Explicit convergence bounds for Metropolis Markov chains

Andi Q. Wang

University of Warwick

Joint with: Christophe Andrieu, Anthony Lee, Sam Power.

Approximation Methods in Bayesian Analysis - CIRM

June 2023

Overview

Quick overview

Focus of the talk: MCMC / sampling.

Quick overview

Focus of the talk: MCMC / sampling.

We will be discussing (new) bounds for the convergence to equilibrium of MCMC algorithms.

Quick overview

Focus of the talk: MCMC / sampling.

We will be discussing (new) bounds for the convergence to equilibrium of MCMC algorithms.

Thus the talk is mostly theoretical, but the purpose of this theory is to guide application!

Quick overview

Focus of the talk: MCMC / sampling.

We will be discussing (new) bounds for the convergence to equilibrium of MCMC algorithms.

Thus the talk is mostly theoretical, but the purpose of this theory is to guide application!

I will present fundamental bounds on the spectral gap of Random Walk Metropolis, which has been an open problem for many years!

Quick overview

Focus of the talk: MCMC / sampling.

We will be discussing (new) bounds for the convergence to equilibrium of MCMC algorithms.

Thus the talk is mostly theoretical, but the purpose of this theory is to guide application!

I will present fundamental bounds on the spectral gap of Random Walk Metropolis, which has been an open problem for many years!

Along the way I will describe our technique for deriving convergence bounds based on isoperimetry and conductance.

Quick overview

Focus of the talk: MCMC / sampling.

We will be discussing (new) bounds for the convergence to equilibrium of MCMC algorithms.

Thus the talk is mostly theoretical, but the purpose of this theory is to guide application!

I will present fundamental bounds on the spectral gap of Random Walk Metropolis, which has been an open problem for many years!

Along the way I will describe our technique for deriving convergence bounds based on isoperimetry and conductance.

See arXiv preprint! `https://arxiv.org/abs/2211.08959`.

Statistical modelling

Suppose we have some dataset $y = \{y_1, y_2, \ldots, y_N\}$.

Statistical modelling

Suppose we have some dataset $y = \{y_1, y_2, \ldots, y_N\}$.

Posit a model (density function) $f_x(y)$ which generated y, which depends upon (unknown) parameters $x \in \mathcal{X} = \mathbb{R}^d$.

Statistical modelling

Suppose we have some dataset $y = \{y_1, y_2, \ldots, y_N\}$.

Posit a model (density function) $f_x(y)$ which generated y, which depends upon (unknown) parameters $x \in \mathcal{X} = \mathbb{R}^d$.

Seek learn or infer values of the parameter x which are commensurate with the observed dataset y.

The Bayesian approach

Encode prior beliefs into a prior distribution $\nu(x)$, and define likelihood $\ell_y(x) := f_x(y)$.

The Bayesian approach

Encode prior beliefs into a prior distribution $\nu(x)$, and define likelihood $\ell_y(x) := f_x(y)$.

Given our observations, our posterior distribution is

$$\pi(x) = \pi(x|y) = \frac{\nu(x)\ell_y(x)}{\int \nu(z)\ell_y(z)\,\mathrm{d}z} \propto \nu(x)\ell_y(x).$$

The Bayesian approach

Encode prior beliefs into a prior distribution $\nu(x)$, and define likelihood $\ell_y(x) := f_x(y)$.

Given our observations, our posterior distribution is

$$\pi(x) = \pi(x|y) = \frac{\nu(x)\ell_y(x)}{\int \nu(z)\ell_y(z)\,\mathrm{d}z} \propto \nu(x)\ell_y(x).$$

We are then interested in quantities of the form

$$I = \pi(f) = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x,$$

e.g. $f(x) = \|x\|^p$ (posterior moments), $f(x) = 1_A(x)$ (credible sets / posterior tail probabilities), etc.

Sampling

So we wish to evaluate integrals

$$I = \pi(f) = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x,$$

where π is a probability density function (our posterior distribution).

Sampling

So we wish to evaluate integrals

$$I = \pi(f) = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x,$$

where π is a probability density function (our posterior distribution).

Direct integration infeasible in high-dimensions (curse of dimensionality), furthermore only have access to π up to a normalizing constant!

Sampling

So we wish to evaluate integrals

$$I = \pi(f) = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x,$$

where π is a probability density function (our posterior distribution).

Direct integration infeasible in high-dimensions (curse of dimensionality), furthermore only have access to π up to a normalizing constant!

So instead, approximate I by sampling $X_1, X_2, \ldots, X_n \sim \pi$ and consider

$$I_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx I = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x.$$

Sampling

So we wish to evaluate integrals

$$I = \pi(f) = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x,$$

where π is a probability density function (our posterior distribution).

Direct integration infeasible in high-dimensions (curse of dimensionality), furthermore only have access to π up to a normalizing constant!

So instead, approximate I by sampling $X_1, X_2, \ldots, X_n \sim \pi$ and consider

$$I_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx I = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x.$$

There are also optimization-based approaches such as Variational Inference, INLA, ...

Monte Carlo

So instead, approximate I by sampling $X_1, X_2, \ldots, X_n \sim \pi$.

Monte Carlo

So instead, approximate I by sampling $X_1, X_2, \ldots, X_n \sim \pi$.

Exact sampling hard (e.g. rejection sampling also suffers from a curse of dimensionality)

Monte Carlo

So instead, approximate I by sampling $X_1, X_2, \ldots, X_n \sim \pi$.

Exact sampling hard (e.g. rejection sampling also suffers from a curse of dimensionality) so instead: build an ergodic Markov chain X which possesses π as its stationary distribution.

Monte Carlo

So instead, approximate I by sampling $X_1, X_2, \ldots, X_n \sim \pi$.

Exact sampling hard (e.g. rejection sampling also suffers from a curse of dimensionality) so instead: build an ergodic Markov chain X which possesses π as its stationary distribution.

