Comparisons theorems for slice sampling

Andi Q. Wang

University of Warwick

Joint with: Daniel Rudolf, Sam Power, Björn Sprungk.

Algorithms seminar, Warwick

30th January 2024

Overview

Overview

Statistical modelling

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

Statistical modelling

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

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

Statistical modelling

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

Posit a model (density function) $f_x(y)$ which generated y, which depends upon 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

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

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

Popular approach: 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

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

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

Random walk Metropolis

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

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

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

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.

In [Andrieu, Lee, Power, W. (2022)], showed that the spectral gap for nice densities decays like $O(d^{-1})$,

Random walk Metropolis

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

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.

In [Andrieu, Lee, Power, W. (2022)], showed that the spectral gap for nice densities decays like $O(d^{-1})$, and RWM will converge very slowly for heavy-tailed and/or multimodal distributions.

Random walk Metropolis

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

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.

In [Andrieu, Lee, Power, W. (2022)], showed that the spectral gap for nice densities decays like $O(d^{-1})$, and RWM will converge very slowly for heavy-tailed and/or multimodal distributions.

Slice Sampling was introduced to try and circumvent these deficiencies.

Slice Sampling [Neal (2003)]

Target: $\pi(\mathrm{d}x) = \varpi(x)\nu(\mathrm{d}x)$ on $\mathcal{X}$, reference measure ν.

$$G(t) := \{x \in \mathcal{X} : \varpi(x) > t\}, \quad \nu_t := \frac{\nu(\cdot \cap G(t))}{\nu(G(t))}.$$

Slice Sampling [Neal (2003)]

$$\text{Target: } \pi(\mathrm{d}x) = \varpi(x)\nu(\mathrm{d}x) \quad \text{on } \mathcal{X}, \quad \text{reference measure } \nu.$$

$$G(t) := \{x \in \mathcal{X} : \varpi(x) > t\}, \quad \nu_t := \frac{\nu(\cdot \cap G(t))}{\nu(G(t))}.$$

Algorithm 3 Ideal Slice Sampling

1: *initialise:* $X_0 = x_0, i = 0$
2: **while** $i < N$ **do**
3: $\quad i \leftarrow i + 1$
4: $\quad$ **Sample** $t \sim \mathrm{Unif}([0, \varpi(x)])$;
5: $\quad$ **Sample** $Y \sim \nu_t$;
6: $\quad$ **Set** $X_i = Y$.
7: **return** $(X_i)_{i=1,\ldots,N}$

Slice Sampling

Slice Sampling defines a π-reversible Markov chain.

Slice Sampling

Slice Sampling defines a π-reversible Markov chain.

It possesses no tuning parameters and is entirely rejection-free.

Slice Sampling

Slice Sampling defines a π-reversible Markov chain.

It possesses no tuning parameters and is entirely rejection-free.

Convergence properties (spectral gap) have been discussed in [Natarovskii et. al. (2021)].

Slice Sampling

Slice Sampling defines a π-reversible Markov chain.

It possesses no tuning parameters and is entirely rejection-free.

Convergence properties (spectral gap) have been discussed in [Natarovskii et. al. (2021)].

Catch: sampling from ν_t is in general intractable.

Slice Sampling

Slice Sampling defines a π-reversible Markov chain.

It possesses no tuning parameters and is entirely rejection-free.

Convergence properties (spectral gap) have been discussed in [Natarovskii et. al. (2021)].

Catch: sampling from ν_t is in general intractable.

But this can be relaxed, if instead we have access to ν_t-reversible kernels!

Slice Sampling

Slice Sampling defines a π-reversible Markov chain.

It possesses no tuning parameters and is entirely rejection-free.

Convergence properties (spectral gap) have been discussed in [Natarovskii et. al. (2021)].

Catch: sampling from ν_t is in general intractable.

But this can be relaxed, if instead we have access to ν_t-reversible kernels!

(Similar to going from Gibbs sampling $\rightsquigarrow$ Metropolis-within-Gibbs.)

Hybrid Slice Sampling

Suppose we have a family of kernels (H_t) where each H_t is ν_t-reversible (e.g. RWM on the slice).

Hybrid Slice Sampling

Suppose we have a family of kernels (H_t) where each H_t is ν_t-reversible (e.g. RWM on the slice).

Algorithm 5 Hybrid Slice Sampling (HSS)

1: *initialise:* $X_0 = x_0, i = 0$
2: **while** $i < N$ **do**
3: $\quad i \leftarrow i + 1$
4: $\quad$ **Sample** $t \sim \mathrm{Unif}([0, \varpi(x)])$;
5: $\quad$ **Sample** $Y \sim H_t(X_{i-1}, \cdot)$;
6: $\quad$ **Set** $X_i = Y$.
7: **return** $(X_i)_{i=1,\dots,n}$

This is still a π-reversible Markov chain.

Hybrid Slice Sampling examples

HSS also defines a π-reversible Markov chain.

- Random Walk Metropolis on the slice
- Hit-and-Run
- Step out and shrinkage
- ...

Hybrid Slice Sampling examples

HSS also defines a π-reversible Markov chain.

- Random Walk Metropolis on the slice
- Hit-and-Run
- Step out and shrinkage
- ...

It is known that in each case, the performance of HSS is worse than the original ISS (which cannot be implemented), since each H_t is really trying to approximate ν_t [Rudolf and Ullrich (2018)].

Hybrid Slice Sampling examples

HSS also defines a π-reversible Markov chain.

- Random Walk Metropolis on the slice
- Hit-and-Run
- Step out and shrinkage
- ...

It is known that in each case, the performance of HSS is worse than the original ISS (which cannot be implemented), since each H_t is really trying to approximate ν_t [Rudolf and Ullrich (2018)].

Question: how much worse?

Example: Metropolis chains

Consider the special case of Metropolis–Hastings where the proposal Q is ν-reversible, for some measure ν.

Example: Metropolis chains

Consider the special case of Metropolis–Hastings where the proposal Q is ν-reversible, for some measure ν.

