Notes on theory of Hamiltonian Monte Carlo and delayed rejection

Alex H. Barnett

May 20, 2021

Abstract

Since the rigorous proof of invariance of HMC for the proposal density is not simply written down anywhere we can find, we summarize it here. Measure theory is needed for mathematical rigor, but we avoid excessively general notation. We do not delve into mixing or convergence rates. We also aim to write down our delayed rejection HMC method rigorously. It also serves as a tutorial, and scratch for our paper.

1 Basics: Markov chains and Metropolis–Hastings

We consider Markov chains on continuous state spaces, using measure theory in order to handle the non-absolutely continuous transition kernels of HMC. Our notation follows, eg Tierney's papers from the 90s, Geyer's and Kennedy's lecture notes. We avoid the very abstract framework of Andrieu et al 2021. Also see graduate-level books such as Hunter & Nachtergaele, or Stein & Shakarchi, for basic measure theory.

1.1 The basics in pdf notation

We consider a continuous state space $S = \mathbb{R}^n$. A state is a point $x \in S$. The goal of MCMC is to sample from a given distribution π over S. In our applications π may be taken to be *absolutely continuous* (AC, with respect to Lebesgue measure dx), meaning that it is described by a nonnegative density *function* (pdf) which without ambiguity we may also call $\pi : S \to \mathbb{R}_{\geq 0}$. The normalization is $\int \pi(x)dx = 1$, although all MCMC methods discussed can handle unnormalized π. All integrals are over S unless otherwise indicated.

A Markov chain is defined by a *transition operator* K, which in the simplest AC case we may also write as a kernel function $k(x, y)$ giving the pdf of the next state y conditioned on the current state x, that is, $p(y|x)$. The convention in probability is that operators act from the left (the opposite of that for usual operators in applied math), so, such as kernel acts on a density π as an integral operator

$$(\pi K)(y) := \int \pi(x)k(x, y)dx \ . \tag{1}$$

A mnemonic for the kernel indices is k(initial,final), which is backwards from the usual in integral equations. $k(x, \cdot)$ is normalized (the Markov property), ie,

$$\int k(x, y)dy = 1 \ , \qquad \forall x \in S \ . \tag{2}$$

Then πK is also normalized: $\int (\pi K)(y)dy = \int \int \pi(x)k(x, y)dxdy = \int \pi(x)[\int k(x, y)dy]dx = \int \pi(x)dx$, where swapping the order of integration is justified by Fubini's theorem because $\pi(x)k(x, y)$ is integrable over $dxdy$.

Invariance (stationarity) of a pdf π with respect to the kernel K is then $\pi K = \pi$ as pdfs, ie

$$\int \pi(x)k(x,y)dx = \pi(y) \ , \qquad \forall y \in S \tag{3}$$

Detailed balance (DB, also called "reversibility") for a kernel (or "chain") is the condition

$$\pi(x)k(x,y) = \pi(y)k(y,x) \ , \qquad \forall x,y \in S \qquad \text{(DB)} \tag{4}$$

The proof that DB implies invariance is just by integration of (4) with respect to dy, then using (2).

M-H involves a *proposal* kernel $q(x,y)$, which we assume is AC for now, and is also normalized so $\int q(x,y)dy = 1$ for all $x \in S$. Starting at x, the M-H step is: draw y from $q(x,y)$ then accept with probability $\alpha(x,y)$ in which case the new x is y, otherwise the new x is x. We could introduce chain notation x_t, $t = 0, 1, \ldots$ but little would be gained at this point. The standard M-H acceptance formula is

$$\alpha(x,y) = \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right) \ , \qquad x,y \in S \ , \tag{5}$$

which obeys

$$\pi(x)q(x,y)\alpha(x,y) = \pi(y)q(y,x)\alpha(y,x) \ , \qquad x,y \in S \ . \tag{6}$$

Note that (6) is more general than (5) since it allows for $q = 0$ for some (x,y). It is easy to show that (6) implies (4) for the case of $y \neq x$, since in that case $k(x,y) = p(y|x) = q(x,y)\alpha(x,y)$, and substituting into (4) and using (6) shows (4) holds for $y \neq x$. However, we have not proved DB, since what is the meaning of $k(x,x)$, which is infinite? What is the meaning of the pointwise equality (4) for $y = x$? These issues do not arise in the discrete state space S case, only the continous one. This motivates definitions using measures (distributions), a generalization of density functions, as in the next section.

However, as a warm-up, one does not need measures to prove that M-H is merely π invariant, avoiding DB, as follows.

