Weak Peskun ordering for approximate MCMC comparison

Florian Maire (Université de Montréal)
florian.maire@umontreal.ca

joint work with:
P. Gagnon (univ. de Montréal)
P. Vandekerkhove, (univ. Paris-Est)

TCD Statistics Seminar Series – April 7th, 2021

Context: MCMC, CLT and Peskun ordering

Markov chains

A Markov chain $\{X_t \,:\, t \in \mathbb{N}\}$ on some state space $(\mathsf{E}, \mathcal{E})$ is characterized by:

- an **initial distribution** π_0 on $(\mathsf{E}, \mathcal{E})$,
- a **transition kernel** $P : (\mathsf{E}, \mathcal{E}) \to [0, 1]$, that is a conditional probability distribution

$$P(x, A) = \mathrm{Pr}(X_t \in A \,|\, X_{t-1} = x)\,, \qquad \forall\, x \in \mathsf{E},\ A \in \mathcal{E}\,.$$

Markov chains

A Markov chain $\{X_t \,:\, t \in \mathbb{N}\}$ on some state space $(\mathsf{E}, \mathcal{E})$ is characterized by:

- an **initial distribution** π_0 on $(\mathsf{E}, \mathcal{E})$,
- a **transition kernel** $P : (\mathsf{E}, \mathcal{E}) \to [0, 1]$, that is a conditional probability distribution

$$P(x, A) = \Pr(X_t \in A \,|\, X_{t-1} = x)\,, \qquad \forall\, x \in \mathsf{E},\, A \in \mathcal{E}\,.$$

Certain Markov chains have the property of having a **limiting distribution**, call it π,

$$\lim_{t\to\infty} \Pr(X_t \in A) \to \pi(A)\,, \quad \forall A \in \mathcal{X}\,,$$

which does not depend on π_0.

Markov chains

A Markov chain $\{X_t \,:\, t \in \mathbb{N}\}$ on some state space $(\mathsf{E}, \mathcal{E})$ is characterized by:

- an **initial distribution** π_0 on $(\mathsf{E}, \mathcal{E})$,
- a **transition kernel** $P : (\mathsf{E}, \mathcal{E}) \to [0, 1]$, that is a conditional probability distribution

$$P(x, A) = \mathrm{Pr}(X_t \in A \,|\, X_{t-1} = x)\,, \qquad \forall\, x \in \mathsf{E},\; A \in \mathcal{E}\,.$$

Certain Markov chains have the property of having a **limiting distribution**, call it π,

$$\lim_{t \to \infty} \mathsf{Pr}(X_t \in A) \to \pi(A)\,, \quad \forall A \in \mathcal{X}\,,$$

which does not depend on π_0.

In this talk:

- we prescribe a probability distribution of reference π on $(\mathsf{E}, \mathcal{E})$, usually called **target distribution**
- we want to sample from π using the Markov P
- we assume that π is the limiting distribution of P
- we assume $\pi_0 = \pi$ (start in stationary regime)

Markov chain Monte Carlo (MCMC)

Consider a **test function** $f : \mathsf{E} \to \mathbb{R}$ (measurable). The typical goal of a MCMC procedure is to estimate quantities of the form

$$\pi f := \int f(x)\pi(\mathrm{d}x)$$

by mean of the **empirical average along the path** of a Markov chain $\{X_t \ : \ t \in \mathbb{N}\}$

$$\widehat{\pi f}_T := (1/T)\sum_{t=1}^{T} f(X_t)\,, \qquad X_0 \sim \pi\,, \quad X_{t+1} \sim P(X_t, \cdot\,)\,,\ t > 0\,.$$

Markov chain Monte Carlo (MCMC)

Consider a **test function** $f : \mathsf{E} \to \mathbb{R}$ (measurable). The typical goal of a MCMC procedure is to estimate quantities of the form

$$\pi f := \int f(x)\pi(\mathrm{d}x)$$

by mean of the **empirical average along the path** of a Markov chain $\{X_t \ : \ t \in \mathbb{N}\}$

$$\widehat{\pi f}_T := (1/T)\sum_{t=1}^{T} f(X_t)\,, \qquad X_0 \sim \pi\,, \quad X_{t+1} \sim P(X_t, \cdot\,)\,,\ t > 0\,.$$

In Statistics and ML applications, we ask that $\{\widehat{\pi f}_T\}$ satisfies a **CLT**

$$\sqrt{T}\left(\widehat{\pi f}_T - \pi f\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{var}(P, f))\,, \quad \text{as } T \to \infty$$

Markov chain Monte Carlo (MCMC)

Consider a **test function** $f : \mathsf{E} \to \mathbb{R}$ (measurable). The typical goal of a MCMC procedure is to estimate quantities of the form

$$\pi f := \int f(x)\pi(\mathrm{d}x)$$

by mean of the **empirical average along the path** of a Markov chain $\{X_t \; : \; t \in \mathbb{N}\}$

$$\widehat{\pi f}_T := (1/T)\sum_{t=1}^{T} f(X_t)\,, \qquad X_0 \sim \pi\,, \quad X_{t+1} \sim P(X_t, \cdot\,)\,,\; t > 0\,.$$

In Statistics and ML applications, we ask that $\{\widehat{\pi f}_T\}$ satisfies a **CLT**

$$\sqrt{T}\left(\widehat{\pi f}_T - \pi f\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{var}(P, f))\,, \quad \text{as } T \to \infty$$

where the **asymptotic variance of the Markov chain**

$$\mathrm{var}(P, f) := \lim_{T\to\infty} \frac{1}{T}\mathbb{V}ar\left\{\sum_{t=1}^{T} f(X_t)\right\}$$

must exist and be finite.

$\Rightarrow$ the lower $\mathrm{var}(P, f)$ the more precise the estimation of πf .

Three examples of Markov chain CLT's

Since we need to have $\mathrm{var}(P, f) < \infty$ and that

$$\mathrm{var}(P, f) = \mathbb{V}ar_{\pi} f(X_0) + 2 \sum_{t=1}^{\infty} \mathbb{C}ov(f(X_0), f(X_t))$$

we need to focus on $f \in \mathsf{L}_2(\pi) = \{f : \mathsf{E} \to \mathbb{R} \ : \ \mathbb{V}ar_{\pi} f(X) < \infty\}$.

Three examples of Markov chain CLT's

Since we need to have $\mathrm{var}(P, f) < \infty$ and that

$$\mathrm{var}(P, f) = \mathbb{V}ar_{\pi} f(X_0) + 2 \sum_{t=1}^{\infty} \mathbb{C}ov(f(X_0), f(X_t))$$

we need to focus on $f \in \mathsf{L}_2(\pi) = \{f : \mathsf{E} \to \mathbb{R} \ : \ \mathbb{V}ar_{\pi} f(X) < \infty\}$.

- (Cogburn, 1976) If P is **uniformly ergodic**, then for each $f \in \mathsf{L}_2(\pi)$

  CLT holds.

Three examples of Markov chain CLT's

Since we need to have $\mathrm{var}(P, f) < \infty$ and that

$$\mathrm{var}(P, f) = \mathbb{V}ar_{\pi} f(X_0) + 2\sum_{t=1}^{\infty} \mathbb{C}ov(f(X_0), f(X_t))$$

we need to focus on $f \in \mathsf{L}_2(\pi) = \{f : \mathsf{E} \to \mathbb{R} \ : \ \mathbb{V}ar_{\pi} f(X) < \infty\}$.

- (Cogburn, 1976) If P is **uniformly ergodic**, then for each $f \in \mathsf{L}_2(\pi)$

  $$\text{CLT holds}\,.$$

- (Kipnis and Varadhan, 1986) If P is π-**reversible** and $\mathrm{var}(P, f) < \infty$ for all $f \in]Ltwo(\pi)$, then

  $$\text{CLT holds for all } \mathsf{L}_2(\pi)\,.$$

Three examples of Markov chain CLT's

Since we need to have $\mathrm{var}(P, f) < \infty$ and that

$$\mathrm{var}(P, f) = \mathbb{V}ar_\pi f(X_0) + 2\sum_{t=1}^{\infty} \mathbb{C}ov(f(X_0), f(X_t))$$

we need to focus on $f \in \mathsf{L}_2(\pi) = \{f : \mathsf{E} \to \mathbb{R} \;:\; \mathbb{V}ar_\pi f(X) < \infty\}$.

- (Cogburn, 1976) If P is **uniformly ergodic**, then for each $f \in \mathsf{L}_2(\pi)$

  $$\text{CLT holds}.$$

- (Kipnis and Varadhan, 1986) If P is π-**reversible** and $\mathrm{var}(P, f) < \infty$ for all $f \in]Ltwo(\pi)$, then

  $$\text{CLT holds for all } \mathsf{L}_2(\pi).$$

- (Roberts and Rosenthal, 2004) If P is **geometrically ergodic** with drift function $V : \mathsf{E} \to [1, \infty)$, then

  $$\text{CLT holds for the subset } \{f \in \mathsf{L}_2(\pi) \;:\; \sup f^2/V < \infty\}.$$

Three examples of Markov chain CLT's

Since we need to have $\mathrm{var}(P, f) < \infty$ and that

$$\mathrm{var}(P, f) = \mathbb{V}ar_\pi f(X_0) + 2\sum_{t=1}^{\infty} \mathbb{C}ov(f(X_0), f(X_t))$$

we need to focus on $f \in \mathsf{L}_2(\pi) = \{f : \mathsf{E} \to \mathbb{R} \ : \ \mathbb{V}ar_\pi f(X) < \infty\}$.

- (Cogburn, 1976) If P is **uniformly ergodic**, then for each $f \in \mathsf{L}_2(\pi)$

  $$\text{CLT holds}\,.$$

- (Kipnis and Varadhan, 1986) If P is π-**reversible** and $\mathrm{var}(P, f) < \infty$ for all $f \in]Ltwo(\pi)$, then

  $$\text{CLT holds for all } \mathsf{L}_2(\pi)\,.$$

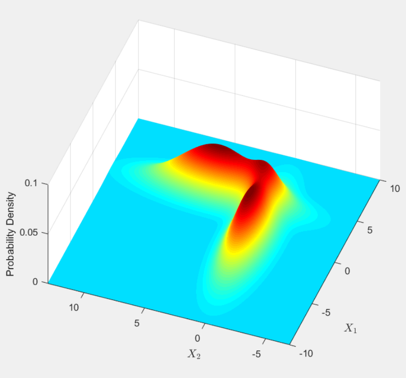
- (Roberts and Rosenthal, 2004) If P is **geometrically ergodic** with drift function $V : \mathsf{E} \to [1, \infty)$, then

  $$\text{CLT holds for the subset } \{f \in \mathsf{L}_2(\pi) \ : \ \sup f^2/V < \infty\}\,.$$

See *On the Markov Chain Central Limit Theorem* (2004) by G. Jones for a comprehensive review.

