# MCMC convergence bounds for reversible chains

Daniel Jerison

University of San Francisco

February 5, 2021

# Sampling problem

Goal: Sample from an intractable probability distribution $\pi$.

Method: Devise a Markov chain whose stationary distribution is $\pi$. Run until it mixes, then sample.

How can we be sure that we are running the chain for long enough?

Convergence diagnostics and Monte Carlo standard error computations can be fooled if the chain converges on two time scales.

# Geometric ergodicity

Let $P$ be a Markov transition kernel on the state space $\mathcal{X}$ with stationary distribution $\pi$.

We say $P$ is *geometrically ergodic* if there is $\rho < 1$ such that

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq F(x)\rho^t \qquad \text{for all } x \in \mathcal{X}.$$

Total variation distance: For probability measures $\mu, \mu'$ on $\mathcal{X}$,

$$\|\mu - \mu'\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\mu(A) - \mu'(A)|.$$

# Geometric ergodicity

Let $P$ be a Markov transition kernel on the state space $\mathcal{X}$ with stationary distribution $\pi$.

We say $P$ is *geometrically ergodic* if there is $\rho < 1$ such that

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq F(x)\rho^t \qquad \text{for all } x \in \mathcal{X}.$$

Total variation distance: For probability measures $\mu, \mu'$ on $\mathcal{X}$,

$$\|\mu - \mu'\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\mu(A) - \mu'(A)|.$$

How to get explicit numerical bounds on $\rho$ and $F(x)$?

# Drift and minorization

The function $V : \mathcal{X} \to [1, \infty)$ is a *drift function* for $P$ with respect to the set $C \subset \mathcal{X}$ if

$$PV(x) \leq \begin{cases} \lambda V(x), & x \in \mathcal{X} \setminus C \\ K, & x \in C \end{cases}$$

for fixed constants $\lambda < 1$, $K < \infty$.

Easy to show: $\mathbf{E}_x[\lambda^{-\tau_C}] \leq V(x)$ where $\tau_C$ is the hitting time of $C$.

# Drift and minorization

The function $V : \mathcal{X} \to [1, \infty)$ is a *drift function* for $P$ with respect to the set $C \subset \mathcal{X}$ if

$$PV(x) \leq \begin{cases} \lambda V(x), & x \in \mathcal{X} \setminus C \\ K, & x \in C \end{cases}$$

for fixed constants $\lambda < 1$, $K < \infty$.

Easy to show: $\mathbf{E}_x[\lambda^{-\tau_C}] \leq V(x)$ where $\tau_C$ is the hitting time of $C$.

The set $C \subset \mathcal{X}$ is *small* if there are $\varepsilon > 0$, $m \geq 1$ and a probability measure $\nu$ on $\mathcal{X}$ such that

$$P^m(x, \cdot) \geq \varepsilon \nu(\cdot) \qquad \text{for all } x \in C.$$

# Drift and minorization

The function $V : \mathcal{X} \to [1, \infty)$ is a *drift function* for $P$ with respect to the set $C \subset \mathcal{X}$ if

$$PV(x) \leq \begin{cases} \lambda V(x), & x \in \mathcal{X} \setminus C \\ K, & x \in C \end{cases}$$

for fixed constants $\lambda < 1$, $K < \infty$.

Easy to show: $\mathbf{E}_x[\lambda^{-\tau_C}] \leq V(x)$ where $\tau_C$ is the hitting time of $C$.

The set $C \subset \mathcal{X}$ is *small* if there are $\varepsilon > 0$, $m \geq 1$ and a probability measure $\nu$ on $\mathcal{X}$ such that

$$P^m(x, \cdot) \geq \varepsilon \nu(\cdot) \qquad \text{for all } x \in C.$$

Names: Nummelin, Tuominen, Athreya, Ney, Meyn, Tweedie, Roberts, Rosenthal, ...

# Regeneration

What does it mean if $P$ has a drift function with respect to a small set $C$?