Proposition 1. *Let $q(x,.)$ be an AC proposal density, then M-H with acceptance probability* (5) *has π an invariant density.*

Proof. At the current state x, the probability of rejection (called s in Andrieu) is

$$r(x) = \int q(x,z)[1 - \alpha(x,z)]dz = 1 - \int q(x,z)\alpha(x,z)dz \ ; \tag{7}$$

here z is a dummy variable. The action of an M-H step on π is a mixture of an α-weighted proposal and the rejection (no change, $y = x$), resulting in the final density as a function of final state y,

$$\begin{aligned} \int \pi(x)q(x,y)\alpha(x,y)dx + \pi(y)r(y) &= \int \pi(y)q(y,x)\alpha(y,x)dx + \pi(y)r(y) \\ &= \pi(y)\int q(y,x)\alpha(y,x)dx + \pi(y)r(y) \\ &= \pi(y)[1 - r(y)] + \pi(y)r(y) \ = \ \pi(y) \end{aligned}$$

where we applied (6) to the integrand, and to get to the third line used (7). Comparing (3) completes the proof. □

Remark 2. *Any form of $\alpha(x,y)$ that satisfies* (6) *is valid in the above; however,* (5) *is the choice that is most efficient in the sense that any other has higher probability of rejection. This is simply because, for each $x, y \in S$, either $\alpha(x,y)$ or $\alpha(y,x)$ is 1, the largest allowed value for a probability.*

1.2 Distributions and measure theory notation

M-H, and in particular HMC, involves transition kernels for which $k(x,\cdot)$ is not AC. One extension of the function notation that can handle this is to use δ, the Dirac delta distribution, or unit point mass, defined by

$$\delta(x) = 0, \quad x \neq 0, \qquad \int \delta(x)dx = 1 \ .$$

For any function f continuous at 0, we have the sifting property $\int f(x)\delta(x)dx = f(0)$. The transformation rules are, in 1D ($n = 1$), with $f : \mathbb{R} \to \mathbb{R}$ a differentiable function,

$$\delta(ax) = |a|^{-1}\delta(x) \ , \qquad a \in \mathbb{R},\ a \neq 0 \ , \tag{8}$$

$$\delta(f(x)) = \sum_{z:f(z)=0} |f'(z)|^{-1}\delta(x-z) \ . \tag{9}$$

Of course the latter is inapplicable when the RHS is not defined. In general dimension ($n \geq 1$), where $F : S \to S$ is a differentiable map,

$$\delta(F(x)) = \sum_{z:F(z)=0} |\det DF(z)|^{-1}\delta(x-z) \ , \tag{10}$$

where DF is the $n \times n$ Jacobean derivative matrix $\partial_{x_j} F_i$, and 0 is the origin in $\mathbb{R}^n$. By replacing $F(x)$ by $F(x) - b$, for some constant translation $b \in \mathbb{R}^n$,

$$\delta(F(x)-b) = \sum_{z:F(z)=b} |\det DF(z)|^{-1}\delta(x-z) \ . \tag{11}$$

Here z sums over the preimages of b. Subscript notation is also used: $\delta_x(y) = \delta(x-y) = \delta(y-x) = \delta_y(x)$.

For example, in the above case of M-H with an AC proposal density $q(x,y)$, the kernel for the M-H step is a mixture of an AC pdf and a (rejected) point mass at $y = x$,

$$k(x,y) = q(x,y)\alpha(x,y) + \delta_x(y)r(x) \ , \tag{12}$$

which one may check obeys $\int k(x,y)dy = 1$ for all x, recalling (7). See Tierney '98, Kennedy Sec. 8.2. Note, however, that not all non-AC measures are sums of point masses; there may be distributions on intermediate dimension subsets, fractals, etc.

One must generalize the notation to measures rather than functions. To recap this, measures are defined over $(S, \mathcal{S})$, where $\mathcal{S}$ is a "σ-field over S". Loosely speaking, $\mathcal{S}$ is the set of all measurable subsets of S (see any textbook on measure theory). A measure π on $(S, \mathcal{S})$ is defined by $\pi(B)$, i.e., probability of being in B, for all (measurable) sets $B \subset S$ (strictly, $B \in \mathcal{S}$). The normalization is $\pi(S) = 1$. Only when a measure π is AC can we write it as a pdf $\pi(x)$ (here we overload the notation; sometimes a distinct symbol is used), in which case for any $B \in \mathcal{S}$,

$$\pi(B) = \int_B \pi(x)dx \ .$$

Here $\pi(x)dx$ may be thought of as a measure, also written $\pi(dx)$. Lebesgue measure dx is a special case of a measure with unit density function. Equivalent notations (see Kennedy notes p. 89) include in the measure case

$$\pi(B) = \int_B d\pi = \int_B d\pi(x) = \int_B \pi(dx) \ .$$

Two measures π and μ are equal if

$$\pi(B) = \mu(B) \qquad \forall B \in \mathcal{S} \ ;$$

note that this is a *weak* definition of equality. This is sometimes summarized by $\pi(dx) = \mu(dx)$, which is useful since it makes explicit an independent variable x; however it is somewhat imprecise, being analogous to writing "$f(x) = g(x)$" to express equality of f and g as functions.

