Some properties of exact approximations of the Metropolis-Hastings algorithm

Christophe Andrieu (joint work with Matti Vihola, University of Jyväskylä)

22nd January 2016

University of BRISTOL

Overview

- Assume we are interested in sampling from a probability distribution of density $\pi(x)$.
- Standard "universal" algorithms require one to evaluate $\pi(x)$.
- Assume for any $x \in \mathsf{X}$, "noisy" unbiased measurements of $\pi(x)$ are available.
- In recent years "novel" MCMC algorithms have been proposed in order to sample from $\pi(x)$ in this context.
- The main idea is to replace $\pi(x)$ with a noisy estimator whenever needed.
- A key point is that these algorithms can still be exact, but can be seen as being (random) approximations of algorithms which make us of $\pi(x)$.
- Here we focus on the theoretical properties of these noisy algorithms.

Overview

- Assume we are interested in sampling from a probability distribution of density $\pi(x)$.
- Standard "universal" algorithms require one to evaluate $\pi(x)$.
- Assume for any $x \in \mathsf{X}$, "noisy" unbiased measurements of $\pi(x)$ are available.
- In recent years "novel" MCMC algorithms have been proposed in order to sample from $\pi(x)$ in this context.
- The main idea is to replace $\pi(x)$ with a noisy estimator whenever needed.
- A key point is that these algorithms can still be exact, but can be seen as being (random) approximations of algorithms which make us of $\pi(x)$.
- Here we focus on the theoretical properties of these noisy algorithms.

Overview

- Assume we are interested in sampling from a probability distribution of density $\pi(x)$.
- Standard "universal" algorithms require one to evaluate $\pi(x)$.
- Assume for any $x \in \mathsf{X}$, "noisy" unbiased measurements of $\pi(x)$ are available.
- In recent years "novel" MCMC algorithms have been proposed in order to sample from $\pi(x)$ in this context.
- The main idea is to replace $\pi(x)$ with a noisy estimator whenever needed.
- A key point is that these algorithms can still be exact, but can be seen as being (random) approximations of algorithms which make us of $\pi(x)$.
- Here we focus on the theoretical properties of these noisy algorithms.

Overview

- Assume we are interested in sampling from a probability distribution of density $\pi(x)$.
- Standard "universal" algorithms require one to evaluate $\pi(x)$.
- Assume for any $x \in \mathsf{X}$, "noisy" unbiased measurements of $\pi(x)$ are available.
- In recent years "novel" MCMC algorithms have been proposed in order to sample from $\pi(x)$ in this context.
- The main idea is to replace $\pi(x)$ with a noisy estimator whenever needed.
- A key point is that these algorithms can still be exact, but can be seen as being (random) approximations of algorithms which make us of $\pi(x)$.
- Here we focus on the theoretical properties of these noisy algorithms.

Overview

- Assume we are interested in sampling from a probability distribution of density $\pi(x)$.
- Standard "universal" algorithms require one to evaluate $\pi(x)$.
- Assume for any $x \in \mathsf{X}$, "noisy" unbiased measurements of $\pi(x)$ are available.
- In recent years "novel" MCMC algorithms have been proposed in order to sample from $\pi(x)$ in this context.
- The main idea is to replace $\pi(x)$ with a noisy estimator whenever needed.
- A key point is that these algorithms can still be exact, but can be seen as being (random) approximations of algorithms which make us of $\pi(x)$.
- Here we focus on the theoretical properties of these noisy algorithms.

Overview

- Assume we are interested in sampling from a probability distribution of density $\pi(x)$.
- Standard "universal" algorithms require one to evaluate $\pi(x)$.
- Assume for any $x \in \mathsf{X}$, "noisy" unbiased measurements of $\pi(x)$ are available.
- In recent years "novel" MCMC algorithms have been proposed in order to sample from $\pi(x)$ in this context.
- The main idea is to replace $\pi(x)$ with a noisy estimator whenever needed.
- A key point is that these algorithms can still be exact, but can be seen as being (random) approximations of algorithms which make us of $\pi(x)$.
- Here we focus on the theoretical properties of these noisy algorithms.

Overview

- Assume we are interested in sampling from a probability distribution of density $\pi(x)$.
- Standard "universal" algorithms require one to evaluate $\pi(x)$.
- Assume for any $x \in \mathsf{X}$, "noisy" unbiased measurements of $\pi(x)$ are available.
- In recent years "novel" MCMC algorithms have been proposed in order to sample from $\pi(x)$ in this context.
- The main idea is to replace $\pi(x)$ with a noisy estimator whenever needed.
- A key point is that these algorithms can still be exact, but can be seen as being (random) approximations of algorithms which make us of $\pi(x)$.
- Here we focus on the theoretical properties of these noisy algorithms.

Latent variables and pseudo-marginals

- Assume interest is in a posterior distribution

$$\pi(x) = p(x|y) \propto p(x)p(y|x) = p(x)\int p(y,z|x)dz$$

where the integral cannot be computed analytically.

- Then with $z_i \overset{\text{iid}}{\sim} Q_x$ and $p(y,z|x)/Q_x(z)$ well defined, consider an IS approximation of the likelihood

$$\frac{1}{N}\sum_{i=1}^{N}\frac{p(y,z_i|x)}{Q_x(z_i)}$$

This is a noisy measurement of the intractable "likelihood" $p(y|x)$.

- One gets a noisy measurement (up to a constant) of the posterior distribution with

$$\begin{aligned}\hat{\pi}^N(x) &\propto p(x)\left[\int p(y,z|x)dz\right] \times \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{p(y,z_i|x)}{Q_x(z_i)}}{\int p(y,z|x)dz} \\ &\propto \pi(x) \times w\end{aligned}$$

Latent variables and pseudo-marginals

- Assume interest is in a posterior distribution

$$\pi(x) = p(x|y) \propto p(x)p(y|x) = p(x) \int p(y,z|x)dz$$

where the integral cannot be computed analytically.

- Then with $z_i \overset{\text{iid}}{\sim} Q_x$ and $p(y,z|x)/Q_x(z)$ well defined, consider an IS approximation of the likelihood

$$\frac{1}{N} \sum_{i=1}^{N} \frac{p(y,z_i|x)}{Q_x(z_i)}$$

This is a noisy measurement of the intractable "likelihood" $p(y|x)$.

- One gets a noisy measurement (up to a constant) of the posterior distribution with

$$\hat{\pi}^N(x) \propto p(x)\left[\int p(y,z|x)dz\right] \times \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{p(y,z_i|x)}{Q_x(z_i)}}{\int p(y,z|x)dz}$$
$$\propto \pi(x) \times w$$

Latent variables and pseudo-marginals

- Assume interest is in a posterior distribution

$$\pi(x) = p(x|y) \propto p(x)p(y|x) = p(x)\int p(y,z|x)dz$$

  where the integral cannot be computed analytically.
- Then with $z_i \overset{\text{iid}}{\sim} Q_x$ and $p(y,z|x)/Q_x(z)$ well defined, consider an IS approximation of the likelihood

$$\frac{1}{N}\sum_{i=1}^{N}\frac{p(y,z_i|x)}{Q_x(z_i)}$$

  This is a noisy measurement of the intractable "likelihood" $p(y|x)$.
- One gets a noisy measurement (up to a constant) of the posterior distribution with

$$\hat{\pi}^N(x) \propto p(x)\left[\int p(y,z|x)dz\right] \times \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{p(y,z_i|x)}{Q_x(z_i)}}{\int p(y,z|x)dz}$$

$$\propto \pi(x) \times w$$

Modelling of the noisy measurements

- Measurements of the form $\pi(x) \times w$ where
  - $w \sim Q_x$, $w \geq 0$, can be thought of as a multiplicative noise,
  - and $\mathbb{E}_{Q_x}[w] = 1$.
- This covers numerous cases of interest
  - latent variable setups,
  - model selection,
  - statistical inference in diffusion models,
  - optimal design,
  - fixed parameter estimation in dynamical systems with particle filters...
  - Bayesian inference/ML estimation when the normalising constant of the likelihood is unknown...
  - Approximate Bayesian Computation (ABC methods).

Modelling of the noisy measurements

- Measurements of the form $\pi(x) \times w$ where
  - $w \sim Q_x$, $w \geq 0$, can be thought of as a multiplicative noise,
  - and $\mathbb{E}_{Q_x}[w] = 1$.
- This covers numerous cases of interest
  - latent variable setups,
  - model selection,
  - statistical inference in diffusion models,
  - optimal design,
  - fixed parameter estimation in dynamical systems with particle filters...
  - Bayesian inference/ML estimation when the normalising constant of the likelihood is unknown...
  - Approximate Bayesian Computation (ABC methods).

Noisy measurements and MCMC

- Unbiased measurements $\pi(x) \times w$ where $w \sim Q_x$, $w \geq 0$ and $\mathbb{E}_{Q_x}[w] = 1$.
- What a standard MH algorithm P would do. Given x, $y \sim q(x, \cdot)$ and use

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} = \min\{1, r(x, y)\}$$

  to accept/reject the transition.
- Naive idea: such measurements could be directly plugged into the standard MH algorithm.
- One could suggest to use the following "noisy" MH algorithm, $\tilde{P}$: $y \sim q(x, \cdot)$, obtain a measurement $\pi(y)u$ of $\pi(y)$ and evaluate

$$\tilde{\alpha}(x, y) = \min\left\{1, \frac{\pi(y) \times u \; q(y, x)}{\pi(x) \times w \; q(x, y)}\right\} = \min\left\{1, r(x, y)\frac{u}{w}\right\}$$

- $\tilde{P}$ approximates P and targets $\pi(x)Q_x(w) \times w \Longrightarrow$ "exact approximation".

Noisy measurements and MCMC

- Unbiased measurements $\pi(x) \times w$ where $w \sim Q_x$, $w \geq 0$ and $\mathbb{E}_{Q_x}[w] = 1$.
- What a standard MH algorithm P would do. Given x, $y \sim q(x, \cdot)$ and use

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} = \min\{1, r(x, y)\}$$

to accept/reject the transition.

- Naive idea: such measurements could be directly plugged into the standard MH algorithm.
- One could suggest to use the following "noisy" MH algorithm, $\tilde{P}$: $y \sim q(x, \cdot)$, obtain a measurement $\pi(y)u$ of $\pi(y)$ and evaluate

$$\tilde{\alpha}(x, y) = \min\left\{1, \frac{\pi(y) \times u\ q(y, x)}{\pi(x) \times w\ q(x, y)}\right\} = \min\left\{1, r(x, y)\frac{u}{w}\right\}$$

- $\tilde{P}$ approximates P and targets $\pi(x)Q_x(w) \times w \implies$ "exact approximation".

Noisy measurements and MCMC

- Unbiased measurements $\pi(x) \times w$ where $w \sim Q_x$, $w \geq 0$ and $\mathbb{E}_{Q_x}[w] = 1$.
- What a standard MH algorithm P would do. Given x, $y \sim q(x, \cdot)$ and use

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} = \min\{1, r(x, y)\}$$

  to accept/reject the transition.
- Naive idea: such measurements could be directly plugged into the standard MH algorithm.
- One could suggest to use the following "noisy" MH algorithm, $\tilde{P}$: $y \sim q(x, \cdot)$, obtain a measurement $\pi(y)u$ of $\pi(y)$ and evaluate

$$\tilde{\alpha}(x, y) = \min\left\{1, \frac{\pi(y) \times u \; q(y, x)}{\pi(x) \times w \; q(x, y)}\right\} = \min\left\{1, r(x, y)\frac{u}{w}\right\}$$

- $\tilde{P}$ approximates P and targets $\pi(x)Q_x(w) \times w \Longrightarrow$ "exact approximation".

Noisy measurements and MCMC

- Unbiased measurements $\pi(x) \times w$ where $w \sim Q_x$, $w \geq 0$ and $\mathbb{E}_{Q_x}[w] = 1$.
- What a standard MH algorithm P would do. Given x, $y \sim q(x, \cdot)$ and use

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\} = \min\{1, r(x,y)\}$$

  to accept/reject the transition.
- Naive idea: such measurements could be directly plugged into the standard MH algorithm.
- One could suggest to use the following "noisy" MH algorithm, $\tilde{P}$: $y \sim q(x, \cdot)$, obtain a measurement $\pi(y)u$ of $\pi(y)$ and evaluate

$$\tilde{\alpha}(x,y) = \min\left\{1, \frac{\pi(y) \times u \;\, q(y,x)}{\pi(x) \times w \;\, q(x,y)}\right\} = \min\left\{1, r(x,y)\frac{u}{w}\right\}$$

- $\tilde{P}$ approximates P and targets $\pi(x)Q_x(w) \times w \implies$ "exact approximation".

