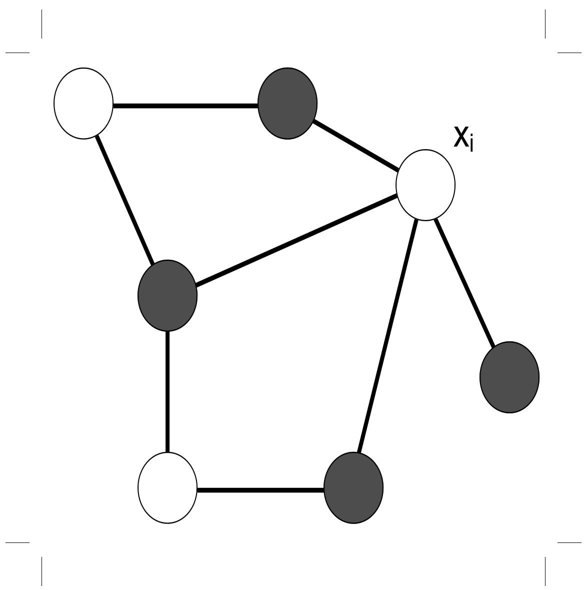
CS281A/Stat241A: Statistical Learning Theory

Sampling Methods (11/30/04)

Lecturer: Michael I. Jordan *Scribe: Jaspal S. Sandhu*

1 Gibbs Sampling



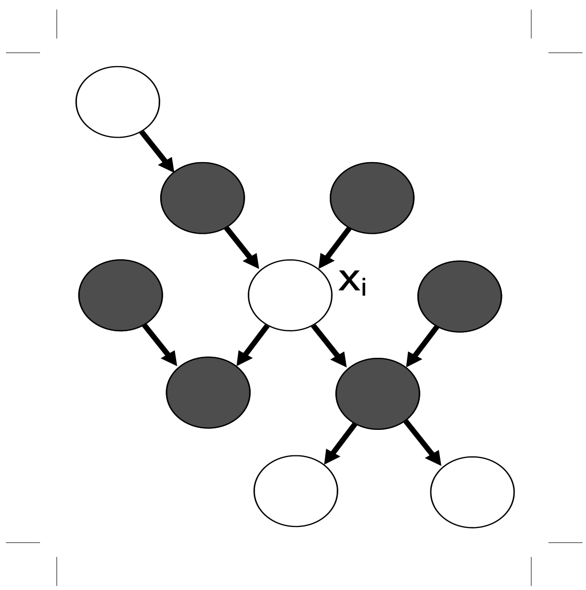


Figure 1: Undirected and directed graphs, respectively, with set of nodes colored that represent the minimum set of nodes necessary to render node x_i conditionally independent of all of the other nodes in the graph.

A *Markov blanket* is the minimum set of nodes that renders node x_i conditionally independent of all other nodes in the directed graph. In the undirected case, simple graph separation suffices (as shown). For the directed case, the Markov blanket (also shown) consists of the parents of x_i, the children of x_i, and the parents of the children of x_i (the "coparents" of x_i).

If interested in this topic, see BUGS (Bayesian inference Using Gibbs Sampling) software project (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml).

When performing Gibbs sampling, for each hidden node X_i, construct a conditional probability $p(x_i|x_{\neg i})$, where $x_{\neg i}$ is the set of nodes in the graph not including x_i. Note that the above conditional independence arguments imply that it suffices to condition on the graph separators of x_i for the undirected graph, and the Markov blanket for the directed case.

Gibbs sampling proceeds by sampling each hidden variable from the appropriate conditional distribution, given the current values of the other variables in the graph. Marginal probabilities can be estimated by summing over the samples.

2 MCMC (Markov Chain Monte Carlo)

State: x_t [x: the state of the entire graph, time t of Markov chain]

Transition matrix: $T(x, x')$ [transition matrix is homogeneous if T is independent of t]

Invariant distribution:

$$p^*(x) = \sum_{x'} T(x', x) p^*(x')$$

[a trivial example of this is the identity matrix, where you stay in the same state no matter the distribution]

Detailed balance:

$$p^*(x)T(x, x') = p^*(x')T(x', x)$$

detailed balance is a sufficient condition for invariance ($\Rightarrow$ but not $\Leftarrow$). [In this context, the term *detailed* - which comes from statistical physics - means *local*.]

