Lecture 0: Markov Chain Monte Carlo Methods

Mandatory readings: Andrieu et al. (2003)

0.0 Introduction: Objective Bayes Approach

This lecture develops Markov Chain Monte Carlo (MCMC) methods as supplementary inference and optimization tools for econometricians, may they be Bayesian of frequentist. This ecumenical conceptualization of the MCMC toolbox is often referred to as the *objective Bayes approach*. To be sure, the objective Bayes approach was originally associated with the search for uniformative priors, but is today (and for our purposes) tantamount to using MCMC methods for computational reasons, and without any particular desire to impose prior knowledge. See Efron and Hastie (2016; p.233 and 251) for a brief discussion of the matter.

The remainder of this lecture is divided as follows. In the first subsection we motivate Monte Carlo sampling, we construct the Metropolis-Hastings sampler "from scratch" in an attempt to best convey the intuition for its workings, and we present other basic Monte Carlo Samplers. In subsection 2, we present the basic theory of Markov chains, so to be able to treat the samplers as mathematical and probabilistic objects. In subsection 3, we investigate more advanced and modern Monte Carlo methods. In subsection 4, we develop more advanced topics in the theory of Markov chains, with a particular focus on mixing times.

0.1 Motivational Problems and Intuitive Solutions

We are interested in characterizing parameter estimates from their likelihood or posterior distributions. This may result in a computationally challenging task because the likelihood can be computationally intractable. Indeed, it is common in practice that

- the likelihood is intractable because the estimation of the normalizing constant requires inordinate computations,

and it likewise occurs that

- computation of the likelihood is intractable, even up to a proportionality constant.

The task at hand may be optimization (point estimation), integration, or inference. We will focus particularly on the latter. In a typical situation, we have a posterior distribution f, and we would like to sample from it. We may also think of sampling directly from the

likelihood if the prior is uniform[1].

There are many ways to sample from a distribution. For instance, denoting by F the cumulative distribution function (cdf) of f, and by X a random variable drawn from that law, we know that $X \sim F^{-1}(U)$, where U is distributed uniformly on the unit interval[2]. This intuition can be carried out in practice, even in the absence of an explicit form for F^{-1}, and constitutes the *inverse sampling transform* method. This approach is generally considered slow. An alternative approach is to find a clever change of variables from random variables we can easily sample directly to the variables we wish to sample from. MCMC methods avoid this need for case-by-case cleverness.

✹

We proceed constructively and begin with a simple yet non-trivial case. We need to sample from the discrete support Θ according to $p(\theta|\mathbf{x})$, but do not know *a priori* how to simulate from $p(\cdot|\mathbf{x})$. A naive approach is then to go through the support, one point at a time, for multiple rounds, and record (accept) each point θ with probability $p(\theta|\mathbf{x})$.

Example 1 *Suppose you have $\mathbf{x}_n = \{x_1, ..., x_{40}\}$ draws from a normal $N(\theta, 1)$, and it is known that $\theta \in \{1, 2, 3, 4\}$. Take a deterministic walk and for $i = 1, 2, 3, 4, 1, 2, 3, ...$ and sample i if $u \leq p(i|\mathbf{x}_n) := \phi(\mathbf{x}_n|\theta = i)/\sum_{j=1}^{4} \phi(\mathbf{x}_n|\theta = j)$. Figure 1 (left) plots 1000 sampled values.*

Another approach, instead of sampling according to a deterministic walk over the support, would be to sample according to a simple random walk.

Example 2 *Suppose you have $\{x_1, ..., x_{40}\}$ draws from a normal $N(\theta, 1)$, and it is known that $\theta \in \{1, 2, 3, 4\}$. Take a random walk and for $\theta_1 = 1$, $\theta_i = \theta_{i-1} + Bin\{-1, +1\}$ mod 4, $i = 2, 3, ...$, and sample θ_i if $u \leq p(i|\mathbf{x}_n) := \phi(\mathbf{x}_n|\theta = i)/\sum_{j=1}^{4} \phi(\mathbf{x}_n|\theta = j)$. Figure 1 (right) plots 1000 sampled values.*

[1] See section notes for a clarification of the connection between the –frequentist– sampling of $\arg\max_\theta f(\theta|\mathbf{x})$, when $\mathbf{x}$ is resampled, and the –Bayesian– sampling according to $\pi(\theta) \propto f(\theta|\mathbf{x})$. The asymptotic coincidence of both distributions is established by the Bernstein von Mises theorem.

[2] We will assume throughout that $U[0, 1]$ random variables are available to the researcher.

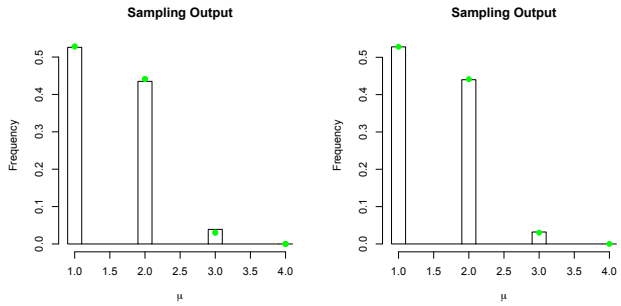

Figure 1 *Posterior distribution (green) and output (histogram) from the samplers described in examples 1 (left) and 2 (right).*

We isolated a first important idea; we can move around the parameter space Θ and sample the parameter θ in proportion to its posterior distribution.

Why is this candid approach insufficient? First, example 1 only addresses the problem for discrete spaces. Second, we can often only evaluate p up to proportionality, because it is the normalizing constant which is the computational bottleneck. Third, intuitively, an effective method should spend less "time" in regions of Θ where the probability law takes lower values. Thinking of the path across Θ as a chain, this suggests that p should inform the movement of the chain such that it spends more time where p has higher value.

This immediately suggests the construction of the Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm will be a unifying principle in the analysis of different MCMC methods.

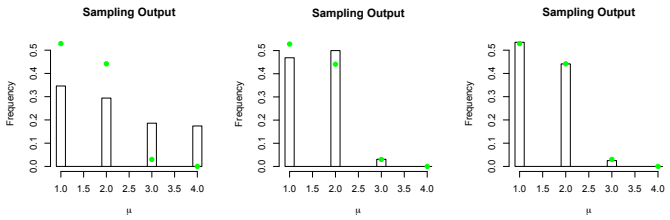

Figure 2 *Posterior distribution (green) and output (histogram) from the samplers a (left), b (center), c (right) described in example 3.*

We need to find a way to get the random walk to "stick around" longer in regions of higher probability density. What could do it, and still sample p correctly?

Example 3 *Consider the following chains for sampling p.*

a. A simple random walk which stays at θ with probability $p(\theta)$, and moves one to

the right or left both with probability $(1 - p(\theta))/2$.

b. A simple random walk which, at θ*, moves to* $\theta - 1$ *with probability* $p(\theta - 1)/\left(p(\theta - 1) + p(\theta + 1)\right)$ *and moves to* $\theta+1$ *with probability* $p(\theta+1)/\left(p(\theta - 1) + p(\theta + 1)\right)$.

c. A random walk which, at θ*, receives a proposal* $\theta' \in \{\theta - 1, \theta + 1\}$ *from a simple random walk and accepts it with probability* $p(\theta)/p(\theta')$*. If it accepts, the chain moves to* θ'*, if it rejects, it stays at* θ *for one more step.*

Only the last intuitively plausible approach works, and that is an implementation of the Metropolis-Hastings (MH) algorithm. See Figure 2.

Once we have acquired more analytical tools, we will be able to prove that 1 and 2 do not sample from p.

Algorithm 1 Metropolis-Hastings

1: **for** $i = 0$ to $N - 1$ **do**
2: Sample $u \sim U[0, 1]$
3: Sample $\theta^* \sim q(\theta^*|\theta^{(i)})$
4: If $u < \frac{p(\theta^*)q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)})q(\theta^*|\theta^{(i)})}$
5: Then $\theta^{(i+1)} = \theta^*$
6: Else $\theta^{(i+1)} = \theta^{(i)}$
7: **end for**

We say that we "propose" a parameter value θ^* according to the **proposal distribution** q and that it is "accepted" with probability given by the **acceptance probability**

$$\alpha = \min\left\{\frac{p(\theta^*)q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)})q(\theta^*|\theta^{(i)})}, 1\right\}. \tag{1}$$

We call the fraction of accepted proposals the **acceptance rate**.

Observe that the target distribution p enters only through the ratio $p(\theta^*)/p(\theta^{(i)})$, meaning its normalization constant cancels out and need not be computed. In the case of a likelihood $f(\theta|\mathbf{x})$ and uniform prior, the target posterior density is $p(\theta) = f(\theta|\mathbf{x})/\int f(\theta|\mathbf{x})d\theta$ and

$$\frac{p(\theta^*)}{p(\theta^{(i)})} = \frac{f(\theta^*|\mathbf{x})/\int f(\theta|\mathbf{x})d\theta}{f(\theta^{(i)}|\mathbf{x})/\int f(\theta|\mathbf{x})d\theta} = \frac{f(\theta^*|\mathbf{x})}{f(\theta^{(i)}|\mathbf{x})}.$$

This is a major computational advantage of Metropolis-Hastings sampling.

The acceptance probability is a key intuitive piece of the algorithm. Suppose for now that $q = U[\Theta]$, the uniform distribution over the support of θ. Then the acceptance probability is simply $\alpha = \min\left\{p(\theta^*)/p(\theta^{(i)}), 1\right\}$. That is, when proposed a new value θ^*, we draw it or draw again the previous sample $\theta^{(i)}$ with probability proportional to the

ratio of their posterior evaluations. This should make us confident that we are drawing from p; in subsection 2 we will be able to prove that this is indeed the case.