Noisy measurements and MCMC

- Unbiased measurements $\pi(x) \times w$ where $w \sim Q_x$, $w \geq 0$ and $\mathbb{E}_{Q_x}[w] = 1$.
- What a standard MH algorithm P would do. Given x, $y \sim q(x, \cdot)$ and use
$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} = \min\{1, r(x, y)\}$$
to accept/reject the transition.
- Naive idea: such measurements could be directly plugged into the standard MH algorithm.
- One could suggest to use the following "noisy" MH algorithm, $\tilde{P}$: $y \sim q(x, \cdot)$, obtain a measurement $\pi(y)u$ of $\pi(y)$ and evaluate
$$\tilde{\alpha}(x, y) = \min\left\{1, \frac{\pi(y) \times u \; q(y, x)}{\pi(x) \times w \; q(x, y)}\right\} = \min\left\{1, r(x, y)\frac{u}{w}\right\}$$
- $\tilde{P}$ approximates P and targets $\pi(x)Q_x(w) \times w \implies$ "exact approximation".

Exactness

- Consider the probability density

$$\pi(x, w) = \pi(x) \times w \times Q_x(w)$$

- From the assumed unbiasedness ($\mathbb{E}_{Q_x}[w] = 1$) its marginal is $\pi(x)$.
- Now consider a MH algorithm targeting this density and proposal distribution

$$q(x, y) \times Q_y(u) \ .$$

- The acceptance probability is

$$\begin{aligned}
\tilde{\alpha}(x, w; y, u) &= \min\left\{1, \frac{\pi(y) \times u \times Q_y(u)}{\pi(x) \times w \times Q_x(w)} \frac{q(y, x) Q_x(w)}{q(x, y) Q_y(u)}\right\} \\
&= \min\left\{1, \frac{\pi(y) \times u \ q(y, x)}{\pi(x) \times w \ q(x, y)}\right\} \ .
\end{aligned}$$

- This is the naive algorithm suggested earlier!

Exactness

- Consider the probability density

$$\pi(x, w) = \pi(x) \times w \times Q_x(w)$$

- From the assumed unbiasedness ($\mathbb{E}_{Q_x}[w] = 1$) its marginal is $\pi(x)$.
- Now consider a MH algorithm targeting this density and proposal distribution

$$q(x, y) \times Q_y(u) \; .$$

- The acceptance probability is

$$\begin{aligned}\tilde{\alpha}(x, w; y, u) &= \min\left\{1, \frac{\pi(y) \times u \times Q_y(u)}{\pi(x) \times w \times Q_x(w)} \frac{q(y, x) Q_x(w)}{q(x, y) Q_y(u)}\right\} \\ &= \min\left\{1, \frac{\pi(y) \times u \; q(y, x)}{\pi(x) \times w \; q(x, y)}\right\} \; .\end{aligned}$$

- This is the naive algorithm suggested earlier!

Exactness

- Consider the probability density

$$\pi(x, w) = \pi(x) \times w \times Q_x(w)$$

- From the assumed unbiasedness ($\mathbb{E}_{Q_x}[w] = 1$) its marginal is $\pi(x)$.
- Now consider a MH algorithm targeting this density and proposal distribution

$$q(x, y) \times Q_y(u) \ .$$

- The acceptance probability is

$$\begin{aligned} \tilde{\alpha}(x, w; y, u) &= \min\left\{1, \frac{\pi(y) \times u \times Q_y(u)}{\pi(x) \times w \times Q_x(w)} \frac{q(y, x) Q_x(w)}{q(x, y) Q_y(u)}\right\} \\ &= \min\left\{1, \frac{\pi(y) \times u \ q(y, x)}{\pi(x) \times w \ q(x, y)}\right\} \ . \end{aligned}$$

- This is the naive algorithm suggested earlier!

Exactness

- Consider the probability density

$$\pi(x, w) = \pi(x) \times w \times Q_x(w)$$

- From the assumed unbiasedness ($\mathbb{E}_{Q_x}[w] = 1$) its marginal is $\pi(x)$.
- Now consider a MH algorithm targeting this density and proposal distribution

$$q(x, y) \times Q_y(u) \ .$$

- The acceptance probability is

$$\begin{aligned}
\tilde{\alpha}(x, w; y, u) &= \min\left\{1, \frac{\pi(y) \times u \times Q_y(u)}{\pi(x) \times w \times Q_x(w)} \frac{q(y, x) Q_x(w)}{q(x, y) Q_y(u)}\right\} \\
&= \min\left\{1, \frac{\pi(y) \times u \ q(y, x)}{\pi(x) \times w \ q(x, y)}\right\} \ .
\end{aligned}$$

- This is the naive algorithm suggested earlier!

Exactness

- Consider the probability density

$$\pi(x, w) = \pi(x) \times w \times Q_x(w)$$

- From the assumed unbiasedness ($\mathbb{E}_{Q_x}[w] = 1$) its marginal is $\pi(x)$.
- Now consider a MH algorithm targeting this density and proposal distribution

$$q(x, y) \times Q_y(u) \ .$$

- The acceptance probability is

$$\begin{aligned}\tilde{\alpha}(x, w; y, u) &= \min\left\{1, \frac{\pi(y) \times u \times Q_y(u)}{\pi(x) \times w \times Q_x(w)} \frac{q(y, x) Q_x(w)}{q(x, y) Q_y(u)}\right\} \\ &= \min\left\{1, \frac{\pi(y) \times u \ q(y, x)}{\pi(x) \times w \ q(x, y)}\right\} \ .\end{aligned}$$

- This is the naive algorithm suggested earlier!

Exact approximation

- $\tilde{P}$ approximates P.
- the more w is concentrated on 1 the better the approximation looks,
- for example if for $x \in \mathsf{X}$ we have N (say independent) noisy measurements of $\pi(x)w_i$ then one could use the following (better) estimator

$$\pi(x)\frac{1}{N}\sum_{i=1}^{N} w_i$$

- Question: is it important to average, or more generally use better approximations of the algorithm we cannot implement?

Exact approximation

- $\tilde{P}$ approximates P.
- the more w is concentrated on 1 the better the approximation looks,
- for example if for $x \in \mathsf{X}$ we have N (say independent) noisy measurements of $\pi(x)w_i$ then one could use the following (better) estimator
$$\pi(x)\frac{1}{N}\sum_{i=1}^{N} w_i$$
- Question: is it important to average, or more generally use better approximations of the algorithm we cannot implement?

Exact approximation

- $\tilde{P}$ approximates P.
- the more w is concentrated on 1 the better the approximation looks,
- for example if for $x \in \mathsf{X}$ we have N (say independent) noisy measurements of $\pi(x)w_i$ then one could use the following (better) estimator

$$\pi(x)\frac{1}{N}\sum_{i=1}^{N} w_i$$

- Question: is it important to average, or more generally use better approximations of the algorithm we cannot implement?

Exact approximation

- $\tilde{P}$ approximates P.
- the more w is concentrated on 1 the better the approximation looks,
- for example if for $x \in \mathsf{X}$ we have N (say independent) noisy measurements of $\pi(x)w_i$ then one could use the following (better) estimator

$$\pi(x)\frac{1}{N}\sum_{i=1}^{N} w_i$$

- Question: is it important to average, or more generally use better approximations of the algorithm we cannot implement?

Toy latent variables example

- We consider here a simple example where the target distribution is

$$\pi(x,z) = \mathcal{N}\left(\left(\begin{array}{c} x \\ z \end{array}\right); \left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left[\begin{array}{cc} 1 & -0.9 \\ -0.9 & 1 \end{array}\right]\right)$$

- Marginal is $\pi(x) = \mathcal{N}(x; 0, 1)$
- Sample with random walk Metropolis algorithm
  - with $q(x,y) = \mathcal{N}(y; x, 2.4^2)$ and $Q_x(Z) = \prod_{i=1}^{N} \mathcal{N}(z_i; 0, 1)$ for IS.
  - $q(x,y) = \mathcal{N}(y; x, 2.4^2)$ is known to be optimal in terms of asymptotic variance.

Toy latent variables example

- We consider here a simple example where the target distribution is

$$\pi(x,z) = \mathcal{N}\left(\left(\begin{array}{c} x \\ z \end{array}\right); \left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left[\begin{array}{cc} 1 & -0.9 \\ -0.9 & 1 \end{array}\right]\right)$$

- Marginal is $\pi(x) = \mathcal{N}(x; 0, 1)$
- Sample with random walk Metropolis algorithm
  - with $q(x,y) = \mathcal{N}(y; x, 2.4^2)$ and $Q_x(Z) = \prod_{i=1}^{N} \mathcal{N}(z_i; 0, 1)$ for IS.
  - $q(x,y) = \mathcal{N}(y; x, 2.4^2)$ is known to be optimal in terms of asymptotic variance.

Toy latent variables example

- We consider here a simple example where the target distribution is

$$\pi(x,z) = \mathcal{N}\left(\left(\begin{array}{c} x \\ z \end{array}\right);\left(\begin{array}{c} 0 \\ 0 \end{array}\right),\left[\begin{array}{cc} 1 & -0.9 \\ -0.9 & 1 \end{array}\right]\right)$$

- Marginal is $\pi(x) = \mathcal{N}(x; 0, 1)$
- Sample with random walk Metropolis algorithm
  - with $q(x,y) = \mathcal{N}(y; x, 2.4^2)$ and $Q_x(Z) = \prod_{i=1}^{N} \mathcal{N}(z_i; 0, 1)$ for IS.
  - $q(x,y) = \mathcal{N}(y; x, 2.4^2)$ is known to be optimal in terms of asymptotic variance.

Toy latent variables example

- We consider here a simple example where the target distribution is

$$\pi(x, z) = \mathcal{N}\left(\left(\begin{array}{c} x \\ z \end{array}\right); \left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left[\begin{array}{cc} 1 & -0.9 \\ -0.9 & 1 \end{array}\right]\right)$$

- Marginal is $\pi(x) = \mathcal{N}(x; 0, 1)$
- Sample with random walk Metropolis algorithm
  - with $q(x, y) = \mathcal{N}\left(y; x, 2.4^2\right)$ and $Q_x(Z) = \prod_{i=1}^N \mathcal{N}(z_i; 0, 1)$ for IS.
  - $q(x, y) = \mathcal{N}\left(y; x, 2.4^2\right)$ is known to be optimal in terms of asymptotic variance.

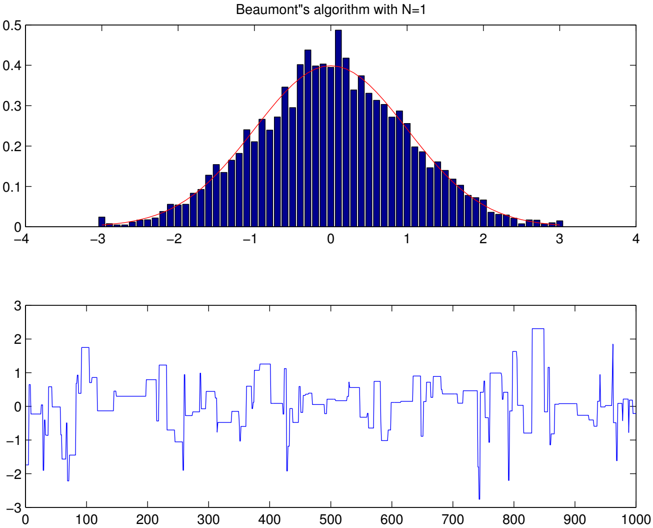
Toy latent variables example

- We consider here a simple example where the target distribution is

$$\pi(x,z) = \mathcal{N}\left(\left(\begin{array}{c} x \\ z \end{array}\right); \left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left[\begin{array}{cc} 1 & -0.9 \\ -0.9 & 1 \end{array}\right]\right)$$

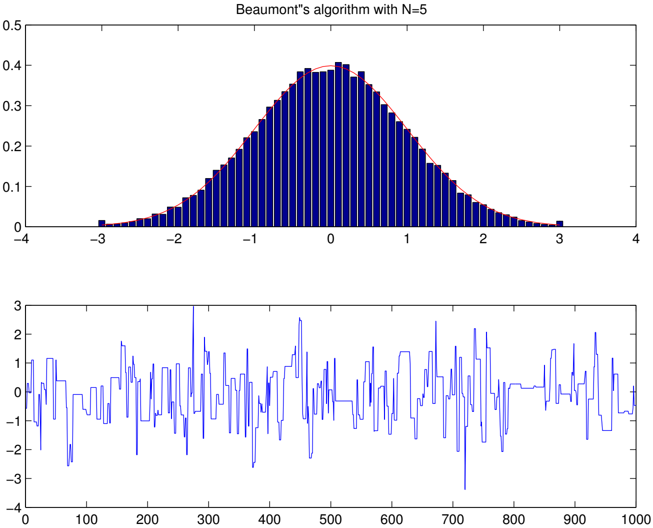
- Marginal is $\pi(x) = \mathcal{N}(x; 0, 1)$
- Sample with random walk Metropolis algorithm
  - with $q(x,y) = \mathcal{N}\left(y; x, 2.4^2\right)$ and $Q_x(Z) = \prod_{i=1}^{N} \mathcal{N}(z_i; 0, 1)$ for IS.
  - $q(x,y) = \mathcal{N}\left(y; x, 2.4^2\right)$ is known to be optimal in terms of asymptotic variance.