A transition (Markov) kernel $K(x, \cdot)$ can be considered a measure that depends on the parameter $x \in S$, namely the initial point. The normalization is $K(x, S) = 1$ for any $x \in S$. Then $K(x, B) = p(y \in B|x)$ is the conditional probability of ending in B after one step, when starting in at x. The action of a kernel K on a measure π is (compare (1)), for any "test" set $B \in \mathcal{S}$,

$$(\pi K)(B) = \int_S \pi(dx) K(x, B) \ .$$

The measure version of (3) is as follows (Kennedy (7.33)).

Definition 3 (Invariance for measures)**.** *A measure π is invariant under the Markov kernel K if $\pi K = \pi$, as measures, that is*

$$\int \pi(dx) K(x, B) = \pi(B) \ , \qquad \forall B \in \mathcal{S} \ . \tag{13}$$

The direct proof of M-H invariance for (12) (Prop. 1) in this sense of invariance is done in, eg, Tierney '94 Sec. 2.3.1 and notes by Geyer, Kennedy. They all avoid the route via DB.

The measure (weak) sense of (4) states: the probability of being in set A then going to set B is the same as the other way around, as follows.

Definition 4 (Detailed balance for measures)**.** *A kernel K has detailed balance (reversible) with respect to a measure π if*

$$\int_A \pi(dx) K(x, B) = \int_B \pi(dx) K(x, A) \qquad \forall A, B \in \mathcal{S} \ . \tag{14}$$

A summary of this is $\pi(dx)K(x, dy) = \pi(dy)K(y, dx)$ as measures on $\mathcal{S} \otimes \mathcal{S}$; this needs the standard idea of a product measure. A less abstract way to write it is to return to weak equality of distributions,

$$\int_B \int_A \pi(x) k(x, y) dy dx = \int_B \int_A \pi(y) k(y, x) dx dy \qquad \forall A, B \subset S \ ,$$

using $k(x, y)dy$ to represent the measure $K(x, dy)$.

Proposition 5. *With respect to a measure π, detailed balance of a Markov operator K implies invariance.*

Proof. Choose $A = S$ in (14). The left hand side is $(\pi K)(B)$. On the right hand side $K(x, S) = 1$ for all x is the Markov property, leaving $\pi(B)$. This holds for any $B \in \mathcal{S}$. □

1.3 M-H with proposals that are measures

Tierney '98 and Andrieu prove that M-H with acceptance a generalization of (5) obeys DB in the measure sense; they are too abstract for our purposes. We now instead show the same in a simpler way.

We need M-H general measure proposal $Q(x, dy)$, which might not be represented by an AC kernel $q(x, y)dy$ for any function $q(x, \cdot)$. In this case the α condition (6) is meaningless and must be replaced by

$$\int_{x \in A} \int_{y \in B} \pi(dx) Q(x, dy) \alpha(x, y) = \int_{x \in A} \int_{y \in B} \pi(dy) Q(y, dx) \alpha(y, x) \ , \qquad \forall A, B \in \mathcal{S} \ . \tag{15}$$

Note that the dy notation is unavoidable because α depends on (x, y) and thus modifies Q by pointwise multiplication; one cannot use the set notation of (14). Here we take α as a function of (x, y), which thus may safely multiply measures pointwise in both variables. (15) can be summarized by an equality of measures on $\mathcal{S} \otimes \mathcal{S}$,

$$\pi(dx)Q(x, dy)\alpha(x, y) = \pi(dy)Q(y, dx)\alpha(y, x) \ , \tag{16}$$

in other words that the left hand side measure of (16) is symmetric with respect to $x \leftrightarrow y$. [1]

The rejection probability is, for each $x \in S$,

$$r(x) = 1 - \int Q(x, dz)\alpha(x, z) \ , \tag{17}$$

and, in terms of this, the M-H transition measure (with respect to y) is

$$K(x, dy) = Q(x, dy)\alpha(x, y) + r(x)\delta_x(dy) \ , \tag{18}$$

or (Kennedy (8.7)),

$$K(x, B) = \int_B Q(x, dy)\alpha(x, y) + r(x)1_{x\in B} \ , \qquad \forall B \in \mathcal{S} \ . \tag{19}$$