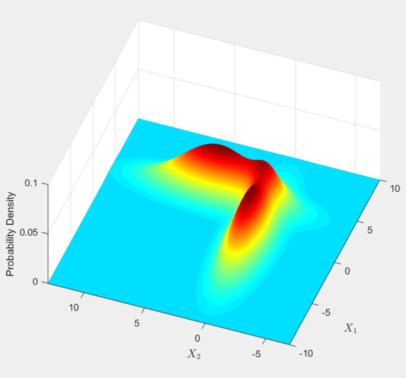
Example 1 and illustration of CLT

Consider that π is the mixture of two gaussians with pdf as follow



Example 1 and illustration of CLT

Consider that π is the mixture of two gaussians with pdf as follow



We are interested in estimating $\mu = \Pr(X_1 < 0)$ so we take $f(x_1, x_2) = \mathbb{1}_{\{x_1<0\}}$ since

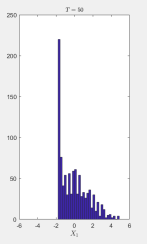
$$\mathbb{E}f(X_1, X_2) = \iint f(x_1, x_2)\pi(\mathrm{d}x_1, \mathrm{d}x_2) = \int_{-\infty}^{0} \pi(\mathrm{d}x_1) = \mu\,,$$

thus with a Markov chain $\{X_{1,t}, X_{2,t}\}_{t=1}^{T}$, $\frac{1}{T}\sum_{t=1}^{T} f(X_{1,t}, X_{2,t}) \to \mu\,.$

Example 1 and illustration of CLT

We chose the Gibbs sampler to define P and got the following distribution of

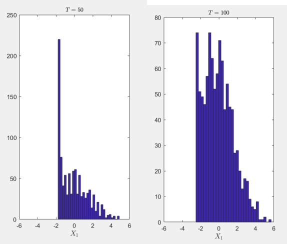
$$\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T} f(X_{1,t}, X_{2,t}) - \mu\right].$$



Example 1 and illustration of CLT

We chose the Gibbs sampler to define P and got the following distribution of

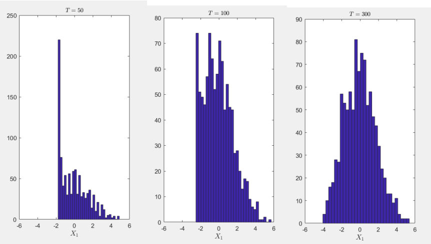
$$\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T} f(X_{1,t}, X_{2,t}) - \mu\right].$$



Example 1 and illustration of CLT

We chose the Gibbs sampler to define P and got the following distribution of

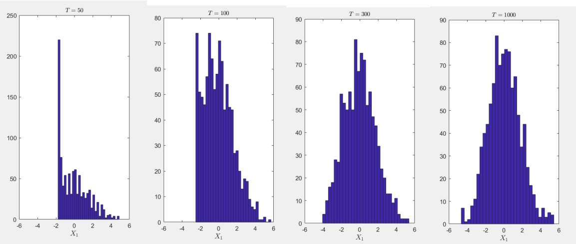
$$\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T} f(X_{1,t}, X_{2,t}) - \mu\right].$$



Example 1 and illustration of CLT

We chose the Gibbs sampler to define P and got the following distribution of

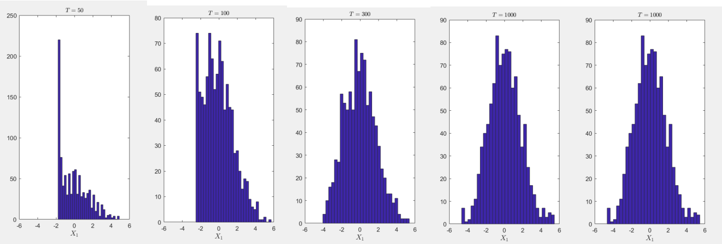
$$\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T} f(X_{1,t}, X_{2,t}) - \mu\right].$$



Example 1 and illustration of CLT

We chose the Gibbs sampler to define P and got the following distribution of

$$\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T} f(X_{1,t}, X_{2,t}) - \mu\right].$$



Example 1 and illustration of CLT

We chose the Gibbs sampler to define P and got the following distribution of

$$\sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T} f(X_{1,t}, X_{2,t}) - \mu\right] .$$

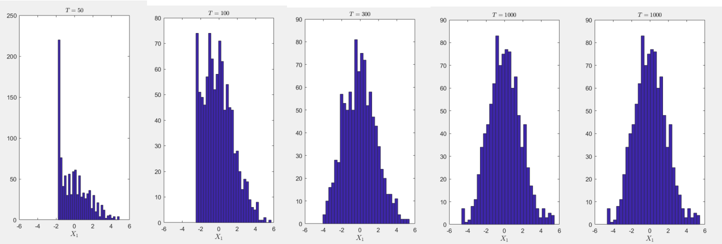


Figure: $T \in \{50, 100, 300, 1000, 10000\}$

$$\Rightarrow \mathrm{var}(P, f) = \underbrace{\mathbb{V}ar\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f(X_{1,t}, X_{2,t})\right]}_{\text{asymptotic variance}} \quad \text{matters for CI's tightness}\,.$$

Comparison of MCMC algorithms

It can be achieved on different grounds, but in this talk we say for two MCMC samplers P_0 and P_1 that

$$P_1 \text{ is better than } P_0 \text{ to estimate } \pi f$$

if $\mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)$.

Comparison of MCMC algorithms

It can be achieved on different grounds, but in this talk we say for two MCMC samplers P_0 and P_1 that

$$P_1 \text{ is better than } P_0 \text{ to estimate } \pi f$$

if $\mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)$.

In the previous example, consider that P_1 is the Gibbs sampler and P_0 is a "modified" Gibbs sampler.

f	$\mathrm{var}(P_1, f)$	$\mathrm{var}(P_0, f)$
$f(x) = \mathbb{1}_{x_1<0}$	3.3	**2.1**
$f(x) = \|x/10\|^2$	1.44	**1.22**
$f(x) = (x_2/10)^3$	0.85	**0.15**
$f(x) = \cos(x)$	**5.1**	5.5

In this example P_0 seems often better than P_1, but not always (see last function).

Comparison of MCMC algorithms

It can be achieved on different grounds, but in this talk we say for two MCMC samplers P_0 and P_1 that

$$P_1 \text{ is better than } P_0 \text{ to estimate } \pi f$$

if $\mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)$.

In the previous example, consider that P_1 is the Gibbs sampler and P_0 is a "modified" Gibbs sampler.

f	$\mathrm{var}(P_1, f)$	$\mathrm{var}(P_0, f)$
$f(x) = \mathbb{1}_{x_1<0}$	3.3	**2.1**
$f(x) = \|x/10\|^2$	1.44	**1.22**
$f(x) = (x_2/10)^3$	0.85	**0.15**
$f(x) = \cos(x)$	**5.1**	5.5

In this example P_0 seems often better than P_1, but not always (see last function).

$\Rightarrow$Is it possible to decide which algorithm to choose to estimate μ without doing any preliminary simulation?

Peskun ordering on reversible Markov chains

More rigorously, we say that P_1 is more efficient than P_0 in terms of **asymptotic statistical efficiency** if

$$\mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,,$$

holds for a rich enough class of test functions $\mathcal{F} \subset \mathsf{L}_2(\pi)$.

Peskun ordering on reversible Markov chains

More rigorously, we say that P_1 is more efficient than P_0 in terms of **asymptotic statistical efficiency** if

$$\mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,,$$

holds for a rich enough class of test functions $\mathcal{F} \subset \mathsf{L}_2(\pi)$.

Peskun-Tierney partial ordering. We note $P_1 \succeq P_0$ if

$$P_1(x, A\backslash\{x\}) \geq P_0(x, A\backslash\{x\})\,, \qquad \text{for all } (x, A) \in \mathsf{E} \times \mathcal{E}$$

Peskun ordering on reversible Markov chains

More rigorously, we say that P_1 is more efficient than P_0 in terms of **asymptotic statistical efficiency** if

$$\mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,,$$

holds for a rich enough class of test functions $\mathcal{F} \subset \mathsf{L}_2(\pi)$.

Peskun-Tierney partial ordering. We note $P_1 \succeq P_0$ if

$$P_1(x, A\backslash\{x\}) \geq P_0(x, A\backslash\{x\})\,, \qquad \text{for all } (x, A) \in \mathsf{E} \times \mathcal{E}$$

Let (P_0, P_1) be π-reversible. Peskun (1973) and Tierney (1998) showed that

$$P_1 \succeq P_0 \Longrightarrow \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \quad \textbf{for each } f \in \mathsf{L}_2(\pi)\,.$$

$\Rightarrow$ When comparing two MCMC (π-rev.) samplers P_0 and P_1, proving that P_1 is better than P_0, one can "try" to show that $P_1 \succeq P_0$.

$\Rightarrow$ In the previous example, P_0 **does not** dominate P_1 according to Peskun ordering, even though it looks better for most test functions we tried.

Relaxed and strong versions of Peskun ordering

Let $\omega > 0$ and assume

$$P_1(x, A\backslash\{x\}) \geq \omega P_0(x, A\backslash\{x\})\,, \qquad \text{for all } (x, A) \in \mathsf{E} \times \mathcal{E}$$

then Andrieu, Lee and Vihola (2018) showed that

$$\mathrm{var}(P_1, f) \leq \frac{1}{\omega}\mathrm{var}(P_0, f) + \left(\frac{1}{\omega} - 1\right) \mathbb{V}ar_\pi f(X)\,,$$

for each $f \in \mathsf{L}_2(\pi)$.

Relaxed and strong versions of Peskun ordering

Let $\omega > 0$ and assume

$$P_1(x, A\backslash\{x\}) \geq \omega P_0(x, A\backslash\{x\})\,, \qquad \text{for all } (x, A) \in \mathsf{E} \times \mathcal{E}$$

then Andrieu, Lee and Vihola (2018) showed that

$$\mathrm{var}(P_1, f) \leq \frac{1}{\omega}\mathrm{var}(P_0, f) + \left(\frac{1}{\omega} - 1\right) \mathbb{V}ar_\pi f(X)\,,$$

for each $f \in \mathsf{L}_2(\pi)$.

- Peskun-Tierney corresponds to $\omega = 1$

Relaxed and strong versions of Peskun ordering

Let $\omega > 0$ and assume

$$P_1(x, A\backslash\{x\}) \geq \omega P_0(x, A\backslash\{x\}), \qquad \text{for all } (x, A) \in \mathsf{E} \times \mathcal{E}$$

then Andrieu, Lee and Vihola (2018) showed that

$$\mathrm{var}(P_1, f) \leq \frac{1}{\omega}\mathrm{var}(P_0, f) + \left(\frac{1}{\omega} - 1\right) \mathbb{V}ar_\pi f(X),$$

for each $f \in \mathsf{L}_2(\pi)$.

- Peskun-Tierney corresponds to $\omega = 1$
- Strong/quantitative version of Peskun-Tierney $\omega > 1$, e.g. $\omega = 2$ and we have
$$\mathrm{var}(P_1, f) \leq \frac{1}{2}\mathrm{var}(P_0, f) - \mathbb{V}ar_\pi f(X)/2 \leq \frac{1}{2}\mathrm{var}(P_0, f)$$

Relaxed and strong versions of Peskun ordering

Let $\omega > 0$ and assume

$$P_1(x, A\backslash\{x\}) \geq \omega P_0(x, A\backslash\{x\}), \qquad \text{for all } (x, A) \in \mathsf{E} \times \mathcal{E}$$

then Andrieu, Lee and Vihola (2018) showed that

$$\mathrm{var}(P_1, f) \leq \frac{1}{\omega}\mathrm{var}(P_0, f) + \left(\frac{1}{\omega} - 1\right) \mathbb{V}ar_\pi f(X),$$

for each $f \in \mathsf{L}_2(\pi)$.