We simulate a π-reversible ergodic Markov chain,

$$X_1, X_2, \ldots$$

where $X_n \to \pi$ in distribution and considering

$$I_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i) \approx I = \int_{\mathcal{X}} f(x)\pi(x)\,\mathrm{d}x.$$

Metropolis–Hastings

Algorithm 1 Metropolis–Hastings (MH)

initialise: $X_0 = x_0, i = 0$
while $i < N$ **do**
  $i \leftarrow i + 1$
  simulate $Y_i \sim Q(X_{i-1}, \cdot)$
  $\alpha(X_{i-1}, Y_i) = 1 \wedge \frac{q(Y_i, X_{i-1})\pi(Y_i)}{q(X_{i-1}, Y_i)\pi(X_{i-1})}$
  with probability $\alpha(X_{i-1}, Y_i)$
    $X_i \leftarrow Y_i$
  else
    $X_i \leftarrow X_{i-1}$
return $(X_i)_{i=1,\ldots,n}$

Random walk Metropolis

We will focus on Random Walk Metropolis (RWM) [Metropolis et. al. (1953)]:
$Q(X_{i-1}, \cdot) = \mathcal{N}(X_{i-1}, \sigma^2 \cdot \mathbf{I})$.

Random walk Metropolis

We will focus on Random Walk Metropolis (RWM) [Metropolis et. al. (1953)]: $Q(X_{i-1}, \cdot) = \mathcal{N}(X_{i-1}, \sigma^2 \cdot \mathbf{I})$.

A 'fundamental' MCMC method – first port of call, benchmark method.

Random walk Metropolis

We will focus on Random Walk Metropolis (RWM) [Metropolis et. al. (1953)]:
$Q(X_{i-1}, \cdot) = \mathcal{N}(X_{i-1}, \sigma^2 \cdot \mathbf{I})$.

A 'fundamental' MCMC method – first port of call, benchmark method.

Very simple to implement, and yet surprisingly robust [Livingstone and Zanella (2022)].

Random walk Metropolis

We will focus on Random Walk Metropolis (RWM) [Metropolis et. al. (1953)]: $Q(X_{i-1}, \cdot) = \mathcal{N}(X_{i-1}, \sigma^2 \cdot \mathbf{I})$.

A 'fundamental' MCMC method – first port of call, benchmark method.

Very simple to implement, and yet surprisingly robust [Livingstone and Zanella (2022)].

But tuning of $\sigma^2 \cdot \mathbf{I}$ is critical for good performance.

Random walk Metropolis

We will focus on Random Walk Metropolis (RWM) [Metropolis et. al. (1953)]: $Q(X_{i-1}, \cdot) = \mathcal{N}(X_{i-1}, \sigma^2 \cdot \mathbf{I})$.

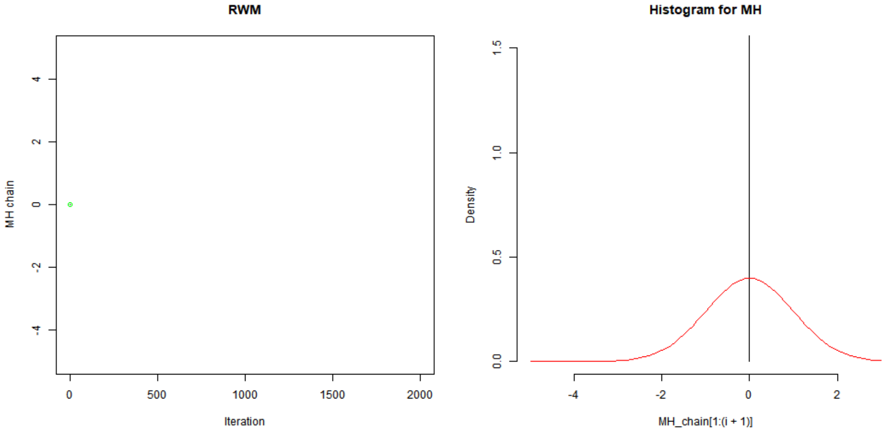
A 'fundamental' MCMC method – first port of call, benchmark method.

Very simple to implement, and yet surprisingly robust [Livingstone and Zanella (2022)].

But tuning of $\sigma^2 \cdot \mathbf{I}$ is critical for good performance.

And suprisingly some things were still unknown! (Spectral gap.)

MH example


RWM
MH chain
4
2
0
-2
-4
0
500
1000
1500
2000
Iteration
Histogram for MH
Density
1.5
1.0
0.5
0.0
-4
-2
0
2
MH_chain[1:(i + 1)]

Tuning of RWM

Tuning proposal variance σ^2 is critical for good performance of RWM.

Tuning of RWM

Tuning proposal variance σ^2 is critical for good performance of RWM.

σ^2 too large $\Rightarrow$ most proposals rejected; wasted computational effort.

Tuning of RWM

Tuning proposal variance σ^2 is critical for good performance of RWM.

σ^2 too large $\Rightarrow$ most proposals rejected; wasted computational effort.

σ^2 too small $\Rightarrow$ proposing tiny moves; wasted computational effort.

Tuning of RWM

Tuning proposal variance σ^2 is critical for good performance of RWM.

σ^2 too large $\Rightarrow$ most proposals rejected; wasted computational effort.

σ^2 too small $\Rightarrow$ proposing tiny moves; wasted computational effort.

One beautiful way to approach this problem is optimal scaling [Roberts, Gelman, Gilks (1997)]:

Tuning of RWM

Tuning proposal variance σ^2 is critical for good performance of RWM.

σ^2 too large $\Rightarrow$ most proposals rejected; wasted computational effort.

σ^2 too small $\Rightarrow$ proposing tiny moves; wasted computational effort.

One beautiful way to approach this problem is optimal scaling [Roberts, Gelman, Gilks (1997)]:

It was shown that for a restricted class of targets π, in the high-dimensional limit, when scaling the variance like $\sigma^2 \sim d^{-1}$, the RWM chain has a stable acceptance ratio, and converges to a Langevin diffusion, and that the cost is like $O(d)$.

