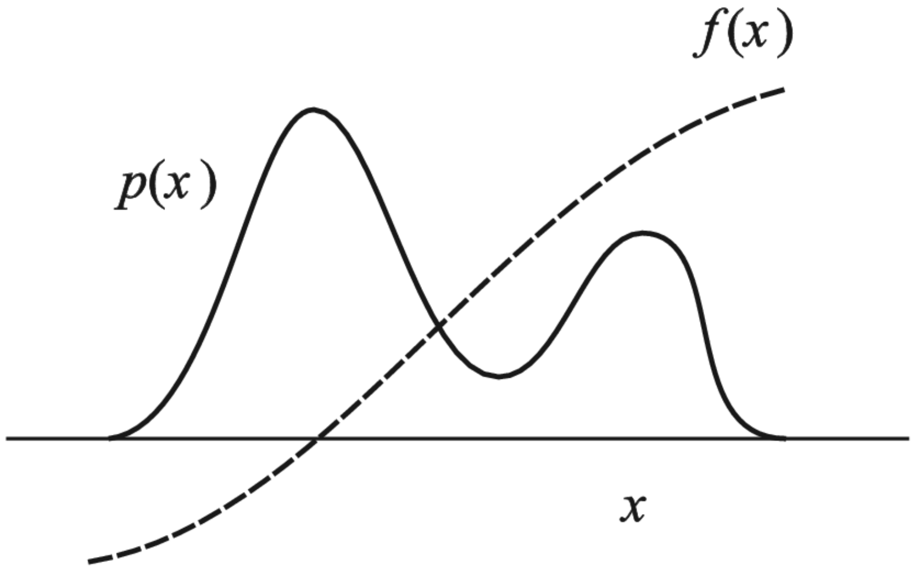
Sampling

Goal: Independent samples $x^{(1)}, x^{(2)}, \ldots$ from target probability distribution p

Estimating expectation $E[f(X)]$ where X from p

$\hat{f} = \frac{1}{M} \sum_{m=1}^{M} f(x^{(m)})$

$E[\hat{f}] = E[f(X)]$ even if samples not independent, but each from p



Markov chain Monte Carlo (MCMC)

- MCMC generates a sequence of samples $X^{(0)}, X^{(1)}, \ldots$
- Distribution of $X^{(n)}$ in limit converges to the target distribution
- But samples not independent

Markov chains

- Sequence of random variables $X^{(0)}, X^{(1)}, \ldots$ such that $\Pr(X^{(t)} \mid X^{(0)}, \ldots, X^{(t-1)}) = \Pr(X^{(t)} \mid X^{(t-1)})$, i.e. value in time t depends only on value in time $t-1$
- For simplicity assume values of $X^{(t)}$ from a finite set (set of states)
- $\Pr(X^{(t)} = y \mid X^{(t-1)} = x)$ given by $p_{x,y}$ from transition matrix P
- $\Pr(X^{(t)} = y \mid X^{(0)} = x)$ obtained from P^t
- Distribution π over set of states is **stationary** for P if for each j we have $\sum_i \pi(i) p_{i,j} = \pi(j)$
  or in matrix notation $\pi P = \pi$
- Ergodic matrices P have exactly one stationary distribution π, and for each x and y we have $\lim_{t\to\infty} p^t_{x,y} = \pi(y)$

Ergodicity of Markov chains

Matrix P is ergodic if for some t has P^t all entries non-zero

Examples: the first three non-ergodic, last ergodic

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 0.5 & 0.5 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0.5 & 0.5 \\ 1 & 0 \end{pmatrix}$$

Markov chain Monte Carlo

- Want to sample from a complex distribution π
- Create ergodic Markov chain with π as stationary distribution
- Start from some $X^{(0)}$, repeatedly sample from $\Pr(X^{(t)} \mid X^{(t-1)})$
- After sufficiently long t, $X^{(t)}$ from distribution similar to π
- But successive samples not independent!
- Still, they can be used to estimate expected values
  $\frac{1}{t}\sum_{i=1}^{t} f(X^{(t)})$ converges to $E_{\pi}[f(X)]$

We will cover two MCMC algorithms:

Gibbs sampling, Metropolis–Hastings algorithm

Gibbs sampling

- Target distribution $\pi(X)$ over vectors $X = (x_1, \ldots, x_n)$
- In each step sample one coordinate x_i from conditional $\Pr(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_n)$
- Other coordinates left from the previous step
- Value i chosen randomly or periodically $i = 1, 2, \ldots, n, 1, \ldots$

