

# STAT 238 - Bayesian Statistics

## Lecture Thirty One

Spring 2026, UC Berkeley

Aditya Guntuboyina

15 April 2026

We will study the MALA and HMC algorithms. MALA stands for Metropolis Adjusted Langevin Algorithm. HMC stands for Hamiltonian Monte Carlo. Before we discuss MALA, let us first recall the basics of markov chains.

### 1 Recap: Markov Chains, detailed balance and the Metropolis-Hastings Algorithm

Let  $S$  be a state space. A sequence of random variables  $\Theta^{(0)}, \Theta^{(1)}, \dots$  taking values in  $S$  is a Markov Chain on  $S$  provided the conditional distribution of  $\Theta^{(t+1)}$  given  $\Theta^{(t)} = \theta^{(t)}, \Theta^{(t-1)} = \theta^{(t-1)}, \dots, \Theta^{(0)} = \theta^{(0)}$  only depends on  $\theta^{(t)}$ . Further, if this conditional distribution is the same for all  $t$ , then the Markov chain is said to be time-homogeneous. We will only deal with time-homogeneous Markov Chains in this course.

A (time-homogeneous) Markov Chain is described via the conditional distribution of  $\Theta^{(t+1)}$  given  $\Theta^{(t)} = x$ , which is described by a probability transition kernel  $P(x, y)$ . If  $S$  is discrete,  $P(x, y)$  is the conditional probability of  $\Theta^{(t+1)} = y$  given  $\Theta^{(t)} = x$ . If  $S$  is an open subset of  $\mathbb{R}^d$ ,  $P(x, y)$  represents the conditional density of  $\Theta^{(t+1)}$  given  $\Theta^{(t)} = x$ .

The Markov chain given by  $P(x, y)$  satisfies **detailed balance** with respect to a probability  $\pi(x)$  on  $S$  if the following condition is satisfied:

$$\pi(y)P(y, x) = \pi(x)P(x, y) \quad \text{for all } x, y \in S. \quad (1)$$

This detailed balance condition implies that  $\pi$  is stationary for the chain  $P$  in the sense that  $\Theta^{(t)} \sim \pi$  implies  $\Theta^{(t+1)} \sim \pi$ .

Stationarity along with the additional condition of irreducibility (which means that for every pair of points  $x$  and  $y$  in the state space, there is a positive probability that the chain goes from  $x$  to  $y$  in finite time) imply ergodicity which means:

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N g(\Theta^{(t)}) = \int g(\theta) d\pi(\theta)$$

almost surely.

Metropolis-Hastings is the most widely used method for constructing markov chains satisfying detailed balance with respect to  $\pi$ . This method starts with an arbitrary transition

kernel  $Q(x, y)$  that may not even have anything to do with  $\pi$  (in particular, it is not at all necessary that  $Q$  satisfies detailed balance with respect to  $\pi$ ). Metropolis-Hastings uses  $Q(\cdot, \cdot)$  and  $\pi(\cdot)$  to construct a new Markov chain  $\{\Theta^{(t)}\}$  which moves as follows. Given  $\Theta^{(t)} = x$ , first use  $Q(x, \cdot)$  to generate  $y$ , then take  $\Theta^{(t+1)}$  to be:

$$\Theta^{(t+1)} = \begin{cases} y, & \text{with probability } \min\left(1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right), \\ x, & \text{with probability } 1 - \min\left(1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right). \end{cases}$$

The resulting Metropolis-Hastings chain satisfies detailed balance with respect to  $\pi$ , as we saw previously in Lecture 25.

## 2 Random Walk Metropolis (RWM)

The Random Walk Metropolis (RWM) algorithm is arguably the simplest example of a Markov chain which satisfies detailed balance with respect to  $\pi$ . It uses the transition kernel  $Q(x, y)$  which generates proposals via:

$$y = x + \sigma z \quad \text{where } z \sim N(0, I_d). \quad (2)$$

In other words, the proposal transition kernel is given by

$$Q(x, y) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right). \quad (3)$$

Clearly  $Q(x, y) = Q(y, x)$  so the resulting acceptance probability (in the Metropolis-Hastings algorithm) is simply

$$\min\left(1, \frac{\pi(y)}{\pi(x)}\right).$$

The choice of  $\sigma$  (in (3)) is crucial to the practical performance of the algorithm. If  $\sigma$  is not small, then  $Q(x, \cdot)$  may result in proposals  $y$  that are far from  $x$ . If  $\pi(\cdot)$  is highly concentrated (as is typically the case with posteriors), then  $\pi(y)$  might be much smaller than  $\pi(x)$  leading to the chain rejecting often and getting stuck. On the other hand if  $\sigma$  is too small, then the chain might take a long time to explore the full distribution  $\pi(\cdot)$ .