Standard AV



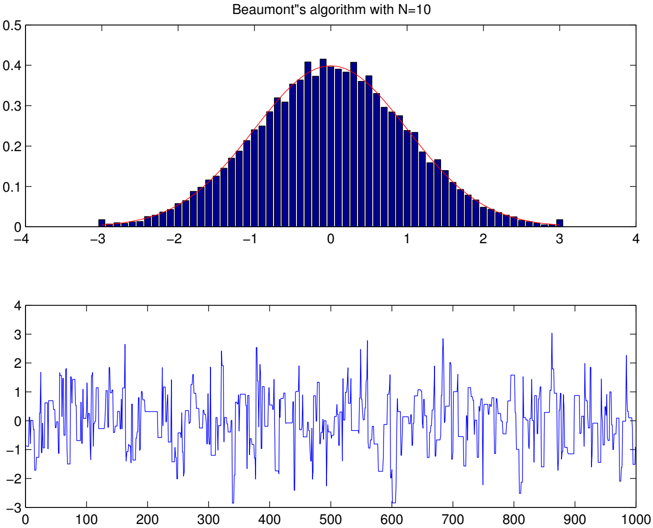
Beaumont"s algorithm with N=1
0.5
0.4
0.3
0.2
0.1
0
−4
−3
−2
−1
0
1
2
3
4
3
2
1
0
−1
−2
−3
0
100
200
300
400
500
600
700
800
900
1000

$N = 5$



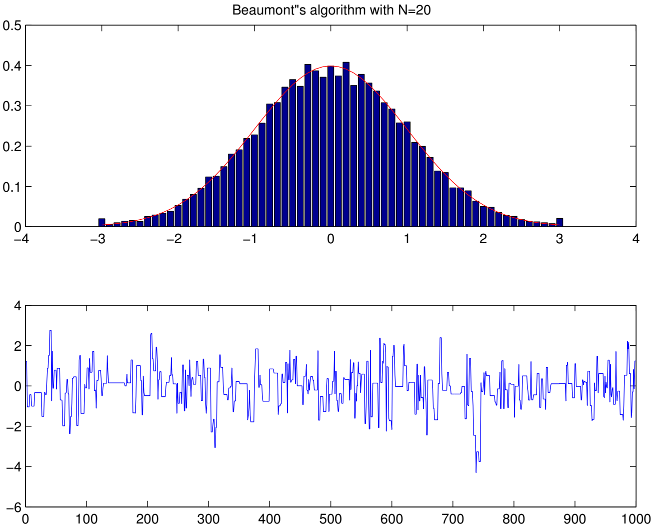
Beaumont"s algorithm with N=5
0.5
0.4
0.3
0.2
0.1
0
−4
−3
−2
−1
0
1
2
3
4
3
2
1
0
−1
−2
−3
−4
0
100
200
300
400
500
600
700
800
900
1000

$N = 10$



Beaumont"s algorithm with N=10
0.5
0.4
0.3
0.2
0.1
0
−4
−3
−2
−1
0
1
2
3
4
4
3
2
1
0
−1
−2
−3
0
100
200
300
400
500
600
700
800
900
1000

$N = 20$



Intuition

- The acceptance probability of the algorithm is

$$\min\left\{1, r(x, y)\frac{u}{w}\right\}$$

- The probability of escaping (x, w) can be made arbitrarily small by increasing w...
- The Markov chain becomes "sticky".

Intuition

- The acceptance probability of the algorithm is

$$\min\left\{1, r(x,y)\frac{u}{w}\right\}$$

- The probability of escaping (x, w) can be made arbitrarily small by increasing w...
- The Markov chain becomes "sticky".

Intuition

- The acceptance probability of the algorithm is

$$\min\left\{1, r(x,y)\frac{u}{w}\right\}$$

- The probability of escaping (x, w) can be made arbitrarily small by increasing w...
- The Markov chain becomes "sticky".

Asymptotic variance and expected acceptance probability

- With Π a Markov transition kernel with invariant distribution μ, letting $X_1 \sim \mu$ and $X_n \sim \Pi(X_{n-1}, \cdot)$,

$$\mathrm{var}\,(f, \Pi) := \lim_{T\to\infty} T\mathbb{E}\left(\frac{1}{T}\sum_{k=1}^{T} f(X_k) - \mu(f)\right)^2 \in [0, \infty].$$

- The expected acceptance probability of a MH algorithm with invariant distribution π is

$$\int \alpha(x, y)\pi(dx)q(x, dy)$$

Asymptotic variance and expected acceptance probability

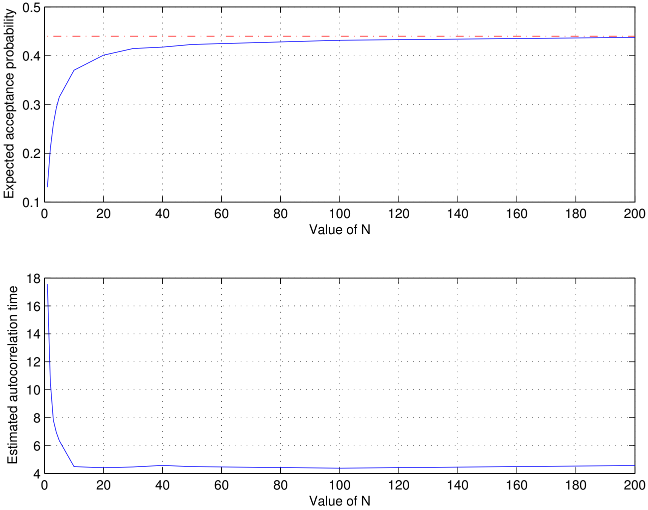
- With Π a Markov transition kernel with invariant distribution μ, letting $X_1 \sim \mu$ and $X_n \sim \Pi(X_{n-1}, \cdot)$,

$$\mathrm{var}\,(f, \Pi) := \lim_{T\to\infty} T\mathbb{E}\left(\frac{1}{T}\sum_{k=1}^{T} f(X_k) - \mu(f)\right)^2 \in [0, \infty].$$

- The expected acceptance probability of a MH algorithm with invariant distribution π is

$$\int \alpha(x, y)\pi(dx)q(x, dy)$$

Performance as a function of N



Comparing pseudo-marginal algorithms?

- A natural question is whether the performance of the algorithm indeed always improves as we increase N?
- Our work is concerned with developing tools for the comparison of the performance of pseudo-marginal algorithms in terms of the choice of Q_x.
- Let $\left\{Q_x^{(1)}\right\}$ and $\left\{Q_x^{(2)}\right\}$ be two families of distributions corresponding to two possible approximations of the marginal density.
- Let $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ be the corresponding competing pseudo-marginal implementations of the MH algorithm
  - targeting $\pi(\cdot)$ marginally
  - sharing the same family of proposal distributions $\{q(x,\cdot), x \in \mathsf{X}\}$.

Comparing pseudo-marginal algorithms?

- A natural question is whether the performance of the algorithm indeed always improves as we increase N?
- Our work is concerned with developing tools for the comparison of the performance of pseudo-marginal algorithms in terms of the choice of Q_x.
- Let $\left\{Q_x^{(1)}\right\}$ and $\left\{Q_x^{(2)}\right\}$ be two families of distributions corresponding to two possible approximations of the marginal density.
- Let $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ be the corresponding competing pseudo-marginal implementations of the MH algorithm
    - targeting $\pi(\cdot)$ marginally
    - sharing the same family of proposal distributions $\{q(x, \cdot), x \in \mathsf{X}\}$.

Comparing pseudo-marginal algorithms?

- A natural question is whether the performance of the algorithm indeed always improves as we increase N?
- Our work is concerned with developing tools for the comparison of the performance of pseudo-marginal algorithms in terms of the choice of Q_x.
- Let $\left\{Q_x^{(1)}\right\}$ and $\left\{Q_x^{(2)}\right\}$ be two families of distributions corresponding to two possible approximations of the marginal density.
- Let $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ be the corresponding competing pseudo-marginal implementations of the MH algorithm
  - targeting $\pi(\cdot)$ marginally
  - sharing the same family of proposal distributions $\{q(x, \cdot), x \in \mathsf{X}\}$.

Comparing pseudo-marginal algorithms?

- A natural question is whether the performance of the algorithm indeed always improves as we increase N?
- Our work is concerned with developing tools for the comparison of the performance of pseudo-marginal algorithms in terms of the choice of Q_x.
- Let $\{Q_x^{(1)}\}$ and $\{Q_x^{(2)}\}$ be two families of distributions corresponding to two possible approximations of the marginal density.
- Let $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ be the corresponding competing pseudo-marginal implementations of the MH algorithm
  - targeting $\pi(\cdot)$ marginally
  - sharing the same family of proposal distributions $\{q(x,\cdot), x \in \mathsf{X}\}$.

Comparing pseudo-marginal algorithms?

- A natural question is whether the performance of the algorithm indeed always improves as we increase N?
- Our work is concerned with developing tools for the comparison of the performance of pseudo-marginal algorithms in terms of the choice of Q_x.
- Let $\left\{Q_x^{(1)}\right\}$ and $\left\{Q_x^{(2)}\right\}$ be two families of distributions corresponding to two possible approximations of the marginal density.
- Let $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ be the corresponding competing pseudo-marginal implementations of the MH algorithm
  - targeting $\pi(\cdot)$ marginally
  - sharing the same family of proposal distributions $\{q(x,\cdot), x \in \mathsf{X}\}$.

Comparing pseudo-marginal algorithms?

- A natural question is whether the performance of the algorithm indeed always improves as we increase N?
- Our work is concerned with developing tools for the comparison of the performance of pseudo-marginal algorithms in terms of the choice of Q_x.
- Let $\left\{Q_x^{(1)}\right\}$ and $\left\{Q_x^{(2)}\right\}$ be two families of distributions corresponding to two possible approximations of the marginal density.
- Let $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ be the corresponding competing pseudo-marginal implementations of the MH algorithm
    - targeting $\pi(\cdot)$ marginally
    - sharing the same family of proposal distributions $\{q(x,\cdot), x \in \mathsf{X}\}$.

Comparison of pseudo-marginal algorithms

- The transition probabilities are, for $i \in \{1, 2\}$,

$$\tilde{P}^{(i)}(x, w; \mathrm{d}y \times \mathrm{d}u) := q(x, \mathrm{d}y) Q_y^{(i)}(\mathrm{d}u) \min\left\{1, r(x, y)\frac{u}{w}\right\} + \delta_{x,w}(\mathrm{d}y \times \mathrm{d}u)\tilde{\rho}^{(i)}(x, w)$$

- They target different distributions, $\tilde{\pi}^{(i)}(\mathrm{d}x \times \mathrm{d}w) = \pi(\mathrm{d}x) Q_x^{(i)}(\mathrm{d}w) w$,
- The natural question we are interested in is to find a useful characterization of $\left\{Q_x^{(1)}\right\}$ and $\left\{Q_x^{(2)}\right\}$ which implies that for $f : \mathsf{X} \to \mathbb{R}$,

$$\mathrm{var}(f, \tilde{P}^{(1)}) \leq \mathrm{var}(f, \tilde{P}^{(2)}) \text{ or } \mathrm{Gap}_R(\tilde{P}^{(1)}) \leq \mathrm{Gap}_R(\tilde{P}^{(2)}) \quad .$$

Comparison of pseudo-marginal algorithms

- The transition probabilities are, for $i \in \{1, 2\}$,

$$\tilde{P}^{(i)}(x, w; \mathrm{d}y \times \mathrm{d}u) := q(x, \mathrm{d}y) Q_y^{(i)}(\mathrm{d}u) \min\left\{1, r(x, y)\frac{u}{w}\right\} + \delta_{x,w}(\mathrm{d}y \times \mathrm{d}u)\tilde{\rho}^{(i)}(x, w)$$

- They target different distributions, $\tilde{\pi}^{(i)}\left(\mathrm{d}x \times \mathrm{d}w\right) = \pi(\mathrm{d}x) Q_x^{(i)}(\mathrm{d}w) w$,
- The natural question we are interested in is to find a useful characterization of $\left\{Q_x^{(1)}\right\}$ and $\left\{Q_x^{(2)}\right\}$ which implies that for $f : \mathsf{X} \to \mathbb{R}$,

$$\mathrm{var}(f, \tilde{P}^{(1)}) \leq \mathrm{var}(f, \tilde{P}^{(2)}) \text{ or } \mathrm{Gap}_R\left(\tilde{P}^{(1)}\right) \leq \mathrm{Gap}_R\left(\tilde{P}^{(2)}\right) \quad .$$

Comparison of pseudo-marginal algorithms

- The transition probabilities are, for $i \in \{1,2\}$,

$$\tilde{P}^{(i)}(x, w; \mathrm{d}y \times \mathrm{d}u) := q(x, \mathrm{d}y) Q_y^{(i)}(\mathrm{d}u) \min\left\{1, r(x,y)\frac{u}{w}\right\} + \delta_{x,w}(\mathrm{d}y \times \mathrm{d}u)\tilde{\rho}^{(i)}(x, w)$$

- They target different distributions, $\tilde{\pi}^{(i)}(\mathrm{d}x \times \mathrm{d}w) = \pi(\mathrm{d}x) Q_x^{(i)}(\mathrm{d}w) w$,
- The natural question we are interested in is to find a useful characterization of $\left\{Q_x^{(1)}\right\}$ and $\left\{Q_x^{(2)}\right\}$ which implies that for $f : \mathsf{X} \to \mathbb{R}$,

$$\mathrm{var}(f, \tilde{P}^{(1)}) \leq \mathrm{var}(f, \tilde{P}^{(2)}) \text{ or } \mathrm{Gap}_R\left(\tilde{P}^{(1)}\right) \leq \mathrm{Gap}_R\left(\tilde{P}^{(2)}\right) \quad .$$

Standard ordering of MH algorithms—Peskun

- Let $\Pi^{(1)}$ and $\Pi^{(2)}$ be two Markov kernel reversible with respect to some common invariant distribution μ on $\left(\mathsf{E}, \mathcal{B}(\mathsf{E})\right)$.
- A well known result due originally to Peskun states that

Theorem (Peskun)

Whenever for any $x \in \mathsf{E}$ and $A \in \mathcal{B}(\mathsf{E})$ such that $x \notin A$, $\Pi^{(1)}(x, A) \geq \Pi^{(2)}(x, A)$ then for any $f : \mathsf{E} \to \mathbb{R}$ such that $\mathrm{var}_\mu(f) < \infty$ then

$$\mathrm{var}(f, \Pi^{(1)}) \leq \mathrm{var}(f, \Pi^{(2)}) \text{ and } \mathrm{Gap}_R(\Pi^{(1)}) \geq \mathrm{Gap}_R(\Pi^{(2)})$$

- therefore leading to a simple and intuitive criterion for the comparison of performance of algorithms.
- Peskun's result is not an "iff" statement (more later), but it is practically useful.
- Clearly Peskun's result does not apply to the comparison of pseudo-marginal algorithms since $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ do not share the same invariant distribution.

