Chapter 5

Stochastic Thermostats

The Metropolis accept/reject step allowed for the use of very naive proposal densities. Errors in the proposal step are corrected at the acceptance step at the cost of a potentially very high rejection probability. In this section we describe a family of methods that use continuous time Markov processes that can be used to sample π without detailed knowledge of its structure, and without an accept/reject step. In practice, the continuous time Markov processes must be discretized which introduces error. This error can be eliminated by the addition of a Metropolis accept/reject step or it can be reduced by decreasing the discretization parameter. These methods (typically without any accept/reject step) are the basis of most MCMC simulations in very high dimensions.

5.1 Overdamped Langevin schemes

In this section we will derive the first of our new, accept/reject free schemes starting from a very simple "isotropic" Metropolis scheme. For any $h > 0$, let X_h be a Markov chain in $\mathbb{R}$ generated according to the Metropolis-Hastings rule described in the previous chapter with proposal density $q(y \,|\, x) = \mathcal{N}(x, 2h)$.

We will begin by computing the generator corresponding to X_h. It will be convenient to scale the generator by a factor of h^{-1} so that

$$\mathcal{L}_h f(x) = \frac{\mathbf{E}_x\left[f(X_h^{(1)})\right] - f(x)}{h}.$$

This rescaling corresponds to associating each discrete step of the chain with a size h increment of a continuous time variable.

Taylor Expanding $f(X_h^{(1)})$ about x we find that

$$\mathcal{L}_h f(x) = \frac{f'(x)\mathbf{E}_x\left[\triangle_0^1 X_h\right] + \frac{1}{2}f''(x)\mathbf{E}_x\left[\left(\triangle_0^1 X_h\right)^2\right] + \mathcal{O}\left(\mathbf{E}_x\left[|\triangle_0^1 X_h|^3\right]\right)}{h}$$

Now notice that, by a change of variables,

$$\begin{aligned} E_x\left[\triangle_0^1 X_h\right] &= \int (y-x)p(y\,|\,x)dy = \int (y-x)q(y\,|\,x)p_{acc}(x,y)dy \\ &= \sqrt{2h}\int z\,p_{acc}(x, x+\sqrt{2h}\,z)\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz. \end{aligned}$$

Examining p_{acc} in more detail we find that

$$p_{acc}(x, x+\sqrt{2h}\,z) = 1 + \min\left\{0, \frac{\pi'(x)}{\pi(x)}\sqrt{2h}\,z + \mathcal{O}\left(h\right)\right\}.$$

When h is small, the sign of the expression in the minimum is determined by the first order (in $\sqrt{h}$) term. Since

$$\int z\,\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz = 0$$

we obtain

$$E_x\left[\triangle_0^1 X_h\right] = 2h\frac{\pi'(x)}{\pi(x)}\int_{\pi'(x)z<0} z^2\,\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz + \mathcal{O}\left(h^{3/2}\right) = h(\log\pi(x))' + \mathcal{O}\left(h^{3/2}\right).$$

Similarly

$$\begin{aligned} E_x\left[\left(\triangle_0^1 X_h\right)^2\right] &= 2h\int z^2 p_{acc}(x, x+\sqrt{2h}z)\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz \\ &= 2h + \mathcal{O}\left(h^{3/2}\right) \end{aligned}$$

and

$$E_x \left[|\triangle_0^1 X_h|^3 \right] = \mathcal{O}\left(h^{3/2}\right)$$

so that

$$\mathcal{L}_h f(x) = \mathcal{L}_O f(x) + \mathcal{O}\left(\sqrt{h}\right)$$

where we have introduced the second order differential operator

$$\mathcal{L}_O f = f'(x)\left(\log \pi(x)\right)' + f''(x) = \frac{1}{\pi(x)}\left(\pi(x) f'(x)\right)' \tag{5.1}$$

Were we to repeat this derivation in higher dimensions for the Metropolis-Hastings scheme with proposal density $q(y \,|\, x) = \mathcal{N}\left(x, 2hS(x)\right)$ for some symmetric $d \times d$ positive definite matrix $S(x)$ we would again find that $\mathcal{L}_h = \mathcal{L}_O + \mathcal{O}\left(\sqrt{h}\right)$, where now

$$\begin{aligned}\mathcal{L}_O f &= \nabla f(x) \frac{1}{\pi(x)} \mathrm{div}\left(\pi(x) S(x)\right) + \mathrm{trace}\left(D^2 f(x)\, S(x)\right) \\ &= \frac{1}{\pi(x)} \mathrm{div}\left(\pi(x) \nabla f(x) S(x)\right). \end{aligned} \tag{5.2}$$

In these formulae, the matrix $D^2 f$ is the $d \times d$ matrix of second derivatives of f and for any n, the divergence of an $n \times d$ matrix valued function $M(x)$ is the $n \times 1$ vector with entries

$$\left(\mathrm{div} M(x)\right)_i = \sum_{j=1}^{d} \partial_j M_{ij}(x).$$

Exercise 45. *Verify the expression* (5.2) *for the limiting generator of the Metropolis-Hastings scheme with* $q(y \,|\, x) = \mathcal{N}\left(x, 2hS(x)\right)$. *Note that this transition density is not symmetric which leads to the additional first order term* $\nabla f(x)\; \mathit{div}\, S(x)$.

Let's examine this operator $\mathcal{L}_O$ in (5.2) in slightly more detail. First, recall that, when μ is a density, the action of $\mathcal{L}_O$ on μ is $\mu \mathcal{L}_O = \mathcal{L}_O^* \mu$ where $\mathcal{L}_O^*$ is the adjoint of $\mathcal{L}_O$ in the inner product $\langle f, g \rangle = \int f(x) g(x) dx$. For this particular operator the adjoint is computed by an integration by parts. We find that

$$\begin{aligned}
\int f(x)\mu\mathcal{L}_O(x)dx &= \int \mathcal{L}_O f(x)\mu(x)dx \\
&= \int \operatorname{div}\left(\pi(x)\nabla f(x)S(x)\right)\frac{\mu(x)}{\pi(x)}dx \\
&= -\int \pi(x)\nabla f(x)S(x)\nabla^{\mathrm{T}}\left(\frac{\mu(x)}{\pi(x)}\right)dx \\
&= \int f(x)\operatorname{div}\left(\pi(x)\nabla\left(\frac{\mu(x)}{\pi(x)}\right)S(x)\right)dx
\end{aligned}$$

for all f, so that

$$\mu\mathcal{L}_O(x) = \operatorname{div}\left(\pi(x)\nabla\left(\frac{\mu(x)}{\pi(x)}\right)S(x)\right). \tag{5.3}$$

Plugging in π in place of μ we see that $\pi\mathcal{L}_O = 0$. In fact, $\mathcal{L}_O$ is reversible with respect to π. In particular,

$$\int f(x)\mathcal{L}_O g(x)\pi(dx) = \int g(x)\mathcal{L}_O f(x)\pi(dx)$$

for all f and g.

Exercise 46. *Verify the formula in the last display.*

We already knew that the Metropolis scheme preserves π. But having used that scheme to derive the operator $\mathcal{L}_O$, and having observed that $\pi\mathcal{L}_O = 0$, we can consider alternative schemes that may not exactly preserve π, but whose generator (after rescaling by h^{-1}), $\mathcal{L}_h$, also approximates $\mathcal{L}_O$. One might expect that the invariant measure (should it exist) for one of these alternative schemes, π_h, while not exactly equal to π, is close to π for small h. Indeed, if we assume that π_h is a probability measure satisfying $\pi_h\mathcal{L}_h = 0$, and if for some test function f, u is the solution of the PDE

$$\mathcal{L}_O u = f - \pi[f], \tag{5.4}$$

then

$$\pi_h[f] - \pi[f] = \pi_h\left(\mathcal{L}_O - \mathcal{L}_h\right)u \tag{5.5}$$

The last term will be small (in h), for example, if u is smooth and bounded with bounded derivatives. For the Metropolis scheme the final bound is $\mathcal{O}(\sqrt{h})$, which we know is too pessimistic (in that case, $\pi_h = \pi$), but for other chains that do not exactly preserve π, this expression can tell us that Monte Carlo estimates produced using the chain X_h will have a bias that is small. Later in this section we use a similar argument to bound the error in the Markov chain Monte Carlo estimator of $\pi[f]$ using a chain introduced below for which $\mathcal{L}_h f = \mathcal{L}_O f + \mathcal{O}(h)$.

