A textbook proof of geometric ergodicity.

Sam Livingstone, Heiko Strathmann

July 27, 2015

Todo list

SJL: the point is you want to write P as a mixture, so you need the mixture weights ϵ and $1-\epsilon$ such that they sum to 1. You can't take ϵ to be zero otherwise the chains will never couple. You can have them exactly 1 but this would mean you are just drawing independent samples, as $P(x,\cdot)=\nu(\cdot)$ then. 2

HS: This is by construction with the two cases right? SJL: Yes exactly 2

1 Notation and Concepts

- We use upper-case letters for random variables X and events $A \in \mathcal{B}$. Lower-case values represent values of random variables, i.e. $X = x$ means the random variable X takes the value x.
- Denote by $\mathbb{P}(X_{i+1} \in A | X_i = x)$ the probability of the event $X_{i+1} \in A$ given that $X_i = x$.
- **Total variation distance**
$$\|\mu(\cdot) - \nu(\cdot)\|_{\mathrm{TV}} := \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)|$$
is the total variation distance between $\mu(\cdot)$ and $\nu(\cdot)$. Informally, this is the largest possible difference between the probabilities that $\mu(\cdot)$ and $\nu(\cdot)$ can assign to the same event.
- **Irreducibility**. When $\mathcal{X}$ is countable, a Markov chain $\{X_t\}_{t\geq 0}$ is called *irreducible* if for any $x, y \in \mathcal{X}$ there exist $n = n(x,y)$ and $m = m(y,x)$ such that $P^n(x,y) > 0$ and $P^m(y,x) > 0$. In the uncountable case, the chain is called φ-irreducible for some measure $\varphi(\cdot)$ if for any $x \in \mathcal{X}$ and any $A \in \mathcal{B}$ there is an $n = n(x,A)$ such that $P^n(x,A) > 0$ whenever $\varphi(A) > 0$.
- **Aperiodicity**. When $\mathcal{X}$ is countable, the period p_x of a state x of a Markov chain is defined as
$$p_x = \gcd\{n \in \mathbb{N} : P^n(x,x) > 0\},$$
Fwhere gcd$A$ denotes the greatest common divisor of the set A. For irreducible chains all states have the same period []. The chain is called *aperiodic* if $p_x = 1$. In the uncountable case, it suffices to note that any chain with a non-zero probability of remaining at the current position is aperiodic (so any Metropolis–Hastings algorithm) (for a full definition see []).

2 Geometric Ergodicity of Markov chains

We describe the Markov chain $\{X_t\}_{t\geq 0}$ on the measurable space $(\mathcal{X}, \mathcal{B})$ through a starting point $X_0 = x$ and a transition kernel $P : \mathcal{X} \times \mathcal{B} \to [0,1]$. For each $x \in \mathcal{X}$, $P(x,\cdot)$ defines a probability measure where $P(x,A) = \mathbb{P}(X_{i+1} \in A | X_i = x)$ for any $A \in \mathcal{B}$. We will also use the shorthand $X_{i+1} \sim P(X_i,\cdot)$ rather than writing 'If $X_i = x_i$ for any $x_i \in \mathcal{X}$ then $X_{i+1} \sim P(x_i,\cdot)$.' The n-step kernel $P^n(x,A)$ is defined similarly for X_{i+n}. We say $\pi(\cdot)$ is an invariant distribution of P if $\int P(x,A)\pi(dx) = \pi(A)$. We assume for

the purpose of this note that each $\pi(\cdot)$ admits a Lebesgue density, and write $\pi(dx) = \pi(x)dx$. The chain is called *geometrically ergodic* if

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{\text{TV}} = \mathcal{O}(\rho^n) \tag{1}$$

for some $\rho < 1$. In words, $P^n(x,\cdot)$ approaches $\pi(\cdot)$ at a geometric rate in n. As the name suggests, a kernel $P(x,\cdot)$, can *only* converge to a $\pi(\cdot)$ which is invariant for that kernel.