Standard ordering of MH algorithms—Peskun

- Let $\Pi^{(1)}$ and $\Pi^{(2)}$ be two Markov kernel reversible with respect to some common invariant distribution μ on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$.
- A well known result due originally to Peskun states that

Theorem (Peskun)

Whenever for any $x \in \mathsf{E}$ and $A \in \mathcal{B}(\mathsf{E})$ such that $x \notin A$, $\Pi^{(1)}(x, A) \geq \Pi^{(2)}(x, A)$ then for any $f : \mathsf{E} \to \mathbb{R}$ such that $\mathrm{var}_\mu(f) < \infty$ then

$$\mathrm{var}(f, \Pi^{(1)}) \leq \mathrm{var}(f, \Pi^{(2)}) \text{ and } \mathrm{Gap}_R(\Pi^{(1)}) \geq \mathrm{Gap}_R(\Pi^{(2)})$$

- therefore leading to a simple and intuitive criterion for the comparison of performance of algorithms.
- Peskun's result is not an "iff" statement (more later), but it is practically useful.
- Clearly Peskun's result does not apply to the comparison of pseudo-marginal algorithms since $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ do not share the same invariant distribution.

Standard ordering of MH algorithms—Peskun

- Let $\Pi^{(1)}$ and $\Pi^{(2)}$ be two Markov kernel reversible with respect to some common invariant distribution μ on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$.
- A well known result due originally to Peskun states that

Theorem (Peskun)

Whenever for any $x \in \mathsf{E}$ and $A \in \mathcal{B}(\mathsf{E})$ such that $x \notin A$, $\Pi^{(1)}(x, A) \geq \Pi^{(2)}(x, A)$ then for any $f : \mathsf{E} \to \mathbb{R}$ such that $\mathrm{var}_\mu(f) < \infty$ then

$$\mathrm{var}(f, \Pi^{(1)}) \leq \mathrm{var}(f, \Pi^{(2)}) \text{ and } \mathrm{Gap}_R(\Pi^{(1)}) \geq \mathrm{Gap}_R(\Pi^{(2)})$$

- therefore leading to a simple and intuitive criterion for the comparison of performance of algorithms.
- Peskun's result is not an "iff" statement (more later), but it is practically useful.
- Clearly Peskun's result does not apply to the comparison of pseudo-marginal algorithms since $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ do not share the same invariant distribution.

Standard ordering of MH algorithms—Peskun

- Let $\Pi^{(1)}$ and $\Pi^{(2)}$ be two Markov kernel reversible with respect to some common invariant distribution μ on $\left(\mathsf{E}, \mathcal{B}(\mathsf{E})\right)$.
- A well known result due originally to Peskun states that

Theorem (Peskun)

Whenever for any $x \in \mathsf{E}$ *and* $A \in \mathcal{B}(\mathsf{E})$ *such that* $x \notin A$, $\Pi^{(1)}(x, A) \geq \Pi^{(2)}(x, A)$ *then for any* $f : \mathsf{E} \to \mathbb{R}$ *such that* $\mathrm{var}_\mu(f) < \infty$ *then*

$$\mathrm{var}(f, \Pi^{(1)}) \leq \mathrm{var}(f, \Pi^{(2)}) \text{ and } \mathrm{Gap}_R\left(\Pi^{(1)}\right) \geq \mathrm{Gap}_R\left(\Pi^{(2)}\right)$$

- therefore leading to a simple and intuitive criterion for the comparison of performance of algorithms.
- Peskun's result is not an "iff" statement (more later), but it is practically useful.
- Clearly Peskun's result does not apply to the comparison of pseudo-marginal algorithms since $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ do not share the same invariant distribution.

Standard ordering of MH algorithms—Peskun

- Let $\Pi^{(1)}$ and $\Pi^{(2)}$ be two Markov kernel reversible with respect to some common invariant distribution μ on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$.
- A well known result due originally to Peskun states that

Theorem (Peskun)

Whenever for any $x \in \mathsf{E}$ and $A \in \mathcal{B}(\mathsf{E})$ such that $x \notin A$, $\Pi^{(1)}(x, A) \geq \Pi^{(2)}(x, A)$ then for any $f : \mathsf{E} \to \mathbb{R}$ such that $\mathrm{var}_\mu(f) < \infty$ then

$$\mathrm{var}(f, \Pi^{(1)}) \leq \mathrm{var}(f, \Pi^{(2)}) \text{ and } \mathrm{Gap}_R(\Pi^{(1)}) \geq \mathrm{Gap}_R(\Pi^{(2)})$$

- therefore leading to a simple and intuitive criterion for the comparison of performance of algorithms.
- Peskun's result is not an "iff" statement (more later), but it is practically useful.
- Clearly Peskun's result does not apply to the comparison of pseudo-marginal algorithms since $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ do not share the same invariant distribution.

An order for variability

- Intuitively performance of pseudo-marginal algorithms should depend on the variability of the approximation.
- Considering the variance is not sufficient : one can construct counterexamples where $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ but $\mathrm{var}(f, \tilde{P}^{(1)}) \geq \mathrm{var}(f, \tilde{P}^{(2)})$ [CA & Vihola, 2015].
- The convex order is a natural way to compare the "variability" or "dispersion" of two random variables or distributions.

Definition

The random variables W_1 and W_2 are *convex ordered* $W_1 \leq_{cx} W_2$ if for any convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(W_1)] \leq \mathbb{E}[\phi(W_2)],$$

whenever the expectations are well-defined.

- Note that $W_1 \leq_{cx} W_2$ implies $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ i.e. an observed order in terms of variance will be a by-product of the convex order.

An order for variability

- Intuitively performance of pseudo-marginal algorithms should depend on the variability of the approximation.
- Considering the variance is not sufficient : one can construct counterexamples where $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ but $\mathrm{var}(f, \tilde{P}^{(1)}) \geq \mathrm{var}(f, \tilde{P}^{(2)})$ [CA & Vihola, 2015].
- The convex order is a natural way to compare the "variability" or "dispersion" of two random variables or distributions.

Definition

The random variables W_1 and W_2 are *convex ordered* $W_1 \leq_{cx} W_2$ if for any convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(W_1)] \leq \mathbb{E}[\phi(W_2)],$$

whenever the expectations are well-defined.

- Note that $W_1 \leq_{cx} W_2$ implies $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ i.e. an observed order in terms of variance will be a by-product of the convex order.

An order for variability

- Intuitively performance of pseudo-marginal algorithms should depend on the variability of the approximation.
- Considering the variance is not sufficient : one can construct counterexamples where $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ but $\mathrm{var}(f, \tilde{P}^{(1)}) \geq \mathrm{var}(f, \tilde{P}^{(2)})$ [CA & Vihola, 2015].
- The convex order is a natural way to compare the "variability" or "dispersion" of two random variables or distributions.

Definition

The random variables W_1 and W_2 are *convex ordered* $W_1 \leq_{cx} W_2$ if for any convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(W_1)] \leq \mathbb{E}[\phi(W_2)],$$

whenever the expectations are well-defined.

- Note that $W_1 \leq_{cx} W_2$ implies $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ i.e. an observed order in terms of variance will be a by-product of the convex order.

An order for variability

- Intuitively performance of pseudo-marginal algorithms should depend on the variability of the approximation.
- Considering the variance is not sufficient : one can construct counterexamples where $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ but $\mathrm{var}(f, \tilde{P}^{(1)}) \geq \mathrm{var}(f, \tilde{P}^{(2)})$ [CA & Vihola, 2015].
- The convex order is a natural way to compare the "variability" or "dispersion" of two random variables or distributions.

Definition

The random variables W_1 and W_2 are *convex ordered* $W_1 \leq_{cx} W_2$ if for any convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(W_1)] \leq \mathbb{E}[\phi(W_2)],$$

whenever the expectations are well-defined.

- Note that $W_1 \leq_{cx} W_2$ implies $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ i.e. an observed order in terms of variance will be a by-product of the convex order.

An order for variability

- Intuitively performance of pseudo-marginal algorithms should depend on the variability of the approximation.
- Considering the variance is not sufficient : one can construct counterexamples where $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ but $\mathrm{var}(f, \tilde{P}^{(1)}) \geq \mathrm{var}(f, \tilde{P}^{(2)})$ [CA & Vihola, 2015].
- The convex order is a natural way to compare the "variability" or "dispersion" of two random variables or distributions.

Definition

The random variables W_1 and W_2 are *convex ordered* $W_1 \leq_{cx} W_2$ if for any convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\phi(W_1)] \leq \mathbb{E}[\phi(W_2)],$$

whenever the expectations are well-defined.

- Note that $W_1 \leq_{cx} W_2$ implies $\mathrm{var}(W_1) \leq \mathrm{var}(W_2)$ i.e. an observed order in terms of variance will be a by-product of the convex order.

Relevance of the convex order?

- An equivalent characterization of the convex order is possible by restricting the subset of convex functions to $t \mapsto -\min\{a, t\}$ for $a \in \mathbb{R}$,
- The algorithm's acceptance ratio is

$$\min\left\{1, r(x,y)\frac{u}{w}\right\}$$

and at a superficial level one may suspect a link...
- Except for a very specific scenario we do not claim that this is the optimal way of ordering algorithms.
- Importantly it allows us to establish practically relevant results.

Relevance of the convex order?

- An equivalent characterization of the convex order is possible by restricting the subset of convex functions to $t \mapsto -\min\{a, t\}$ for $a \in \mathbb{R}$,
- The algorithm's acceptance ratio is

$$\min\left\{1, r(x, y)\frac{u}{w}\right\}$$

and at a superficial level one may suspect a link...

- Except for a very specific scenario we do not claim that this is the optimal way of ordering algorithms.
- Importantly it allows us to establish practically relevant results.

Relevance of the convex order?

- An equivalent characterization of the convex order is possible by restricting the subset of convex functions to $t \mapsto -\min\{a, t\}$ for $a \in \mathbb{R}$,
- The algorithm's acceptance ratio is

$$\min\left\{1, r(x,y)\frac{u}{w}\right\}$$

  and at a superficial level one may suspect a link...
- Except for a very specific scenario we do not claim that this is the optimal way of ordering algorithms.
- Importantly it allows us to establish practically relevant results.

Relevance of the convex order?

- An equivalent characterization of the convex order is possible by restricting the subset of convex functions to $t \mapsto -\min\{a, t\}$ for $a \in \mathbb{R}$,
- The algorithm's acceptance ratio is

  $$\min\left\{1, r(x, y)\frac{u}{w}\right\}$$

  and at a superficial level one may suspect a link...
- Except for a very specific scenario we do not claim that this is the optimal way of ordering algorithms.
- Importantly it allows us to establish practically relevant results.