Optimal scaling

So optimal scaling tells us that for certain targets π, we should choose $\sigma^2 \sim d^{-1}$ to get a stable acceptance ratio in high dimensions, and even that we should aim for average acceptances rates of 0.234.

Optimal scaling

So optimal scaling tells us that for certain targets π, we should choose $\sigma^2 \sim d^{-1}$ to get a stable acceptance ratio in high dimensions, and even that we should aim for average acceptances rates of 0.234.

But optimal scaling is purely asymptotic and does not say anything about any particular algorithm.

Optimal scaling

So optimal scaling tells us that for certain targets π, we should choose $\sigma^2 \sim d^{-1}$ to get a stable acceptance ratio in high dimensions, and even that we should aim for average acceptances rates of 0.234.

But optimal scaling is purely asymptotic and does not say anything about any particular algorithm.

For example, suppose I am doing Bayesian logistic regression in $d = 1000$ and I have chosen $\sigma^2 = 5 \times 10^{-4}$. How long should I run my chain for?

Our approach

Instead, we take a different perspective to analyse the high-dimensional properties of RWM:

Our approach

Instead, we take a different perspective to analyse the high-dimensional properties of RWM:

We seek to explicitly give bounds on the convergence rate of RWM (via spectral gap) in arbitrary dimensions d and for any value of the proposal variance σ^2.

Our approach

Instead, we take a different perspective to analyse the high-dimensional properties of RWM:

We seek to explicitly give bounds on the convergence rate of RWM (via spectral gap) in arbitrary dimensions d and for any value of the proposal variance σ^2.

For appropriately regular targets, we will show that scaling $\sigma^2 \sim d^{-1}$ does indeed imply a spectral gap of order d^{-1}, and that this is optimal.

Our approach

Instead, we take a different perspective to analyse the high-dimensional properties of RWM:

We seek to explicitly give bounds on the convergence rate of RWM (via spectral gap) in arbitrary dimensions d and for any value of the proposal variance σ^2.

For appropriately regular targets, we will show that scaling $\sigma^2 \sim d^{-1}$ does indeed imply a spectral gap of order d^{-1}, and that this is optimal.

Unlike previous work, we do not need to restrict the state space to a compact set [Belloni and Chernozhukov (2009), Dwivedi et. al. (2019), Chen et. al. (2019)].

Our approach

Instead, we take a different perspective to analyse the high-dimensional properties of RWM:

We seek to explicitly give bounds on the convergence rate of RWM (via spectral gap) in arbitrary dimensions d and for any value of the proposal variance σ^2.

For appropriately regular targets, we will show that scaling $\sigma^2 \sim d^{-1}$ does indeed imply a spectral gap of order d^{-1}, and that this is optimal.

Unlike previous work, we do not need to restrict the state space to a compact set [Belloni and Chernozhukov (2009), Dwivedi et. al. (2019), Chen et. al. (2019)].

However we are restricted to considering RWM, as opposed to MALA/HMC [Dwivedi et. al. (2019), Chen et. al. (2019)].

Spectral gap

Recall that a reversible π-invariant Markov kernel P defines an operator on $L^2(\pi)$, and its convergence to equilibrium can be bounded by the spectral gap γ (and this is the best rate):

Spectral gap

Recall that a reversible π-invariant Markov kernel P defines an operator on $L^2(\pi)$, and its convergence to equilibrium can be bounded by the spectral gap γ (and this is the best rate):

$$\|P^n f - \pi(f)\|_2 \leq (1-\gamma)^n \|f\|_2.$$

Spectral gap

Recall that a reversible π-invariant Markov kernel P defines an operator on $L^2(\pi)$, and its convergence to equilibrium can be bounded by the spectral gap γ (and this is the best rate):

$$\|P^n f - \pi(f)\|_2 \leq (1-\gamma)^n \|f\|_2.$$

Under the commmon assumptions of L-smoothness and m-strong convexity of the potential U [Dwivedi et. al. (2019), Chen et. al. (2019)], we can give straightforward results (but framework more general!).

Spectral gap

Recall that a reversible π-invariant Markov kernel P defines an operator on $L^2(\pi)$, and its convergence to equilibrium can be bounded by the spectral gap γ (and this is the best rate):

$$\|P^n f - \pi(f)\|_2 \leq (1-\gamma)^n \|f\|_2.$$

Under the common assumptions of L-smoothness and m-strong convexity of the potential U [Dwivedi et. al. (2019), Chen et. al. (2019)], we can give straightforward results (but framework more general!).

Such densities can be sandwiched between $\mathcal{N}(x_*, L^{-1}\mathbf{I}_d)$ and $\mathcal{N}(x_*, m^{-1}\mathbf{I}_d)$ densities.

Main result

Theorem ([Andrieu, Lee, Power, W. (2022)])

For an L-smooth and m-strongly convex and twice differential potential U on $\mathbb{R}^d$, RWM targeting $\pi \propto \exp(-U)$ with proposal increments $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ has spectral gap γ satisfying

$$C \cdot L \cdot d \cdot \sigma^2 \cdot \exp(-2Ld\sigma^2) \cdot \frac{m}{L} \cdot \frac{1}{d} \leq \gamma \leq \frac{L \cdot \sigma^2}{2} \wedge (1 + m \cdot \sigma^2)^{-d/2},$$

where $C = 1 \times 10^{-4}$.

Main result

Theorem ([Andrieu, Lee, Power, W. (2022)])

For an L-smooth and m-strongly convex and twice differential potential U on $\mathbb{R}^d$, RWM targeting $\pi \propto \exp(-U)$ with proposal increments $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ has spectral gap γ satisfying

$$C \cdot L \cdot d \cdot \sigma^2 \cdot \exp(-2Ld\sigma^2) \cdot \frac{m}{L} \cdot \frac{1}{d} \leq \gamma \leq \frac{L \cdot \sigma^2}{2} \wedge (1 + m \cdot \sigma^2)^{-d/2},$$

where $C = 1 \times 10^{-4}$.

