Chapter 6

Conditioning and affine invariance

In this chapter we consider design choices for the Markov chain Monte Carlo methods introduced so far. In many high dimensional problems, the MCMC schemes that we have introduced will require very small perturbations, $X^{(k+1)} - X^{(k)}$, at each step to avoid either large step size errors and instabilities (for non-Metropolized schemes) or high rejection rates (for Metropolized schemes). Fortunately, as we will see, carefully choosing the design parameters of the scheme (e.g. the matrix S in (5.9)) allows the methods to be used with much larger step sizes, which in turn yield faster convergence.

6.1 Badly scaled measures

Consider constructing an MCMC sampling scheme for the Gaussian density

$$\pi(x) \propto e^{-\frac{1}{2}x_1^2 - \frac{1}{2\epsilon^2}x_2^2}. \tag{6.1}$$

Suppose that we attempt to use a Metropolis scheme with proposal density $q(y \,|\, x) = \mathcal{N}(x, \sigma^2 I)$ to sample from π. It is possible to show that in this case

as $\epsilon \to 0$, the average acceptance probability satisfies

$$\int (1 - p_{rej}(x))\pi(dx) = \mathcal{O}\left(\frac{\epsilon}{\sigma}\right).$$

Exercise 67. *In 1 dimension for the density $\pi(x) \propto e^{-\frac{1}{2\epsilon^2}x^2}$ and proposal density $q(y \,|\, x) = \mathcal{N}(x, \sigma^2)$, show that*

$$\lim_{\epsilon \to 0} \frac{\sigma}{\epsilon} \int p_{acc}(x, y) q(dy \,|\, x) \pi(dx) = K$$

where the constant K is independent of σ and ϵ.

So, in order to maintain a reasonable acceptance probability as $\epsilon \to 0$, we need to shrink σ at the same rate as ϵ. But the x and y coordinates are independent under π and the variance of the x coordinate is 1. We cannot hope to estimate, for example, the variance of the x coordinate under π without at least $\mathcal{O}(1/\epsilon)$ steps of our MCMC method (many more in fact). It may already be intuitively clear that the solution to this problem is to modify the proposal density so that larger steps are proposed in the x direction than in the y direction. This modification is a simple but informative representative of the strategies we will discuss in this section.

The problems caused by multiple scales can be very clearly understood within the framework of the overdamped Langevin methods introduced in the last chapter (which are also, the reader will recall, small proposal size limits of Metropolis schemes). In particular, consider sampling the general Gaussian density

$$\pi(x) \propto e^{-\frac{1}{2}x^{\mathrm{T}} M^{-1} x}$$

using the simple overdamped Langevin scheme in (5.8), i.e. using

$$X_h^{(k+1)} = \left(I - hSM^{-1}\right) X_h^{(k)} + \sqrt{2hS}\, \xi^{(k+1)} \tag{6.2}$$

where we will assume that $\xi^{(k)} \sim \mathcal{N}(0, 1)$. As long as $X_h^{(0)}$ is Gaussian, $X^{(k)}$ will remain Gaussian for all iterations. We can therefore compute its invariant measure by finding values for the mean and covariance that are left invariant by the iteration. The vector 0 of all zeros is the invariant mean. Suppose

that $X_h^{(k)} \sim \mathcal{N}(0, Q)$ for some covariance matrix Q. Then the covariance of $X_h^{(k+1)}$ is

$$\mathbf{E}\left[X_h^{(k+1)}\left(X_h^{(k+1)}\right)^{\mathrm{T}}\right] = \left(I - hSM^{-1}\right)\mathbf{E}\left[X_h^{(k)}\left(X_h^{(k)}\right)^{\mathrm{T}}\right]\left(I - hSM^{-1}\right)^{\mathrm{T}} + 2h\sqrt{S}\mathbf{E}\left[\xi^{(k+1)}\left(\xi^{(k+1)}\right)^{\mathrm{T}}\right]\sqrt{S}^{\mathrm{T}}$$

where we have used the fact that $\xi^{(k+1)}$ is independent of $X_h^{(k)}$ and has mean 0. Plugging in the assumed covariances of $X_h^{(k)}$ and $\xi^{(k+1)}$ and recalling that M and S are both symmetric, we find that

$$\mathbf{E}\left[X_h^{(k+1)}\left(X_h^{(k+1)}\right)^{\mathrm{T}}\right] = \left(I - hSM^{-1}\right)Q\left(I - hSM^{-1}\right)^{\mathrm{T}} + 2hS.$$

