Chapter MCMC

Aldous - Fill

January 8, 2001

This book is intended primarily as "theoretical mathematics", focusing on ideas that can be encapsulated in theorems. *Markov Chain Monte Carlo* (MCMC), which has grown explosively since the early 1990s, is in a sense more of an "engineering mathematics" field – a suite of techniques which attempt to solve applied problems, the design of the techniques being based on intuition and physical analogies, and their analysis being based on experimental evaluation. In such a field, the key insights do not correspond well to theorems.

In section 1 we give a verbal overview of the field. Section 2 describes the two basic schemes (Metropolis and line-sampling), and section 3 describes a few of the many more complex chains which have been suggested. The subsequent sections are fragments of theory, indicating places where MCMC interfaces with topics treated elsewhere in this book. Liu [23] gives a comprehensive textbook treatment of the field.

1 Overview of Applied MCMC

1.1 Summary

We give a brisk summary here, and expand upon some main ideas (the **boldface phrases**) in section 1.2.

Abstractly, we start with the following type of problem.

> Given a probability distribution π on a space S, and a numerical quantity associated with π (for instance, the mean $\bar{g} := \sum_x \pi(x)g(x)$ or $:= \int g \, d\pi$, for specified $g : S \to R$), how can one estimate the numerical quantity using Monte Carlo (i.e. randomized algorithm) methods?

Asking such a question implicitly assumes we do not have a solution using mathematical analysis or efficient deterministic numerical methods. *Exact Monte Carlo sampling* presumes the ability to sample exactly from the target distribution π, enabling one to simulate an i.i.d. sequence (X_i) and then use classical statistical estimation, e.g. estimate $\bar{g}$ by $n^{-1}\sum_{i=1}^{n} g(X_i)$. Where implementable, such exact sampling will typically be the best randomized algorithm. For **one-dimensional distributions** and a host of special distributions on higher-dimensional space or combinatorial structures, exact sampling methods have been devised. But it is unrealistic to expect there to be any exact sampling method which is effective in all settings. *Markov Chain Monte Carlo sampling* is based on the following idea.

> First devise a Markov chain on S whose stationary distribution is π. Simulate n steps $X_1, \ldots, X_n$ of the chain. Treat $X_{\tau^*}, X_{\tau^*+1}, \ldots, X_n$ as dependent samples from π (where τ^* is some estimate of some mixing time) and then use these samples in a statistical estimator of the desired numerical quantity, where the confidence interval takes the dependence into account.

Variations of this basic idea include running multiple chains and introducing auxiliary variables (i.e. defining a chain on some product space $S \times A$). The basic scheme and variations are what make up the field of MCMC. Though there is no a priori reason why one must use *reversible* chains, in practice the need to achieve a target distribution π as stationary distribution makes general constructions using reversibility very useful.

MCMC originated in statistical physics, but mathematical analysis of its uses there are too sophisticated for this book, so let us think instead of Bayesian statistics with high-dimensional data as the prototype setting for MCMC. So imagine a point $x \in R^d$ as recording d numerical characteristics of an individual. So data on n individuals is represented as a $n \times d$ matrix $\mathbf{x} = (x_{ij})$. As a model, we first take a parametric family $\phi(\theta, x)$ of probability densities; that is, $\theta \in R^p$ is a p-dimensional parameter and for each θ the function $x \to \phi(\theta, x)$ is a probability density on R^d. Finally, to make a Bayes model we take θ to have some probability density $h(\theta)$ on R^p. So the probability model for the data is: first choose θ according to $h(\cdot)$, then choose $(x_{i\cdot})$ i.i.d. with density $\phi(\theta, x)$. So there is a posterior distribution on θ specified by

$$f_{\mathbf{x}}(\theta) := \frac{h(\theta)\ \prod_{i=1}^{n} \phi(\theta, x_{i\cdot})}{z_{\mathbf{x}}} \tag{1}$$

where $z_{\mathbf{x}}$ is the normalizing constant. Our goal is to sample from $f_{\mathbf{x}}(\cdot)$, for purposes of e.g. estimating posterior means of real-valued parameters. An explicit instance of (1) is the **hierarchical Normal model**, but the general form of (1) exhibits features that circumscribe the type of chains it is feasible to implement in MCMC, as follows.

(i) Though the underlying functions $\phi(\cdot,\cdot), h(\cdot)$ which define the model may be mathematically simple, our target distribution $f_{\mathbf{x}}(\cdot)$ depends on actual numerical data (the data matrix $\mathbf{x}$), so it is hard to predict, and dangerous to assume, global regularity properties of $f_{\mathbf{x}}(\cdot)$.

(ii) The normalizing constant $z_{\mathbf{x}}$ is hard to compute, so we want to define chains which can be implemented without calculating $z_{\mathbf{x}}$.

The wide range of issues arising in MCMC can loosely be classified as "design" or "analysis" issues. Here "design" refers to deciding which chain to simulate, and "analysis" involves the interpretation of results. Let us start by discussing design issues. The most famous general-purpose method is the *Metropolis scheme*, of which the following is a simple implementation in setting (1). Fix a length scale parameter l. Define a step $\theta \to \theta^{(1)}$ of a chain as follows.

> Pick i uniformly from $\{1, 2, \ldots, p\}$.
> Pick U uniformly from $[\theta_i - l, \theta_i + l]$.
> Let θ' be the p-vector obtained from θ by changing the i'th coordinate to U.
> With probability $\min(1, f_{\mathbf{x}}(\theta')/f_{\mathbf{x}}(\theta))$ set $\theta^{(1)} = \theta'$; else set $\theta^{(1)} = \theta$.

The target density enters the definition only via the ratios $f_{\mathbf{x}}(\theta')/f_{\mathbf{x}}(\theta)$, so the value of $z_{\mathbf{x}}$ is not needed. The essence of a Metropolis scheme is that there is a *proposal chain* which *proposes* a move $\theta \to \theta'$, and then an *acceptance/rejection step* which accepts or rejects the proposed move. See section 2.1 for the general definition, and proof that the stationary distribution is indeed the target distribution. There is considerable flexibility in the choice of proposal chain. One might replace the uniform proposal step by a Normal or symmetrized exponential or Cauchy jump; one might instead choose a random (i.e. isotropic) direction and propose to step some random distance in that direction (to make an isotropic Normal step, or a step uniform within a ball, for instance). There is no convincing theory to say which of these choices is better in general. However, in each proposal chain there is some length scale parameter l: there is a trade-off between

making l too small (proposals mostly accepted, but small steps imply slow mixing) and making l too large (proposals rarely accepted), and in section 5 we give some theory (admittedly in an artificial setting) which does give guidance on choice of l.

The other well-known general MCMC method is exemplified by the *Gibbs sampler*. In the setting of (1), for $\theta = (\theta_1, \ldots, \theta_p)$ and $1 \le j \le p$ write

$$f_{\mathbf{x},j,\theta}(v) = f_{\mathbf{x}}(\theta_1, \ldots, \theta_{j-1}, v, \theta_{j+1}, \ldots, \theta_p).$$

A step $\theta \to \theta^{(1)}$ of the Gibbs sampler is defined as follows.

> Pick j uniformly from $\{1, 2, \ldots, p\}$.
> Pick V from the density on R^1 proportional to $f_{\mathbf{x},j,\theta}(v)$.
> Let $\theta^{(1)}$ be θ with its j'th coordinate replaced by V.

The heuristic appeal of the Gibbs sampler, compared to a Metropolis scheme, is that in the latter one typically considers only small proposal moves (lest proposals be almost always rejected) whereas in the Gibbs sampler one samples over an infinite line, which may permit larger moves. The disadvantage is that sampling along the desired one-dimensional line may not be easy to implement (see section 1.2). Closely related to the Gibbs sampler is the *hit-and-run* sampler, where one takes a random (isotropic) direction line instead of a coordinate line; section 2.2 abstracts the properties of such *line samplers*, and section 3 continues this *design* topic to discuss more complex designs of chains which attain a specified target distribution as their stationary distribution.

