Gradient-based Monte Carlo sampling methods

Johannes von Lindheim

31. May 2016

Abstract

Notes for a 90-minute presentation on gradient-based Monte Carlo sampling methods for the Uncertainty Quantification seminar at Freie Universität Berlin, held on Tuesday, the 31st of May, 2016. Mostly based on resources from Wikipedia and papers by Girolami and Calderhead [2] and Neal [3].

1 Introduction

Let $(\Theta, \mathcal{A}, \pi)$ be probability space, and in this talk, we will have $\Theta := \mathbb{R}^D$. We will assume throughout that $\pi > 0$. We are in the situation that π is analytically intractable and we can only evaluate $\pi(\theta)$ for any $\theta \in \Theta$. Therefore, we want a finite sequence of independent samples $(\theta_1, \ldots, \theta_N) \in \Theta^N$ distributed according to π, which can be used to approximate the distribution (e.g. to generate a histogram or to approximate expectations of the form $\mathbb{E}_\pi[X] \approx \frac{1}{N} \sum_{i=1}^N X(\theta_i)$ where $X \sim \pi$).

1.1 Rejection sampling

It may be natural to generate independent samples $(\theta, \xi) \in \Theta \times [0, 1]$ uniformly and keep only those, for which $\xi \leq \pi(\theta)$. The projections θ are then distributed according to π – an analogy is the throwing darts uniformly onto a dart board and take of (reject) the darts, that did not hit the board. The remaining darts will be distributed uniformly within the board.

The problem is that the probability of $\xi \leq \pi(\theta)$ decreases exponentially with n. It is therefore a good idea to not just throw in random guesses, but to stay in about the same region, if "got under" the distribution, i.e. to walk around according to a Markov chain, that has the stationary distribution π. If we then walk a big enough number of steps along this Markov chain, we will have one sample from π.

1.2 Metropolis-Hastings-Algorithm

The most basic idea to do this goes as follows: Define some proposal density function $\omega(\theta^*|\theta)$ we can sample from, e.g. $\omega(\theta^*|\theta) = \mathcal{N}(\theta^*|\theta, \Lambda)$ (i.e. a random walk) and some initial configuration θ^0. For one Metropolis-Hastings step, proceed according to the following algorithm:

Draw new proposal state θ^* according to $\omega(\theta^*|\theta^k)$
Calculate the acceptance probability $\alpha = \min\{1, \frac{\pi(\theta^*)\omega(\theta^k|\theta^*)}{\pi(\theta^k)\omega(\theta^*|\theta^k)}\}$
Sample $u \in [0, 1]$ uniformly
if $u \leq \alpha$ **then**
  $\theta^{k+1} := \theta^*$ (accept)
else
  $\theta^{k+1} := \theta^k$ (reject)
end

1.3 Detailed Balance

One can easily verify, that the Markov chain constructed by this algorithm is reversible with respect to π, i.e. π satisfies detailed balance. What we want want to show is that $\pi(\theta)\pi(\theta^*|\theta) = \pi(\theta^*)\pi(\theta|\theta^*)$ (where $\pi(\theta^*|\theta)$ is the transition probability we control), i.e. we have a symmetry in θ and θ^*. We can calculate this as follows:

$$\begin{aligned}\pi(\theta)\pi(\theta^*|\theta) &= \pi(\theta)\omega(\theta^*|\theta)\min\left\{1, \frac{\pi(\theta^*)\omega(\theta|\theta^*)}{\pi(\theta)\omega(\theta^*|\theta)}\right\} \\ &= \min\{\pi(\theta)\pi(\theta^*|\theta), \pi(\theta^*)\pi(\theta|\theta^*)\}\end{aligned}$$

since all probabilities are greater or equal to zero and we have the desired symmetry.

Now of course, detailed balance implies, that the Markov chain has the unique equlilibrium distribution π:

$$\int \pi(\theta)\pi(\theta^*|\theta)d\theta \overset{\text{DB}}{=} \int \pi(\theta^*)\pi(\theta|\theta^*)d\theta = \pi(\theta^*)\int \pi(\theta|\theta^*)d\theta = \pi(\theta^*).$$

Furthermore, our Markov chain is obviously aperiodic and irreducible and therefore ergocic – hence, if we do T such steps for T large, we will get one sample from approximately π.