## 3 Metropolis Adjusted Langevin Algorithm (MALA)

In MALA, the RWM proposal (2) is modified with an expression that depends on  $\pi$ . To find this formula, let us first consider the following proposal:

$$y = x + g(x) + \sigma z \quad \text{with } z \sim N(0, I_d) \quad (4)$$

for some function  $g(x)$  depending on  $\pi(x)$ . Clearly (2) can be seen as a special case of (4) corresponding to  $g(x) = 0$ . The proposal transition kernel corresponding to (4) is given by:

$$Q(x, y) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(-\frac{1}{2\sigma^2}\|y - x - g(x)\|^2\right).$$

Note that  $Q(x, y)$  is no longer equal to  $Q(y, x)$ . The key question is how to choose  $g(x)$ . It is natural to choose  $g(x)$  so that the detailed balance condition holds with respect to  $\pi$ :

$$\pi(x)Q(x, y) \approx \pi(y)Q(y, x)$$

at least approximately when  $y$  is close to  $x$ . Check that

$$\frac{Q(y, x)}{Q(x, y)} = \exp\left(-\frac{1}{\sigma^2} \langle y - x, g(x) + g(y) \rangle + \frac{1}{2\sigma^2} (\|g(x)\|^2 - \|g(y)\|^2)\right).$$

As a result:

$$\begin{aligned} & \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)} \\ &= \exp(\log \pi(y) - \log \pi(x)) \exp\left(-\frac{1}{\sigma^2} \langle y - x, g(x) + g(y) \rangle + \frac{1}{2\sigma^2} (\|g(x)\|^2 - \|g(y)\|^2)\right). \end{aligned}$$

For  $y$  close to  $x$ , we use the approximations:

$$\begin{aligned} \log \pi(y) - \log \pi(x) &\approx \langle y - x, \nabla \log \pi(x) \rangle \\ \frac{1}{\sigma^2} \langle y - x, g(x) + g(y) \rangle &\approx \frac{2}{\sigma^2} \langle y - x, g(x) \rangle \\ \frac{1}{2\sigma^2} (\|g(x)\|^2 - \|g(y)\|^2) &\approx 0. \end{aligned}$$

This leads to

$$\frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)} \approx \exp\left(\left\langle y - x, \nabla \log \pi(x) - \frac{2g(x)}{\sigma^2} \right\rangle\right)$$

when  $y$  is close to  $x$ . This clearly suggests taking

$$g(x) = \frac{\sigma^2}{2} \nabla \log \pi(x).$$

This leads to the MALA proposal:

$$y = x + \frac{1}{2}\sigma^2 \nabla \log \pi(x) + \sigma z \quad \text{where } z \sim N(0, I_d). \quad (5)$$

When  $\sigma$  is small,  $y$  is automatically close to  $x$ , so the above argument can be used to show that the Metropolis-Hastings acceptance ratio for this proposal is nearly equal to 1.

## 4 Mechanics Interpretation

We can interpret the RWM proposal (2) in physical terms as follows:

$$\begin{aligned} x(0) &= x, \quad \dot{x}(0) = z \sim N(0, I_d) \\ \text{and, motion with constant velocity } z &\text{ for time } [0, \sigma] \\ \implies x(\sigma) &= \text{RWM proposal} \end{aligned} \quad (6)$$

Similarly, we can interpret the MALA proposal (5) in physical terms as follows:

$$\begin{aligned} x(0) &= x, \quad \dot{x}(0) = z \sim N(0, I_d) \\ \text{and, motion with constant acceleration } \nabla \log \pi(x) &\text{ for time } [0, \sigma] \\ \implies x(\sigma) &= \text{MALA proposal} \end{aligned} \quad (7)$$

The physical analogy makes the difference between RWM and MALA transparent. In both cases the particle starts at  $x$  with a random velocity  $z \sim N(0, I_d)$ , and the proposal  $y = x(\sigma)$  is the position after time  $\sigma$ . The distinction lies in the *dynamics*:

- **RWM** is a free particle: no force acts, so the particle travels in a straight line at constant velocity. The proposal is purely isotropic and *blind* to the target  $\pi$  — it has no tendency to move toward regions of high probability.
- **MALA** subjects the particle to a constant acceleration  $\nabla \log \pi(x)$ , i.e. a force pointing in the direction of steepest ascent of  $\log \pi$ . By Newton's second law,

$$x(\sigma) = x + \sigma z + \frac{\sigma^2}{2} \nabla \log \pi(x),$$

which is precisely the Langevin proposal (5). The drift term  $\frac{\sigma^2}{2} \nabla \log \pi(x)$  biases every proposal *uphill* in  $\log \pi$ , so the chain is steered toward high-probability regions before the Metropolis correction is even applied.

As a consequence, MALA proposals are systematically better aligned with  $\pi$  than RWM proposals. In practice this allows MALA to take much larger steps  $\sigma$  while maintaining a reasonable acceptance rate, leading to faster exploration of the target distribution.

In the coming lectures, we shall look at HMC, which is a further modification of (7). Specifically, the HMC proposal is obtained via:

$$\begin{aligned} x(0) &= x, \quad \dot{x}(0) = z \sim N(0, I_d) \\ \ddot{x}(t) &= \nabla \log \pi(x(t)) \\ \implies x(\sigma) &= \text{HMC proposal} \end{aligned} \tag{8}$$

The step from MALA to HMC is precisely the step from *constant* acceleration to *position-dependent* acceleration. In MALA the acceleration  $\nabla \log \pi(x)$  is evaluated once at the current position  $x$  and held fixed for the entire flight time  $[0, \sigma]$ . In HMC the acceleration is updated continuously along the trajectory: at every time  $t$  the particle feels  $\nabla \log \pi(x(t))$ . One downside to this time-dependent acceleration is that  $x(\sigma)$  usually cannot be written in closed form as a function of  $x$  and  $z$  (unlike RWM and MALA), and some numerical integration scheme has to be used to figure out  $x(\sigma)$ .