3 Establishing Geometric Ergodicity

In fact, all Markov chains with a finite state space ($|\mathcal{X}| < \infty$) which are irreducible and aperiodic have a unique invariant measure [**?**], and are geometrically ergodic. A straightforward way of establishing this is through the *coupling inequality*

$$\|\mu(\cdot) - \nu(\cdot)\|_{\text{TV}} \leq \mathbb{P}(X \neq Y), \tag{2}$$

where (X, Y) is some pair of random variables such that marginally $X \sim \mu(\cdot)$ and $Y \sim \nu(\cdot)$ (the proof is straightforward, see e.g. [**?**]). We can use the coupling inequality to establish (1) by constructing two Markov chains $\{Y_t\}_{t\geq 0}$ and $\{X_t\}_{t\geq 0}$ such that (i) $Y_n \sim \pi(\cdot)$ for all n, and (ii) the upper bound on the TV, $\mathbb{P}(X_n \neq Y_n)$, decreases geometrically with n.

Imagine that we can write P as a mixture, i.e. assume there exists a measure $\nu(\cdot)$, a transition kernel $R(x, A)$, and a $0 < \epsilon \leq 1$,

$$P(x,\cdot) = \epsilon\nu(\cdot) + (1-\epsilon)R(x, A). \tag{3}$$

In fact, this 'splitting' technique can be done *in certain regions of* $\mathcal{X}$ whenever the Markov chain is recurrent (see below) []. While we avoid a full defnition, any Markov chain which is π-irreducible and π-invariant is recurrent (so these are necessary requirements for $P(x,\cdot)$ to converge to $\pi(\cdot)$). The following procedure gives such a bound (assuming $\pi(\cdot)$ is the limiting distribution for the chain)

1. Start with $Y_0 \sim \pi(\cdot)$ and $X_0 = x$
2. For each iteration n of the chain, sample $b_n \sim \texttt{Bernoulli}(\epsilon)$.
   - If $b_n = 1$ then sample $Y_n \sim \nu(\cdot)$ and set $X_n = Y_n$, and then make the chains take equal values forever more. We say the chains 'couple'.
   - If $b_n = 0$ then sample $X_{n+1} \sim R(X_n,\cdot)$ and $Y_{n+1} \sim R(Y_n,\cdot)$ independently.

Note two things

- $\mathbb{P}(X_n \neq Y_n) = (1-\epsilon)^n$. For $n = 0$, we clearly have $\mathbb{P}(X_0 \neq Y_0) = 1$ as $\pi(\cdot)$ is continuous (since it admits a density) and $X_0 = x$. For $n = 1$, since the probability of 'coupling' the chains is ϵ and they have zero probability of hitting the same value otherwise (since they are both continuous and in this case independent random variables), we have $\mathbb{P}(X_1 \neq Y_1) = \mathbb{P}(X_0 \neq Y_0)\mathbb{P}(\epsilon = 0) = 1 - \epsilon$. And so on.
- At each iteration $Y_{n+1}|Y_n$ is marginally sampled from P, so $Y_{n+1} \sim \pi(\cdot)$ for all n.

This means that at the point at which the chains 'couple', we have $X_n \sim \pi(\cdot)$. The kernel $R(x,\cdot)$ represents the part of $P(x,\cdot)$ in which there is depedence on the current point x, and the independent measure $\nu(\cdot)$ represents the part of $P(x,\cdot)$ that can be written as independent of x. The key is to break the dependence on the past, so that the chain can 'regenerate' and hence be coupled with another. This approach was first introduced by Doeblin (1936), and (3) is sometimes called the Doeblin condition. It can be extended to the case where P^m can be decomposed as in (3) for some fixed m straightforwardly.

