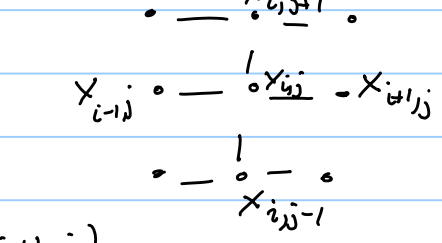
Lecture 16.

Markov Chain Monte Carlo (MCMC)

Note Title 5/2/2011

Suppose we have a graphical model with closed loops



E.G. Potts Model in 2D

$$E(\underline{x}) = \sum_{ij} \sum_{k \in N(ij)} \varphi(x_{ij}, x_{kl})$$

$$N(ij) = \{(i-1,j), (i+1,j), (i,j+1), (i,j-1)\}$$

neighborhood.

$$\underline{x} = \{(x_{ij})\}$$

$$\pi(\underline{x}) = \frac{1}{Z} e^{-E(\underline{x})}.$$

How to sample from $\pi(\underline{x})$?

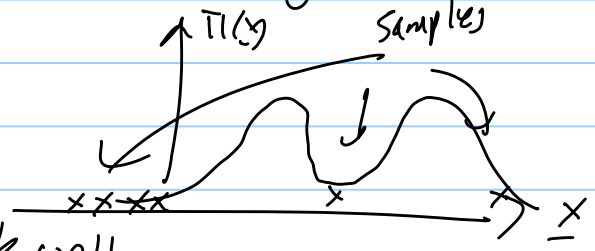
We could try SIS – i.e. choose an order $x_1, \ldots, x_N$ of the variables $\{x_{ij}\}$.

Select trial distributions $g(x_t | x_1 \ldots x_{t-1})$ and approximate distributions $\pi_t(\underline{x}_t)$ $\underline{x}_t = (x_1, \ldots, x_t)$

Such that $\pi_N(\underline{x}_N) = \pi(\underline{x}_N)$.

The problem is that we do not know how to specify $g(\cdot | \ldots)$ and $\pi_t(\underline{x}_t)$. Unless these relate closely to $\pi(\underline{x})$ the SIS samples will be in the wrong parts of the distribution and have low weights.

So we may need an enormous number of samples in order to get a good estimator.



So SIS will usually not work well.

Markov Chain Monte Carlo. (MCMC)

Note Title 5/7/2006

MCMC is a way to sample from any distribution $\pi(\underline{x})$.

It does not sample directly from $\pi(\underline{x})$. Instead it proceeds by defining a Markov chain that converges to samples from $\pi(\underline{x})$.

The original MCMC is the Metropolis algorithm (Metropolis, Rosenbluth & Rosenbluth, Teller & Teller 1953).

Later generalized to the Metropolis-Hastings algorithm.

Strictly speaking, these are not algorithms. They are design principles for algorithms.

For any distribution $\pi(\underline{x})$, there are many different Metropolis-Hastings algorithms that we can design to obtain samples from $\pi(\underline{x})$. Some will be much more efficient than others.

Not all MCMC are Metropolis-Hastings.

First we will specify the most general MCMC.

Then we will introduce Metropolis and Metropolis-Hastings.

Markov Chain.

Let $P(\underline{x}^{t+1} | \underline{x}^t) = K(\underline{x}^{t+1} | \underline{x}^t)$ transition kernel.

$K(\cdot|\cdot)$ must obey $K(\underline{x}^{t+1}|\underline{x}^t) \geq 0$, for all $\underline{x}^t, \underline{x}^{t+1}$

and $\sum_{\underline{x}^{t+1}} K(\underline{x}^{t+1}|\underline{x}^t) = 1$, for all $\underline{x}^t$

We can obey a set of samples from this chain.

$\underline{x}^0$ – randomly initialised.

$\underline{x}^1$ – sampled from $K(\underline{x}^1|\underline{x}^0)$

$\underline{x}^2$ – sampled from $K(\underline{x}^2|\underline{x}^1)$

$\underline{x}^N$ – sampled from $K(\underline{x}^N|\underline{x}^{N-1})$

MCMC is a special MC choosen so that the transition kernel $K(\underline{x}|\underline{y})$ satisfies the

fixed point condition $\sum_{\underline{y}} K(\underline{x}|\underline{y})\pi(\underline{y}) = \pi(\underline{x})$ (1)

or the detailed balance condition

$$K(\underline{x}|\underline{y})\pi(\underline{y}) = K(\underline{y}|\underline{x})\pi(\underline{x}) \quad (2)$$

Note: detailed balance implies fixed point

because $\sum_{\underline{y}} K(\underline{x}|\underline{y})\pi(\underline{y}) = \sum_{\underline{y}} K(\underline{y}|\underline{x})\pi(\underline{x}) = \pi(\underline{x}) \sum_{\underline{y}} K(\underline{y}|\underline{x}) = \pi(\underline{x})$, where $\sum_{\underline{y}} K(\underline{y}|\underline{x}) = 1$

Note: fixed point condition means intuitively that if we sample $\underline{y}$ from $\pi(\underline{y})$, next sample $\underline{x}$ from $K(\underline{x}|\underline{y})$, then $\underline{x}$ is a sample from $\pi(\underline{x})$

Note: In practice, most transition kernels are choosen to obey Detailed Balance (simpler to check).