- Peskun-Tierney corresponds to $\omega = 1$
- Strong/quantitative version of Peskun-Tierney $\omega > 1$, e.g. $\omega = 2$ and we have
$$\mathrm{var}(P_1, f) \leq \frac{1}{2}\mathrm{var}(P_0, f) - \mathbb{V}ar_\pi f(X)/2 \leq \frac{1}{2}\mathrm{var}(P_0, f)$$
- Relaxed version $\omega < 1$, e.g $\omega = 1/2$ and we have
$$\mathrm{var}(P_1, f) \leq 2\mathrm{var}(P_0, f) + \mathbb{V}ar_\pi f(X)$$

Recent papers discussing Peskun-Tierney ordering

Methodology: constructing MCMC samplers that improve existing ones by showing a Peskun-Tierney ordering

- Informed proposals for local MCMC in discrete spaces, Zanella (2020)
- A Metropolis-class sampler for targets with non-convex support, Moriarty et al. (2020)
- Markov chain Monte Carlo algorithms with sequential proposals, Park and Atchadé (2020)
- Nonreversible Jump Algorithms for Bayesian Nested Model Selection, Gagnon and Doucet (2020)

Theoretical side:

- Peskun-Tierney ordering for Markov chain and process Monte Carlo: beyond the reversible scenario, Andrieu and Livingstone (2020)

Main result

Definition and notation

(1) Function space

- Let the subset of centered functions be $\mathsf{L}_{2,0}(\pi) := \{f \in \mathsf{L}_2(\pi) \;:\; \mathbb{E}_\pi f(X) = 0\}$
- Define the weighted inner product on $\mathsf{L}_2(\pi)$ as

$$\langle f, g \rangle = \int_{\mathsf{E}} \pi(\mathrm{d}x) f(x) g(x) \,, \qquad (f, g) \in \mathsf{L}_2(\pi)^2$$

- The (weighted) L_2-norm of $f \in \mathsf{L}_{2,0}(\pi)$ induced by this inner product satisfies

$$\|f\|_2^2 = \int f(x)^2 \pi(\mathrm{d}x) = \mathbb{V}ar_\pi f(X) \,.$$

Definition and notation

(1) Function space

- Let the subset of centered functions be
  $\mathrm{L}_{2,0}(\pi) := \{f \in \mathrm{L}_2(\pi) \; : \; \mathbb{E}_\pi f(X) = 0\}$
- Define the weighted inner product on $\mathrm{L}_2(\pi)$ as

$$\langle f, g \rangle = \int_{\mathsf{E}} \pi(\mathrm{d}x) f(x) g(x) \,, \qquad (f, g) \in \mathrm{L}_2(\pi)^2$$

- The (weighted) L_2-norm of $f \in \mathrm{L}_{2,0}(\pi)$ induced by this inner product satisfies

$$\|f\|_2^2 = \int f(x)^2 \pi(\mathrm{d}x) = \mathbb{V}ar_\pi f(X) \,.$$

(2) Markov chain kernel P as an operator on $\mathrm{L}_{2,0}(\pi)$

- the operator is defined by

$$P : f \mapsto \int_E P(\,\cdot\,, \mathrm{d}x) f(x) = \mathbb{E}\left\{ f(X_1) \,|\, X_0 = \cdot \right\}$$

- if $f \in \mathrm{L}_{2,0}(\pi)$, the inner product between f and Pf is

$$\langle f, Pf \rangle = \mathbb{C}ov(f(X_0), f(X_1)) \,.$$

- P is a positive operator on $\mathrm{L}_2(\pi)$ is for each $f \in \mathrm{L}_2(\pi)$,

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

Why is a Peskun-Tierney ordering hard to verify?

Like the ordering induced by positivity for symmetric matrices, Peskun-Tierney ordering is a partial ordering:

$$\text{there exist } (P_0, P_1) \qquad : \qquad P_0 \not\succeq P_1 \quad \text{and} \quad P_1 \not\succeq P_0 \,.$$

Why is a Peskun-Tierney ordering hard to verify?

Like the ordering induced by positivity for symmetric matrices, Peskun-Tierney ordering is a partial ordering:

$$\text{there exist } (P_0, P_1) \qquad : \qquad P_0 \not\succeq P_1 \quad \text{and} \quad P_1 \not\succeq P_0 \,.$$

It has strong necessary conditions:

- (as we saw)

$$\mathrm{var}(f, P_1) \leq \mathrm{var}(f, P_0)\,, \qquad \textbf{for all } \; f \in \mathsf{L}_{2,0}(\pi)$$

Why is a Peskun-Tierney ordering hard to verify?

Like the ordering induced by positivity for symmetric matrices, Peskun-Tierney ordering is a partial ordering:

$$\text{there exist } (P_0, P_1) \qquad : \qquad P_0 \not\succeq P_1 \quad \text{and} \quad P_1 \not\succeq P_0 \,.$$

It has strong necessary conditions:

- (as we saw)

$$\mathrm{var}(f, P_1) \leq \mathrm{var}(f, P_0)\,, \qquad \textbf{for all } f \in \mathsf{L}_{2,0}(\pi)$$

- positivity of $P_0 - P_1$

$$\langle f, (P_0 - P_1) f \rangle \geq 0\,, \qquad \textbf{for all } f \in \mathsf{L}_{2,0}(\pi)$$

Why is a Peskun-Tierney ordering hard to verify?

Like the ordering induced by positivity for symmetric matrices, Peskun-Tierney ordering is a partial ordering:

$$\text{there exist } (P_0, P_1) \qquad : \qquad P_0 \not\succeq P_1 \quad \text{and} \quad P_1 \not\succeq P_0 \,.$$

It has strong necessary conditions:

- (as we saw)

$$\mathrm{var}(f, P_1) \leq \mathrm{var}(f, P_0)\,, \qquad \textbf{for all } f \in \mathrm{L}_{2,0}(\pi)$$

- positivity of $P_0 - P_1$

$$\langle f, (P_0 - P_1) f \rangle \geq 0\,, \qquad \textbf{for all } f \in \mathrm{L}_{2,0}(\pi)$$

- as is well known

$$\mathrm{gap}_R(P_1) \geq \mathrm{gap}_R(P_0)\,,$$

  where $\mathrm{gap}_R(P_i)$ is the right spectral gap of P_i

$$\mathrm{gap}_R(P_i) = \inf_{\mathbf{f} \in \mathrm{L}_{2,0}(\pi)} \frac{\langle f, (\mathrm{Id} - P_i) f \rangle}{\|f\|^2}\,.$$

Weak version of Peskun-Tierney result?

Recall that if P_0 and P_1 are two π-reversible Markov kernels

$$P_1 \succeq P_0 \quad \Longrightarrow \quad \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \text{ for each } f \in \mathsf{L}_{2,0}(\pi)$$

A natural question would be

- can we find a (partial) ordering weaker than $P_1 \succeq P_0$, say $P_1 \succsim P_0$, easier to verify while nevertheless satisfying something like

$$\begin{array}{c} P_1 \succsim P_0 \\ \\ + \text{ perhaps additional assumptions} \end{array} \quad \Longrightarrow \quad \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \text{ for each } f \in \mathcal{F}$$

for a smaller, yet sufficiently rich, class of functions $\mathcal{F} \subset \mathsf{L}_{2,0}(\pi)$.

Attempt #1: marginal Peskun-Tierney ordering

Setup:

- Consider a product space $(\mathsf{E} \times \mathsf{E}, \mathcal{E} \otimes \mathcal{E})$

$$\pi(\mathrm{d}x) = \pi(\mathrm{d}x_1, \mathrm{d}x_2)$$

- Suppose interest lies intrinsically on the marginal $\pi_1(\mathrm{d}x) := \pi(\mathrm{d}x, \mathsf{E})$
- The test functions of interest are

$$\mathcal{F} := \{f \in \mathsf{L}_{2,0}(\pi) \ : \ \forall\, (x_1, x_2) \in \mathsf{E} \times \mathsf{E},\ f(x_1, x_2) = \bar{f}(x_1)\}$$

Attempt #1: marginal Peskun-Tierney ordering

Setup:

- Consider a product space $(\mathsf{E} \times \mathsf{E}, \mathcal{E} \otimes \mathcal{E})$

$$\pi(\mathrm{d}x) = \pi(\mathrm{d}x_1, \mathrm{d}x_2)$$

- Suppose interest lies intrinsically on the marginal $\pi_1(\mathrm{d}x) := \pi(\mathrm{d}x, \mathsf{E})$
- The test functions of interest are

$$\mathcal{F} := \{f \in \mathsf{L}_{2,0}(\pi) \ : \ \forall\,(x_1, x_2) \in \mathsf{E} \times \mathsf{E},\ f(x_1, x_2) = \bar{f}(x_1)\}$$

- Denote $P_1 \succsim P_0$ if

$$\forall\,(x_1, x_2, A) \in \mathsf{E} \times \mathsf{E} \times \mathcal{E}\,, \quad P_1(x_1, x_2; A\backslash\{x_1\}, \mathsf{E}) \geq P_0(x_1, x_2; A\backslash\{x_1\}, \mathsf{E})$$

Attempt #1: marginal Peskun-Tierney ordering

Setup:

- Consider a product space $(\mathsf{E} \times \mathsf{E}, \mathcal{E} \otimes \mathcal{E})$

$$\pi(\mathrm{d}x) = \pi(\mathrm{d}x_1, \mathrm{d}x_2)$$

- Suppose interest lies intrinsically on the marginal $\pi_1(\mathrm{d}x) := \pi(\mathrm{d}x, \mathsf{E})$
- The test functions of interest are

$$\mathcal{F} := \{f \in \mathsf{L}_{2,0}(\pi) \ : \ \forall\,(x_1, x_2) \in \mathsf{E} \times \mathsf{E},\ f(x_1, x_2) = \bar{f}(x_1)\}$$

- Denote $P_1 \gtrsim P_0$ if

$$\forall\,(x_1, x_2, A) \in \mathsf{E} \times \mathsf{E} \times \mathcal{E}\,, \quad P_1(x_1, x_2; A\backslash\{x_1\}, \mathsf{E}) \geq P_0(x_1, x_2; A\backslash\{x_1\}, \mathsf{E})$$

Do we have

$$P_1 \gtrsim P_0 \quad \Longrightarrow \quad \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \quad \text{for all } f \in \mathcal{F}\,?$$

Attempt #1: marginal Peskun-Tierney ordering

Setup:

- Consider a product space $(\mathsf{E} \times \mathsf{E}, \mathcal{E} \otimes \mathcal{E})$

$$\pi(\mathrm{d}x) = \pi(\mathrm{d}x_1, \mathrm{d}x_2)$$

- Suppose interest lies intrinsically on the marginal $\pi_1(\mathrm{d}x) := \pi(\mathrm{d}x, \mathsf{E})$
- The test functions of interest are

$$\mathcal{F} := \{f \in \mathsf{L}_{2,0}(\pi) \; : \; \forall\, (x_1, x_2) \in \mathsf{E} \times \mathsf{E},\; f(x_1, x_2) = \bar{f}(x_1)\}$$

- Denote $P_1 \succsim P_0$ if

$$\forall\, (x_1, x_2, A) \in \mathsf{E} \times \mathsf{E} \times \mathcal{E}\,, \quad P_1(x_1, x_2; A \backslash \{x_1\}, \mathsf{E}) \geq P_0(x_1, x_2; A \backslash \{x_1\}, \mathsf{E})$$

Do we have

$$P_1 \succsim P_0 \quad \Longrightarrow \quad \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \quad \text{for all } f \in \mathcal{F}\,?$$

Proposition 1

In the previous setup, assume that for $i \in \{0, 1\}$

$$\pi(\mathrm{d}x_1 \,|\, x_2) P_i(x_1, x_2; \mathrm{d}x_1', \mathsf{E}) = \pi(\mathrm{d}x_1' \,|\, x_2) P_i(x_1', x_2; \mathrm{d}x_1, \mathsf{E})\,, \quad \textit{for all } x_2 \in \mathsf{E}$$

then

$$P_1 \succsim P_0 \Rightarrow P_0 - P_1 \textit{ is positive on } \mathcal{F} \not\Rightarrow \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \quad \textit{for all } f \in \mathcal{F}$$

Attempt #2: conditional Peskun-Tierney ordering

Setup:

- Let $\tilde{\mathsf{E}} \subset \mathsf{E}$ with $\pi(\tilde{\mathsf{E}}) > 0$ and let $\tilde{\mathcal{E}}$ be a σ-algebra on $\tilde{\mathsf{E}}$
- Interest lies specifically on the conditional distribution

$$\pi_{\tilde{\mathsf{E}}}(\mathrm{d}x) \propto \pi(\mathrm{d}x)\mathbb{1}_{\{x \in \tilde{\mathsf{E}}\}}$$

- Test functions are

$$\mathcal{F} := \{f \in \mathsf{L}_2(\pi) \; : \; \forall\, x \in \mathsf{E},\, f(x) = f(x)\mathbb{1}_{\{x \in \tilde{\mathsf{E}}\}}\}$$

Attempt #2: conditional Peskun-Tierney ordering

Setup:

- Let $\tilde{\mathsf{E}} \subset \mathsf{E}$ with $\pi(\tilde{\mathsf{E}}) > 0$ and let $\tilde{\mathcal{E}}$ be a σ-algebra on $\tilde{\mathsf{E}}$
- Interest lies specifically on the conditional distribution

$$\pi_{\tilde{\mathsf{E}}}(\mathrm{d}x) \propto \pi(\mathrm{d}x)\mathbb{1}_{\{x \in \tilde{\mathsf{E}}\}}$$