In general, the proposal distribution q is not uniform. Note how its evaluation appears in the ratio that is the inverse of that in p; this is because the ratio $q(\theta^{(i)}|\theta^*)/q(\theta^*|\theta^{(i)})$ *corrects* for the frequency of proposal, i.e., it decreases/increases the acceptance probability of values which are overproposed/underproposed.

A key element in the implementation of the Metropolis-Hastings algorithm is the design of a proposal distribution q. A proposal distribution resembling p will yield a high acceptance rate of draws with low serial correlation –hence carrying more information.

Exercise 1 *Suppose that $q = p$, show that the acceptance rate of the Metropolis-Hastings sampler is 1.*

It is sometimes the case that evaluating q is computationally costly. A popular trick is then to use a symmetric proposal, i.e., $q(\theta|\theta') = q(\theta'|\theta)$, $\forall\ \theta, \theta' \in \Theta$. In that case, even when the proposal is not uniform, it vanishes from the acceptance proposal,

$$\frac{p(\theta^*)q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)})q(\theta^*|\theta^{(i)})} = \frac{p(\theta^*)}{p(\theta^{(i)})}.$$

In that case, Metropolis-Hastings requires the computational effort of simulating from q, but never of evaluating q.

We will see in the next section that it is very easy to prove that Metropolis-Hastings samples from p, but that it has this property, almost by construction, due to a a sufficient but not necessary condition called detailed balance or reversibility.

✹

Before moving on, we should mention two popular "competitors" to Markov Chain Monte Carlo techniques: importance sampling and rejection sampling. They are sometimes "collaborators" of Monte Carlo methods, as we will see below with the case of pseudo-marginal MCMC, where importance sampling may be used within a Metropolis-Hastings algorithm.

The straightforward approach of **rejection sampling** relies on the idea that if we know the distribution p we want to sample from (at least up to a proportionality constant) as well as some "envelope" density q satisfying $p \leq Mq$ for some $M < \infty$ and *from which it is easy to sample*, then we may sample from p by drawing some θ^* from q and accepting the draw as a sample from p if it is relatively – to its probability of being proposed– likely, that is if $p(\theta^*)/q(\theta^*)$ is high. Specifically, Algorithm 2 produces a rejection sampler.

Algorithm 2 Rejection Sampling

1: Initialise $i = 1$
2: **while** $i \leq N$ **do**
3: Sample $\theta^i \sim q(\theta)$ and $u \sim U[0,1]$
4: If $u < \frac{p(\theta^*)}{Mq(\theta^{(i)})}$ then accept $\theta^{(i)}$ and increment the counter i by one. Otherwise, reject.
5: **end while**

Another very useful tool is **importance sampling**, which is taylored to the problem of approximating integrals. The MCMC strategy for approximating integrals is to do "sampling instead of summing", i.e.,

$$\int f(\theta)p(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i),\ \theta_i \overset{iid}{\sim} p,$$

for N large. It may be that p is difficult to sample from, but we have access to some easy-to-sample q which does not vanish on the support of p. Then the approximation

$$\int f(\theta)p(\theta)d\theta \equiv \int f(\theta)\omega(\theta)q(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)\omega(\theta_i),\ \theta_i \overset{iid}{\sim} q,$$

where $\omega(\theta) = p(\theta)/q(\theta)$, may be much more tractable.

Remark that we do not cover adaptive rejection sampling, sampling importance resampling or adaptive importance sampling, although they are well established generalizations.

0.2 Basic Theory for MCMC Methods

In this subsection, we model the samplers as probabilistic objects, called Markov chains. We will find that two properties often satisfied by Markov chains used to model samplers of practical interest deliver an array of powerful results. These two properties are irreducibility (every state is eventually visited with some probability) and aperiodicity (there are no cycles). With these, we will be able to establish the existence and uniqueness of a target distribution, as well as the convergence of the Markov chain to that distribution. This is of practical importance because we want to build samplers which will sample from a distribution of interest and, equipped with the results herein, we will be able to obtain guarantees that if we design a transition probability with respect to which the distribution of interest is a natural target (i.e., with respect to which it is "stationary", see below), then the Markov chain will eventually sample from that distribution, regardless of its

starting distribution.

In subsection 0.1, we built samplers using common sense and seat-of-the-pants intuition. We would like to be able to treat these samplers analytically, so to study and develop their properties. The key object to study is the random sequence of sampled values, which we call a chain. In fact it is a special kind of chain. Suppose we built our chain over the elements of χ by picking our next point, moving from $x \in \chi$, according to the transition probability $P(\cdot, x)$. Likewise, the probability of moving from x to y in t steps will be denoted $P^t(y, x)$.

Definition 1 *A sequence of random variables $\{X_0, X_1, ...\}$ is a **Markov chain with state space χ and transition matrix** P if for all $x, y \in \chi$, all $t \geq 1$, and all events $H_{t-1} = \bigcap_{s=0}^{t-1} \{X_s = x_s\}$ satisfying $P\left(H_{t-1} = \bigcap_{s=0}^{t-1} \{X_s = x_s\}\right) > 0$, we have*

$$P(X_{t+1} = y | H_{t-1} \cap \{X_t = x\}) = P(X_{t+1} = y \,|\{X_t = x\}\,).$$

The equation in the display above is referred to as the **Markov property**, it says that conditional on the immediate past, more ancient realizations are not informative.

We now define two simple and powerful assumptions pertaining to Markov chains which are often satisfied in practice.

Definition 2 *A Markov chain with transition matrix P is called **irreducible** if for any two states $x, y \in \chi$ there exists an integer t such that $P^t(x, y) > 0$.*
*Let $\mathcal{T}(x) = \{t \geq 1 \ : \ P^t(x, x) > 0\}$ be the set of times when it is possible for the chain to return to its starting position x. The period of state x is defined to be the greatest common divisor of $\mathcal{T}(x)$. The chain is called **aperiodic** if all states have period 1.*

The dynamics of the chain are best articulated in matrix form. For any distribution over χ at time t, say μ_t, we obtain the distribution one step ahead in the chain by applying the transition matrix to the right,

$$\mu_{k+1} = \mu_k P.$$

Note that we can iterate the above and obtain $\mu_{k+t} = \mu_k P^t$.

Exercise 2 *Give the interpretation "in English" of $\mu_t P$ and $P\mu_t$.*

One of the big results ahead is that under minimal conditions on P, for any given initial distribution μ_0, the distribution $\mu_0 P^k$ will stabilize at some limiting (in k) distribution, hence delivering a sampler for that distribution. This distribution will be called the

stationary distribution, and our first order of business is to define it, show that it exists and is unique, and show that it indeed obtains as the limit in t of $\mu_0 P^t$. We will find that the long-term fraction of time the chain spends in each state coincides with its stationary distribution.

Definition 3 *A probability distribution π satisfying*

$$\pi = \pi P$$

is called a ***stationary distribution*** *of the Markov chain defined by P.*

Are we guaranteed that such a distribution even exists? There is a guarantee and in fact, with a little bit of work, we can even provide a constructive proof. First, we need to define a new object which will be at the heart of the construction.

Definition 4 *For any $x \in \chi$, the* ***hitting time*** *for x is*

$$\tau_x := \min \{t \geq 0 \ : \ X_t = x\},$$

the first time at which the chain visits state x. Likewise, the ***first return time*** *for x is*

$$\tau_x^+ := \min \{t \geq 1 \ : \ X_t = x\}.$$

We can intuit the candidate distribution before beginning construction. If we expect that the time to return to z from z, i.e., $E_z[\tau_z^+]$, is relatively long then it must be that the chain is spending proportionally less time in state z. Hence a good candidate for the stationary distribution evaluated at z would be commensurate to $(E_z[\tau_z^+])^{-1}$. In fact, it will verify that $\pi(z) = (E_z[\tau_z^+])^{-1}$ integrates to 1.

The construction of the candidate distribution will rely on a related but different observation. Take an arbitrary $z \in \chi$ and consider the sojourn of the chain from z to itself. That is, since visits to z break up the trajectory of the chain into identically distributed segments, it should not be surprising that the average fraction of time per segment spent in each state y coincides with the long-term fraction of time spent in y (Levin and Peres, 2017).

To construct the candidate this way, we need to following object. Define

$$\begin{aligned}\tilde{\pi}(y) &:= E_z\left[\text{number of visits to } y \text{ before to } z\right] \\ &= \sum_{t=0}^{\infty} E_z\left[\mathbf{1}\left\{X_t = y,\ \tau_z^+ > t\right\}\right] \\ &= \sum_{t=0}^{\infty} P_z\left(X_t = y,\ \tau_z^+ > t\right).\end{aligned}$$

Theorem 1: Existence of Stationary Distribution *Let $\tilde{\pi}$ be the measure on χ defined above.*

(i) If $P_z\left(\tau_z^+ < \infty\right) = 1$, then $\tilde{\pi}$ satisfies $\tilde{\pi}P = \tilde{\pi}$.

(ii) If $E_z\left[\tau_z^+\right] < \infty$, then $\pi := \frac{\tilde{\pi}}{E_z\left[\tau_z^+\right]}$ is a stationary distribution.