For example: $\nu = \mathrm{Leb}$, $Q(x, \cdot) = \mathcal{N}(x, \sigma^2 \cdot \mathbf{I})$, Random Walk Metropolis.

Example: Metropolis chains

Consider the special case of Metropolis–Hastings where the proposal Q is ν-reversible, for some measure ν.

For example: $\nu = \mathrm{Leb}$, $Q(x,\cdot) = \mathcal{N}(x, \sigma^2 \cdot \mathbf{I})$, Random Walk Metropolis.
$\nu = \mathcal{N}(0, \mathbf{C})$, $Q(x,\cdot) = \mathcal{N}\left(\rho x, (1-\rho^2) \cdot \mathbf{C}\right)$, preconditioned Crank–Nicolson.

Example: Metropolis chains

Consider the special case of Metropolis–Hastings where the proposal Q is ν-reversible, for some measure ν.

For example: $\nu = \mathrm{Leb}$, $Q(x, \cdot) = \mathcal{N}(x, \sigma^2 \cdot \mathbf{I})$, Random Walk Metropolis.
$\nu = \mathcal{N}(0, \mathbf{C})$, $Q(x, \cdot) = \mathcal{N}\left(\rho x, (1-\rho^2) \cdot \mathbf{C}\right)$, preconditioned Crank–Nicolson.

It turns out that such chains can actually be seen as examples of Hybrid Slice Samplers.

Example: Metropolis chains

Consider the special case of Metropolis–Hastings where the proposal Q is ν-reversible, for some measure ν.

For example: $\nu = \mathrm{Leb}$, $Q(x, \cdot) = \mathcal{N}(x, \sigma^2 \cdot \mathbf{I})$, Random Walk Metropolis.
$\nu = \mathcal{N}(0, \mathbf{C})$, $Q(x, \cdot) = \mathcal{N}(\rho x, (1 - \rho^2) \cdot \mathbf{C})$, preconditioned Crank–Nicolson.

It turns out that such chains can actually be seen as examples of Hybrid Slice Samplers.
Intuition: the ideal SS first draws an acceptance region, then a proposed point conditional on being in this acceptance region.

Example: Metropolis chains

Consider the special case of Metropolis–Hastings where the proposal Q is ν-reversible, for some measure ν.

For example: $\nu = \mathrm{Leb}$, $Q(x,\cdot) = \mathcal{N}(x, \sigma^2 \cdot \mathbf{I})$, Random Walk Metropolis.
$\nu = \mathcal{N}(0, \mathbf{C})$, $Q(x,\cdot) = \mathcal{N}(\rho x, (1-\rho^2)\cdot \mathbf{C})$, preconditioned Crank–Nicolson.

It turns out that such chains can actually be seen as examples of Hybrid Slice Samplers.
Intuition: the ideal SS first draws an acceptance region, then a proposed point conditional on being in this acceptance region.

Thus our subsequent results allow us to study such chains.

Summary

We will present a comparison result which can quantify precisely how much worse the HSS is compared to the ISS.

Summary

We will present a comparison result which can quantify precisely how much worse the HSS is compared to the ISS.

Intuitively, it relates the convergence of HSS to the convergence rate of ISS in terms of the convergence rates of the (H_t) kernels.

Summary

We will present a comparison result which can quantify precisely how much worse the HSS is compared to the ISS.

Intuitively, it relates the convergence of HSS to the convergence rate of ISS in terms of the convergence rates of the (H_t) kernels.

We will do this comparison using the framework of weak Poincaré inequalities.

Summary

We will present a comparison result which can quantify precisely how much worse the HSS is compared to the ISS.

Intuitively, it relates the convergence of HSS to the convergence rate of ISS in terms of the convergence rates of the (H_t) kernels.

We will do this comparison using the framework of weak Poincaré inequalities.

This will enable us to give quantitative bounds, convering cases when there is no spectral gap, i.e. the convergence is subgeometric, substantially extending previous results [Qin et. al. (2023), Łatuszyński & Rudolf (2014)].

Overview

Convergence of Markov chains

Recall that a reversible π-invariant Markov kernel P defines a (self-adjoint) operator on $L^2(\pi)$.

Convergence of Markov chains

Recall that a reversible π-invariant Markov kernel P defines a (self-adjoint) operator on $L^2(\pi)$.

Its convergence to equilibrium can be bounded by the spectral gap γ (and this is the best rate):

Convergence of Markov chains

Recall that a reversible π-invariant Markov kernel P defines a (self-adjoint) operator on $L^2(\pi)$.

Its convergence to equilibrium can be bounded by the spectral gap γ (and this is the best rate):

$$\|P^n f - \pi(f)\|_2 \leq (1-\gamma)^n \|f\|_2, \quad \forall f \in L^2(\pi).$$

Convergence of Markov chains

Recall that a reversible π-invariant Markov kernel P defines a (self-adjoint) operator on $L^2(\pi)$.

Its convergence to equilibrium can be bounded by the spectral gap γ (and this is the best rate):

$$\|P^n f - \pi(f)\|_2 \leq (1-\gamma)^n \|f\|_2, \quad \forall f \in L^2(\pi).$$

However some chains have 0 spectral gap and have only subgeometric convergence;

$$\|P^n f - \pi(f)\|_2^2 \leq \gamma(n)\Phi(f),$$

where

$$\Phi(f) = \|f\|_{\mathrm{osc}}^2 = (\operatorname{ess\,sup} f - \operatorname{ess\,inf} f)^2.$$

Standard Poincaré inequalities

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

Standard Poincaré inequalities

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

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

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

Standard Poincaré inequalities

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

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

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

Standard Poincaré inequality (SPI)

A SPI holds if there exists a constant $C_{\mathrm{P}} > 0$ such that for all $f \in \mathrm{L}^2_0(\mu)$,

$$C_{\mathrm{P}}\|f\|_2^2 \leq \mathcal{E}(P^*P, f).$$

Geometric convergence / spectral gap

$$C_{\mathrm{P}} \|f\|_2^2 \leq \mathcal{E}(P^* P, f).$$

Geometric convergence / spectral gap

$$C_{\mathrm{P}} \|f\|_2^2 \le \mathcal{E}(P^* P, f).$$

Theorem (Geometric convergence)

