Handout 8

STAT5710 Applied Stochastic Processes

Here we consider practical implementation of Metropolis–Hastings to work, and Gibbs sampling for multivariate targets.

Implementing Metropolis-Hastings

Take $f(x) \propto \exp\{-\frac{1}{2}(x-\mu)^2\}$ with a Gaussian random walk proposal.

1. Fill in the algorithm. Initialize X_0; for $n = 0, 1, 2, \ldots, N$:

   (a) $Y = X_n + \varepsilon$ where $\varepsilon \sim$, so that $q(y \mid x) =$

   (b)

$$X_{n+1} \;=\; \begin{cases} Y & \text{w.p. } \alpha(X_n, Y) \\ & \\ & \text{w.p.} \end{cases}$$

   where the acceptance probability is

$$\alpha(x, y) \;=\; \min\,(1, \qquad\qquad\qquad\qquad\qquad\qquad)$$

   Return $X_0, \ldots, X_N$. The proposal tuning parameter is _____.

2. Discarding the first m draws, estimate the mean and a tail probability.

$$\hat{\mu} =$$

A note on implementation. We need to return Y with probability α and X_n otherwise.

3. Let $U \sim \text{unif}(0, 1)$ and let $r > 0$.

   (a) $P(U < r) =$, which is exactly _____.

   (b) Since log is increasing, $U < r$ if and only if

   (c) So the update is: if _____ return Y, else return X_n.

We want two things from the chain.

4. i.
   ii.

5. The step size trades these against each other.

   If s is large, but

   If s is small, but

Mixing

A general way to characterize mixing is through the distance from stationarity after n steps. Write $P^n(x, \cdot)$ for the n-step transition kernel started at x.

7. **Definition.** The **total variation distance** from stationarity is

   $$\|P^n(x,\cdot) - \pi\|_{TV} :=$$

8. **Definition.** A chain is **geometrically ergodic** with coefficient ρ if

   Informally, a large mixing coefficient ρ means…

Stronger dependence reduces estimation efficiency. Suppose $X_1, \dots, X_n$ are dependent and stationary, and write $\gamma(h) = \mathrm{cov}(X_0, X_h)$.

9. Derive the variance of the sample mean.

   (a) Starting from the definition,

   $$\mathrm{var}(\bar{X}_n) \;=$$

   (b) For each $h \geq 0$ there are _____ pairs (X_i, X_j) with $|i-j| = h$, so collecting terms by lag,

   $$\mathrm{var}(\bar{X}_n) \;=$$

   (c) Let $n \to \infty$ and assume $\sum_h |\gamma(h)| < \infty$. Writing $\rho(h)$ for the autocorrelation,

   $$\mathrm{var}(\bar{X}_n) \;\approx$$

   This gives the **variance inflation**

   $$\tau =$$

   For independent draws $\tau =$

10. Each piece is estimable from the chain itself.

    $\hat{\gamma}(h) =$ $\qquad$ $\hat{\rho}(h) =$ $\qquad$ $\hat{\tau} =$

    (a) A standard error for $\bar{X}$ is then $\quad \mathrm{SE}(\bar{X}) =$

    (b) We call $n/\hat{\tau}$ the _____.

Returning to transition matrix estimation (if time)

Recall the Dirichlet-multinomial model with a stationary start, whose posterior we could not integrate. On the log scale,

$$\ell(P) \;=\; \sum_i \sum_j (N_{ij} + \alpha_{ij} - 1) \log p_{ij} \;+\; \log \pi(x_0; P) \;+\; C$$

A random walk on P itself would leave the simplex, so we move to an unconstrained parameter instead.

11. Let $\Theta \in \mathbb{R}^{k \times (k-1)}$ be a latent parameter with

$$p_{ij} = \operatorname{softmax}(\theta_i)_j = \qquad\qquad \text{where } \theta_{ik} := 0$$

(a) Why is one column fixed at zero?