Due to the drift function, the Markov chain visits $C$ frequently.

Every visit, with probability $\varepsilon$ the chain will *regenerate* at the measure $\nu$ after $m$ subsequent steps.

# Regeneration

What does it mean if $P$ has a drift function with respect to a small set $C$?

Due to the drift function, the Markov chain visits $C$ frequently.

Every visit, with probability $\varepsilon$ the chain will *regenerate* at the measure $\nu$ after $m$ subsequent steps.

## Definition

A randomized stopping time $T$ for the Markov chain $(X_t)$ with $X_0 \sim \mu$ is a *regeneration time* with measure $\nu$ if

$$\mathbf{P}_\mu(X_n \in S \mid T = n, X_0, \ldots, X_{n-1}) = \nu(S) \qquad \text{for all } n \geq 0,\ S \subset \mathcal{X}.$$

# Strong random times

Intuition: Using drift and minorization, we can define a sequence of regeneration times that partition the sample path of the chain into iid tours started from $\nu$.

This works when $m = 1$, i.e. $P(x, \cdot) \geq \varepsilon\nu(\cdot)$ for $x \in C$.

When $m > 1$ in the minorization, the tours are not iid and the random times do not satisfy the regeneration time definition.

# Strong random times

Intuition: Using drift and minorization, we can define a sequence of regeneration times that partition the sample path of the chain into iid tours started from $\nu$.

This works when $m = 1$, i.e. $P(x, \cdot) \geq \varepsilon\nu(\cdot)$ for $x \in C$.

When $m > 1$ in the minorization, the tours are not iid and the random times do not satisfy the regeneration time definition.

## Definition (Miclo)

A randomized stopping time $T$ for the Markov chain $(X_t)$ with $X_0 \sim \mu$ is a *strong random time* with measure $\nu$ if

$$\mathbf{P}_\mu(X_n \in S \mid T = n) = \nu(S) \qquad \text{for all } n \geq 0,\ S \subset \mathcal{X}.$$

# Aperiodicity

**Theorem**

*Suppose the Markov transition kernel $P$ has a drift function with respect to a small set. Then there is a strong random time $T$ for the associated Markov chain such that for each $x \in \mathcal{X}$,*

$$\mathbf{P}_x(T > t) \leq F(x)\rho^t \qquad \text{for all } t \geq 0,$$

*where the function $F(x)$ and the constant $\rho < 1$ have explicit formulas in terms of the drift and minorization data.*

# Aperiodicity

**Theorem**

*Suppose the Markov transition kernel $P$ has a drift function with respect to a small set. Then there is a strong random time $T$ for the associated Markov chain such that for each $x \in \mathcal{X}$,*

$$\mathbf{P}_x(T > t) \leq F(x)\rho^t \qquad \text{for all } t \geq 0,$$

*where the function $F(x)$ and the constant $\rho < 1$ have explicit formulas in terms of the drift and minorization data.*

In order for this to imply convergence to the stationary distribution, we need an aperiodicity condition.

# Almost periodic example

Consider the transition matrix $P$ on the cycle $\mathbf{Z}/n\mathbf{Z}$ given by:

- $P(j, j-1) = 1$ for $j \neq 0$
- $P(0,0) = P(0, n-1) = 1/2$

# Almost periodic example

Consider the transition matrix $P$ on the cycle $\mathbf{Z}/n\mathbf{Z}$ given by:

- $P(j, j-1) = 1$ for $j \neq 0$
- $P(0,0) = P(0, n-1) = 1/2$

The Markov chain $(X_t)$ regenerates every $n$ steps, but takes order $n^3$ steps to converge to its stationary distribution $\pi$.

If we want an approximate sample from $\pi$, we could sample at a random time $X_N$ where $N \sim \mathrm{Uniform}(1, \ldots, n)$. This takes much less time than waiting for the chain to converge!