However, there are several disadvantages of this method:

- Although convergence to π is guaranteed, the initial samples may follow a very different distribution, especially if θ^0 lies in a region of low density. As a consequence, typically one has to throw an initial number of samples away (*burn-in period*).
- θ^k and θ^{k+1} are correlated. Although we still have $\frac{1}{n}\sum_{i=1}^n f(X_i) \overset{\text{a.s.}}{\to} \mathbb{E}[f(X)]$ for $X \sim \pi$ and some quantity of interest f, if we want independent samples, we need to throw away the results of $T-1$ intermediate steps between samples for T large.

Especially the last downside is a problem of the Metropolis-Hastings algorithm: If we want to get low autocorrelation times, we need to make big steps (i.e. choose a large variance for ω). But since we perform a random walk on the probability space Θ, we will get low acceptance probabilities and explore the space very slowly.

2 Metropolis-adjusted Langevin algorithm (MALA)

2.1 Langevin Dynamics

The idea of MALA is to exploit a somewhat "smooth" structure of π: If we have accepted a step, it might be a good idea to keep on walking into that direction, assuming that we get into regions of even higher probabilities.

Since we are viewing the Markov chain as "moving" through Θ and now want to do that in a more elaborate way, it makes sense to define these dynamics in terms of some dynamics. MALA is based on a Langevin diffusion, defined by the stochastic differential equation (SDE)

$$d\theta(t) = \frac{1}{2}\nabla_\theta \mathcal{L}(\theta(t))dt + db(t) \tag{1}$$

where $\mathcal{L}(\theta) \equiv \log \pi(\theta)$ and b denotes a D-dimensional Brownian motion.

2.2 Discretization

A first-order Euler discretization of the SDE gives the proposal mechanism

$$\theta^* = \theta^k + \frac{1}{2}\varepsilon \nabla_\theta \mathcal{L}(\theta^k) + \sqrt{\varepsilon} z^k \tag{2}$$

where $z \sim \mathcal{N}(z|0, I)$ and ε is the integration step size. Since we introduce a first-order integration error with this, we need to perform a Metropolis accept-reject step, since otherwise convergence to π is not guaranteed. The acceptance probability takes the standard form

$$\alpha = \min \left\{ 1, \frac{\pi(\theta^*)\omega(\theta^k|\theta^*)}{\pi(\theta^k)\omega(\theta^*|\theta^k)} \right\} \tag{3}$$

2.3 Optimal Scaling for $D \to \infty$

In practice, as with MH, we need to decide which stepsize ε, i.e. which "scaling" to use: if the step size, is too large, we will have low acceptance probability, because we will most likely end in a region of much smaller probability, but if it is too small, then we will explore the sample space Θ very slowly and have to throw away most of the samples because of high correlations. As the dimension D increases, one would intuitively guess that ε should decrease, as D increases, which is the case. In fact, it needs to be chosen proportionally to $D^{-\frac{1}{3}}$ and one can derive an asymptotic average acceptance rate of 0.574. The exact result can be found in Roberts and Rosenthal [5].

This yields some practical guidance, how to choose ε in medium to high dimensions D: One should tune the proposal variance, so that the average acceptance rate is about 0.574.

So even in these very simple spaces, MALA scales better than Metropolis-Hastings. We can expect even greater differences in mixing behaviors, when our space is more complicated, e.g. we have strongly correlated variables.

2.4 Benefits and Downsides

Since now a drift term drift term in the proposal mechanism based on the gradient information is introduced, if we have some smoothness condition fulfilled for our target density π, we will much more likely get proposals in directions of higher probabilities, and therefore higher acceptance probabilities.

We will also see, that MALA scales much better than MH: Asymptotically, MH has an optimal acceptance rate of 0.234. We will derive this below, and more formally this is obtained by Roberts et al. [4].