Proof of Gibbs sampling correctness for ergodic chains

- Def.: P and π satisfy **detailed balance** if for each x and y we have $\pi(x)p_{x,y} = \pi(y)p_{y,x}$
- Lemma: If P and π satisfy detailed balance, π is stationary for P.
- Proof: $\sum_x \pi(x)p_{x,y} = \sum_x \pi(y)p_{y,x} = \pi(y)\sum_x p_{y,x} = \pi(y)$.
- Lemma: Gibbs sampling chain satisfies detailed balance for target distribution π (and thus π stationary as needed)
- Proof: let x and y are successive vectors differing in i-th coordinate
- Let x_{-i} be values of all coordinates except x_i
- $\pi(x)p_{x,y} = \pi(x)\Pr(y_i \mid x_{-i}) = \Pr(x_{-i})\Pr(x_i \mid x_{-i})\Pr(y_i \mid x_{-i}) = \pi(y)\Pr(x_i \mid x_{-i}) = \pi(y)\Pr(x_i \mid y_{-i}) = \pi(y)p_{y,x}$

Example of Gibbs sampling: motif finding in DNA

Matrix W (size $|\Sigma| \times L$) of frequencies in the motif

Background frequencies q outside the motif

Position o_i of motif in sequence S_i

A	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
C	0.01	0.01	0.01	0.39	0.19	0.97	0.01	0.01	0.89
G	0.01	0.01	0.01	0.59	0.79	0.01	0.97	0.97	0.09
T	0.97	0.97	0.97	0.01	0.01	0.01	0.01	0.01	0.01

$q[A] = 0.3$, $q[C] = 0.2$, $q[G] = 0.2$, $q[T] = 0.3$

Model for motif finding in DNA

Matrix W (size $|\Sigma| \times L$) of frequencies in the motif

Background frequencies q outside the motif

Position o_i of motif in sequence S_i

Model defines probability distribution $\Pr(S \mid W, q, O)$

$$\Pr(S_i \mid W, q, o_i) = \prod_{j=1}^{L} W[S_i[j+o_i-1], j] \prod_{j=1}^{o_i-1} q[S_i[j]] \prod_{j=o_i+L}^{m} q[S_i[j]]$$

$$\Pr(S \mid W, q, O) = \prod_{i=1}^{n} \Pr(S_i \mid W, q, o_i)$$

+ added priors on W and O to get $\Pr(S, W, O \mid q)$

Gibbs sampling for motifs PhyloGibbs (Siddharthan et al. 2005)

Model defines probability distribution Pr(S, W, O)

$S = (S_1, \dots S_n)$: DNA sequences, each of length m

W: matrix of frequencies in the motif (size $|\Sigma| \times L$)

$O = (o_1, \dots, o_n)$: positions of motif occurrences

Algorithm

Sample from Pr$(O \mid S)$, marginalize out W

In step $t + 1$ select one sequence S_i

For each position o_i' compute Pr$(o_i' \mid O_{-i}^{(t)}, S)$

Sample a particular o_i' proportional to these probabilities

$O^{(t+1)}$ obtained from $O^{(t)}$ by substituting o_i' for o_i

Computation of $\Pr(o_i \mid O_{-i}, S)$

$\Pr(o_i \mid O_{-i}, S) = \Pr(O \mid S) / \Pr(O_{-i} \mid S) \propto \Pr(O \mid S)$

$\Pr(O \mid S) = \Pr(S \mid O)\Pr(O) / \Pr(S) \propto \Pr(S \mid O)$ (if $\Pr(O)$ uniform)

$\Pr(S \mid W, O)$ is easy to compute, but we need $\Pr(O \mid S)$

Let $S_{(W)}$ be parts of sequences generated from W, $S_{(q)}$ the rest

$\Pr(S \mid O) = \Pr(S_{(W)} \mid O)\Pr(S_{(q)} \mid O)$

$\Pr(S_{(q)} \mid O) = \Pr(S_{(q)})$ easy to compute

Computation of $\Pr(o_i \mid O_{-i}, S)$ (cont.)