# Agenda

Outline of our plan:

1. Find a drift function with respect to a small set.
2. This gives a strong random time.
3. Deduce geometric ergodicity with explicit constants.
4. Now we know how long to run the Markov chain.

If the chain exhibits (almost) periodic behavior, then the constants in step 3 will be very bad. And, we might be able to wash out the periodicity by sampling at a random time, in which case we wouldn't care that the chain takes much longer to converge.

# Agenda

Outline of our plan:

1. Find a drift function with respect to a small set.
2. This gives a strong random time.
3. Deduce geometric ergodicity with explicit constants.
4. Now we know how long to run the Markov chain.

If the chain exhibits (almost) periodic behavior, then the constants in step 3 will be very bad. And, we might be able to wash out the periodicity by sampling at a random time, in which case we wouldn't care that the chain takes much longer to converge.

Conclusion: This plan is most likely to give useful results if we restrict our attention to chains that display no periodic behavior.

# Reversible chains with nonnegative spectrum

We say $P$ is *reversible* with respect to $\pi$ if $\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$ as measures on $\mathcal{X} \times \mathcal{X}$, or equivalently, if the operator $f \mapsto Pf$ is self-adjoint on the space $L^2(\pi)$ with inner product

$$\langle f, g \rangle_\pi = \int_{\mathcal{X}} f(x)g(x)\pi(dx).$$

If $P$ is reversible, its spectrum is contained in the real interval $[-1, 1]$.

# Reversible chains with nonnegative spectrum

We say $P$ is *reversible* with respect to $\pi$ if $\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$ as measures on $\mathcal{X} \times \mathcal{X}$, or equivalently, if the operator $f \mapsto Pf$ is self-adjoint on the space $L^2(\pi)$ with inner product

$$\langle f, g \rangle_\pi = \int_{\mathcal{X}} f(x)g(x)\pi(dx).$$

If $P$ is reversible, its spectrum is contained in the real interval $[-1, 1]$.

Examples of Markov chains that are reversible with nonnegative spectrum (contained in $[0, 1]$): many Metropolis–Hastings chains, random scan Gibbs samplers, two-variable deterministic scan Gibbs samplers.

# Lack of periodic behavior

Heuristically, a Markov chain with near-periodic behavior of period $k$ should have an eigenvalue close to a $k$-th root of unity. If $P$ is reversible with nonnegative spectrum, this cannot happen.

Indeed, for such $P$, the "sample at a random time" trick doesn't work: if $N$ is a random variable supported on $\{0, 1, \ldots, n\}$, then the law of $X_n$ is closer to $\pi$ (in an appropriate distance) than the law of $X_N$.

# Convergence result

## Theorem (J., 2019)

*Let $P$ be a Markov transition kernel that is reversible with respect to $\pi$ and has nonnegative spectrum. Suppose that $T$ is a strong random time with measure $\nu$ for the associated Markov chain, such that $\mathbf{P}_\nu(T = 0) = 0$ and $\mathbf{E}_\nu[T] < \infty$. Then for all $t \geq 0$,*

$$\|P^t(\nu, \cdot) - \pi\|_{\mathrm{TV}}^2 \leq \frac{1}{4} \sum_{n=2t+1}^{\infty} \mathbf{P}_\nu(T > n).$$

The result can easily be extended to bound $\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}}$ for any $x \in \mathcal{X}$ by a coupling argument.

# Drift and minorization consequences

## Corollary

*Let $P$ be a Markov transition kernel that is reversible with respect to $\pi$ and has nonnegative spectrum. Suppose that $P$ has a drift function with respect to a small set, leading to a strong random time $T$ satisfying*

$$\mathbf{P}_x(T > t) \leq F(x)\rho^t \qquad \text{for all } t \geq 0$$

*where $F(x)$ and $\rho < 1$ have explicit formulas. Then there are explicit functions $G(x), H(x)$ such that for all $x \in \mathcal{X}$,*

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq [G(x)t + H(x)]\rho^t \qquad \text{for all } t \geq 0.$$

The exponential convergence rate is the same as the tail decay rate of the strong random time. In the case of one-step minorization, this was previously shown by Baxendale (2005) using analytic methods.