Proof Sketch

The irreducibility of the chain implies that expected time of return is bounded, which in turn implies boundedness of $\tilde{\pi}$, i.e., $\tilde{\pi}(z) \leq E_z\left[\tau_z^+\right] < \infty$. Remains to show stationarity:

$$\begin{aligned}\sum_{x\in\chi} \tilde{\pi}(x)P(x,y) &= \sum_{x\in\chi}\sum_{t=0}^{\infty} P_z\left(X_t = x,\ \tau_z^+ > t\right) P(x,y) \\ &= \sum_{x\in\chi}\sum_{t=0}^{\infty} P_z\left(X_{t+1} = y,\ X_t = x,\ \tau_z^+ > t\right) \\ &= \sum_{t=0}^{\infty} P_z\left(X_{t+1} = y,\ \tau_z^+ > t\right) = \sum_{t=1}^{\infty} P_z\left(X_t = y,\ \tau_z^+ \geq t\right) \\ &= \tilde{\pi}(y) - P_z\left(X_0 = y,\ \tau_z^+ > 0\right) + \sum_{t=1}^{\infty} P_z\left(X_t = y,\ \tau_z^+ = t\right) \\ &= \tilde{\pi}(y) - P_z\left(X_0 = y\right) + P_z\left(X_{\tau_z^+} = y\right) = \tilde{\pi}(y),\end{aligned}$$

where the last equality follows from case analysis. If $y = z$, then $P_z\left(X_0 = y\right) = 1 = P_z\left(X_{\tau_z^+} = y\right)$, and $y \neq z$, $P_z\left(X_0 = y\right) = 0 = P_z\left(X_{\tau_z^+} = y\right)$. In either case, the stationarity is established.

□

Exercise 3 *Show that $E_z\left[\tau_z^+\right] = \sum_{x\in\chi} \tilde{\pi}(x)$. hint: $\tau_z^+ > 0$ so you can use a special expression for the expectation formula.*

We now know that some stationary distribution exists. This is crucial for practice as we will be designing algorithms targeting the stationary distribution. In practice, we will have a distribution μ we want to sample from, and we will be able to design a P such

that μ is a stationary distribution of P. It is however essential, as we will be sampling from *some* stationary distribution of P, that this distribution is unique.

Theorem 2: Uniqueness of Stationary Distribution *If P is an irreducible transition matrix then it has a unique probability distribution π solving $\pi P = \pi$, and for all z, $\pi(z) = \frac{1}{E_z\left[\tau_z^+\right]}$.*

Proof Sketch

It is straightforward to show that $P - I$ has kernel rank 1 (exercise 14). Therefore, the probability distribution satisfying $\pi P = \pi$, since it is normalized to sum to 1, is unique.

Index $\tilde{\pi}$ by the choice of starting state, and observe that for for all $z \in \chi$,

$$\tilde{\pi}_z(z) := \frac{\tilde{\pi}(z)}{E_z\left[\tau_z^+\right]} = \frac{E_z\left[\text{number of visits to } y \text{ before to } z\right]}{E_z\left[\tau_z^+\right]} = \frac{1}{E_z\left[\tau_z^+\right]}.$$

Since all $\tilde{\pi}_z, \tilde{\pi}_{z'}$ are stationary, and the stationary distribution is unique, it must be that the stationary distribution is $\pi(z) = \frac{1}{E_z\left[\tau_z^+\right]},\ z \in \chi$.

□

We have established existence and uniqueness of the stationary distribution, but that is useless in practice if we can't "get to it". Convergence results guarantee that running the Markov chain long enough insures we will eventually be sampling from its stationary distribution, i.e., that X_t for t large enough will be a draw from π, the stationary distribution of P.

Convergence is established rigorously in subsection 0.4, as it is best studied in conjunction with mixing times. It would, however, be in poor taste to cultivate the suspense until then, and we should convince ourselves of the plausibility of convergence before moving on.

The transition probability P has biggest left eigenvalue equal to 1. See Lemma 1 (biggest right eigenvalue is 1) and 8 (equality of the left and right spectrum). It is well established in practice and in theory that the power method (under regularity conditions) produces the largest eigenvalue and corresponding eigenvector. It should therefore be no surprise that many steps of the Markov chain, thought of as repeated applications of the matrix P, converges to the stationary distribution,

$$\mu P^t \to \pi.$$

In fact, this is close to the intuition underpinning the convergence proof of Theorem 3.

It is now time to put the "Markov Chain" in Markov Chain Monte Carlo. We derive the transition probability P_{MH} of Metropolis-Hastings. Redefine the acceptance probability as

$$\alpha(\theta^{(i)}, \theta^{(i+1)}) = \min\left\{1, \frac{p(\theta^*)q(\theta^{(i)}|\theta^*)}{p(\theta^{(i)})q(\theta^*|\theta^{(i)})}\right\}$$

to ensure it is a proper probability. Then

$$P_{MH}\left(\theta^{(i+1)} \middle| \theta^{(i)}\right) = q\left(\theta^{(i+1)} \middle| \theta^{(i)}\right)\alpha(\theta^{(i)}, \theta^{(i+1)}) + \delta_{\theta^{(i)}}(\theta^{(i+1)})r(\theta^{(i)}), \tag{2}$$

where $r(\theta^{(i)})$ is the term associated with rejection

$$r(\theta^{(i)}) = \int_{\mathcal{X}} q(\theta^*|\theta^{(i)})\left(1 - \alpha(\theta^{(i)}, \theta^*)\right)d\theta^*. \tag{3}$$

Intuitively, $P_{MH}\left(\theta^{(i+1)} \middle| \theta^{(i)}\right)$ gives the probability that, when the chain is at $\theta^{(i)}$, the value $\theta^{(i+1)}$ is proposed by q and accepted, and if $\theta^{(i)} = \theta^{(i+1)}$, we add the probability that any value is proposed by q and rejected.

In what sense then, is the Metropolis-Hastings algorithm implicitly designing an admissible transition probability? There is a sufficient condition, called detailed balance, which is sufficient to guarantee that a chain will have a given p as its target distribution. Metropolis-Hastings is in a sense reverse-engineered to satisfy this condition.

We say the chain corresponding to a transition kernel P is reversible when

$$P\left(\theta^{(i-1)} \middle| \theta^{(i)}\right)p(\theta^{(i)}) = P\left(\theta^{(i)} \middle| \theta^{(i-1)}\right)p(\theta^{(i-1)}), \tag{4}$$

which says that if you move by taking a draw from p and then doing one transition according to P, the probability of moving from $\theta^{(i)}$ to $\theta^{(i-1)}$ is then equal to that of moving from $\theta^{(i-1)}$ to $\theta^{(i)}$. Summing over both sides we get

$$p(\theta^{(i)}) = \sum_{\theta^{(i-1)}} P\left(\theta^{(i)} \middle| \theta^{(i-1)}\right)p(\theta^{(i-1)}).$$

The distribution p with respect to which the transition probability P fulfills the detailed balance condition is thus the invariant distribution of P. The implication that detailed balance delivers p as the stationary distribution is almost immediate, that's because the condition is strong. How easy it is to verify varies across applications.

The Metropolis-Hastings transition probability (1) satisfies detailed balance. In fact, it can be thought of as constructed to satisfy this condition. Suppose for a brief moment that $P_{q(\theta^{(i+1)}|\theta^{(i)})}\left(\delta_{\theta^{(i)}}(\theta^{(i+1)})\right) = 0$; for instance the support Θ is continuous and q is uniformly continuous with respect to Lebesgue measure. Then we can ignore the term

$\delta_{\theta^{(i)}}(\theta^{(i+1)})r(\theta^{(i)})$ in (2) and (4) takes the form

$$q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)\alpha(\theta^{(i)},\theta^{(i-1)})p(\theta^{(i)}) = q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)\alpha(\theta^{(i)},\theta^{(i-1)})p(\theta^{(i)}).$$

If $q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)p(\theta^{(i)}) > q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)p(\theta^{(i)})$, then detailed balance is trivially brought about by setting $\alpha(\theta^{(i)},\theta^{(i-1)})$ to $\frac{q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)p(\theta^{(i)})}{q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)p(\theta^{(i)})}$ and $\alpha(\theta^{(i)},\theta^{(i-1)})$ equal to 1. brought about, if $q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)p(\theta^{(i)}) > q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)p(\theta^{(i)})$, then detailed balance is trivially reinstated by setting $\alpha(\theta^{(i)},\theta^{(i-1)})$ to $\frac{q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)p(\theta^{(i)})}{q\left(\theta^{(i-1)}\left|\theta^{(i)}\right.\right)p(\theta^{(i)})}$ and $\alpha(\theta^{(i)},\theta^{(i-1)})$ equal to 1. In general, we therefore solved for

$$\alpha(\theta,\theta') = \min\left\{1, \frac{q\left(\theta\left|\theta'\right.\right)p(\theta')}{q\left(\theta'\left|\theta\right.\right)p(\theta)}\right\}.$$

Exercise 4 *Show that the Metropolis-Hastings transition probability satisfies detailed balance in the general case.*

0.3 Advanced MCMC Methods

0.3.0 Gibbs

It is sometimes the case that, even though it is difficult to sample from a joint distribution $p(\theta_1,...,\theta_n)$, it is easy to sample from full conditionals, $p\left(\theta_j\left|\theta_1,...,\theta_{j-1},\theta_{j+1},...,\theta_n\right.\right)$, $j = 1,...,n$. These carry all the information of the joint, and it turns out that drawing from the full conditionals one by one delivers an admissible sampler, i.e., the procedure described in Algorithm 3 samples from p.