Lemma 6 (M-H for proposals that are measures)**.** *For any acceptance probability function $\alpha(x, y)$ obeying* (15)*, M-H with a proposal measure $Q(x, \cdot)$ has detailed balance with respect to the measure π.*

Proof. Let $A, B \in \mathcal{S}$. Inserting (19) into the LHS of (14) gives

$$\int_A \pi(dx)K(x, B) = \int_{x\in A} \pi(dx) \int_{y\in B} Q(x, dy)\alpha(x, y) + \int_{A\cap B} \pi(dx)r(x)$$

Instead inserting into the RHS gives

$$\int_B \pi(dx)K(x, A) = \int_{x\in B} \pi(dx) \int_{y\in A} Q(x, dy)\alpha(x, y) + \int_{A\cap B} \pi(dx)r(x)$$

The second terms are equal. By (15) (swapping the roles $x \leftrightarrow y$) the first terms are equal, so (14) holds for all A, B. □

Note that the formula for $r(x)$ was not needed here. Combining with Prop. 5 shows that for general proposal measures, M-H with α condition (15) has the correct π-invariance.

Distribution notation also gives a proof of Lemma 6 that looks simpler, when π is AC. Write $Q(x, dy) = q(x, y)dy$, keeping in mind that $q(x, \cdot)$ is now a distribution. First substituting (12), then using the α condition (6) now understanding it as a statement about distributions on $S \times S$, gives

$$\pi(x)q(x, y) = \pi(x)q(x, y)\alpha(x, y) + \pi(x)r(x)\delta_x(y) = \pi(y)q(y, x)\alpha(y, x) + \pi(y)r(y)\delta_y(x) = \pi(y)q(y, x)$$

where we also used the symmetry $f(x)\delta_x(y) = f(y)\delta_x(y) = f(y)\delta_y(x)$ for any continuous function f. Thus we proved the needed symmetry (4) interpreting both sides *as distributions*, resolving our question about the case $y = x$. Since $r(x)$ is the strength of the point mass along the diagonal, and any delta distribution along the diagonal is symmetric, the form of $r(x)$ was not needed (as in the above proof too). This may explain Tierney '98's cryptic comment that "the diagonal component does not matter" on p.2.

[1] Andrieu states this way more abstractly as: the ratio $\alpha(y, x)/\alpha(x, y)$ is the Radon–Nikodym derivative of $\pi(dx)Q(x, dy)$ and its $x \leftrightarrow y$ transpose.

2 M-H with a deterministic proposal from a map

One component of Hamiltonian MC is an M-H step using a proposal measure defined by a map. Let $F : S \to S$ be a deterministic smooth map (this will be given by, eg, leapfrog steps). Let the proposal measure $Q_F(x, \cdot)$ be defined by

$$Q_F(x, B) = 1_{F(x)\in B} \qquad \forall B \in \mathcal{S} \ ,$$

which simply places a unit point mass at $F(x)$. This can also be written

$$Q_F(x, dy) = \delta_{F(x)}(dy) = \delta(y - F(x))dy \ .$$

Remark 7. *Acting this proposal directly as a Markov kernel (ie, always accepting) onto a measure π gives πQ_F, the so-called "pushforward", defined by*

$$(\pi Q_F)(B) = \int \pi(dx) Q_F(x, B) = \pi(F^{-1}(B)) \qquad \forall B \in \mathcal{S} \ ,$$

and often denoted by $F_*\pi$, or by π^F (Andrieu).

Definition 8 (Involution)**.** *A map $F : S \to S$ is an involution if $F^{-1} = F$, that is, $F^2 = I$ where I is the identity.*

Andrieu, building on Tierney '98 p.4, emphasizes involutions, although this is confused for this reader with the $x \leftrightarrow y$ flip (see Andrieu eq. (4)).

Definition 9 (Liouville)**.** *A map $F : S \to S$ is volume-preserving if it preserves Lebesgue measure, that is,*

$$\int_B dx = \int_{F(B)} dx \qquad \forall B \subset S \ ,$$

where $F(B) := \{F(x) : x \in B\}$ is the image of the set B.

Volume preservation is equivalent to F having everywhere unit Jacobean, $|\det DF(x)| = 1$ for all $x \in S$.