We now turn to *analysis* issues, and focus on the simplest type of problem, obtaining an estimate for an expectation $\bar{g} = \sum g(\mathbf{x})\pi(\mathbf{x})$ using an irreducible chain (X_t) designed to have stationary distribution π. How do we obtain an estimate, and how accurate is it? The most straightforward approach is *single-run* estimation. The asymptotic variance rate is

$$\sigma^2 := \lim_{t \to \infty} t^{-1} \mathrm{var} \left(\sum_{s=1}^{t} g(X_s) \right) = \sum_{s=-\infty}^{\infty} \mathrm{cov}_\pi(g(X_0), g(X_s)). \qquad (2)$$

So simulate a single run of the chain, from some initial state, for some large number t of steps. Estimate $\bar{g}$ by

$$\hat{g} = \frac{1}{t - t_0} \sum_{i=t_0+1}^{t} g(X_i) \qquad (3)$$

and estimate the variance of $\hat{g}$ by $(t-t_0)^{-1}\hat{\sigma}^2$, and report a confidence interval for $\hat{g}$ by assuming $\hat{g}$ has Normal distribution with mean $\bar{g}$ and the estimated variance. Here $\hat{\sigma}^2$ is an estimate of σ^2 obtained by treating the sample covariances $\hat{\gamma}_s$ (i.e. the covariance of the data-set $(g(X_i), g(X_{i+s})); 0 \le i \le t-s$) as estimators of $\gamma_s = \mathrm{cov}_\pi(g(X_0), g(X_s))$. And the *burn-in* time t_0 is chosen as a time after which the $\hat{\gamma}_s$ become small.

Though the **practical relevance of theoretical mixing time parameters** is debatable, one can say loosely that single-run estimates based on t steps will work fine if t is large compared to the relaxation time τ_2. The difficulty is that in practical MCMC problems we do not know, or have reasonable upper bounds on, τ_2, nor can we estimate τ_2 rigorously from simulations. The difficulty in diagnosing convergence from simulations is the possibility of *metastability error* caused by multimodality. Using **statistical physics imagery**, the region around each mode is a *potential well*, and the stationary distribution conditioned to a potential well is a *metastable distribution*. Believing that a simulation reaches the stationary distribution when in fact it only reaches a metastable distribution is the metastability error.

The simplest way to try to guard against metastability error is the *multiple trials diagnostic*. Here we run k independent copies of the chain from different starting states, each for t steps. One diagnostic is to calculate the k sample averages $\hat{g}_j$, and check that the empirical s.d. of these k averages is consistent with the estimated s.d. $(t-t_0)^{-1/2}\hat{\sigma}$. Intuitively, one chooses the initial states to be "overdispersed", i.e. more spread out than we expect the target distribution to be; passing the diagnostic test gives us some reassurance against metastability error (if there were different potential wells, we hope our runs would find more than one well, and that different behavior of g on different wells would be manifest).

Of course, if one intends to perform such diagnostics it makes sense to start out doing the k multiple runs. A more elaborate procedure is to divide $[0,t]$ into L successive blocks, and seek to check whether the kL blocks "look similar". This can be treated as a classical topic in statistics ("analysis of variance"). In brief, we compute the sample mean $\hat{g}_{i,j}$ and sample variance $\hat{\sigma}^2_{i,j}$ for the j'th block of the i'th simulation, and see if this data (perhaps after deleting the first few blocks of each simulation) is consistent with the blocks being i.i.d.. If so, we use the overall average as an estimator of $\bar{g}$, and estimate the accuracy of this estimator by assuming the blocks were independent.

If a multiple-runs diagnostic fails, or if one lacks confidence in one's

ability to choose a small number of starting points which might be attracted to different nodes (if such existed), then one can seek schemes specially adapted to multimodal target densities. Because it is easy to find *local* maxima of a target density f, e.g. by a deterministic hill-climbing algorithm, one can find modes by repeating such an algorithm from many initial states, to try to find an exhaustive list of modes with relatively high f-values. This is *mode-hunting*; one can then design a chain tailored to jump between the wells with non-vanishing probabilities. Such methods are highly problem-specific; more general methods (such as the multi-level or multi-particle schemes of sections 3.3 and 3.4) seek to automate the search for relevant modes within MCMC instead of having a separate mode-hunting stage.

In seeking theoretical analysis of MCMC one faces an intrinsic difficulty: MCMC is only needed on "hard" problems, but such problems are difficult to study. In comparing effectiveness of different variants of MCMC it is natural to say "forget about theory – just see what works best on real examples". But such experimental evaluation is itself conceptually difficult: **pragmatism is easier in theory than in practice**!

1.2 Further aspects of applied MCMC

Sampling from one-dimensional distributions. Consider a probability distribution μ on R^1 with density function f and and distribution function F. In one sense, sampling from μ is easy, because of the elementary result that $F^{-1}(U)$ has distribution μ, where U is uniform on $[0,1]$ and $x = F^{-1}(u)$ is the inverse function of $u = F(x)$. In cases where we have an explicit formula for F^{-1}, we are done. Many other cases can be done using *rejection sampling*. Suppose there is some other density g from which we can sample by the inverse distribution function method, and suppose we know a bound $c \geq \sup_x f(x)/g(x)$. Then the algorithm

> propose a sample x from $g(\cdot)$;
> accept x with probability $\frac{f(x)}{cg(x)}$; else propose a new sample from g

produces an output with density $f(\cdot)$ after mean c steps. By combining these two methods, libraries of algorithms for often-encountered one-dimensional distributions can be built, and indeed exist in statistical software packages.

But what about a general density $f(x)$? If we need to sample many times from the same density, it is natural to use deterministic numerical methods.

First probe f at many values of x. Then either
(a) build up a numerical approximation to F and thence to F^{-1}; or
(b) choose from a library a suitable density g and use rejection sampling.
The remaining case, which is thus the only "hard" aspect of sampling from one-dimensional distributions, is where we only need one sample from a general distribution. In other words, where we want many samples which are all from different distributions. This is exactly the setting of the Gibbs sampler where the target multidimensional density is complicated, and thus motivates some of the variants we discuss in section 3.

Practical relevance of theoretical mixing time parameters. Standard theory from Chapter 4 (yyy cross-refs) relates τ_2 to the asymptotic variance rate $\sigma^2(g)$ at (2) for the "worst-case" g:

$$\tau_2 = \frac{1}{1-\lambda_2} \approx \frac{1+\lambda_2}{1-\lambda_2} = \sup_g \frac{\sigma^2(g)}{\text{var}_{\ \pi} g}. \tag{4}$$

Moreover Proposition 29 of Chapter 4 (yyy 10/11/94 version) shows that $\sigma^2(g)$ also appears in an upper bound on variances of finite-time averages from the *stationary* chain. So in asking how long to run MCMC simulations, a natural principle (not practical, of course, because we typically don't know τ_2) is

> base estimates on t steps, where t is a reasonable large multiple of τ_2.

But this principle can be attacked from opposite directions. It is sometimes argued that worrying about τ_2 (corresponding to the *worst-case* g) is overly pessimistic in the context of studying some specific g. For instance, Sokal [37] p. 8 remarks that in natural statistical physics models on the infinite lattice near a phase transition in a parameter θ, as θ tends to the critical point the growth exponent of $\sigma^2(g)$ for "interesting" g is typically different from the growth exponent of τ_2. Madras and Slade [25] p. 326 make similar remarks in the context of the pivot algorithm for self-avoiding walk. But we do not know similar examples in the statistical R^d setting. In particular, in the presence of multimodality such counterexamples would require that g be essentially "orthogonal" to the differences between modes, which seems implausible.