Under a standard Poincaré inequality, we have for all $f \in \mathrm{L}^2_0(\mu)$, $n \in \mathbb{N}_0$,

$$\|P^n f\|_2^2 \le (1 - C_{\mathrm{P}})^n \|f\|_2^2.$$

Geometric convergence / spectral gap

$$C_{\mathrm{P}} \|f\|_2^2 \le \mathcal{E}(P^*P, f).$$

Theorem (Geometric convergence)

Under a standard Poincaré inequality, we have for all $f \in \mathrm{L}_0^2(\mu)$, $n \in \mathbb{N}_0$,

$$\|P^n f\|_2^2 \le (1 - C_{\mathrm{P}})^n \|f\|_2^2.$$

Proof. Rewriting the SPI, see $\mathcal{E}(P^*P, f)$ behaves like a discrete derivative:

$$C_{\mathrm{P}} \|f\|_2^2 \le \mathcal{E}(P^* P, f).$$

Theorem (Geometric convergence)

Under a standard Poincaré inequality, we have for all $f \in \mathrm{L}^2_0(\mu)$, $n \in \mathbb{N}_0$,

$$\|P^n f\|_2^2 \le (1 - C_{\mathrm{P}})^n \|f\|_2^2.$$

Proof. Rewriting the SPI, see $\mathcal{E}(P^* P, f)$ behaves like a discrete derivative:

$$\begin{aligned} C_{\mathrm{P}} \|f\|_2^2 \le \mathcal{E}(P^* P, f) &= \|f\|_2^2 - \langle P^* P f, f \rangle \\ &= \|f\|_2^2 - \|Pf\|_2^2 \\ \Rightarrow \|Pf\|_2^2 \le (1 - C_{\mathrm{P}}) \|f\|_2^2. \end{aligned}$$

Geometric convergence / spectral gap

$$C_{\mathrm{P}}\|f\|_2^2 \le \mathcal{E}(P^*P, f).$$

Theorem (Geometric convergence)

Under a standard Poincaré inequality, we have for all $f \in \mathrm{L}_0^2(\mu)$, $n \in \mathbb{N}_0$,

$$\|P^n f\|_2^2 \le (1 - C_{\mathrm{P}})^n \|f\|_2^2.$$

Proof. Rewriting the SPI, see $\mathcal{E}(P^*P, f)$ behaves like a discrete derivative:

$$\begin{aligned} C_{\mathrm{P}}\|f\|_2^2 \le \mathcal{E}(P^*P, f) &= \|f\|_2^2 - \langle P^*Pf, f\rangle \\ &= \|f\|_2^2 - \|Pf\|_2^2 \\ \Rightarrow \|Pf\|_2^2 &\le (1 - C_{\mathrm{P}})\|f\|_2^2. \end{aligned}$$

The rest is by induction. □

SPI to weak Poincaré inequality

$$C_{\mathrm{P}}\|f\|_2^2 \le \mathcal{E}(P^*P, f).$$

SPI to weak Poincaré inequality

$$C_{\mathrm{P}} \|f\|_2^2 \le \mathcal{E}(P^* P, f).$$

We now generalize this to allow for subgeometric rates of convergence:

SPI to weak Poincaré inequality

$$C_{\mathrm{P}}\|f\|_2^2 \le \mathcal{E}(P^*P, f).$$

We now generalize this to allow for subgeometric rates of convergence:

Require $\beta : (0,\infty) \to [0,\infty)$ decreasing with $\beta(s) \downarrow 0$ as $s \to \infty$ and $\Phi : \mathrm{L}^2(\mu) \to [0,\infty]$ given by $\Phi(f) = \|f\|_{\mathrm{osc}}^2 = (\mathrm{ess}_\mu \sup f - \mathrm{ess}_\mu \inf f)^2$.

SPI to weak Poincaré inequality

$$C_{\mathrm{P}} \|f\|_2^2 \leq \mathcal{E}(P^* P, f).$$

We now generalize this to allow for subgeometric rates of convergence:

Require $\beta : (0, \infty) \to [0, \infty)$ decreasing with $\beta(s) \downarrow 0$ as $s \to \infty$ and $\Phi : \mathrm{L}^2(\mu) \to [0, \infty]$ given by $\Phi(f) = \|f\|_{\mathrm{osc}}^2 = (\mathrm{ess}_\mu \sup f - \mathrm{ess}_\mu \inf f)^2$.

Weak Poincaré inequality (WPI) (*c.f.* [Röckner and Wang (2001)])

A WPI holds if:

SPI to weak Poincaré inequality

$$C_{\mathrm{P}} \|f\|_2^2 \leq \mathcal{E}(P^* P, f).$$

We now generalize this to allow for subgeometric rates of convergence:

Require $\beta : (0, \infty) \to [0, \infty)$ decreasing with $\beta(s) \downarrow 0$ as $s \to \infty$ and $\Phi : \mathrm{L}^2(\mu) \to [0, \infty]$ given by $\Phi(f) = \|f\|_{\mathrm{osc}}^2 = (\mathrm{ess}_\mu \sup f - \mathrm{ess}_\mu \inf f)^2$.

Weak Poincaré inequality (WPI) (*c.f.* [Röckner and Wang (2001)])

A WPI holds if: for some such β, Φ, $\forall s > 0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\|f\|_2^2 \leq s\, \mathcal{E}(P^* P, f) + \beta(s) \Phi(f).$$

SPI to weak Poincaré inequality

$$C_{\mathrm{P}} \|f\|_2^2 \leq \mathcal{E}(P^* P, f).$$

We now generalize this to allow for subgeometric rates of convergence:

Require $\beta : (0, \infty) \to [0, \infty)$ decreasing with $\beta(s) \downarrow 0$ as $s \to \infty$ and $\Phi : \mathrm{L}^2(\mu) \to [0, \infty]$ given by $\Phi(f) = \|f\|_{\mathrm{osc}}^2 = (\mathrm{ess}_\mu \sup f - \mathrm{ess}_\mu \inf f)^2$.