A bit of algebra reveals that $Q = M + hR$ for R solving the equation

$$SM^{-1}R + RM^{-1}S = SM^{-2}S$$

satisfies

$$Q = \left(I - hSM^{-1}\right)Q\left(I - hSM^{-1}\right)^{\mathrm{T}} + 2hS,$$

i.e. is an invariant covariance for the process in (6.2). So the bias in averages with respect to the invariant measure $\mathcal{N}(0, Q)$ of (6.2) will be small when h is small. In fact, we can make a small perturbation of the diffusion coefficient in this Gaussian setting and arrive at a slightly modified iteration

$$X_h^{(k+1)} = \left(I - hSM^{-1}\right)X_h^{(k)} + \sqrt{2hS\left(I - \frac{h}{2}M^{-1}S\right)}\xi^{(k+1)} \tag{6.3}$$

that preserves π exactly. Though such a modification would not be possible for a general non-Gaussian π, it will simplify comparisons below and the general conclusions of the discussion in this section will be applicable to realistic problems.

Because S and M are both positive definite, so is the matrix SM^{-1}, i.e. all of the eigenvalues of SM^{-1} are positive. Suppose that λ_{max} is the largest eigenvalue of the matrix SM^{-1} and v_{max} is the corresponding right eigenvector. Then if $X_h^{(0)} = v_{max}$,

$$\mathbf{E}\left[X_h^{(k)}\right] = (1 - h\lambda_{max})^k v_{max}.$$

Clearly we need to choose h so that $|1 - h\lambda_{max}| < 1$ to ensure that $\mathbf{E}\left[X_h^{(k)}\right]$ converges to the invariant mean vector 0. This restriction on the maximum step size is called a linear stability condition. In particular we need to choose $h < 2/\lambda_{max}$.

Intuitively we know that a smaller value of h should correspond to slower convergence. To quantify our intuition, observe that

$$\begin{aligned}\mathbf{cov}_\pi\left[X_h^{(k)}, X_h^{(0)}\right] &= \mathbf{E}\left[\mathbf{E}\left[X_h^{(k)}\left(X_h^{(0)}\right)^{\mathrm{T}} \mid X_h^{(k-1)}\right]\right] \\ &= \left(I - hSM^{-1}\right)\mathbf{cov}_\pi\left[X_h^{(k-1)}, X_h^{(0)}\right] \\ &= \left(I - hSM^{-1}\right)^k M.\end{aligned}$$

Taking an arbitrary vector v and computing the integrated autocorrelation time corresponding to the test function $f(x) = v^{\mathrm{T}}x$, we find that

$$\mathbf{cov}_\pi\left[f(X_h^{(k)}), f(X_h^{(0)})\right] = v^{\mathrm{T}}\left(I - hSM^{-1}\right)^k Mv$$

and

$$\tau_f = 1 + 2\left(v^{\mathrm{T}}Mv\right)^{-1} v^{\mathrm{T}}\left(\sum_{k=1}^{\infty}\left(I - hSM^{-1}\right)^k\right)Mv.$$

Assuming that our linear stability condition is satisfied, we can rewrite this expression as

$$\tau_f = 1 + 2\left(v^{\mathrm{T}}Mv\right)^{-1} v^{\mathrm{T}}\left(h^{-1}MS^{-1} - I\right)Mv = \frac{2\,v^{\mathrm{T}}MS^{-1}Mv}{h\,v^{\mathrm{T}}Mv} - 1.$$

Plugging in $v = v_{min}$, the left eigenvector corresponding to λ_{min}, the smallest eigenvalue of SM^{-1}, i.e. one over largest eigenvalue of MS^{-1} we find that, for the particular observable $f_*(x) = v_{min}^{\mathrm{T}}x$,

$$\tau_{f_*} = \frac{2}{h\,\lambda_{min}} - 1. \tag{6.4}$$

Recalling our linear stability bound $h < 2/\lambda_{max}$ we find that

$$\tau_{f_*} > \kappa(SM^{-1}) - 1.$$

where $\kappa(SM^{-1})$ is the ratio of the largest to smallest eigenvalues of the matrix SM^{-1} and is usually referred to as the condition number of the matrix.

