by reducing the original factors in our (Directed/undirected) graphical model by the evidence **E=e**.
- Let x[m] be an instance that is consistent with evidence E=e.
- Define x[m](**X**-Xi) as x[m] excluding variable Xi.

Protocol for Gibbs sampling of P(X|E=e):

- Set x[m]=x[m-1]
- For each variable $X_i \in$ **X-E**
  - Set $\mathbf{x}_{-i}$ = x[m](**X**-X_i)
  - Sample from $P(X_i \mid \mathbf{x}_{-i})$
  - Save sample x[m](X_i) = sampled value
- Return x[m] (a new sample is generated)

How do we evaluate P(Xi | x_{-i}) in GS?

let $\mathbf{x}_{-i} = \{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n\}$

$$P_\Phi(\boldsymbol{X}) = \frac{1}{Z}\prod_j \phi_j(\boldsymbol{D}_j) = \frac{1}{Z}\prod_{j\,:\,X_i\in\boldsymbol{D}_j}\phi_j(\boldsymbol{D}_j)\prod_{j\,:\,X_i\notin\boldsymbol{D}_j}\phi_j(\boldsymbol{D}_j)$$

(Both BN and MRF)

Let $\boldsymbol{x}_{j,-i}$ denote the assignment in $\boldsymbol{x}_{-i}$ to $\boldsymbol{D}_j - \{X_i\}$, noting that when $X_i \notin \boldsymbol{D}_j$, $\boldsymbol{x}_{j,-i}$ is a full assignment to $\boldsymbol{D}_j$.

$$\begin{aligned}
P(x_i' \mid \boldsymbol{x}_{-i}) &= \frac{P(x_i', \boldsymbol{x}_{-i})}{\sum_{x_i''} P(x_i'', \boldsymbol{x}_{-i})} \\
&= \frac{\frac{1}{Z}\prod_{\boldsymbol{D}_j\ni X_i}\phi_j(x_i', \boldsymbol{x}_{j,-i})\prod_{\boldsymbol{D}_j\not\ni X_i}\phi_j(x_i', \boldsymbol{x}_{j,-i})}{\frac{1}{Z}\sum_{x_i''}\prod_{\boldsymbol{D}_j\ni X_i}\phi_j(x_i'', \boldsymbol{x}_{j,-i})\prod_{\boldsymbol{D}_j\not\ni X_i}\phi_j(x_i'', \boldsymbol{x}_{j,-i})} \\
&= \frac{\prod_{\boldsymbol{D}_j\ni X_i}\phi_j(x_i', \boldsymbol{x}_{j,-i})\prod_{\boldsymbol{D}_j\not\ni X_i}\phi_j(\boldsymbol{x}_{j,-i})}{\sum_{x_i''}\prod_{\boldsymbol{D}_j\ni X_i}\phi_j(x_i'', \boldsymbol{x}_{j,-i})\prod_{\boldsymbol{D}_j\not\ni X_i}\phi_j(\boldsymbol{x}_{j,-i})} \\
&= \frac{\prod_{\boldsymbol{D}_j\ni X_i}\phi_j(x_i', \boldsymbol{x}_{j,-i})}{\sum_{x_i''}\prod_{\boldsymbol{D}_j\ni X_i}\phi_j(x_i'', \boldsymbol{x}_{j,-i})}.
\end{aligned}$$

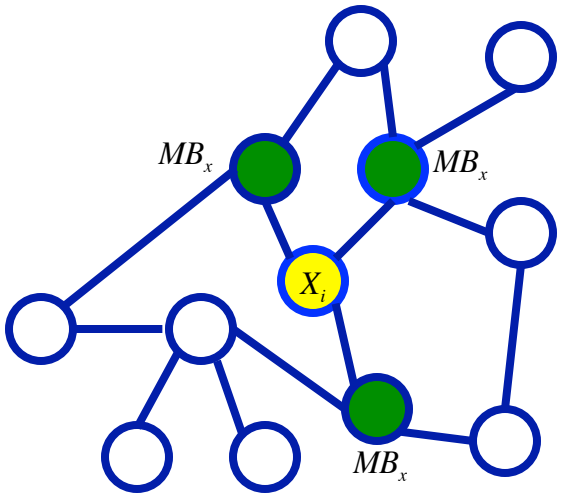
Remark: In GM, P(Xi | $\mathbf{x}_{-i}$) is easily computed via Markov blanket.