Burn-in, the time t_0 excluded from the estimator (3) to avoid undue influence of initial state, is conceptually more problematic. Theory says that taking t_0 as a suitable multiple of τ_1 would guarantee reliable estimates. The

general fact $\tau_1 \geq \tau_2$ then suggests that allowing sufficient burn-in time is a stronger requirement than allowing enough "mixing" for the stationary chain – so the principle above is overly optimistic. On the other hand, because it refers to worst-case initial state, requiring a burn-in time of τ_1 seems far too conservative in practice. The bottom line is that one cannot eliminate the possibility of metastability error; in general, all one gets from multiple-runs and diagnostics is confidence that one is sampling from a single potential well, in the imagery below (though section 6.2 indicates a special setting where we can do better).

Statistical physics imagery. Any probability distribution π can be written as

$$\pi(x) \propto \exp(-H(x)).$$

One can call H a *potential function*; note that a mode (local maximum) of π is a local minimum of H. One can envisage a realization of a Markov chain as a particle moving under the influence of both a potential function (the particle responds to some "force" pushing it towards lower values of H) and random noise. Associated with each local minimum y of H is a *potential well*, which we envisage as the set of points which under the influence of the potential only (without noise) the particle would move to y (in terms of π, states from which a "steepest ascent" path leads to y).

A fundamental intuitive picture is that the main reason why a reversible chain may relax slowly is that there is more than one potential well, and the chain takes a long time to move from one well to another. In such a case, π conditioned to a single potential well will be a *metastable* (i.e. almost-stationary) distribution. One expects the chain's distribution, from any initial state, to reach fairly quickly one (or a mixture) of these metastable distributions, and then the actual relaxation time to stationarity is dominated by the times taken to move between wells. In more detail, if there are w wells then one can consider, as a coarse-grained approximation, a w-state continuous-time chain where the transition rates $w_1 \to w_2$ are the rates of moving from well w_1 to well w_2. Then τ_2 for the original chain should be closely approximated by τ_2 for the coarse-grained chain.

The hierarchical Normal model. As a very simple instance of (1), take $d = 1, p = 2$ and $x \to \phi(\mu, \sigma^2, x)$ the Normal(μ, σ^2) density. Then let (μ, σ) be chosen independently for each individual from some joint density $h(\mu, \sigma)$ on $R \times R^+$. The data is an n-vector $\mathbf{x} = (x_1, \ldots, x_n)$ and the full posterior

distribution is

$$f_{\mathbf{x}}(\mu_1, \ldots, \mu_n, \sigma_1, \ldots, \sigma_n) = z_{\mathbf{x}}^{-1} \prod_{i=1}^{n} h(\mu_i, \sigma_i) \phi(\mu_i, \sigma_i^2, x_i).$$

Typically we are interested in a posterior mean of μ_i for fixed i, that is $\bar{g}$ for

$$g(\mu_1, \ldots, \mu_n, \sigma_1, \ldots, \sigma_n) := \mu_i.$$

Pragmatism is easier in theory than in practice. In comparing MCMC methods experimentally, one obvious issue is the choice of example to study. Another issue is that, if we measure "time" as "number of steps", then a step of one chain may not be comparable with a step of another chain. For instance, a Metropolis step is typically easier to implement than a Gibbs step. More subtlely, in combinatorial examples there may be different ways to set up a data structure to represent the current state in a way that permits easy computation of π-values. The alternative of measuring "time" as CPU time introduces different problems – details of coding matter.

2 The two basic schemes

We will present general definitions and discussion in the context of finite-state chains on a state space S; translating to continuous state space such as R^d involves slightly different notation without any change of substance.

2.1 Metropolis schemes

Write $K = (k_{xy})$ for a *proposal* transition matrix on S. The simplest case is where K is symmetric ($k_{xy} \equiv k_{yx}$). In this case, given π on S we define a step $x \to x'$ of the associated *Metropolis chain* in words by

- pick y from $k(x, \cdot)$ and propose a move to y;
- accept the move (i.e. set $x' = y$) with probability $\min(1, \pi_y/\pi_x)$, otherwise stay ($x' = x$).

This recipe defines the transition matrix P of the Metropolis chain to be

$$p_{xy} = k_{xy} \min(1, \pi_y/\pi_x), \quad y \neq x.$$

Assuming K is irreducible and π strictly positive, then clearly P is irreducible. Then since $\pi_x p_{xy} = k_{xy} \min(\pi_x, \pi_y)$, symmetry of K implies P

satisfies the detailed balance equations and so is reversible with stationary distribution π.

The general case is where K is an arbitrary transition matrix, and the acceptance rule becomes

- accept a proposed move $x \to y$ with probability $\min(1, \frac{\pi_y k_{yx}}{\pi_x k_{xy}})$.

The transition matrix of the Metropolis chain becomes

$$p_{xy} = k_{xy} \min\left(1, \frac{\pi_y k_{yx}}{\pi_x k_{xy}}\right), \quad y \neq x. \tag{5}$$

To ensure irreducibility, we now need to assume connectivity of the graph on S whose edges are the (x, y) such that $\min(k_{xy}, k_{yx}) > 0$. Again detailed balance holds, because

$$\pi_x p_{xy} = \min(\pi_x k_{xy}, \pi_y k_{yx}), \quad y \neq x.$$

The general case is often called *Metropolis-Hastings* – see Notes for terminological comments.

2.2 Line-sampling schemes

The abstract setup described below comes from Diaconis [10]. Think of each S_i as a line, i.e. the set of points in a line.

Suppose we have a collection (S_i) of subsets of state space S, with $\cup_i S_i = S$. Write $I(x) := \{i : x \in S_i\}$. Suppose for each $x \in S$ we are given a probability distribution $i \to w(i, x)$ on $I(x)$, and suppose

$$\text{if } x, y \in S_i \text{ then } w(i, x) = w(i, y). \tag{6}$$

Write $\pi^{[i]}(\cdot) = \pi(\cdot | S_i)$. Define a step $x \to y$ of the *line-sampling chain* in words by

- choose i from $w(\cdot, x)$;
- then choose y from $\pi^{[i]}$.

So the chain has transition matrix

$$p_{xy} = \sum_{i \in I(x)} w(i, x) \pi_y^{[i]}, \quad y \neq x.$$

We can rewrite this as

$$p_{xy} = \sum_{i \in I(x) \cap I(y)} w(i,x)\pi_y / \pi(S_i)$$

and then (6) makes it clear that $\pi_x p_{xy} = \pi_y p_{yx}$. For irreducibility, we need the condition

$$\text{the union over } i \text{ of the edges in the complete graphs}$$
$$\text{on } S_i \text{ form a connected graph on } S. \tag{7}$$

Note in particular we want the S_i to be overlapping, rather than a partition.

This setting includes many examples of random walks on combinatorial sets. For instance, card shuffling by random transpositions (yyy cross-ref) is essentially the case where the collection of subsets consists of all 2-card subsets. In the R^d setting, with target density f, the Gibbs sampler is the case where the collection consists of all lines parallel to some axis. Taking instead all lines in all directions gives the *hit-and-run* sampler, for which a step from x is defined as follows.

- Pick a direction uniformly at random, i.e. a point y on the surface on the unit ball.
- Step from x to $x + Uy$, where $-\infty < U < \infty$ is chosen with density proportional to
$$u^{d-1} f(x + uy).$$

The term u^{d-1} here arises as a Jacobean; see Liu [23] Chapter 8 for explanation and more examples in R^d.

3 Variants of basic MCMC

3.1 Metropolized line sampling

Within the Gibbs or hit-and-run scheme, at each step one needs to sample from a one-dimensional distribution, but a different one-dimensional distribution each time. As mentioned in section 1.2, this is in general not easy to implement efficiently. An alternative is *Metropolized line sampling*, where one instead takes a single step of a Metropolis (i.e. propose/accept) chain with the correct stationary distribution. To say the idea abstractly, in the general "line sampling" setting of section 2.2, assume also:

for each i we have an irreducible transition matrix K^i on S_i whose stationary distribution is $\pi^{[i]}$.