- Test functions are

$$\mathcal{F} := \{f \in \mathsf{L}_2(\pi) \ : \ \forall x \in \mathsf{E},\, f(x) = f(x)\mathbb{1}_{\{x \in \tilde{\mathsf{E}}\}}\}$$

- Denote $P_1 \succsim P_0$ if for all $x \in \tilde{\mathsf{E}}$ and $B \in \tilde{\mathcal{E}}$

$$P_1(x, B \setminus \{x\}) \geq P_0(x, B \setminus \{x\})$$

Attempt #2: conditional Peskun-Tierney ordering

Setup:

- Let $\tilde{\mathsf{E}} \subset \mathsf{E}$ with $\pi(\tilde{\mathsf{E}}) > 0$ and let $\tilde{\mathcal{E}}$ be a σ-algebra on $\tilde{\mathsf{E}}$
- Interest lies specifically on the conditional distribution

$$\pi_{\tilde{\mathsf{E}}}(\mathrm{d}x) \propto \pi(\mathrm{d}x)\mathbb{1}_{\{x \in \tilde{\mathsf{E}}\}}$$

- Test functions are

$$\mathcal{F} := \{f \in \mathsf{L}_2(\pi) \;:\; \forall\, x \in \mathsf{E},\, f(x) = f(x)\mathbb{1}_{\{x \in \tilde{\mathsf{E}}\}}\}$$

- Denote $P_1 \succsim P_0$ if for all $x \in \tilde{\mathsf{E}}$ and $B \in \tilde{\mathcal{E}}$

$$P_1(x, B\backslash\{x\}) \geq P_0(x, B\backslash\{x\})$$

Do we have

$$P_1 \succsim P_0 \quad \Longrightarrow \quad \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \quad \text{for all } f \in \mathcal{F}\,?$$

Attempt #2: conditional Peskun-Tierney ordering

Setup:

- Let $\tilde{\mathsf{E}} \subset \mathsf{E}$ with $\pi(\tilde{\mathsf{E}}) > 0$ and let $\tilde{\mathcal{E}}$ be a σ-algebra on $\tilde{\mathsf{E}}$
- Interest lies specifically on the conditional distribution

$$\pi_{\tilde{\mathsf{E}}}(\mathrm{d}x) \propto \pi(\mathrm{d}x)\mathbb{1}_{\{x \in \tilde{\mathsf{E}}\}}$$

- Test functions are

$$\mathcal{F} := \{f \in \mathrm{L}_2(\pi) \ : \ \forall x \in \mathsf{E},\ f(x) = f(x)\mathbb{1}_{\{x \in \tilde{\mathsf{E}}\}}\}$$

- Denote $P_1 \succsim P_0$ if for all $x \in \tilde{\mathsf{E}}$ and $B \in \tilde{\mathcal{E}}$

$$P_1(x, B\backslash\{x\}) \geq P_0(x, B\backslash\{x\})$$

Do we have

$$P_1 \succsim P_0 \quad \Longrightarrow \quad \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \quad \text{for all } f \in \mathcal{F}\,?$$

Proposition 2

In the previous setup, assume that for $i \in \{0, 1\}$

$$P_1(x, \mathsf{E}\backslash\tilde{\mathsf{E}}) \leq P_0(x, \mathsf{E}\backslash\tilde{\mathsf{E}})\,, \qquad x \in \mathsf{E}$$

then

$$P_1 \succsim P_0 \Rightarrow P_0 - P_1 \text{ is positive on } \mathcal{F} \not\Rightarrow \mathrm{var}(P_1, f) \leq \mathrm{var}(P_0, f)\,, \quad \text{for all } f \in \mathcal{F}$$

The Peskun-Tierney route

Peskun and Tierney have used the following fact in their proof

$$f \in \mathsf{L}_{2,0}(\pi)\,, \qquad \mathrm{var}(P,f) = \lim_{\lambda\to 0} \mathbb{V}ar_\pi f(X)\left[1 + 2\sum_{t=1}^{\infty}\lambda^t \langle f, P^t f\rangle\right].$$

It is remarkable to be able to show that

$$\mathrm{var}(P_1,f) \le \mathrm{var}(P_0,f)\,, \quad \text{for all } f \in \mathsf{L}_{2,0}(\pi)$$

given *only* that

- $\langle f, P_1 f\rangle \le \langle f, P_0 f\rangle$ (for all $f \in \mathsf{L}_{2,0}(\pi)$)

and in particular by not caring whether or not

- $\langle f, P_1^t f\rangle \le \langle f, P_0^t f\rangle\,, \quad t > 1$

The Peskun-Tierney route

Peskun and Tierney have used the following fact in their proof

$$f \in \mathsf{L}_{2,0}(\pi)\,, \qquad \mathrm{var}(P, f) = \lim_{\lambda \to 0} \mathbb{V}ar_\pi f(X) \left[1 + 2 \sum_{t=1}^{\infty} \lambda^t \langle f, P^t f \rangle \right] .$$

It is remarkable to be able to show that

$$\mathrm{var}(P_1, f) \le \mathrm{var}(P_0, f)\,, \quad \text{for all } f \in \mathsf{L}_{2,0}(\pi)$$

given *only* that

- $\langle f, P_1 f \rangle \le \langle f, P_0 f \rangle$ (for all $f \in \mathsf{L}_{2,0}(\pi)$)

and in particular by not caring whether or not

- $\langle f, P_1^t f \rangle \le \langle f, P_0^t f \rangle\,, \quad t > 1$

A careful analysis of Peskun and Tierney's proof show that it is *essential* that

$$P_0 - P_1 \quad \text{be a positive operator on the } \underline{\text{whole}}\ \mathsf{L}_{2,0}(\pi)$$

Application 1: a weighted random scan Gibbs sampler

Consider $\mathsf{E} = \{1, \ldots, m\}^d$ the d-dimensional hypercube with m-length side and

$$\pi = (1-\sigma)\mathsf{unif}(\tilde{\mathsf{E}}) + \sigma\mathsf{unif}(\mathsf{E}\backslash\tilde{\mathsf{E}}), \quad \sigma \in [0,1)$$

where $\tilde{\mathsf{E}}$ is the sequence of neighbouring states along edges as follow :

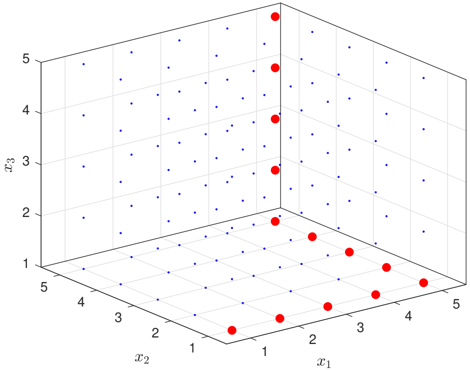


Figure: Illustration of π with $d = 3$ and $m = 5$.

Here, sampling from π is easy but pretend it is not.

Application 1: a weighted random scan Gibbs sampler

Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly at random $I \sim \text{unif}(1, \ldots, d)$

Application 1: a weighted random scan Gibbs sampler

Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly at random $I \sim \text{unif}(1, \dots, d)$
2. $Z \sim \pi_I(\,\cdot\,|\,X_t(1), \dots, X_t(i-1), X_t(i+1), \dots, X_t(d))$

Application 1: a weighted random scan Gibbs sampler

Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

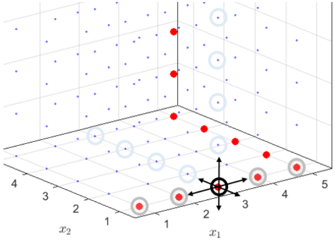
1. draw a direction uniformly at random $I \sim \text{unif}(1, \ldots, d)$
2. $Z \sim \pi_I(\,\cdot \mid X_t(1), \ldots, X_t(i-1), X_t(i+1), \ldots, X_t(d))$
3. set $X_{t+1} = (X_t(1), \ldots, X_t(i-1), Z, X_t(i+1), \ldots, X_t(d))$

Application 1: a weighted random scan Gibbs sampler

Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly at random $I \sim \text{unif}(1, \dots, d)$
2. $Z \sim \pi_I(\,\cdot \mid X_t(1), \dots, X_t(i-1), X_t(i+1), \dots, X_t(d))$
3. set $X_{t+1} = (X_t(1), \dots, X_t(i-1), Z, X_t(i+1), \dots, X_t(d))$

Possible next states from $X_t = (3, 1, 1)$ are circled

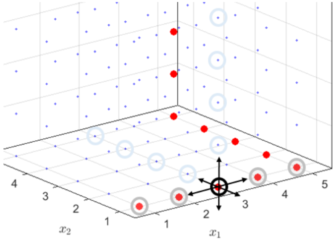


Application 1: a weighted random scan Gibbs sampler

Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly at random $I \sim \text{unif}(1, \ldots, d)$
2. $Z \sim \pi_I(\,\cdot\,|\,X_t(1), \ldots, X_t(i-1), X_t(i+1), \ldots, X_t(d))$
3. set $X_{t+1} = (X_t(1), \ldots, X_t(i-1), Z, X_t(i+1), \ldots, X_t(d))$

Possible next states from $X_t = (3, 1, 1)$ are circled



We have

$$\Pr[X_t = X_{t+1}] \geq \frac{2}{3}\frac{1-\sigma}{1+3\sigma} = \frac{d-1}{d}\frac{1-\sigma}{1+(m-2)\sigma}$$

consider the case $d \to \infty$?

Application 1: a weighted random scan Gibbs sampler

Define for all $i \in \{1, \ldots, d\}$

$$p_i(x) \propto \sum_{a=1}^{d} \pi(x(1), x(2), \ldots, a, \ldots, x(d)) \,, \quad x = (x(1), \ldots, x(d)) \in \mathsf{E}$$

such that $(p_1(x), \ldots, p_d(x))$ is a probability on $\{1, \ldots, d\}$.

Application 1: a weighted random scan Gibbs sampler

Define for all $i \in \{1, \ldots, d\}$

$$p_i(x) \propto \sum_{a=1}^{d} \pi(x(1), x(2), \ldots, a, \ldots, x(d)), \quad x = (x(1), \ldots, x(d)) \in \mathsf{E}$$

such that $(p_1(x), \ldots, p_d(x))$ is a probability on $\{1, \ldots, d\}$.

Locally weighted Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

Application 1: a weighted random scan Gibbs sampler

Define for all $i \in \{1, \ldots, d\}$

$$p_i(x) \propto \sum_{a=1}^{d} \pi(x(1), x(2), \ldots, a, \ldots, x(d)), \quad x = (x(1), \ldots, x(d)) \in \mathsf{E}$$

such that $(p_1(x), \ldots, p_d(x))$ is a probability on $\{1, \ldots, d\}$.

Locally weighted Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly according to $I \sim (p_1(X_t), p_2(X_t), \ldots, p_d(X_t))$

Application 1: a weighted random scan Gibbs sampler

Define for all $i \in \{1, \ldots, d\}$

$$p_i(x) \propto \sum_{a=1}^{d} \pi(x(1), x(2), \ldots, a, \ldots, x(d)), \quad x = (x(1), \ldots, x(d)) \in \mathsf{E}$$

such that $(p_1(x), \ldots, p_d(x))$ is a probability on $\{1, \ldots, d\}$.

Locally weighted Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly according to $I \sim (p_1(X_t), p_2(X_t), \ldots, p_d(X_t))$
2. $Z \sim \pi_I(\cdot \,|\, X_t(1), \ldots, X_t(i-1), X_t(i+1), \ldots, X_t(d))$

Application 1: a weighted random scan Gibbs sampler

Define for all $i \in \{1, \dots, d\}$

$$p_i(x) \propto \sum_{a=1}^{d} \pi(x(1), x(2), \dots, a, \dots, x(d)), \quad x = (x(1), \dots, x(d)) \in \mathsf{E}$$

such that $(p_1(x), \dots, p_d(x))$ is a probability on $\{1, \dots, d\}$.