On unbounded state spaces, however, it is often difficult to satisfy the Doeblin condition with a uniform ϵ for all $x \in \mathcal{X}$. Instead we tend to establish (3) for any $x \in C$, where C is called a 'small set'. Under

SJL: the point is you want to write P as a mixture, so you need the mixture weights ϵ and $1-\epsilon$ such that they sum to 1. You can't take ϵ to be zero otherwise the chains will never couple. You can have them exactly 1 but this would mean you are just drawing independent samples, as $P(x,\cdot) = \nu(\cdot)$ then.

HS: This is by construction with the two cases right? SJL: Yes exactly

mild conditions any compact set is small. To establish a geometric bound here entails showing (3) inside C, together with another condition, the stipulation that

$$\tau_C := \inf\{t \geq 1 : X_t \in C\} \tag{4}$$

the return time to C when leaving it, follows a distribution with geometric tails. In fact, Meyn & Tweedie showed that this is equivalent to the condition that there exists a function $V : \mathcal{X} \to [1, \infty)$ such that

$$\int V(y)P(x, dy) \leq \lambda V(x) + b1_C(x), \tag{5}$$

or some $\lambda < 1$, $b < \infty$, with V called a *Lyapunov* function. Intuitively, $\{V(X_t)\}_{t\geq 0}$ can be thought of as a one-dimensional projection of the chain. The condition effectively states that to establish 4 we only need look at the one-dimensional projections of $\{X_t\}_{t\geq 0}$.

Roberts & Tweedie [] further simplified matters by showing that if all compact sets are small then we need not explicitly find a C, but instead show that

$$\limsup_{|x|\to\infty} \int \frac{V(y)}{V(x)} P(x, dy) < 1. \tag{6}$$

Effectively, showing (6) establishes geometric ergodicity. In the case where P is a Metropolis–Hastings kernel, then we can equivalently write

$$\limsup_{|x|\to\infty} \int \left[\frac{V(y)}{V(x)} - 1\right] \alpha(x, y)Q(x, dy) < 0, \tag{7}$$

where Q is the proposal kernel and α the acceptance rate. The skill in establishing the result in a given scenario is find a suitable way to bound α and choosing an appropriate V such that (7) can be established.

4 Geometric Ergodicity of the Random Walk Metropolis in 1D

For the Random Walk Metropolis the kernel choice is such that $Q(x, dy) = q(|x - y|)dy$, meaning $\alpha(x, y) = 1 \wedge \pi(y)/\pi(x)$. Since the acceptance rate is just the ratio of target densities, it lends itself quite nicely to a simple bound. If we assume $\pi(x)$ is log-concave in the tails, then $\pi(y)/\pi(x) \leq \exp(-a(|y| - |x|))$ for large enough x. With this, a sensible choice of Lyapunov function would seem to be $V(x) = e^{s|x|}$, for some $0 < s < a$. Let's consider the positive tail, i.e. the case $x \to \infty$. In this instance we can re-write the integral in (7) as

$$\begin{aligned}\int_{-\infty}^{0} [e^{s(|y|-x)} - 1]\alpha(x, y)Q(x, dy) + \int_{0}^{x} [e^{s(y-x)} - 1]\alpha(x, y)Q(x, dy) \\ + \int_{x}^{2x} [e^{s(y-x)} - 1]\alpha(x, y)Q(x, dy) + \int_{2x}^{\infty} [e^{s(y-x)} - 1]\alpha(x, y)Q(x, dy).\end{aligned}$$

In the first and last terms can be made arbitrarily small by taken x large. In the first case this is because

$$\int_{-\infty}^{0} [e^{s(|y|-x)} - 1]\alpha(x, y)Q(x, dy) \leq \int_{-x}^{0} [e^{s(|y|-x)} - 1]Q(x, dy) + \int_{-\infty}^{-x} [e^{s(|y|-x)} - 1]e^{-a(|y|-x)}Q(x, dy),$$

in the latter case because $\alpha(x, y) \leq e^{-a(|y|-|x|)}$ for $|y| \geq |x|$. The first integral on the right-hand side is strictly negative for any x and the second is bounded above by $Q(x, (-\infty, -x))$, which will clearly become negligibly small as x grows. For the last term, we can again use the log-concave restriction to bound the integral with

$$\int_{2x}^{\infty} [e^{(s-a)(y-x)} - e^{-a(y-x)}]Q(x, dy) \leq Q(x, (2x, \infty)) \to 0.$$