Then define a step $x \to y$ of the Metropolized line sampler as

- choose i from $w(\cdot, x)$;
- then choose y from $k^i(x, \cdot)$.

It is easy to check that the chain has stationary distribution π, and is reversible if the K^i are reversible, so in particular if the K^i are defined by a Metropolis-type propose-accept scheme. In the simplest setting where the line sampler is the Gibbs sampler and we use the same one-dimensional proposal step distribution each time, this scheme is *Metropolis-within-Gibbs*. In that context is seems intuitively natural to use a long-tailed proposal distribution such as the Cauchy distribution. Because we might encounter wildly different one-dimensional target densities, e.g. one density with s.d. $1/10$ and another with two modes separated by 10, and using a $U(-L, L)$ step proposal would be inefficient in the latter case if L is small, and inefficient in the former case if L is large. Intuitively, a long-tailed distribution avoids these worst cases, at the cost of having the acceptance rate be smaller in good cases.

3.2 Multiple-try Metropolis

In the setting (section 2.1) of the Metropolis scheme, one might consider making several draws from the proposal distribution and choosing one of them to be the proposed move. Here is one way, suggested by Liu et al [24], to implement this idea. It turns out that to ensure the stationary distribution is the target distribution π, we need extra samples which are used only to adjust the acceptance probability of the proposed step.

For simplicity, we take the case of a symmetric proposal matrix K. Fix $m \geq 2$. Define a step from x of the *multiple-try Metropolis* (MTM) chain as follows.

- Choose $y_1, \ldots, y_m$ independently from $k(x, \cdot)$;
- Choose y_i with probability proportional to $\pi(y_i)$;
- Choose $x_1, \ldots, x_{m-1}$ independently from $k(y_i, \cdot)$, and set $x_m = x$;
- Accept the proposed move $x \to y_i$ with probability $\min\left(1, \frac{\sum_i \pi(y_i)}{\sum_i \pi(x_i)}\right)$.

Irreducibility follows from irreducibility of K. To check detailed balance, write the acceptance probability as $\min(1, q)$. Then

$$p_{xy} = mk_{xy} \sum \prod_{i=1}^{m-1} k_{x,y_i} \prod_{i=1}^{m-1} k_{y,x_i} \frac{\pi_y}{\sum_i \pi_{y_i}} \min(1, q)$$

where the first sum is over ordered $(2m-2)$-tuples $(y_1, \ldots, y_{m-1}, x_1, \ldots, x_{m-1})$. So we can write

$$\pi_x p_{xy} = mk_{xy}\pi_x\pi_y \sum \prod_{i=1}^{m-1} k_{x,y_i} \prod_{i=1}^{m-1} k_{y,x_i} \min\left(\frac{1}{\sum_i \pi_{y_i}}, \frac{q}{\sum_i \pi_{y_i}}\right).$$

The choice of q makes the final term become $\min(\frac{1}{\sum_i \pi_{y_i}}, \frac{1}{\sum_i \pi_{x_i}})$. One can now check $\pi_x p_{xy} = \pi_y p_{yx}$, by switching the roles of x_j and y_j.

To compare MTM with single-try Metropolis, consider the $m \to \infty$ limit, in which the empirical distribution of $y_1, \ldots, y_m$ will approach $k(x, \cdot)$, and so the distribution of the chosen y_i will approach $k(x, \cdot)\pi(\cdot)/a_x$ for $a_x := \sum_y k_{xy}\pi_y$. Thus for large m the transition matrix of MTM will approximate

$$p^\infty_{xy} = \frac{k_{xy}\pi_y}{a_x} \min(1, a_x/a_y), \quad y \neq x.$$

To compare with single-try Metropolis P, rewrite both as

$$\begin{aligned} p^\infty_{xy} &= k_{xy}\pi_y \min\left(\tfrac{1}{a_x}, \tfrac{1}{a_y}\right), \quad y \neq x \\ p_{xy} &= k_{xy}\pi_y \min\left(\tfrac{1}{\pi_x}, \tfrac{1}{\pi_y}\right), \quad y \neq x. \end{aligned}$$

Thinking of a step of the proposal chain as being in a random direction unrelated to the behavior of π, from a π-typical state x we expect a proposed move to tend to make π decrease, so we expect $a_x < \pi_x$ for π-typical x. In this sense, the equations above show that MTM is an improvement. Of course, if we judge "cost" in terms of the number of evaluations of π_x, then a step of MTM costs $2m-1$ times the cost of single-step Metropolis. By this criterion it seems implausible that MTM would be cheaper than single-step. On the other hand one can envisage settings where there is substantial cost in updating a data structure associated with the current state x, and in such a setting MTM may be more appealing.

3.3 Multilevel sampling

Writing $\pi(x) \propto \exp(-H(x))$, as in the statistical physics imagery (section 1.2), suggests defining a one-parameter family of probability distributions by

$$\pi_\theta(x) \propto \exp(-\theta H(x)).$$

(In the physics analogy, θ corresponds to 1/temperature). If π is multimodal we picture π_θ, as θ increases from 0 to 1, interpolating between the uniform distribution and π by making the potential wells grow deeper. Fix a proposal matrix K, and let P_θ be the transition matrix for the Metropolized chain (5) associated with K and π_θ. Now fix L and values $0 = \theta_1 < \theta_2 < \ldots < \theta_L = 1$. The idea is that for small θ the P_θ-chain should have less difficulty moving between wells; for $\theta = 1$ we get the correct distribution within each well; so by varying θ we can somehow sample accurately from all wells. There are several ways to implement this idea. *Simulated tempering* [26] defines a chain on state space $S \times \{1, \ldots, L\}$, where state (x, i) represents configuration x and parameter θ_i, and where each step is either of the form

- $(x, i) \to (x', i)$; $x \to x'$ a step of P_{θ_i}

or of the form

- $(x, i) \to (x, i')$; where $i \to i'$ is a proposed step of simple random walk on $\{1, 2, \ldots, L\}$.

However, implementing this idea is slightly intricate, because normalizing constants z_θ enter into the desired acceptance probabilities. A more elegant variation is the *multilevel exchange chain* suggested by Geyer [16] and implemented in statistical physics by Hukushima and Nemoto [21]. First consider L independent chains, where the i'th chain $X_t^{(i)}$ has transition matrix P_{θ_i}. Then introduce an interaction; propose to switch configurations $X^{(i)}$ and $X^{(i+1)}$, and accept with the appropriate probability. Precisely, take state space S^L with states $\mathbf{x} = (x_1, \ldots, x_L)$. Fix a (small) number $0 < \alpha < 1$.

- With probability $1 - \alpha$ pick i uniformly from $\{1, \ldots, L\}$, pick x_i' according to $P_{\theta_i}(x_i, \cdot)$ and update $\mathbf{x}$ by changing x_i to x_i'.

- With probability α, pick uniformly an adjacent pair $(i, i+1)$, and propose to update $\mathbf{x}$ by replacing (x_i, x_{i+1}) by (x_{i+1}, x_i). Accept this proposed move with probability

$$\min\left(1, \frac{\pi_{\theta_i}(x_{i+1})\pi_{\theta_{i+1}}(x_i)}{\pi_{\theta_i}(x_i)\pi_{\theta_{i+1}}(x_{i+1})}\right).$$

To check that the product $\pi = \pi_{\theta_1} \times \ldots \times \pi_{\theta_L}$ is indeed a stationary distribution, write the acceptance probability as $\min(1, q)$. If $\mathbf{x}$ and $\mathbf{x}'$ differ only by interchange of (x_i, x_{i+1}) then

$$\frac{\pi(\mathbf{x})}{\pi(\mathbf{x}')}\frac{p(\mathbf{x}, \mathbf{x}')}{p(\mathbf{x}', \mathbf{x})} = \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}')}\frac{\frac{\alpha}{L-1}\min(1, q)}{\frac{\alpha}{L-1}\min(1, q^{-1})} = \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}')}q$$

and the definition of q makes the expression $= 1$. The case of steps where only one component changes is easier to check.