Locally weighted Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly according to $I \sim (p_1(X_t), p_2(X_t), \dots, p_d(X_t))$
2. $Z \sim \pi_I(\cdot \,|\, X_t(1), \dots, X_t(i-1), X_t(i+1), \dots, X_t(d))$
3. let $X' = (X_t(1), \dots, X_t(i-1), Z, X_t(i+1), \dots, X_t(d))$

Application 1: a weighted random scan Gibbs sampler

Define for all $i \in \{1, \ldots, d\}$

$$p_i(x) \propto \sum_{a=1}^{d} \pi(x(1), x(2), \ldots, a, \ldots, x(d)), \quad x = (x(1), \ldots, x(d)) \in \mathsf{E}$$

such that $(p_1(x), \ldots, p_d(x))$ is a probability on $\{1, \ldots, d\}$.

Locally weighted Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly according to $I \sim (p_1(X_t), p_2(X_t), \ldots, p_d(X_t))$
2. $Z \sim \pi_I(\cdot \mid X_t(1), \ldots, X_t(i-1), X_t(i+1), \ldots, X_t(d))$
3. let $X' = (X_t(1), \ldots, X_t(i-1), Z, X_t(i+1), \ldots, X_t(d))$
4. set $X_{t+1} = X'$ w.p. $\min(1, p_I(X')/p_I(X_t))$ and $X_{t+1} = X_t$ otherwise

Application 1: a weighted random scan Gibbs sampler

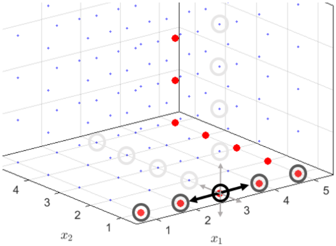
Define for all $i \in \{1, \ldots, d\}$

$$p_i(x) \propto \sum_{a=1}^{d} \pi(x(1), x(2), \ldots, a, \ldots, x(d)), \quad x = (x(1), \ldots, x(d)) \in \mathsf{E}$$

such that $(p_1(x), \ldots, p_d(x))$ is a probability on $\{1, \ldots, d\}$.

Locally weighted Random scan Gibbs sampler to sample from π: a Markov chain $\{X_t\}$ whose transition $X_t \mapsto X_{t+1}$ is as follows:

1. draw a direction uniformly according to $I \sim (p_1(X_t), p_2(X_t), \ldots, p_d(X_t))$
2. $Z \sim \pi_I(\,\cdot\,|\,X_t(1), \ldots, X_t(i-1), X_t(i+1), \ldots, X_t(d))$
3. let $X' = (X_t(1), \ldots, X_t(i-1), Z, X_t(i+1), \ldots, X_t(d))$
4. set $X_{t+1} = X'$ w.p. $\min(1, p_I(X')/p_I(X_t))$ and $X_{t+1} = X_t$ otherwise



Application 1: a weighted random scan Gibbs sampler

We have two competing algorithms to sample from π

- P_0 the usual random scan Gibbs sampler,
- P_1 the locally weighted random scan Gibbs sampler, which draws the direction in an "informed' way.

Application 1: a weighted random scan Gibbs sampler

We have two competing algorithms to sample from π

- P_0 the usual random scan Gibbs sampler,
- P_1 the locally weighted random scan Gibbs sampler, which draws the direction in an "informed' way.

Proposition 3

There is a conditional Peskun ordering, conditionally on $\tilde{\mathsf{E}}$ *:*

$$(x, y) \in \tilde{\mathsf{E}}^2 , \quad x \neq y , \qquad P_1(x, y) \geq P_0(x, y) .$$

Application 1: a weighted random scan Gibbs sampler

We have two competing algorithms to sample from π

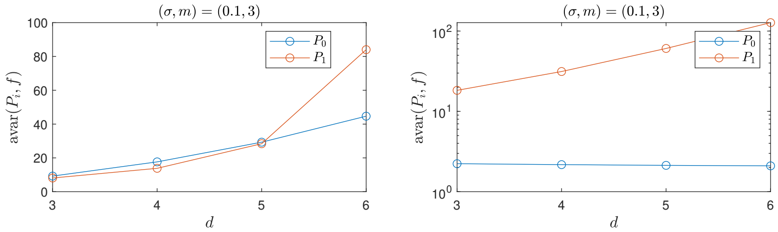
- P_0 the usual random scan Gibbs sampler,
- P_1 the locally weighted random scan Gibbs sampler, which draws the direction in an "informed' way.

Proposition 3

There is a conditional Peskun ordering, conditionally on $\tilde{\mathsf{E}}$:

$$(x, y) \in \tilde{\mathsf{E}}^2, \quad x \neq y, \qquad P_1(x, y) \geq P_0(x, y).$$

Plot of $\mathrm{var}(P, f)$ for different d and two test functions $f(x) = \mathbb{1}_{\{x=(1,1,1)\}}$ (left) and $f(x) = \mathbb{1}_{\{x=(1,2,1)\}}$ (right).



Application 1: a weighted random scan Gibbs sampler

- The previous results show that $\mathrm{var}(P_1, f) \nleq \mathrm{var}(P_0, f)$ for all $f \in \mathsf{L}_{2,0}(\pi)$ and in fact it seems that for a lot of those functions $\mathrm{var}(P_1, f) \gg \mathrm{var}(P_0, f)$!

Application 1: a weighted random scan Gibbs sampler

- The previous results show that $\mathrm{var}(P_1, f) \nleq \mathrm{var}(P_0, f)$ for all $f \in \mathsf{L}_{2,0}(\pi)$ and in fact it seems that for a lot of those functions $\mathrm{var}(P_1, f) \gg \mathrm{var}(P_0, f)$!
- However the following somewhat 'degenerate' case is interesting.

Application 1: a weighted random scan Gibbs sampler

- The previous results show that $\mathrm{var}(P_1, f) \not\leq \mathrm{var}(P_0, f)$ for all $f \in \mathsf{L}_{2,0}(\pi)$ and in fact it seems that for a lot of those functions $\mathrm{var}(P_1, f) \gg \mathrm{var}(P_0, f)$!
- However the following somewhat 'degenerate' case is interesting.

Proposition 4

Let $\sigma = 0$. Then for all $f \in L_{2,0}(\pi)$,

$$\mathrm{var}(P_1, f) \leq \frac{2}{d}\mathrm{var}(P_0, f)\,.$$

Application 1: a weighted random scan Gibbs sampler

- The previous results show that $\mathrm{var}(P_1, f) \not\leq \mathrm{var}(P_0, f)$ for all $f \in \mathsf{L}_{2,0}(\pi)$ and in fact it seems that for a lot of those functions $\mathrm{var}(P_1, f) \gg \mathrm{var}(P_0, f)$!
- However the following somewhat 'degenerate' case is interesting.

Proposition 4

Let $\sigma = 0$. Then for all $f \in L_{2,0}(\pi)$,

$$\mathrm{var}(P_1, f) \leq \frac{2}{d}\mathrm{var}(P_0, f)\,.$$

$\Rightarrow$ continuity as $\sigma \to 0$?

Application 1: a weighted random scan Gibbs sampler

- The previous results show that $\mathrm{var}(P_1, f) \not\leq \mathrm{var}(P_0, f)$ for all $f \in \mathsf{L}_{2,0}(\pi)$ and in fact it seems that for a lot of those functions $\mathrm{var}(P_1, f) \gg \mathrm{var}(P_0, f)$!
- However the following somewhat 'degenerate' case is interesting.

Proposition 4

Let $\sigma = 0$. Then for all $f \in L_{2,0}(\pi)$,

$$\mathrm{var}(P_1, f) \leq \frac{2}{d}\mathrm{var}(P_0, f)\,.$$

$\Rightarrow$ continuity as $\sigma \to 0$?

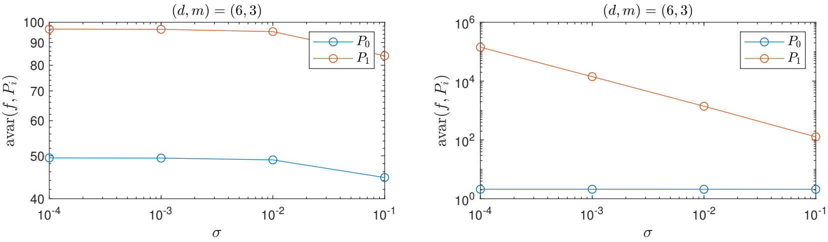


Figure: For the two functions $f(x) = \mathbb{1}_{\{x=(1,1,1)\}}$ (left) and $f(x) = \mathbb{1}_{\{x=(1,2,1)\}}$ (right).

Main result

Idea: working not on the whole $f \in \mathsf{L}_{2,0}(\pi)$ (too difficult), not on some fixed subset $\mathcal{F}$ which ignores parts of $\mathsf{L}_{2,0}(\pi)$ (not enough) but on some subset $\mathsf{L}_{2,0}(\tilde{\pi}_\theta)$ which eventually contains all the interesting functions in a limit sense.

Main result

Idea: working not on the whole $f \in \mathrm{L}_{2,0}(\pi)$ (too difficult), not on some fixed subset $\mathcal{F}$ which ignores parts of $\mathrm{L}_{2,0}(\pi)$ (not enough) but on some subset $\mathrm{L}_{2,0}(\tilde{\pi}_\theta)$ which eventually contains all the interesting functions in a limit sense.

- state-space and/or statistical model has a controllable parameter $\theta > 0$:

$$\mathsf{E} \equiv \mathsf{E}_\theta, \ \pi \equiv \pi_\theta, \ P_i \equiv P_{i,\theta}$$

Main result

Idea: working not on the whole $f \in \mathsf{L}_{2,0}(\pi)$ (too difficult), not on some fixed subset $\mathcal{F}$ which ignores parts of $\mathsf{L}_{2,0}(\pi)$ (not enough) but on some subset $\mathsf{L}_{2,0}(\tilde{\pi}_\theta)$ which eventually contains all the interesting functions in a limit sense.

- state-space and/or statistical model has a controllable parameter $\theta > 0$:

$$\mathsf{E} \equiv \mathsf{E}_\theta,\ \pi \equiv \pi_\theta,\ P_i \equiv P_{i,\theta}$$

- define restricted kernels to some subset $\tilde{\mathsf{E}}_\theta \subset \mathsf{E}_\theta$

$$\tilde{P}_{i,\theta}(x, B) := P_{i,\theta}(x, B \cap \tilde{\mathsf{E}}_\theta) + \delta_x(B) P_{i,\theta}(x, B \backslash \tilde{\mathsf{E}}_\theta)\,, \qquad (x, B) \in \tilde{\mathsf{E}}_\theta \times \tilde{\mathcal{E}}_\theta$$

Main result

Idea: working not on the whole $f \in \mathsf{L}_{2,0}(\pi)$ (too difficult), not on some fixed subset $\mathcal{F}$ which ignores parts of $\mathsf{L}_{2,0}(\pi)$ (not enough) but on some subset $\mathsf{L}_{2,0}(\tilde{\pi}_\theta)$ which eventually contains all the interesting functions in a limit sense.

- state-space and/or statistical model has a controllable parameter $\theta > 0$:

$$\mathsf{E} \equiv \mathsf{E}_\theta, \; \pi \equiv \pi_\theta, \; P_i \equiv P_{i,\theta}$$

- define restricted kernels to some subset $\tilde{\mathsf{E}}_\theta \subset \mathsf{E}_\theta$

$$\tilde{P}_{i,\theta}(x, B) := P_{i,\theta}(x, B \cap \tilde{\mathsf{E}}_\theta) + \delta_x(B) P_{i,\theta}(x, B \backslash \tilde{\mathsf{E}}_\theta) , \qquad (x, B) \in \tilde{\mathsf{E}}_\theta \times \tilde{\mathcal{E}}_\theta$$

If we ask that

- π_θ concentrates on $\tilde{\mathsf{E}}_\theta$,

$$\lim_{\theta \to \infty} \pi_\theta(\tilde{\mathsf{E}}_\theta) = 1$$

- a Peskun-Tierney ordering holds for the restricted kernels

$$\tilde{P}_{1,\theta}(x, B \backslash \{x\}) \geq \tilde{P}_{0,\theta}(x, B \backslash \{x\})$$

  for each $\theta > 0$.