Nevertheless, it is clear that the isotropic diffusion will be inefficient for strongly correlated variables with widely differing variances forcing the step size ε to accomodate the variate with smallest variance. This issue can be circumvented by using a preconditioning matrix $M \in \mathbb{R}^{D \times D}$, such that

$$\theta^* = \theta^k + \frac{1}{2}\varepsilon M \nabla_\theta \mathcal{L}(\theta^k) + \sqrt{\varepsilon M} z^k. \tag{4}$$

One can obtain $\sqrt{M}$ by diagonalization of M or Cholesky decomposition.

The problem is, that there is no systematic way or guideline, how to choose that matrix in a principled manner – indeed, it might be even inappropriate for the starting phase of the Markov chain (see [1]).

3 Hamiltonian Monte Carlo (HMC)

Hamiltonian dynamics has a physical interpretation that can provide useful intuitions. In two dimensions, we can visualize the dynamics as that of a frictionless ball that rolls over a surface of varying height. The state of this system consists of the position of the ball, given by a two-dimensional vector q, and the momentum of the ball (its mass times its velocity), given by a two-dimensional vector p. The potential energy, $U(q)$, of the ball is proportional to the height of the surface at its current position, and its kinetic energy, $K(p)$, is equal to $\frac{|p|^2}{2m}$, where m is the mass of the ball. On a level part of the surface, the ball moves at a constant velocity, equal to $\frac{p}{m}$. If it encounters a rising slope, the ball's momentum allows it to continue, with its kinetic energy decreasing and its potential energy increasing, until the kinetic energy (and hence p) is zero, at which point it will roll back down (with kinetic energy increasing and potential energy decreasing). In nonphysical MCMC applications of Hamiltonian dynamics, the position will correspond to the variables of interest. The potential energy will be minus the log of the probability density for these variables. Momentum variables, one for each position variable, will be introduced artificially.

3.1 Hamilton Dynamics

Hamiltonian dynamics operates on a D-dimensional position vector, q,and a D-dimensional momentum vector, p, so that the full state space has $2D$ dimensions. The system is described by a function of q and p known as the Hamiltonian, $H(q, p)$.

The partial derivatives of the Hamiltonian determine how q and p change over time, t, according to Hamilton's equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \tag{5}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \tag{6}$$

for $i = 1, \ldots, D$.

For HMC, we want to use a Hamiltonian function, that can be written as

$$H(q, p) = U(q) + K(p) \tag{7}$$

where $U(q)$ is called the *potential energy*, defined as

$$U(q) := -\log \pi(q)$$

where π is the distribution we want to sample from (plus a convenient constant) and $K(p)$ is called the *kinetic energy*, usually defined as

$$K(p) := \frac{1}{2} p^T M^{-1} p$$

Therefore, we can write (5) as

$$\frac{dq_i}{dt} = [M^{-1}p]_i,$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}.$$

In practical applications, one typically chooses $M = \text{diag}(m_1, \ldots, m_D)$.

3.2 Discretization

The most frequent used method of discretizing this dynamics for calculations (which is hard to beat in practice) works as follows:

$$
\begin{aligned}
p_i\left(t+\frac{\varepsilon}{2}\right) &= p_i(t) - \frac{\varepsilon}{2}\frac{\partial U}{\partial q_i}(q(t)) \\
q_i(t+\varepsilon) &= q_i(t) + \varepsilon\frac{p_i\left(t+\frac{\varepsilon}{2}\right)}{m_i} \\
p_i\left(t+\varepsilon\right) &= p_i\left(t+\frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2}\frac{\partial U}{\partial q_i}(q(t+\varepsilon))
\end{aligned}
$$

If we want to go to $t + 2\varepsilon$, we do not have to apply this whole update scheme twice, but instead we can combine the last half step of the first update with the first half step of the second update. The method then looks very similar to an Euler approximation, except that the momentum variables computed are shifted by $\frac{\varepsilon}{2}$, which makes this method second order: Its local error is of order ε^3 and the global error of order ε^2. The Euler instead would have order ε^2 error and order ε error.