This leaves the middle two terms. We can combine these by writing $y = x + Z$, for $Z \sim \mu(\cdot)$, meaning $\mu(\cdot)$ denotes zero mean proposal 'increment' distribution. Typically $\mu(\cdot)$ might be a zero mean Gaussian, if $Q(x, \cdot) = \mathcal{N}(x, h\sigma^2)$. We can then bound the middle two integrals with

$$\int_0^x [e^{-sz} - 1 + e^{(s-a)z} + e^{-az}]\mu(dz) = -\int_0^x (1 - e^{(s-a)z})(1 - e^{-sz})\mu(dz), \tag{8}$$

which is strictly negative. Since for large x the entire integral will be comprised of terms which can be made arbitrarily small and terms which are strictly negative, this establishes (7) as $x \to \infty$, and the log-concave restriction means an equivalent argument holds as $x \to -\infty$.

5 Geometric Ergodicity of KMC-lite in 1D

5.1 Notation

If x denotes the current position in the chain, we denote the next candidate move as y. We can write for the KMC lite kernel

$$y(x, p) = x + c(x, p) + P, \quad P \sim \mathcal{N}(0, 1),$$

where $c(x, p)$ is basically a sequence of gradient steps, which depends on the current point x and the random variable P. For brevity, we write $y(p) = x + c(x) + P$. We also write $\mu(\cdot)$ to denote a standard Gaussian measure.

5.2 Extra Assumptions

We assumed that $c(x) \xrightarrow{p} 0$ as $|x| \to \infty$, which seems valid given the KMC-lite kernel, and that $c(x) < M$ for all $x \in \mathcal{X}$.

We also assumed the conditions of Theorem 2.2 in []. These are technical conditions, but well-known among the Markov chain community. They are extremely loose, and assert that

- The Markov chain is μ^L-irreducible, where $\mu^L(\cdot)$ denotes Lebesgue measure, and aperiodic
- All compact sets are small

This means that we can use (6) to establish geometric ergodicity here too. There is not a direct proof of this assumption, but noone would question that it is true in the cases of interest to us (I actually have a rough proof which will go in the geometric ergodicity of HMC paper we are writing, but the paper is still in progress at the moment).

Theorem 2.2 (Roberts & Tweedie). *Assume that $\pi(\cdot)$ and Q admit densities $\pi(x)$ and $q(y|x)$, that $\pi(x)$ is bounded away from 0 on compact sets, and that there are ϵ_q and delta$_q$ such that*

$$|x - y| \leq \delta_q \implies q(y|x) \geq \epsilon_q.$$

Then the Metropolis–Hastings chain with proposal kernel Q is μ^L-irreducible and aperiodic, and every compact set is small.

5.3 Details

We extend the Random Walk result, using the intuition the KMC-lite approaches a random walk in the tails.

The result can straightforwardly be adjusted to an integral over any set $(-x^\delta, x^\delta)$ for some $\delta \in (0, 1]$. This will be crucial to the approach here, as the entire set $(x - x^\delta, x + x^\delta)$ can be pushed arbitrarily far into the tails of the distribution for $\delta < 1$, whereas $(0, 2x)$ always contains the centre of the space. So denoting

$$I_a^b := \int_a^b \left[e^{s(|y|-|x|)} - 1\right] \alpha(x, y) Q(x, dy),$$

we divide the integral of interest into

$$I_{-\infty}^{x-x^\delta} + I_{x-x^\delta}^{x+x^\delta} + I_{x+x^\delta}^{\infty}.$$

