Plan MCMC and General State Chains

0. A few notational notes
1. Markov Chain Monte Carlo (MCMC) for finding posteriors
2. General-state MC theory: similarities and differences
3. Back to MCMC

Next Time: Spring Break

Midterm Exam: Tuesday 28 March in class
Homework 5 (sample exam) coming soon

Minor Notes on Notation 1.

- Comment: The set D in the state decomposition $D \cup \bigcup C_r$ is a countable union of transient sets. Such a set is called *dissipative*, which explains the "D."
- A measure ν on a set $\mathcal{S}$ is σ-finite if $\mathcal{S}$ can be written as a countable union of ν-finite sets. That is, $\mathcal{S} = \bigcup_i F_i$ with $\nu(F_i) < \infty$. Examples? Cases where it does not hold? The "σ" in this term and in σ-field refers to countable unions (think σ for sum).
- If P_1 and P_2 are two probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$, then we can define a distance metric, called total variation distance, by

$$d_{\text{TV}}(P_1, P_2) = \sup_{A \in \mathcal{F}} |P_1(A) - P_2(A)|. \tag{1}$$

- If ρ is a measure on a set $\mathcal{S}$, then for any suitable function h on $\mathcal{S}$, we can write

$$\rho h = \int_{\mathcal{S}} h(s)\rho(ds). \tag{2}$$

This should evoke the matrix-vector multiplication. In particular, for a Markov transition kernel $P(s, A)$, recall that we have

$$(\rho P)(A) = \int P(s, A)\rho(ds) \tag{3}$$

$$(Ph)(s') = \int h(s')P(s, ds'). \tag{4}$$

A Problem 2. Consider a statistical model where the data $\boldsymbol{Y}$ are drawn from one of the distributions in the collection $\{f_\theta : \theta \in \mathcal{H}\}$. Each f_θ is a probability distribution indexed by a *parameter* θ from some *parameter space* $\mathcal{H}$.

In Bayesian inference, we use probability as a calculus of uncertainty. We put a *prior distribution* on the unknown parameter, treating it as a random variable Θ. The *posterior distribution*, the conditional distribution of Θ given $\boldsymbol{Y}$, updates our uncertainties about the parameter and contains all the information we need for inference.

Suppose $f_\theta(y)$ is the conditional density of $Y \mid \Theta \text{ near } \theta$, and the prior p is the marginal distribution of Θ. We want to update our uncertainty for Θ in light of the observed data Y. This can be expressed by the *posterior distribution* π:

$$\pi(A) = \mathsf{P}\{\Theta \in A \mid Y \text{ near } y\} \tag{5}$$

$$= \frac{\mathsf{P}\{\Theta \in A \text{ and } Y \text{ near } y\}}{\mathsf{P}\{Y \text{ near } y\}} \tag{6}$$

$$= \frac{\int_{\theta \in A} \mathsf{P}\{Y \text{ near } y \text{ and } \Theta \text{ near } \theta\}}{\int_{\theta \in \mathcal{H}} \mathsf{P}\{Y \text{ near } y \text{ and } Theta \text{ near } \theta\}} \tag{7}$$

$$= \frac{\int_{\theta \in A} \mathsf{P}\{Y \text{ near } y \mid \Theta \text{ near } \theta\}\, \mathsf{P}\{\Theta \text{ near } \theta\}}{\int_{\theta \in \mathcal{H}} \mathsf{P}\{Y \text{ near } y \mid Theta \text{ near } \theta\}\, \mathsf{P}\{\Theta \text{ near } \theta\}} \tag{8}$$

$$= \frac{\int_{\theta \in A} f_\theta(y) p(d\theta)}{\int_{\theta \in \mathcal{H}} f_\theta(y) p(d\theta)}. \tag{9}$$

If we take π and p to be densities (i.e., $\pi(d\theta) = \pi(\theta)d\theta$ and $p(d\theta) = p(\theta)d\theta$), then we have

$$\pi(\theta) = \frac{f_\theta(y)p(\theta)}{\int_{\psi \in \mathcal{H}} f_\theta(y) p(d\psi)}. \tag{10}$$

This is straightforward to express but in general is difficult to calculate because the normalizing constant in the numerator is hard to compute.

One Approach 3. Importance Sampling

Suppose we find a probability density g on $\mathcal{H}$ that approximates the density π to some degree and for whic Let $\Theta_1, \Theta_2, \ldots$ be a random sample drawn from g.

Then for any measurable function h on $\mathcal{H}$,

$$\frac{1}{n}\sum_{k=1}^{n} h(\Theta_i) \frac{f_{\Theta_k}(y)p(\Theta_k)}{g(\Theta_k)} \to \mathsf{E}_g h(\Theta) \frac{f_\Theta(y)p(\Theta)}{g(\Theta)} \tag{11}$$

$$= \left(\int_{\psi \in \mathcal{H}} f_\theta(y) p(d\psi)\right) \mathsf{E}_\pi h(\Theta). \tag{12}$$

Taking $h \equiv 1$ gives us a way to estimate the normalizing constant.