Main result

Theorem

Let π be a probability distribution on some measurable space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ and $\tilde{P}_1$ and $\tilde{P}_2$ be two implementations of pseudo-marginal algorithms to sample from π sharing the family of proposal distributions $\{q(x, \cdot), x \in \mathsf{X}\}$ but noise distributions $\{Q_x^{(1)}, x \in \mathsf{X}\}$ and $\{Q_x^{(2)}, x \in \mathsf{X}\}$ such that for any $x \in \mathsf{X}$ $W_x^{(1)} \leq_{cx} W_x^{(2)}$. Then for any $f \in L^2(\mathsf{X}, \pi)$ we have the following orders for the

1. *asymptotic variances:* $\mathrm{var}(f, \tilde{P}_2) \geq \mathrm{var}(f, \tilde{P}_1)$,
2. *spectral gaps:* $\mathrm{Gap}_R(\tilde{P}_i) \leq \mathrm{Gap}_R(P)$ *and more...*

Extremal distributions (I)

Theorem

For $\mu, a, b \in \mathbb{R}$ ($a \le \mu \le b$) let $\mathscr{P}(\mu, [a,b])$ be the set of probability distributions Q on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that for $W \sim Q$, $\mathbb{E}_Q\big[W\big] = \mu$ and $Q\big(W \in [a,b]\big) = 1$. Then for any $Q \in \mathscr{P}(\mu, [a,b])$

$$Q^{\mathsf{min}} \le_{cx} Q \le_{cx} Q^{\max}$$

$$\begin{aligned} Q^{\min}(\mathrm{d}w) &:= \delta_\mu(\mathrm{d}w), \\ Q^{\max}(\mathrm{d}w) &:= \frac{b-\mu}{b-a}\delta_a(\mathrm{d}w) + \frac{\mu-a}{b-a}\delta_b(\mathrm{d}w) \quad . \end{aligned}$$

Extremal distributions (II)

Theorem

Let $a_x, b_x : \mathsf{X}^2 \to [0, \infty)$ *(*$a_x \leq 1 \leq b_x$*). Consider the class of pseudo marginal algorithms* $\tilde{P}$ *such that for any* $x \in \mathsf{X}$ *the weight distribution* Q_x *is such that* $Q_x \in \mathscr{P}(1, [a_x, b_x])$*. Then for any* $f \in L^2(\mathsf{X}, \pi)$*,*

$$\mathrm{var}\left(P, f\right) \leq \mathrm{var}\left(\tilde{P}, f\right) \leq \mathrm{var}\left(\tilde{P}_{\max}, f\right) \quad ,$$

where $\tilde{P}_{\max}$ *is the pseudo-marginal algorithm with distribution*

$$Q_x^{\max}(\mathrm{d}w) = \frac{1 - a_x}{b_x - a_x}\delta_{a_x}(\mathrm{d}w) + \frac{b_x - 1}{b_x - a_x}\delta_{b_x}(\mathrm{d}w) \quad .$$

Furthermore

$$\mathrm{var}\left(\tilde{P}_{\max}, f\right) \leq \sup_{x \in \mathsf{X}} b_x \mathrm{var}\left(P, f\right) + (\sup_{x \in \mathsf{X}} b_x - 1)\mathrm{var}_\pi(f) \quad .$$

Every sample counts in pseudo-marginal MCMC

- As mentioned earlier a suggestion in order to improve the performance of such algorithms one can suggest averaging, i.e. use an average of (say independent) estimates of the density

$$\pi(x)W^N := \pi(x)\frac{1}{N}\sum_{i=1}^{N} W_i$$

- Intuitively this should help since we are reducing the variance. But we know that the variance is not necessarily a good indicator (counterexample).
- However... for exchangeable random variables, it is known that for any $N \geq 1$

$$\frac{1}{N+1}\sum_{i=1}^{N+1} W_i \leq_{cx} \frac{1}{N}\sum_{i=1}^{N} W_i$$

- Which from our results immediately implies that for any $f \in L^2(\mathsf{X}, \pi)$ and any $N \geq 2$

$$\mathrm{var}(f, \tilde{P}_{N-1}) \geq \mathrm{var}(f, \tilde{P}_N) \ldots$$

Every sample counts in pseudo-marginal MCMC

- As mentioned earlier a suggestion in order to improve the performance of such algorithms one can suggest averaging, i.e. use an average of (say independent) estimates of the density

$$\pi(x)W^N := \pi(x)\frac{1}{N}\sum_{i=1}^{N} W_i$$

- Intuitively this should help since we are reducing the variance. But we know that the variance is not necessarily a good indicator (counterexample).
- However... for exchangeable random variables, it is known that for any $N \geq 1$

$$\frac{1}{N+1}\sum_{i=1}^{N+1} W_i \leq_{cx} \frac{1}{N}\sum_{i=1}^{N} W_i$$

- Which from our results immediately implies that for any $f \in L^2(\mathsf{X},\pi)$ and any $N \geq 2$

$$\mathrm{var}(f,\tilde{P}_{N-1}) \geq \mathrm{var}(f,\tilde{P}_N)\ldots$$

Every sample counts in pseudo-marginal MCMC

- As mentioned earlier a suggestion in order to improve the performance of such algorithms one can suggest averaging, i.e. use an average of (say independent) estimates of the density

$$\pi(x)W^N := \pi(x)\frac{1}{N}\sum_{i=1}^{N} W_i$$

- Intuitively this should help since we are reducing the variance. But we know that the variance is not necessarily a good indicator (counterexample).
- However... for exchangeable random variables, it is known that for any $N \geq 1$

$$\frac{1}{N+1}\sum_{i=1}^{N+1} W_i \leq_{cx} \frac{1}{N}\sum_{i=1}^{N} W_i$$

- Which from our results immediately implies that for any $f \in L^2(\mathsf{X}, \pi)$ and any $N \geq 2$

$$\mathrm{var}(f, \tilde{P}_{N-1}) \geq \mathrm{var}(f, \tilde{P}_N) \ldots$$

Every sample counts in pseudo-marginal MCMC

- As mentioned earlier a suggestion in order to improve the performance of such algorithms one can suggest averaging, i.e. use an average of (say independent) estimates of the density

$$\pi(x)W^N := \pi(x)\frac{1}{N}\sum_{i=1}^{N} W_i$$

- Intuitively this should help since we are reducing the variance. But we know that the variance is not necessarily a good indicator (counterexample).
- However... for exchangeable random variables, it is known that for any $N \geq 1$

$$\frac{1}{N+1}\sum_{i=1}^{N+1} W_i \leq_{cx} \frac{1}{N}\sum_{i=1}^{N} W_i$$

- Which from our results immediately implies that for any $f \in L^2(\mathsf{X}, \pi)$ and any $N \geq 2$

$$\mathrm{var}(f, \tilde{P}_{N-1}) \geq \mathrm{var}(f, \tilde{P}_N)...$$

Éléments de preuve

- As pointed out earlier the main difficulty when trying to establish an order here stems from the fact that $\tilde{P}_1$ and $\tilde{P}_2$ do not share the same invariant distribution since for $i \in \{1, 2\}$

$$\tilde{\pi}^{(i)}\big(\mathrm{d}x \times \mathrm{d}w\big) = \pi(\mathrm{d}x)Q_x^{(i)}(\mathrm{d}w)w$$

- The central idea of the proof is to embed these two probability distributions into one, $\breve{\pi}$
- With this idea in mind (and say, $W_x^{(1)}$ "less noisy" than $W_x^{(2)}$) we consider

$$\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x)Q_x^{(1)}(\mathrm{d}w)w \times K_{x,w}(\mathrm{d}m)m \quad ,$$

where we have the properties

1. $\int Q_x^{(1)}(\mathrm{d}w)K_{x,w}(\mathrm{d}m)\mathbb{I}\{w \times m \in A\} = Q_x^{(2)}(A)$ for all $(x, A) \in \mathsf{X} \times \mathcal{B}(\mathbb{R}_+)$
2. $\int K_{x,w}(\mathrm{d}m)m = 1$

- m can be thought of as a Martingale multiplicative increment which "adds" noise to w

Éléments de preuve

- As pointed out earlier the main difficulty when trying to establish an order here stems from the fact that $\tilde{P}_1$ and $\tilde{P}_2$ do not share the same invariant distribution since for $i \in \{1, 2\}$

$$\tilde{\pi}^{(i)}\big(\mathrm{d}x \times \mathrm{d}w\big) = \pi(\mathrm{d}x)Q_x^{(i)}(\mathrm{d}w)w$$

- The central idea of the proof is to embed these two probability distributions into one, $\breve{\pi}$
- With this idea in mind (and say, $W_x^{(1)}$ "less noisy" than $W_x^{(2)}$) we consider

$$\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x)Q_x^{(1)}(\mathrm{d}w)w \times K_{x,w}(\mathrm{d}m)m \quad ,$$

  where we have the properties
  1. $\int Q_x^{(1)}(\mathrm{d}w)K_{x,w}(\mathrm{d}m)\mathbb{I}\{w \times m \in A\} = Q_x^{(2)}(A)$ for all $(x, A) \in \mathsf{X} \times \mathcal{B}(\mathbb{R}_+)$
  2. $\int K_{x,w}(\mathrm{d}m)m = 1$
- m can be thought of as a Martingale multiplicative increment which "adds" noise to w

Éléments de preuve

- As pointed out earlier the main difficulty when trying to establish an order here stems from the fact that $\tilde{P}_1$ and $\tilde{P}_2$ do not share the same invariant distribution since for $i \in \{1, 2\}$

$$\tilde{\pi}^{(i)}(\mathrm{d}x \times \mathrm{d}w) = \pi(\mathrm{d}x) Q_x^{(i)}(\mathrm{d}w) w$$

- The central idea of the proof is to embed these two probability distributions into one, $\breve{\pi}$
- With this idea in mind (and say, $W_x^{(1)}$ "less noisy" than $W_x^{(2)}$) we consider

$$\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x^{(1)}(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m \quad ,$$

where we have the properties

1. $\int Q_x^{(1)}(\mathrm{d}w) K_{x,w}(\mathrm{d}m) \mathbb{I}\{w \times m \in A\} = Q_x^{(2)}(A)$ for all $(x, A) \in \mathsf{X} \times \mathcal{B}(\mathbb{R}_+)$
2. $\int K_{x,w}(\mathrm{d}m) m = 1$

- m can be thought of as a Martingale multiplicative increment which "adds" noise to w

Éléments de preuve

- As pointed out earlier the main difficulty when trying to establish an order here stems from the fact that $\tilde{P}_1$ and $\tilde{P}_2$ do not share the same invariant distribution since for $i \in \{1, 2\}$

$$\tilde{\pi}^{(i)}(\mathrm{d}x \times \mathrm{d}w) = \pi(\mathrm{d}x) Q_x^{(i)}(\mathrm{d}w) w$$

- The central idea of the proof is to embed these two probability distributions into one, $\breve{\pi}$
- With this idea in mind (and say, $W_x^{(1)}$ "less noisy" than $W_x^{(2)}$) we consider

$$\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x^{(1)}(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m \quad ,$$

where we have the properties

1. $\int Q_x^{(1)}(\mathrm{d}w) K_{x,w}(\mathrm{d}m) \mathbb{I}\{w \times m \in A\} = Q_x^{(2)}(A)$ for all $(x, A) \in \mathsf{X} \times \mathcal{B}(\mathbb{R}_+)$
2. $\int K_{x,w}(\mathrm{d}m) m = 1$

- m can be thought of as a Martingale multiplicative increment which "adds" noise to w

Éléments de preuve

- As pointed out earlier the main difficulty when trying to establish an order here stems from the fact that $\tilde{P}_1$ and $\tilde{P}_2$ do not share the same invariant distribution since for $i \in \{1, 2\}$

$$\tilde{\pi}^{(i)}(\mathrm{d}x \times \mathrm{d}w) = \pi(\mathrm{d}x) Q_x^{(i)}(\mathrm{d}w) w$$

- The central idea of the proof is to embed these two probability distributions into one, $\breve{\pi}$
- With this idea in mind (and say, $W_x^{(1)}$ "less noisy" than $W_x^{(2)}$) we consider

$$\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x^{(1)}(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m \quad ,$$

  where we have the properties

  1. $\int Q_x^{(1)}(\mathrm{d}w) K_{x,w}(\mathrm{d}m) \mathbb{I}\{w \times m \in A\} = Q_x^{(2)}(A)$ for all $(x, A) \in \mathsf{X} \times \mathcal{B}(\mathbb{R}_+)$
  2. $\int K_{x,w}(\mathrm{d}m) m = 1$

- m can be thought of as a Martingale multiplicative increment which "adds" noise to w

Éléments de preuve

- As pointed out earlier the main difficulty when trying to establish an order here stems from the fact that $\tilde{P}_1$ and $\tilde{P}_2$ do not share the same invariant distribution since for $i \in \{1, 2\}$

$$\tilde{\pi}^{(i)}(\mathrm{d}x \times \mathrm{d}w) = \pi(\mathrm{d}x) Q_x^{(i)}(\mathrm{d}w) w$$

- The central idea of the proof is to embed these two probability distributions into one, $\breve{\pi}$
- With this idea in mind (and say, $W_x^{(1)}$ "less noisy" than $W_x^{(2)}$) we consider

$$\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x^{(1)}(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m \quad ,$$

where we have the properties

1. $\int Q_x^{(1)}(\mathrm{d}w) K_{x,w}(\mathrm{d}m) \mathbb{I}\{w \times m \in A\} = Q_x^{(2)}(A)$ for all $(x, A) \in \mathsf{X} \times \mathcal{B}(\mathbb{R}_+)$
2. $\int K_{x,w}(\mathrm{d}m) m = 1$

- m can be thought of as a Martingale multiplicative increment which "adds" noise to w

Strassen's characterisation

- One of the miracles in this work is that Strassen's characterisation of the convex order tells us that for $x \in \mathsf{X}$, $W_x^{(1)} \leq_{cx} W_x^{(2)}$ "less noisy" then

Theorem (Strassen)

Suppose that $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$ are well-defined. Then, $W_1 \leq_{cx} W_2$ if and only if there exists a probability space with random variables $\check{W}_1$ and $\check{W}_2$ coinciding with W_1 and W_2 in distribution, respectively, such that $(\check{W}_1, \check{W}_2)$ is a martingale pair, that is, $\mathbb{E}[\check{W}_2 \mid \check{W}_1] = \check{W}_1$ (a.s.).