To maximise the lower bound, take $\sigma = \varsigma \cdot L^{-1/2} \cdot d^{-1/2}$, and then

$$C \cdot \varsigma^2 \cdot \exp(-2\varsigma^2) \cdot \frac{m}{L} \cdot \frac{1}{d} \leq \gamma \leq \frac{\varsigma^2}{2} \cdot \frac{1}{d}.$$

Comments

Theorem ([Andrieu, Lee, Power, W. (2022)])

For an L-smooth and m-strongly convex and twice differential potential U on $\mathbb{R}^d$, RWM targeting $\pi \propto \exp(-U)$ with proposal increments $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ with $\sigma^2 = \varsigma \cdot L^{-1/2} \cdot d^{-1/2}$ has spectral gap γ satisfying

$$C \cdot \varsigma^2 \cdot \exp(-2\varsigma^2) \cdot \frac{m}{L} \cdot \frac{1}{d} \leq \gamma \leq \frac{\varsigma^2}{2} \cdot \frac{1}{d}.$$

Comments

Theorem ([Andrieu, Lee, Power, W. (2022)])

For an L-smooth and m-strongly convex and twice differential potential U on $\mathbb{R}^d$, RWM targeting $\pi \propto \exp(-U)$ with proposal increments $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ with $\sigma^2 = \varsigma \cdot L^{-1/2} \cdot d^{-1/2}$ has spectral gap γ satisfying

$$C \cdot \varsigma^2 \cdot \exp(-2\varsigma^2) \cdot \frac{m}{L} \cdot \frac{1}{d} \leq \gamma \leq \frac{\varsigma^2}{2} \cdot \frac{1}{d}.$$

So indeed we see the spectral gap of RWM is $O(d^{-1})$.

Comments

Theorem ([Andrieu, Lee, Power, W. (2022)])

For an L-smooth and m-strongly convex and twice differential potential U on $\mathbb{R}^d$, RWM targeting $\pi \propto \exp(-U)$ with proposal increments $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ with $\sigma^2 = \varsigma \cdot L^{-1/2} \cdot d^{-1/2}$ has spectral gap γ satisfying

$$C \cdot \varsigma^2 \cdot \exp(-2\varsigma^2) \cdot \frac{m}{L} \cdot \frac{1}{d} \leq \gamma \leq \frac{\varsigma^2}{2} \cdot \frac{1}{d}.$$

So indeed we see the spectral gap of RWM is $O(d^{-1})$.

Note that this applies for any d and for any ς, i.e. it actually says something about the algorithm you are running!

Overview

Conductance

Definition: Conductance

The conductance of a π-invariant Markov kernel P is

$$\Phi_P^* := \inf\left\{\frac{(\pi\otimes P)(A\times A^{\complement})}{\pi(A)} : \pi(A)\leq\frac{1}{2}\right\},\quad v\in(0,1/2].$$

Conductance

Definition: Conductance

The conductance of a π-invariant Markov kernel P is

$$\Phi_P^* := \inf \left\{ \frac{(\pi \otimes P)(A \times A^{\complement})}{\pi(A)} : \pi(A) \leq \frac{1}{2} \right\}, \quad v \in (0, 1/2].$$

Theorem (Cheeger inequalities)

For a positive chain, such as RWM, we have the bounds on the spectral gap,

$$\frac{1}{2} \cdot [\Phi_P^*]^2 \leq \gamma \leq \Phi_P^*.$$

Conductance

Definition: Conductance

The conductance of a π-invariant Markov kernel P is

$$\Phi_P^* := \inf\left\{\frac{(\pi \otimes P)(A \times A^\complement)}{\pi(A)} : \pi(A) \leq \frac{1}{2}\right\}, \quad v \in (0, 1/2].$$

Theorem (Cheeger inequalities)

For a positive chain, such as RWM, we have the bounds on the spectral gap,

$$\frac{1}{2} \cdot [\Phi_P^*]^2 \leq \gamma \leq \Phi_P^*.$$

Thus our goal is to lower bound the conductance.

Isoperimetry

Fix target density π on metric space (E, d).

Definition: isoperimetric profile / minorant, c.f. [Milman (2009)]

Given a measurable set A, define the r-enlargment of A via $A_r := \{x \in \mathrm{E} : \mathrm{d}(x, A) \le r\}$, and set

$$\pi^+(A) := \liminf_{r \downarrow 0} \frac{\pi(A_r) - \pi(A)}{r}.$$

Then the isoperimetric profile of π is

$$I_\pi(p) := \inf\{\pi^+(A) : A \in \mathcal{E}, \pi(A) = p\}, \quad p \in (0, 1).$$

Isoperimetry

Fix target density π on metric space (E, d).

Definition: isoperimetric profile / minorant, c.f. [Milman (2009)]

Given a measurable set A, define the r-enlargment of A via $A_r := \{x \in \mathsf{E} : \mathsf{d}(x, A) \leq r\}$, and set

$$\pi^+(A) := \liminf_{r \downarrow 0} \frac{\pi(A_r) - \pi(A)}{r}.$$

Then the isoperimetric profile of π is

$$I_\pi(p) := \inf\{\pi^+(A) : A \in \mathcal{E}, \pi(A) = p\}, \quad p \in (0, 1).$$

A function $\tilde{I}_\pi : (0, 1) \to (0, \infty)$ is a regular isoperimetric minorant of π if $\tilde{I}_\pi$ is continuous, monotone increasing, symmetric about $1/2$ and $\tilde{I}_\pi \leq I_\pi$.