3.4 Multiparticle MCMC

Consider the setting of section 2.2. There is a target distribution π on S and a collection of subsets (S_i). Write $\pi^{[i]} = \pi(\cdot|S_i)$ and $I(x) = \{i : x \in S_i\}$. Now fix $m \geq 2$. We can use the line-sampling scheme of section 2.2 to define (recall Chapter 4 section 6.2) (yyy 10/11/94 version) a product chain on S^m with stationary distribution $\pi \times \pi \times \ldots \times \pi = \pi^k$. For this product chain, picture m particles, at each step picking a random particle and making it move as a step from the line-sampling chain. Now let us introduce an interaction: the line along which a particle moves may depend on the positions of the other particles.

Here is a precise construction. Suppose that for each $(x, \hat{\mathbf{x}}) \in S \times S^{m-1}$ we are given a probability distribution $w(\cdot, x, \hat{\mathbf{x}})$ on $I(x)$ satisfying the following analog of (6):

$$\text{if } x, y \in S_i \text{ then } w(i, x, \hat{\mathbf{x}}) = w(i, y, \hat{\mathbf{x}}). \tag{8}$$

A step of the chain from (x_i) is defined by

- Pick k uniformly from $\{1, 2, \ldots, m\}$
- Pick i from $w(\cdot, x_k, (x_i, i \neq k))$
- Pick x'_k from $\pi^{[i]}(\cdot)$
- Update (X_i) by replacing x_k by x'_k.

It is easy to check that π^m is indeed a stationary distribution; and the chain is irreducible under condition (7). Of course we could, as in section 3.1, use a Metropolis step instead of sampling from $\pi^{[K]}$.

Constructions of this type in statistical applications on R^d go back to Gilks et al [18], under the name *adaptive directional sampling*. In particular

they suggested picking a distinct pair (j,k) of the "particles" and taking the straight line through x_j and x_k as the line to sample x'_k from. Liu et al [24] suggest combining this idea with mode-hunting. Again pick a distinct pair (j,k) of "particles"; but now use some algorithm to find a local maximum $m(x_j)$ of the target density starting from x_j, and sample x'_k from the line through x_k and $m(x_j)$.

4 A little theory

The chains designed for MCMC in previous sections are reversible, and therefore the theory of reversible chains developed in this book is available. Unfortunately there is very little extra to say – in that sense, there is no "theory of MCMC". What follows is rather fragmentary observations.

4.1 Comparison methods

Consider the Metropolis chain

$$p_{xy}^{\text{Metro}} = k_{xy} \min\left(1, \frac{\pi_y k_{yx}}{\pi_x k_{xy}}\right), \quad y \neq x.$$

The requirement that a step of a chain be constructible as a proposal from K followed by acceptance/rejection, is the requirement that $p_{xy} \leq k_{xy}$, $y \neq x$. Recall the asymptotic variance rate

$$\sigma^2(P,f) := \lim_t t^{-1} \text{var} \sum_{s=1}^{t} f(X_s).$$

Lemma 1 (Peskun's Theorem [30]) *Given K and π, let P be a reversible chain with $p_{xy} \leq k_{xy}$, $y \neq x$ and with stationary distribution π. Then $\sigma^2(P,f) \geq \sigma^2(P^{\text{Metro}}, f)\ \forall f$.*

Proof. Reversibility of P implies

$$p_{xy} = \frac{\pi_y p_{yx}}{\pi_x} \leq \frac{\pi_y k_{yx}}{\pi_x} = k_{xy} \frac{\pi_y k_{yx}}{\pi_x k_{xy}}$$

and hence

$$p_{xy} = p_{xy}^{\text{Metro}} \beta_{xy}$$

where $\beta_{xy} = \beta_{yx} \leq 1$, $y \neq x$. So the result follows directly from Peskun's lemma (yyy Lemma 5, to be moved elsewhere). □

This result can be interpreted as saying that the Metropolis rates (5) are the optimal way of implementing a proposal-rejection scheme. Loosely speaking, a similar result holds in any natural Metropolis-like construction of a reversible chain using a $\max(1, \cdot)$ acceptance probability.

It is important to notice that Lemma 1 does not answer the following question, which (except for highly symmetric graphs) seems intractable.

Question. Given a connected graph and a probability distribution π on its vertices, consider the class of reversible chains with stationary distribution π and with transitions only across edges of the graph. Within that class, which chain has smallest relaxation time?

Unfortunately, standard comparison theorems don't take us much further in comparing MCMC methods. To see why, consider Metropolis on R^d with isotropic Normal$(0, \sigma^2 \mathbf{I}_d)$ proposal steps. This has some relaxation time $\tau_2(f, \sigma)$, where f is the target density. For $\sigma_1 < \sigma_2$, the normal densities $g_\sigma(x)$ satisfy $g_{\sigma_2}(x)/g_{\sigma_1}(x) \geq (\sigma_1/\sigma_2)^d$. So the comparison theorem (Chapter 3 Lemma 29) (yyy 9/2/94 version) shows

$$\tau_2(f, \sigma_2) \geq (\sigma_1/\sigma_2)^d \tau_2(f, \sigma_2), \quad \sigma_1 < \sigma_2.$$

But this is no help in determining the optimal σ.

4.2 Metropolis with independent proposals

Though unrealistic in practical settings, the specialization of the Metropolis chain to the case where the proposal chain is i.i.d., that is where $k_{xy} = k_y$, is mathematically a natural object of study. In this setting the transition matrix (5) becomes

$$p_{xy} = k_y \min(1, w_y/w_x), \quad y \neq x$$

where $w_x := \pi_w/k_x$. It turns out there is a simple and sharp coupling analysis, based on the trick of labeling states as $1, 2, \ldots, n$ so that $w_1 \geq w_2 \geq \ldots \geq w_n$ (Liu [22] used this trick to give an eigenvalue analysis, extending part (b) below). Let ρ be the chance that a proposed step from state 1 is rejected (count a proposed step from state 1 to 1 as always accepted). So

$$\rho = \sum_{i=1}^{n} k_i (1 - \tfrac{w_i}{w_1}) < 1.$$

Proposition 2 *For the Metropolis chain over independent proposals, with states ordered as above,*
(a) $\bar{d}(t) \le \rho^t$
(b) The relaxation time $\tau_2 = (1-\rho)^{-1}$.

Proof. For the chain started at state 1, the time T of the first acceptance of a proposed step satisfies

$$P(T > t) = \rho^t.$$

Recall from (yyy Chapter 4-3 section 1; 10/11/99 version) the notion of *coupling*. For this chain a natural coupling is obtained by using the same $U(0,1)$ random variable to implement the accept/reject step (accept if $U < P(\text{accept})$) in two versions of the chain. It is easy to check this coupling (X_t, X'_t) respects the ordering: if $X_0 \le X'_0$ then $X_t \le X'_t$. At time T the fact that a proposed jump from 1 is accepted implies that a jump from any other state must be accepted. So T is a coupling time, and the coupling inequality (yyy Chapter 4-3 section 1.1; 10/11/99 version) implies $\bar{d}(t) \le P(T > t)$. This establishes (a), and the general inequality $\bar{d}(t) = \Omega(\lambda_2^t)$ implies $\lambda_2 \le \rho$. On the other hand, for the chain started at state 1, on $\{T = 1\}$ the time-1 distribution is π; in other words

$$P_1(X_1 \in \cdot) = \rho\delta_1(\cdot) + (1-\rho)\pi(\cdot).$$

But this says that ρ is an eigenvalue of P (corresponding to the eigenvector $\delta_1 - \pi$), establishing (b). □

In the continuous-space setting, with a proposal distribution uniform on $[0,1]$ and target density f with $f^* := \max_x f(x)$, part (b) implies the relaxation time τ_2 equals f^*. So (unsurprisingly) Metropolis-over-independent is comparable to the basic rejection sampling scheme (section 1.2), which gives an exact sample in mean f^* steps.