We can also write (6) as an integral with respect to the Gaussian measure $\mu(\cdot)$ rather than $Q(x,\cdot)$, as P is the only random variable, giving

$$\int \left[e^{s(|y(p)|-|x|)} - 1\right] \alpha(x, y(p))\mu(dp)$$

The crux of the proof is basically to show that in the set $y \in (x - x^\delta, x + x^\delta)$, this integral can be made arbitrarily close to (8), which is strictly negative (meaning this integral will be too). Specifically, on this set the integral reduces to

$$\int \left[e^{s(c(x)+p)} - 1\right] \alpha(x, x + c(x) + p)\mu(dp). \tag{9}$$

Note that $c(x)$ can be made arbitrarily small for any choice of $p \in (-x^\delta, x^\delta)$, and also that therefore $\alpha(x, y(p))$ can be made arbitrarily close to $1 \wedge \pi(y(p))/\pi(x)$. Because of these, we can say that (9) can be made arbitrarily close to (8) on this set, and hence will be strictly negative for large enough x.

For proposals y outside of the set $(x - x^\delta, x + x^\delta)$, then in the tails the two integrals $I_{-\infty}^{x-x^\delta}$ and $I_{x+x^\delta}^{\infty}$ can be made arbitrarily close to zero because $\mu(dp)$ is a Gaussian. Specifically, the dominant term in the integrals will be the Gaussian density $e^{-(y-x)^2}$, and for large enough x this will be $\mathcal{O}(e^{-x^{2\delta}})$ or smaller. Since the Lyapunov function term is $\mathcal{O}(e^{x^\delta})$ by the same argument, then the combination will be $\mathcal{O}(e^{x^\delta - x^{2\delta}})$, so the integrals can be made arbitrarily small by choosing x large enough, for any $\delta > 0$. Note that the assumption that $c(x) < M$ for all x is important here for these order based arguments to be valid.

5.4 Proof in paper

I am sure this is not the most elegant way to make the above arguments precise, but I was a little pushed for time before submission. Here is the proof in the paper reproduced in full for convenience. Hopefully makes a bit more sense after the above explanation.

Notation Denote by $\alpha(x_t, x^*(p'))$ is the probability of accepting a (p', x^*) proposal at state x_t. Let $a \wedge b = \min(a, b)$. Define $c(x^{(0)}) := \epsilon^2 \sum_{i=0}^{L-1} \nabla f(x^{(i\epsilon)})/2$ and $d(x^{(0)}) := \epsilon(\nabla f(x^{(0)}) + \nabla f(x^{(L\epsilon)}))/2 + \epsilon \sum_{i=1}^{L-1} \nabla f(x^{(i\epsilon)})$, where $x^{(i\epsilon)}$ is the i-th point of the leapfrog integration from $x = x^{(0)}$.

We assumed $\pi(x)$ is log-concave in the tails, meaning $\exists x_U > 0$ s.t. for $x^* > x_t > x_U$, we have $\pi(x^*)/\pi(x_t) \leq e^{-\alpha_1(\|x^*\|_2 - \|x_t\|_2)}$ and for $x_t > x^* > x_U$, we have $\pi(x^*)/\pi(x_t) \geq e^{-\alpha_1(\|x^*\|_2 - \|x_t\|_2)}$, and a similar condition holds in the negative tail. Furthermore, we assumed fixed HMC parameters: L leapfrog steps of size ϵ, and wlog the identity mass matrix I. Following [**?**, **?**], it is sufficient to show

$$\limsup_{\|x_t\|_2 \to \infty} \int \left[e^{s(\|x^*(p')\|_2 - \|x_t\|_2)} - 1\right] \alpha(x_t, x^*(p'))\mu(dp') < 0,$$

for some $s > 0$, where $\mu(\cdot)$ is a standard Gaussian measure. Denoting the integral $I_{-\infty}^{\infty}$, we split it into