Algorithm 3 Gibbs Sampler

1: Initialise θ^0
2: **for** $i = 0$ to $N-1$ **do**
3: Sample $\theta_1^{(i+1)} \sim p\left(\theta_1\left|\theta_2^{(i)},...,\theta_n^{(i)}\right.\right)$
4: Sample $\theta_2^{(i+1)} \sim p\left(\theta_2\left|\theta_1^{(i+1)},\theta_3^{(i)},...,\theta_n^{(i)}\right.\right)$
5: ...
6: Sample $\theta_n^{(i+1)} \sim p\left(\theta_2\left|\theta_1^{(i+1)},\theta_2^{(i+1)},...,\theta_{n-1}^{(i+1)}\right.\right)$
7: **end for**

One way to establish that p is indeed the stationary distribution of the Gibbs sampler is to recognize it as a Metropolis-Hastings sampler (Andrieu et al., 2003), which implies Gibbs satisfies detailed balance with respect to p. It is likewise straightforward to directly show that the detailed balance condition is fulfilled.

Exercise *5* *Show that the Gibbs sampler satisfies detailed balance.*

Note that if some full conditional is difficult to simulate from, one can use Metropolis-Hastings within Gibbs to draw from the full conditional. Further note that θ_i can itself be a (sub) vector, in which case we talk of a block Gibbs algorithm.

0.3.1 Simulated Annealing

A candid way of using MCMC methods to estimate the maximum of a posterior distribution is to sample from the distribution and then report the draw with the largest evaluation of the target distribution,

$$\hat{\theta} = \arg \max_{\theta^{(i)}, i=1,\dots,n} p(\theta^{(i)}).$$

However, this lets the chain vagabond far from the maximum, for lengthy sojourns that have high computational cost and no information return. We would like to keep the chain close to the optimum, and one way to do this is to "exaggerate" the optimum. We start "exaggerating" once we are confident the chain is not "too far" from it, i.e. once we have plausibly reached stationarity, and do so gradually. More precisely, we use a non-homogeneous target

$$p_i(\theta) \propto p^{1/T_i}(\theta),$$

which gradually concentrates around its optimum as $i \to \infty$ and $T_i \to 0$, according to the –delicately chosen– **cooling schedule** $\{T_1, T_2, ...\}$.

Algorithm 4 Simulated Annealing

1: **for** $i = 0$ to $N - 1$ **do**
2: Sample $u \sim U[0, 1]$
3: Sample $\theta^* \sim q(\theta^*|\theta^i)$
4: If $u < \frac{p^{1/T_i}(\theta^*) q(\theta^{(i)}|\theta^*)}{p^{1/T_i}(\theta^{(i)}) q(\theta^*|\theta^{(i)})}$
5: Then $\theta^{(i+1)} = \theta^*$
6: Else $\theta^{(i+1)} = \theta^{(i)}$
7: **end for**

0.3.2 Auxiliary Variable Strategies

General ideas. Classical papers. Including the Neil Shephard paper.

0.3.3 Approximate Targets and Pseudo-Marginal MCMC

Suppose you want to run a Metropolis-Hastings algorithm, but the target distribution is computationally intractable. Suppose however that it may be estimated using importance

sampling or some other unbiased method. It is then tempting to replace the target distribution (in the formula for the Metropolis-Hatings acceptance ratio) by its estimate. Intuitively, if the estimate is very accurate, the resulting draws should approximate draws from the target distribution. This is approach is called Markov Chain Within Metropolis (MCWM).

But using a noisy estimate of the target via the naice MCWN approach delivers an inadmissible sampler. Indeed, MCWM will typically not have the "correct" target distribution. See Andrieu and Roberts (2009; p.699).

The motivation for pseudo-marginal MCMC is that by instead treating the importance sampling as the marginalization of an auxiliary variable z, we can treat z as "just another" random variable whose (implicitly defined) posterior is sampled with the Metropolis-Hastings sampler.

As can be observed from Table 1 of Andrieu and Roberts (2009), the difference between the implementation of MCWM and pseudo-marginal MCMC (there called group independence MH, or GIMH), although crucial, is very small.

The traditional justification of pseudo-marginal MCMC as a Metropolis-Hastings sampler relies on thinking of the approximation to the posterior as coming from an importance sampler. We provide the traditional derivation in Appendix A but give a simpler derivation for now.

The key idea: introducing and marginalizing an auxiliary variable

The classical derivation can be simplified to better convey the point that the success of pseudo-marginal MCMC does not rely on importance sampling *per se*, but rather on having access to draws of an unbiased estimate of the target distribution and using them as auxiliary variables, to be marginalized from a joint posterior.

Suppose that we wish to sample from

$$\pi(\theta)$$

using a Metropolis-Hastings sampler with acceptance ratio

$$\alpha = \frac{\pi(\theta^*)q(\theta|\theta^*)}{\pi(\theta)q(\theta^*|\theta)}.$$

However, we do not have access to $\pi(\theta)$, but only draws $\widehat{\pi(\theta)} \geq 0$, which are unbiased estimates, i.e., $E\left[\widehat{\pi(\theta)}\right] = \pi(\theta)$, $\forall\ \theta$.

The claim is that by running the sampler described in Algorithm 5, we will draw from π. The key observation is that Algorithm 5 may be represented as a *bona fide* (without

Algorithm 5 Pseudo-Marginal MCMC

1: **for** $k = 1$ to K **do**
2: Sample $u \sim U[0,1]$
3: Sample $\theta^* \sim q(\theta^*|\theta^{(k)})$
4: Sample unbiased estimate $\left(\hat{\pi}\left(\theta^*\right)\right)^*$
5: If $u < \frac{\left(\hat{\pi}(\theta^*)\right)^* q(\theta^{(k)}|\theta^*)}{\left(\hat{\pi}\left(\theta^{(k)}\right)\right)^{(k)} q(\theta^*|\theta^{(k)})}$
6: Then $\theta^{(k+1)} = \theta^*$ and $\left(\hat{\pi}\left(\theta^{(k+1)}\right)\right)^{(k+1)} = \left(\hat{\pi}\left(\theta^*\right)\right)^*$
7: Else $\theta^{(k+1)} = \theta^{(k)}$ and $\left(\hat{\pi}\left(\theta^{(k+1)}\right)\right)^{(k+1)} = \left(\hat{\pi}\left(\theta^{(k)}\right)\right)^{(k)}$
8: **end for**

approximation) Metropolis-Hastings sampler whose (pseudo) marginal posterior is the target distribution of interest, π. Nevertheless observe that Algorithm 5, on the face of it, simply reads as a Metropolis-Hastings sampler where the target distribution is approximated.

For emotional comfort, write $\hat{\pi}_\theta^k := \left(\hat{\pi}\left(\theta^{(k)}\right)\right)^{(k)}$ and $\hat{\pi}_\theta^* := \left(\hat{\pi}\left(\theta^*\right)\right)^*$, more easily thought of as auxiliary random variables. The approximations are distributed according to some law,

$$\hat{\pi}_\theta^{(k)} \sim p(\cdot|\theta^{(k)}),$$

likewise $\hat{\pi}_\theta^* \sim p(\cdot|\theta^*)$. Consider the one-to-one mapping defined $\hat{\pi}_\theta^{(k)} \mapsto \hat{\pi}_\theta^{(k)}/\pi(\theta^{(k)})$ and $\hat{\pi}_\theta^* \mapsto \hat{\pi}_\theta^*/\pi(\theta^*)$; the distribution of the newly formed random variables readily obtains by the change-of-variable formula.

Under these considerations, we may rewrite the acceptance ratio of Algorithm 5 as follows,

$$\begin{aligned}
&\frac{\hat{\pi}_\theta^* q(\theta^{(k)}|\theta^*)}{\hat{\pi}_\theta^k q(\theta^*|\theta^{(k)})} \\
=&\frac{\frac{\hat{\pi}_\theta^*}{\pi(\theta^*)}\pi\left(\theta^*\right) q(\theta^{(k)}|\theta^*)}{\frac{\hat{\pi}_\theta^k}{\pi(\theta^k)}\pi\left(\theta^{(k)}\right) q(\theta^*|\theta^{(k)})} \\
=&\frac{\omega^*\pi\left(\theta^*\right) p(\omega^*|\theta^*) q(\theta^{(k)}|\theta^*) p(\omega^{(k)}|\theta^{(k)})}{\omega^k\pi\left(\theta^{(k)}\right) p(\omega^{(k)}|\theta^{(k)}) q(\theta^*|\theta^{(k)}) p(\omega^*|\theta^*)}
\end{aligned}$$

where $\omega^* = \frac{\hat{\pi}_\theta^*}{\pi(\theta^*)}$ and $\omega^{(k)} = \frac{\hat{\pi}_\theta^{(k)}}{\pi\left(\theta^{(k)}\right)}$.

Therefore, Algorithm 5 proposes (ω, θ) from $p(\omega|\theta)q(\theta|\theta^{(k-1)})$ and targets a density proportional to $\omega\pi\left(\theta\right)p(\omega|\theta)$. Marginalizing over the auxiliary variable ω, we obtain the pseudo marginal

$$\pi\left(\theta\right) = \int_\omega \omega\pi\left(\theta\right)p(\omega|\theta)d\omega,$$

where we used unbiasedness, $\int_\omega \omega p(\omega|\theta)d\omega = E[\hat{\pi}_\theta/\pi(\theta)] = 1$. We see that Algorithm 5 is indeed a Metropolis-Hastings sampling from a posterior having the distribution of interest as its marginal.