Lemma 10 (M-H using a deterministic volume-preserving involution)**.** *Let π be an AC target density. Let F be a volume-preserving involution. Then M-H with the proposal measure Q_F, and acceptance probability α obeying*

$$\pi(x)\alpha(x, y) = \pi(y)\alpha(y, x) \qquad \forall x, y \in S \tag{20}$$

has DB with respect to π, and π as an invariant measure.

Proof. The conditions on F render Q_F symmetric (with respect to Lebesgue) as a measure on $\mathcal{S}\otimes\mathcal{S}$:

$$dxQ_F(x, dy) = \delta(F(x)-y)dxdy = |\det DF(F^{-1}(y))|^{-1}\delta(x-F^{-1}(y))dxdy = \delta(x-F(y))dxdy = dyQ_F(y, dx) \ .$$

Here, the 2nd step used the transformation rule (11), there being only one term in the sum because F is bijective, and the next step uses the involution and unit Jacobean. Combining this and (20) shows that (15) holds, so that DB holds by Lemma 6. Then invariance follows by Prop. 5. □

There is probably an integral change of variable version of this proof too, which might be even simpler. In any case, this formalizes verbal arguments such as those of Neal '11 using infinitesimal volumes.

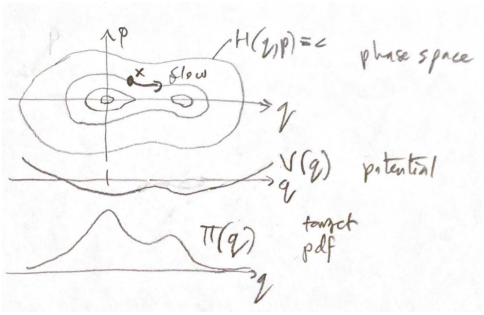

Figure 1: Hamiltonian dynamics in phase space $x = (q, p)$ sketched in $d = 1$ dimension, also showing the target pdf $\pi(q)$ and its resulting potential $V(q)$.

3 Hamiltonian Monte Carlo

Here we use the above tools to prove that HMC has the correct invariant pdf.

The goal is to sample from a pdf $\pi(q)$ over a d-dimensional Euclidean state space $q \in \mathbb{R}^d$. (π is assumed AC.) The MCMC state space is augmented to $x = (q, p)$ where $q \in \mathbb{R}^d$ is now interpreted as position and $p \in \mathbb{R}^d$ as momentum. Thus $S = \mathbb{R}^{2d}$, which is called phase space. In the simplest (scalar mass) case the Hamiltonian $H : S \to \mathbb{R}$ is

$$H(x) = H(q, p) = U(q) + \frac{1}{2}\|p\|^2$$

where the potential is constructed as $U(q) = -\log \pi(q)$, that is, $\pi(q) = e^{-U(q)}$, and $\|.\|$ is the 2-norm in $\mathbb{R}^d$. This assumes π is everywhere positive. Since H is the sum of terms in q and terms in p, Gibbs (aka Boltzmann) measure γ, defined as follows, is separable:

$$\gamma(x) := Z^{-1} e^{-H(x)} = Z^{-1} e^{-U(q)} e^{-\|p\|^2/2} = Z^{-1} \pi(q) e^{-\|p\|^2/2} \ , \tag{21}$$

where $Z > 0$ is some normalizing constant. Thus the q-marginal of γ is the desired π. The goal is to show that the steps in HMC are γ-invariant.

Hamiltonian flow with respect to a continuous time variable t is then (where dot means ∂_t)

$$\left\{ \begin{array}{rcccl} \dot{q} &=& \nabla_p H(q(t), p(t)) &=& p(t) \\ \dot{p} &=& -\nabla_q H(q(t), p(t)) &=& -\nabla U(q(t)) \end{array} \right. \tag{22}$$

See Fig. 1. Since HMC does not use this flow, we skip any proofs about it for now (it is H-preserving, volume-preserving, symplectic, time-reversible, etc).

Definition 11 (Shear)**.** *Any map on $\mathbb{R}^{2d}$ of the form $(q, p) \mapsto (q+G(p), p)$, or $(q, p) \mapsto (q, p+G(q))$, where $G : \mathbb{R}^d \to \mathbb{R}^d$ is some differentiable map, is called a* shear.