Gibbs Sampling in Graphical Models

A system with K variables:

- $X=x^0$
- For t=1 to N

$$
\begin{aligned}
x_1^{(t+1)} &\sim P(x_1 \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\
x_2^{(t+1)} &\sim P(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\
x_3^{(t+1)} &\sim P(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_K^{(t)}), \\
&\vdots \\
x_K^{(t+1)}
\end{aligned}
$$

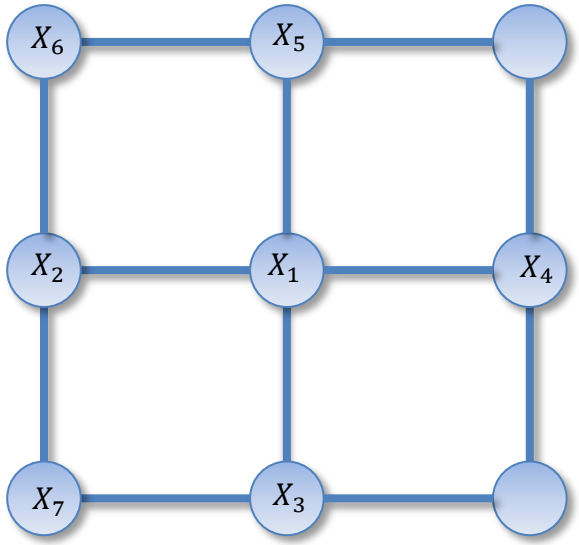


For a BN, the Markov Blanket of X_i is the set containing its parents, children, and co-parents

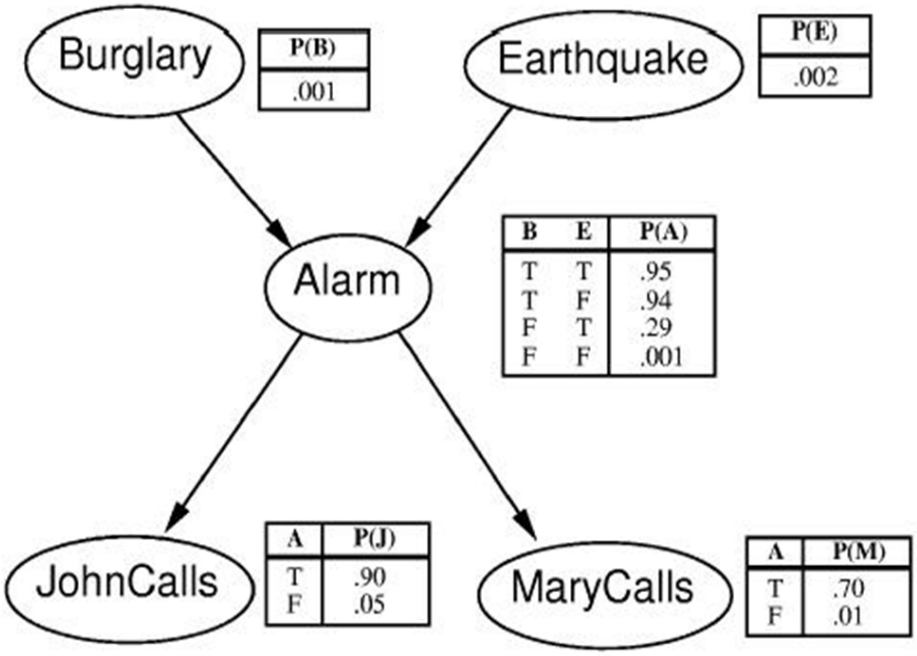
- Variations of GS:
  - Randomly pick a variable to sample
  - Sample block by block

Example: Gibbs Sampling of MRF

- Pairwise Markov random field
  - $P(X_1, \dots, X_k) \propto \exp\left(\sum_{ij} \theta_{ij} X_i X_j + \sum_i \theta_i X_i\right)$
  - In the Gibbs sampling step, we need conditional $P(X_1 | X_2, \dots X_k)$
  - $P(X_1 | X_2, \dots X_k) \propto$
    $\exp(\theta_{12} X_1 X_2 + \theta_{13} X_1 X_3 + \theta_{14} X_1 X_4 + \theta_{15} X_1 X_5 + \theta_1 X_1)$



Example: Gibbs Sampling of BN (I)