To find other candidates for Markov chains that might have invariant probability measures approximating π, we return to the key features of the Metropolis scheme responsible for the convergence of $\mathcal{L}_h$ to $\mathcal{L}_O$. Can we identify those same properties in other chains? As we will see in a moment, the answer to this question is yes. The most compelling argument for looking for alternatives to the Metropolis–Hastings framework is the slow convergence of those schemes in high dimensions, an unfortunate characteristic that we have already explored. The so called Langevin schemes that we will derive in this Chapter can avoid the accept reject step (at the cost of a small in h bias) and are generally much more effective in very high dimensional settings than Metropolis–Hastings schemes.

Looking back at the convergence argument, the key properties that lead to the derivation of the limiting generator (5.1) are that

$$E_x\left[\triangle_0^1 X_h\right] = h\left(\log \pi(x)\right)' + o(h) \quad \text{and} \quad E_x\left[\left(\triangle_0^1 X_h\right)^2\right] = 2h + o(h)$$

or, in higher dimensions,

$$E_x\left[\triangle_0^1 X_h\right] = h\frac{1}{\pi(x)}\operatorname{div}\left(\pi(x)S(x)\right) + o(h)$$

and

$$E_x\left[\left(\triangle_0^1 X_h\right)\left(\triangle_0^1 X_h\right)^{\mathrm{T}}\right] = 2hS(x) + o(h).$$

More generally, our manipulations suggest that a discrete time Markov chain X_h satisfying

$$E_x\left[\triangle_0^1 X_h\right] = b(x)h + o(h) \quad \text{and} \quad E_x\left[\left(\triangle_0^1 X_h\right)^2\right] = \sigma(x)\sigma^{\mathrm{T}}(x)h + o(h) \tag{5.6}$$

should have limiting generator

$$\mathcal{L}f(x) = \nabla f(x)\, b(x) + \frac{1}{2}\operatorname{trace}\left(\sigma(x)\sigma^{\mathrm{T}}(x)D^2 f(x)\right). \tag{5.7}$$

With some additional restrictions, this is in fact the case.

The simple recursion

$$X_h^{(k+1)} = X_h^{(k)} + h\, b(X_h^{(k)}) + \sqrt{h}\, \sigma(X_h^{(k)})\, \xi^{(k+1)} \tag{5.8}$$

where each $\xi^{(k)}$ is, for example, a vector of independent random variables with $\mathbf{P}\left[(\xi^{(k)})_i = 1\right] = \mathbf{P}\left[(\xi^{(k)})_i = -1\right] = 1/2$ or a vector of independent standard Gaussian random variables, has limiting generator in (5.7).

Exercise 47. *Show that for the process in* (5.8) *and any smooth function f with bounded derivatives,*

$$\mathcal{L}_h f - \mathcal{L} f = \mathcal{O}(h)$$

for $\mathcal{L}$ in (5.7).

When we plug the choices $b = \operatorname{div}(\pi S)/\pi$ and $\sigma\sigma^{\mathrm{T}} = 2S$ obtained above into (5.8) we obtain a discretization of the overdamped Langevin sampler with updates

$$\begin{aligned} X_h^{(k+1)} = X_h^{(k)} &+ h\, S(X_h^{(k)}) \nabla^{\mathrm{T}} \log \pi(X_h^{(k)}) \\ &+ h\, \mathrm{div} S(X_h^{(k)}) + \sqrt{2hS(X_h^{(k)})}\, \xi^{(k+1)} \end{aligned} \tag{5.9}$$

Note that to implement (5.9), one needs to compute the matrix square root of S e.g. by Choleski factorization. The chain in (5.9) will not preserve π exactly and for h too large will not have an invariant probability measure at all (it will be transient). However, under restrictions on π and for h small the chain will have an invariant probability measure π_h and, combining (5.5) and the result Exercise 47, we expect that $\pi_h[f] = \pi[f] + \mathcal{O}(h)$.

Exercise 48. *Compute the invariant measure π_h of the process* (5.9) *in the case when $S = I$ and the target density is $\pi = \mathcal{N}(\mu, M)$. Confirm that $\pi_h[f] = \pi[f] + \mathcal{O}(h)$.*

The operator $\mathcal{L}_O$ in (5.2) is an example of a Kolmogorov operator and is the generator of a continuous time Markov process that is the limit of our Metropolis scheme, thus justifying our use of the term limiting generator. That limiting Markov process is called a diffusion process. Diffusion processes always have generators of the form (5.7) where the vector valued function b and the matrix valued function σ can be more general than the functions that we have derived above by considering the limit of the Metropolis scheme.

At the moment we have no particular interest in diffusion processes. On the other hand, in this Chapter we are precisely interested in MCMC schemes with diffusion limits. And because, as we have already seen, there are many discrete processes with the same diffusion limit (i.e. same limiting functions b and σ), it will sometimes be useful to characterize our sampling schemes by their limiting drift and diffusion coefficients or, equivalently, by their limiting generators. Diffusion processes are used to model many physical processes and will receive more attention in Part II of these notes.

The approach represented by (5.9) and its many generalizations is the most common method by which very high dimensional averages and integrals are computed. To get a basic feeling for the qualitative properties of (5.9) (and for those of the Metropolis scheme that we have just learned is intimately related), assume that S is constant and replace (5.9) by the deterministic iteration

$$x_h^{(k+1)} = x_h^{(k)} + hS\nabla^{\mathrm{T}} \log \pi(x_h^{(k)}). \tag{5.10}$$

The generator of this process (after rescaling by h^{-1}) converges to $\mathcal{L}$ defined by $\mathcal{L}f = \nabla f S \nabla^{\mathrm{T}} \log \pi$. The operator $\mathcal{L}$ is just the Liouvillian corresponding to the ODE $\frac{d}{dt}y^{(t)} = S\nabla^{\mathrm{T}} \log \pi(y^{(t)})$, which implies that $x_h^{(k)}$ converges to $y^{(kh)}$ as h decreases (with kh held constant). That ODE is a gradient ascent for the function $\log \pi$. In other words

$$\frac{d}{dt} \log \pi \left(y^{(t)}\right) = \nabla(\log \pi(y^{(t)}))S\nabla^{\mathrm{T}}(\log \pi(y^{(t)})).$$

The expression on the right hand side is always positive because S is a symmetric positive definite matrix. In fact, if $\lambda_1 > 0$ is the smallest eigenvalue of S then

$$\frac{d}{dt} \log \pi \left(y^{(t)}\right) \geq \lambda_1 \|\nabla(\log \pi(y^{(t)}))\|_2^2$$

so that as y evolves it can only increase the value of π. Similarly, the Markov chain X generated by (5.8) tends to move toward higher π-probability regions. The added noise in (5.9) prevents $X_h^{(k)}$ from converging to a local maximum of π. Because of its tendency to reduce the "energy" $-\log \pi$, any term of the form $S\nabla^{\mathrm{T}} \log \pi$ for a symmetric positive semi-definite matrix S

is referred to a dissipative or damping term. In fact, the subscript O in $\mathcal{L}_O$ references the fact that the operator $\mathcal{L}_O$ can be derived as a highly damped (or "overdamped") limit of a more general family of operators that we will introduce later in this chapter. We will refer to any scheme with a limiting generator of the forms considered in this section as an overdamped scheme.