Proposition 12. *Any shear is volume-preserving.*

Proof. Let F be a shear. Computing its Jacobean with $d \times d$ blocks, $DF = \left[\begin{smallmatrix} I_d & DG \\ 0 & I_d \end{smallmatrix}\right]$, or $DF = \left[\begin{smallmatrix} I_d & 0 \\ DG & I_d \end{smallmatrix}\right]$. In either case $\det DF \equiv 1$. □

Fixing a timestep $\varepsilon > 0$, the leapfrog (Verlet) map L, acting on an "initial" state $x_k := (q_k, p_k)$, to give a "final" state $(q_{k+1}, p_{k+1}) = L(q_k, p_k)$, comprises three sequential steps:

1. $p' \ \leftarrow \ p_k - \frac{\varepsilon}{2} \nabla U(q_k)$
2. $q_{k+1} \ \leftarrow \ q_k + \varepsilon p'$

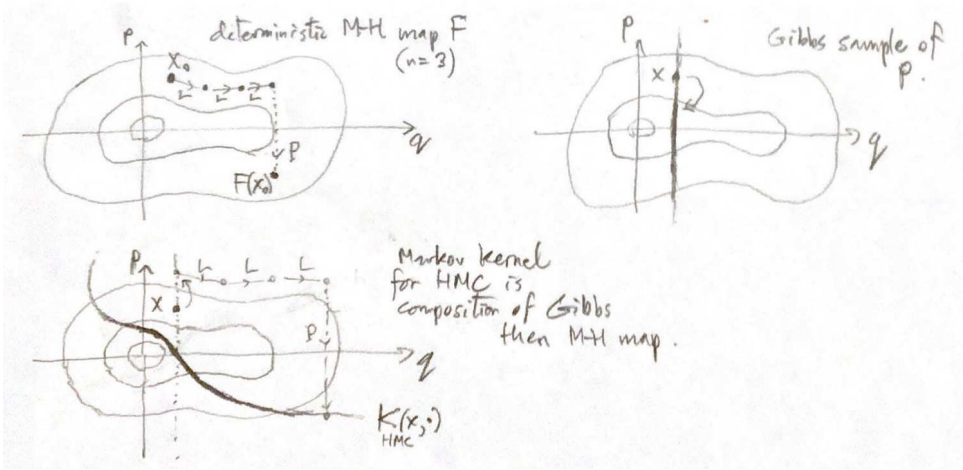

Figure 2: Steps comprising a single HMC trajectory for the Hamiltonian of Fig. 1. Top left: deterministic map F in phase space (sketch in $d = 1$ dimension) for $n = 3$ leapfrogs plus a p-flip. Top right: Gibbs resample (complete randomization to a Gaussian) of p holding q fixed. Bottom: kernel resulting from composition of Gibbs then M-H; note it is a singular measure on a distorted d-dimensional surface in $\mathbb{R}^{2d}$. One example route is sketched.

3. $p_{k+1} \leftarrow p' - \frac{\varepsilon}{2}\nabla U(q_{k+1})$

Let P be the momentum-flip operator

$$P(q,p) := (q,-p) \ .$$

Lemma 13. *Let $n \in \{0,1,\dots\}$. The map L^nP is a volume-preserving involution. (Here, as above, we compose operators to the right, so this means L^n followed by P, although it happens not to matter here.)*

Proof. L is the composition of three steps each of which is a shear, and thus volume-preserving by Prop. 12. P is obviously volume-preserving. Thus L^nP is volume-preserving. L is time-reversible in the sense that, if $L(q_k,p_k) = (q_{k+1},p_{k+1})$, one may verify $L(q_{k+1},-p_{k+1}) = (q_k,-p_k)$ by checking the three steps in reverse order (p' is negated relative to its forward value). The same is true for L^n, by so reversing each leapfrog. Stating this algebraically, $PL^nP = L^{-n}$. Using $P = P^{-1}$ and rearranging, $(L^nP)^2 = I$, so L^nP is an involution. □

Remark 14. *L^n happens to be an $\mathcal{O}(\varepsilon^2)$-accurate approximation to integrating the flow (22) to time $t = n\varepsilon$, although this plays no role in the correctness of HMC. There exist other step schemes that are also volume-preserving involutions, for instance "modified FE" in Neal '11 (but not plain FE = forward Euler), which is $\mathcal{O}(\varepsilon)$-accurate.*

Vanilla HMC has algorithm parameters $\varepsilon > 0$ the timestep, and n a number of leapfrog steps per trajectory. A single "trajectory" composes two steps in sequence:

1. Gibbs-sample (randomize or partially so) p in a way that leaves the Gaussian marginal $e^{-\|p\|^2/2}$ invariant (without affecting q).

2. Perform one M-H MCMC step to the state $x = (q,p)$ using the deterministic proposal Q_F generated by the map $F = L^nP$, with acceptance probability α obeying (20), such as the standard
$$\alpha(x,y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right) \ .$$

This either changes $(q,p) = x$ to y, or rejects leaving it unchanged.