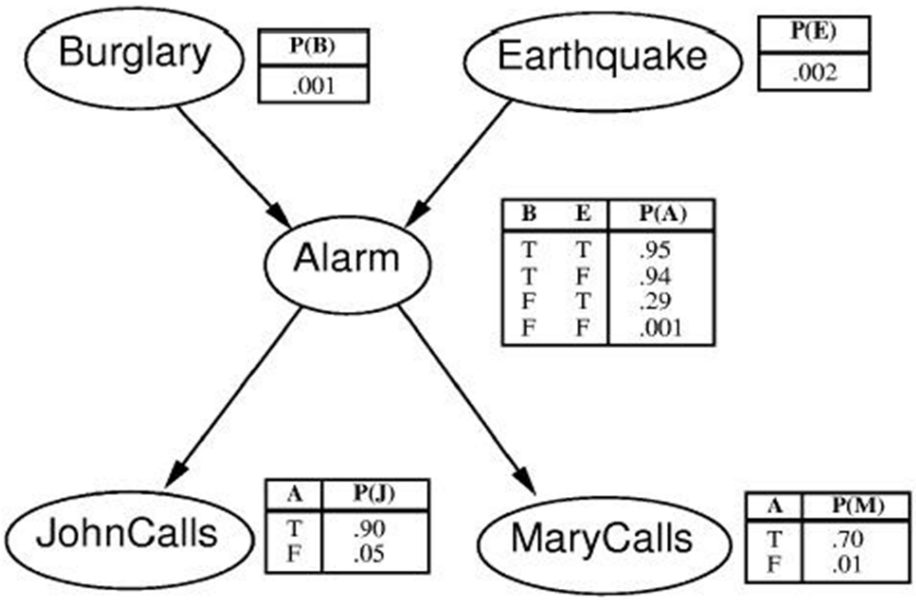


t	B	E	A	J	M
0	F	F	F	F	F
1					
2					
3					
4					

- Consider the alarm network
  - Assume we sample variables in the order B,E,A,J,M
  - Initialize all variables at t = 0 to False

-Xing

Example: Gibbs Sampling of BN (II)



t	B	E	A	J	M
0	F	F	F	F	F
1	F				
2					
3					
4					

- Sampling P(B|A,E) at t = 1: Using Bayes Rule,

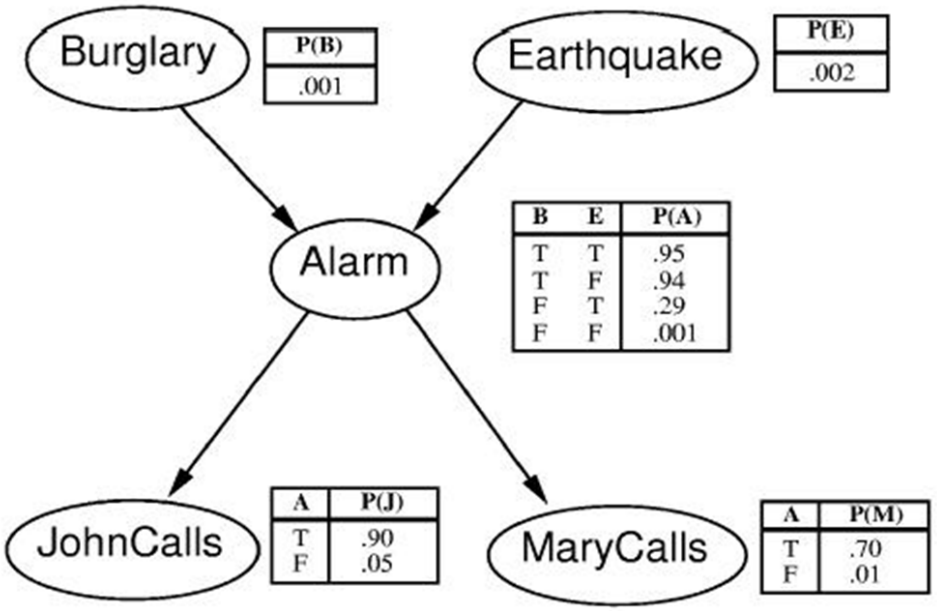
$$P(B \mid A, E) \propto P(A \mid B, E)P(B)$$

- (A,E) = (F,F), so we compute the following, and sample B = F

$$P(B = T \mid A = F, E = F) \propto (0.06)(0.01) = 0.0006$$

$$P(B = F \mid A = F, E = F) \propto (0.999)(0.999) = 0.9980$$

Example: Gibbs Sampling of BN (III)



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T			
2					
3					
4					

- Sampling P(E|A,B): Using Bayes Rule,

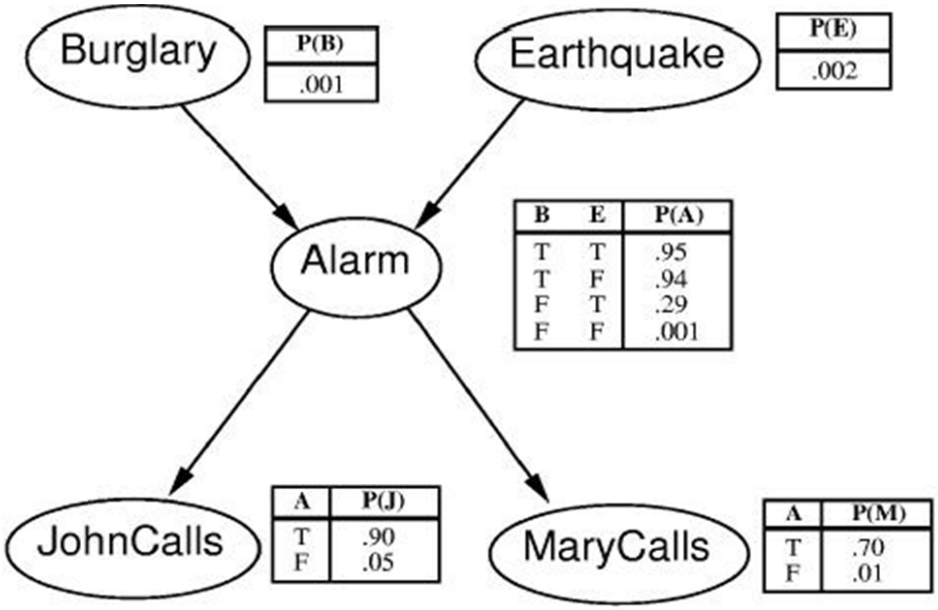
$$P(E \mid A, B) \propto P(A \mid B, E)P(E)$$