In Chapter **??** we will consider how the choice of S can be used to speed convergence of the MCMC estimator corresponding to (5.9) with $b = \operatorname{div}(\pi S)/\pi$ and $\sigma\sigma^{\mathrm{T}} = 2S$. Before moving on to introducing further improvements to this estimator we return to the question of how large an error we should expect when we estimate $\pi[f]$ by

$$\overline{f}_N = \frac{1}{N}\sum_{k=1}^{N} f(X_h^{(k)})$$

recalling that $\mathcal{L}_h$ only approximately preserves π.

In fact, if we again make the (strong) assumptions that the PDE (5.4) has a smooth solution u, which, along with S and $\log\pi$, is bounded with bounded derivatives we can bound the error of $\overline{f}_N$ directly. To see this note that

$$\frac{u(X_h^{(k+1)}) - u(X_h^{(1)})}{h\,k} = \frac{1}{k}\sum_{\ell=1}^{k} \mathcal{L}_h u(X_h^{(\ell)}) + \frac{1}{h\,k} M_h^{(k)}$$

where $M_h^{(k)}$ is the martingale

$$M_h^{(k)} = \sum_{\ell=1}^{k} u(X_h^{(\ell+1)}) - \mathbf{E}_{X_h^{(\ell)}}\left[u(X_h^{(\ell+1)})\right].$$

Using the PDE solved by u, we can write

$$\frac{u(X_h^{(k+1)}) - u(X_h^{(1)})}{h\,k} = \frac{1}{k}\sum_{\ell=1}^{k} f(X_h^{(\ell)}) - \pi[f] + \frac{1}{k}R_h^{(k)} + \frac{1}{h\,k}M_h^{(k)} \qquad (5.11)$$

where

$$R_h^{(k)} = \sum_{\ell=1}^{k} (\mathcal{L}_h - \mathcal{L})\, u(X_h^{(\ell)}).$$

Note that, if u is bounded with bounded derivatives and if $X_h^{(k)}$ is generated by (5.9), then $R_h^{(k)}$ is of size $\mathcal{O}(kh)$ so that

$$\overline{f}_k - \pi[f] = \frac{1}{h\,k} M_h^{(k)} + \mathcal{O}((hk)^{-1}) + \mathcal{O}(h).$$

Now consider the expected square of the martingale $M_h^{(k)}$,

$$\mathbf{E}\left[(M_h^{(k)})^2\right] = \sum_{\ell=1}^{k} \mathbf{E}\left[\left(u(X_h^{(\ell+1)}) - \mathbf{E}_{X_h^{(\ell)}}\left[u(X_h^{(\ell+1)})\right]\right)^2\right]. \qquad (5.12)$$

The cross terms on the right hand side have vanished because, if $\mathcal{F}_\ell$ is the sigma algebra generated by $X_h^{(0)}, X_h^{(1)}, \dots, X_h^{(\ell)}$, then, for $r < \ell$,

$$\begin{aligned}
&\mathbf{E}\left[\left(u(X_h^{(\ell+1)}) - \mathbf{E}_{X_h^{(\ell)}}\left[u(X_h^{(\ell+1)})\right]\right)\left(u(X_h^{(r+1)}) - \mathbf{E}_{X_h^{(r)}}\left[u(X_h^{(r+1)})\right]\right)\right] \\
&\quad = \mathbf{E}\left[\mathbf{E}\left[u(X_h^{(\ell+1)}) - \mathbf{E}_{X^{(\ell)}}\left[u(X_h^{(\ell+1)})\right] \mid \mathcal{F}_\ell\right]\left(u(X_h^{(r+1)}) - \mathbf{E}_{X_h^{(r)}}\left[u(X_h^{(r+1)})\right]\right)\right] \\
&\quad = 0.
\end{aligned}$$

Returning to (5.12), taylor expansion of u reveals that (again assuming $X_h^{(k)}$ is generated by (5.9)) the expected square of $M_h^{(k)}$ is of size $\mathcal{O}(kh)$ so that

$$\mathbf{E}\left[(\overline{f}_k - \pi[f])^2\right] = \mathcal{O}\left(\frac{1}{h\,k}\right) + \mathcal{O}(h^2).$$

Note that this is exactly the qualitative behavior you would expect. As h is decreased, larger k is required to maintain a fixed accuracy because the chain is perturbed less at each step. On the other hand, when k increases with h fixed, an error of size $\mathcal{O}(h)$ persists.

5.2 Hamilton's ODEs

The basis of the family of MCMC schemes built from continuous time dynamics that we will describe is a simple, but very important, set of ordinary differential equations. For $x \in \mathbb{R}^d$, any sufficiently smooth real valued function $H(x)$ and skew-symmetric, smooth, $d \times d$ matrix valued function $J(x)$ (i.e. $J^{\mathrm{T}} = -J$), the ODE

$$\frac{d}{dt} y^{(t)} = -J(y^{(t)})\nabla^{\mathrm{T}} H(y^{(t)}) + \mathrm{div} J(y^{(t)}) \qquad (5.13)$$

is referred to as a Hamiltonian system of ODE. The function $H(x)$ is called the Hamiltonian. We will assume that

$$\liminf_{x\to\infty} \frac{H(x)}{\|x\|} > 0.$$

This condition guarantees that the function $\exp\left(-H(x)\right)$ is integrable for any $\beta > 0$. In the context of statistical mechanics, the density

$$\pi_H(x) \propto \exp\left(-H(x)\right) \tag{5.14}$$

is called a Boltzmann density. In terms of the Boltzmann density we can rewrite the Hamiltonian system of ODE (5.13) as

$$\frac{d}{dt}y^{(t)} = \frac{1}{\pi_H(y^{(t)})}\operatorname{div}\left(\pi_H(y^{(t)})J(y^{(t)})\right) \tag{5.15}$$

Example 23. *In the special case that* $x = (\hat{x}, \tilde{x})$ *and*

$$H(x) = \frac{1}{2}\tilde{x}^{\scriptscriptstyle T} M^{-1}\tilde{x} + U(\hat{x})$$

where $\tilde{x}$ *and* $\hat{x}$ *are the positions and velocities of a system of particles with mass matrix* M *and experiencing the potential field* U *and*

$$J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix},$$

the Hamiltonian system can be rewritten

$$M\frac{d^2}{dt^2}\hat{y}^{(t)} = -\nabla^{\scriptscriptstyle T} U(\hat{y}^{(t)}),$$

which is just Newton's equations of motion i.e. force equals mass times acceleration.

The system (5.13) has several important characteristics. To derive these characteristics, recall that the generator corresponding to the ODE (5.13) is

$$\mathcal{L}_H f = \nabla f \frac{1}{\pi_H}\operatorname{div}\left(\pi_H J\right)$$

The action of $\mathcal{L}_H$ on a density μ is found (by an integration by parts) to be

$$\mu\mathcal{L}_H = -\mathrm{div}\left(\frac{\mu}{\pi_H}\,\mathrm{div}^{\mathrm{T}}(\pi_H J)\right)$$

Expanding this expression we find that

$$\mu\mathcal{L}_H = -\nabla\left(\frac{\mu}{\pi_H}\right)\mathrm{div}\,(\pi_H J) - \frac{\mu}{\pi_H}\mathrm{div}\,(\mathrm{div}^{\mathrm{T}}\,(\pi_H J))$$

The second term of the last display vanishes because J is antisymmetric, leaving

$$\mu\mathcal{L}_H = -\nabla\left(\frac{\mu}{\pi_H}\right)\mathrm{div}\,(\pi_H J) \tag{5.16}$$

Exercise 49. *Check this.*

Plugging in $\mu = \pi_H$ we find that

$$\pi_H\mathcal{L}_H = 0.$$

In fact, with respect to π_H, the operator $\mathcal{L}_H$ satisfies the even stronger property

$$\int g(x)\mathcal{L}_H f(x)\pi_H(dx) = -\int f(x)\mathcal{L}_H g(x)\pi_H(dx) \tag{5.17}$$

for any test functions f and g. Condition (5.17) closely resembles the reversibility condition introduced in the Chapter (**??**) and will be referred to here as skew-reversibility with respect to π_H. Like reversibility, skew-reversibility with respect to π_H implies that $\pi_H\mathcal{L}_H = 0$ (though we had already verified this).