- Here there are some subtle measurability issues since Strassen's theorem can be applied for any $x \in \mathsf{X}$ but we require

$$\check{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x^{(1)}(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m$$

to define a probability distribution...

- An extension of Strassen's theorem is required in practice. This is possible, highly technical and skipped here.

Strassen's characterisation

- One of the miracles in this work is that Strassen's characterisation of the convex order tells us that for $x \in \mathsf{X}$, $W_x^{(1)} \leq_{cx} W_x^{(2)}$ "less noisy" then

Theorem (Strassen)

Suppose that $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$ are well-defined. Then, $W_1 \leq_{cx} W_2$ if and only if there exists a probability space with random variables $\check{W}_1$ and $\check{W}_2$ coinciding with W_1 and W_2 in distribution, respectively, such that $(\check{W}_1, \check{W}_2)$ is a martingale pair, that is, $\mathbb{E}[\check{W}_2 \mid \check{W}_1] = \check{W}_1$ (a.s.).

- Here there are some subtle measurability issues since Strassen's theorem can be applied for any $x \in \mathsf{X}$ but we require

$$\check{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x^{(1)}(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m$$

  to define a probability distribution...

- An extension of Strassen's theorem is required in practice. This is possible, highly technical and skipped here.

Strassen's characterisation

- One of the miracles in this work is that Strassen's characterisation of the convex order tells us that for $x \in \mathsf{X}$, $W_x^{(1)} \leq_{cx} W_x^{(2)}$ "less noisy" then

Theorem (Strassen)

Suppose that $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$ are well-defined. Then, $W_1 \leq_{cx} W_2$ if and only if there exists a probability space with random variables $\check{W}_1$ and $\check{W}_2$ coinciding with W_1 and W_2 in distribution, respectively, such that $(\check{W}_1, \check{W}_2)$ is a martingale pair, that is, $\mathbb{E}[\check{W}_2 \mid \check{W}_1] = \check{W}_1$ (a.s.).

- Here there are some subtle measurability issues since Strassen's theorem can be applied for any $x \in \mathsf{X}$ but we require

$$\check{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x^{(1)}(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m$$

to define a probability distribution...

- An extension of Strassen's theorem is required in practice. This is possible, highly technical and skipped here.

Strassen's characterisation

- One of the miracles in this work is that Strassen's characterisation of the convex order tells us that for $x \in \mathsf{X}$, $W_x^{(1)} \leq_{cx} W_x^{(2)}$ "less noisy" then

Theorem (Strassen)

Suppose that $\mathbb{E}[W_1]$ and $\mathbb{E}[W_2]$ are well-defined. Then, $W_1 \leq_{cx} W_2$ if and only if there exists a probability space with random variables $\check{W}_1$ and $\check{W}_2$ coinciding with W_1 and W_2 in distribution, respectively, such that $(\check{W}_1, \check{W}_2)$ is a martingale pair, that is, $\mathbb{E}[\check{W}_2 \mid \check{W}_1] = \check{W}_1$ (a.s.).

- Here there are some subtle measurability issues since Strassen's theorem can be applied for any $x \in \mathsf{X}$ but we require

$$\check{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x^{(1)}(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m$$

to define a probability distribution...

- An extension of Strassen's theorem is required in practice. This is possible, highly technical and skipped here.

Working on the embedding space

- Now we consider two Markov transition probabilities $\breve{P}^{(1)}$ and $\breve{P}^{(2)}$ reversible with respect to $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m)$

- For $f, g \in L^2(\mathsf{E}, \mu)$ define $\langle f, g \rangle_\mu := \int f(z)g(z)\mu(\mathrm{d}z)$

- One can establish that with $\bar{f} = f - \mu(f)$,
$\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2\sum_{k=1}^{\infty} \langle \bar{f}, \Pi^k \bar{f} \rangle_\mu$,

- We aim to construct $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ such that for $k \geq 0$ and $g : \mathsf{X} \to \mathbb{R}$

$$\left\langle g, \left[\breve{P}^{(1)}\right]^k g \right\rangle_{\breve{\pi}} = \left\langle g, \left[\tilde{P}^{(1)}\right]^k g \right\rangle_{\tilde{\pi}^{(1)}} \quad ,$$

$$\left\langle g, \left[\breve{P}^{(2)}\right]^k g \right\rangle_{\breve{\pi}} = \left\langle g, \left[\tilde{P}^{(2)}\right]^k g \right\rangle_{\tilde{\pi}^{(2)}} \quad .$$

- So $\mathrm{var}(g, \breve{P}^{(1)}) = \mathrm{var}(g, \tilde{P}^{(1)})$ $\mathrm{var}(g, \breve{P}^{(2)}) = \mathrm{var}(g, \tilde{P}^{(2)})$ and it is sufficient to establish the sought result on the "fictitious" kernels in order to deduce the result on the kernels of interest.

Working on the embedding space

- Now we consider two Markov transition probabilities $\breve{P}^{(1)}$ and $\breve{P}^{(2)}$ reversible with respect to $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m)$
- For $f, g \in L^2(\mathsf{E}, \mu)$ define $\langle f, g \rangle_\mu := \int f(z)g(z)\mu(\mathrm{d}z)$
- One can establish that with $\bar{f} = f - \mu(f)$, $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2\sum_{k=1}^{\infty} \langle \bar{f}, \Pi^k \bar{f} \rangle_\mu$,
- We aim to construct $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ such that for $k \geq 0$ and $g : \mathsf{X} \to \mathbb{R}$

$$\begin{aligned} \left\langle g, \left[\breve{P}^{(1)}\right]^k g \right\rangle_{\breve{\pi}} &= \left\langle g, \left[\tilde{P}^{(1)}\right]^k g \right\rangle_{\tilde{\pi}^{(1)}} \quad , \\ \left\langle g, \left[\breve{P}^{(2)}\right]^k g \right\rangle_{\breve{\pi}} &= \left\langle g, \left[\tilde{P}^{(2)}\right]^k g \right\rangle_{\tilde{\pi}^{(2)}} \quad . \end{aligned}$$

- So $\mathrm{var}(g, \breve{P}^{(1)}) = \mathrm{var}(g, \tilde{P}^{(1)})$ $\mathrm{var}(g, \breve{P}^{(2)}) = \mathrm{var}(g, \tilde{P}^{(2)})$ and it is sufficient to establish the sought result on the "fictitious" kernels in order to deduce the result on the kernels of interest.

Working on the embedding space

- Now we consider two Markov transition probabilities $\check{P}^{(1)}$ and $\check{P}^{(2)}$ reversible with respect to $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m)$
- For $f, g \in L^2(\mathsf{E}, \mu)$ define $\langle f, g \rangle_\mu := \int f(z) g(z) \mu(\mathrm{d}z)$
- One can establish that with $\bar{f} = f - \mu(f)$, $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2 \sum_{k=1}^{\infty} \left\langle \bar{f}, \Pi^k \bar{f} \right\rangle_\mu$,
- We aim to construct $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ such that for $k \geq 0$ and $g : \mathsf{X} \to \mathbb{R}$

$$\left\langle g, \left[\check{P}^{(1)}\right]^k g \right\rangle_{\breve{\pi}} = \left\langle g, \left[\tilde{P}^{(1)}\right]^k g \right\rangle_{\tilde{\pi}^{(1)}} \quad ,$$

$$\left\langle g, \left[\check{P}^{(2)}\right]^k g \right\rangle_{\breve{\pi}} = \left\langle g, \left[\tilde{P}^{(2)}\right]^k g \right\rangle_{\tilde{\pi}^{(2)}} \quad .$$

- So $\mathrm{var}(g, \check{P}^{(1)}) = \mathrm{var}(g, \tilde{P}^{(1)})$ $\mathrm{var}(g, \check{P}^{(2)}) = \mathrm{var}(g, \tilde{P}^{(2)})$ and it is sufficient to establish the sought result on the "fictitious" kernels in order to deduce the result on the kernels of interest.

Working on the embedding space

- Now we consider two Markov transition probabilities $\breve{P}^{(1)}$ and $\breve{P}^{(2)}$ reversible with respect to $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m)$

- For $f, g \in L^2(\mathsf{E}, \mu)$ define $\langle f, g\rangle_\mu := \int f(z)g(z)\mu(\mathrm{d}z)$

- One can establish that with $\bar{f} = f - \mu(f)$,
$\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2\sum_{k=1}^{\infty} \left\langle \bar{f}, \Pi^k \bar{f}\right\rangle_\mu$,

- We aim to construct $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ such that for $k \geq 0$ and $g : \mathsf{X} \to \mathbb{R}$

$$\left\langle g, \left[\breve{P}^{(1)}\right]^k g\right\rangle_{\breve{\pi}} = \left\langle g, \left[\tilde{P}^{(1)}\right]^k g\right\rangle_{\tilde{\pi}^{(1)}} \quad ,$$
$$\left\langle g, \left[\breve{P}^{(2)}\right]^k g\right\rangle_{\breve{\pi}} = \left\langle g, \left[\tilde{P}^{(2)}\right]^k g\right\rangle_{\tilde{\pi}^{(2)}} \quad .$$

- So $\mathrm{var}(g, \breve{P}^{(1)}) = \mathrm{var}(g, \tilde{P}^{(1)})$ $\mathrm{var}(g, \breve{P}^{(2)}) = \mathrm{var}(g, \tilde{P}^{(2)})$ and it is sufficient to establish the sought result on the "fictitious" kernels in order to deduce the result on the kernels of interest.

Working on the embedding space

- Now we consider two Markov transition probabilities $\breve{P}^{(1)}$ and $\breve{P}^{(2)}$ reversible with respect to $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m)$
- For $f, g \in L^2(\mathsf{E}, \mu)$ define $\langle f, g \rangle_\mu := \int f(z) g(z) \mu(\mathrm{d}z)$
- One can establish that with $\bar{f} = f - \mu(f)$, $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2 \sum_{k=1}^{\infty} \left\langle \bar{f}, \Pi^k \bar{f} \right\rangle_\mu$,
- We aim to construct $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ such that for $k \geq 0$ and $g : \mathsf{X} \to \mathbb{R}$

$$\left\langle g, \left[\breve{P}^{(1)}\right]^k g \right\rangle_{\breve{\pi}} = \left\langle g, \left[\tilde{P}^{(1)}\right]^k g \right\rangle_{\tilde{\pi}^{(1)}} \quad ,$$
$$\left\langle g, \left[\breve{P}^{(2)}\right]^k g \right\rangle_{\breve{\pi}} = \left\langle g, \left[\tilde{P}^{(2)}\right]^k g \right\rangle_{\tilde{\pi}^{(2)}} \quad .$$

- So $\mathrm{var}(g, \breve{P}^{(1)}) = \mathrm{var}(g, \tilde{P}^{(1)})$ $\mathrm{var}(g, \breve{P}^{(2)}) = \mathrm{var}(g, \tilde{P}^{(2)})$ and it is sufficient to establish the sought result on the "fictitious" kernels in order to deduce the result on the kernels of interest.

Sampling in two different ways

- Two ways to think about the target
  $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x)Q_x(\mathrm{d}w)w \times K_{x,w}(\mathrm{d}m)m$ or
  $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x)Q_x(\mathrm{d}w)K_{x,w}(\mathrm{d}m)(w \times m)$

- The transitions are defined as follows
  1. $\breve{P}^{(1)}(x, w, m_w; \mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u) =$

$$q(x, \mathrm{d}y)Q_y(\mathrm{d}u)\min\left\{1, r(x,y)\frac{u}{w}\right\} K_{y,u}(\mathrm{d}m_u)m_u + \delta_{x,w,m_w}(\mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u)\breve{\rho}^{(1)}(x, w)$$

  1. $\breve{P}^{(2)}(x, w, m_w; \mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u) =$

$$q(x, \mathrm{d}y)Q_y(\mathrm{d}u)K_{y,u}(\mathrm{d}m_u)\min\left\{1, r(x,y)\frac{u\, m_u}{w\, m_w}\right\} + \delta_{x,w,m_w}(\mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u)\breve{\rho}^{(2)}(x, wm_w)$$

- It should be at least believable that there is a correspondence between $\breve{P}^{(1)}$, $\breve{P}^{(2)}$ and $\tilde{P}^{(1)}$, $\tilde{P}^{(2)}$.