- (A,B) = (F,F), so we compute the following, and sample E = T

$$P(E = T \mid A = F, B = F) \propto (0.71)(0.02) = 0.0142$$

$$P(E = F \mid A = F, B = F) \propto (0.999)(0.998) = 0.9970$$

Example: Gibbs Sampling of BN (IV)



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F		
2					
3					
4					

- Sampling P(A|B,E,J,M): Using Bayes Rule,

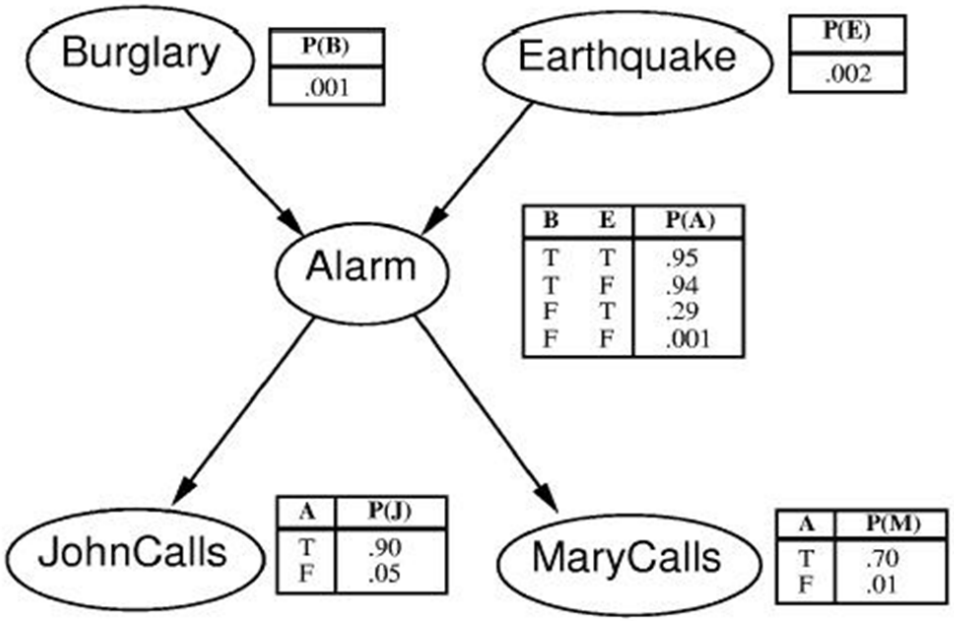
$$P(A \mid B, E, J, M) \propto P(J \mid A) P(M \mid A) P(A \mid B, E)$$

- (B,E,J,M) = (F,T,F,F), so we compute the following, and sample A = F

$$P(A = T \mid B = F, E = T, J = F, M = F) \propto (0.1)(0.3)(0.29) = 0.0087$$

$$P(A = F \mid B = F, E = T, J = F, M = F) \propto (0.95)(0.99)(0.71) = 0.6678$$

Example: Gibbs Sampling of BN (V)