Weak Poincaré inequality (WPI) (*c.f.* [Röckner and Wang (2001)])

A WPI holds if: for some such β, Φ, $\forall s > 0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\|f\|_2^2 \leq s\, \mathcal{E}(P^* P, f) + \beta(s)\Phi(f).$$

E.g. $\beta(s) = c_0 s^{-c_1}$.

Subgeometric convergence

$$\|f\|_2^2 \le s\,\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0, f \in \mathrm{L}_0^2(\mu).$$

Subgeometric convergence

$$\|f\|_2^2 \leq s\,\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0, f \in \mathrm{L}_0^2(\mu).$$

Define

$$\begin{aligned} K(u) &:= u\beta(1/u), \qquad u \geq 0, \\ K^*(v) &:= \sup_{u \geq 0}\{uv - K(u)\}, \quad v \geq 0, \\ F(x) &:= \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \qquad 0 < x \leq 1. \end{aligned}$$

Subgeometric convergence

$$\|f\|_2^2 \leq s\,\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0, f \in \mathrm{L}_0^2(\mu).$$

Define

$$\begin{aligned} K(u) &:= u\beta(1/u), && u \geq 0, \\ K^*(v) &:= \sup_{u \geq 0}\{uv - K(u)\}, && v \geq 0, \\ F(x) &:= \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, && 0 < x \leq 1. \end{aligned}$$

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

Under a weak Poincaré inequality, we have, $\forall n \in \mathbb{N}_0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\|P^n f\|_2^2 \leq \Phi(f)F^{-1}(n).$$

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

Under a weak Poincaré inequality, we have, $\forall n \in \mathbb{N}_0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\|P^n f\|_2^2 \leq \Phi(f) F^{-1}(n).$$

If $\beta(s) = c_0 s^{-c_1}$, we can bound

$$F^{-1}(n) \leq C n^{-c_1}.$$

Examples

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

Under a weak Poincaré inequality, we have, $\forall n \in \mathbb{N}_0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\|P^n f\|_2^2 \leq \Phi(f) F^{-1}(n).$$

If $\beta(s) = c_0 s^{-c_1}$, we can bound

$$F^{-1}(n) \leq C n^{-c_1}.$$

If $\beta(s) = \eta_0 \exp(-\eta_1 s^{\eta_2})$, we can bound

$$F^{-1}(n) \leq C' \exp\left(-(Cn)^{\eta_2/(1+\eta_2)}\right).$$

Examples

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

Under a weak Poincaré inequality, we have, $\forall n \in \mathbb{N}_0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\|P^n f\|_2^2 \leq \Phi(f) F^{-1}(n).$$

If $\beta(s) = c_0 s^{-c_1}$, we can bound

$$F^{-1}(n) \leq C n^{-c_1}.$$

If $\beta(s) = \eta_0 \exp(-\eta_1 s^{\eta_2})$, we can bound

$$F^{-1}(n) \leq C' \exp\left(-(Cn)^{\eta_2/(1+\eta_2)}\right).$$

Intuition: the faster β decays, the faster the rate of convergence.

Comparisons of Markov chains

Furthermore we can compare Markov chains. Suppose we have chains $P, \tilde{P}$ which are both π-reversible (think of $\tilde{P}$ as an approximation of P).

Comparisons of Markov chains

Furthermore we can compare Markov chains. Suppose we have chains $P, \tilde{P}$ which are both π-reversible (think of $\tilde{P}$ as an approximation of P).

Suppose $\forall s > 0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\mathcal{E}(P, f) \leq s\,\mathcal{E}(\tilde{P}, f) + \beta(s)\Phi(f).$$

Comparisons of Markov chains

Furthermore we can compare Markov chains. Suppose we have chains $P, \tilde{P}$ which are both π-reversible (think of $\tilde{P}$ as an approximation of P).

Suppose $\forall s > 0$, $f \in \mathrm{L}^2_0(\mu)$,

$$\mathcal{E}(P, f) \leq s\,\mathcal{E}(\tilde{P}, f) + \beta(s)\Phi(f).$$

This implies that $\tilde{P}$ converges at a rate governed by β, relative to P. In other words, β controls the degradation in convergence when we move from P to $\tilde{P}$.

Comparisons of Markov chains

Furthermore we can compare Markov chains. Suppose we have chains $P, \tilde{P}$ which are both π-reversible (think of $\tilde{P}$ as an approximation of P).

Suppose $\forall s > 0$, $f \in \mathrm{L}^2_0(\mu)$,

$$\mathcal{E}(P, f) \leq s\,\mathcal{E}(\tilde{P}, f) + \beta(s)\Phi(f).$$

This implies that $\tilde{P}$ converges at a rate governed by β, relative to P. In other words, β controls the degradation in convergence when we move from P to $\tilde{P}$.

In [Andrieu, Lee, Power, W. (2022)], we used this machinery to study pseudo-marginal MCMC.

Comparisons of Markov chains II

Suppose we are able to establish the comparison: $\forall s > 0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\mathcal{E}(P, f) \leq s\,\mathcal{E}(\tilde{P}, f) + \beta(s)\Phi(f).$$

Comparisons of Markov chains II

Suppose we are able to establish the comparison: $\forall s > 0$, $f \in \mathrm{L}^2_0(\mu)$,

$$\mathcal{E}(P, f) \leq s\,\mathcal{E}(\tilde{P}, f) + \beta(s)\Phi(f).$$

Suppose the 'ideal' chain P possesses a spectral gap, i.e. it satisfies a SPI:

$$C_{\mathrm{P}}\|f\|_2^2 \leq \mathcal{E}(P, f).$$

Comparisons of Markov chains II

Suppose we are able to establish the comparison: $\forall s > 0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\mathcal{E}(P, f) \leq s\,\mathcal{E}(\tilde{P}, f) + \beta(s)\Phi(f).$$

Suppose the 'ideal' chain P possesses a spectral gap, i.e. it satisfies a SPI:

$$C_{\mathrm{P}}\|f\|_2^2 \leq \mathcal{E}(P, f).$$

Then by combining these inequalities,

$$\|f\|_2^2 \leq s\,\mathcal{E}(\tilde{P}, f) + C_{\mathrm{P}}^{-1}\beta(C_{\mathrm{P}} \cdot s)\Phi(f),$$