Sampling in two different ways

- Two ways to think about the target
  $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m$ or
  $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x(\mathrm{d}w) K_{x,w}(\mathrm{d}m)(w \times m)$

- The transitions are defined as follows
  1. $\breve{P}^{(1)}(x, w, m_w; \mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u) =$

$$q(x, \mathrm{d}y) Q_y(\mathrm{d}u) \min\left\{1, r(x,y)\frac{u}{w}\right\} K_{y,u}(\mathrm{d}m_u) m_u + \delta_{x,w,m_w}(\mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u)\breve{\rho}^{(1)}(x, w)$$

  1. $\breve{P}^{(2)}(x, w, m_w; \mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u) =$

$$q(x, \mathrm{d}y) Q_y(\mathrm{d}u) K_{y,u}(\mathrm{d}m_u) \min\left\{1, r(x,y)\frac{u\, m_u}{w\, m_w}\right\} + \delta_{x,w,m_w}(\mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u)\breve{\rho}^{(2)}(x, w m_w)$$

- It should be at least believable that there is a correspondence between $\breve{P}^{(1)}$, $\breve{P}^{(2)}$ and $\tilde{P}^{(1)}$, $\tilde{P}^{(2)}$.

Sampling in two different ways

- Two ways to think about the target
  $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m$ or
  $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x(\mathrm{d}w) K_{x,w}(\mathrm{d}m)(w \times m)$

- The transitions are defined as follows
  1. $\breve{P}^{(1)}(x, w, m_w; \mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u) =$

$$q(x, \mathrm{d}y) Q_y(\mathrm{d}u) \min\left\{1, r(x,y)\frac{u}{w}\right\} K_{y,u}(\mathrm{d}m_u) m_u$$
$$+ \delta_{x,w,m_w}(\mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u) \breve{\rho}^{(1)}(x, w)$$

  1. $\breve{P}^{(2)}(x, w, m_w; \mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u) =$

$$q(x, \mathrm{d}y) Q_y(\mathrm{d}u) K_{y,u}(\mathrm{d}m_u) \min\left\{1, r(x,y)\frac{u\, m_u}{w\, m_w}\right\}$$
$$+ \delta_{x,w,m_w}(\mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u) \breve{\rho}^{(2)}(x, w m_w)$$

- It should be at least believable that there is a correspondence between $\breve{P}^{(1)}$, $\breve{P}^{(2)}$ and $\tilde{P}^{(1)}$, $\tilde{P}^{(2)}$.

Sampling in two different ways

- Two ways to think about the target
  $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x(\mathrm{d}w) w \times K_{x,w}(\mathrm{d}m) m$ or
  $\breve{\pi}(\mathrm{d}x, \mathrm{d}w, \mathrm{d}m) := \pi(\mathrm{d}x) Q_x(\mathrm{d}w) K_{x,w}(\mathrm{d}m)(w \times m)$

- The transitions are defined as follows
  1. $\breve{P}^{(1)}(x, w, m_w; \mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u)=$
  $$q(x, \mathrm{d}y) Q_y(\mathrm{d}u) \min\left\{1, r(x,y)\frac{u}{w}\right\} K_{y,u}(\mathrm{d}m_u) m_u \\ + \delta_{x,w,m_w}(\mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u)\breve{\rho}^{(1)}(x, w)$$
  1. $\breve{P}^{(2)}(x, w, m_w; \mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u)=$
  $$q(x, \mathrm{d}y) Q_y(\mathrm{d}u) K_{y,u}(\mathrm{d}m_u) \min\left\{1, r(x,y)\frac{u\, m_u}{w\, m_w}\right\} \\ + \delta_{x,w,m_w}(\mathrm{d}y, \mathrm{d}u, \mathrm{d}m_u)\breve{\rho}^{(2)}(x, w m_w)$$

- It should be at least believable that there is a correspondence between $\breve{P}^{(1)}$, $\breve{P}^{(2)}$ and $\tilde{P}^{(1)}$, $\tilde{P}^{(2)}$.

Hilbert space techniques I

- Let μ be a probability distribution on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ and Π a Markov kernel reversible w.r.t. μ.
- One can establish that with $\bar{f} = f - \mu(f)$, $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2\sum_{k=1}^{\infty} \left\langle f, \Pi^k f \right\rangle_\mu$,
- Then for $\lambda \in [0, 1)$

$$\mathrm{var}(f, \lambda\Pi) := 2\left\langle f, (I - \lambda\Pi)^{-1} f \right\rangle_\mu - \|f\|_\mu^2.$$

  where $(I - \lambda\Pi)^{-1} := \sum_{k=0}^{\infty} \lambda^k \Pi^k$.
- Define the "Dirichlet forms" $\mathcal{E}_{\lambda\Pi}(f) := \langle f, (I - \lambda\Pi) f \rangle_\mu$ [related to the first order autocovariance coefficient of the chain]
- Now for Π_1 and Π_2 reversible w.r.t μ the property underpinning Peskun's result is essentially

$$\left[\forall f \in L^2(\mathsf{E}, \mu) \quad \left\langle f, (I - \lambda\Pi_2)^{-1} f \right\rangle_\mu \geq \left\langle f, (I - \lambda\Pi_1)^{-1} f \right\rangle_\mu\right]$$

$$\iff \left[\forall g \in L^2(\mathsf{E}, \mu) \quad \langle g, (I - \lambda\Pi_2) g \rangle_\mu \leq \langle g, (I - \lambda\Pi_1) g \rangle_\mu\right]$$

Hilbert space techniques I

- Let μ be a probability distribution on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ and Π a Markov kernel reversible w.r.t. μ.
- One can establish that with $\bar{f} = f - \mu(f)$, $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2\sum_{k=1}^{\infty} \left\langle f, \Pi^k f \right\rangle_\mu$,
- Then for $\lambda \in [0, 1)$
$$\mathrm{var}(f, \lambda\Pi) := 2\left\langle f, (I - \lambda\Pi)^{-1} f \right\rangle_\mu - \|f\|_\mu^2.$$
where $(I - \lambda\Pi)^{-1} := \sum_{k=0}^{\infty} \lambda^k \Pi^k$.
- Define the "Dirichlet forms" $\mathcal{E}_{\lambda\Pi}(f) := \langle f, (I - \lambda\Pi) f \rangle_\mu$ [related to the first order autocovariance coefficient of the chain]
- Now for Π_1 and Π_2 reversible w.r.t μ the property underpinning Peskun's result is essentially
$$\left[\forall f \in L^2(\mathsf{E}, \mu) \quad \left\langle f, (I - \lambda\Pi_2)^{-1} f \right\rangle_\mu \geq \left\langle f, (I - \lambda\Pi_1)^{-1} f \right\rangle_\mu\right]$$
$$\iff \left[\forall g \in L^2(\mathsf{E}, \mu) \quad \langle g, (I - \lambda\Pi_2) g \rangle_\mu \leq \langle g, (I - \lambda\Pi_1) g \rangle_\mu\right]$$

Hilbert space techniques I

- Let μ be a probability distribution on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ and Π a Markov kernel reversible w.r.t. μ.
- One can establish that with $\bar{f} = f - \mu(f)$,
$\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2\sum_{k=1}^{\infty} \left\langle f, \Pi^k f \right\rangle_\mu$,
- Then for $\lambda \in [0, 1)$
$$\mathrm{var}(f, \lambda\Pi) := 2\left\langle f, (I - \lambda\Pi)^{-1} f \right\rangle_\mu - \|f\|_\mu^2.$$
where $(I - \lambda\Pi)^{-1} := \sum_{k=0}^{\infty} \lambda^k \Pi^k$.
- Define the "Dirichlet forms" $\mathcal{E}_{\lambda\Pi}(f) := \langle f, (I - \lambda\Pi) f\rangle_\mu$ [related to the first order autocovariance coefficient of the chain]
- Now for Π_1 and Π_2 reversible w.r.t μ the property underpinning Peskun's result is essentially
$$\left[\forall f \in L^2(\mathsf{E}, \mu) \quad \left\langle f, (I - \lambda\Pi_2)^{-1} f \right\rangle_\mu \geq \left\langle f, (I - \lambda\Pi_1)^{-1} f \right\rangle_\mu\right]$$
$$\Longleftrightarrow \left[\forall g \in L^2(\mathsf{E}, \mu) \quad \langle g, (I - \lambda\Pi_2) g\rangle_\mu \leq \langle g, (I - \lambda\Pi_1) g\rangle_\mu\right]$$

Hilbert space techniques I

- Let μ be a probability distribution on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ and Π a Markov kernel reversible w.r.t. μ.
- One can establish that with $\bar{f} = f - \mu(f)$, $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2\sum_{k=1}^{\infty} \left\langle f, \Pi^k f \right\rangle_\mu$,
- Then for $\lambda \in [0, 1)$
$$\mathrm{var}(f, \lambda\Pi) := 2\left\langle f, (I - \lambda\Pi)^{-1} f \right\rangle_\mu - \|f\|_\mu^2 .$$
where $(I - \lambda\Pi)^{-1} := \sum_{k=0}^{\infty} \lambda^k \Pi^k$.
- Define the "Dirichlet forms" $\mathcal{E}_{\lambda\Pi}\left(f\right) := \langle f, (I - \lambda\Pi) f \rangle_\mu$ [related to the first order autocovariance coefficient of the chain]
- Now for Π_1 and Π_2 reversible w.r.t μ the property underpinning Peskun's result is essentially
$$\left[\forall f \in L^2(\mathsf{E}, \mu) \quad \left\langle f, (I - \lambda\Pi_2)^{-1} f \right\rangle_\mu \geq \left\langle f, (I - \lambda\Pi_1)^{-1} f \right\rangle_\mu \right]$$
$$\iff \left[\forall g \in L^2(\mathsf{E}, \mu) \quad \langle g, (I - \lambda\Pi_2) g \rangle_\mu \leq \langle g, (I - \lambda\Pi_1) g \rangle_\mu \right]$$

Hilbert space techniques I

- Let μ be a probability distribution on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ and Π a Markov kernel reversible w.r.t. μ.
- One can establish that with $\bar{f} = f - \mu(f)$, $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f) + 2\sum_{k=1}^{\infty} \left\langle f, \Pi^k f \right\rangle_\mu$,
- Then for $\lambda \in [0, 1)$
$$\mathrm{var}(f, \lambda\Pi) := 2 \left\langle f, (I - \lambda\Pi)^{-1} f \right\rangle_\mu - \|f\|_\mu^2.$$
where $(I - \lambda\Pi)^{-1} := \sum_{k=0}^{\infty} \lambda^k \Pi^k$.
- Define the "Dirichlet forms" $\mathcal{E}_{\lambda\Pi}(f) := \langle f, (I - \lambda\Pi) f \rangle_\mu$ [related to the first order autocovariance coefficient of the chain]
- Now for Π_1 and Π_2 reversible w.r.t μ the property underpinning Peskun's result is essentially
$$\left[\forall f \in L^2(\mathsf{E}, \mu) \quad \left\langle f, (I - \lambda\Pi_2)^{-1} f \right\rangle_\mu \geq \left\langle f, (I - \lambda\Pi_1)^{-1} f \right\rangle_\mu \right]$$
$$\iff \left[\forall g \in L^2(\mathsf{E}, \mu) \quad \langle g, (I - \lambda\Pi_2) g \rangle_\mu \leq \langle g, (I - \lambda\Pi_1) g \rangle_\mu \right]$$

Explicit bounds

Theorem (Tierney)

Let Π_1 and Π_2 be two Markov transition probabilities defined on some measurable space $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ and reversible with respect to some common invariant distribution μ. Then for any $f \in L^2(\mathsf{E}, \mu)$ and any $\lambda \in [0, 1)$

$$\mathcal{E}_{\lambda\Pi_1}(\hat{f}_1^\lambda) - \mathcal{E}_{\lambda\Pi_2}(\hat{f}_1^\lambda) \leq \frac{1}{2}\Big[\mathrm{var}(f, \lambda\Pi_2) - \mathrm{var}(f, \lambda\Pi_1)\Big] \leq \mathcal{E}_{\lambda\Pi_1}(\hat{f}_2^\lambda) - \mathcal{E}_{\lambda\Pi_2}(\hat{f}_2^\lambda) \quad ,$$

where $\hat{f}_i^\lambda := \left(I - \lambda\Pi_i\right)^{-1} f$.