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	
2					
3					
4					

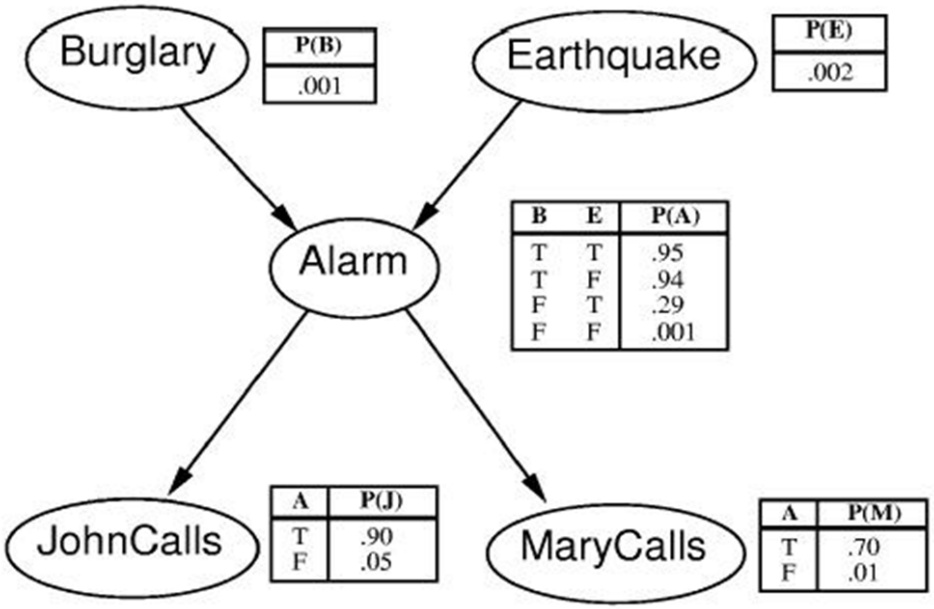
- Sampling P(J|A): No need to apply Bayes Rule

- A = F, so we compute the following, and sample J = T

$$P(J = T \mid A = F) \propto 0.05$$

$$P(J = F \mid A = F) \propto 0.95$$

Example: Gibbs Sampling of BN (VI)



t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	F
2					
3					
4					

- Sampling P(M|A): No need to apply Bayes Rule

- A = F, so we compute the following, and sample M = F

$$P(M = T \mid A = F) \propto 0.01$$

$$P(M = F \mid A = F) \propto 0.99$$

Example: Gibbs Sampling of BN (VII)

- Now t = 2, and we repeat the procedure to sample new values of B,E,A,J,M …

t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	F
2	F	T	T	T	T
3					
4					

- Likewise for t = 3, 4,….

t	B	E	A	J	M
0	F	F	F	F	F
1	F	T	F	T	F
2	F	T	T	T	T
3	T	F	T	F	T
4	T	F	T	F	F

Gibbs Sampling Properties

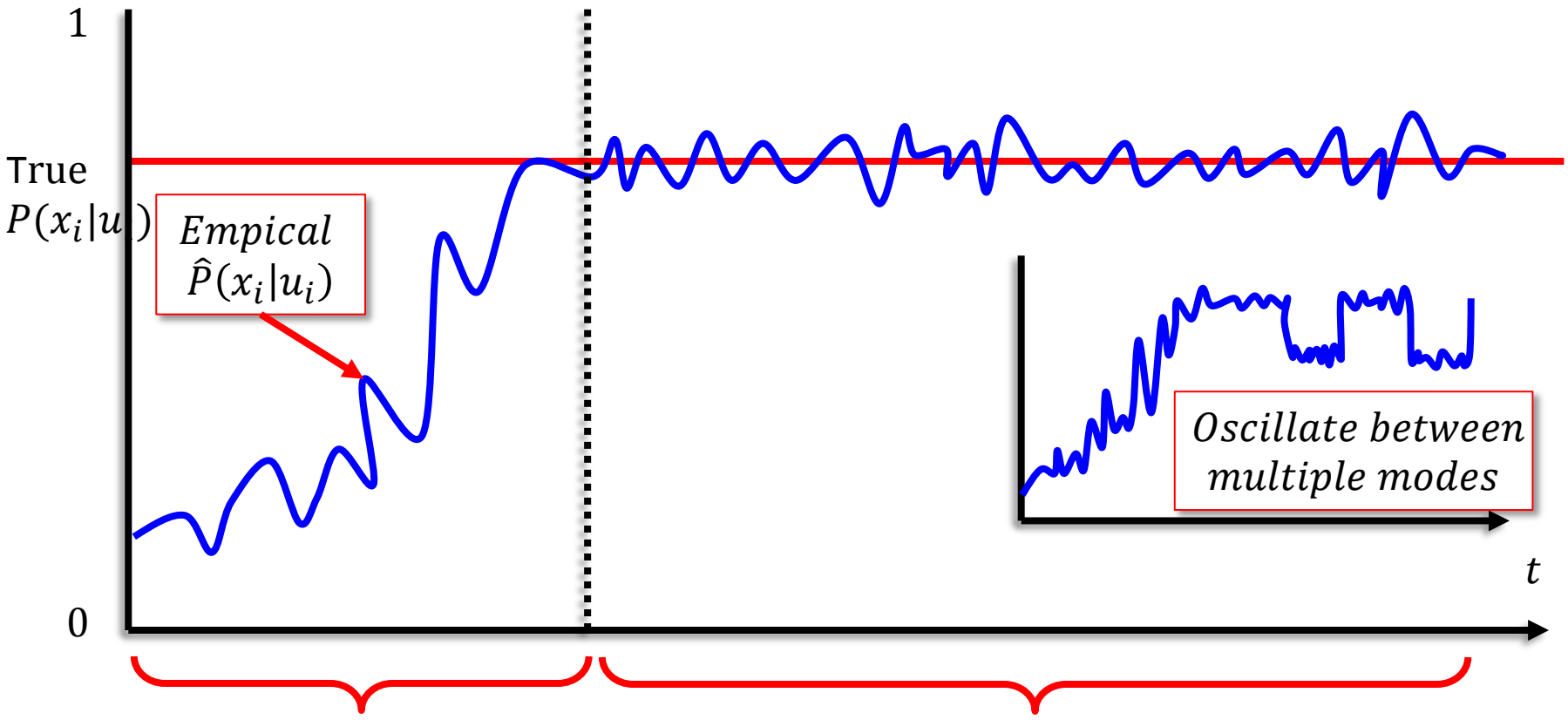
- Fairly easy to derive for many graphical models (e.g., mixture models, Latent Dirichlet allocation, etc.)
- Have reasonable computation and memory requirements, because they sample one variable at a time
- *Can be used in Rao-Blackwellized particle filtering* (i.e., integrate out some r.v.s) to decrease the sampling variance