This *pair* of steps may then be repeated to generate a Markov chain. Note that this is not a Markov chain generated by any M-H rule: the Markov operator for the chain is a Gibbs move plus a M-H update. See Fig. 2.

Theorem 15 (Invariant pdf of HMC)**.** *Let π be a continuous positive pdf. The Markov operator defined by one trajectory of vanilla HMC as described above has γ, defined by* (21)*, as an invariant pdf.*

Proof. It is sufficient to show that each step in the pair is γ-invariant. This holds for step 1 (a Gibbs update) since it preserves the conditional over p, which is identical at each fixed q, while leaving q unaffected. It holds for step 2 (one deterministic M-H step) since by Lemma 13, $F = L^n P$ is a volume-preserving involution, so one can apply Lemma 10. □

We have not discussed uniqueness, or mixing rates, at all. However, assuming that the Markov chain *converges* to the unique invariant pdf γ, the fact that marginalizing (21) over p leaves $\pi(q)$ immediately give the following.

Corollary 16. *Assuming a converged chain, vanilla HMC generates a sequence of q values drawn from π.*

Note that if the Gibbs move (step 1) does not fully randomize p, then one cannot write vanilla HMC as *any* Markov chain in q alone: p has a memory that destroys the Markov property in q.

Remark 17. *The point of the Gibbs moves alternating with the M-H moves is to increase mixing with respect to the value of H. A chain of M-H moves alone, while also γ-invariant would approximately (to accuracy $\mathcal{O}(\varepsilon^2)$) conserve H, because* (22) *conserves H exactly.*

Remark 18. *In what sense does the overall Markov step (Gibbs followed by M-H) for vanilla HMC obey DB, as claimed by Sohl-Dickstein et al 2016? It appears not to, since while the Gibbs sampler and the M-H step each separately obey DB, they do not commute as operators. As Geyer reminds, DB is equiv to self-adjointness under the L^2_π metric, and given $A^* = A$ and $B^* = B$, $(AB)^* = BA \neq AB$ unless they commute. (Mira's thesis p. 32 validates this.)*

Questions:

1. What goes wrong in the proof of Q a symmetric product measure if $F = L^n$ which is not an involution?

2. Understand Campos & Sanz-Serna.

4 Delayed rejection

The best source on this is Mira's thesis from 1998, Sec. 5.2. We use the distribution notation from the end of Sec. 1.3, keeping in mind that it summarizes weak statements involving equality of integrals for all measurable subsets. Starting at $x \in S$ we allow a second proposal $q_2(x,s,y)$ to y if the first one $q_1(x,s)$ is rejected. The proposals q_1 and q_2 are designed to mix efficiently, then the acceptances α_1 and α_2 are derived from them to have DB. Fig. 3 shows the idea.

The transition kernel analogous to (12) for delayed rejection is, marginalizing over all possible rejected first tries s,

$$k(x,y) = q_1(x,y)\alpha_1(x,y) + \int q_1(x,s)[1-\alpha_1(x,s)][q_2(x,s,y)\alpha_2(x,s,y) + r_2(x,s)\delta_x(y)]ds \qquad (23)$$

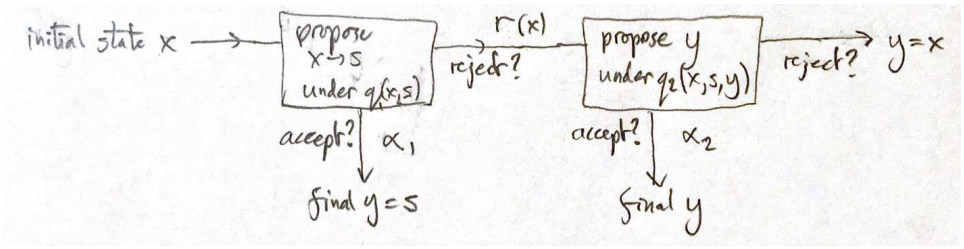

Figure 3: Delayed rejection, simplest case, for general proposals, taking x to y.