3.3 Properties

Several properties of Hamiltonian dynamics are crucial to its use in constructing MCMC updates.

3.3.1 Time Reversibility

First, Hamiltonian dynamics is reversible: The mapping T_s from the state at time t,$(q(t), p(t))$, to the state at time $t + s$, $(q(t + s), p(t + s))$, is one-to-one, and hence has an inverse, T_s. This inverse mapping is obtained by simply negating the time derivatives in equations (5) and (6).

This reversibility is important for showing that MCMC updates that use the dynamics leave the desired distribution π invariant, since this is most easily proved by showing reversibility of the Markov chain transitions, which requires reversibility of the dynamics used to propose a state.

3.3.2 Conservation of Hamiltonian

A second property of the dynamics is that the Hamiltonian stays invariant. This is easily seen by looking at the time derivative of H:

$$
\begin{aligned}
\frac{dH}{dt} &= \sum_{i=1}^{D} \frac{dq_i}{dt}\frac{\partial H}{\partial q_i} + \frac{dp_i}{dt}\frac{\partial H}{\partial p_i} \\
&= \sum_{i=1}^{D} \frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i} - \frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i} \\
&= 0.
\end{aligned}
$$

For Metropolis updates using a proposal found by Hamiltonian dynamics, which form part of the HMC method, the acceptance probability is one if H is kept invariant. We will see later, however, that in practice we can only make H approximately invariant, and hence we will not quite be able to achieve this.

3.3.3 Volume Preservation

A third fundamental property of Hamiltonian dynamics is that it preserves volume in (q, p) space: If we apply the mapping T_s to the points in some region R of (q, p) space, with volume V, the image of R under T_s will also have volume V. This can be proved in several ways, one is to note that the divergence if the vector field defined by (5) and (6) is zero:

$$\begin{aligned}\sum_{i=1}^{D} \frac{\partial}{\partial q_i}\frac{dq_i}{dt} + \frac{\partial}{\partial p_i}\frac{dp_i}{dt} &= \sum_{i=1}^{D} \frac{\partial}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial}{\partial p_i}\frac{\partial H}{\partial q_i} \\ &= \sum_{i=1}^{D} \frac{\partial H^2}{\partial p_i \partial q_i} - \frac{\partial^2 H}{\partial p_i \partial q_i} \\ &= 0.\end{aligned}$$

As the divergence is the rate of change in volume, a vector field has zero divergence exactly when the flow is volume preserving.

3.4 MCMC using Hamiltonian Dynamics

Using Hamiltonian dynamics to sample from a distribution requires translating the density function for this distribution to a potential energy function and introducing "momentum" variables to go with the original variables of interest (now seen as "position" variables). We can then simulate a Markov chain in which each iteration resamples the momentum and then does a Metropolis update with a proposal found using Hamiltonian dynamics.

3.4.1 Canonical Distribution

The distribution we wish to sample from can be related to a potential energy function via the *canonical distribution* from statistical mechanics, which is

$$P(x) = \frac{1}{Z} \exp\left(-\frac{E(x)}{T}\right)$$

where in our case, $E(x) = H(q, p)$ and we choose $T = 1$. If $H(q, p) = U(q) + K(p)$, the joint density is

$$P(q, p) = \frac{1}{Z} \exp\left(-U(q)\right) \exp\left(-K(p)\right) \tag{8}$$

and we see that q and p are independent, and each have canonical distributions. We will use q to represent the variables of interest and introduce p just artificially to let the Hamiltonian dynamics operate.

3.4.2 The HMC Algorithm

Now that we know Hamilton Dynamics, we can present the Hamiltonian Monte Carlo Algorithm.