Remember that the choice of S is at the users discretion. If we choose $S = I$ we find that the integrated autocorrelation time for $f_*(x) = v_{min}^{\mathrm{T}} x$ satisfies the bound

$$\tau_{f_*} > \kappa(M^{-1}) - 1,$$

so that, even when we choose the largest time step h possible, convergence of the MCMC estimate of this observable will be slow whenever M has a wide range of eigenvalues. On the other hand, we also observe from (6.4) that if we choose $S = M$, then $\kappa(SM^{-1}) = 1$ and the worst case integrated autocorrelation (among functions of form $f(x) = v^{\mathrm{T}} x$) is

$$\tau_{f_*} = \frac{2}{h} - 1$$

with stability condition $h < 2$. In particular, the performance of the over-dammped scheme (6.3) is completely independent of M if we choose $S = M$.

6.2 Affine invariance and Newton-type schemes

In the last section we saw that, when π is Gaussian we can choose S in (5.8) to yield substantially faster convergence when the covariance of π has a wide range of eigenvalues. But does this observation apply to more general densities? Clearly we will not be able to answer this question by exactly computing the integrated autocorrelation function. Instead we will use the notion of affine invariance. To introduce that idea, given a specific density, π, assume that the goal is to sample the distribution with density

$$\pi_\varphi(x) \propto \pi(\varphi(x))$$

where $\varphi(x) = Ax + a$ with A, any invertible matrix, and a fixed vector a. The mapping φ is called an affine transformation. Note that the Jacobian of the transformation φ is constant. First, we will consider the application (5.8) to the density π_φ which generates the chain

$$X_h^{(k+1)} = X_h^{(k)} + hS_\varphi \nabla^{\mathrm{T}} \log \pi_\varphi(X_h^{(k)}) + \sqrt{2hS_\varphi}\, \xi^{(k+1)}$$

where S_φ is a symmetric positive definite matrix to be chosen by the user and possibly depending on the target density. Writing the derivative of π_φ

in terms of a derivative of π we find that

$$X_h^{(k+1)} = X_h^{(k)} + hS_\varphi A^{\mathrm{T}} \nabla^{\mathrm{T}} \log \pi(\varphi(X_h^{(k)})) + \sqrt{2hS_\varphi}\, \xi^{(k+1)}.$$

Setting $Y_h^{(k)} = \varphi(X_h^{(k)})$ the last recursion becomes

$$\varphi^{-1}(Y_h^{(k+1)}) = \varphi^{-1}(Y_h^{(k)}) + hS_\varphi A^{\mathrm{T}} \nabla^{\mathrm{T}} \log \pi(Y_h^{(k)}) + \sqrt{2hS_\varphi}\, \xi^{(k+1)}$$

or, after using the formula $\varphi^{-1}(y) = A^{-1}(y - a)$ and multiplying both sides of the equation by A,

$$Y_h^{(k+1)} = Y_h^{(k)} + hAS_\varphi A^{\mathrm{T}} \nabla^{\mathrm{T}} \log \pi(Y_h^{(k)}) + A\sqrt{2hS_\varphi}\, \xi^{(k+1)}.$$

Now notice that if we fix some symmetric positive definite matrix S and choose

$$S_\varphi = A^{-1} S A^{-\mathrm{T}} \tag{6.5}$$

we can assume that $\sqrt{S_\varphi} = A^{-1}\sqrt{S}$ so that

$$Y_h^{(k+1)} = Y_h^{(k)} + hS \nabla^{\mathrm{T}} \log \pi(Y_h^{(k)}) + \sqrt{2hS}\, \xi^{(k+1)}.$$

This iteration is completely independent of the choice of A. In particular, if we apply the scheme with $A = I$ and $a = 0$, i.e. to the original density, the scheme becomes exactly the iteration in the last display. Finally, notice that

$$\mathbf{cov}_{\pi_\varphi} = A^{-1} \mathbf{cov}_\pi\, A^{-T}.$$

Thus the choice $S_\varphi = \mathbf{cov}_{\pi_\varphi}$ has exactly the form required.

