
Additional notes on Metropolis–Hastings

This document contains a few notes on Metropolis–Hastings and Markov chains that are not in the Hoff book or the mathematicalmonk videos.

Metropolis–Hastings

Metropolis–Hastings is a generalization of both Gibbs sampling and the Metropolis algorithm. It is similar to Metropolis, but allows for an asymmetric proposal distribution. Consequently, it is more flexible, and can obtain better mixing if one uses a well-chosen proposal distribution, however, it requires that you be able to evaluate the density of the proposal distribution.

Suppose $\pi(x)$ is a p.m.f. on a countable set \mathcal{X} (this is the “target distribution”). For each $x \in \mathcal{X}$, let $q_x(x^*)$ be a p.m.f. on values $x^* \in \mathcal{X}$ (this is the “proposal distribution”). For $x, y \in \mathcal{X}$, define the “acceptance ratio” $\alpha(x, y) = \pi(y)q_y(x)/(\pi(x)q_x(y))$.

Metropolis–Hastings algorithm

- Initialize $x_1 \in \mathcal{X}$.
- For $i = 1, \dots, N - 1$,
 1. Sample $x^* \sim q_{x_i}(x^*)$.
 2. Sample $u \sim \text{Uniform}(0, 1)$.
 3. Set $x_{i+1} = x^*$ if $\alpha(x_i, x^*) > u$, otherwise set $x_{i+1} = x_i$.

Note that steps 2 and 3 can equivalently be written: With probability $\min\{1, \alpha(x_i, x^*)\}$, set $x_{i+1} = x^*$, and otherwise, set $x_{i+1} = x_i$. Thus, the probability of accepting the proposal is $\min\{1, \alpha(x_i, x^*)\}$.

Therefore, the Metropolis–Hastings algorithm defines a Markov chain (X_1, X_2, \dots, X_N) with transition matrix T where $T_{xy} = q_x(y) \min\{1, \alpha(x, y)\}$ when $x \neq y$, and $T_{xx} = 1 - \sum_{y \neq x} T_{xy}$. Assuming $q_x(y) > 0$ and $\pi(x) > 0$ for all $x, y \in \mathcal{X}$, it is easy to verify that this has detailed balance with respect to π :

$$\begin{aligned}\pi(x)T_{xy} &= \pi(x)q_x(y) \min\{1, \alpha(x, y)\} = \pi(x)q_x(y) \min\left\{1, \frac{\pi(y)q_y(x)}{\pi(x)q_x(y)}\right\} \\ &= \min\{\pi(x)q_x(y), \pi(y)q_y(x)\} = \pi(y)q_y(x) \min\left\{\frac{\pi(x)q_x(y)}{\pi(y)q_y(x)}, 1\right\} \\ &= \pi(y)q_y(x) \min\{\alpha(y, x), 1\} = \pi(y)T_{yx}.\end{aligned}$$

Combining moves in Markov chains

Products of transition matrices (deterministic cycle of moves)

If $T^{(1)}, T^{(2)}, \dots, T^{(k)}$ have stationary distribution π , then the product $T = T^{(1)}T^{(2)} \dots T^{(k)}$ has stationary distribution π , since

$$\pi T = \pi T^{(1)}T^{(2)} \dots T^{(k)} = \pi T^{(2)} \dots T^{(k)} = \dots = \pi T^{(k)} = \pi.$$

This fact is used in “fixed-scan” Gibbs sampling, i.e., updating the variables in a deterministically chosen order. This is the version of Gibbs sampling that we have discussed.

Mixtures of transition matrices (random choice of move)

If $T^{(1)}, T^{(2)}, \dots, T^{(k)}$ have stationary distribution π , and $w_1, w_2, \dots, w_k \geq 0$ with $\sum w_i = 1$, then the mixture $T = \sum_{i=1}^k w_i T^{(i)}$ has stationary distribution π , since

$$\pi T = \sum_{i=1}^k w_i \pi T^{(i)} = \sum_{i=1}^k w_i \pi = \pi \sum_{i=1}^k w_i = \pi.$$

This fact is used in “random-scan” Gibbs sampling, i.e., randomly choosing which variable to update at each step (here, w_i is the probability of updating variable i at a given step).

State-dependent moves are typically invalid

Careful! The choice of move (transition matrix) to use at a given iteration should not depend on the current state of the Markov chain, x_i . Using such a state-dependent move can result in a failure to converge to the correct stationary distribution.