Do we have that for a large enough θ,

$$\mathrm{var}(P_{1,\theta}, f) \leq \mathrm{var}(P_{0,\theta}, f) , \quad \text{for all } f \in \mathcal{F}$$

for a rich enough class of functions $\mathcal{F}$?

Main result

Assume that

- π_θ concentrates on $\tilde{\mathsf{E}}_\theta$,

$$\lim_{\theta \to \infty} \pi_\theta(\tilde{\mathsf{E}}_\theta) = 1$$

- a Peskun-Tierney ordering holds for the restricted kernels

$$\tilde{P}_{1,\theta}(x, B\backslash\{x\}) \geq \omega(\theta)\tilde{P}_{0,\theta}(x, B\backslash\{x\})$$

  for each $\theta > 0$ with

$$\lim_{\theta \to \infty} \omega(\theta) = 1\,.$$

Main result

Assume that

- π_θ concentrates on $\tilde{\mathsf{E}}_\theta$,

$$\lim_{\theta\to\infty} \pi_\theta(\tilde{\mathsf{E}}_\theta) = 1$$

- a Peskun-Tierney ordering holds for the restricted kernels

$$\tilde{P}_{1,\theta}(x, B\backslash\{x\}) \geq \omega(\theta)\tilde{P}_{0,\theta}(x, B\backslash\{x\})$$

  for each $\theta > 0$ with

$$\lim_{\theta\to\infty} \omega(\theta) = 1\,.$$

Theorem 1

Under the previous assumptions, assume in addition that the right spectral gaps of $(P_{i,\theta}, \tilde{P}_{i,\theta})$, $i \in \{0, 1\}$ are bounded away from zero. Then for all $\varepsilon > 0$, for certain collection of functions $\{f_\theta\} \in \mathcal{F}$, there exists $\theta_0 \equiv \theta_0(f_\theta) > 0$, such that

$$\theta > \theta_0 \Rightarrow \mathrm{var}(P_{1,\theta}, f_\theta) \leq \frac{1}{1-\varepsilon}\mathrm{var}(P_{0,\theta}, f_\theta) + \varepsilon\,.$$

Remarks:

Main result

Assume that

- π_θ concentrates on $\tilde{\mathsf{E}}_\theta$,
$$\lim_{\theta\to\infty} \pi_\theta(\tilde{\mathsf{E}}_\theta) = 1$$
- a Peskun-Tierney ordering holds for the restricted kernels
$$\tilde{P}_{1,\theta}(x, B\backslash\{x\}) \geq \omega(\theta)\tilde{P}_{0,\theta}(x, B\backslash\{x\})$$
for each $\theta > 0$ with
$$\lim_{\theta\to\infty} \omega(\theta) = 1\,.$$

Theorem 1

Under the previous assumptions, assume in addition that the right spectral gaps of $(P_{i,\theta}, \tilde{P}_{i,\theta})$, $i \in \{0, 1\}$ are bounded away from zero. Then for all $\varepsilon > 0$, for certain collection of functions $\{f_\theta\} \in \mathcal{F}$, there exists $\theta_0 \equiv \theta_0(f_\theta) > 0$, such that

$$\theta > \theta_0 \Rightarrow \mathrm{var}(P_{1,\theta}, f_\theta) \leq \frac{1}{1-\varepsilon}\mathrm{var}(P_{0,\theta}, f_\theta) + \varepsilon\,.$$

Remarks:

- here $\mathcal{F}$ is a class of functions for which $\|f_\theta\|_{2+\delta}$ does not grow too fast comparatively to $1/(1 - \pi(\tilde{\mathsf{E}}_\theta))$

Main result

Assume that

- π_θ concentrates on $\tilde{\mathsf{E}}_\theta$,

$$\lim_{\theta\to\infty} \pi_\theta(\tilde{\mathsf{E}}_\theta) = 1$$

- a Peskun-Tierney ordering holds for the restricted kernels

$$\tilde{P}_{1,\theta}(x, B\backslash\{x\}) \geq \omega(\theta)\tilde{P}_{0,\theta}(x, B\backslash\{x\})$$

  for each $\theta > 0$ with

$$\lim_{\theta\to\infty} \omega(\theta) = 1\,.$$

Theorem 1

Under the previous assumptions, assume in addition that the right spectral gaps of $(P_{i,\theta}, \tilde{P}_{i,\theta})$, $i \in \{0,1\}$ are bounded away from zero. Then for all $\varepsilon > 0$, for certain collection of functions $\{f_\theta\} \in \mathcal{F}$, there exists $\theta_0 \equiv \theta_0(f_\theta) > 0$, such that

$$\theta > \theta_0 \Rightarrow \mathrm{var}(P_{1,\theta}, f_\theta) \leq \frac{1}{1-\varepsilon}\mathrm{var}(P_{0,\theta}, f_\theta) + \varepsilon\,.$$

Remarks:

- here $\mathcal{F}$ is a class of functions for which $\|f_\theta\|_{2+\delta}$ does not grow too fast comparatively to $1/(1-\pi(\tilde{\mathsf{E}}_\theta))$
- we can relax the spectral gap assumptions, but result holds for a smaller class of functions $\mathcal{F}$ and needs π to concentrates sufficiently fast on E_θ

Application 1: a weighted random scan Gibbs sampler

- Unsurprisingly, the spectral gap of the locally weighted RSGS P_1 goes to zero when $\sigma \to 0$, so it is not covered by Theorem 1.

Application 1: a weighted random scan Gibbs sampler

- Unsurprisingly, the spectral gap of the locally weighted RSGS P_1 goes to zero when $\sigma \to 0$, so it is not covered by Theorem 1.
- We consider a slight modification of P_1, denoted P_1^* whose weights are now defined as

$$p^*(x) = (p_1^*(x), p_2^*(x), \ldots, p_d^*(x)), \quad p_i^*(x) = \max(p_i(x), 1/d^2).$$

Application 1: a weighted random scan Gibbs sampler

- Unsurprisingly, the spectral gap of the locally weighted RSGS P_1 goes to zero when $\sigma \to 0$, so it is not covered by Theorem 1.
- We consider a slight modification of P_1, denoted P_1^* whose weights are now defined as

$$p^*(x) = (p_1^*(x), p_2^*(x), \ldots, p_d^*(x)), \quad p_i^*(x) = \max(p_i(x), 1/d^2).$$

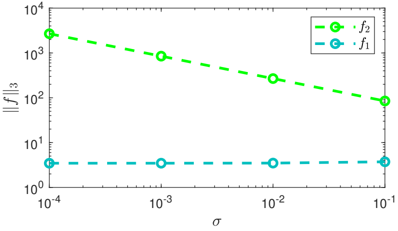
- The spectral gap of P_1^* is now bounded away from 0 and Theorem 1 applies for some $\mathcal{F}$.

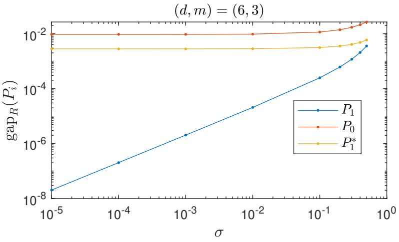
Application 1: a weighted random scan Gibbs sampler

- Unsurprisingly, the spectral gap of the locally weighted RSGS P_1 goes to zero when $\sigma \to 0$, so it is not covered by Theorem 1.
- We consider a slight modification of P_1, denoted P_1^* whose weights are now defined as

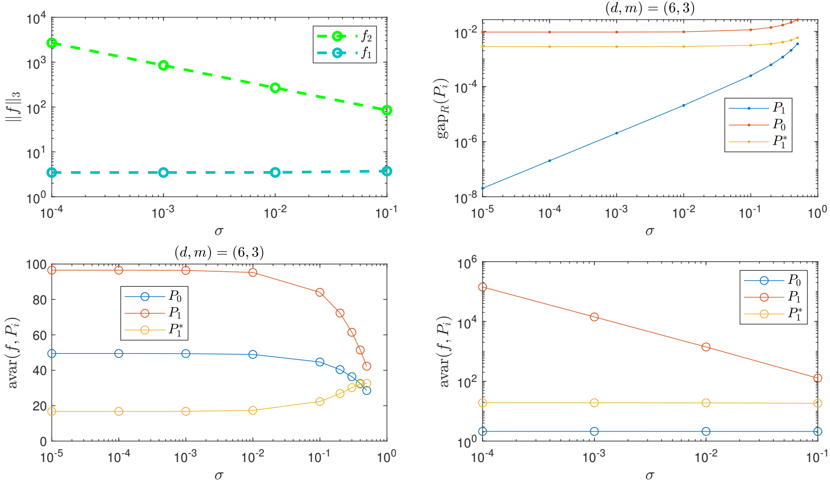
$$p^*(x) = (p_1^*(x), p_2^*(x), \ldots, p_d^*(x)), \quad p_i^*(x) = \max(p_i(x), 1/d^2).$$

- The spectral gap of P_1^* is now bounded away from 0 and Theorem 1 applies for some $\mathcal{F}$.





Application 1: a weighted random scan Gibbs sampler



Discussion

When are non-reversible Markov chains better?

- Gustafson's Guided Walk (GW) (1998): a very easy way to construct a non-reversible version of Random Walk Metropolis-Hastings (MH) samplers in one dimension.

When are non-reversible Markov chains better?

- Gustafson's Guided Walk (GW) (1998): a very easy way to construct a non-reversible version of Random Walk Metropolis-Hastings (MH) samplers in one dimension.
- Andrieu and Livingstone (2020) have shown that GW can never increase the asymptotic variance of a MH.

When are non-reversible Markov chains better?

- Gustafson's Guided Walk (GW) (1998): a very easy way to construct a non-reversible version of Random Walk Metropolis-Hastings (MH) samplers in one dimension.
- Andrieu and Livingstone (2020) have shown that GW can never increase the asymptotic variance of a MH.
- Gagnon and Maire (2020) propose a generalization of Gustafson's Guided-Walk for sampling from distributions defined on

$$\mathsf{E} = \{-1, 1\}^n$$

- similar algorithms have been proposed by Kamatani and Song (2020) and Power and Goldman (2020).

When are non-reversible Markov chains better?

- Gustafson's Guided Walk (GW) (1998): a very easy way to construct a non-reversible version of Random Walk Metropolis-Hastings (MH) samplers in one dimension.
- Andrieu and Livingstone (2020) have shown that GW can never increase the asymptotic variance of a MH.
- Gagnon and Maire (2020) propose a generalization of Gustafson's Guided-Walk for sampling from distributions defined on

$$\mathsf{E} = \{-1, 1\}^n$$

- similar algorithms have been proposed by Kamatani and Song (2020) and Power and Goldman (2020).
- these algorithms are all lifted non-reversible MCMC but the theory of Andrieu and Livingstone (2020) does not guarantee that they inherit from GW their superiority over MH.

We showed that a weak Peskun ordering holds between MH and our generalized Guided-Walk and thus we obtain conditions on which our generalized Guided-Walk is better than MH. Without those conditions, the Guided-Walk do not always dominate MH!

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

- needs a controllable parameter θ: noise level, state-space dimension, sample size, etc.