Exercise 50. *Verify equation* (5.17).

Expression (5.17) in turn implies that

$$\int f(x)g(y^{(t)}(x))\pi_H(dx) = \int g(x)f(y^{(-t)}(x))\pi_H(dx). \tag{5.18}$$

Exercise 51. *Establish the expression in the last display. Hint: fix $s \in [0,t]$ and let $w(s) = \int f(y^{(s-t)}(x))g(y^{(s)}(x))\pi_H(dx)$ and then show that the derivative of w is zero.*

We will use these facts to derive a number of sampling schemes based loosely on the ODE in (5.13).

Some final useful properties of the solution to (5.13) can be derived if we make additional assumptions on J and H. We will assume that the variable x can be decomposed into two vectors $x = (\hat{x}, \tilde{x})$ with $\hat{x} \in \mathbb{R}^{\hat{d}}$ and $\tilde{x} \in \mathbb{R}^{\tilde{d}}$ and that first, $H(\hat{x}, \tilde{x})$ is an even function of $\tilde{x}$, i.e that

$$H(\hat{x}, \tilde{x}) = H(\hat{x}, -\tilde{x}), \tag{5.19}$$

and second that J has the particular form

$$J(x) = \begin{bmatrix} 0 & -\hat{J}(\hat{x}) \\ \hat{J}^{\mathrm{T}}(\hat{x}) & 0 \end{bmatrix} \tag{5.20}$$

where $\hat{J}$ is a $\hat{d} \times \tilde{d}$ matrix valued function of only the $\hat{x}$ variables. Under these assumptions (5.13) becomes,

$$\frac{d}{dt} \begin{pmatrix} \hat{y}^{(t)} \\ \tilde{y}^{(t)} \end{pmatrix} = \begin{pmatrix} \hat{J}(\hat{y}^{(t)}) \nabla_{\tilde{x}}^{\mathrm{T}} H(y^{(t)}) \\ -\hat{J}^{\mathrm{T}}(\hat{y}^{(t)}) \nabla_{\hat{x}}^{\mathrm{T}} H(y^{(t)}) + \mathrm{div} \hat{J}^{\mathrm{T}}(\hat{y}^{(t)}) \end{pmatrix}.$$

Exercise 52. *Check that for J of the form in* (5.20),

$$\mathrm{div} J = \begin{pmatrix} 0 \\ \mathrm{div} \hat{J}^{T} \end{pmatrix}.$$

Under assumptions (5.19) and (5.20) the action of the operator $\mathcal{L}_H$ on functions becomes

$$\mathcal{L}_H f = \nabla_{\hat{x}} f \hat{J} \nabla_{\tilde{x}}^{\mathrm{T}} H - \nabla_{\tilde{x}} f \hat{J}^{\mathrm{T}} \nabla_{\hat{x}}^{\mathrm{T}} H + \nabla_{\tilde{x}} f \, \mathrm{div} \hat{J}^{\mathrm{T}}$$

and its action on probability densities becomes

$$\mu \mathcal{L}_H = -\nabla_{\hat{x}} \mu \, \hat{J} \nabla_{\tilde{x}}^{\mathrm{T}} H + \nabla_{\tilde{x}} \mu \, \hat{J}^{\mathrm{T}} \nabla_{\hat{x}}^{\mathrm{T}} H - (\mu \, \nabla_{\tilde{x}} H + \nabla_{\tilde{x}} \mu) \, \mathrm{div} \hat{J}^{\mathrm{T}}$$

Notice that if for some test function f we set $f_-(x) = f(\hat{x}, -\tilde{x})$ then

$$\mathcal{L}_H f_-(\hat{x}, -\tilde{x}) = -\mathcal{L}_H f(x). \tag{5.21}$$

A similar formula holds if we apply $\mathcal{L}_H$ to $\mu_-(x) = \mu(\hat{x}, -\tilde{x})$. Relation (5.21) has several remarkable and useful ramifications. For example, it implies that if f is an even (resp. odd) function of $\tilde{x}$ then $\mathcal{L}_H f$ is an odd (resp. even) function of $\tilde{x}$. In particular, if f and g are both even functions of $\tilde{x}$ then

$$\int g(x)\mathcal{L}_H f(x)dx = 0 \tag{5.22}$$

Exercise 53. *Establish expression* (5.22)*.*

Expression (5.21) also implies that the functions

$$y^{(-t)}(x) \qquad \text{and} \qquad \big(\hat{y}^{(t)}(\hat{x}, -\tilde{x}), -\tilde{y}^{(t)}(\hat{x}, -\tilde{x})\big)$$

both solve (5.13) with the sign of the right hand side reversed and with initial condition x. By the uniqueness of solutions to the ODE, we find therefore that the two functions are equal, i.e.

$$y^{(-t)}(x) = \big(\hat{y}^{(t)}(\hat{x}, -\tilde{x}), -\tilde{y}^{(t)}(\hat{x}, -\tilde{x})\big) \tag{5.23}$$

Exercise 54. *Assuming uniques of solutions to* (5.13)*, establish* (5.23)*.*

Equation (5.23) is called time reversal symmetry and tells us that the inverse of the flow map at time t can also be written as a forward-in-time integration using an $\tilde{x}$ initial condition with reversed sign. Time reversal symmetry is not to be confused with the notion of reversibility that we have introduced early and indeed, we have already seen that solutions to Hamilton's ODE are skew-reversible. But time reversal symmetry will have important implications in the next section where it will be used to show that a Markov process incorporating Hamilton's ODE are indeed reversible.

Finally, time reversal symmetry combined with expression (5.18) implies that if f and g are even functions of $\tilde{x}$ then

$$\int g(x)f(y^{(t)}(x))\pi_H(dx) = \int f(x)g(y^{(t)}(x))\pi_H(dx). \tag{5.24}$$

Exercise 55. *Establish equation* (5.24)*.*

Before closing this section we make a few additional observations in the case when J is a constant matrix. Notice that, in this case, $\mathcal{L}_H H = 0$ so that the value of H is exactly preserved by the solution to (5.13). Moreover, when J is constant, $\mu\mathcal{L}_H = 0$ for an density μ of the form $\mu(x) = \rho(H(x))$ for some function ρ. In words, the flow map $y^{(t)}(x)$ preserves any density of this form (including the constant density). However, it is also clear that if the value of H is preserved, the solutions to (5.13) cannot be ergodic (they cannot be irreducible). We will see in the next section that with some modification, (5.13) can be used to build effective MCMC algorithms.

5.3 Hamiltonian based MCMC schemes

In the last section we learned that solutions to the ODE (5.13) preserve the Boltzmann density. Since we have wide latitude in defining H, it is natural to ask if, given a target density $\pi(x)$, the unusual properties of solutions to (5.13) can be put to use in efficiently generating samples from π. The answer is yes, but, in general, requires the addition of some source of randomness. Indeed, as we have already mentioned, we do not expect solutions to (5.13) to be irreducible (much less ergodic).