Practical Aspects of Gibbs Sampling (I)

- We must wait until the burn-in time has ended; i.e., number of steps until we take samples from the stationary chain
  - Must wait until the sampling distribution is close to stationary dist.
  - Hard to characterize the 'mixing time'
  - Once the burn-in time ended, all samples are from the stationary distribution.
  - We can examine burn-in time by comparing the estimates (comparable variances for K chains, see page 523 of textbook for details).

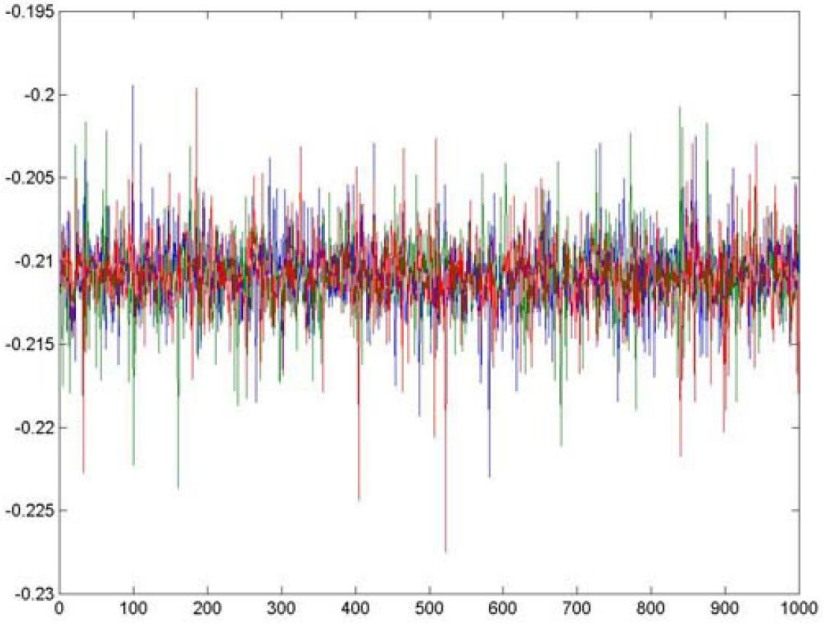
$$\hat{E}_{\mathrm{k}}(f) \approx \frac{1}{M}\sum_{m=1}^{M} f(x^k[m], e)$$ from multiple chains 1,...,K