Need $\Pr(S_{(W)} \mid O)$:

$\Pr(S_{(W)} \mid O) = \int \Pr(S_{(W)} \mid O, W) \Pr(W) dW$,

integral over W where $w_{a,j} \geq 0$ and $\sum_a w_{a,j} = 1$

$\Pr(W)$ is a constant for uniform prior

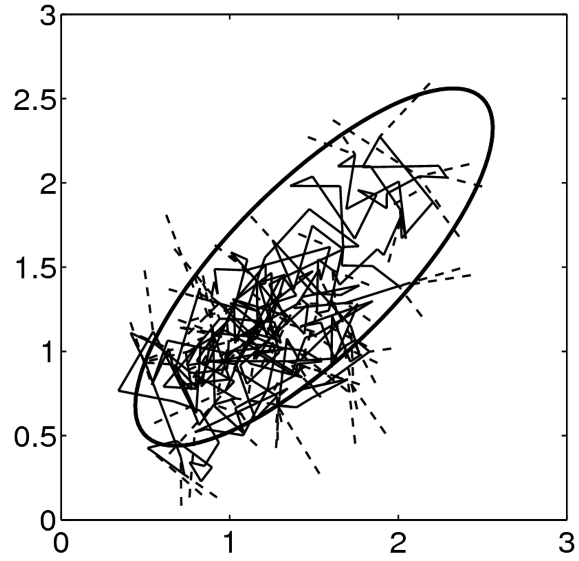
$\Pr(S_{(W)} \mid O, W) = \prod_{i=1}^{L} \prod_a (w_{a,j})^{n_{a,j}}$

$n_{a,j}$ is the number of occurrences of a at position j in windows O

$\Pr(S_{(W)} \mid O) = \prod_{j=1}^{L} 3!/(n+3)! \prod_a n_{a,j}! \propto \prod_{j=1}^{L} n_{S_i[o_i+j-1],j}$

Metropolis–Hastings algorithm

- Proposal distribution $q(x \mid x^{(t)})$
- Sample x from $q(x \mid x^{(t)})$
- Compute $q(x \mid x^{(t)})$, $q(x^{(t)} \mid x)$, $p(x^{(t)})$, $p(x)$ (up to a constant factor)
- Accept x as $x^{(t+1)}$ with probability min $\left(1, \frac{p(x)q(x^{(t)}|x)}{p(x^{(t)})q(x|x^{(t)})}\right)$
- If rejected, set $x^{(t+1)} = x^{(t)}$

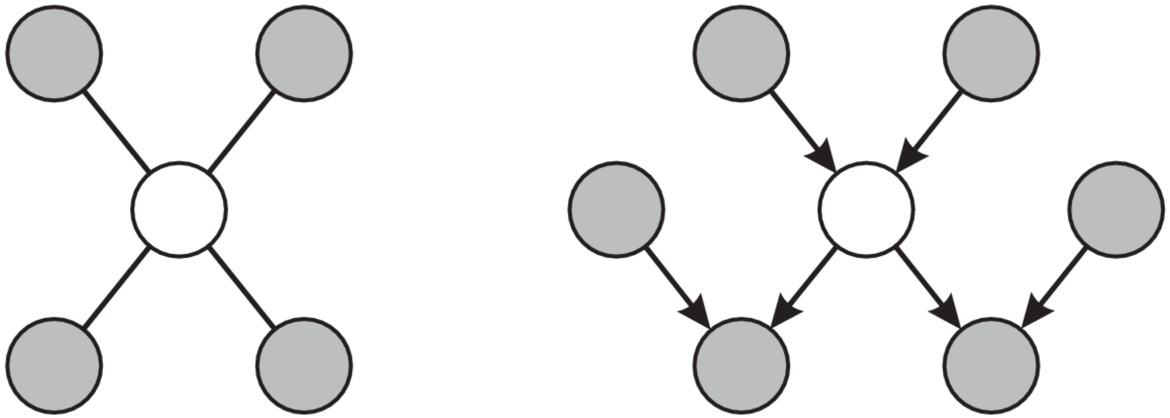


MCMC notes

- Typically discard start of each chain (burn-in)
- If “independent” samples desired, use every kth sample for a large k
- Possible problems: slow convergence (slow mixing), high rejection rate
- Many tricks for improving / monitoring convergence

Gibbs sampling in graphical models

Sample a node conditioning on its Markov blanket



(a) (b)

Potentially sample groups of nodes with tracktable structure