Designing a Markov process based on Hamilton's ODE that is irreducible generally requires adding some additional "conjugate" dimensions to the system. In other words, if the target density is $\pi(\hat{x})$ where $\hat{x} \in \mathbb{R}^{\hat{d}}$, we set $x = (\hat{x}, \tilde{x})$ where $\tilde{x} \in \mathbb{R}^{\tilde{d}}$ and $d = \hat{d} + \tilde{d}$. If we choose H of the form

$$H(\hat{x}, \tilde{x}) = -\log \pi(\hat{x}) + K(\tilde{x}) \tag{5.25}$$

then the marginal of the $\hat{x}$ variables under the density

$$\pi_H(\hat{x}, \tilde{x}) \propto e^{-H(\hat{x}, \tilde{x})}$$

will be exactly π, i.e.

$$\int \pi_H(x) d\tilde{x} = \pi(\hat{x}).$$

In fact, under π_H, the $\hat{x}$ and $\tilde{x}$ variables are independent. We will take advantage of this observation by designing (higher dimensional) Markov chains to sample from π_H instead of π. Samples from π can be recovered from the $\hat{x}$ components of the resulting trajectory.

Assuming that the density proportional to $\exp(-K(\tilde{x}))$ can be sampled easily (we chose K after all), then a simple Markov chain sampling π_H can be constructed as follows: Fix an $s > 0$, draw $\tilde{X}^{(0)}$ from $\exp(-K(\tilde{x}))/\tilde{\mathcal{Z}}$, and proceed from a sample $X^{(k)} = (\hat{X}^{(k)}, \tilde{X}^{(k)})$ to generate $X^{(k+1)}$ by,

Algorithm 1. *Hamilton's ODE's with randomized conjugate variables*

```
1. Independently sample a variable $\tilde{Y}^{(k)}$ from the density proportional
   to $e^{-K(\tilde{x})}$.

2. Set $X^{(k+1)} = y^{(s)}(\hat{X}^{(k)}, \tilde{Y}^{(k)})$ where $y^{(s)}(x)$ is solution to (5.13)
   at time $s$ with initial condition $x$.
```

That the Markov chain $X^{(t)}$ preserves π_H follows from the fact that $y^{(s)}$ preserves π_H.

When we enforce that J has the structure in (5.20) and that K is an even function (i.e. $K(-\tilde{x}) = K(\tilde{x})$), we can prove that the Markov chain $\hat{X}$ generated by Algorithm 1 is reversible with respect to π. To see this note that the transition operator for the Markov chain $\hat{X}^{(t)}$ generated by Algorithm 1 is given by

$$\mathcal{T}_s f(\hat{x}) = \frac{\int f(\hat{y}^{(s)}(\hat{x}, \tilde{x}))e^{-K(\tilde{x})}d\tilde{x}}{\tilde{\mathcal{Z}}}.$$

Appealing to time reversal symmetry and the fact that K is an even function we can apply expression (5.24) to find that

$$\begin{aligned}\int g(\hat{x})\mathcal{T}_s f(\hat{x})\pi(d\hat{x}) &= \int g(\hat{x})f(\hat{y}^{(s)}(x))\pi_H(dx)\\ &= \int f(\hat{x})g(\hat{y}^{(s)}(x))\pi_H(dx)\\ &= \int f(\hat{y})\mathcal{T}_s g(\hat{y})\pi(d\hat{y}),\end{aligned}$$

i.e. that the $\hat{X}^{(k)}$ process is reversible with respect to π.

There is one major issue that we have so far avoided. We need a way to approximate solutions of (5.13). Unfortunately it is not be possible to design a scheme that exactly preserves π_H as the exact solutions do. However, when

J is constant we can find simple discrete time approximations to (5.13) whose solutions are both symplectic and have time reversal symmetry. Assuming that J is constant and has the structure in (5.20), one such discretization is the Velocity Verlet scheme:

$$\begin{aligned}
\tilde{y}_h' &= \tilde{y}_h^{(\ell)} + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}}\log\pi(\hat{y}_h^{(\ell)}) \\
\hat{y}_h^{(\ell+1)} &= \hat{y}_h^{(\ell)} + h\hat{J}\nabla^{\mathrm{T}}K(\tilde{y}_h') \\
\tilde{y}_h^{(\ell+1)} &= \tilde{y}_h' + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}}\log\pi(\hat{y}_h^{(\ell+1)})
\end{aligned} \tag{5.26}$$

where h is a small time-discretization parameter. Beginning with $y_h^{(0)} = x$, after $n = \lfloor s/h \rfloor$ iterations $y_h^{(n)}$ will approximate $y^{(s)}(x)$ up to an $\mathcal{O}(h^2)$ error. Modifications to this approximation are required when $\hat{J}$ is allowed to depend on $\hat{x}$.

Exercise 56. *Show that this is a consistent integration scheme with truncation error of order 3, i.e.*

$$\frac{y_h^{(1)}(x) - x}{h} = -J\nabla^T H(x) + \mathcal{O}(h^2)$$

Exercise 57. *Show that the Velocity Verlet scheme is symplectic (show that the jacobian of the map $x \to y_h^{(1)}(x)$ is 1), and time reversible in the same sense as the Hamiltonian ODE.*

With the discretization of (5.13) in (5.26), the practical alternative to (1) generates a chain $X_h^{(k+1)}$ according to the rule:

Algorithm 2. *Velocity Verlet with randomized conjugate variables*

1. Independently sample variable $\tilde{Y}_h^{(k)}$ from the density proportional to $\exp(-K(\tilde{x}))$

2. Set $X_h^{(k+1)} = y_h^{(n)}\left(\hat{X}_h^{(k)}, \tilde{Y}_h^{(k)}\right)$ where $y_h^{(n)}(x)$ is the solution of (5.26) after n steps with initial conditions x and n chosen by the user.

If we choose n very large in Algorithm 2, we will expend substantial effort to generate a single update of the chain $X^{(k)}$ and the scheme will become inefficient. On the other hand, consider the opposite extreme in which we choose $n = 1$ in Algorithm 2. Assume that we also make the typical choice $K(\tilde{x}) = \|\tilde{x}\|_2^2/2$. Then,

$$\begin{aligned}\hat{X}_h^{(k+1)} &= \hat{X}_h^{(k)} + h\hat{J}\left(\tilde{Y}_h^{(k)} + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}}\log\pi(\hat{X}_h^{(k)})\right)\\ &= \hat{X}_h^{(k)} + \frac{h^2}{2}\hat{J}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}}\log\pi(\hat{X}_h^{(k)}) + h\hat{J}\,\tilde{Y}_h^{(k)}.\end{aligned}$$

Setting $\delta = h^2/2$ and noting that $Y_h^{(k)}$ has mean 0 and identity covariance, we see that the above iteration is exactly of the overdamped form in (5.9). We therefore expect, that when n is small, the performance of this scheme is similar to the corresponding overdamped scheme. On the other hand, when n is very large, we expect (2) to converge slowly because H is nearly conserved by (5.26). It is often the case however, that for intermediate choices of n, the scheme in (2) outperforms its overdamped analogue (even accounting for the additional cost of the multiple evaluations of $\nabla\log\pi$ in Step 2).

An alternative approach to deriving possibly ergodic schemes based on the Hamiltonian ODE is to add appropriate random terms at each integration step. Indeed, adding $\mathcal{L}_H$ to the limiting generator $\mathcal{L}_O$ in (5.2) with π replaced by π_H yields a new limiting generator $\mathcal{L}_U = \mathcal{L}_H + \mathcal{L}_O$ or

$$\mathcal{L}_U f(x) = \frac{1}{\pi_H}\mathrm{div}\left(\pi_H(x)\nabla f(x)(S+J)(x)\right) \tag{5.27}$$

that satisfies

$$\pi_H\mathcal{L}_U = \pi_H\mathcal{L}_H + \pi_H\mathcal{L}_O = 0.$$

The subscript U in $\mathcal{L}_U$ stands for "underdamped."