Proof:

$$\begin{aligned}
\sum_{x'} p^*(x')T(x', x) &= \sum_{x'} p^*(x)T(x, x') \\
&= p^*(x) \sum_{x'} T(x, x') \\
&= p^*(x)
\end{aligned}$$

Ergodicity: This is when: (1) there is a non-zero probability of getting from any state to any other state and (2) there are no (deterministic) cycles. Refer to the book for a more rigorous definition.

Ergodicity $\Rightarrow$ we have an equilibrium distribution (unique, invariant).

Book recommendation:

Norris, James R. et al. *Markov Chains*, Cambridge U. Press: 1998. (neither too elementary nor too advanced)

2.1 Metropolis-Hastings (M-H)

Current state: x

Propose to move to x' according to proposal distribution $q_t(x, x')$

Accept the proposal with probability:

$$A_t(x', x) = min\left(1, \frac{p(x')}{p(x)} \frac{q_t(x', x)}{q_t(x, x')}\right)$$

M-H ”loves to go uphill, is willing to go downhill”.

Evaluating $p(x)$, $p(x')$ is easy.

q_t is a simple distribution, e.g., Gaussian, so it is easy to compute by definition. In general, the normalizer terms for $p(x), p(x')$ will cancel in this ratio.

“Uphill” steps will be accepted with probability 1; downhill steps will be accepted with probability $A_t(x', x)$.

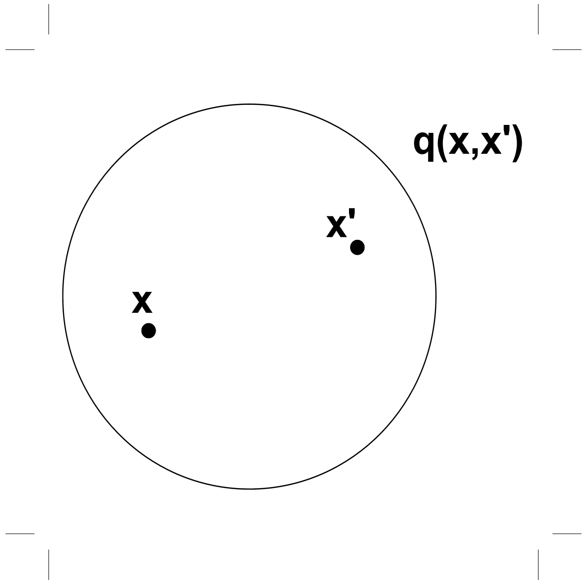


Figure 2:

If the transition is symmetric, the $q_t(\cdot)$ terms cancel. This is the "Metropolis" algorithm:

Metropolis, N., Rosenbluth, A. W., Rosenbluth, N., Teller, A. H.,and Teller, E. (1953) Equation of state calculation by fast computing machines. *Journal of Chemical Physics* 21: 1087-1092.

Note: for Ising model and q="flip one bit", M-H $\equiv$ Gibbs

M-H satisfies detailed balance.

Proof:

$$\begin{aligned}
p(x)q_t(x,x')A_t(x',x) &= \min(p(x)q_t(x,x'), p(x')q_t(x',x)) \\
&= \min(p(x')q_t(x',x), p(x)q_t(x,x')) \\
&= p(x')q_t(x',x)A_t(x,x')
\end{aligned}$$

Ergodicity is based on a reasonable choice for proposal distribution (e.g., Gaussian), to ensure that all paths from any state to any other state are non-zero.

2.2 Gibbs as M-H

Fix $x_{\neg i}$. Consider updating node x_i (x' is x with $x_{\neg i}$ held fixed and x_i variable) according to proposal distribution:

$$q(x,x') = p(x_i|x_{\neg i})$$

by Bayes' Theorem:

$$\begin{aligned}\frac{p(x')}{p(x)}\frac{q(x',x)}{q(x,x')} &= \frac{p(x'_i|x'_{-i})}{p(x_i|x_{-i})}\frac{p(x'_{-i})}{p(x_{-i})}\frac{p(x_i|x'_{-i})}{p(x'_i|x_{-i})} \\ &= 1\end{aligned}$$

(i.e. always accept proposal)

Book recommendations:

Robert, Christian P. and George Casella. *Monte Carlo Statistical Methods*, Springer-Verlag: 2004.

Liu, Jun S. *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag: 2001.

Gilks, W.R. et al. *Markov Chain Monte Carlo in Practice*, Chapman-Hall: 1995.