0.3.4 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo

0.3.5 ABC and Bayesian Inference on Manifolds

As mentioned in the introduction, in some cases even evaluation of the likelihood up to the normalization constant may be too computationally taxing. In that case, one may employ a set of tools referred to as likelihood free inference. Its MCMC representative is Approximate Bayesian Inference (ABC). ABC may be employed when the likelihood is intractable but a generative model is available. That is, the statistic Y is generated from a generative model g which takes as its argument the parameter of interest θ and a random element z of known distribution,

$$Y_\theta = g(\theta, z),\ \theta \in \Theta,\ z \sim F_z.$$

The idea is that if we generate a Y_θ with θ close to the true parameter θ_0, then Y_θ and Y_{θ_0} must themselves be close to each other. This suggests an acceptance criterion for a proposed θ^*. Precisely, the ABC algorithm is as stated in Algorithm 6, for a general distance function d between statistics.

Algorithm 6 ABC

for $i = 0$ to $N - 1$ **do**
  Sample $u \sim U[0, 1]$
  Sample $\theta^* \sim q(\theta^*|\theta^i)$
  Sample $z \sim F_z$
  Compute $Y_{\theta^*} = g(\theta^*, z)$
  If $u < \mathbf{1}\left\{d\left(\hat{Y}, Y_{\theta^*}\right) < \epsilon\right\} \frac{q(\theta^{(i)}|\theta^*)}{q(\theta^*|\theta^{(i)})}$
    Then $\theta^{(i+1)} = \theta^*$
    Else $\theta^{(i+1)} = \theta^{(i)}$
end for

ABC can be thought of as the Bayesian inference equivalent of the indirect inference. For a deep comparison of the two methods, see the beautiful paper Forneron and Ng (2016). Some of the same issues arise in both ABC and indirect inference. In particular, the choice of the statistic or auxiliary variable is delicate and important. Bernton et al.

(2017) is a good example of attempting to circumvent this problem by looking directly at the distance between true and simulated datas sets.

It is clear from the algorithm that proposals close to the true distribution are crucial to the practicality of the algorithm. For that reason, ABC algorithms are often used with adaptive priors, such as sequential Monte Carlo.

Nevertheless, it is very difficult to provide diagnostics in order to assess whether ϵ is small enough to be considered negligible, and ABC output is often considered as approximative.

Exercise 7 *Give the expression for the exact distribution, called the ABC posterior, which is sampled by the ABC sampler.*

We can however do ABC inference with $\epsilon = 0$. Indeed, setting $Y_\theta = g(\theta, z)$ defines a nonlinear manifold in ambient space Θ, and a distribution implied on it by F_z. In practice, the key piece when implementing a Monte Carlo sampling algorithm on such a manifold is to get a projection for projecting back on the manifold after a "proposal step" which may take you outside of the manifold. Refer to Graham and Storkey (2016) for ABC inference with $\epsilon = 0$, Diaconis et al. (2013) for a more thorough investigation of inference on manifolds, and Federer (2014) for the relevant mathematics.

0.4 Advanced MCMC Theory

We think of a good transition probability P as one which will make for a short burn-in period, amongst other things. That is, a "good" chain is one we do not have to run for too long before we obtain a draw from the stationary distribution. Without theoretical guidance, the analyst usually runs the chain until convergence is assessed using some heuristic or *ad hoc* method. There is therefore much need to provide theoretical guarantees on the convergence rates of given MCMC algorithms.

Many distances are used to talk about how close the distribution of MCMC draws is to the stationary distribution after running the chain for some time. Different ones are chosen to accommodate different proof strategies (see Diaconis, 2013, for discussion) and to provide different assumptions to verify. Common choices are *total variation distance*, which is proportional to the ℓ_1 norm on finite spaces, and the ℓ_2 norm with respect to some base measure.

0.4.0 Spectral Bounds

When studying nonasymptotic rates of convergence for Markov chains, it is useful to treat the transition probability as a linear algebraic object. In fact, we will find that its

second biggest eigenvalue (the first corresponds to the stationary distribution) governs the convergence rate.

As a linear operator, P applies to a measure on $\mathcal{X}$ from the right. That is, after applying one step of the chain to a measure μ (a horizontal vector), we obtain the measure μP. Whence the constraint in P that its *rows* sum up to 1.

Before talking about convergence to a stationary distribution π, we must pin it down[3]! Since the rows of P sum up to one, we have that

$$P\mathbf{1}_n = \mathbf{1}_n,$$

which is to say that $\mathbf{1}_n$ is a right eigenvector of P with eigenvalue 1.

Exercise 8 *Let $\Lambda_R(A)$ and $\Lambda_L(A)$ be the set of eigenvalues corresponding to right and left eigenvectors, respectively. Show that $\Lambda_R(A) = \Lambda_L(A) = \Lambda(A)$. hint: use the characteristic polynomial. challenge: can you give a proof without using the characteristic polynomial, but instead via matrix operations and by treating A as a linear operator?*

We will denote the set of eigenvalues of P by $\Lambda(P) = \{\lambda_0, \lambda_1, ..., \lambda_{n-1}\}$ and assume, without loss of generality, that $\lambda_0 \geq \lambda_1 \geq ... \geq \lambda_{n-1}$.

Because $\mathbf{1}_n$ is a right eigenvector with eigenvalue 1, there is a left eigenvector π, with eigenvalue 1,

$$\pi P = \pi,$$

in other words, a stationary distribution. Note that this linear algebraic argument is very direct, but does not have the pedagogical quality of being constructive. Compare with theorem 1.

Is it unique? Uniqueness guarantees for stationary distribution in finite Markov Chains are offered by the Frobenius-Perron theorem (Saloff-Coste, 1997; Gantmacher, 1959). Under slightly lenient assumptions (we point them out below), a more conceptual and pedagogically efficient argument can be laid out.

Lemma 1 *If P has positive elements, i.e., $P_{i,j} > 0$ for all i, j, then $\lambda_1 < \lambda_0 \equiv 1$.*

Proof

Consider the case in which P is diagonalizable. Suppose $Pv = \lambda v$. Pick x such that

[3]This was done in Theorem 1, but here we give a streamlined, nonconstructive, linear algebraic proof of the flavor of the convergence theorem to come.

$|v(x)| \geq |v(y)|$, $\forall\ y \in \mathcal{X}$. Then

$$|\lambda v(x)| = |\left(Pv\right)_x| = \left|\sum_y P_{x,y} v(y)\right| \leq \sum_y |v(y)|\, P_{x,y} \leq \sum_y |v(x)|\, P_{x,y} = |v(x)|\,.$$

The inequality can only be an inequality if v is constant. Thus λ_0 is the only eigenvalue equal to 1. Therefore, if P is diagonalizable (and all generalized eigenvectors are in fact eigenvectors), then the eigenvalue 1 has multiplicity 1.

If P is not diagonalizable, then it is straightforward to show that $\lambda_0 = 1$ is not part of a larger Jordan block. See Rosenthal (1995) Fact 1. See the mathematical supplement for a review of Jordan blocks. □

We have pinned down a unique stationary distribution π. The question of utmost practical importance is then: can we quantify *ex ante* whether a Monte Carlo chain can reach it within reasonable time? Yes. In fact, a short and elegant argument delivers a nonasymptotic bound on the distance between the current iterate and the stationary distribution.

Theorem 3 *Suppose $\lambda_1(P) < 1$ and P is diagonalizable, then for any given initial distribution μ_0,*

$$\|\pi - \mu_l\|_2^2 \leq C\lambda_1^l,$$

where $\mu_l = \mu_o P^l$.

Proof

Write μ_0 in terms of the orthonormal basis of left eigenvectors of P,

$$\mu_0 = a_0\pi + a_1 v_1 + \cdots + a_{n-1} v_{n-1}.$$

Since $1 > \lambda_j$, $j = 1, ..., n-1$, we have that

$$\mu_l = \mu_0 P^l = a_0\pi + a_1\lambda_1^l v_1 + \cdots + a_{n-1}\lambda_{n-1}^l v_{n-1} \to a_0\pi.$$

and thus $a_0 = 1$. We can then quantify the rate of this convergence,

$$\begin{aligned}\|\pi - \mu_l\|_2^2 &= \left\|\pi - \pi - a_1\lambda_1^l v_1 - \cdots - a_{n-1}\lambda_{n-1}^l v_{n-1}\right\|_2^2 \\ &= \left\|a_1\lambda_1^l v_1 + \cdots + a_{n-1}\lambda_{n-1}^l v_{n-1}\right\|_2^2 \\ &\leq \sum_{m=1}^{n-1} |a_m|^2 |\lambda_m|^{2l} \\ &\leq \left(\sum_{m=1}^{n-1} |a_m|^2\right) \lambda_1^l.\end{aligned}$$

□

The result immediately generalizes to non-diagonalizable matrices by accounting for the size of the Jordan blocks. See the mathematical supplement for a review of Jordan blocks. Further note that using the ℓ_1 instead of the ℓ_2 norm in the argument would have delivered a bound on $\|\pi - \mu_l\|_1$, and thus, by Proposition 1 below, a bound on total variation, which is detailed in Definition 5 below.

The discussion above presented a streamlined version of the existence Theorem 1. Likewise, Lemma 1 presented a streamlined version of the uniqueness Theorem 2. The convergence Theorem 3 was of the linear algebraic flavor of the streamlined theorems. We now give a more probabilistic proof which hopefully will convey greater, or at least additional, intuition.

0.4.1 Distance in Total Variation

We may be interested in bounding the distance to the stationary distribution using a different notion of distance. One reason to do this is that it provides an alternative way to compute the bound which may be easier in some applications. Another reason is that a different distance measure calls for a different convergence proof which may convey a different intuition. In this subsection, we study convergence of the Markov chain to its stationary distribution as measured by the total variation distance.