(b) Proposing $\operatorname{vec} \Theta^* = \operatorname{vec} \Theta + \varepsilon$ with $\varepsilon \sim N(0, \sigma^2 I)$ gives the elementwise proposal distribution

$$q(y \mid x) =$$

elementwise.

(c) The acceptance ratio simplifies to

$$\log r(\Theta^*, \Theta) \;=$$

12. Initialize $P^{(0)} = \left\{ \dfrac{N_{ij} + \alpha_{ij}}{N_{i\cdot} + \alpha_{i\cdot}} \right\}$ and $\theta_{ij}^{(0)} \leftarrow$.

Then for $n = 0, 1, \ldots, N$:

i. propose

ii. update

iii. record

Then estimate $\hat{P} =$

Gibbs sampler

Now let the target be multivariate, $f(x)$ for $x \in \mathbb{R}^d$ with $d > 1$.

13. The **full conditionals** are

Assume these are tractable up to a constant.

14. **The Gibbs sampler.** At the nth MCMC iteration, do a coordinate sweep.

    For $j = 1, \ldots, d$:

    $$X_j^{(n+1)} \sim$$

    Note that previous coordinates have already been updated when we draw $X_j^{(n+1)}$.

15. **Lemma.** The Gibbs sampler is a Metropolis–Hastings algorithm.

    Proof. It suffices to show that…

    Without loss of generality consider updating the first coordinate, and let

    $$i = (x_1, x_2, \ldots, x_d) \qquad j = (x_1', x_2, \ldots, x_d)$$

    (a) Coordinate-wise, the proposal is

    $$q(\cdot \mid i) =$$

    (b) Substitute into the acceptance ratio and simplify.

    $$r(i, j) \;=$$

16. **Finding conditionals.** If $f(x, y)$ is a joint density or mass function and $f(x, y) = g(x, y)\, h(y)$, then

    $$f(x \mid y) \;=$$

    So a conditional can be read off, up to a constant, by _____.

17. **Example.** Let $X \mid n, p \sim \text{binomial}(n, p)$, with $n \sim \text{Poisson}(\lambda)$ and $p \sim \text{beta}(1, 1)$.

Recall: the kernel of a beta(a, b) density is $p^{a-1}(1-p)^{b-1}$.

(a) Write the joint mass up to a constant.

$$f(x, n, p) \ \propto$$

(b) Treating the other arguments as fixed, identify each conditional.

$$p(n \mid x, p) \propto \qquad \implies n - x \mid x, p \ \sim$$

$$f(p \mid x, n) \propto \qquad \implies p \mid x, n \ \sim$$

(c) Write the MCMC algorithm.

18. **Abundance estimation.** Extend the model in (17) with hyperparameters on the beta prior.

$$X \mid N, p \sim \text{binomial}(N, p) \qquad N \sim \text{Poisson}(\lambda) \qquad p \sim \text{beta}(\alpha, \beta)$$

We interpret the model components as follows:

i.

ii.

iii.

Usually the goal is to estimate N, p. For an example application, suppose you do a quick scan and count $X = x$ students in the library; N would estimate the total number, and p would estimate the quality of your survey.

19. Write the posterior and full conditionals.

(a) The posterior is

$$f(x, N, p) \;\propto\; p(x \mid N, p)\, p(N)\, f(p) \;=$$

(b) The full conditionals are

$$f(p \mid x, N) \propto \qquad\qquad \Longrightarrow\; p \mid x, N \sim$$

$$p(N \mid x, p) \propto \qquad\qquad \Longrightarrow\; N - x \mid x, p \sim$$

20. Write the MCMC algorithm.

21. For vague priors take $\lambda \gg x$ and $\alpha = \beta = 1$. With $x = 15$ the script gives

	$\hat{p}$	$\hat{N}$
$\lambda = 100$	0.160	98.97
$\lambda = 50$	0.320	48.97

The estimate of N has essentially tracked the prior mean. What went wrong?