We can summarize what we have just learned by saying that, when applied to the density π_φ, the scheme (5.8) with S_φ chosen to be the covariance of the target density, results in a chain that is equal in distribution to $\varphi(X_h^{(k)})$ where $X_h^{(k)}$ is the chain generated by the scheme applied to π. In other words, the scheme (5.8) with S equal to the covariance of the target density is *affine invariant*. Note that we have phrased the scheme without any reference to a particular affine transformation. Indeed, the sampling problems that we face in applications are not associated with any affine transformation (at least that we are aware of). To fully appreciate the impact of affine invariance in practice, observe that any MCMC scheme with the affine invariance property will converge exactly as quickly on the target density as it will on any affine

transformation of the target density, including transformations that alleviate conditioning issues. For example, an affine invariant scheme will have the same performance on the poorly scaled density in (6.1) as it will have on the well scaled density

$$\pi(x) \propto e^{-\frac{1}{2}x_1^2 - \frac{1}{2}x_2^2}.$$

More generally, affine invariance can be phrased as the requirement that, for any affine transformation φ and any test function f

$$E_{\varphi^{-1}(y)}\left[f(\varphi(X_\varphi^{(k)}))\right] = E_x\left[f(X^{(k)})\right] \tag{6.6}$$

where $X_\varphi^{(k)}$ is the Markov chain generated by an MCMC scheme applied to sample the density π_φ and $X^{(k)}$ is the Markov chain generated by applying the same scheme to π. In terms of the transition operators $\mathcal{T}_\varphi$ and $\mathcal{T}$ for the two chains (assuming that they are time-homogenous) affine invariance becomes

$$\mathcal{T}_\varphi[f \circ \varphi] \circ \varphi^{-1} = \mathcal{T} f. \tag{6.7}$$

Our goal in this Chapter is to explore Markov chain Monte Carlo methods satisfying (6.7).

Exercise 68. *Show that the Metropolis-Hastings scheme with* $q(y|x) = \mathcal{N}(x, \mathbf{cov}_\pi)$ *satisfies* (6.7)*.*

We will begin by considering the stochastic thermostat schemes that we developed in the previous chapter. The important properties of these schemes were largely (but not completely) encoded in their limiting generators. This is also the case for affine invariance. From (6.7) we see that a scheme will be affine invariant when the generator $\mathcal{L}_\varphi$ of the Markov chain generated by the scheme when applied to the transformed density π_φ satisfies

$$\mathcal{L}_\varphi[f \circ \varphi] \circ \varphi^{-1} = \mathcal{L} f. \tag{6.8}$$

Of course if a scheme is affine invariant then its limiting generator should also satisfy (6.8). And conversely, if the limiting generator of a scheme satisfies (6.8) then the scheme should be affine invariant up to a small (in the step size h) deviation. In the following, to avoid dealing with the particular details of any one discrete time approximation we will consider affine invariance in terms of the limiting generator rather than in terms of the discrete time

scheme directly. Of course, if possible, one should design a discrete time chain that is affine invariant (and not just in the continuous time limit).

The transformation formulas

$$\nabla[f \circ \varphi] = \nabla f \circ \varphi \, A \qquad \text{and} \qquad \operatorname{div}[M \circ \varphi] = \operatorname{div}[MA^{\mathrm{T}}] \circ \varphi \tag{6.9}$$

will be useful for verifying (6.8) for limiting generators.

Exercise 69. *Check the above formulas.*

For example, if $\mathcal{L}_\varphi$ is the limiting generator of the scheme (5.9) applied to π_φ with S replaced by a φ dependent symmetric positive definite matrix S_φ then

$$\begin{aligned}\mathcal{L}_\varphi[f \circ \varphi] &= \frac{1}{\pi \circ \varphi} \operatorname{div}[\pi \circ \varphi \, \nabla[f \circ \varphi] \, S_\varphi] \\ &= \frac{1}{\pi \circ \varphi} \left(\operatorname{div}\left[\pi \, \nabla f \, A(S_\varphi \circ \varphi^{-1})A^{\mathrm{T}}\right]\right) \circ \varphi.\end{aligned}$$

If our choice of S_φ has the transformation property

$$S_\varphi(x) = A^{-1} S(\varphi(x)) A^{-\mathrm{T}} \tag{6.10}$$

then we find that

$$\mathcal{L}_\varphi[f \circ \varphi] = (\mathcal{L}_O f) \circ \varphi$$

which is equivalent to (6.8).

We have already seen that the constant matrix $S_\varphi = \mathbf{cov}_{\pi_\varphi}$ satisfies the required transformation property (6.10) for $S = \mathbf{cov}_\pi$. For an example of a choice of S that depends on position (and on the target density) and yields an affine invariant scheme, consider (5.9) with $S = -\left(D^2 \log \pi\right)^{-1}$. The matrix $-D^2 \log \pi$ is called a Hessian and is always symmetric (as long as $\log \pi_\varphi$ is twice continuously differentiable) though it need not be positive definite (in which case the Hessian can be replaced by a positive definite square root of the square of the Hessian). The inverse Hessian $S = -\left(D^2 \log \pi(x)\right)^{-1}$ satisfies (6.10) and therefore yields an affine invariant limiting generator.