5 The diffusion heuristic for optimal scaling of high dimensional Metropolis

In any Metropolis scheme for sampling from a target distribution on R^d, there arises the question of how large to take the steps of the proposal chain. One can answer this for isotropic-proposal schemes in high dimensions, in the setting where the target is a product distribution, and the result in this (very artificial) setting provides a heuristic for more realistic settings exemplified by (1).

5.1 Optimal scaling for high-dimensional product distribution sampling

Fix a probability density function f on R^1. For large d consider the i.i.d. product distribution $\pi_f(d\mathbf{x}) = \prod_{i=1}^d f(x_i)\, dx_i$ for $\mathbf{x} = (x_i) \in R^d$. Suppose we want to sample from π_f using Metropolis or Gibbs; what is the optimal scaling (as a function of d) for the step size of the proposal chain, and how does the relaxation time scale?

For the Gibbs sampler this question is straightforward. Consider the one-dimensional case, and take the proposal step increments to be Normal$(0, \sigma^2)$. Then (under technical conditions on f – we omit technical conditions here and in Theorem 3) the Gibbs chain will have some finite relaxation time depending on f and σ, and choosing the optimal σ^* gives a relaxation time $\tau_2(f)$, say. The Gibbs sampler chain in which we choose a random coordinate and propose changing only that coordinate (using the optimal σ^* above) is a product chain in the sense of Chapter 4 section 6.2 (yyy 10/11/94 version), and so the relaxation time of this product chain is $\tau_2^{\text{Gibbs}}(f) = \tau_2(f)\ d$.

Though the argument above is very simple, it is unsatisfactory because there is no simple expression for relaxation time as a function of σ or for the optimal σ^*. It turns out that this difficulty is eliminated in the isotropic-proposal Metropolis chain. In the Gibbs sampler above, the variance of the length of a proposed step is σ^2, so we retain this property by specifying the steps of the proposal chain to have Normal$(0, \sigma^2 d^{-1} \mathbf{I}_d)$ distribution. One expects the relaxation time to grow linearly in d in this setting also. The following result of Roberts et al [34] almost proves this, and has other useful corollaries.

Theorem 3 *Fix $\sigma > 0$. Let $(\mathbf{X}(t), t = 0, 1, 2, \ldots)$ be the Metropolis chain for sampling from product measure π_f on R^d based on a proposal random walk with step distribution Normal$(0, \sigma^2 d^{-1} \mathbf{I}_d)$. Write $X^{(1)}(t)$ for the first coordinate of $\mathbf{X}(t)$, and let $Y_d(t) := X^{(1)}(\lfloor td \rfloor)$ be this coordinate process speeded up by a factor d, for continuous $0 \le t < \infty$. Suppose $\mathbf{X}(0)$ has the stationary distribution π_f. Then*

$$(Y_d(t), 0 \le t < \infty) \stackrel{d}{\to} (Y(t), 0 \le t < \infty) \text{ as } d \to \infty \tag{9}$$

where the limit process is the stationary one-dimensional diffusion

$$dY_t = \theta^{1/2} dW_t + \theta \mu(Y_t) dt \tag{10}$$

for standard Brownian motion W_t, where

$$\begin{aligned}\mu(y) &:= \tfrac{f'(y)}{2f(y)}\\ \theta &:= 2\sigma^2\Phi(-\sigma\kappa/2) \text{ where } \Phi \text{ is the Normal distribution function}\\ \kappa &:= \left(\int \tfrac{(f'(x))^2}{f(x)}\, dx\right)^{1/2}.\end{aligned}$$

Moreover, as $d \to \infty$ the proportion of accepted proposals in the stationary chain tends to $2\Phi(-\sigma\kappa/2)$.

We outline the proof in section 5.3. The result may look complicated, so one piece of background may be helpful. Given a probability distribution on the integers, there is a Metropolis chain for sampling from it based on the simple random walk proposal chain. As a continuous-space analog, given a density f on R^1 there is a "Metropolis diffusion" with stationary density f based on $\theta^{1/2}W_t$ (for arbitrary constant θ) as "proposal diffusion", and this Metropolis diffusion is exactly the diffusion (10): see Notes to (yyy final Chapter).

Thus the appearance of the limit diffusion Y is not unexpected; what is important is the explicit formula for θ in terms of σ and f. Note that the parameter θ affects the process (Y_t) only as a speed parameter. That is, if Y_t^* is the process (10) with $\theta = 1$ then the general process can be represented as $Y_t = Y^*_{\theta t}$. In particular, the relaxation time scales as $\tau_2(Y) = \theta^{-1}\tau_2(Y^*)$. Thus we seek to maximize θ as a function of the underlying step variance σ, and a simple numerical calculation shows this is maximized by taking $\sigma = 2.38/\kappa$, giving $\theta = 1.3/\kappa^2$.

Thus Theorem 3 suggests that for the Metropolis chain $\mathbf{X}$, the optimal variance is $2.38^2\kappa^{-2}d^{-1}\mathbf{I}_d$, and suggests that the relaxation time $\tau_2(f,d)$ scales as

$$\tau_2(f,d) \sim \frac{d\kappa^2}{1.3}\ \tau_2(Y^*). \tag{11}$$

In writing (11) we are pretending that the Metropolis chain is a product chain (so that its relaxation time is the relaxation time of its individual components) and that relaxation time can be passed to the limit in (9). Making a rigorous proof of (11) seems hard.

5.2 The diffusion heuristic.

Continuing the discussion above, Theorem 3 says that the long-run proportion of proposed moves which are accepted is $2\Phi(-\kappa\sigma/2)$. At the optimal

value $\sigma = 2.38/\kappa$ we find this proportion is a "pure number" 0.23, which does not depend on f. To quote [34]

> This result gives rise to the useful heuristic for random walk Metropolis in practice:
>
> *Tune the proposal variance so that the average acceptance rate is roughly* $1/4$.

We call this the *diffusion heuristic* for proposal-step scaling. Intuitively one might hope that the heuristic would be effective for fairly general *unimodel* target densities on R^d, though it clearly has nothing to say about the problem of passage between modes in a multimodal target. Note also that to invoke the diffusion heuristic in a combinatorial setting, where the proposal chain is random walk on a graph, one needs to assume that the target distribution is "smooth" in the sense that $\pi(v)/\pi(w) \approx 1$ for a typical edge (v, w). In this case one can make a Metropolis chain in which the proposal chain jumps σ edges in one step, and seek to optimize σ. See Roberts [33] for some analysis in the context of smooth distributions on the d-cube. However, such smoothness assumptions seem inapplicable to most practical combinatorial MCMC problems.

5.3 Sketch proof of Theorem

Write a typical step of the proposal chain as

$$(x_1, x_2, \ldots, x_d) \to (x_1 + \xi_1, x_2 + \xi_2, \ldots, x_d + \xi_d).$$

Write

$$J = \log \frac{f(x_i + \xi_1)}{f(x_1)}; \quad S = \log \prod_{i=2}^{d} \frac{f(x_i + \xi_i)}{f(x_i)}.$$

The step is accepted with probability $\min(1, \prod_{i=1}^{d} \frac{f(x_i+\xi_i)}{f(x_i)}) = \min(1, e^{J+S})$. So the increment of the first coordinate of the Metropolis chain has mean and mean-square $E\xi_1 \ \min(1, e^{J+S})$ and $E\xi_1^2 \ \min(1, e^{J+S})$. The essential issue in the proof is to show that, for "typical" values of $(x_2, \ldots, x_n)$,

$$E\xi_1 \ \min(1, e^{J+S}) \ \sim \ \theta\mu(x_1)/d \tag{12}$$

$$E\xi_1^2 \ \min(1, e^{J+S}) \ \sim \ \theta/d. \tag{13}$$

This identifies the asymptotic drift and variance rates of $Y_d(t)$ with those of $Y(t)$.