# Nuclear pump example

Well-studied toy example of a Gibbs sampler for a Poisson–Gamma Bayesian model that predicts the failure rate of pumps in a nuclear power plant. Rosenthal (1995) proposed a simple drift function.

Let $\tau_{0.01} = \min\{t : \|P^t(x_0, \cdot) - \pi\|_{\mathrm{TV}} \leq 0.01\}$ for a fixed starting value $x_0$.

# Nuclear pump example

Well-studied toy example of a Gibbs sampler for a Poisson–Gamma Bayesian model that predicts the failure rate of pumps in a nuclear power plant. Rosenthal (1995) proposed a simple drift function.

Let $\tau_{0.01} = \min\{t : \|P^t(x_0, \cdot) - \pi\|_{\mathrm{TV}} \leq 0.01\}$ for a fixed starting value $x_0$.

Rosenthal (1995, using bivariate drift approach): $\tau_{0.01} \leq 192$

Baxendale (2005, purposely ignoring that $P$ is reversible with nonnegative eigenvalues): $\tau_{0.01} \leq 1.0 \cdot 10^7$

Baxendale (2005, using analytic methods): $\tau_{0.01} \leq 212$

J. (2019, using strong random times): $\tau_{0.01} \leq 83$

# Nuclear pump example

Well-studied toy example of a Gibbs sampler for a Poisson–Gamma Bayesian model that predicts the failure rate of pumps in a nuclear power plant. Rosenthal (1995) proposed a simple drift function.

Let $\tau_{0.01} = \min\{t : \|P^t(x_0, \cdot) - \pi\|_{\mathrm{TV}} \leq 0.01\}$ for a fixed starting value $x_0$.

Rosenthal (1995, using bivariate drift approach): $\tau_{0.01} \leq 192$

Baxendale (2005, purposely ignoring that $P$ is reversible with nonnegative eigenvalues): $\tau_{0.01} \leq 1.0 \cdot 10^7$

Baxendale (2005, using analytic methods): $\tau_{0.01} \leq 212$

J. (2019, using strong random times): $\tau_{0.01} \leq 83$

The truth: $\tau_{0.01} = 2$

# Proof of main theorem

Let $f = d\nu/d\pi$ be the Radon–Nikodym derivative, and consider the sequence $a_t = \mathbf{E}_\nu[f(X_t)]$.

Since $P$ is reversible with nonnegative spectrum, and the chain converges to $\pi$, it can be shown that $(a_t)$ is a decreasing sequence with limit 1.

In addition, using reversibility again, we have

$$4\|P^t(\nu,\cdot) - \pi\|_{\mathrm{TV}}^2 \leq \|P^t(\nu,\cdot) - \pi\|_{L^2(\pi)}^2 = a_{2t} - 1.$$

# Proof of main theorem

Let $f = d\nu/d\pi$ be the Radon–Nikodym derivative, and consider the sequence $a_t = \mathbf{E}_\nu[f(X_t)]$.

Since $P$ is reversible with nonnegative spectrum, and the chain converges to $\pi$, it can be shown that $(a_t)$ is a decreasing sequence with limit 1.

In addition, using reversibility again, we have

$$4\|P^t(\nu,\cdot) - \pi\|_{\mathrm{TV}}^2 \leq \|P^t(\nu,\cdot) - \pi\|_{L^2(\pi)}^2 = a_{2t} - 1.$$

As $T$ is a strong random time with measure $\nu$, the definition of $a_t$ yields

$$\sum_{j=0}^{n-1} \mathbf{P}_\nu(T = n - j)a_j \leq a_n$$

and applying summation by parts to this inequality finishes the proof.

Thank you!