Draw new proposal state $p_i \sim \mathcal{N}(0, m_i)$
Simulate Hamiltonian Dynamics for L leapfrog steps, negate p, which yields (q^*, p^*)
Calculate the acceptance probability $\alpha = \min\{1, \exp(-H(q^*, p^*) + H(q, p)\}$
Sample $u \in [0, 1]$ uniformly
if $u \leq \alpha$ **then**
  $(q, p)^{k+1} := (q^*, p^*)$ (accept)
else
  $(q, p)^{k+1} := (q, p)^k$ (reject)
end

3.5 Detailed Balance

Since the proposal distribution ω is analytically not as easily accessible as in the MALA or MH-Algorithm, we have to check that the acceptance probability is chosen right to satisfy detailed balance, or in other words, the joint density $H(q,p)$ (and therefore the marginal density of q) is left invariant.

Let us therefore partition the (q,p)-space into regions A_l, each with small volume V. Let the image of A_l under L leapfrog steps and negation of p be B_l. Due to reversibility of the leapfrog steps and negation, the B_l will also partition the space and because of volume preservation of the leapfrog steps, each B_l also has volume V. Detailed balance holds, if $\forall i,j$

$$p(A_i)\rho(B_j|A_i) = p(B_j)\rho(A_i|B_j), \tag{9}$$

where p is the probability under the canonical distribution and $\rho(X|Y) = (\alpha \circ \omega)(X|Y)$ is the conditional probability of proposing *and* accepting a move to region X when being in state Y (where the proposal probability equals to 1, since we are calculating the Hamilton dynamics deterministically). Clearly, when $i \neq j$, $\rho(B_j|A_i) = \rho(A_i|B_j) = 0$ and equation (9) is satisfied. Otherwise, since H is continuous almost everywhere, in the limit as the regions A_l and B_l become smaller, H becomes effectively constant within each region, with value H_X in region X, and hence the canonical probability density and the transition probabilities become effectively constant within each region as well. We can now (for infinitesimal A_l, B_l) rewrite the left side of (9) for $i = j$ (say, both equal to l) as

$$\begin{aligned}\frac{V}{Z}\exp(-H_{A_l})\min\{1,\exp(-H_{B_l}+H_{A_l})\} &= \frac{V}{Z}\min\{\exp(-H_{A_l}),\exp(-H_{B_l})\}\\ &= \frac{V}{Z}\exp(-H_{B_l})\min\{\exp(-H_{A_l}+H_{B_l}),1\},\end{aligned}$$

which is precisely equation (9).

We also see now, how the volume preservation property is coming in handy here: Otherwise, the acceptance probability would need to be dependent on these volume changes, since we would not have $p(X) = \frac{V}{Z}\exp(-H_X)$ and $p(Y) = \frac{V}{Z}\exp(-H_Y)$ for the same V.

3.6 Optimal Scaling for $D \to \infty$

3.6.1 Choosing the Stepsize

Now, let us not state a theorem, but derive some practical guidance for scaling HMC right in comparison to MH at least briefly and informally in the asymptotics for the special case as in MALA, so we are now assuming $U(q) = \sum u_i(q_i)$ for our potential energy function, where the functions u_i are drawn independently from some distribution, that is the u_i are iid. Let us write Δ_1 for the energy difference $E(x^*) - E(x)$ of a single variable ($x = (q_i, p_i)$ and $E(x) = u_i(q_i) + \frac{p_i^2}{2}$) or Δ_D for $E(x^*) - E(x)$ for the whole system ($x = (q,p)$ and $E(x) = U(q) + K(p)$).

Let us first note, that because of volume preservation, we have $dqdp = dq^*dp^*$ for infinitesimal volume elements and we can derive

$$\begin{aligned}1 &= \frac{1}{Z}\int \exp(-E(q^*,p^*))\; dq^*dq^*\\ &= \frac{1}{Z}\int \exp(-(E(q^*,p^*) - E(q,p)))\exp(-E(q,p))\; dqdp\\ &= \mathbb{E}_{(q,p)\sim H}[\exp(-(E(q^*,p^*) - E(q,p)))]\\ &= \mathbb{E}_{x\sim H}[\exp(-\Delta)].\end{aligned}$$