Exercise 70. *Check that the inverse Hessian satisfies* (6.10)*, i.e. that*

$$-\left(D^2 \log \pi_\varphi(x)\right)^{-1} = -A^{-1}\left(D^2 \log \pi(x)\right)^{-1} A^{-T}.$$

In fact, (5.9) with this choice for S_φ is affine invariant even for finite h. If we set $V = -\log \pi$ and $H = D^2V$, (5.9) applied to the target measure π using $S = H^{-1}$ yields the iteration

$$X_h^{(k+1)} = X_h^{(k)} - h\, H^{-1}(X_h^{(k)})\nabla^{\mathrm{T}} V(X_h^{(k)}) + h\,\mathrm{div} H^{-1}(X_h^{(k)}) + \sqrt{2h\, H^{-1}(X_h^{(k)})}\,\xi^{(k+1)}. \quad (6.11)$$

In fact, if we neglect the last two terms we get the deterministic iteration

$$x_h^{(k+1)} = x_h^{(k)} - hH^{-1}(x_h^{(k)})\nabla^{\mathrm{T}} V(x_h^{(k)})$$

which is just Newton's method for finding the minimum of V. The fact that Newton's method is affine invariant is largely responsible for its success as an optimization tool for complex multivariable functions. Because of the close relationship between the iteration in (6.11) and Newton's method we refer to (6.11) as the stochastic Newton iteration. As for the scheme using $S = \mathbf{cov}_\pi$, we can phrase the stochastic Newton iteration without reference to any specific affine transformation. Because of the transformation properties of the Hessian, the scheme is "automatically" affine invariant.

We have seen that stochastic Newton, like its deterministic counterpart, is affine invariant. Unfortunately, also like its deterministic counterpart, it suffers from several drawbacks on high dimensional problems. The first and most severe of these drawbacks is that for many problems it is not practical to ask that the Hessian matrix H be available. Moreover, even if the Hessian is available it will often not be positive definite as required for the scheme to function properly. In the contexts of deterministic optimization and root finding the solution to these problems has been to replace the Hessian by an available and well behaved approximation. If such a matrix is available it can also replace H in the stochastic Newton iteration in (6.11). The resulting schemes are unlikely to be affine invariant, but often perform substantially better than the simple overdamped scheme with $S = I$.

Exercise 71. *Use a Metropolized version of the overdamped stochastic Newton scheme, to sample from the Rosenbrock density,*

$$\pi(x) \propto \exp\left(-\frac{100(x_2 - x_1^2)^2 + (1 - x_1)^2}{20}\right).$$

It may be useful to note that when S is a 2×2 matrix with $trace(S) \neq 0$, the matrix

$$R = \frac{S + \sqrt{det(S)}\, I}{\sqrt{trace(S) + 2\sqrt{det(S)}}}$$

is a square root of S. The inverse Hessian, $S = (-D^2 \log \pi(x))^{-1}$, will not be positive definite for all x. When it is not you can replace the inverse Hessian by the square of the matrix obtained by taking absolute values before each square root in the formula for R above. Since you are Metropolizing, you can omit the divergence of S term in the overdamped proposal step without introducing additional bias (though this may increase the rejection rate). Compare the performance of this preconditioned scheme to the Metropolized overdamped scheme with $S = I$.

In the next section we will discus another, particularly simple family of schemes satisfying the affine invariance property. Before closing this section, however, we pause to mention that one can also define affine invariant underdamped Langevin sampling schemes. Recall that those schemes sample the density

$$\pi_H(x) \propto \pi(\hat{x}) e^{-K(\tilde{x})}$$

where $x = (\hat{x}, \tilde{x})$ with $\hat{x} \in \mathbb{R}^{\hat{d}}$ and $\tilde{x} \in \mathbb{R}^{\tilde{d}}$. The limiting generator of an underdamped scheme (e.g. (5.30)) is of the form

$$\mathcal{L}_U f = (\mathcal{L}_H + \mathcal{L})\, f = \frac{1}{\pi_H} \mathrm{div}\, (\pi_H \nabla f (J + S))$$

introduced in the last chapter, where $J(x)$ is an antisymmetric matrix valued function and $S(x)$ is a symmetric matrix valued function. We will assume that J has the particular structure

$$J(x) = \begin{bmatrix} 0 & -\hat{J}(\hat{x}) \\ \hat{J}^{\mathrm{T}}(\hat{x}) & 0 \end{bmatrix} \tag{6.12}$$

where $\hat{J}(\hat{x}) \in \mathbb{R}^{\hat{d} \times \tilde{d}}$ and that

$$S(x) = \begin{bmatrix} 0 & 0 \\ 0 & \gamma \tilde{I} \end{bmatrix} \tag{6.13}$$

for some $\gamma > 0$ where $\tilde{I}$ is the $\tilde{d} \times \tilde{d}$ identity matrix.