where r_2 is the probability of rejection of the 2nd proposal; it's form will be irrelevant because as a measure it lies on the diagonal $x = y$. Note we use slightly different symbols than Mira. Say α_1 satisfies the usual M-H acceptance condition (6). The goal is to choose $\alpha_2(x, s, y)$ to achieve DB (4), ie $\pi(x)k(x, y)$ is $x \leftrightarrow y$ symmetric, distributionally. Mira's mechanism to guarantee this is to make the integrand in (23) $x \leftrightarrow y$ symmetric once multiplied by $\pi(x)$. This symmetry obviously is then maintained by integrating ds. The first (plain M-H acceptance) term $\pi(x)q_1(x, y)\alpha_1(x, y)$ is already symmetric so can be ignored, and we can also ignore the (always symmetric) rejected point mass. This leaves symmetry of the integrands

$$\pi(x)q_1(x, s)[1 - \alpha_1(x, s)]q_2(x, s, y)\alpha_2(x, s, y) = \pi(y)q_1(y, s)[1 - \alpha_1(y, s)]q_2(y, s, x)\alpha_2(y, s, x) \ , \forall s \in S \tag{24}$$

One may choose the Metropolis form obeying this with the least rejection,

$$\alpha_2(x, s, y) = \min\left[\frac{\pi(y)q_1(y, s)[1 - \alpha_1(y, s)]q_2(y, s, x)}{\pi(x)q_1(x, s)[1 - \alpha_1(x, s)]q_2(x, s, y)}, 1\right] \ ,$$

being the simplest DR acceptance probability. I do not derive the higher-order formulae here (see Mira's thesis).

4.1 DR for proposals given by deterministic maps

Our application of DR to adaptive HMC is in its M-H step. The first proposal is $q_1(x, y) = \delta(y - F_1(x))$, and the 2nd $q_2(x, s, y) = \delta(y - F_2(x))$ which in our case is independent of the rejected proposed s. For us $F_1 = L_\varepsilon^n P$ for some stepsize ε, then $F_2 = L_{\varepsilon/a}^{an} P$ for some stepsize $a \in \{2, 3, \dots\}$ times smaller; however, all we need to know is that they are both volume-preserving involutions. Then by Lemma 10 we already know $\alpha_1(x, s) = \min[\pi(s)/\pi(x), 1]$. Naively substituting the map proposal kernels into (24) gives, for almost all $s \in S$,

$$\pi(x)\delta(s-F_1(x))[1-\alpha_1(x, s)]\delta(y-F_2(x))\alpha_2(x, s, y) = \pi(y)\delta(s-F_1(y))[1-\alpha_1(y, s)]\delta(x-F_2(y))\alpha_2(y, s, x).$$

Yet this cannot hold as s-dependent measures on $S \times S$, since for the LHS to be nonzero $F_1(x) = s$, but for the RHS $F_1(y) = s$; thus (since F_1 is injective) it could only hold for $x = y$.

Thus Mira's mechanism fails, and one must instead impose symmetry *after* integrating over ds. Writing symmetry for the non-rejected part of the integral from (23) gives the looser condition

$$\pi(x)\int q_1(x, s)[1-\alpha_1(x, s)]q_2(x, s, y)\alpha_2(x, s, y)ds = \pi(y)\int q_1(y, s)[1-\alpha_1(y, s)]q_2(y, s, x)\alpha_2(y, s, x)ds$$

Inserting the map proposal kernels gives

$$\pi(x)\int \delta(s-F_1(x))[1-\alpha_1(x, s)]\delta(y-F_2(x))\alpha_2(x, s, y)ds = \pi(y)\int \delta(s-F_1(y))[1-\alpha_1(y, s)]\delta(x-F_2(y))\alpha_2(y, s, x)ds$$

which simplifies (using that F_2 is an involution to "cancel" the associated δ's),

$$\pi(x)[1-\alpha_1(x,F_1(x))]\alpha_2(x,F_1(x),y) = \pi(y)[1-\alpha_1(y,F_1(y))]\alpha_2(y,F_1(y),x) \ , \tag{25}$$

a condition for α_2 which is only defined on the 2nd proposal surface $y = F_2(x)$. Thus for $y = F_2(x)$ we get the rule, simplifying using α_1,

$$\begin{aligned} \alpha_2(x,F_1(x),F_2(x)) &= \min\left[\frac{\pi(F_2(x))[1-\alpha_1(F_2(x),F_1(F_2(x)))]}{\pi(x)[1-\alpha_1(x,F_1(x))]},1\right] \\ &= \min\left\{\frac{\max[\pi(F_2(x))-\pi(F_1(F_2(x))),0]}{\pi(x)-\pi(F_1(x))},1\right\} \ . \end{aligned} \tag{26}$$

Here the max in denominator was removed because it is always positive if the 1st proposal is not accepted. The middle argument of α_2, being deterministically given by the first argument, is slightly redundant. The rule involves four densities to check, at locations: x, $F_1(x)$, $F_2(x)$, and $F_1(F_2(x))$ a new ghost (pre)image of the 2nd proposal.