Close coupling

Definition: close coupling

Given $\epsilon, \delta > 0$, we say that a Markov kernel P is $(\mathsf{d}, \delta, \epsilon)$-close coupling if

$$\mathsf{d}(x, y) \leq \delta \Rightarrow \|P(x, \cdot) - P(y, \cdot)\|_{\mathrm{TV}} \leq 1 - \epsilon, \quad \forall x, y \in \mathsf{E}.$$

Close coupling

Definition: close coupling

Given $\epsilon, \delta > 0$, we say that a Markov kernel P is $(\mathrm{d}, \delta, \epsilon)$-close coupling if

$$\mathrm{d}(x, y) \le \delta \Rightarrow \|P(x, \cdot) - P(y, \cdot)\|_{\mathrm{TV}} \le 1 - \epsilon, \quad \forall x, y \in \mathrm{E}.$$

Lemma: close coupling for Metropolis chains

For Metropolis chains, we have the bound:

$$\|P(x, \cdot) - P(y, \cdot)\|_{\mathrm{TV}} \le \|Q(x, \cdot) - Q(y, \cdot)\|_{\mathrm{TV}} + 1 - \alpha_0,$$

$$\alpha_0 := \inf_{x \in \mathrm{E}} \alpha(x), \quad \alpha(x) := \int \alpha(x, y) Q(x, \mathrm{d}y).$$

Close coupling

Definition: close coupling

Given $\epsilon, \delta > 0$, we say that a Markov kernel P is $(\mathsf{d}, \delta, \epsilon)$-close coupling if

$$\mathsf{d}(x, y) \le \delta \Rightarrow \|P(x, \cdot) - P(y, \cdot)\|_{\mathrm{TV}} \le 1 - \epsilon, \quad \forall x, y \in \mathsf{E}.$$

Lemma: close coupling for Metropolis chains

For Metropolis chains, we have the bound:

$$\|P(x, \cdot) - P(y, \cdot)\|_{\mathrm{TV}} \le \|Q(x, \cdot) - Q(y, \cdot)\|_{\mathrm{TV}} + 1 - \alpha_0,$$

$$\alpha_0 := \inf_{x \in \mathsf{E}} \alpha(x), \quad \alpha(x) := \int \alpha(x, y) Q(x, \mathrm{d}y).$$

Thus we can choose δ such that $|x - y| \le \delta \Rightarrow \|Q(x, \cdot) - Q(y, \cdot)\|_{\mathrm{TV}} \le \alpha_0/2$ to obtain P is close coupling with $\epsilon \ge \alpha_0/2$, provided we can bound α_0!

Close coupling, conductance and isoperimetry

Theorem: Conductance lower bound; c.f. [Dwivedi et. al. (2019)]

Suppose $\tilde{I}_\pi$ is a regular, concave isoperimetric minorant of π. Let P be $(\mathrm{d}, \delta, \epsilon)$-close coupling. Then

$$\Phi_P^* \geq \frac{1}{4} \cdot \epsilon \cdot 1 \wedge \left(\frac{\delta}{2} \cdot \frac{\tilde{I}_\pi(1/4)}{1/4} \right).$$

Close coupling, conductance and isoperimetry

Theorem: Conductance lower bound; c.f. [Dwivedi et. al. (2019)]

Suppose $\tilde{I}_\pi$ is a regular, concave isoperimetric minorant of π. Let P be $(\mathrm{d}, \delta, \epsilon)$-close coupling. Then

$$\Phi_P^* \geq \frac{1}{4} \cdot \epsilon \cdot 1 \wedge \left(\frac{\delta}{2} \cdot \frac{\tilde{I}_\pi(1/4)}{1/4} \right).$$

So we have a lower bound on the conductance Φ_P^*, and hence on the spectral gap.

Close coupling, conductance and isoperimetry

Theorem: Conductance lower bound; c.f. [Dwivedi et. al. (2019)]

Suppose $\tilde{I}_\pi$ is a regular, concave isoperimetric minorant of π. Let P be $(\mathrm{d}, \delta, \epsilon)$-close coupling. Then

$$\Phi_P^* \geq \frac{1}{4} \cdot \epsilon \cdot 1 \wedge \left(\frac{\delta}{2} \cdot \frac{\tilde{I}_\pi(1/4)}{1/4} \right).$$

So we have a lower bound on the conductance Φ_P^*, and hence on the spectral gap.

This result thus breaks the problem into two pieces:

- For a given target π, establish a regular concave isoperimetric minorant $\tilde{I}_\pi$.
- For the chain P, establish close coupling.

Overview

Isoperimetric minorants for π

There are various ways to establish isoperimetric minorants: for example, they can be derived from functional inequalities, e.g. Poincaré inequalities, log-Sobolev inequalities, c.f. [Bobkov (1999)].

Isoperimetric minorants for π

There are various ways to establish isoperimetric minorants: for example, they can be derived from functional inequalities, e.g. Poincaré inequalities, log-Sobolev inequalities, c.f. [Bobkov (1999)].

The specific case of interest for this talk:

Lemma (Strongly convex case)

Suppose $\pi \propto \exp(-U)$ possesses an m-strongly convex potential U. Then

$$I_\pi(p) \geq m^{1/2} \cdot \varphi(\Phi^{-1}(p)) =: \tilde{I}_\pi(p),$$

where φ, Φ are the standard Gaussian p.d.f. and c.d.f., and furthermore

$$\tilde{I}_\pi(1/4) = m^{1/2} \cdot C_{\mathrm{g}},$$

where $C_{\mathrm{g}} \geq 0.317776$.

Close coupling for RWM

Prevously: provided we can choose δ such that $|x - y| \leq \delta \Rightarrow \|Q(x, \cdot) - Q(y, \cdot)\|_{\mathrm{TV}} \leq \alpha_0/2$, we obtain that P is close coupling with $\epsilon \geq \alpha_0/2$.

Close coupling for RWM

Prevously: provided we can choose δ such that $|x - y| \leq \delta \Rightarrow \|Q(x,\cdot) - Q(y,\cdot)\|_{\mathrm{TV}} \leq \alpha_0/2$, we obtain that P is close coupling with $\epsilon \geq \alpha_0/2$.