Definition 5 *The **total variation distance** between two probability distributions μ and ν on χ is defined*

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A \subseteq \chi} |\mu(A) - \nu(A)|.$$

There are very useful alternative formulations of the total variation distance.

Proposition 1 *Let μ and ν be two probability distributions on χ, then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\chi} |\mu(x) - \nu(x)| = \sum_{x:\mu(x)\geq\nu(x)} |\mu(x) - \nu(x)| .$$

Another alternative formulation follows from the following observation. Instead of optimizing by picking weights in the domain of the probability measures, we may pick weights in the range.

Proposition 2 *Let μ and ν be two probability distributions on χ, then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\sup\left\{\sum_{x\in\chi} f(x)\mu(x) - \sum_{x\in\chi} f(x)\nu(x) \ : \ \max_{x\in\chi}|f(x)| \leq 1\right\}.$$

A very powerful and interesting formulation of total variation distance is via couplings.

Definition 6 *A **coupling** of two probability distributions μ and ν is a pair of random variables (X, Y) defined on a single probability space such that the marginal distribution of X is μ and the marginal distribution of Y is ν. That is, a coupling (X, Y) satisfies $\mathbf{P}\{X = x\} = \mu(x)$ and $\mathbf{P}\{Y = y\} = \nu(y)$.*

Intuition gathered from the optimal transport problem suggests that the "distance" between two random variables X and Y may be measured by how "close" a feasible joint distribution can get to the distribution with respect to which X and Y equal in probability, i.e., $P(\{X = Y\}) = 1$.

Exercise 9 *Write the computation of total variance as an optimal transport problem. In particular, display explicitly the cost matrix.*

For example, if the marginals of both X and Y are obtained from the flip a fair coin, i.e. $X, Y \in \{0, 1\}$ and $P(X = 1) = P(Y = 1) = 1/2$, then we may couple them so that they are equal in probability, i.e. $P(X = i, Y = i) = 1/2$, $i = 0, 1$, and $P(\{X \neq Y\}) = 0$.

However, when ν and μ are not identical, it is not possible for them to be equal in probability. But how close to identical can they be? As it turns out, total variation distance quantifies and measures precisely this.

Proposition 3 *Let μ and ν be two probability distributions on χ, then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \inf\left\{\mathbf{P}\left\{X \neq Y\right\} \ : \ (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\right\}, \tag{5}$$

and the minimum is achieved by some joint distribution.

Definition 7 *The joint distribution achieving the minimum in (5) is called the* ***optimal coupling****.*

Proof of Proposition

It immediately obtains that for any coupling (X, Y) of μ and ν and any event $A \subset \chi$, we have

$$\begin{aligned}\mu(A) - \nu(A) &= \mathbf{P}\{X \in A\} - \mathbf{P}\{Y \in A\} \\ &\leq \mathbf{P}\{X \in A, Y \notin A\} \\ &\leq \mathbf{P}\{X \neq Y\},\end{aligned}$$

from which it immediately follows that

$$\|\mu - \nu\|_{\mathrm{TV}} \leq \inf\{\mathbf{P}\{X \neq Y\} \ : \ (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}.$$

The challenge is to find a coupling, the optimal coupling, for which the above holds with equality. We give the recipe for constructing an optimal coupling.

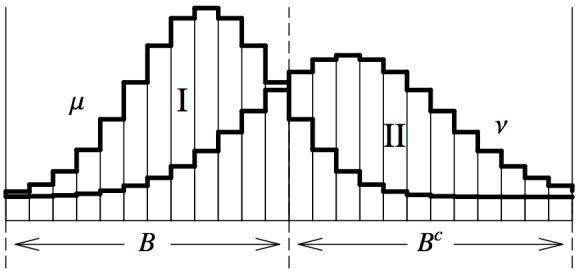

Figure 3 *Since each of regions I and II has area* $\|\mu - \nu\|_{\mathrm{TV}}$ *and* μ *and* ν *are probability measures, region III has area* $1 - \|\mu - \nu\|_{\mathrm{TV}}$*. Picture from Levin and Peres (2017).*

We construct the optimal coupling by forcing X and Y to be equal as often as they possibly can. Consider Figure 3. Pick a point uniformly at random in the union of regions I and III and set X to be the x-coordinate at this point. If the point is in III, set $Y = X$. If the point is in I, then choose a point uniformly at random in II and set Y to be its x-coordinate. It is clear that $X = Y$ if and only if the first randomly drawn point

falls in III. Therefore,

$$\begin{aligned}\mathbf{P}\{X \neq Y\} &= 1 - \sum_{x\in\chi} \mu(x) \wedge \nu(x) \\ &= 1 - \sum_{x:\mu(x)\leq\nu(x)} \mu(x) - \sum_{x:\nu(x)<\mu(x)} \nu(x) \\ &= \sum_{x:\nu(x)<\mu(x)} \mu(x) - \nu(x) \\ &= \|\mu - \nu\|_{\mathrm{TV}}\,,\end{aligned}$$

where proposition 1 was invoked for the last equality.

□

Exercise 10 *Are optimal couplings unique? Give a proof or counterexample.*

We can give a distance to the stationary distribution which is independent of the starting point by looking at the worst case. Define

$$d(t) := \max_{x\in\chi} \left\| P^t(x,\cdot) - \pi \right\|_{\mathrm{TV}}.$$

The definition above makes sense because it suffices to look only at degenerate initial distributions.

Exercise 11 *Show that* $d(t) = \sup_{\mu\in\mathcal{P}} \|\mu P^t - \pi\|_{\mathrm{TV}}$*, where* $\mathcal{P}$ *is the collection of all probability distributions on* χ*.*

A most precious quantifier of the performance of a Markov chain is its mixing time.

Definition 8 *The* ***mixing time*** *is defined by*

$$t_{\mathrm{mix}}(\varepsilon) := \min\{t \ : \ d(t) \leq \varepsilon\}$$

and

$$t_{\mathrm{mix}} := t_{\mathrm{mix}}(1/4).$$

Proposition 4 *The following holds,*

$$t_{\mathrm{mix}}(\varepsilon) \leq \left\lceil \log_2 \varepsilon^{-1} \right\rceil t_{\mathrm{mix}}.$$

✹

The mixing time measures the running time of a Markov chain passed which the starting distribution does not matter much anymore, regardless of the distribution. A "Cauchy-flavored" way of assessing that the starting distribution does not matter is to look at the distance of any two chains with different starting distributions. In order to do this, we need to specify a joint distribution –a coupling– for any two such chains. The boon from such an approach is that any joint distribution will yield an upper bound. That is, for *any* given coupling of the P-chain $X_{x_0,t}$ and $X_{x_0',t}$ started at x_0 and x_0', respectively, the maximum waiting time over *all* pairs x_0, x_0' until the chains are indistinguishable ought to inform the mixing time. We now formalize this intuition.

Definition 9 *A **coupling of Markov chains** with transition matrix P is a process $(X_t, Y_t)_{t=1}^{\infty}$ with the property that both (X_t) and (Y_t) are Markov chains with transition matrix P, although the two chains may possibly have different starting distributions.*

Definition 10 *Given a Markov chain on χ with transition matrix P, a **Markovian coupling** of two P-chains is a Markov chain $\{(X_t, Y_t)\}_{t\geq 0}$ with state space $\chi \times \chi$ which satisfies, for all x,x',y,y',*

$$\mathbf{P}\{X_{t+1} = x' \,|X_t = x,\ Y_t = y\} = P(x, x'),$$

$$\mathbf{P}\{Y_{t+1} = y' \,|X_t = x,\ Y_t = y\} = P(y, y').$$

Proposition 3 would suggest to compute bounds on the probability that the chains are unequal. Perhaps surprisingly, we will often find it easier to build bounds in terms of the time until chains are *exactly* equal –obviously, this will yield an upper bound. The reason for this is that it is "easy" to build Markovian couplings according to which the two coupled chains are forever equal after they meeting, i.e., such that

$$\text{if } X_s = Y_s, \text{ then } X_t = Y_t \text{ for } t \geq s. \tag{6}$$

To construct a coupling satisfying (6), simply run the chains according to whatever original coupling until they meet, then run them together.

Theorem 4 *Let $\{(X_t, Y_t)\}$ be a coupling satisfying (6) for which $X_0 = x$ and $Y_0 = y$. Let τ_{couple} be the coalescence time of the chains:*

$$\tau_{\text{couple}} := \min\{t \ : \ X_s = Y_s \text{ for all } s \geq t\}. \tag{7}$$

Then,

$$\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\text{TV}} \leq \mathbf{P}_{x,y}\left\{\tau_{\text{couple}} > t\right\}. \tag{8}$$

PROOF

The coupling (X, Y) implies a coupling for (X_t, Y_t) with marginals $P^t(x,z) = P_{x,z}\{X_t = z\}$ and $P^t(y,z) = P_{x,z}\{Y_t = z\}$. Therefore, Proposition 3 implies that

$$\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\text{TV}} \leq \mathbf{P}_{x,y}\left\{X_t \neq Y_t\right\}.$$

But $\mathbf{P}_{x,y}\left\{X_t \neq Y_t\right\} = \mathbf{P}_{x,y}\left\{\tau_{\text{couple}} > t\right\}$, and the desired expression obtains.

□

As will be clear in example 4, the following corollary allows us to build bounds on the mixing time.