In this underdamped context we will also restrict the set of affine transformations that we allow so that $\varphi(x) = (\hat{\varphi}(\hat{x}), \tilde{x}) = (\hat{A}\hat{x} + \hat{a}, \tilde{x})$,

$$A = \begin{bmatrix} \hat{A} & 0 \\ 0 & \tilde{I} \end{bmatrix},$$

and $\hat{A}$ is a $\hat{d} \times \hat{d}$ invertible matrix. Notice that for an affine transformation of this form and for J and S as in (6.12) and (6.13),

$$AJ(x)A^{\mathrm{T}} = \begin{bmatrix} 0 & -\hat{A}\hat{J}(\hat{x}) \\ \hat{J}^{\mathrm{T}}(\hat{x})\hat{A}^{\mathrm{T}} & 0 \end{bmatrix} \qquad \text{and} \qquad ASA^{\mathrm{T}} = S.$$

As a consequence of the second identity in the last display, if $\mathcal{L}_\varphi$ is the limiting generator of an underdamped scheme applied to π_φ with S as in (6.13), but with $\hat{J}$ replaced by a φ dependent matrix $\hat{J}_\varphi$ (and J_φ formed from $\hat{J}_\varphi$ as in (6.12)) then

$$\begin{aligned} \mathcal{L}_\varphi[f \circ \varphi] &= \frac{1}{\pi_H \circ \varphi} \mathrm{div}\left[\pi_H \circ \varphi \, \nabla[f \circ \varphi] \, (J_\varphi + S)\right] \\ &= \frac{1}{\pi_H \circ \varphi} \left(\mathrm{div}\left[\pi_H \nabla f (A(J_\varphi \circ \varphi^{-1})A^{\mathrm{T}} + S)\right]\right) \circ \varphi \end{aligned}$$

Now we see that if our choice of $\hat{J}_\varphi$ has the transformation property

$$\hat{J}_\varphi = \hat{A}^{-1} \hat{J}(\hat{\varphi}(\hat{x})) \tag{6.14}$$

(note that square roots of the covariance of $\pi_{\hat{\varphi}}$ and square roots of the inverse Hessian of $-\log \pi_{\hat{\varphi}}$ have this property) then we find that

$$\mathcal{L}_\varphi[f \circ \varphi] = \frac{1}{\pi_H \circ \varphi} \left(\mathrm{div}\left[\pi_H \nabla f (J + S)\right]\right) \circ \varphi = (\mathcal{L}_U f) \circ \varphi$$

or, that the limiting generator of an underdamped Langevin scheme has the affine invariance property (6.8). At least in the small h limit, when applied to the density $\pi_{\hat{\varphi}}$, an underdamped Langevin scheme with S chosen as in (6.13) and an appropriate choice of $\hat{J}$ (e.g. the square root of the covariance matrix or inverse Hessian corresponding to π), results in a chain that is equal in distribution to $\varphi(X_h^{(k)})$ where $X_h^{(k)}$ is the chain generated by the scheme applied to sampling π.

Exercise 72. *Show that with the above choices of J and S, the discrete time underdamped Langevin scheme in* (5.30) *is affine invariant (i.e. not just in the small h limit).*

It is important to keep in mind that we have only shown that our underdamped scheme is invariant to affine transformations of the $\hat{x}$ variables alone. However, the $\tilde{x}$ variables were only introduced as a device to speed the sampling of the $\hat{x}$ variables, and we do not need to worry that their distribution will be badly scaled. So, from a practical point of view, affine invariance with respect to the $\hat{x}$ variables is all that matters.