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

- needs a controllable parameter θ: noise level, state-space dimension, sample size, etc.
- key is to define a set $\tilde{\mathsf{E}}_\theta$ on which a (relaxed) Peskun-Tierney ordering holds between the kernels restricted to $\tilde{\mathsf{E}}_\theta$

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

- needs a controllable parameter θ: noise level, state-space dimension, sample size, etc.
- key is to define a set $\tilde{\mathsf{E}}_\theta$ on which a (relaxed) Peskun-Tierney ordering holds between the kernels restricted to $\tilde{\mathsf{E}}_\theta$
- results holds in a θ limit sense, provided π_θ concentrates on $\tilde{\mathsf{E}}_\theta$

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

- needs a controllable parameter θ: noise level, state-space dimension, sample size, etc.
- key is to define a set $\tilde{\mathsf{E}}_\theta$ on which a (relaxed) Peskun-Tierney ordering holds between the kernels restricted to $\tilde{\mathsf{E}}_\theta$
- results holds in a θ limit sense, provided π_θ concentrates on $\tilde{\mathsf{E}}_\theta$
- Not the whole $\mathsf{L}_{2,0}(\pi_\theta)$ is covered but only a subset of it

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

- needs a controllable parameter θ: noise level, state-space dimension, sample size, etc.
- key is to define a set $\tilde{\mathsf{E}}_\theta$ on which a (relaxed) Peskun-Tierney ordering holds between the kernels restricted to $\tilde{\mathsf{E}}_\theta$
- results holds in a θ limit sense, provided π_θ concentrates on $\tilde{\mathsf{E}}_\theta$
- Not the whole $\mathsf{L}_{2,0}(\pi_\theta)$ is covered but only a subset of it

We have applied this technique successfully to

- Defining a scalable locally-weighted Gibbs sampler which dominates the random scan Gibbs sampler useful for noise-vanishing distributions, here $\theta = \sigma$ the noise level

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

- needs a controllable parameter θ: noise level, state-space dimension, sample size, etc.
- key is to define a set $\tilde{\mathsf{E}}_\theta$ on which a (relaxed) Peskun-Tierney ordering holds between the kernels restricted to $\tilde{\mathsf{E}}_\theta$
- results holds in a θ limit sense, provided π_θ concentrates on $\tilde{\mathsf{E}}_\theta$
- Not the whole $\mathsf{L}_{2,0}(\pi_\theta)$ is covered but only a subset of it

We have applied this technique successfully to

- Defining a scalable locally-weighted Gibbs sampler which dominates the random scan Gibbs sampler useful for noise-vanishing distributions, here $\theta = \sigma$ the noise level
- Compare a generalization of the one-dimensional Lifted MCMC with MH, here $\mathsf{E} = \{-1, 1\}^n$ and $\theta = n$

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

- needs a controllable parameter θ: noise level, state-space dimension, sample size, etc.
- key is to define a set $\tilde{\mathsf{E}}_\theta$ on which a (relaxed) Peskun-Tierney ordering holds between the kernels restricted to $\tilde{\mathsf{E}}_\theta$
- results holds in a θ limit sense, provided π_θ concentrates on $\tilde{\mathsf{E}}_\theta$
- Not the whole $\mathsf{L}_{2,0}(\pi_\theta)$ is covered but only a subset of it

We have applied this technique successfully to

- Defining a scalable locally-weighted Gibbs sampler which dominates the random scan Gibbs sampler useful for noise-vanishing distributions, here $\theta = \sigma$ the noise level
- Compare a generalization of the one-dimensional Lifted MCMC with MH, here $\mathsf{E} = \{-1, 1\}^n$ and $\theta = n$

This research is related to a number of questions dealing with the analysis of Markov chains that are only (super)-efficient on a portion of the state-space

- Approximate spectral gaps for Markov chains mixing times in high dimensions, Atchadé (2019)

Discussion and related problems

We have presented a new technique to compare, up to an arbitrary low approximation error, the asymptotic variance of two π-rev. MCMC samplers

- needs a controllable parameter θ: noise level, state-space dimension, sample size, etc.
- key is to define a set $\tilde{\mathsf{E}}_\theta$ on which a (relaxed) Peskun-Tierney ordering holds between the kernels restricted to $\tilde{\mathsf{E}}_\theta$
- results holds in a θ limit sense, provided π_θ concentrates on $\tilde{\mathsf{E}}_\theta$
- Not the whole $\mathsf{L}_{2,0}(\pi_\theta)$ is covered but only a subset of it

We have applied this technique successfully to

- Defining a scalable locally-weighted Gibbs sampler which dominates the random scan Gibbs sampler useful for noise-vanishing distributions, here $\theta = \sigma$ the noise level
- Compare a generalization of the one-dimensional Lifted MCMC with MH, here $\mathsf{E} = \{-1, 1\}^n$ and $\theta = n$

This research is related to a number of questions dealing with the analysis of Markov chains that are only (super)-efficient on a portion of the state-space

- Approximate spectral gaps for Markov chains mixing times in high dimensions, Atchadé (2019)
- Complexity Results for MCMC derived from Quantitative Bounds, Yang and Rosenthal (2019)

References

Main work can be found (soon on arXiv)

- Maire and Vandekerkhove, Locally weighted aggregation of Markov kernels and applications to noise vanishing distribution sampling, (2021)
- Gagnon and Maire, Lifted samplers for partially ordered discrete state-space, (2021)

Other references:

Andrieu, Lee and Vihola (2018) Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers
Andrieu and Livingstone (2020) Peskun-Tierney ordering for Markov chain and process Monte Carlo: beyond the reversible scenario.
Andrieu and Vihola (2016) Establishing some order amongst exact approximations of MCMCs
Atchade (2019) Approximate spectral gaps for Markov chains mixing times in high dimensions.
Cogburn (1976) Asymptotic properties of stationary sequences.
Gagnon and Doucet (2019) Non-reversible jump algorithms for Bayesian nested model selection.
Gustafson (1998) A guided walk Metropolis algorithm
Jones (2004) On the Markov chain central limit theorem.
Kamatani and Song (2020) Non-reversible guided Metropolis-Hastings kernel.
Kipnis and Varadhan (1986) CLT for additive functionals of reversible Markov processes and applications to simple exclusions
Moriarty, Vogrinc and Zocca (2020) A Metropolis-class sampler for targets with non-convex support.
Peskun (1973) Optimum Monte-Carlo sampling using Markov chains.
Power and Goldman (2020) Accelerated sampling on discrete spaces with non-reversible Markov processes.
Roberts and Rosenthal (2004) General state space Markov chains and MCMC algorithms.
Tierney (1998) A note on Metropolis-Hastings kernels for general state spaces.
Yang and Rosenthal (2019) Complexity results for MCMC derived from quantitative bounds
Zanella (2020) Informed proposals for local MCMC in discrete spaces.

More on Lifted MCMC

Metropolis-Hastings: Random Walk on E

Two components:

Metropolis-Hastings: Random Walk on E

Two components:

(1) For all $x \in \mathrm{E}$, let $\mathfrak{N}(x)$ be a **neighborhood structure** on E s.t.

$$y \in \mathfrak{N}(x) \Rightarrow x \in \mathfrak{N}(y)\,.$$

Metropolis-Hastings: Random Walk on E

Two components:

(1) For all $x \in \mathrm{E}$, let $\mathfrak{N}(x)$ be a **neighborhood structure** on E s.t.

$$y \in \mathfrak{N}(x) \Rightarrow x \in \mathfrak{N}(y)\,.$$

(2) Let $\{R(x,\cdot)\}$ be a **collection of conditional dist.** on E s.t.

$$R(x, \mathfrak{N}(x)) = 1\,.$$

Metropolis-Hastings: Random Walk on E

Two components:

(1) For all $x \in \mathrm{E}$, let $\mathfrak{N}(x)$ be a **neighborhood structure** on E s.t.

$$y \in \mathfrak{N}(x) \Rightarrow x \in \mathfrak{N}(y)\,.$$

(2) Let $\{R(x,\cdot)\}$ be a **collection of conditional dist.** on E s.t.

$$R(x, \mathfrak{N}(x)) = 1\,.$$

Metropolis-Hastings   Set $X_0 \in \mathrm{E}$, $t = 0$

(i) propose a move:
- $\tilde{X} \sim R(X_t, \cdot)$

(ii) accept/reject of the move: set $X_{t+1} = \tilde{X}$ w.p.

$$\alpha_0(X_t, \tilde{X}) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R(\tilde{X}, X_t)}{R(X_t, \tilde{X})}$$

and set $X_{t+1} = X_t$ otherwise. Set $t \leftarrow t + 1$.

Repeat (i)-(ii) to generate $\{X_t\}$.

Lifted MCMC: generalization of the Guided Walk on E

Three components:

Lifted MCMC: generalization of the Guided Walk on E

Three components:

(1)–(2) A **neighborhood structure** and a **collection of conditional dist.**

$$\mathfrak{N}(x) \quad \text{and} \quad R(x,\cdot)\,, \qquad \forall\, x \in \mathsf{E}\,.$$

Lifted MCMC: generalization of the Guided Walk on E

Three components:

(1)–(2) A **neighborhood structure** and a **collection of conditional dist.**

$$\mathfrak{N}(x) \quad \text{and} \quad R(x, \cdot)\,, \qquad \forall\, x \in \mathsf{E}\,.$$

(3) A **neighborhood splitting mechanism**

$$\mathfrak{N}(x) = \mathfrak{N}_1(x) \cup \mathfrak{N}_{-1}(x) \quad \text{with} \quad \mathfrak{N}_1(x) \cap \mathfrak{N}_{-1}(x) = \emptyset\,.$$

Lifted MCMC: generalization of the Guided Walk on E

Three components:

(1)–(2) A **neighborhood structure** and a **collection of conditional dist.**

$$\mathfrak{N}(x) \quad \text{and} \quad R(x,\cdot)\,, \qquad \forall\, x \in \mathsf{E}\,.$$

(3) A **neighborhood splitting mechanism**

$$\mathfrak{N}(x) = \mathfrak{N}_1(x) \cup \mathfrak{N}_{-1}(x) \quad \text{with} \quad \mathfrak{N}_1(x) \cap \mathfrak{N}_{-1}(x) = \emptyset\,.$$

which also induces

$$R_1(x,\,\cdot\,) \propto R(x, \cdot \cap \mathfrak{N}_1(x)) \quad \text{and} \quad R_{-1}(x,\,\cdot\,) \propto R(x, \cdot \cap \mathfrak{N}_{-1}(x))$$

Lifted MCMC: generalization of the Guided Walk on E

Three components:

(1)–(2) A **neighborhood structure** and a **collection of conditional dist.**

$$\mathfrak{N}(x) \quad \text{and} \quad R(x,\cdot)\,, \qquad \forall\, x \in \mathsf{E}\,.$$

(3) A **neighborhood splitting mechanism**

$$\mathfrak{N}(x) = \mathfrak{N}_1(x) \cup \mathfrak{N}_{-1}(x) \quad \text{with} \quad \mathfrak{N}_1(x) \cap \mathfrak{N}_{-1}(x) = \emptyset\,.$$

which also induces

$$R_1(x,\,\cdot\,) \propto R(x, \cdot \cap \mathfrak{N}_1(x)) \quad \text{and} \quad R_{-1}(x,\,\cdot\,) \propto R(x, \cdot \cap \mathfrak{N}_{-1}(x))$$

Lifted MCMC — Set $X_0 \in \mathsf{E}$, $\xi_0 \in \{-1, 1\}$, $t = 0$

(i) propose a move:
- $\tilde{X} \sim R_{\zeta_t}(X_t, \cdot)$

(ii) accept/reject of the move: set $(X_{t+1}, \zeta_{t+1}) = (\tilde{X}, \zeta_t)$ w.p.

$$\alpha_{\mathrm{LIF}}(X_t, \tilde{X} \,|\, \zeta_t) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_0)} \times \frac{R_{-\zeta_t}(\tilde{X}, X_t)}{R_{\zeta_t}(X_t, \tilde{X})}\,.$$

and $(X_{t+1}, \zeta_{t+1}) = (\tilde{X}, -\zeta_t)$. Set $t \leftarrow t + 1$.

Repeat (i)-(ii) to generate $\{X_t, \xi_t\}$.