Corollary 1 *Suppose that for each pair of states $x, y \in \chi$ there is a coupling (X_t, Y_t) with $X_0 = x$ and $Y_0 = y$. For each such coupling, let τ_{couple} be the coalescence time of the chains. Then*

$$d(t) \leq \max_{x,y\in\chi} \mathbf{P}_{x,y}\left\{\tau_{\text{couple}} > t\right\} \tag{9}$$

and therefore

$$t_{\text{mix}} \leq 4 \max_{x,y\in\chi} E_{x,y}\left[\tau_{\text{couple}}\right]. \tag{10}$$

In practice, we may get an upper bound if we find explicit coupling, and the optimal coupling will give the tightest upper bound.

Example 4: Random Walk on Cycle *Consider a random walk on $\mathbb{Z}_n$ which stays where it is with probability $1/2$, and moves down and up (mod n) with probability $p/2$ and $q/2$, respectively. Naturally, $p + q = 1$. The key exercise is to build a coupling for the chains X and Y, possibly started at different points. Our coupling is as follows, at each step we flip a fair coin to decide which of the two chain moves. This takes care of the 1/2 probability of staying put in the respective marginal, and eliminates the possibility that the two chains "jump over each other" –recall, we are trying to get them to coalesce as early as possible. The selected chain then moves down with probability p and up with probability q. Once the chains coalesce, they both move in sync according to the marginal law.*

Let D_t be the clockwise distance from X_t to Y_t. Note that D_t is a simple random walk on the interior vertices of $\{0, 1, ..., n\}$ and gets absorbed at either 0 or n. It is an instance of the gambler's ruin problem (Levin and Peres, 2017), whence we know that $E_{x,y}[\tau] = k(n-k)$ where $\tau = \min\{t \geq 0 \ : \ D_t \in \{0, n\}\}$ and k is the clockwise distance between x and y. Of course, $\tau = \tau_{\text{couple}}$, therefore, using Markov's

inequality,

$$d(t) \leq \max_{x,y \in \mathbb{Z}_n} P_{x,y}\{\tau > t\} \leq \frac{\max_{x,y} E_{x.y}[\tau]}{t} \leq \frac{n^2}{4t}.$$

The right-hand side equals $1/4$ *for* $t = n^2$*, whence* $t_{\text{mix}} \leq n^2$*.*
We can obtain a lower bound by working on $\mathbb{Z}$*. Let* S_t *be the lazy* $(p$*–*$q)$ *random walk on* $\mathbb{Z}$*, and* $X_t = S_t \mod n$*. Let* ρ *be the distance on the cycle, and write the mean drift term* $\mu_t = t(p-q)/2$*. Define*

$$A_t = \{k \ : \ \rho\left(k, \lfloor x_0 + \mu_t \rfloor \mod n\right) \geq n/4\}.$$

Note that $\pi(A_t) \geq 1/2$*. Using Chebyshev's inequality, since* $Var(S_t) = t(\frac{1}{4} + pq) \leq t/2$*,*

$$P\{X_t \in A_t\} \leq P\{|S_t - \mu_t| \geq n/4\} \leq \frac{8t}{n^2} < \frac{1}{4}$$

for $t < n^2/32$*. Thus, for* $t < n^2/32$*, directly applying the definition of total variation,*

$$d(t) \geq \pi(A_t) - P(X_t \in A_t) > \frac{1}{2} - \frac{1}{4},$$

hence $t_{\text{mix}} \geq n^2/32$*.*

In the end, whether you should work out bounds using spectral methods or couplings is a practical matter, it might be easier to compute –perhaps a bound on– the second biggest eigenvalue than it is to find a coupling yielding a good bound, or it might be the converse. And there are yet more methods for bounds which we have not explored yet.

0.5 Further Readings and Open Questions

Rosenthal (1995a) gives a nice overview of the use of minorization conditions[4] to construct a coupling. Rosenthal (1995b) further develops on the topic. Jones and Hobert (2004) borrow this strategy to derive an upper bound on the burn in time for a Gibbs sampler of the Bayesian hierarchical version of the one way random effects model.

Yang et al. (2016) provide a bound on the mixing time for an MCMC algorithm for high-dimensional Bayesian linear regression under sparsity constraint. Their proof "controls the spectral gap of the Markov chain by constructing a canonical path ensemble that is inspired by the steps taken by greedy algorithms for variable selection."

Belloni and Chernozhukov (2009) use the Berstein von Mises theorem and resulting normal approximation to "establish polynomial bounds on the computational complexity

[4] Precisely,

$$P^k(x, A) \geq \beta\zeta(A), \ x \in \chi, \ A \subset \chi,$$

for k a positive integer, $0 < \beta < 1$, and ζ some probability distribution on χ.

of general Metropolis random walks methods in large samples."

We only treated the case of discrete probability spaces. Continous spaces are of course common in practice, but the theory becomes substantially more convoluted. For exposition of some results, consult Levin and Peres (2017) and Roberts and Rosenthal (2004). The latter presents applications of coupling and minorisation are presented.

References

Andrieu, Christophe, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan. "An introduction to MCMC for machine learning." *Machine learning* 50, no. 1-2 (2003): 5-43.

Belloni, Alexandre, and Victor Chernozhukov. "On the computational complexity of MCMC-based estimators in large samples." *The Annals of Statistics* (2009): 2011-2055.

Bernton, Espen, Pierre E. Jacob, Mathieu Gerber, and Christian P. Robert. "Inference in generative models using the Wasserstein distance." *arXiv preprint arXiv*:1701.05146 (2017).

Diaconis, Persi. "Some things we've learned (about Markov chain Monte Carlo)." Bernoulli 19, no. 4 (2013): 1294-1305.

Diaconis, Persi, Susan Holmes, and Mehrdad Shahshahani. "Sampling from a manifold." In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, pp. 102-125. Institute of Mathematical Statistics, 2013.

Efron, Bradley, and Trevor Hastie. *Computer Age Statistical Inference*. Vol. 5. Cambridge University Press, 2016.

Federer, Herbert. *Geometric measure theory*. Springer, 2014.

Forneron and Ng (2016)

Gantmacher, Feliks R. "Matrix theory." *Chelsea, New York* 21 (1959).

Graham, Matthew M., and Amos Storkey. "Asymptotically exact inference in likelihood-free models." *arXiv preprint arXiv:1605.07826* (2016).

Hoff, Peter D. *A first course in Bayesian statistical methods*. Springer Science & Business Media, 2009.

Jones, Galin L., and James P. Hobert. "Sufficient burn-in for Gibbs samplers for a hierarchical random effects model." *The Annals of Statistics* 32, no. 2 (2004): 784-817. Harvard

Levin, David A., and Yuval Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

Neal, R.M., 2011. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).

Rosenthal, Jeffrey S. "Convergence rates for Markov chains." *Siam Review* 37, no. 3 (1995a): 387-405.

Rosenthal, Jeffrey S. "Minorization conditions and convergence rates for Markov chain Monte Carlo." *Journal of the American Statistical Association* 90, no. 430 (1995b): 558-566.

Saloff-Coste, Laurent. "Lectures on finite Markov chains." In *Lectures on probability theory and statistics*, pp. 301-413. Springer Berlin Heidelberg, 1997.

Yang, Yun, Martin J. Wainwright, and Michael I. Jordan. "On the computational complexity of high-dimensional Bayesian variable selection." *The Annals of Statistics* 44, no. 6 (2016): 2497-2532.

Exercises

Exercise 14 **a**. Give an example of a pair P and π where P samples from π but is not aperiodic. **b**. Give an example of a pair P and π where P samples from π and is aperiodic but does not satisfy detailed balance.

Exercise 15 Let P be a transition probability matrix. Show that the null space of $P - I$ has dimension 1, i.e., $\text{rank}(\text{null}(P - I)) = 1$.

Exercise 16 Let $P^{\infty} = \lim P^l$, and show that $P^{\infty}_{i,\cdot} = \pi$, for $i = 1., ..., n$.

Exercise17 *C*onsider this "simplest nontrivial example" with

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

where $0 < p, q < 1$. Let $\mu_0 = (1, 0)$. You can verify that $\pi = \frac{1}{p+q}(q, p)$.
Show directly that

$$\mu_l(0) = \frac{q}{p+q} + \left(1 - \frac{q}{p+q}\right)(1-p-q)^l.$$

Conclude that $\|u_l - \pi\| = \left(1 - \frac{q}{p+q}\right)|1-p-q|^l$.

This provides an instance in which the bound offered by theorem 3 is tight. The second eigenvector of P from the exercise above is $v_1 = (1, -1)$, which has $\ell_2(\pi)$ norm equal to 1. The decomposition gives

$$\mu_0 = \pi + a_1 v_1,$$

implying $a_1 = \left(1 - \frac{q}{p+q}\right)$. Invoking theorem yields $|\mu_l(x) - \pi(x)| \leq \left(1 - \frac{q}{p+q}\right)|1-p-q|^l$ and thus

$$\|\mu_l(x) - \pi(x)\|_{\text{TV}} \leq \sum_x \left(1 - \frac{q}{p+q}\right)|1-p-q|^l \pi(x) = \left(1 - \frac{q}{p+q}\right)|1-p-q|^l .$$

Research Questions

Question 1 Can you find an economic problem whose generative distribution has a tractable optimal coupling? Can you give a closed form for optimal coupling? This should be very useful for indirect inference.