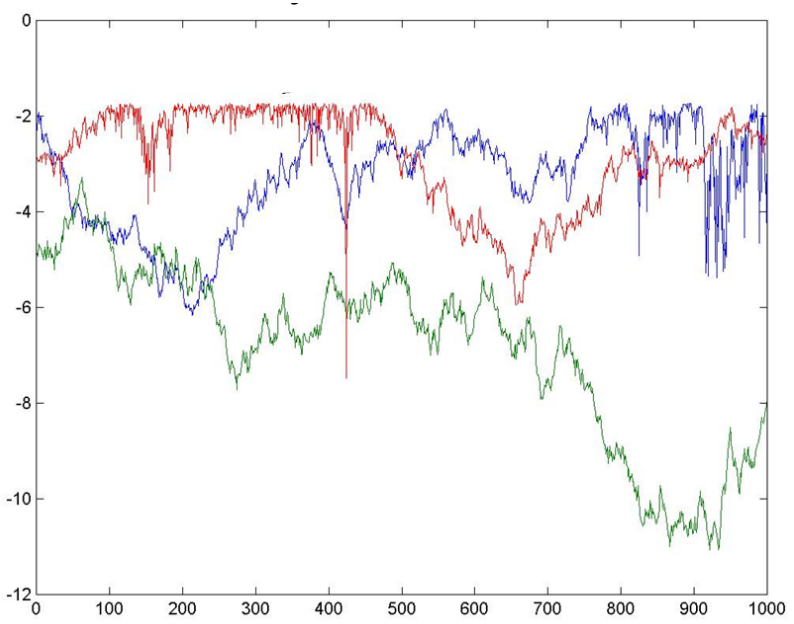
Burning-Time in Gibbs Sampling



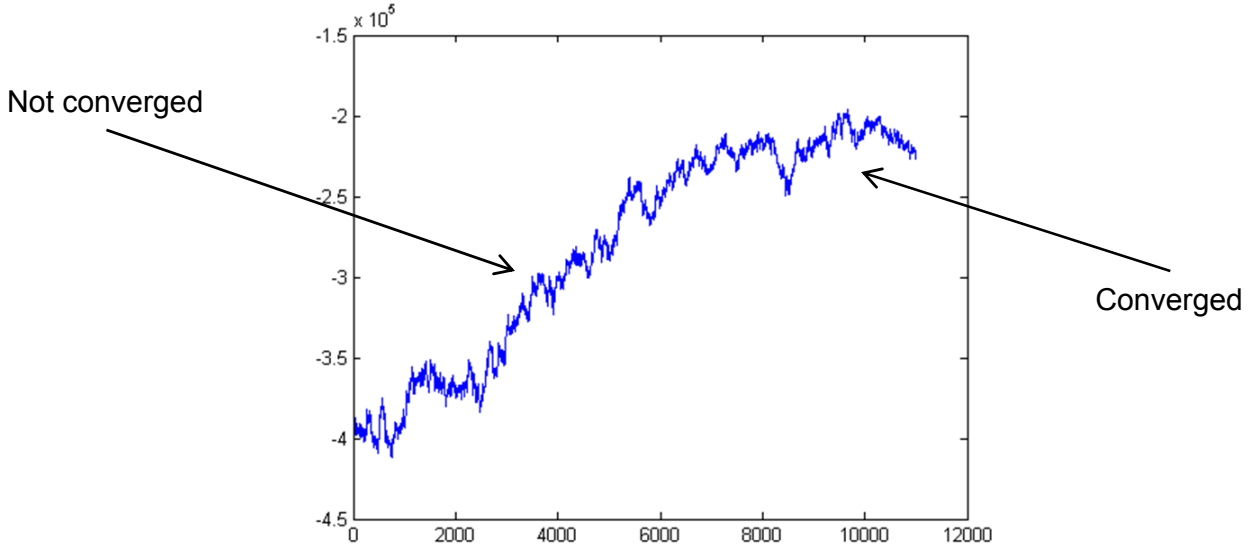
Evaluating Convergence

- Monitor convergence by plotting samples (of r.v.s) from multiple runs of the chain.
  - If the chains are well-mixed (left), they are probably converged
  - If the chains are poorly-mixed (right), we should continue burn in the time.





Log-likelihood vs Time



- Many graphical models are high-dimensional
  - Hard to visualize all r.v. chains at once
- Instead, plot the complete log-likelihood vs. time
  - The complete log-likelihood is an r.v. that depends on all model r.v.s
  - Generally, the log-likelihood will climb, then eventually plateau

-Xing

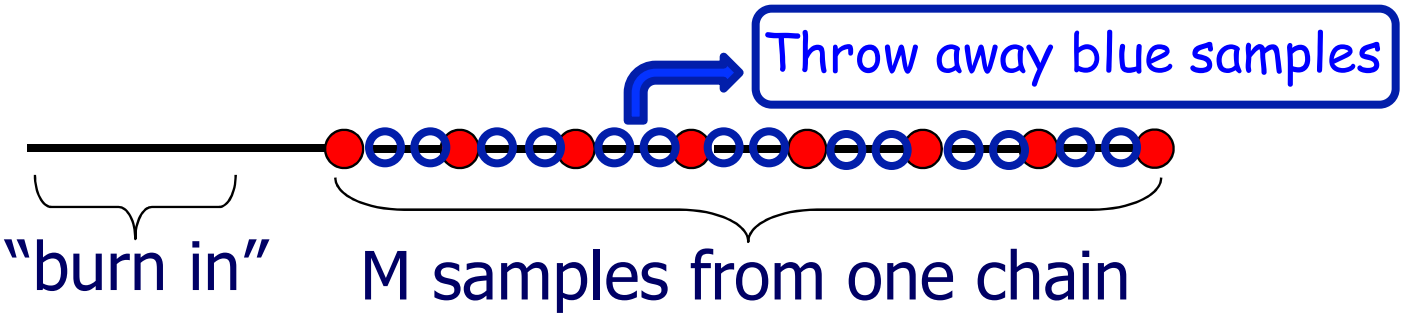
Practical Aspects of Gibbs Sampling (II)