Since we have Gaussian $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ proposals, we can use Pinsker's inequality to obtain

Lemma

For $v > 0$,

$$|x - y| \leq v \cdot \sigma \Rightarrow \|Q(x,\cdot) - Q(y,\cdot)\|_{\mathrm{TV}} \leq v/2.$$

Close coupling for RWM

Prevously: provided we can choose δ such that $|x-y| \leq \delta \Rightarrow \|Q(x,\cdot) - Q(y,\cdot)\|_{\mathrm{TV}} \leq \alpha_0/2$, we obtain that P is close coupling with $\epsilon \geq \alpha_0/2$.

Since we have Gaussian $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ proposals, we can use Pinsker's inequality to obtain

Lemma

For $v > 0$,

$$|x-y| \leq v \cdot \sigma \Rightarrow \|Q(x,\cdot) - Q(y,\cdot)\|_{\mathrm{TV}} \leq v/2.$$

Thus by taking $v = \alpha_0$, i.e. $\delta = \alpha_0 \sigma$, we have that P is close coupling with $\epsilon = \alpha_0/2$.

Close coupling for RWM

Prevously: provided we can choose δ such that $|x-y| \le \delta \Rightarrow \|Q(x,\cdot)-Q(y,\cdot)\|_{\mathrm{TV}} \le \alpha_0/2$, we obtain that P is close coupling with $\epsilon \ge \alpha_0/2$.

Since we have Gaussian $\mathcal{N}(0,\sigma^2 \mathbf{I}_d)$ proposals, we can use Pinsker's inequality to obtain

Lemma

For $v > 0$,

$$|x-y| \le v \cdot \sigma \Rightarrow \|Q(x,\cdot)-Q(y,\cdot)\|_{\mathrm{TV}} \le v/2.$$

Thus by taking $v = \alpha_0$, i.e. $\delta = \alpha_0 \sigma$, we have that P is close coupling with $\epsilon = \alpha_0/2$.

So all that remains is to get a handle on α_0.

Controlling acceptance probabilities

We now assume that the potential U is *m-strongly convex* and *L-smooth*:

$$\frac{m}{2}|h|^2 \leq U(x+h) - U(x) - \langle \nabla U(x), h \rangle \leq \frac{L}{2}|h|^2, \quad x, h \in \mathsf{E}.$$

Controlling acceptance probabilities

We now assume that the potential U is m-strongly convex and L-smooth:

$$\frac{m}{2}|h|^2 \leq U(x+h) - U(x) - \langle \nabla U(x), h \rangle \leq \frac{L}{2}|h|^2, \quad x, h \in \mathrm{E}.$$

Then through a direct calculation, we obtain:

Lemma

Let $\sigma = \varsigma \cdot d^{-1/2} \cdot L^{-1/2}$, some $\varsigma > 0$. Then

$$\alpha_0 \geq \frac{1}{2} \cdot \exp\left(-\frac{\varsigma^2}{2}\right).$$

Main result

Putting together all of these pieces, we obtain the main result.

Theorem

We obtain the lower bound on the spectral gap of RWM, for $\sigma = \varsigma \cdot d^{-1/2} \cdot L^{-1/2}$

$$\gamma \geq 2^{-9} C_{\mathrm{g}}^2 \cdot \varsigma^2 \cdot \exp(-2\varsigma^2) \cdot d^{-1} \cdot \frac{m}{L}.$$

Main result

Putting together all of these pieces, we obtain the main result.

Theorem

We obtain the lower bound on the spectral gap of RWM, for $\sigma = \varsigma \cdot d^{-1/2} \cdot L^{-1/2}$

$$\gamma \geq 2^{-9} C_{\mathrm{g}}^2 \cdot \varsigma^2 \cdot \exp(-2\varsigma^2) \cdot d^{-1} \cdot \frac{m}{L}.$$

The upper bound on the spectral gap is derived through direct calculations.

Main result

Putting together all of these pieces, we obtain the main result.

Theorem

We obtain the lower bound on the spectral gap of RWM, for $\sigma = \varsigma \cdot d^{-1/2} \cdot L^{-1/2}$

$$\gamma \geq 2^{-9} C_{\mathrm{g}}^2 \cdot \varsigma^2 \cdot \exp(-2\varsigma^2) \cdot d^{-1} \cdot \frac{m}{L}.$$

The upper bound on the spectral gap is derived through direct calculations.

In the strongly convex, smooth case had a nice isoperimetric minorant; but can be applied in other cases too.

Using the full conductance profile can get much more intricate analysis of the mixing times.

Overview

Concluding remarks

I have presented explicit lower and upper bounds on the spectral gap of the RWM algorithm, focussing on the case of strongly convex and smooth potentials.

Concluding remarks

I have presented explicit lower and upper bounds on the spectral gap of the RWM algorithm, focussing on the case of strongly convex and smooth potentials.

However the general framework developed is applicable much more broadly!

Concluding remarks

I have presented explicit lower and upper bounds on the spectral gap of the RWM algorithm, focussing on the case of strongly convex and smooth potentials.

However the general framework developed is applicable much more broadly!

Furthermore the full conductance profile can give much more detailed mixing time bounds (not presented today; see paper).

Concluding remarks

I have presented explicit lower and upper bounds on the spectral gap of the RWM algorithm, focussing on the case of strongly convex and smooth potentials.

However the general framework developed is applicable much more broadly!

Furthermore the full conductance profile can give much more detailed mixing time bounds (not presented today; see paper).

Our paper also discusses the preconditioned Crank–Nicolson (pCN) algorithm a popular MCMC method for Bayesian Inverse Problems, which can be analysed in an analogous manner.

Concluding remarks

I have presented explicit lower and upper bounds on the spectral gap of the RWM algorithm, focussing on the case of strongly convex and smooth potentials.

However the general framework developed is applicable much more broadly!

Furthermore the full conductance profile can give much more detailed mixing time bounds (not presented today; see paper).

Our paper also discusses the preconditioned Crank–Nicolson (pCN) algorithm a popular MCMC method for Bayesian Inverse Problems, which can be analysed in an analogous manner.

Natural next steps would be to consider more advanced algorithms such as MALA, HMC, etc...