6.3 Affine invariant ensemble schemes

At this point we have seen that both overdamped and underdamped Langevin schemes can be constructed with the affine invariance property (at least in the small h limit). The schemes introduced so far require access to a matrix, depending on the target density, that transforms appropriately (according to (6.5) and (6.14)) when the target density is composed with an affine transformation. We have learned that such a matrix can be built from the covariance of the target density or from its Hessian (for a log-concave target density). Unforfortunately, in practice we very rarely have access to either of these matrices. The covariance matrix is generally only known after we have successfully sampled the target density, and, if we are fortunate enough to be presented with a log-concave target density, we are very rarely presented with its Hessian matrix. The most obvious response to these realities is to try to provide the scheme with a reasonable approximation to the covariance or Hessian matrices. In fact, one can regard any Metropolis-Hastings scheme with an effective, problem specific, proposal density as a variant of this approach. Of course our ability to design a useful approximation of the covariance or Hessian is highly problem dependent and the resulting schemes will not typically exactly satisfy the affine invariance property.

A simple, very general, and practically useful alternative is to use an affine invariant ensemble MCMC technique, a simple version of which we will present below. Instead of estimating the average of an observable function f against

a target density π using the single trajectory average

$$\langle f \rangle = \overline{f}_N = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(X^{(k)}),$$

an affine invariant ensemble MCMC scheme builds a set of L trajectories $\left\{X^{(k,j)}\right\}_{j=1}^{L}$, each of which samples the target density, and estimates the average via the formula

$$\langle f \rangle = \overline{f}_{L,N} = \lim_{N\to\infty} \frac{1}{LN} \sum_{k=1}^{N} \sum_{j=1}^{L} f(X^{(k,j)}).$$

We will refer to each of the individual processes $X^{(k,j)}$ as a "walker" and use the symbol

$$\vec{X}^{(k)} = \left(X^{(k,1)}, X^{(k,2)}, \ldots, X^{(k,L)}\right)$$

for the vector of all walkers and refer to $\vec{X}^{(k)}$ as the "ensemble." In general, the individual walkers will not be Markov chains; the rule governing their transitions depends on the positions of the other walkers. The ensemble $\vec{X}$ will be a Markov chain. We ensure that each individual walker samples from π by constructing the transition rule for $\vec{X}$ so that it preserves the product density

$$\Pi(\vec{x}) = \pi(x^{(1)})\pi(x^{(2)}) \cdots \pi(x^{(L)})$$

where here $x^{(j)} \in \mathbb{R}^d$ is a dummy variable representing the position of the jth walker. Clearly, the marginal density of $x^{(j)}$ under Π is π and each individual walker will sample from π any time $\vec{X}$ samples from Π. Moreover, though the transitions of each walker will depend on the other walkers, the family of walkers will be independent in the large k limit.

Now we need to construct a transition rule for $\vec{X}$ that results in an affine invariant chain preserving Π. There are many possibilities. Assuming that the chain has generated a sample $\vec{X}^{(k)}$ at step k, a simple and effective choice proceeds to generate $\vec{X}^{(k+1)}$ as follows.

1. `Repeat Steps 2--4 for each walker` i.

2. `Randomly select another walker index` $j \neq i$.

3. Construct a proposal for a new position of the ith walker according to
$$Y^{(k+1,i)} = X^{(k,j)} + Z\left(X^{(k,i)} - X^{(k,j)}\right) \tag{6.15}$$
where Z is an independent draw from some density g on $(0,\infty)$.

4. With probability
$$p_{acc}(X^{(k,i)}, Y^{(k+1,i)}) = \min\left\{1, Z^{d-2}\frac{\pi(Y^{(k+1,i)})g\left(\frac{1}{Z}\right)}{\pi(X^{(k,i)})g(Z)}\right\} \tag{6.16}$$
set $X^{(k+1,i)} = Y^{(k+1,i)}$, otherwise set $X^{(k+1,i)} = X^{(k,i)}$.

The principle of partial resampling tells us that if the move of each individual walker preserves the correct conditional distribution then the transitions will preserve Π. Given that we have selected index j in Step 2 of the scheme, the proposal move for walker i is confined to the ray
$$\left\{y \in \mathbb{R}^d : \, y - X^{(k,j)} = z\left(X^{(k,i)} - X^{(k,j)}\right), \, z > 0\right\}. \tag{6.17}$$
The conditional density for walker i along this ray is proportional to
$$\|y - X^{(k,j)}\|_2^{d-1}\pi(y).$$

Exercise 73. *Check this. Hint: use the co-area formula to write any integral over all of $\mathbb{R}^d$ as an integral over $r \in (0,\infty)$ of the integral over a sphere of radius r.*

Noting that the inverse of the map from Z to proposal $Y^{(k+1,i)}$ in (6.15) is
$$z = \frac{\|y^{(i)} - X^{(k,j)}\|_2}{\|X^{(k,i)} - X^{(k,j)}\|_2}.$$
we find, by change of variables, that the proposal density is
$$q(y^{(i)} \mid X^{(k,i)}, X^{(k,j)}) = \frac{1}{\|X^{(k,i)} - X^{(k,j)}\|_2} g\left(\frac{\|y^{(i)} - X^{(k,j)}\|_2}{\|X^{(k,i)} - X^{(k,j)}\|_2}\right)$$
for $y^{(i)}$ on the ray in (6.17). Combining these observations yields the acceptance criterion in (6.16).