Much research has gone into Importance Sampling, but he short story is that in general, it is difficult to get right. One reason is that the tails of g must be at least as thick as the tails of π if the estimate is to be stable.

A Better(?) Approach 4. Markov Chain Monte Carlo

Here's the idea: Create a Markov chain with limiting distribution π. Run the chain and then (after a while to achieve equilibrium), read off the values as a sample from π. From that we can estimate the distribution or any functional thereof.

Issues:

A. If we can't compute π, how do we make a chain that converges to it?
B. What conditions on the chain do we need to make this work?
C. When is equilibrium reached to sufficient approximation?
D. How accurate are the approximations derived from the chain, given especially that the samples are dependent?
E. And wait, this will not usually be on a countable state space, does what we know still work?

Comment 5. I'm going to assume that everything in site has a density over the line and that $\pi(x) > 0$ everywhere. All this can be done more carefully using weaker conditions and other base measures.

Algorithm 6. Metropolis-Hastings

Let Q be a probability transition kernel defined by

$$Q(x, dy) = q(x, y)dy. \tag{13}$$

Define a Markov chain $(X_n)_{n\geq 0}$ with transition probabilities

$$P(x, dy) = p(x, y)dy + r(x)\delta_x(dy) \tag{14}$$

where

$$p(x, y) = \begin{cases} q(x, y)\alpha(x, y) & \text{if } y \neq x \\ 0 & \text{if } y = x \end{cases} \tag{15}$$

$$r(x) = 1 - \int p(x, t)dt, \tag{16}$$

and where

$$\alpha(x, y) = \begin{cases} \min\left\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\} & \text{if } \pi(x)q(x, y) > 0 \\ 0 & \text{if } \pi(x)q(x, y) = 0. \end{cases} \tag{17}$$

Describe the behavior of this chain? What does $\alpha(x, y)$ represent?

Notice that the transition probabilities are defined only in terms of ratios of π, so the unknown normalizing constant cancels. This gives one answer to question A above.

Notice first that

$$\pi(x)p(x, y) = \pi(y)p(y, x). \tag{18}$$

(This implies the reversibility condition that you read about.)

Now, we can show that π is an invariant distribution for the chain. For a measurable set $A \subset \mathcal{H}$:

$$\pi P(A) = \int P(x, A)\pi(dx) \tag{19}$$

$$= \int \left[\int_{y\in A} p(x, y)dy\right] \pi(x)dx + \int r(x)\delta_x(A)\pi(x)dx \tag{20}$$

$$= \int \left[\int_{y \in A} \pi(x) p(x,y) dy \right] dx + \int_{x \in A} r(x) \pi(x) dx \tag{21}$$

$$= \int \left[\int_{y \in A} \pi(y) p(y,x) dy \right] dx + \int_{x \in A} r(x) \pi(x) dx \tag{22}$$

$$= \int_{y \in A} \left[\int p(y,x) dx \right] \pi(y) dy + \int_{x \in A} r(x) \pi(x) dx \tag{23}$$

$$= \int_{y \in A} (1 - r(y)) \pi(y) dy + \int_{x \in A} r(x) \pi(x) dx \tag{24}$$

$$= \int_{y \in A} \pi(y) dy \tag{25}$$

$$= \pi(A). \tag{26}$$

Examples 7. A few Metropolis-Hastings Chains

1. Independence chains: $q(x,y) = f(y)$ for some density f
2. Random Walk chains: $q(,y) = f(y-x)$ for some density f.
3. Symmetrix Candidate distribution: $q(x,y) = q(y,x)$

Algorithm 8. Gibbs Sampling

Suppose X has distribution π and h is a function on $\mathcal{H}$. If $Y = h(X)$, then define

$$P(x,A) = \mathsf{P}\{ X \in A \mid Y \text{ near } h(x) \}. \tag{27}$$

We sample X_{n+1} from the conditional distribution of $X \mid Y = h(X_n)$. This produces a Markov chain $(X_n)_{n \geq 0}$.

Notice that π is an invariant distribution for the chain.

$$\pi P(A) = \int_{x \in \mathcal{H}} P(x,A) \pi(dx) \tag{28}$$

$$= \int_{x \in \mathcal{H}} \int_{t \in A} P(x,dt) \pi(dx) \tag{29}$$

$$= \int_{x \in \mathcal{H}} \int_{\substack{t \in A \\ h(t) = h(x)}} P(x,dt) \pi(dx) \tag{30}$$

$$= \int_{x \in \mathcal{H}} \int_{\substack{t \in A \\ h(t) = h(x)}} \mathsf{P}\{ X \text{ near } t \mid Y \text{ near } h(x) \} \, \mathsf{P}\{ X \text{ near } x \} \tag{31}$$

$$= \int_{x \in \mathcal{H}} \int_{\substack{t \in A \\ h(t) = h(x)}} \mathsf{P}\{ Y \text{ near } h(t) \mid X \text{ near } t \} \frac{\mathsf{P}\{ X \text{ near } x \}}{\mathsf{P}\{ Y \text{ near } h(x) \}} \mathsf{P}\{ X \text{ near } t \} \tag{32}$$