Exercise 58. *Verify the above formula for the action of $\mathcal{L}_U$ on a function f.*

In fact, when J of form in (5.20) and H is an even function of $\tilde{x}$, expression (5.21) and the reversibility of $\mathcal{L}_O$ with respect to π_H imply that if f and g are even functions of $\tilde{x}$ then

$$\int f(x)\mathcal{L}_U g(x)\pi_H(dx) = \int g(x)\mathcal{L}_U f(x)\pi_H(dx). \tag{5.28}$$

Therefore we expect that discrete Markov chains corresponding (in the sense of Section **??**) to the limiting generator $\mathcal{L}_U$ should approximately preserve π_H. For example, the scheme

$$\begin{aligned} X_h^{(k+1)} = X_h^{(k)} - h\ (J+S)\,(X_h^{(k)})\nabla^{\mathrm{T}} H(X_h^{(k)}) \\ + h\,\mathrm{div}\,(J+S)\,(X_h^{(k)}) + \sqrt{2h\,S(X_h^{(k)})}\,\xi^{(k)} \end{aligned} \quad (5.29)$$

for independent $\xi^{(k)}$ with $\mathbf{E}\left[\xi^{(k)}\right] = 0$ and $\mathbf{cov}\left[\xi^{(k)}\right] = I$ (and finite higher moments) has limiting generator $\mathcal{L}_U$. Here S is assumed to be a symmetric positive semi-definite matrix and J is assumed to be anti-symmetric.

In the most common setup, one uses an J of form in (5.20) and H of form in (5.25) with

$$K(\tilde{x}) = \frac{\|\tilde{x}\|_2^2}{2},$$

and

$$S = \begin{bmatrix} 0 & 0 \\ 0 & \gamma\tilde{I} \end{bmatrix}$$

where $\gamma > 0$ and $\tilde{I}$ is the $\tilde{d} \times \tilde{d}$ identity matrix. With these choices and when $\hat{J}$ is constant, a discrete process with favorable properties (compared to (5.29)) is

$$\begin{aligned} \tilde{X}_h' &= \tilde{X}_h^{(\ell)} + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}} \log \pi(\hat{X}_h^{(\ell)}) \\ \hat{X}_h' &= \hat{X}_h^{(\ell)} + \frac{h}{2}\hat{J}\tilde{X}_h' \\ \tilde{X}_h'' &= e^{-\gamma h}\tilde{X}_h' + \sqrt{(1-e^{-2\gamma h})}\,\xi^{(\ell+1)} \\ \hat{X}_h^{(\ell+1)} &= \hat{X}_h' + \frac{h}{2}\hat{J}\tilde{X}_h'' \\ \tilde{X}_h^{(\ell+1)} &= \tilde{X}_h'' + \frac{h}{2}\hat{J}^{\mathrm{T}}\nabla^{\mathrm{T}} \log \pi(\hat{X}_h^{(\ell+1)}) \end{aligned} \quad (5.30)$$

where ξ is a sequence of independent $\tilde{d}$ dimensional Gaussian random vectors with mean zero and identity covariance (and finite higher moments).

Exercise 59. *Check that the discrete time dynamics in* (5.30) *corresponds to the limiting generator $\mathcal{L}_U$ as claimed. How large (in terms of h) is the difference between the rescaled discrete generator $\mathcal{L}_h f$ and $\mathcal{L}_U f$ for smooth bounded test functions f with bounded derivatives?*

Finally, observe that if the "friction" coefficient γ is set to 0 this scheme reduces to the Velocity Verlet scheme (5.26).

5.4 Hybrid-MC and Metroplized Langevin schemes

The methods that we have described in this chapter all introduce some systematic error in our estimate of $\pi[f]$. For these schemes we expect that $\overline{f}_N$ converges not to $\pi[f]$ but to another quantity that converges to $\pi[f]$ when h is small. Given the many physical approximations often already inherent in the specification of π (e.g. approximate models), this additional systematic error can frequently be safely ignored. However, there are settings in which very high accuracy estimates of $\pi[f]$ are needed and the systematic error can be larger than the sampling (finite N) error. In these cases it may be worth while to "Metropolize" one of the schemes introduced in this chapter. By adding a Metropolis-Hastings type accept/reject step we can guarantee that $\overline{f}_N \to \pi[f]$ exactly as N increases. The cost for this improvement in accuracy is slower convergence (e.g. increased integrated autocorrelation time). For high dimensional problems it is often better to reduce the step size parameter h than to introduce Metroplization.

We begin by describing how our overdamped Langevin scheme in (5.9) can be modified so that it exactly preserves a target density π. As long as the density of noise variables, $\xi^{(k)}$, that we choose is non-zero everywhere, and the matrix S is positive definite, a single step of (5.9) defines a transition density $q(y \,|\, x)$ for which the ratio $q(x \,|\, y)/q(y \,|\, x)$ is finite and which can be therefore be used within the general Metropolis Hastings framework. For example, when the $\xi^{(k)}$ are Gaussian random variables and the matrix S is constant, the transition density is

$$q(y \,|\, x) \propto \exp\left(-\frac{(y - x - hS\nabla \log \pi(x))^{\mathrm{T}}\, S^{-1}\, (y - x - hS\nabla \log \pi(x))^{\mathrm{T}}}{4h}\right)$$

as can be verified by writing down the density for the Gaussian random variable $\xi^{(k)}$ and then changing variables from $\xi^{(k)}$ to $X_h^{(k)}$. Note that the formula is more complicated when S depends on position because the change of variables from $\xi^{(k)}$ to $X_h^{(k)}$ is non-linear. In any case, with q in hand we can use the standard Metropolis-Hastings procedure:

Algorithm 3. *Metropolized overdamped Langevin*

1. Let $Y^{(k+1)}$ be the result of a single step of (5.9) starting from initial point $X^{(k)}$.
2. With probability
$$p_{acc}(X^{(k)}, Y^{(k+1)}) = \min\left\{1, \frac{q(X^{(k)}\,|\,Y^{(k+1)})\pi(\hat{Y}^{(k+1)})}{q(Y^{(k+1)}\,|\,X^{(k)})\pi(\hat{X}^{(k)})}\right\}$$
set $X^{(k+1)} = Y^{(k+1)}$. Otherwise set $X^{(k+1)} = X^{(k)}$.

Recalling that (5.9) is approximately reversible with respect to π and that the rejection probability is a measure of the distance between the densities $q(y\,|\,x)\pi(x)$ and $q(x\,|\,y)\pi(y)$, we should expect the rejection probability for a Metropolis-Hastings scheme with proposal density $q(y\,|\,x)$ corresponding to (5.9) to be very small. To see that this is indeed the case, consider the one dimensional case with $S = 1$. Letting $V(x) = -\log\pi(x)$, and using the change of variables
$$y \to x - hV'(x) + \sqrt{2h}\,\xi,$$
we obtain
$$\begin{aligned}\int p_{rej}(x)\pi(dx) &= \int \left|1 - \frac{q(x\,|\,y)\pi(y)}{q(y\,|\,x)\pi(x)}\right| q(y\,|\,x)\pi(x)dydx \\ &= \int \Bigg|1 - e^{-V(x-hV'(x)+\sqrt{2h}\,\xi)+V(x)} \\ &\qquad \times e^{\frac{-(\xi-\sqrt{h/2}V'(x)-\sqrt{h/2}V'(x-hV'(x)+\sqrt{2h}\,\xi))^2+\xi^2}{2}}\Bigg| e^{-\frac{\xi^2}{2}}\pi(x)dydx.\end{aligned}$$