Comparisons of Markov chains II

Suppose we are able to establish the comparison: $\forall s > 0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\mathcal{E}(P, f) \leq s\,\mathcal{E}(\tilde{P}, f) + \beta(s)\Phi(f).$$

Suppose the 'ideal' chain P possesses a spectral gap, i.e. it satisfies a SPI:

$$C_{\mathrm{P}}\|f\|_2^2 \leq \mathcal{E}(P, f).$$

Then by combining these inequalities,

$$\|f\|_2^2 \leq s\,\mathcal{E}(\tilde{P}, f) + C_{\mathrm{P}}^{-1}\beta(C_{\mathrm{P}} \cdot s)\Phi(f),$$

from which a convergence bound for $\tilde{P}$ can be obtained.

Overview

Main result

We have $\pi(\mathrm{d}x) = \varpi(x)\nu(\mathrm{d}x)$

$$m_t := \nu(G(t)),$$

recall $\nu_t = \nu(\cdot \cap G(t))/m_t$. Let's write U for Ideal Slice Sampling, H for Hybrid Slice Sampling.

Main result

We have $\pi(\mathrm{d}x) = \varpi(x)\nu(\mathrm{d}x)$

$$m_t := \nu(G(t)),$$

recall $\nu_t = \nu(\cdot \cap G(t))/m_t$. Let's write U for Ideal Slice Sampling, H for Hybrid Slice Sampling.

Theorem (Rudolf, Power, Sprungk, W. (2023+))

Suppose each H_t is ν_t-reversible, positive and satisfies a WPI with function β_t.

Main result

We have $\pi(\mathrm{d}x) = \varpi(x)\nu(\mathrm{d}x)$

$$m_t := \nu(G(t)),$$

recall $\nu_t = \nu(\cdot \cap G(t))/m_t$. Let's write U for Ideal Slice Sampling, H for Hybrid Slice Sampling.

Theorem (Rudolf, Power, Sprungk, W. (2023+))

Suppose each H_t is ν_t-reversible, positive and satisfies a WPI with function β_t. We have the comparison: $\forall s > 0$, $f \in \mathrm{L}_0^2(\mu)$,

$$\mathcal{E}(U, f) \leq s \cdot \mathcal{E}(H, f) + \beta(s)\Phi(f),$$

where $\beta : (0, \infty) \to [0, \infty)$ *is given by*

$$\beta(s) := \int_0^{\|\varpi\|_\infty} \beta_t(s)\, m_t \,\mathrm{d}t,$$

Comments

$$\mathcal{E}(U, f) \leq s \cdot \mathcal{E}(H, f) + \beta(s)\Phi(f), \qquad \beta(s) := \int_0^{\|\pi\|_\infty} \beta_t(s)\, m_t \,\mathrm{d}t.$$

We see that the convergence rate of U is a kind of weighted average of the convergence rates of each H_t.

Comments

$$\mathcal{E}(U, f) \leq s \cdot \mathcal{E}(H, f) + \beta(s)\Phi(f), \qquad \beta(s) := \int_0^{\|\pi\|_\infty} \beta_t(s)\, m_t \,\mathrm{d}t.$$

We see that the convergence rate of U is a kind of weighted average of the convergence rates of each H_t.

For instance, if each H_t has a uniform spectral gap bound, then so does H (relative to U).

Comments

$$\mathcal{E}(U,f) \le s \cdot \mathcal{E}(H,f) + \beta(s)\Phi(f), \qquad \beta(s) := \int_0^{\|\pi\|_\infty} \beta_t(s)\, m_t \,\mathrm{d}t.$$

We see that the convergence rate of U is a kind of weighted average of the convergence rates of each H_t.

For instance, if each H_t has a uniform spectral gap bound, then so does H (relative to U).

However if ‘enough’ of the H_t are subgeometric, or the spectral gaps decay to 0, H will only be subgeometric (relative to U).

Comments

$$\mathcal{E}(U, f) \le s \cdot \mathcal{E}(H, f) + \beta(s)\Phi(f), \qquad \beta(s) := \int_0^{\|\pi\|_\infty} \beta_t(s)\, m_t \, \mathrm{d}t.$$

We see that the convergence rate of U is a kind of weighted average of the convergence rates of each H_t.

For instance, if each H_t has a uniform spectral gap bound, then so does H (relative to U).

However if 'enough' of the H_t are subgeometric, or the spectral gaps decay to 0, H will only be subgeometric (relative to U).

If we further have a convergence estimate for U e.g. [Natarovskii et. al. (2021)], then these can be combined to give a convergence bound for H.

Comments

$$\mathcal{E}(U, f) \leq s \cdot \mathcal{E}(H, f) + \beta(s)\Phi(f), \qquad \beta(s) := \int_0^{\|\pi\|_\infty} \beta_t(s)\, m_t \,\mathrm{d}t.$$

We see that the convergence rate of U is a kind of weighted average of the convergence rates of each H_t.

For instance, if each H_t has a uniform spectral gap bound, then so does H (relative to U).

However if 'enough' of the H_t are subgeometric, or the spectral gaps decay to 0, H will only be subgeometric (relative to U).

If we further have a convergence estimate for U e.g. [Natarovskii et. al. (2021)], then these can be combined to give a convergence bound for H.

This significantly extends the previous work of [Łatuszyński & Rudolf (2014)].

Example: Independent Metropolis–Hastings (IMH)

Recall the IMH(π, q) chain: at each iteration with $X_n = x$, propose $Y \sim q$, and accept this proposal with probability

$$\alpha(x, Y) = 1 \wedge \frac{\pi(Y)q(x)}{\pi(x)q(Y)}.$$

It is known that the IMH(π, q) satisfies an SPI with constant $\left\|\mathrm{d}\pi/\mathrm{d}q\right\|_\infty^{-1}$ [Mengersen and Tweedie (1996)].