$$I_{-\infty}^{-x_t^\delta} + I_{-x_t^\delta}^{x_t^\delta} + I_{x_t^\delta}^{\infty},$$

for some $\delta \in (0, 1)$. We show that the first and third terms decay to zero whilst the second remains strictly negative as $x_t \to \infty$ (a similar argument holds as $x_t \to -\infty$). We detail the case $\nabla f(x) \uparrow 0$ as $x \to \infty$ here, the other is analogous. Taking $I_{-x_t^\delta}^{x_t^\delta}$, we can choose an x_t large enough that $x_t - C - L\epsilon x_t^\delta > x_U$, $-\gamma_1 < c(x_t - x_t^\delta) < 0$ and $-\gamma_2 < d(x_t - x_t^\delta) < 0$. So for $p' \in (0, x_t^\delta)$ we have

$$L\epsilon p' > x^* - x_t > L\epsilon p' - \gamma_1 \implies e^{-\alpha_1(-\gamma_1 + L\epsilon p')} \geq e^{-\alpha_1(x^* - x_t)} \geq \pi(x^*)/\pi(x_t),$$

where the last inequality is from (i). For $p' \in (\gamma_2^2/2, x_t^\delta)$

$$\alpha(x_t, x^*) \leq 1 \wedge \frac{\pi(x^*)}{\pi(x_t)} \exp\left(p'\gamma_2/2 - \gamma_2^2/2\right) \leq 1 \wedge \exp\left(-\alpha_2 p' + \alpha_1\gamma_1 - \gamma_2^2/2\right),$$

where x_t is large enough that $\alpha_2 = \alpha_1 L\epsilon - \gamma_2/2 > 0$. Similarly for $p' \in (\gamma_1/L\epsilon, x_t^\delta)$

$$e^{sL\epsilon p'} - 1 \geq e^{s(x^* - x_t)} - 1 \geq e^{s(L\epsilon p' - \gamma_1)} - 1 > 0.$$

Because γ_1 and γ_2 can be chosen to be arbitrarily small, then for large enough x_t we will have

$$\begin{aligned} 0 < I_0^{x_t^\delta} &\leq \int_{\gamma_1/L\epsilon}^{x_t^\delta} [e^{sL\epsilon p'} - 1] \exp\left(-\alpha_2 p' + \alpha_1\gamma_1 - \gamma_2^2/2\right) \mu(dp') + I_0^{\gamma_1/L\epsilon} \\ &= e^{c_1} \int_{\gamma_1/L\epsilon}^{x_t^\delta} [e^{s_2 p'} - 1] e^{-\alpha_2 p'} \mu(dp') + I_0^{\gamma_1/L\epsilon}, \end{aligned} \tag{10}$$

where $c_1 = \alpha_1\gamma_1 - \gamma_2^2/2 > 0$ for large enough x_t, as γ_1 and γ_2 are of the same order. Now turning to $p' \in (-x_t^\delta, 0)$, we can use an exact rearrangement of the same argument (noting that c_1 can be made arbitrarily small) to get

$$I_{-x_t^\delta}^0 \leq e^{c_1} \int_{\gamma_1/L\epsilon}^{x_t^\delta} [e^{-s_2 p'} - 1] \mu(dp') < 0. \tag{11}$$

Combining (10) and (11) and rearranging as in [**?**, Theorem 3.2] shows that $I_{-x_t^\delta}^{x_t^\delta}$ is strictly negative in the limit if $s_2 = sL\epsilon$ is chosen small enough, as $I_0^{\gamma_2/L\epsilon}$ can also be made arbitrarily small.

For $I_{-\infty}^{-x_t^\delta}$ it suffices to note that the Gaussian tails of $\mu(\cdot)$ will dominate the exponential growth of $e^{s(\|x^*(p')\|_2 - \|x_t\|_2)}$ meaning the integral can be made arbitrarily small by choosing large enough x_t, and the same argument holds for $I_{x_t^\delta}^{\infty}$.