The change of variables has re-expressed the average rejection probability as an integral of a non-negative quantity over two independent random variables both of which are independent of h. First, notice that
$$V'(x - hV'(x) + \sqrt{2h}\,\xi) = V'(x) + \sqrt{2h}V''(x)\xi + \mathcal{O}\left(h\right).$$
Expanding the terms in the integrand we find that
$$\begin{aligned}&-(\xi - \sqrt{h/2}V'(x) - \sqrt{h/2}V'(x - hV'(x) + \sqrt{2h}\,\xi))^2 + \xi^2 \\ &\qquad = -(\xi - \sqrt{2h}V'(x) - hV''(x)\xi + \mathcal{O}(h^{3/2}))^2 + \xi^2 \\ &\qquad = 2\sqrt{2h}V'(x)\xi + 2hV''(x)\xi^2 - 2h\left(V'(x)\right)^2 + \mathcal{O}(h^{3/2})\end{aligned}$$

and

$$\begin{aligned} V(x) - V(x - hV'(x) + \sqrt{2h}\,\xi) &= h(V'(x))^2 - \sqrt{2h}V'(x)\xi \\ &\quad - \frac{1}{2}V''(x)\left(\sqrt{2h}\xi - hV'(x)\right)^2 + \mathcal{O}(h^{3/2}) \\ &= h(V'(x))^2 - \sqrt{2h}V'(x)\xi - hV''(x)\xi^2 + \mathcal{O}(h^{3/2}). \end{aligned}$$

So we see that

$$\int p_{rej}(x)\pi(dx) = \int \left|1 - e^{\mathcal{O}(h^{3/2})}\right| e^{-\frac{\xi^2}{2}}\pi(x)dydx = \mathcal{O}(h^{3/2})$$

and we expect the rejection rate to be very small at least when h is small.

Now let

$$H(\hat{x}, \tilde{y}) = -\log \pi(\hat{x}) + K(\tilde{y})$$

where $K(-\tilde{x}) = K(\tilde{x})$ and let π_H be the density proportional to e^{-H}. Suppose that we wish to modify Algorithm 2 so that it still preserves π_H, even though the exact solution to the Hamiltonian ODE (5.15), $y^{(t)}$, is replaced by $y_h^{(k)}$, the Velocity Verlet discrete time approximation in (5.26). The difference between the trajectories of $y^{(t)}$ and $y_h^{(k)}$ over a finite time interval (so over $\mathcal{O}(h^{-1})$ steps) is $\mathcal{O}\left(h^2\right)$. This small error translates into a small error in the sampling scheme (2). The error can either be reduced to tolerable levels by decreasing h (and correspondingly increasing n) or it can be eliminated altogether by introducing a Metropolis accept-reject step: sample $\tilde{X}^{(0)}$ from $\exp\left(-K(\tilde{x})\right)/\tilde{\mathcal{Z}}$ and, given a sample $X^{(k)} = (\hat{X}^{(k)}, \tilde{X}^{(k)})$, generate $X^{(k+1)}$ by

Algorithm 4. *Hybrid Monte Carlo*

1. Independently sample variable $\tilde{Y}(k)$ from the density proportional to $\exp\left(-K(\tilde{x})\right)$
2. Set $Y^{(k+1)} = y_h^{(n)}\left(\hat{X}^{(k)}, \tilde{Y}^{(k)}\right)$ where $y_h^{(\ell)}(x)$ solves (5.26) with initial conditions x and n is chosen by the user.
3. With probability

   $$p_{acc}(X^{(k)}, Y^{(k+1)}) = \min\left\{1, \frac{\pi_H(Y^{(k+1)})}{\pi_H(\hat{X}^{(k)}, \tilde{Y}^{(k)})}\right\}$$

   set $X^{(k+1)} = Y^{(k+1)}$. Otherwise set $X^{(k+1)} = X^{(k)}$.

Just as in Algorithm 1, Step 1 exactly preserves π_H. However, unlike Algorithm 1, Step 2 does not exactly preserve π_H. The purpose of Step 3 is to correct the error introduced in Step 2. Examining this procedure, the first question the reader should ask is "why does the transition density $q(y \,|\, x)$ not appear in the acceptance probability?" The proposal density corresponding to Step 2 is

$$q(y \,|\, x) = \delta(y - y_h^{(t_n)}(x)) \tag{5.31}$$

which certainly does not satisfy $q(y \,|\, x) = q(x \,|\, y)$.

In fact, Algorithm 1 is an example of another kind of Metropolis framework. Suppose that φ is a smooth involution, i.e. $\varphi = \varphi^{-1}$ and, given a target density μ and a proposal density $q(y|x)$ define the transformed densities

$$\mu_\varphi(x) = |D\varphi(x)|\, \mu(\varphi(x))$$

and

$$q_\varphi(y \,|\, x) = |D\varphi(y)|\, q(\varphi(y)|\varphi(x))$$

where we have used the fact that, for a smooth involution,

$$|D\varphi(x)| = \frac{1}{|D\varphi(\varphi(x))|}.$$

Note that the mapping $\varphi(x) = (\hat{x}, -\tilde{x})$ is a smooth involution. In fact, when H is an even function of the $\tilde{x}$ variables as we have assumed, this involution preserves the density π_H.

Now consider the following slight generalization of the Metropolis-Hastings scheme:

Algorithm 5. *Metroplis-Hastings with Involution*

1. Generate a random variable $Y^{(k+1)}$ from the proposal distribution $q(y \,|\, X^{(k)})$.

2. With probability

$$p_{acc}(X^{(k)}, Y^{(k+1)}) = \min\left\{1, \frac{q_\varphi(X^{(k)} \,|\, Y^{(k+1)})\, \mu_\varphi(Y^{(k+1)})}{q(Y^{(k+1)} \,|\, X^{(k)})\, \mu(X^{(k)})}\right\}$$

set $X^{(k+1)} = Y^{(k+1)}$. Otherwise set $X^{(k+1)} = \varphi(X^{(k)})$.

Instead of enforcing reversibility, Algorithm 5 results in a chain with transition operator $\mathcal{T}$ satisfying

$$\int f_\varphi(x)\mathcal{T}g_\varphi(x)\mu(dx) = \int g(x)\mathcal{T}f(x)\mu(dx) \tag{5.32}$$

for any test functions f and g and with $f_\varphi(x) = f(\varphi(x))$. Like reversibility, this condition implies that the resulting Markov chain preserves μ.

Exercise 60. *Show that for any density μ, if the transition density for a Markov chain satisfies* (5.32) *then the Markov chain preserves μ.*

In order to show that Algorithm 5 satisfies (5.32), first observe that

$$\begin{aligned}
&p_{acc}(\varphi(x), \varphi(y))\, q_\varphi(y\,|\,x)\, \mu_\varphi(x)\\
&\qquad = \min\left\{q_\varphi(y|x)\mu_\varphi(x), \frac{q_\varphi(\varphi(x)|\varphi(y))\mu_\varphi(\varphi(y))q_\varphi(y\,|\,x)\,\mu_\varphi(x)}{q(\varphi(y)\,|\,\varphi(x))\mu(\varphi(x))}\right\}\\
&\qquad = \min\left\{q_\varphi(y|x)\mu_\varphi(x), q(x|y)\mu(y)\right\}\\
&\qquad = p_{acc}(y,x)\, q(x\,|\,y)\, \mu(y)
\end{aligned}$$

where, to obtain the third equality we have used the fact that

$$\frac{q_\varphi(\varphi(x)\,|\,\varphi(y))\mu_\varphi(\varphi(y))}{q(\varphi(y)|\varphi(x))\mu(\varphi(x))} = \frac{q(x|y)\mu(y)}{q_\varphi(y|x)\mu_\varphi(x)}.$$

The transition operator resulting from Algorithm 5 is

$$\mathcal{T}f(x) = p_{rej}(x)f_\varphi(x) + \int f(y)q(y\,|\,x)p_{acc}(x,y)dy$$

where

$$p_{rej}(x) = 1 - \int q(y\,|\,x)p_{acc}(x,y)dy.$$

From these expressions we see that

$$\begin{aligned}\int f_\varphi(x)\mathcal{T}g_\varphi(x)\mu(dx) &= \int f_\varphi(x)p_{rej}(x)g(x)\mu(dx)\\ &\qquad + \int f_\varphi(x)g_\varphi(y)p_{acc}(x,y)q(y|x)\mu(x)dydx\\ &= \int f_\varphi(x)p_{rej}(x)g(x)\mu(dx)\\ + \int f(x)g(y)\, p_{acc}(\varphi(x),\varphi(y))\, &q_\varphi(y|x)\mu_\varphi(x)dxdy\\ &= \int f_\varphi(x)p_{rej}(x)g(x)\mu(dx)\\ &\qquad + \int g(x)f(y)p_{acc}(x,y)q(y\,|\,x)\mu(x)dxdy\\ &= \int g(x)\mathcal{T}f(x)\mu(dx).\end{aligned}$$