Comparison Metropolis-Hastings vs Lifted MCMC

- Metropolis-Hastings P_0

  (i) $\tilde{X} \sim R(X_t, \cdot)$

  $$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_0(X_t, \tilde{X}) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R(\tilde{X}, X_t)}{R(X_t, \tilde{X})}$$

- Lifted MCMC P_1

  (i) $\tilde{X} \sim R_{\xi_t}(X_t, \cdot)$

  $$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_1(X_t, \tilde{X} \mid \xi_t) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R_{-\xi_t}(\tilde{X}, X_t)}{R_{\xi_t}(X_t, \tilde{X})}$$

Comparison Metropolis-Hastings vs Lifted MCMC

- Metropolis-Hastings P_0

$$\text{(i) } \tilde{X} \sim R(X_t, \cdot)$$

$$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_0(X_t, \tilde{X}) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R(\tilde{X}, X_t)}{R(X_t, \tilde{X})}$$

- Lifted MCMC P_1

$$\text{(i) } \tilde{X} \sim R_{\xi_t}(X_t, \cdot)$$

$$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_1(X_t, \tilde{X} \mid \xi_t) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R_{-\xi_t}(\tilde{X}, X_t)}{R_{\xi_t}(X_t, \tilde{X})}$$

- introduce a reversibilisation of P_1, called P_1^{rev}

$$\text{(i) } \xi_t \sim \text{unif}(-1, 1)\,,\ \tilde{X} \sim R_{\xi_t}(X_t, \cdot)$$

$$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_1(X_t, \tilde{X} \mid \xi_t) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R_{-\xi_t}(\tilde{X}, X_t)}{R_{\xi_t}(X_t, \tilde{X})}$$

Comparison Metropolis-Hastings vs Lifted MCMC

- Metropolis-Hastings P_0

  (i) $\tilde{X} \sim R(X_t, \cdot)$

  $$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_0(X_t, \tilde{X}) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R(\tilde{X}, X_t)}{R(X_t, \tilde{X})}$$

- Lifted MCMC P_1

  (i) $\tilde{X} \sim R_{\xi_t}(X_t, \cdot)$

  $$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_1(X_t, \tilde{X} \,|\, \xi_t) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R_{-\xi_t}(\tilde{X}, X_t)}{R_{\xi_t}(X_t, \tilde{X})}$$

- introduce a reversibilisation of P_1, called P_1^{rev}

  (i) $\xi_t \sim \text{unif}(-1, 1)$, $\tilde{X} \sim R_{\xi_t}(X_t, \cdot)$

  $$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_1(X_t, \tilde{X} \,|\, \xi_t) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R_{-\xi_t}(\tilde{X}, X_t)}{R_{\xi_t}(X_t, \tilde{X})}$$

From Andrieu and Livingstone (2020), we know that for all $f \in \mathrm{L}_2(\pi)$

$$\mathrm{var}(P_1, f) \leq \mathrm{var}(P_1^{\text{rev.}}, f)$$

Comparison Metropolis-Hastings vs Lifted MCMC

- Metropolis-Hastings P_0

  (i) $\tilde{X} \sim R(X_t, \cdot)$

  $$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_0(X_t, \tilde{X}) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R(\tilde{X}, X_t)}{R(X_t, \tilde{X})}$$

- Lifted MCMC P_1

  (i) $\tilde{X} \sim R_{\xi_t}(X_t, \cdot)$

  $$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_1(X_t, \tilde{X} \mid \xi_t) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R_{-\xi_t}(\tilde{X}, X_t)}{R_{\xi_t}(X_t, \tilde{X})}$$

- introduce a reversibilisation of P_1, called P_1^{rev}

  (i) $\xi_t \sim \text{unif}(-1, 1)$, $\tilde{X} \sim R_{\xi_t}(X_t, \cdot)$

  $$\text{(ii) accept/reject } \tilde{X} \text{ w.p. } \alpha_1(X_t, \tilde{X} \mid \xi_t) = 1 \wedge \frac{\pi(\tilde{X})}{\pi(X_t)} \times \frac{R_{-\xi_t}(\tilde{X}, X_t)}{R_{\xi_t}(X_t, \tilde{X})}$$

From Andrieu and Livingstone (2020), we know that for all $f \in \mathrm{L}_2(\pi)$

$$\operatorname{var}(P_1, f) \leq \operatorname{var}(P_1^{\text{rev.}}, f) \overset{?}{\gtreqless} \operatorname{var}(P_0, f)$$

Comparison Metropolis-Hastings vs Lifted MCMC

Recall $\mathsf{E}_n = \{-1, 1\}^n$, define

- $R(x, \cdot) = \mathrm{unif}(\mathfrak{N}(x))$, $R_1(x, \cdot) = \mathrm{unif}(\mathfrak{N}_1(x))$, $R_{-1}(x, \cdot) = \mathrm{unif}(\mathfrak{N}_{-1}(x))$
- set $\mathfrak{N}(x) = \{y \in \mathsf{E} \; : \; \sum_{k=1}^{n} |x_k - y_k| = 2\}$ and $\mathfrak{N}_1(x) = \{y \in \mathfrak{N}(x) \; : \; y_k \geq x_k\}$.

Comparison Metropolis-Hastings vs Lifted MCMC

Recall $\mathsf{E}_n = \{-1, 1\}^n$, define

- $R(x, \cdot) = \text{unif}(\mathfrak{N}(x))$, $R_1(x, \cdot) = \text{unif}(\mathfrak{N}_1(x))$, $R_{-1}(x, \cdot) = \text{unif}(\mathfrak{N}_{-1}(x))$
- set $\mathfrak{N}(x) = \{y \in \mathsf{E} \, : \, \sum_{k=1}^n |x_k - y_k| = 2\}$ and $\mathfrak{N}_1(x) = \{y \in \mathfrak{N}(x) \, : \, y_k \geq x_k\}$.

We define the collection of sets $\{\tilde{\mathsf{E}}_n\}$

$$\tilde{\mathsf{E}}_n := \{x \in \mathsf{E}_n \, : n/2 - \varphi_n \leq |\mathfrak{N}_{-1}(x)|, |\mathfrak{N}_{+1}(x)| \leq n/2 + \varphi_n\},$$

where $\{\varphi_n\}$ is a sequence such that $\varphi_n = o(n)$.
$\Rightarrow \tilde{\mathsf{E}}_n$ contains those states for which $|\mathfrak{N}_{-1}(x)| \approx |\mathfrak{N}_{-1}(x)|$, when n is large.

Comparison Metropolis-Hastings vs Lifted MCMC

Recall $\mathsf{E}_n = \{-1, 1\}^n$, define

- $R(x, \cdot) = \text{unif}(\mathfrak{N}(x))$, $R_1(x, \cdot) = \text{unif}(\mathfrak{N}_1(x))$, $R_{-1}(x, \cdot) = \text{unif}(\mathfrak{N}_{-1}(x))$
- set $\mathfrak{N}(x) = \{y \in \mathsf{E} \, : \, \sum_{k=1}^{n} |x_k - y_k| = 2\}$ and $\mathfrak{N}_1(x) = \{y \in \mathfrak{N}(x) \, : \, y_k \geq x_k\}$.

We define the collection of sets $\{\tilde{\mathsf{E}}_n\}$

$$\tilde{\mathsf{E}}_n := \{x \in \mathsf{E}_n \, : \, n/2 - \varphi_n \leq |\mathfrak{N}_{-1}(x)|, |\mathfrak{N}_{+1}(x)| \leq n/2 + \varphi_n\},$$

where $\{\varphi_n\}$ is a sequence such that $\varphi_n = o(n)$.
$\Rightarrow$ $\tilde{\mathsf{E}}_n$ contains those states for which $|\mathfrak{N}_{-1}(x)| \approx |\mathfrak{N}_{-1}(x)|$, when n is large.

Theorem 2

Assume that

1. $\{\varphi_n\}$ *can be chosen so that* $\pi_n(\tilde{\mathsf{E}}_n) \to 1$
2. *the spectral gaps of* $(P_{i,n}, \tilde{P}_{i,n})$, $i \in \{0, 1\}$ *are bounded away from 0*

Then for any $\varepsilon > 0$ *and for certain collections of functions* $\{f_n\} \in \mathcal{F} \subset L_{2,0}(\pi_n)$ *with* $\|f_n\|_2 = 1$,

$$\text{var}(P_{1,n}, f_n) \leq \frac{1}{1-\varepsilon} \text{var}(P_{0,n}, f_n) + \varepsilon$$

holds for a sufficiently large $n \equiv n(f_n)$.

Ising model

We consider an Ising model defined on a two-dimensional lattice $\{1, \dots, \eta\} \times \{1, \dots, \eta\}$. A state is a lattice whose vertices are $\{-1, 1\}$ such that

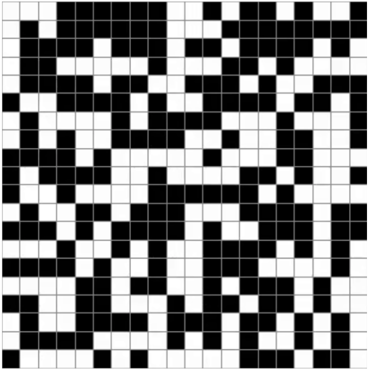


Figure: Realisation of an Ising lattice with $\eta = 20$, the lattice vertices are in $\{1, \dots, 20\} \times \{1, \dots, 20\}$. Black filling indicates that a vertice is 1 and white filling is -1.

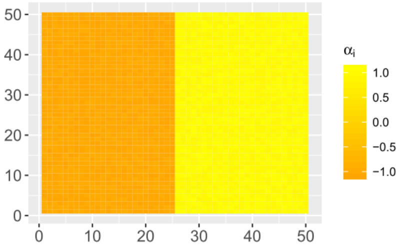
Ising model

We see each Ising lattice as a $n = \eta^2$-dimensional vector in $\{-1, 1\}^n$. The general model is

$$\pi(x) = \frac{1}{Z} \exp\left\{ \sum_{i=1}^{n} \alpha_i x_i + \lambda \sum_{i \sim j} x_i x_j \right\}$$

with $\alpha_i \in \mathbb{R}$, $\lambda \geq 0$ and $i \sim j$ is set of neighboring vertices on the lattice (typically North-South-West-East).

- λ can be seen as an interaction parameter and control the size of color patches
- α_i can be seen as a mean value field, for instance



$\Rightarrow$ λ and $\{\alpha_i\}$ are known (drawn $\alpha_i \sim_{iid} \mathcal{N}(\mu = 1, \tau)$), the goal is to sample from π.

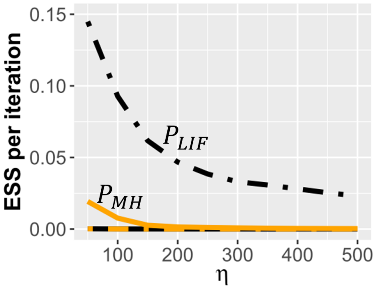
Empirical results 1

Here, we fix $\lambda = 1/2$ and $\{\alpha_i\}$ as in the previous figure and increase η, i.e. n.

Again, we define

$$\mathfrak{N}(x) = \left\{ y \in \mathrm{E} \,:\, \sum_{k=1}^{n} |x_k - y_k| = 2 \right\}, \qquad \mathfrak{N}_1(x) = \{ y \in \mathfrak{N}(x) \,:\, y_k \geq x_k \} \,.$$

We use the uniform and locally balanced proposal (Zanella, 2020) for $\{R(x, \cdot)\}$. Function of interest is $f_n(x) = \sum_{k=1}^{n} x_k$ (system magnetisation)

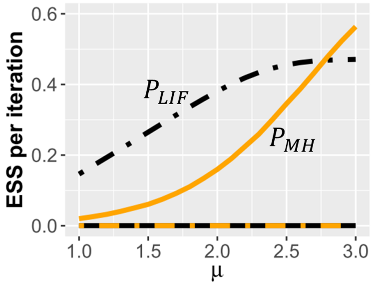


Quantitatively, we have that

$$\mathrm{var}(P_0, f_n)/\mathrm{var}(P_1, f_n) \in \{7, 10, 20\} \qquad \text{when} \qquad n \in \{50^2, 50^2 \times 10, 50^2 \times 100\}$$

Empirical results 2

We now fix $n = 50^2$ and $\lambda = 1/2$ but change $\{\alpha_i \sim \mathcal{N}(\mu, \tau)\}$ to increase the contrast between the two sides of the lattice: the higher μ the larger the contrast.



Increasing μ essentially increases the roughness of π_n which concentrates on a few configurations only, disabling the Lifed MCMC persistent move feature.