Form Jensen's inequality, we get

$$\begin{aligned} 1 = \mathbb{E}[\exp(-\Delta)] &\geq \exp(\mathbb{E}[-\Delta]) \\ \Leftrightarrow \mathbb{E}[\Delta] &\geq 0. \end{aligned}$$

Now as $D \to \infty$, as each Δ_1 has positive mean and Δ_D is the sum of Δ_1 for each variable, $\Delta_D \to \infty$ if we fix the stepsize ε for HMC resp. the standard deviation ζ of the proposal distribution for MH. Therefore, the acceptance probability $\min\{1, \exp(-\Delta_D))\}$ decreases. The only hope is that, if we do not decrease the stepsize by to much, we have a large enough variance of Δ, so that would have sufficiently many proposals with negative energy difference, which are accepted automatically. But this hope is destroyed right away: As $D \to \infty$ and $\zeta, \varepsilon \to 0$, $\Delta \to 0$ as well. Using a second-order approximation of $\exp(-\Delta_1)$ as $1 - \Delta - 1 + \frac{\Delta_1^2}{2}$, we find

$$\begin{aligned} 1 = \mathbb{E}[\exp(-\Delta_1)] &\approx \mathbb{E}[1 - \Delta - 1 + \frac{\Delta_1^2}{2}] \\ \Leftrightarrow \mathbb{E}[\Delta_1] &\approx \frac{1}{2}\mathbb{E}[\Delta_1^2] \end{aligned}$$

so that, for small Δ_1 and $\mathbb{E}[\Delta_1]^2 \approx 0$, by summing over all variables, we get

$$2\mathbb{E}[\Delta_D] \approx \mathrm{Var}[\Delta_D]. \tag{10}$$

Therefore, to have enough accepted proposals, we must keep the mean of Δ_D near 1.

We can now see, how we need to scale ζ by directly averaging Δ_1 for a symmetric proposal. Suppose the proposal for one variable $x^* = x + c$ for $c = a$ and $c = -a$ equally likely. Then we might approximate $U(x^*)$ to second order as

$$U(x^*) \approx U(x) + cU'(x) + \frac{c^2 U''(x)}{2}$$

and we find

$$\begin{aligned} \mathbb{E}_{c\in\{a,-a\}}[\Delta_1] &= \mathbb{E}_{c\in\{a,-a\}}[U(x^*) - U(x)] \\ &\approx \mathbb{E}_{c\in\{a,-a\}}[cU'(x) + \frac{c^2 U''(x)}{2}] \\ &= \frac{a^2}{2}U''(x). \end{aligned}$$

Averaging over a, we see that

$$\begin{aligned} \mathbb{E}_a[\Delta_1] &= \mathbb{E}_a[\frac{a^2}{2}U''(x)] \\ &= \frac{U''(x)}{2}\mathbb{E}_a[a^2] \\ &= \frac{U''(x)}{2}\mathrm{Var}[a^2] \\ &\sim \zeta^2 \end{aligned}$$

so that

$$\mathbb{E}[\Delta_D] \sim D\zeta^2. \tag{11}$$

Therefore, to maintain a reasonable acceptance rate, we must choose $\zeta \sim D^{-\frac{1}{2}}$. Since the number of iterations needed to reach a nearly independent point is proportional to $\zeta^{-2} \sim D$ (from the definition of the autocorrelation function), we have a computation time $= O(D^2)$.

Turning to HMC, we saw that the error in H when using the leapfrog method to simulate a trajectory of a fixed length is proportional to ε^2. The error in H for a single pair (q_i, p_i) is Δ_1, so $\Delta_1^2 \sim \varepsilon^4$. Therefore,

$$\mathbb{E}[\Delta_D] \sim D\varepsilon^4 \tag{12}$$

so that we need to choose $\varepsilon \sim D^{-\frac{1}{4}}$. The number of leapfrog updates to reach a nearly independent point scales as $\varepsilon^{-1} \sim D^{\frac{1}{4}}$ and we have a computation time $= O(D^{\frac{5}{4}})$.