Write $h(u) := E\min(1, e^{u+S})$. Since

$$J \approx \log\left(1 + \tfrac{f'(x_1)}{f(x_1)}\xi_1\right) \approx \tfrac{f'(x_1)}{f(x_1)}\xi_1 = 2\mu(x_1)\xi_1,$$

the desired estimates (12,13) can be rewritten as

$$EJh(J) \sim 2\theta\mu^2(x_1)/d \tag{14}$$
$$EJ^2h(J) \sim 4\theta\mu^2(x_1)/d. \tag{15}$$

Now if J has Normal$(0, \beta^2)$ distribution then for sufficiently regular $h(\cdot)$ we have

$$EJh(J) \sim \beta^2 h'(0); \quad EJ^2h(J) \sim \beta^2 h(0) \text{ as } \beta \to 0.$$

Since J has approximately Normal$(0, 4\mu^2(x_1)\text{var } \xi_1 = 4\mu^2(x_1)\sigma^2/d)$ distribution, proving (14,15) reduces to proving

$$h'(0) \to \frac{\theta}{2\sigma^2} \tag{16}$$
$$h(0) \to \frac{\theta}{\sigma^2}. \tag{17}$$

We shall argue

$$\text{dist}(S) \text{ is approximately Normal}(-\kappa^2\sigma^2/2, \kappa^2\sigma^2). \tag{18}$$

Taking the first two terms in the expansion of $\log(1+u)$ gives

$$\log \tfrac{f(x_i+\xi_i)}{f(x_i)} \approx \tfrac{f'(x_i)}{f(x_i)}\ \xi_i - \tfrac{1}{2}\left(\tfrac{f'(x_i)}{f(x_i)}\right)^2 \xi_i^2.$$

Write $K(\mathbf{x}) = d^{-1}\sum_{i=2}^{d}(\tfrac{f'(x_i)}{f(x_i)})^2$. Summing the previous approximation over i, the first sum on the right has approximately Normal$(0, \sigma^2 K(\mathbf{x}))$ distribution, and (using the weighted law of large numbers) the second term is approximately $-\frac{1}{2}\sigma^2 K(\mathbf{x})$. So the distribution of S is approximately Normal$(-K(\mathbf{x})\sigma^2/2, K(\mathbf{x})\sigma^2)$. But by the law of large numbers, for a typical $\mathbf{x}$ drawn from the product distribution π_f we have $K(\mathbf{x}) \approx \kappa^2$, giving (18).

To argue (17) we pretend S has exactly the Normal distribution at (18). By a standard formula, if S has Normal(α, β^2) distribution then

$$E\max(1, e^S) = \Phi(\alpha/\beta) + e^{\alpha+\beta^2/2}\Phi(-\beta - \alpha/\beta).$$

This leads to

$$h(0) = 2\Phi(-\kappa\sigma/2)$$

which verifies (17). From the definition of $h(u)$ we see

$$h'(0) = Ee^S 1_{(S \le 0)} = h(0) - P(S \ge 0) = h(0) - \Phi(-\kappa\sigma/2) = \Phi(-\kappa\sigma/2)$$

which verifies (16).

6 Other theory

6.1 Sampling from log-concave densities

As mentioned in Chapter 9 section 5.1 (yyy version 9/1/99) there has been intense theoretical study of the problem of sampling uniformly from a convex set in R^d, in the $d \to \infty$ limit. This problem turns out to be essentially equivalent to the problem of sampling from a *log-concave* density f, that is a density of the form $f(x) \propto \exp(-H(x))$ for convex H. The results are not easy to state; see Bubley et al [6] for discussion.

6.2 Combining MCMC with slow exact sampling

Here is a special setting in which one can make rigorous inferences from MCMC without rigorous bounds on mixing times. Suppose we have a guess $\hat{\tau}$ at the relaxation time of a Markov sampler from a traget distribution π; suppose we have some separate method of sampling exactly from π, but where the cost of one exact sample is larger than the cost of $\hat{\tau}$ steps of the Markov sampler. In this setting it is natural to take m exact samples and use them as initial states of m multiple runs of the Markov sampler. It turns out (see [1] for precise statement) that one can obtain confidence intervals for a mean $\bar{g}$ which are always rigorously correct (without assumptions on τ_2) and which, if $\hat{\tau}$ is indeed approximately τ_2, will have optimal length, that is the length which would be implied by this value of τ_2.

7 Notes on Chapter MCMC

Liu [23] provides a nice combination of examples and carefully-described methodology in MCMC, emphasizing statistical applications but also covering some statistical physics. Other statistically-oriented books include [7, 17, 32]. We should reiterate that most MCMC "design" ideas originated

in statistical physics; see the extensive discussion by Sokal [37]. Neal [28] focuses on neural nets but contains useful discussion of MCMC variants.

Section 1.1. In the single-run setting, the variance of sample means (3) could be estimated by classical methods of time series [5].

The phrase *metastability error* is our coinage – though the idea is standard, there seems no standard phrase.

Elaborations of the multiple-runs method are discussed by Gelman and Rubin [14]. The applied literature has paid much attention to diagnostics: for reviews see Cowles and Carlin [8] or Robert [31].

Section 1.2. Devroye [9] gives the classical theory of sampling from one-dimensional and other specific distributions.

Section 2.1. The phrase "Metropolis algorithm" is useful shorthand for "MCMC sampling, where the Markov chain is based on a proposal-acceptance scheme like those in section 2.1". The idea comes from the 1953 paper by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller [27] in the context of statistical physics, and the variant with general proposal matrix is from the 1970 paper of Hastings [20]. Of course the word "algorithm" means a *definite* rule for attaining some goal; the arbitraryness of proposal matrix, and vagueness about when to stop, makes it an extreme stretch to use the word for the Metropolis scheme.

The map $K \to P$ in the Metropolis-Hastings construction (5) has an interpretation as a minimum-length projection in a certain L^1 space of matrices – see Billera and Diaconis [3].

Section 2.2. The Gibbs sampler was popularized in 1984 by Geman and Geman [15] in the context of Bayesian image analysis. The idea is older in statistical physics, under the name *heat bath*. Hit-and-run was introduced in 1984 by Smith [36]. General line-sampling schemes go back to Goodman and Sokal [19].

Section 3.1. Terminology for this type of construction is not standard. What we call "Metropolized line sampling" is what Besag and Greene [2] call an auxiliary variable construction, and this type of construction goes back to Edwards and Sokal [12] in statistical physics.

Section 3.2. One can also define MTM using a general proposal matrix K [24], though (in contrast to Metropolis) the specialization of the general case to the symmetric case is different from the symmetric case described in the text. Liu et al [24] discuss the use of MTM as an ingredient in other variations of MCMC.

Other MCMC variations. In statistical physics, it is natural to think of particles in R^d having position and velocity. This suggests MCMC schemes

in which velocity is introduced as an auxiliary variable. In particular one can use deterministic equations of motion to generate proposal steps for Metropolis, an idea called *hybrid Monte Carlo* – see Neal [29].

Section 4. The survey by Diaconis and Saloff-Coste [11] has further pieces of theory, emphasizing the low-dimensional discrete setting. For target densities on R^d one needs some regularity conditions to ensure τ_2 is finite; see Roberts and Tweedie [35] for results of this type.

Section 4.1. As background to Peskun's theorem, one might think (by vague physical analogy) that it would be desirable to have acceptance probabilities behave as some "smooth" function; e.g. in the symmetric-proposal case, instead of $\min(1, \pi_y/\pi_x)$ take $\frac{\pi_y}{\pi_x+\pi_y}$. Lemma 1 shows this intuition is wrong, at least using asymptotic variance rate or relaxation time as a criterion. Liu [23] section 12.3 gives further instances where Peskun's Theorem can be applied. As usual, it is hard to do such comparison arguments for τ_1.