The acceptance rate simplifies somewhat to

$$p_{acc}(X^{(k,i)}, Y^{(k+1,i)}) = \min\left\{1, Z^{d-1}\frac{\pi(Y^{(k+1,i)})}{\pi(X^{(k,i)})}\right\} \tag{6.18}$$

when we assume the symmetry condition

$$g\left(\frac{1}{z}\right) = zg(z).$$

The choice of g that is used most often in practice, and which satisfies the symmetry condition, is

$$g(z) \propto \begin{cases} \frac{1}{\sqrt{z}} & \text{if } z \in \left\{\frac{1}{\alpha}, \alpha\right\}, \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha > 1$ (a common choice is $\alpha = 2$) is a user chosen parameter determining the typical size of the proposal move.

The scheme is invariant under affine transformations of π. For any affine transformation φ of $\mathbb{R}^d$, if we use this ensemble method to sample from

$$\Pi_\varphi(\vec{x}) = \pi(\varphi(x^{(1)})) \cdots \pi(\varphi(x^{(L)})),$$

then the distribution of the resulting chain is also the distribution of $\left(\varphi(X^{(k,1)}), \varphi(X^{(k,2)}), \ldots, \varphi(X^{(k,L)})\right)$ where $\vec{X}$ is the chain generated by the scheme when applied to Π.

Exercise 74. *Show that the ensemble scheme above is affine invariant in the sense just described.*

Because the scheme satisfies this affine invariance property we expect it to perform well on badly scaled densities. But how does one compare the performance of an ensemble scheme to a standard MCMC scheme in practice? To answer this question, first recall that, for a standard MCMC scheme the mean squared error after N_s steps is

$$\mathbf{var}\left[\overline{f}_{N_s}\right] \approx \frac{\mathbf{var}_\pi(f(X))\tau_s}{N_s}$$

where τ_s is the integrated autocorrelation time corresponding to the observable f for the standard MCMC scheme. Now observe that, in terms of the

larger chain $\vec{X}$, we only use the ensemble scheme to compute averages with respect to observables of the form

$$F(\vec{x}) = \frac{1}{L}\sum_{j=1}^{L} f(x^{(j)}).$$

The equilibrium variance of $F(\vec{x})$ under Π is

$$\mathbf{var}_{\Pi}\left[F(\vec{X})\right] = \frac{1}{L}\mathbf{var}_{\pi}\left[f(X)\right]$$

and so the mean squared error in the ensemble schemes estimate of $\langle f \rangle$ with L walkers and N_e iterations is

$$\mathbf{var}\left[\overline{f}_{L,N_e}\right] \approx \frac{\mathbf{var}_{\pi}\left[f(X)\right]\tau_e}{LN_e}$$

where τ_e is the integrated autocorrelation time corresponding to the obserbable F for the ensemble scheme.

Programmed in serial, the cost of the two schemes will be very similar if we set $N_s = LN_e$. In this case, the above formulas reveal that the ensemble scheme will be preferred over a single trajectory MCMC scheme when $\tau_e < \tau_s$. To program the scheme in parallel it is useful to divide the population into two or more sets of walkers and to select j in Step 2 of the algorithm uniformly from the groups *not* containing i. With this structure the updates (and in particular the evaluations of π) in each group can be carried out completely in parallel. However, it is essential to note that any speedup in this scheme is due to affine invariance and *not* due to parallelism. Indeed, we could chose $N_s = N_e$, run L independent single trajectory MCMC simulations completely in parallel, and average the results to produce an estimate of $\langle f \rangle$ with asymptotic variance $\mathbf{var}_{\pi}\left[f(X)\right]\tau_s/(LN_e)$.

Exercise 75. *Use the affine invariant ensemble scheme to sample from the Rosenbrock density. Experiment with different values of the parameters α and L in the ensemble scheme and compare to the results in Exercise 71.*

6.4 Bibliography