3.6.2 Optimal Acceptance Rates

Knowing how to choose ε and ζ in the asymptotics, we can determine the average acceptance rate, if that optimal choice is used. This is helpful when tuning the algorithm (see below) – provided, of course, that the distribution sampled is high-dimensional and has properties that are adequately modeled by a distribution with replicated variables.

To find this acceptance rate, we remind ourselves that we have detailed balance (which uses, that we have reached equilibrium – for the first (transient) phase, our results will not apply, see [1]). Therefore, the probability of making an accepted move with Δ_D negative is the same as making an accepted move with Δ_D positive. Since moves with negative Δ_D are always accepted, we observe that simply $P(\text{accept}) = 2P(\Delta_D \leq 0)$.

For large D, we have that $\Delta_D = \sum \Delta_1$ is Gaussian because of the Central Limit Theorem, so $P(\Delta_D \leq 0)$ is essentially the CDF of a Gaussian, evaluated at zero. We know from (10) that the variance of Δ_D is twice its mean $\mathbb{E}[\Delta_D] = \mu$, so

$$P(\text{accept}) = 2P(\Delta_D \leq 0) = 2\Phi\left(\frac{0-\mu}{\sqrt{2\mu}}\right) = 2\Phi\left(-\sqrt{\frac{\mu}{2}}\right) = a(\mu) \tag{13}$$

where $\Phi(z)$ is the CDF of a standard Gaussian.

For random-walk Metropolis, we have a computation time proportional to $\frac{1}{a\zeta^2}$ where a is the acceptance rate. As $\mu = \mathbb{E}[\Delta_D] \sim \zeta^2$, we have a computation cost C_{MH} with a proportionality of

$$C_{\text{MH}} \sim \frac{1}{a(\mu)\mu}. \tag{14}$$

Numerical calculation shows that this is minimized when $\mu = 2.8$ and $a(\mu) = 0.23$. Therefore, in high dimensions, it is best to tune MH such that about every fourth move is accepted.

Looking at HMC, the cost of obtaining an independent point will be proportional to $\frac{1}{a\varepsilon}$, and as we saw above that μ is proportional to ε^4. From this we obtain

$$C_{\text{HMC}} \sim \frac{1}{a(\mu)\mu^{\frac{1}{4}}}. \tag{15}$$

Numerical calculation shows that this is minimized when $\mu = 0.41$ and $a(\mu) = 0.65$.

3.7 Benefits and Downsides

Since we are doing L (deterministic) leapfrog steps, we are avoiding an inefficient exploration of the sample space, which one has with random walks: Note that the variance of the position after n steps in MH grows proportionally to n (until the amount of movement becomes comparable to the size of the sample space Θ), since the position is the sum of mostly independent movements for each iteration. The *standard deviation*, which gives the typical amount of movement, is therefore proportional to $\sqrt{n}$.

The stepsize of HMC is similarly limited by the most constrained direction, but the movement will be in the same direction for many steps. The distance moved after n steps will therefore tend to be proportional to n, until the distance moved becomes comparable to the overall width of the distribution. The advantage compared to movement by a random

walk will be a factor roughly equal to the ratio of the standard deviations in the least confined direction and most confined direction.

Furthermore, similar to MALA, HMC has a much better scaling when going up in dimensions.

Unfortunately, although we know somewhat how to tune ε using some pilot runs and looking at the acceptance rate, there is again no way for choosing the mass matrix M in a principled manner. One typically takes diagonal matrices with entries m_i, but even choosing these is sort of a "black magic". In practice, this problem is often handled by using some pilot runs and trace plots, or looking at the autocorrelation function.

References

[1] Christensen, O. F., Roberts, G. O. and Rosenthal, J. S. (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. Journal of the Royal Statistical Society B, 67, 253–268

[2] Girolami, M., Calderhead, B. (2011). Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. Journal of the Royal Statistical Society B, 73(2), 123-214.

[3] Neal, R. M. (2011). MCMC using Hamiltonian Dynamics. Handbook of Markov Chain Monte Carlo, 2, 113-162.

[4] Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Probab., 7, 110-120.

[5] Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. Journal of the Royal Statistical Society B, 60, 255–268.