Explicit example: Exponential distributions

Consider the case when $\mathcal{X} = [0, \infty)$, $\pi(x) = \exp(-x)$, and $\nu(\mathrm{d}x) = \lambda \exp(-\lambda x)\,\mathrm{d}x$ for some $\lambda > 0$.

Explicit example: Exponential distributions

Consider the case when $\mathcal{X} = [0, \infty)$, $\pi(x) = \exp(-x)$, and $\nu(\mathrm{d}x) = \lambda \exp(-\lambda x)\,\mathrm{d}x$ for some $\lambda > 0$.

So

$$\pi(x) = \lambda^{-1} \exp(-(1-\lambda)x)\,\nu(\mathrm{d}x).$$

Explicit example: Exponential distributions

Consider the case when $\mathcal{X} = [0, \infty)$, $\pi(x) = \exp(-x)$, and $\nu(\mathrm{d}x) = \lambda \exp(-\lambda x)\,\mathrm{d}x$ for some $\lambda > 0$.

So

$$\pi(x) = \lambda^{-1} \exp(-(1-\lambda)x)\,\nu(\mathrm{d}x).$$

The level sets $G(t)$ are determined by $\exp(-(1-\lambda)x)$, and the Slice Sampler will sample from ν restricted to $G(t)$.

Explicit example: Exponential distributions

Consider the case when $\mathcal{X} = [0, \infty)$, $\pi(x) = \exp(-x)$, and $\nu(\mathrm{d}x) = \lambda \exp(-\lambda x)\,\mathrm{d}x$ for some $\lambda > 0$.

So

$$\pi(x) = \lambda^{-1} \exp(-(1-\lambda)x)\,\nu(\mathrm{d}x).$$

The level sets $G(t)$ are determined by $\exp(-(1-\lambda)x)$, and the Slice Sampler will sample from ν restricted to $G(t)$.

Lemma

This Ideal Slice Sampler has a spectral gap which can be lower-bounded explicitly:

- *When $\lambda \in (0, 1)$, the slice sampler has a spectral gap of at least $\frac{1+\lambda}{2}$;*
- *When $\lambda > 1$, the slice sampler has a spectral gap of at least $(2\lambda - 1)^{-2}$.*

Explicit example: Exponential distributions

Consider the case when $\mathcal{X} = [0, \infty)$, $\pi(x) = \exp(-x)$, and $\nu(\mathrm{d}x) = \lambda \exp(-\lambda x)\,\mathrm{d}x$ for some $\lambda > 0$.

So

$$\pi(x) = \lambda^{-1} \exp(-(1-\lambda)x)\, \nu(\mathrm{d}x).$$

The level sets $G(t)$ are determined by $\exp(-(1-\lambda)x)$, and the Slice Sampler will sample from ν restricted to $G(t)$.

Lemma

This Ideal Slice Sampler has a spectral gap which can be lower-bounded explicitly:

- *When $\lambda \in (0,1)$, the slice sampler has a spectral gap of at least $\frac{1+\lambda}{2}$;*
- *When $\lambda > 1$, the slice sampler has a spectral gap of at least $(2\lambda - 1)^{-2}$.*

Derived using a contractivity argument.

Hybrid case

Consider doing now performing an IMH with proposal ν on each slice:

$$H_t = \mathrm{IMH}\left(\nu_t, \nu\right).$$

Hybrid case

Consider doing now performing an IMH with proposal ν on each slice:

$$H_t = \mathrm{IMH}\left(\nu_t, \nu\right).$$

When $\lambda \in (0,1)$, we obtain the following comparison for the IMH:

$$\beta(s) = \frac{1}{4} + \frac{1}{4}(1 - s^{-1})^{1/\lambda}\left(\frac{1}{\lambda(s-1)} - 1\right) \sim s^{-1},$$

and a WPI for the IMH-within-Slice Sampler with $\tilde{\beta}(s) = \frac{2\beta\big((1+\lambda)s/2\big)}{1+\lambda} \sim s^{-1}$.

Hybrid case

Consider doing now performing an IMH with proposal ν on each slice:

$$H_t = \mathrm{IMH}\left(\nu_t, \nu\right).$$

When $\lambda \in (0,1)$, we obtain the following comparison for the IMH:

$$\beta(s) = \frac{1}{4} + \frac{1}{4}(1 - s^{-1})^{1/\lambda}\left(\frac{1}{\lambda(s-1)} - 1\right) \sim s^{-1},$$

and a WPI for the IMH-within-Slice Sampler with $\tilde{\beta}(s) = \frac{2\beta\left((1+\lambda)s/2\right)}{1+\lambda} \sim s^{-1}$.

When $\lambda > 1$, we obtain the following comparison for the IMH:

$$\beta(s) = \frac{\lambda - 1}{4\lambda} s^{-1/\lambda},$$

and a WPI for the IMH-within-Slice Sampler with $\tilde{\beta}(s) = (2\lambda - 1)^2 \beta\left((2\lambda - 1)^{-2} s\right)$.

Example: Hit-and-Run

An example of a HSS: on each slice, pick a random direction, then sample a uniform point on this line intersected with the slice.

Example: Hit-and-Run

An example of a HSS: on each slice, pick a random direction, then sample a uniform point on this line intersected with the slice.

From [Lovàsz and Vempala (2004)], we deduce that for L-smooth and m-strongly concave potentials the spectral gap of hit-and-run on a bounded set in dimension d is at least

$$2^{-33} d^{-2} \kappa^{-2},$$

where κ is the condition number.

Example: Hit-and-Run

An example of a HSS: on each slice, pick a random direction, then sample a uniform point on this line intersected with the slice.

From [Lovàsz and Vempala (2004)], we deduce that for L-smooth and m-strongly concave potentials the spectral gap of hit-and-run on a bounded set in dimension d is at least

$$2^{-33} d^{-2} \kappa^{-2},$$

where κ is the condition number.

Thus we arrive at the comparison

$$\mathcal{E}(U, f) \leq (2^{33} \kappa^2 d^2)\, \mathcal{E}(H, f).$$

I.e. spectral gap of H is at least $2^{-33} \kappa^{-2} d^{-2}$ times the spectral gap of U.