2.3 Particle Filtering

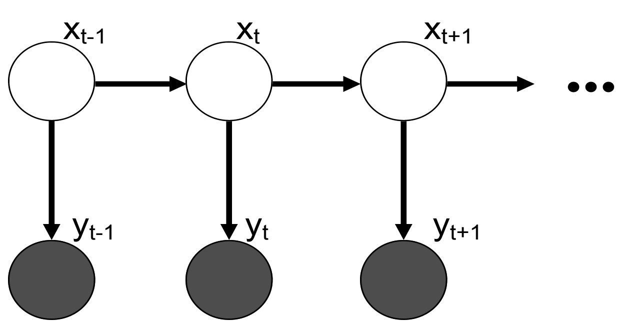


Figure 3: Graphical model on which particle filtering is often used.

Interest is in $p(x_t|y_{(t)})$, $y_{(t)} = (y_1, ..., y_t)$ i.e. - filtering.

Consider some function $f(x_t)$

$$\begin{aligned}\langle f(x_t)\rangle &= \int (f(x_t)p(x_t|y_{(t)})dx_t \\ &= \int (f(x_t)p(x_t|y_t, y_{(t-1)})dx_t \\ &= \frac{\int f(x_t)p(y_t|x_t)p(x_t|y_{(t-1)})dx_t}{\int p(y_t|x_t)p(x_t|y_{(t-1)})dx_t} \\ &\approx \sum_{m=1}^{M} w_t^{(m)} f(x_t^{(m)})\end{aligned}$$

where $\{x_t\}$ are samples drawn from $p(x_t|y_{(t+1)})$ and

$$w_t^{(m)} = \frac{p(y_t|x_t^{(m)})}{\sum_m p(y_t|x_t^{(m)})}$$

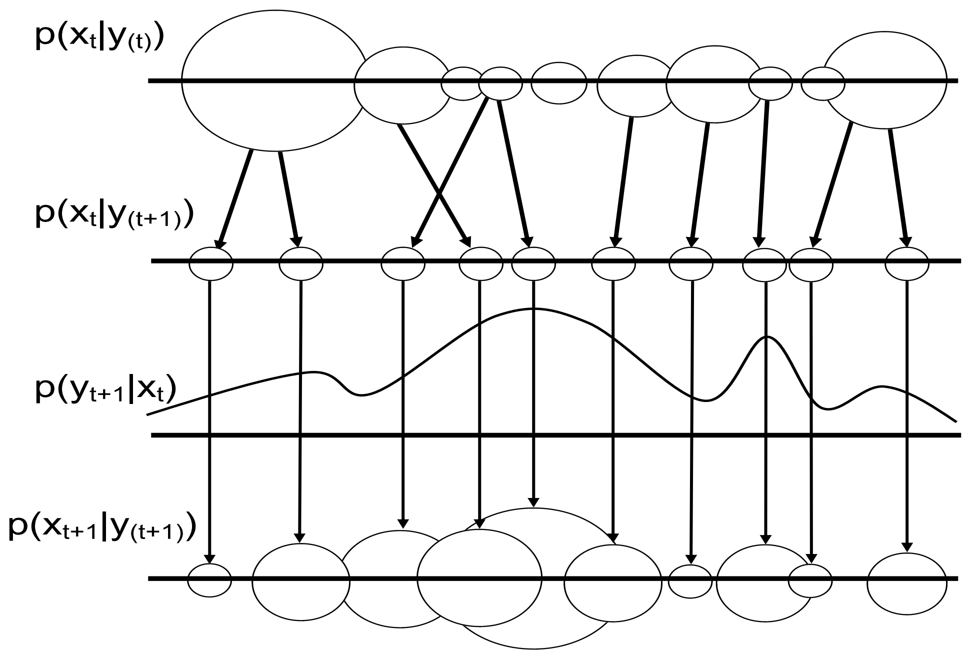


Figure 4: Graphical illustration of particle filter updates.

This schematic of the particle filter shows the posterior represented as a mixture model at time t. M samples are drawn from this distribution and $p(y_{t+1}|x_{t+1}^{(m)})$ is used to determine the new weights $w_{(t+1)}^{m}$.

$$\begin{aligned}
p(x_{t+1}|y_{(t)}) &= \int p(x_{t+1}|x_t, y_{(t)})p(x_t|y_{(t)})dx_t \\
&= \int p(x_{t+1}|x_t)p(x_t|y_{t,(t-1)})dx_t \\
&= \frac{\int p(x_{t+1}|x_t)p(y_t|x_t)p(x_t|y_{(t-1)})dx_t}{\int p(y_t|x_t)p(x_t|y_{(t-1)})dx_t} \\
&\approx \sum_M w_t^{(m)} p(x_{t+1}|x_t^{(m)})
\end{aligned}$$