Overview

Thanks for listening! I

- Andrieu, C., Lee, A., Power, S., Wang, A. Q. (2022). Poincaré inequalities for Markov chains: a meeting with Cheeger, Lyapunov and Metropolis. *Technical report*. `https://doi.org/10.48550/arxiv.2208.05239`.
- **Andrieu, C., Lee, A., Power, S., Wang, A. Q. (2022)**. Explicit convergence bounds for Metropolis Markov chains: isoperimetry, spectral gaps and profiles. `https://doi.org/10.48550/arxiv.2211.08959`.
- Baxendale, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.*, 15(1B), 700738.
- Belloni, A., Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.*, 37(4), 20112055.
- Bobkov, S. G. (1999). Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27(4), 19031921.
- Chen, Y., Dwivedi, R., Wainwright, M. J., Yu, B. (2019). Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21.
- Dwivedi, R., Chen, Y., Wainwright, M. J., Yu, B. (2019). Log-concave sampling: Metropolis-Hastings algorithms are fast. *J. Mach. Learn. Res.*, 20, 142.
- Goel, S., Montenegro, R., Tetali, P. (2006). Mixing time bounds via the spectral profile. *Elec. J. Probab.*, 11(2000), 126.
- Jarner, S. F., Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stoc. Proc. Appl.*, 85(2), 341361.

Thanks for listening! II

Livingstone, S., Zanella, G. (2022). The Barker proposal: Combining robustness and efficiency in gradient-based MCMC. *J. Roy. Statist. Soc. Ser. B: Statist. Meth.*, 84(2), 496523.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21(6), 10871092.

Milman, E. (2009). On the role of convexity in isoperimetry, spectral gap and concentration. *Invent. Math.*, 177(1), 143.

Roberts, G. O., Gelman, A., Gilks, W. R. (1997). Weak Convergence and Optimal Scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1), 110120.

Roberts, G., Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1), 95110.

Convergence framework

We focus now on lower bounding the spectral gap γ of the RWM.

Convergence framework

We focus now on lower bounding the spectral gap γ of the RWM.

Recall a reversible chain P is positive if for any $f \in \mathrm{L}^2(\pi)$,

$$\langle Pf, f\rangle \geq 0.$$

Convergence framework

We focus now on lower bounding the spectral gap γ of the RWM.

Recall a reversible chain P is positive if for any $f \in \mathrm{L}^2(\pi)$,

$$\langle Pf, f \rangle \geq 0.$$

Lemma ([Baxendale (2005)])

RWM with Gaussian proposals is a positive chain.

Convergence of MCMC

What is the criteria for an MCMC chain to be ‘good’?

Convergence of MCMC

What is the criteria for an MCMC chain to be ‘good’?

Classically, MCMC is good if it converges fast to equilibrium and mixes well.

Convergence of MCMC

What is the criteria for an MCMC chain to be 'good'?

Classically, MCMC is good if it converges fast to equilibrium and mixes well.

One measure of the former is to look at rates of convergence:

Convergence of MCMC

What is the criteria for an MCMC chain to be 'good'?

Classically, MCMC is good if it converges fast to equilibrium and mixes well.

One measure of the former is to look at rates of convergence:

Theorem ([Roberts and Tweedie (1996), Jarner and Hansen (2000)])

RWM converges to equilibrium exponentially fast if and only if π has an exponential moment (e.g. $\pi(x) \propto \exp(-\|x-\mu\|^{\alpha})$, $\alpha \geq 1$.). Otherwise, the chain converges at a subgeometric (e.g. polynomial) rate.*

L^2 convergence and Dirichlet forms

We work on $\mathrm{L}^2(\pi) = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_2^2 < \infty\}$, $\quad \langle f, g \rangle := \int fg \, \mathrm{d}\pi$,
$\mathrm{L}^2_0(\pi) := \{f \in \mathrm{L}^2(\pi) : \pi(f) = 0\}$.

L^2 convergence and Dirichlet forms

We work on $\mathrm{L}^2(\pi) = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_2^2 < \infty\}$, $\quad \langle f, g\rangle := \int fg\,\mathrm{d}\pi$,
$\mathrm{L}_0^2(\pi) := \{f \in \mathrm{L}^2(\pi) : \pi(f) = 0\}$.

For a π-invariant Markov transition kernel P with $\mathrm{L}^2(\pi)$-adjoint P^*, define the Dirichlet form $\mathcal{E}(P^*P, f)$, for $f \in \mathrm{L}_0^2(\pi)$:

$$\mathcal{E}(P^*P, f) := \langle (I - P^*P)f, f\rangle = \|f\|^2 - \|Pf\|^2.$$

L^2 convergence and Dirichlet forms

We work on $\mathrm{L}^2(\pi) = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_2^2 < \infty\}$, $\quad \langle f, g\rangle := \int fg \, \mathrm{d}\pi$,
$\mathrm{L}_0^2(\pi) := \{f \in \mathrm{L}^2(\pi) : \pi(f) = 0\}$.

For a π-invariant Markov transition kernel P with $\mathrm{L}^2(\pi)$-adjoint P^*, define the Dirichlet form $\mathcal{E}(P^*P, f)$, for $f \in \mathrm{L}_0^2(\pi)$:

$$\mathcal{E}(P^*P, f) := \langle (I - P^*P)f, f\rangle = \|f\|^2 - \|Pf\|^2.$$

This acts like a discrete derivative, and we will seek to lower bound it.

L^2 convergence and Dirichlet forms

We work on $\mathrm{L}^2(\pi) = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_2^2 < \infty\}, \quad \langle f, g \rangle := \int fg \, \mathrm{d}\pi$,
$\mathrm{L}_0^2(\pi) := \{f \in \mathrm{L}^2(\pi) : \pi(f) = 0\}$.

For a π-invariant Markov transition kernel P with $\mathrm{L}^2(\pi)$-adjoint P^*, define the Dirichlet form $\mathcal{E}(P^*P, f)$, for $f \in \mathrm{L}_0^2(\pi)$:

$$\mathcal{E}(P^*P, f) := \langle (I - P^*P)f, f \rangle = \|f\|^2 - \|Pf\|^2.$$

This acts like a discrete derivative, and we will seek to lower bound it.