Overview

1. Introduction: MCMC
2. Convergence of Markov chains and comparisons
3. Comparisons of Slice Sampling
4. Conclusion

Concluding remarks

We have derived quantitative comparison theorems relating Simple Slice Sampling with Hybrid Slice Sampling, which covers the subgeometric setting.

Concluding remarks

We have derived quantitative comparison theorems relating Simple Slice Sampling with Hybrid Slice Sampling, which covers the subgeometric setting.

We made use of the framework of Weak Poincaré Inequalities, introduced in [Andrieu, Lee, Power, W. (2022)].

Concluding remarks

We have derived quantitative comparison theorems relating Simple Slice Sampling with Hybrid Slice Sampling, which covers the subgeometric setting.

We made use of the framework of Weak Poincaré Inequalities, introduced in [Andrieu, Lee, Power, W. (2022)].

This significantly extends prior work e.g. [Łatuszyński & Rudolf (2014)] which was mostly qualitative.

Concluding remarks

We have derived quantitative comparison theorems relating Simple Slice Sampling with Hybrid Slice Sampling, which covers the subgeometric setting.

We made use of the framework of Weak Poincaré Inequalities, introduced in [Andrieu, Lee, Power, W. (2022)].

This significantly extends prior work e.g. [Łatuszyński & Rudolf (2014)] which was mostly qualitative.

We have further applied this to look at IMH on the slice and Hit-and-run.

Concluding remarks

We have derived quantitative comparison theorems relating Simple Slice Sampling with Hybrid Slice Sampling, which covers the subgeometric setting.

We made use of the framework of Weak Poincaré Inequalities, introduced in [Andrieu, Lee, Power, W. (2022)].

This significantly extends prior work e.g. [Łatuszyński & Rudolf (2014)] which was mostly qualitative.

We have further applied this to look at IMH on the slice and Hit-and-run.

Preprint forthcoming!

Thanks for listening! I

- Andrieu, C., Lee, A., Power, S., Wang, A. Q. (2022). Comparison of Markov chains via weak Poincaré inequalities with application to pseudo-marginal MCMC. *Ann. Statist.*, 50(6), 3592-3618.
- 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`.
- Łatuszyński, K., Rudolf, D. (2014). Convergence of hybrid slice sampling via spectral gap. `https://arxiv.org/abs/1409.2709`.
- 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), 496-523.
- Lovász, L., Vempala, S. (2004). Hit-and-run from a corner. *Conf. Proc. Annual ACM Symp. Th. Comput.*, 310314.
- Mengersen, K.L., Tweedie, R.L. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101 121, 1996.
- 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), 1087-1092.
- Natarovskii, V., Rudolf, D., Sprungk, B. (2021). Quantitative spectral gap estimate and Wasserstein contraction of simple slice sampling. *Ann. Appl. Probab.*, 31(2), 806-825.
- Neal, R. M. (2003). Slice sampling. *Ann. Statist.*, 31(3), 705-767.
- Qin, Q., Ju, N., Wang, G. (2023). Spectral gap bounds for reversible hybrid Gibbs chains. `https://arxiv.org/abs/2312.12782`.

Thanks for listening! II

Röckner, M., Wang, F.-Y. (2001). Weak Poincare Inequalities and L^2 Convergence Rates of Markov Semigroups. *J. Funct. Anal.*, 185, 564-603.

Rudolf, D., Ullrich, M. (2018). Comparison of hit-and-run, slice sampler and random walk Metropolis. Journal of Applied Probability, 55(4), 11861202.

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 ([?, ?])

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

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

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 $f \in \mathrm{L}^2_0(\pi)$:

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

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 and spectral profiles

Lemma ([?])

For nonconstant nonnegative $g \in \mathrm{L}_0^2(\pi)$, 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

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

Proof of convergence bound (I)

Fix $f \in \mathrm{L}_0^2(\mu)$. Have that

$$\|f\|_2^2 \leq s\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0$$

Proof of convergence bound (I)

Fix $f \in \mathrm{L}^2_0(\mu)$. Have that

$$\|f\|_2^2 \leq s\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0$$

$$\Rightarrow \frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq \frac{\|f\|_2^2}{s\Phi(f)} - \frac{\beta(s)}{s}.$$

Proof of convergence bound (I)

Fix $f \in \mathrm{L}_0^2(\mu)$. Have that

$$\|f\|_2^2 \le s\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0$$

$$\Rightarrow \frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \ge \frac{\|f\|_2^2}{s\Phi(f)} - \frac{\beta(s)}{s}.$$

Set $u := 1/s$, $K(u) := u\beta(1/u)$.

Proof of convergence bound (I)

Fix $f \in \mathrm{L}^2_0(\mu)$. Have that

$$\|f\|_2^2 \leq s\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0$$

$$\Rightarrow \frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq \frac{\|f\|_2^2}{s\Phi(f)} - \frac{\beta(s)}{s}.$$

Set $u := 1/s$, $K(u) := u\beta(1/u)$.

$$\frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq u \cdot \frac{\|f\|_2^2}{\Phi(f)} - K(u), \quad \forall u > 0.$$

Proof of convergence bound (I)

Fix $f \in \mathrm{L}^2_0(\mu)$. Have that

$$\|f\|_2^2 \leq s\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0$$

$$\Rightarrow \frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq \frac{\|f\|_2^2}{s\Phi(f)} - \frac{\beta(s)}{s}.$$

Set $u := 1/s$, $K(u) := u\beta(1/u)$.

$$\frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq u \cdot \frac{\|f\|_2^2}{\Phi(f)} - K(u), \quad \forall u > 0.$$

$$\frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq \sup_{u>0} \left\{ u \cdot \frac{\|f\|_2^2}{\Phi(f)} - K(u) \right\}$$

Proof of convergence bound (I)

Fix $f \in \mathrm{L}^2_0(\mu)$. Have that

$$\|f\|_2^2 \le s\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0$$

$$\Rightarrow \frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \ge \frac{\|f\|_2^2}{s\Phi(f)} - \frac{\beta(s)}{s}.$$