$$= \int_{x \in \mathcal{H}} \int_{\substack{t \in A \\ h(t) = h(x)}} \frac{\mathsf{P}\{ X \text{ near } x, Y \text{ near } h(x) \}}{\mathsf{P}\{ Y \text{ near } h(x) \}} \mathsf{P}\{ X \text{ near } t \} \tag{33}$$

$$= \int_{x \in \mathcal{H}} \int_{\substack{t \in A \\ h(t) = h(x)}} \frac{\mathsf{P}\{ X \text{ near } x, Y \text{ near } h(t) \}}{\mathsf{P}\{ Y \text{ near } h(t) \}} \mathsf{P}\{ X \text{ near } t \} \tag{34}$$

$$= \int_{x \in \mathcal{H}} \int_{\substack{t \in A \\ h(t) = h(x)}} P(t,dx) \pi(dt) \tag{35}$$

$$= \int_{t\in A} \int_{\substack{x\in\mathcal{H} \\ h(x)=h(t)}} P(t, dx)\pi(dt) \tag{36}$$

$$= \int_{t\in A} \pi(dt) \tag{37}$$

$$= \pi(A). \tag{38}$$

In general, $P(x, A)$ will not be irreducible, but we can find a set of functions $h_1, ldots, h_m$ with corresponding kernels $P_1, \ldots, P_m$ and then construct a new kernel by choosing one at random at each stage or cycling between them.

The Gibbs sampler uses this strategy with

$$h_i(x) \equiv h_i(x_1, \ldots, x_m) = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m). \tag{39}$$

How does this chain operate?

A Quick Tour of General State Markov Chains

Tour Stop 9. Irreducibility

Irreducibility in the general state case cannot define a relation like $\leftrightarrow$ because the probability of hitting particular states might be zero.

The replacement notion is in some sense even simpler.

Definition. Let ϕ be a measure on the state space $\mathcal{S}$ (now equipped with a σ-field of measurable sets) of a general-state Markov chain X. Then X is ϕ-irreducible if whenever $\phi(A) > 0$, then $R(s, A) > 0$ for all $s \in \mathcal{S}$.

Equivalently, X is ϕ-irreducible if there exists an $n > 0$ (possibly depending on A and s) such that $P^n(s, A) > 0¿$

This tells us that "big" sets (as measured by ϕ are reached from any starting point, precluding "reducible" type behavior.

But what about the reverse? If $\phi(B) = 0$, will the chain avoid B? (Consider a countable-state chain and $\phi = \delta_{s_0}$; such that $s \to s_0$ for every s. That is weaker than what we expect from the countable state case.)

So we want "bigger" measures. In fact, we can find a maximal such measure.

Theorem. If X is ϕ-irreducible for some measure ϕ, then there exists a probability measure ψ such that

1. X is ψ-irreducible.
2. For any measure ν, X is ν irreducible if and only if $\psi(A) = 0$ implies $\nu(A) = 0$ for all A.
3. If $\psi(A) = 0$, then $\psi\{s\colon R(s, A) > 0\} = 0$.

   We have then that if $\psi(A^c) = 0$, then A contains an absorbing set and that any absorbing set A has $\psi(A^c) = 0$.

   Such a chain is called ψ-irreducible.

   Decompositions exactly like the one we found in the countable-state space do not exist, but there are several approximations that are quite close.

Tour Stop 10. Recurrence and Transience
A ψ-irreducible chain is recurrent if $O(s, A) = \infty$ for all s and A such that $\psi(A) > 0$. Otherwise, the chain is transient.
A recurrent chain satisfies $\mathsf{P}_s\{X_n \in A i.o.\} = 1$ for all s in a set of ψ probability 1.
A stronger property is *Harris recurrence* which says that $\mathsf{P}_s\{X_n \in A i.o.\} = 1$ for all $s \in \mathcal{S}$. Much more convenient. A recurrent chain is Harris recurrent if and only if every bounded V with $\Delta V = 0$ is constant, where Δ is the drift operator.
There are also a variety of drift criteria similar to what we've seen.

Tour Stop 11. Periodicity
A ψ-irreducible chain is aperiodic if there do not exist the cyclic classes that we've seen. If there exist d cyclic classes (with the remainder of the space ψ-null), then the chain has period d.

Tour Stop 12. Invariant Measures
Define just as before $\pi P = \pi$, but of course, meaning of πP is as above not a matrix-vector multiplication.
Theorem Let π be a probability measure on $\mathcal{S}$. If P is ϕ-irreducible for some measure ϕ on $\mathcal{S}$ and if $\pi P = \pi$, then P is π-irreducible, positive recurrent, and π is the unique invariant distribution of the chain.
If P is also aperiodic, then for all s outside a set of π-measure 0,

$$d_{\mathrm{TV}}(\pi, P^n(s, \cdot)) \to 0. \tag{40}$$

If the chain is also Harris recurrent, then this convergence occurs for all s.

Tour Stop 13. Atoms and Small Sets
The discrete states in countable state chain have no exact analogue in the general-state case. But we can find sets of states that act much the same way. Atoms and small sets are two important kinds. The definitions of these sets rely on finding lower bounds on $P^n(s, \cdot)$ by some measure for s in the set.