We also require an MCMC to be irreducible, so that for any $\underline{x}, \underline{y}$ we can find a sequence $\underline{x}_1, \ldots \underline{x}_n$ such that

$K(\underline{x}|\underline{x}_1) K(\underline{x}_1|\underline{x}_2) \ldots K(\underline{x}_n|\underline{y}) > 0$. (for some n)

i.e. there is a set of moves that take us from any point $\underline{x}$ to any other point $\underline{y}$.

(Equivalently $\sum_{\underline{y}_1 \ldots \underline{y}_n} K(\underline{x}|\underline{y}_1) K(\underline{y}_1|\underline{y}_2) \ldots K(\underline{y}_n|\underline{y}) > 0$

These conditions ensure that samples from the MCMC will eventually tend to sample from $\Pi(\underline{x})$.

i.e. $\underline{x}_0, \underline{x}_1, \ldots \underline{x}_t, \ldots \underline{x}_N, \ldots$ with $\underline{x}_t$ sampled from $K(\underline{x}_t|\underline{x}_{t-1})$

then, for large enough N, $\underline{x}_N$ is a sample $\Pi(\underline{x})$

(Proof will be given in a later lecture).

This means that we can sample from any distribution $\Pi(\underline{x})$.

This is too good to be true. We could solve NP-complete problems.

The difficulty is how big N must be in order for $\underline{x}_N$ to be a sample from $\Pi(\underline{x})$.

The Metropolis Algorithm

Suppose $\pi(\underline{x}) = \frac{1}{Z} e^{-E(\underline{x})}$ Gibbs distribution.

Z unknown. / hard to calculate

For each $\underline{x}$ define a neighborhood $N(\underline{x})$ such that $\underline{y} \in N(\underline{x})$ implies $\underline{x} \in N(\underline{y})$ (for all $\underline{x}, \underline{y}$) and the size $|N(\underline{x})|$ of the neighborhood is independent of $\underline{x}$

Basic Metropolis:

Loop over t,

state $\underline{x}^t$ at time t.

(1) propose a move/transition from $\underline{x}^t$ to $\underline{x}^{t+1} \in N(\underline{x}^t)$ with uniform probability – i.e. $p(\underline{x}^{t+1}) = 1/|N(\underline{x}^t)|$.

(2) accept the move with probability

$$\min\left\{1, \frac{\pi(\underline{x}^{t+1})}{\pi(\underline{x}^t)}\right\} = \min\left\{1, \underbrace{e^{-E(\underline{x}^{t+1}) + E(\underline{x}^t)}}_{\text{note: independent of } Z}\right\}$$

This has transition kernel:

$K(\underline{x}^{t+1} | \underline{x}^t) = 0$, if $\underline{x}^{t+1} \notin N(\underline{x}^t)$.

$$K(\underline{x}^{t+1} | \underline{x}^t) = \frac{1}{|N(\underline{x}^t)|} \min\left\{1, \frac{\pi(\underline{x}^{t+1})}{\pi(\underline{x}^t)}\right\}, \quad \underline{x}^{t+1} \in N(\underline{x}^t), \ (\underline{x}^{t+1} \neq \underline{x}^t)$$

$$K(\underline{x}^t | \underline{x}^t) = 1 - \sum_{\underline{x}^{t+1} \in N(\underline{x}^t)} K(\underline{x}^{t+1} | \underline{x}^t)$$

Metropolis obeys detailed balance because for $\underline{x}^{t+1} \neq \underline{x}^t$

$$K(\underline{x}^{t+1} | \underline{x}^t)\pi(\underline{x}^t) = \frac{1}{|N(\underline{x}^t)|} \min\{\pi(\underline{x}^t), \pi(\underline{x}^{t+1})\} = K(\underline{x}^t | \underline{x}^{t+1})\pi(\underline{x}^{t+1})$$

(with equality also if $\underline{x}^t = \underline{x}^{t+1}$). (recall $|N(\underline{x}^{t+1})| = |N(\underline{x}^t)| = \text{constant}$).

Hence Metropolis will generate samples from $\pi(\underline{x})$ (assuming irreducibility).