Set $u := 1/s$, $K(u) := u\beta(1/u)$.

$$\frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \ge u \cdot \frac{\|f\|_2^2}{\Phi(f)} - K(u), \quad \forall u > 0.$$

$$\frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \ge \sup_{u>0} \left\{ u \cdot \frac{\|f\|_2^2}{\Phi(f)} - K(u) \right\} =: K^* \left(\frac{\|f\|_2^2}{\Phi(f)} \right).$$

Proof of convergence bound (I)

Fix $f \in \mathrm{L}^2_0(\mu)$. Have that

$$\|f\|_2^2 \leq s\mathcal{E}(P^*P, f) + \beta(s)\Phi(f), \quad \forall s > 0$$

$$\Rightarrow \frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq \frac{\|f\|_2^2}{s\Phi(f)} - \frac{\beta(s)}{s}.$$

Set $u := 1/s$, $K(u) := u\beta(1/u)$.

$$\frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq u \cdot \frac{\|f\|_2^2}{\Phi(f)} - K(u), \quad \forall u > 0.$$

$$\frac{\mathcal{E}(P^*P, f)}{\Phi(f)} \geq \sup_{u>0} \left\{ u \cdot \frac{\|f\|_2^2}{\Phi(f)} - K(u) \right\} =: K^* \left(\frac{\|f\|_2^2}{\Phi(f)} \right).$$

Call this final inequality optimized WPI (oWPI).

Proof of convergence bound (II)

Now define

$$F(x) := \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \quad x \in (0, a], \qquad h_n := \frac{\|P^n f\|_2^2}{\Phi(f)}.$$

Proof of convergence bound (II)

Now define

$$F(x) := \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \quad x \in (0, a], \qquad h_n := \frac{\|P^n f\|_2^2}{\Phi(f)}.$$

Want to bound convergence of $h_n \to 0$.

Proof of convergence bound (II)

Now define

$$F(x) := \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \quad x \in (0, a], \qquad h_n := \frac{\|P^n f\|_2^2}{\Phi(f)}.$$

Want to bound convergence of $h_n \to 0$.

$$F(h_n) - F(h_{n-1}) = \int_{h_n}^{h_{n-1}} \frac{\mathrm{d}v}{K^*(v)}$$

Proof of convergence bound (II)

Now define

$$F(x) := \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \quad x \in (0, a], \qquad h_n := \frac{\|P^n f\|_2^2}{\Phi(f)}.$$

Want to bound convergence of $h_n \to 0$.

$$\begin{aligned} F(h_n) - F(h_{n-1}) &= \int_{h_n}^{h_{n-1}} \frac{\mathrm{d}v}{K^*(v)} \\ &\geq (h_{n-1} - h_n)/K^*(h_{n-1}) \end{aligned}$$

Proof of convergence bound (II)

Now define

$$F(x) := \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \quad x \in (0, a], \qquad h_n := \frac{\|P^n f\|_2^2}{\Phi(f)}.$$

Want to bound convergence of $h_n \to 0$.

$$\begin{aligned} F(h_n) - F(h_{n-1}) &= \int_{h_n}^{h_{n-1}} \frac{\mathrm{d}v}{K^*(v)} \\ &\geq (h_{n-1} - h_n)/K^*(h_{n-1}) \\ &= \frac{\mathcal{E}(P^*P, P^{n-1}f)/\Phi(f)}{K^*(h_{n-1})} \end{aligned}$$

Proof of convergence bound (II)

Now define

$$F(x) := \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \quad x \in (0, a], \qquad h_n := \frac{\|P^n f\|_2^2}{\Phi(f)}.$$

Want to bound convergence of $h_n \to 0$.

$$\begin{aligned} F(h_n) - F(h_{n-1}) &= \int_{h_n}^{h_{n-1}} \frac{\mathrm{d}v}{K^*(v)} \\ &\geq (h_{n-1} - h_n)/K^*(h_{n-1}) \\ &= \frac{\mathcal{E}(P^*P, P^{n-1}f)/\Phi(f)}{K^*(h_{n-1})} \\ &\geq K^*(h_{n-1})/K^*(h_{n-1}) = 1. \qquad \text{(oWPI)} \end{aligned}$$

Proof of convergence bound (II)

Now define

$$F(x) := \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \quad x \in (0, a], \qquad h_n := \frac{\|P^n f\|_2^2}{\Phi(f)}.$$

Want to bound convergence of $h_n \to 0$.

$$\begin{aligned} F(h_n) - F(h_{n-1}) &= \int_{h_n}^{h_{n-1}} \frac{\mathrm{d}v}{K^*(v)} \\ &\geq (h_{n-1} - h_n)/K^*(h_{n-1}) \\ &= \frac{\mathcal{E}(P^*P, P^{n-1}f)/\Phi(f)}{K^*(h_{n-1})} \\ &\geq K^*(h_{n-1})/K^*(h_{n-1}) = 1. \qquad \text{(oWPI)} \\ &\Rightarrow F(h_n) - F(h_0) \geq n. \end{aligned}$$

Proof of convergence bound (II)

Now define

$$F(x) := \int_x^1 \frac{\mathrm{d}v}{K^*(v)}, \quad x \in (0, a], \qquad h_n := \frac{\|P^n f\|_2^2}{\Phi(f)}.$$

Want to bound convergence of $h_n \to 0$.

$$\begin{aligned}
F(h_n) - F(h_{n-1}) &= \int_{h_n}^{h_{n-1}} \frac{\mathrm{d}v}{K^*(v)} \\
&\geq (h_{n-1} - h_n)/K^*(h_{n-1}) \\
&= \frac{\mathcal{E}(P^*P, P^{n-1}f)/\Phi(f)}{K^*(h_{n-1})} \\
&\geq K^*(h_{n-1})/K^*(h_{n-1}) = 1. \qquad \text{(oWPI)} \\
&\Rightarrow F(h_n) - F(h_0) \geq n.
\end{aligned}$$

So we invert this to obtain

$$\|P^n f\|_2^2 \leq \Phi(f) F^{-1}(n). \quad \square$$