Returning to Algorithm 4 and the particular transition density $q(y\,|\,x)$ in (5.31), observe that, for any test function f of $x \in \mathbb{R}^{\hat{d}}$ and $y \in \mathbb{R}^{\tilde{d}}$, volume preservation and time reversibility of the flow map $x \to y_h^{(n)}(x)$ imply that

$$\begin{aligned}\int f(x,y)q(x\,|\,y)dxdy &= \int f(x, y_h^{(n)}(x))dx\\ &= \int f(y_h^{(-n)}(y), y)dy\\ &= \int f((\hat{y}_h^{(n)}(\hat{y},-\tilde{y}), -\tilde{y}_h^{(n)}(\hat{y},-\tilde{y}), y)dy\\ &= \int f(x,y)q(\hat{x},-\tilde{x}\,|\,\hat{y},-\tilde{y})dxdy.\end{aligned}$$

Since this is true for any f we can conclude that

$$q_\varphi(x\,|\,y) = q(y\,|\,x)$$

for the particular involution $\varphi(x) = (\hat{x}, -\tilde{x})$, which explains why we can omit q in the acceptance probability in Algorithm 4. Since this involution also preserves π_H, we find that the acceptance probability in Algorithm 4 is indeed of the form in Algorithm 5. Finally note that, since the $\tilde{x}$ variables are randomized at each iteration of Algorithm 4, we could set $X^{(k+1)} = \varphi(X^{(k)})$ upon rejection rather than $X^{(k+1)} = X^{(k)}$ without changing the distribution of resulting chain.

Exercise 61. *Though the full Markov chain X generated by Algorithm 4 does not satisfy detailed balance, the process $\hat{X}$ is also a Markov chain and is reversible. Show this.*

Before moving on we point out that the arguments above did not actually require that the Hamiltonian appearing in the acceptance probability in Algorithm 4 be the same as the one used to define the evolution (5.26). In fact, as long as the mapping $x \to y_h^{(n)}(x)$ is volume preserving time reversible and H is even in $\tilde{x}$, $y_h^{(n)}(x)$ need have no relationship at all with H. Of course the reason we expect Hybrid Monte Carlo to be an effective method is that, for sufficiently small h, the solution of (5.26) should very nearly conserve the value of H and therefore result in a high acceptance probability. Certainly if the error introduced is not controllable (by reducing some user defined parameter) then one cannot justify omitting the accept/reject step. So introducing an evolution that does not nearly conserve H may be counter-productive.

Finally, we turn our attention to Metropolizing the underdamped Langevin scheme in (5.30). The transition density for a single step of (5.30) is

$$q(y \,|\, x) \propto \delta\left(\hat{y} - \hat{x} - \frac{h}{2}\hat{J}(\tilde{y} + \tilde{x}) + \frac{h^2}{4}\hat{J}\hat{J}^{\mathrm{T}}\left(\nabla^{\mathrm{T}} \log \pi(\hat{y}) - \nabla^{\mathrm{T}} \log \pi(\hat{x})\right)\right) \times r(x, y) \quad (5.33)$$

where we have introduced the function

$$r(x, y) = \exp\left(-\frac{\|\tilde{y} - e^{-\gamma h}\tilde{x} - \frac{1}{2}h\hat{J}^{\mathrm{T}}\left(e^{-\gamma h}\nabla^{\mathrm{T}} \log \pi(\hat{x}) + \nabla^{\mathrm{T}} \log \pi(\hat{y})\right)\|_2^2}{2(1 - e^{-2h\gamma})}\right).$$

Exercise 62. *Show that the transition density for a single step of* (5.30) *is the one given in* (5.33).

The transition operator corresponding to the proposal distribution $q(y \,|\, x)$ in (5.33) satisfies (5.32) with $\mu = \pi_H$ and $\varphi(x) = (\hat{x}, -\tilde{x})$ to within an $\mathcal{O}(h^2)$ error.

Exercise 63. *Show this.*

Plugging the proposal density in (5.33) into Algorithm 5, we find that the delta functions exactly cancel and we obtain

Algorithm 6. *Metropolized underdamped Langevin*

```
1. Let $Y^{(k+1)}$ be the result of a single step of (5.30) starting from
   initial point $X^{(k)} = (\hat{X}^{(k)}, \tilde{X}^{(t)})$.

2. With probability

   $$p_{acc}(X^{(k)}, Y^{(k+1)}) = \min\left\{1, \frac{\pi_H(Y^{(k+1)})\, r(\hat{Y}^{(k+1)}, -\tilde{Y}^{(k+1)}, \hat{X}^{(k)}, -\tilde{X}^{(k)})}{\pi_H(X^{(k)})\, r(X^{(k)}, Y^{(k+1)})}\right\}$$

   set $X^{(k+1)} = Y^{(k+1)}$. Otherwise set $X^{(k+1)} = \left(\hat{X}^{(k)}, -\tilde{X}^{(k)}\right)$.
```

Exercise 64. *Consider a set of 2-dimensional vectors, $\vec{\sigma}_i \in \mathbb{R}^2$ indexed by the 1-dimensional periodic lattice $\mathbb{Z}_L$ and with $\|\vec{\sigma}_i\|_2 = 1$. The nearest neighbor XY model of statistical physics assigns to these vectors the density*

$$\pi(\vec{\sigma}) = \frac{e^{\beta \sum_{i \leftrightarrow j} \vec{\sigma}_i \cdot \vec{\sigma}_j}}{\mathcal{Z}}.$$

In terms of the angles $\theta_i \in [-\pi, \pi)$ of the vectors $\vec{\sigma}_i$, this density becomes

$$\pi(\theta) = \frac{e^{\beta \sum_{i \leftrightarrow j} \cos(\theta_i - \theta_j)}}{\mathcal{Z}}.$$

Write a routine to sample the XY model using both (5.8) and a Metropolized version of (5.8). Compare the Metropolized and un-Metropolized schemes for different values of h (but note that the total number of time-steps you use should scale like h^{-1}). Make your comparisons in terms of the integrated autocorrelation time of the variable

$$\frac{M_1(\sigma)}{\|M(\sigma)\|_2}$$

which is the cosine of the angle of the magnetization vector,

$$M(\sigma) = \sum_{i=0}^{L-1} \vec{\sigma}_i \in \mathbb{R}^2.$$

What do you observe when L increases.

Exercise 65. *Write a routine to sample the XY model described in Exercise 64 using Algorithm 4 both with and without the Metropolis accept/reject step. You're free to choose the other parameters of the algorithm (K, J, n, and h) as you like (but be clear about your choices). Compare the results to those of Exercise 64 using integrated autocorrelation time of the cosine of the angle of magnetization as your measure of efficiency. Make sure you are accounting correctly for the cost to generate each step of the chain (e.g. as measured by the number of evaluations of* $\nabla \log \pi$.

Exercise 66. *Sample the XY model described in Exercise 64 using* (5.30) *with $\hat{J} = \hat{I}$. Compare to the Metropolized scheme in Algorithm 6 for different values of h and for different values of γ using integrated autocorrelation time of the cosine of the angle of magnetization as your measure of efficiency. Compare to your results from Exercises 64 and 65. Which of the schemes do you prefer for sampling the XY model?*

5.5 bibliography