Furthermore if P is reversible and positive (so its spectrum $\sigma(P) \subset [0, 1]$), we have that

$$\mathcal{E}(P^*P, f) = \mathcal{E}(P^2, f) \geq \mathcal{E}(P, f).$$

L^2 convergence and Dirichlet forms

We work on $\mathrm{L}^2(\pi) = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_2^2 < \infty\}$, $\quad \langle f, g \rangle := \int fg \, \mathrm{d}\pi$,
$\mathrm{L}_0^2(\pi) := \{f \in \mathrm{L}^2(\pi) : \pi(f) = 0\}$.

For a π-invariant Markov transition kernel P with $\mathrm{L}^2(\pi)$-adjoint P^*, define the Dirichlet form $\mathcal{E}(P^*P, f)$, for $f \in \mathrm{L}_0^2(\pi)$:

$$\mathcal{E}(P^*P, f) := \langle (I - P^*P)f, f \rangle = \|f\|^2 - \|Pf\|^2.$$

This acts like a discrete derivative, and we will seek to lower bound it.

Furthermore if P is reversible and positive (so its spectrum $\sigma(P) \subset [0, 1]$), we have that

$$\mathcal{E}(P^*P, f) = \mathcal{E}(P^2, f) \geq \mathcal{E}(P, f).$$

So it will be sufficient to lower bound $\mathcal{E}(P, f)$.

Conductance

Definition: Conductance

The conductance profile of a π-invariant Markov kernel P is

$$\Phi_P(v) := \inf \left\{ \frac{(\pi \otimes P)(A \times A^\complement)}{\pi(A)} : \pi(A) \leq v \right\}.$$

The conductance of P is $\Phi_P^* := \Phi_P(1/2)$.

Conductance

Definition: Conductance

The conductance profile of a π-invariant Markov kernel P is

$$\Phi_P(v) := \inf \left\{ \frac{(\pi \otimes P)(A \times A^\complement)}{\pi(A)} : \pi(A) \le v \right\}.$$

The conductance of P is $\Phi_P^* := \Phi_P(1/2)$.

Theorem (Cheeger inequalities)

For a positive chain, such as RWM, we have the bounds on the spectral gap,

$$\frac{1}{2} \cdot [\Phi_P^*]^2 \le \gamma \le \Phi_P^*.$$

Conductance

Definition: Conductance

The conductance profile of a π-invariant Markov kernel P is

$$\Phi_P(v) := \inf \left\{ \frac{(\pi \otimes P)(A \times A^\complement)}{\pi(A)} : \pi(A) \leq v \right\}.$$

The conductance of P is $\Phi_P^* := \Phi_P(1/2)$.

Theorem (Cheeger inequalities)

For a positive chain, such as RWM, we have the bounds on the spectral gap,

$$\frac{1}{2} \cdot [\Phi_P^*]^2 \leq \gamma \leq \Phi_P^*.$$

Thus our goal is to lower bound the conductance.

Conductance and spectral profiles

Lemma ([Goel et. al. (2006)])

For nonconstant nonnegative $g \in \mathrm{L}^2_0(\pi)$*, we have the lower bound*

$$\mathcal{E}(P, g) \geq \mathrm{Var}_\pi(g) \cdot \frac{1}{2} \cdot \Lambda_P\left(\frac{4[\pi(g)]^2}{\mathrm{Var}_\pi(g)}\right),$$

where Λ_P *is the spectral profile of* P.

Lemma

For π*-reversible* P*, we have the further lower bound*

$$\Lambda_P(v) \geq \begin{cases} \frac{1}{2}\Phi_P(v)^2 & 0 < v \leq 1/2, \\ \frac{1}{2}[\Phi_P^*]^2 & v > 1/2. \end{cases}$$

Close coupling, conductance and isoperimetry

Theorem: Conductance lower bound; c.f. [Dwivedi et. al. (2019)]

Suppose $\tilde{I}_\pi$ is a regular, concave isoperimetric minorant of π. Let P be $(\mathrm{d}, \delta, \epsilon)$-close coupling. Then for any $v \in (0, 1/2]$,

$$\Phi_P(v) \geq \frac{1}{4} \cdot \epsilon \cdot 1 \wedge \left(\frac{\delta}{2} \cdot \frac{\tilde{I}_\pi(v/2)}{v/2} \right).$$

Close coupling, conductance and isoperimetry

Theorem: Conductance lower bound; c.f. [Dwivedi et. al. (2019)]

Suppose $\tilde{I}_\pi$ is a regular, concave isoperimetric minorant of π. Let P be $(\mathrm{d}, \delta, \epsilon)$-close coupling. Then for any $v \in (0, 1/2]$,

$$\Phi_P(v) \geq \frac{1}{4} \cdot \epsilon \cdot 1 \wedge \left(\frac{\delta}{2} \cdot \frac{\tilde{I}_\pi(v/2)}{v/2} \right).$$

Taking $v = 1/2$ immediately gives a lower bound on the conductance Φ_P^*, and hence on the spectral gap.

Close coupling, conductance and isoperimetry

Theorem: Conductance lower bound; c.f. [Dwivedi et. al. (2019)]

Suppose $\tilde{I}_\pi$ is a regular, concave isoperimetric minorant of π. Let P be $(\mathrm{d}, \delta, \epsilon)$-close coupling. Then for any $v \in (0, 1/2]$,

$$\Phi_P(v) \geq \frac{1}{4} \cdot \epsilon \cdot 1 \wedge \left(\frac{\delta}{2} \cdot \frac{\tilde{I}_\pi(v/2)}{v/2} \right).$$

Taking $v = 1/2$ immediately gives a lower bound on the conductance Φ_P^*, and hence on the spectral gap.

This result thus breaks the problem into two pieces:

- For a given target π, establish a regular concave isoperimetric minorant $\tilde{I}_\pi$.
- For the chain P, establish close coupling.