- Remark: after the burn-in time, samples are correlated, i.e., consecutive samples from the same trajectory (the same starting state) are correlated.
- Quantify via the autocorrelation function of a r.v. x:

$$R_x(k) = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n-k} (x_t - \bar{x})^2}$$

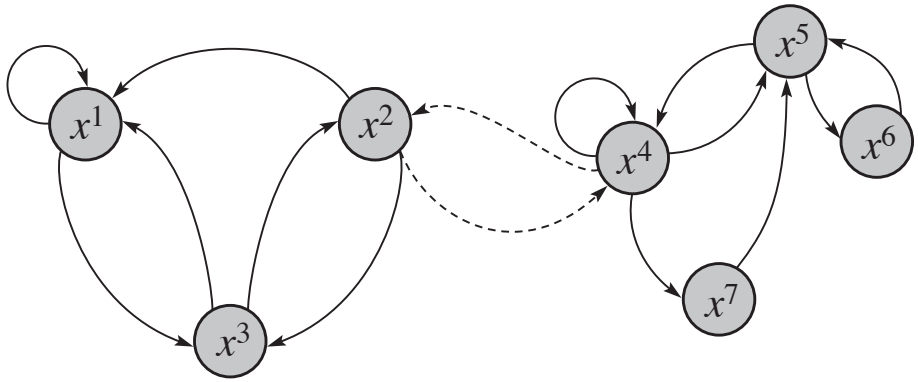
Strategies for Dealing with Correlation

- Strategy I:
  - Run the chain M times, each run for N steps
    - Each run starts from a different state in MC
  - Collect one sample from each run (at the terminating steps)
- Strategy II:
  - Run a single chain for a long time
  - After some "burn in" period, sample points at every some fixed number of steps (Samples might be still weakly correlated)



Limitations of Gibbs sampling

- The chain in Gibbs uses only very local moves over the state space; moves that change only one variable at a time.
- High-probability states will form strong point of attraction, and the chain will be very unlikely to move away from such a state, hence the chain will mix very slowly.



- Possible Solutions:
  - Block Gibbs sampling (partition variables into disjoint blocks of variables and Iteratively sample blocks of variables)
  - Metropolis-Hastings algorithm

Distributional Particles (Rao-Blackwellized Particles)

- In sampling (i.e., creating particles), we used full assignment on all random variables.
- Sampling in high dimensional spaces causes high variance in the estimate.
- Idea: use partial assignments to a subset of the variables, combined with closed form representation of a distribution over the rest (i.e., compute expected value over the rest of variable analytically):
  - $\mathbf{X}_p$ – Variables whose assignments are set by samples/ particles (using any sampling method)
  - $\mathbf{X}_d$ – Variables over which we maintain a distribution

Rao-Blackwellized Particles

- Estimation proceeds as:

$$E_{p(X|e)}[f(X)] = \int p(x_p, x_d \mid e) f(x_p, x_d) dx_p dx_d$$

$$= \int_{x_p} p(x_p \mid e) \left(\int_{x_d} p(x_d \mid x_p, e) f(x_p, x_d) dx_d \right) dx_p$$

$$= \int_{x_p} p(x_p \mid e) E_{p(X_d|x_p,e)}[f(x_p, X_d)] dx_p$$

$$= \frac{1}{M} \sum_m E_{p(X_d|x_p^m,e)}[f(x_p^m, X_d)], \qquad x_p^m \sim p(x_p \mid e)$$

We assume that we can compute the internal expectation efficiently

Is there any benefit of using partial assignment to do the estimation?

- Consider (**law of total variance**) identity:

$$\mathrm{var}[\tau(X_p, X_d)] = \mathrm{var}[E[\tau(X_p, X_d) \mid X_p]] + E[\mathrm{var}[\tau(X_p, X_d) \mid X_p]]$$

- The second term on the right side is always positive,

  hence: $\mathrm{var}[E[\tau(X_p, X_d) \mid X_p]] \le \mathrm{var}[\tau(X_p, X_d)]$. Therefore, we have:

  $\tau(X_p, X_d) = E[f(X_p, X_d) \mid X_p]$ is a lower variance estimator.

Rao-Blackwellized Particles (II)

- Distributional Likelihood Weighting
  - Use M Samples over only a subset $\mathbf{X_p}$ of the variables:

Example:
Probability of an event

$$P(y \mid e) \approx \frac{\sum_{m=1}^{M} w[m]\left(E_{P(X_d \mid x_p[m],e)}\left[\mathbf{1}\{x[m](y) = y\}\right]\right)}{\sum_{m=1}^{M} w[m]}$$

- Distributional Gibbs Sampling
  - Sample only a subset of the variables
  - Transition probability is as before:

$$T((\mathbf{u_i}, x_i) \rightarrow (\mathbf{u_i}, x'_i)) = P(x'_i \mid \mathbf{u_i})$$

  - Computation may require an extra step of inference