Appendix A: Pseudo-Marginal MCMC *à la* Andrieu and Roberts (2009)

We want to sample the distribution $\pi(\theta)$ but its evaluation is too costly or intractable. We have access to an importance sample estimate

$$\tilde{\pi}^N(\theta) := \frac{1}{N}\sum_{k=1}^{N} \frac{\pi(\theta, z(k))}{q_\theta(z(k))} \text{ with } z(k)|\theta \overset{\text{iid}}{\sim} q_\theta(\cdot).$$

We want a to find the right way to draw the $z(k)$'s such that, regardless of the value of $N \in \mathbb{N}$, we can regard

$$\tilde{r}^N(\theta^*, \theta) := \frac{\tilde{\pi}^N(\theta^*)}{\tilde{\pi}^N(\theta)} \cdot \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}$$

as the acceptance ratio of a *bona fide* Metropolis-Hastings algorithm.

The insight of Beaumont (2003) was that by using the $z(k)$'s as auxiliary variables in the Metropolis-Hastings scheme, we can work out the following expansion, which will have the desired interpretation:

$$\tilde{r}^N(\theta^*, \theta) = \frac{\frac{1}{N}\sum_{k=1}^{N} \frac{\pi(\theta^*, z^*(k))}{q_\theta(z^*(k))}}{\frac{1}{N}\sum_{k=1}^{N} \frac{\pi(\theta, z(k))}{q_\theta(z(k))}} \cdot \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}$$

$$= \frac{\frac{1}{N}\sum_{k=1}^{N} \frac{\pi(\theta^*, z^*(k))}{q_\theta(z^*(k))} \cdot \frac{\prod_{i=1}^{N} q_\theta(z^*(i))}{\prod_{i=1}^{N} q_\theta(z^*(i))}}{\frac{1}{N}\sum_{k=1}^{N} \frac{\pi(\theta, z(k))}{q_\theta(z(k))} \cdot \frac{\prod_{i=1}^{N} q_\theta(z(i))}{\prod_{i=1}^{N} q_\theta(z(i))}} \cdot \frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}$$

$$= \frac{\frac{1}{N}\sum_{k=1}^{N} \pi(\theta^*, z^*(k)) \cdot \prod_{i\neq k} q_\theta(z^*(i))}{\frac{1}{N}\sum_{k=1}^{N} \pi(\theta, z(k)) \cdot \prod_{i\neq k} q_\theta(z(i))} \cdot \frac{\prod_{i=1}^{N} q_\theta(z(i)) q(\theta|\theta^*)}{\prod_{i=1}^{N} q_\theta(z^*(i)) \cdot q(\theta^*|\theta)}$$

$$= \frac{\tilde{\pi}^N(\theta^*, Z^*)}{\tilde{\pi}^N(\theta, Z)} \cdot \frac{Q_\theta^N(Z) q(\theta|\theta^*)}{Q_\theta^N(Z^*) \cdot q(\theta^*|\theta)},$$

where $\tilde{\pi}^N(\theta, Z) := \frac{1}{N}\sum_{k=1}^{N} \pi(\theta, z(k)) \cdot \prod_{i\neq k} q_\theta(z(i))$, $Q_\theta^N(Z) = \prod_{i=1}^{N} q_\theta(z(i))$.

That is, $\tilde{r}^N(\theta^*, \theta)$ is the acceptance ratio of a bona fide Metropolis-Hastings algorithm with target $\tilde{\pi}^N(\theta, Z)$ and proposal $Q_\theta^N(Z) \cdot q(\theta|\nu)$, where ν is the previous draw of the chain in Θ.

Of course, the joint posterior is a contrived and perhaps rather awkward distribution. The kicker, however, is that *we sample from the exact marginal $\pi(\theta)$*. Indeed,

$$\int_Z \tilde{\pi}^N(\theta, Z) dZ = \frac{1}{N}\sum_{k=1}^{N} \int_Z \pi(\theta, z(k)) \cdot \prod_{i\neq k} q_\theta(z(i)) dZ$$

$$= \frac{1}{N}\sum_{k=1}^{N}\int_{z(k)} \pi(\theta, z(k))dz(k) \cdot \int_{Z^{-k}} \prod_{i\neq k} q_\theta(z(i))dZ^{-k}$$

$$= \frac{1}{N}\sum_{k=1}^{N}\pi(\theta) \cdot \prod_{i\neq k}\int_{z(i)} q_\theta(z(i))dz(i) = \pi(\theta),$$

where we used $\int_{z(i)} q_\theta(z(i))dz(i) = 1$.

Appendix B: Mathematical Supplement for Lecture 0

Econometricians are used to dealing with positive semidefinite matrices, such as covariance matrices. The eigen structure of such matrices is nicely characterized by the decomposition of the spectral theorem, with each eigenvalue (allowing for multiplicity) corresponding to an eigenvector, and these eigenvectors forming a basis for the space the operator/matrix of interest acts on.

When dealing with discrete markov chains, the econometrician must deal with nonsymmetric matrices, and needs to understand its eigenstructure to proceed with investigations such as the analysis of nonasymptotic rates of convergence. The spectral theorem does not obtain anymore, in particular there are not "enough" eigenvectors, leaving some eigenvalues are left "unclaimed", and a generalization of the concept of eigenvectors is required. In this suppleent, we develop the pertinent linear algebra results.

The path to generalizing eigenvectors is through invariant subspaces.

Recall that for $T \in \mathcal{L}(V)$, and $U \subset V$ be a subspace. If $T|_U$ maps U into itself (i.e., is an operator on U), then we say that U is **invariant** under T. We will be looking for decompositions of V into invariant subspaces, generally speaking

$$V = U_1 \oplus \cdots \oplus U_m.$$

Such a decomposition, where each U_j is a one dimensional subspace of V invariant nuder T is possible *if and only if* V has a basis consisting of eigenvectors of T. This happens *if and only if*

$$V = \operatorname{null}(T - \lambda_1 I) \oplus \cdots \oplus \operatorname{null}(T - \lambda_m I),$$

where $\lambda_1, ..., \lambda_m$ are distinct.

The **goal** is to show that if V is a complex vector space and $T \in \mathcal{L}(V)$, then

$$V = \operatorname{null}(T - \lambda_1 I)^{\dim V} \oplus \cdots \oplus \operatorname{null}(T - \lambda_m I)^{\dim V},$$

where $\lambda_1, ..., \lambda_m$ are distinct. The $(T - \lambda_j I)$ are nilpotent operators on subspaces of V, and the nullspace of their power will identify the generalization of eigenvectors "claiming" the eigenvalues.

For $T \in \mathcal{L}(V)$, and λ an eigenvalue of T, $v \in V$ is a **generalized eigenvector** of T corresponding to λ if

$$(T - \lambda I)^j = 0,$$

for some integer j.

Note that $\text{null}(T-\lambda I)^j = \text{null}(T-\lambda I)^{\dim V}$ for all $j > \dim V$. Therefore, the set of generalized eigenvectors of T corresponding to λ equals $\text{null}(T-\lambda I)^{\dim V}$.

Recall the matrix decompositon result (analogous to the spectral decomposition) allowing to "read out" eigenvalues and their multiplicities in the non-symmetric case. A number λ appears on the diagonal of an upper-triangular matrix of T precisely $\dim \text{null}(T-\lambda I)^{\dim V}$ times.

See Alder (1997) 5.12, 5.13, 5.18 for upper-triangular representations.

We can obtain a sparser representation. The key operational result is the following.

Theorem Let $T \in \mathcal{L}(V)$, V a complex vector space. Let λ_1,...,λ_m be distinct eigenvalues of T, and $U_1, \ldots, U_m$ the corresponding subspaces of generalized eigenvectors. Then

a. $V = U_1 \oplus \cdots \oplus U_m$

b. each U_j is invariant under T

c. each $(T-\lambda_j I)|_{U_j}$ is nilpotent.

See 2.9 to go from sum to direct sum.

This delivers a nice, sparse matrix representation for the operator T.

Theorem Let $T \in \mathcal{L}(V)$, V a complex vector space. Let λ_1,...,λ_m be distinct eigenvalues of T. Then there exists a basis of V with respect to which T has block diagonal matrix of the form

$$\begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{bmatrix},$$

where each A_j is an upper-triangular matrix of the form

$$\begin{bmatrix} \lambda_j & & * \\ & \ddots & \\ 0 & & \lambda_j \end{bmatrix}.$$

This is a very nice and sparse matrix representation for T. We can in fact get zeros everywhere except on the diagonal and the line directly above.

Suppose N is nilpotent. For each nonzero $v \in V$, let $m(v)$ be the largest nonegstive integer such that $N^{m(v)}v \neq 0$.

Lemma Take $N \in \mathcal{L}(V)$, and assume N is nilpotent. Then there exist v_1,...,v_k such that

a. $(v_1, Nv_1, ..., N^{m(v_1)}, ..., v_k, Nv_k, ..., N^{m(v_k)}v_k)$ is a basis of V

b. $(N^{m(v_1)}v_1, ..., N^{m(v_k)}v_k)$ is a basis of null$N$.

A basis v is called a **Jordan basis** for T if with respect to this basis T is has a block

diagonal structure

$$\begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{bmatrix},$$

where each A_j is an upper-triangular matrix of the form

$$\begin{bmatrix} \lambda_j & 1 & & 0 \\ & \lambda_j & 1 & \\ & & \ddots & 1 \\ 0 & & & \lambda_j \end{bmatrix}.$$

λ_j an eigenvalue of T. See 5.18.

Theorem Let $T \in \mathcal{L}(V)$, V a complex vector space. Then there is a basis of V that is a Jordan basis.

References

Axler, Sheldon Jay. *Linear Algebra Done Right.* Vol. 2. New York: Springer, 1997.