Section 4.2. The coupling here is an instance of a one-dimensional *monotone coupling*, which exists for any stochastically monotone chain.

Section 5.2. Discussion of practical aspects of the diffusion heuristic can be found in Roberts et al [13], and discussion in the more complicated setting of Gibbs distributions of $(X_v; v \in Z^d)$ is in Breyer and Roberts [4].

References

[1] D.J. Aldous and A. Bandyopadhyay. How to combine fast heuristic Markov chain Monte Carlo with slow exact sampling. Unpublished, 2001.

[2] J. Besag and P.J. Greene. Spatial statistics and Bayesian computation. *J. Royal Statist. Soc. (B)*, 55:25–37, 1993. Followed by discussion.

[3] L.J. Billera and P. Diaconis. A geometric interpretation of the Metropolis algorithm. Unpublished, 2000.

[4] L.A. Breyer and G.O. Roberts. From Metropolis to diffusions: Gibbs states and optimal scaling. Technical report, Statistical Lab., Cambridge U.K., 1998.

[5] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods.* Springer–Verlag, 1987.

[6] R. Bubley, M. Dyer, and M. Jerrum. An elementary analysis of a procedure for sampling points in a convex body. *Random Struct. Alg.*, 12:213–235, 1998.

[7] M.-H. Chen, Q.-M. Shao, and J.G. Ibrahim. *Monte Carlo Methods in Bayesian Computation.* Springer–Verlag, 2000.

[8] M.K. Cowles and B.P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Amer. Statist. Assoc.*, 91:883–904, 1996.

[9] L. Devroye. *Nonuniform Random Number Generation.* Springer–Verlag, 1986.

[10] P. Diaconis. Notes on the hit-and-run algorithm. Unpublished, 1996.

[11] P. Diaconis and L. Saloff-Coste. What do we know about the Metropolis algorithm? *J. Comput. System Sci.*, 57:20–36, 1998.

[12] R.G. Edwards and A.D. Sokal. Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Phys. Rev. D*, 38:2009–2012, 1988.

[13] A. Gelman, G.O. Roberts, and W.R. Gilks. Efficient Metropolis jumping rules. In *Bayesian Statistics*, volume 5, pages 599–608. Oxford University Press, 1996.

[14] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992. With discussion.

[15] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741, 1984.

[16] C.J. Geyer. Markov chain Monte Carlo maximum likelihood. In E. Keramigas, editor, *Computing Science and Statistics. Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, 1991.

[17] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*, London, 1996. Chapman and Hall.

[18] W.R. Gilks, G.O. Roberts, and E.I. George. Adaptive direction sampling. *The Statistician*, 43:179–189, 1994.

[19] J. Goodman and A. Sokal. Multigrid Monte Carlo method: Conceptual foundations. *Phys. Rev. D*, 40:2037–2071, 1989.

[20] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[21] K. Hukushima and K. Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *J. Physics Soc. Japan*, 65:1604–1608, 1996.

[22] J.S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6:113–119, 1996.

[23] J.S. Liu. *Monte Carlo Techniques in Scientific Computing.* xxx, 2001. To appear.

[24] J.S. Liu, F. Liang, and W.H. Wong. The use of multiple-try method and local optimization in Metropolis sampling. *JASA*, xxx:xxx, xxx.

[25] N. Madras and G. Slade. *The Self-Avoiding Walk*. Birkhauser, 1993.

[26] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19:451–458, 1992.

[27] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *J. Chem. Phys.*, 21:1087–1091, 1953.

[28] R.M. Neal. *Bayesian Learning for Neural Networks*. Number 118 in Lecture Notes in Statistics. Springer–Verlag, 1996.

[29] R.M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6:353–366, 1996.

[30] P. Peskun. Optimal Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

[31] C.P. Robert, editor. *Discretization and MCMC Convergence Assesment*. Number 135 in Lecture Notes in Statistics. Springer–Verlag, 1998.

[32] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer–Verlag, 2000.

[33] G.O. Roberts. Optimal Metropolis algorithms for product measures on the vertices of a hypercube. *Stochastics Stochastic Rep.*, 62:275–283, 1998.

[34] G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7:110–120, 1997.

[35] G.O. Roberts and R.L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110, 1996.

[36] R.L. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32:1296–1308, 1984.

[37] A.D. Sokal. Monte Carlo methods in statistical mechanics: Foundations and new algorithms. In *Cours de Troisieme Cycle de la Physique en Suisse Romande, Lausanne*, 1989.

8 Belongs in other chapters

yyy: add to what's currently sec. 10.2 of Chapter 2, version 9/10/99, but which may get moved to the new Chapter 8.

Where π does not vary with the parameter α we get a simple expression for $\frac{d}{d\alpha}\mathbf{Z}$.

Lemma 4 *In the setting of (yyy Chapter 2 Lemma 37), suppose π does not depend on α. Then*

$$\tfrac{d}{d\alpha}\mathbf{Z} = \mathbf{ZRZ}.$$

xxx JF: I see this from the series expansion for $\mathbf{Z}$ – what to do about a proof, I delegate to you!

8.1 Pointwise ordered transition matrices

yyy: belongs somewhere in Chapter 3.

Recall from Chapter 2 section 3 (yyy 9/10/99 version) that for a function $f : S \to R$ with $\sum_i \pi_i f_i = 0$, the asymptotic variance rate is

$$\sigma^2(\mathbf{P}, f) := \lim_t t^{-1} \mathrm{var} \sum_{s=1}^{t} f(X_s) = f\Gamma f \tag{19}$$

where $\Gamma_{ij} = \pi_i Z_{ij} + \pi_j Z_{ji} + \pi_i \pi_j - \pi_i \delta_{ij}$. These individual-function variance rates can be compared between chains with the same stationary distribution, under a very strong "coordinatewise ordering" of transition matrices.

Lemma 5 (Peskun's Lemma [30]) *Let $\mathbf{P}$ and $\mathbf{Q}$ be reversible with the same stationary distribution π. Suppose $p_{ij} \le q_{ij}$ $\forall j \ne i$. Then $\sigma^2(\mathbf{P}, f) \ge \sigma^2(\mathbf{Q}, f)$ for all f with $\sum_i \pi_i f_i = 0$.*

Proof. Introduce a parameter $0 \le \alpha \le 1$ and write $\mathbf{P}^\alpha = (1-\alpha)\mathbf{P} + \alpha\mathbf{Q}$. Write $(\cdot)'$ for $\frac{d}{d\alpha}(\cdot)$ at $\alpha = 0$. It is enough to show

$$(\sigma^2(\mathbf{P}, f))' \le 0.$$

By (19)

$$(\sigma^2(\mathbf{P}, f))' = f\Gamma' f = 2 \sum_i \sum_j f_i \pi_i z'_{ij} f_j.$$

By (yyy Lemma 4 above) $\mathbf{Z}' = \mathbf{ZP'Z}$. By setting

$$g_i = \pi_i f_i; \quad a_{ij} = z_{ij}/\pi_j; \quad w_{ij} = \pi_i p_{ij}$$

we can rewrite the equality above as

$$(\sigma^2(\mathbf{P}, f))' = 2\ g\mathbf{A}\mathbf{W}'\mathbf{A}g.$$

Since $\mathbf{A}$ is symmetric wsith row-sums equal to zero, it is enough to show that $\mathbf{W}'$ is non-negative definite. By hypothesis $\mathbf{W}'$ is symmetric and $w'_{ij} \geq 0$ for $j \neq i$. These properties imply that, ordering states arbitrarily, we may write

$$\mathbf{W}' = \sum \sum_{i<j} w'_{ij}\mathbf{M}^{ij}$$

where $\mathbf{M}^{ij}$ is the matrix whose only non-zero entries are $m(i,i) = m(j,j) = -1$; $m(i,j) = m(j,i) = 1$. Plainly $\mathbf{M}^{ij}$ is non-negative definite, hence so is $\mathbf{W}'$.