Back to $\breve{P}_i$

- The important point for us is that

$$\mathcal{E}_{\breve{P}^{(1)}}(\hat{f}_1) - \mathcal{E}_{\breve{P}^{(2)}}(\hat{f}_1) \leq \frac{1}{2}\left[\mathrm{var}(f, \breve{P}^{(2)}) - \mathrm{var}(f, \breve{P}^{(1)})\right] \quad .$$

- And

1. $\hat{f}_1 := \left(I - \breve{P}^{(1)}\right)^{-1} f$ is a function of x, w (not m) only if $f : \mathrm{X} \to \mathbb{R}$
2. it is easy to show (Jensen's inequality) that for $g(x, w) : \mathrm{X} \times \mathbb{R}_+ \to \mathbb{R}$

$$\mathcal{E}_{\breve{P}^{(1)}}(g) \geq \mathcal{E}_{\breve{P}^{(2)}}(g) \quad .$$

Back to $\breve{P}_i$

- The important point for us is that

$$\mathcal{E}_{\breve{P}^{(1)}}(\hat{f}_1) - \mathcal{E}_{\breve{P}^{(2)}}(\hat{f}_1) \leq \frac{1}{2}\left[\mathrm{var}(f, \breve{P}^{(2)}) - \mathrm{var}(f, \breve{P}^{(1)})\right] \quad .$$

- And
  1. $\hat{f}_1 := \left(I - \breve{P}^{(1)}\right)^{-1} f$ is a function of x, w (not m) only if $f : \mathsf{X} \to \mathbb{R}$
  2. it is easy to show (Jensen's inequality) that for $g(x, w) : \mathsf{X} \times \mathbb{R}_+ \to \mathbb{R}$

$$\mathcal{E}_{\breve{P}^{(1)}}(g) \geq \mathcal{E}_{\breve{P}^{(2)}}(g) \quad .$$

Ordering of Dirichlet forms

- The Dirichlet form for $\breve{P}^{(2)}$ and $g(x, w)$ [NOT dependent on m] is

$$\int \Big\{ [g(x,w) - g(y,u)]^2 \min\left\{1, r(x,y)\frac{u \times m_u}{w \times m_w}\right\} \times \\ \times \pi(\mathrm{d}x) Q_x(\mathrm{d}w) K_{x,w}(\mathrm{d}m_w) m_w q(x, \mathrm{d}y) Q_y(\mathrm{d}u) K_{y,u}(\mathrm{d}m_u) \Big\}$$

- For $x, y \in \mathrm{X}$ and $w, u \in \mathbb{R}_+$ we have from Jensen's inequality,

$$\begin{aligned} &\int \min\left\{1, r(x,y)\frac{u \times m_u}{w \times m_w}\right\} K_{x,w}(\mathrm{d}m_w) m_w K_{y,u}(\mathrm{d}m_u) \\ &\quad \le \min\left\{1, r(x,y)\int \frac{u \times m_u}{w \times m_w} K_{x,w}(\mathrm{d}m_w) m_w K_{y,u}(\mathrm{d}m_u)\right\} \\ &\quad = \min\left\{1, r(x,y)\frac{u}{w}\right\} \end{aligned}$$

- So $\mathcal{E}_{\breve{P}^{(1)}}(g) \ge \mathcal{E}_{\breve{P}^{(2)}}(g)$ and the conclusion follows.

Ordering of Dirichlet forms

- The Dirichlet form for $\check{P}^{(2)}$ and $g(x, w)$ [NOT dependent on m] is

$$\int \left\{ [g(x,w) - g(y,u)]^2 \min\left\{1, r(x,y)\frac{u \times m_u}{w \times m_w}\right\} \times \right.$$
$$\left. \times\ \pi(\mathrm{d}x) Q_x(\mathrm{d}w) K_{x,w}(\mathrm{d}m_w)\, m_w q(x, \mathrm{d}y) Q_y(\mathrm{d}u) K_{y,u}(\mathrm{d}m_u) \right\}$$

- For $x, y \in \mathrm{X}$ and $w, u \in \mathbb{R}_+$ we have from Jensen's inequality,

$$\begin{aligned}
&\int \min\left\{1, r(x,y)\frac{u \times m_u}{w \times m_w}\right\} K_{x,w}(\mathrm{d}m_w)\, m_w K_{y,u}(\mathrm{d}m_u) \\
&\quad \leq \min\left\{1, r(x,y) \int \frac{u \times m_u}{w \times m_w} K_{x,w}(\mathrm{d}m_w)\, m_w K_{y,u}(\mathrm{d}m_u)\right\} \\
&\quad = \min\left\{1, r(x,y)\frac{u}{w}\right\}
\end{aligned}$$

- So $\mathcal{E}_{\check{P}^{(1)}}(g) \geq \mathcal{E}_{\check{P}^{(2)}}(g)$ and the conclusion follows.

Ordering of Dirichlet forms

- The Dirichlet form for $\breve{P}^{(2)}$ and $g(x,w)$ [NOT dependent on m] is

$$\int \left\{ [g(x,w) - g(y,u)]^2 \min\left\{1, r(x,y)\frac{u \times m_u}{w \times m_w}\right\} \times \right.$$
$$\left. \times\ \pi(\mathrm{d}x) Q_x(\mathrm{d}w) K_{x,w}(\mathrm{d}m_w)\, m_w\, q(x,\mathrm{d}y) Q_y(\mathrm{d}u) K_{y,u}(\mathrm{d}m_u) \right\}$$

- For $x, y \in \mathsf{X}$ and $w, u \in \mathbb{R}_+$ we have from Jensen's inequality,

$$\begin{aligned}
&\int \min\left\{1, r(x,y)\frac{u \times m_u}{w \times m_w}\right\} K_{x,w}(\mathrm{d}m_w)\, m_w K_{y,u}(\mathrm{d}m_u) \\
&\quad \leq \min\left\{1, r(x,y) \int \frac{u \times m_u}{w \times m_w} K_{x,w}(\mathrm{d}m_w)\, m_w K_{y,u}(\mathrm{d}m_u)\right\} \\
&\quad = \min\left\{1, r(x,y)\frac{u}{w}\right\}
\end{aligned}$$

- So $\mathcal{E}_{\breve{P}^{(1)}}(g) \geq \mathcal{E}_{\breve{P}^{(2)}}(g)$ and the conclusion follows.

Conclusion

- Developed tools to compare pseudo-marginal and related MCMC algorithms,
- The convex order seems to be natural order + literature on the topic is rich,
- Effectively develop some sort of extension of Peskun's result...
- Other applications of these ideas.

Conclusion

- Developed tools to compare pseudo-marginal and related MCMC algorithms,
- The convex order seems to be natural order + literature on the topic is rich,
- Effectively develop some sort of extension of Peskun's result...
- Other applications of these ideas.

Conclusion

- Developed tools to compare pseudo-marginal and related MCMC algorithms,
- The convex order seems to be natural order + literature on the topic is rich,
- Effectively develop some sort of extension of Peskun's result...
- Other applications of these ideas.

Conclusion

- Developed tools to compare pseudo-marginal and related MCMC algorithms,
- The convex order seems to be natural order + literature on the topic is rich,
- Effectively develop some sort of extension of Peskun's result...
- Other applications of these ideas.

Rates of convergence of Markov chains

- Denote by $\mathcal{L}_x(\Phi_n)$ the law of a Markov chain Φ_n with
  1. transition probability Π and invariant distribution $\mu\Pi = \mu$,
  2. initial state $\Phi_0 \equiv x$.
- Recall the Markov chain convergence rates

$$\|\mathcal{L}_x(\Phi_n) - \mu\|_* \le \begin{cases} M\rho^n & \text{if uniformly ergodic} \\ MV(x)\rho^n & \text{if geometrically ergodic} \\ MV(x)n^{-p} & \text{if polynomially ergodic} \\ r^{-1}(n) & r(n) \to \infty \text{if ergodic.} \end{cases}$$

Some negative results

Theorem (CA and Roberts, 2009)

If the weight distributions are not (essentially) bounded, then the pseudo-marginal algorithm cannot be geometrically ergodic.

[The pseudo-marginal algorithm has a zero spectral gap if the set below has a positive π-mass,

$$\left\{ x \in \mathsf{X} : \int_M^\infty Q_x(w)\mathrm{d}w > 0 \text{ for all } M < \infty \right\} \quad]$$

Corollary

Even when P is geometrically ergodic if

1. *the noise is unbounded the approximation cannot be geometric,*
2. *for any $N \in \mathbb{N} \setminus \{0\}$, $\left\{ x \in \mathsf{X} : \int_M^\infty Q_x^N(w)\mathrm{d}w > 0 \text{ for all } M < \infty \right\}$ has a positive π-mass, then $\tilde{P}_N$ cannot be geometrically ergodic for any $N \in \mathbb{N} \setminus \{0\}$.*

Some negative results

Theorem (CA and Roberts, 2009)

If the weight distributions are not (essentially) bounded, then the pseudo-marginal algorithm cannot be geometrically ergodic.
[The pseudo-marginal algorithm has a zero spectral gap if the set below has a positive π-mass,

$$\left\{x \in \mathsf{X} : \int_M^\infty Q_x(w)\mathrm{d}w > 0 \text{ for all } M < \infty\right\} \quad]$$

Corollary

Even when P is geometrically ergodic if

1. *the noise is unbounded the approximation cannot be geometric,*
2. *for any $N \in \mathbb{N} \setminus \{0\}$, $\left\{x \in \mathsf{X} : \int_M^\infty Q_x^N(w)\mathrm{d}w > 0 \text{ for all } M < \infty\right\}$ has a positive π-mass, then $\tilde{P}_N$ cannot be geometrically ergodic for any $N \in \mathbb{N} \setminus \{0\}$.*

Some negative results

Theorem (CA and Roberts, 2009)

If the weight distributions are not (essentially) bounded, then the pseudo-marginal algorithm cannot be geometrically ergodic.
[The pseudo-marginal algorithm has a zero spectral gap if the set below has a positive π-mass,

$$\left\{x \in \mathsf{X} : \int_M^\infty Q_x(w)\mathrm{d}w > 0 \text{ for all } M < \infty\right\} \quad]$$

Corollary

Even when P is geometrically ergodic if

1. *the noise is unbounded the approximation cannot be geometric,*
2. *for any $N \in \mathbb{N} \setminus \{0\}$, $\left\{x \in \mathsf{X} : \int_M^\infty Q_x^N(w)\mathrm{d}w > 0 \text{ for all } M < \infty\right\}$ has a positive π-mass, then $\tilde{P}_N$ cannot be geometrically ergodic for any $N \in \mathbb{N} \setminus \{0\}$.*

Intuition

- The acceptance probability of the algorithm is

$$\min\left\{1, \frac{\pi(y) \times u \;\; q(y,x)}{\pi(x) \times w \;\; q(x,y)}\right\}$$

- The probability of escaping (x, w) can be made arbitrarily small...
- The Markov chain becomes "sticky".

Intuition

- The acceptance probability of the algorithm is

$$\min\left\{1, \frac{\pi(y) \times u \; q(y,x)}{\pi(x) \times w \; q(x,y)}\right\}$$

- The probability of escaping (x, w) can be made arbitrarily small...
- The Markov chain becomes "sticky".

Intuition

- The acceptance probability of the algorithm is

$$\min\left\{1, \frac{\pi(y) \times u \;\, q(y,x)}{\pi(x) \times w \;\, q(x,y)}\right\}$$

- The probability of escaping (x, w) can be made arbitrarily small...
- The Markov chain becomes "sticky".

Bounded weights

- One may wonder what happens when the support W of the weights is bounded?
- One can consider the spectral gaps of P and $\tilde{P}$ (remember that $1 - \mathrm{Gap}(\Pi)$ is the second largest eigenvalue of Π).

Theorem (CA and M. Vihola, 2012)

With P the idealised algorithm and $\tilde{P}$ its exact approximation, if the support of the weights is $\mathsf{W} = [0, \bar{w}]$ for some $\bar{w} > 1$ and $\pi(\{x\}) = 0$ for all $x \in \mathsf{X}$ then

$$1 - \mathrm{Gap}(\tilde{P}) \leq 1 - \bar{w}^{-1}\mathrm{Gap}(P) \quad .$$

Remark

Say that we have a sequence $W^N \sim Q_x^N$ and that for all $N \in \mathbb{N} \setminus \{0\}$ and any $x \in \mathsf{X}$, $\epsilon > 0$, $\int_{\bar{w}-\epsilon}^{\bar{w}} Q_x^N(w)\mathrm{d}w > 0$ then it is not possible in general to achieve the rate of convergence of the marginal chain P, even though we may have $\mathrm{var}_{Q_x^N}(W^N) \to 0$ as $N \to \infty$ for all $x \in \mathsf{X}$ (counter-example).

Bounded weights

- One may wonder what happens when the support W of the weights is bounded?
- One can consider the spectral gaps of P and $\tilde{P}$ (remember that $1 - \mathrm{Gap}(\Pi)$ is the second largest eigenvalue of Π).

Theorem (CA and M. Vihola, 2012)

With P the idealised algorithm and $\tilde{P}$ its exact approximation, if the support of the weights is $\mathsf{W} = [0, \bar{w}]$ for some $\bar{w} > 1$ and $\pi(\{x\}) = 0$ for all $x \in \mathsf{X}$ then

$$1 - \mathrm{Gap}(\tilde{P}) \leq 1 - \bar{w}^{-1}\mathrm{Gap}(P) \quad .$$

Remark

Say that we have a sequence $W^N \sim Q_x^N$ and that for all $N \in \mathbb{N} \setminus \{0\}$ and any $x \in \mathsf{X}$, $\epsilon > 0$, $\int_{\bar{w}-\epsilon}^{\bar{w}} Q_x^N(w)\mathrm{d}w > 0$ then it is not possible in general to achieve the rate of convergence of the marginal chain P, even though we may have $\mathrm{var}_{Q_x^N}(W^N) \to 0$ as $N \to \infty$ for all $x \in \mathsf{X}$ (counter-example).

Bounded weights

- One may wonder what happens when the support W of the weights is bounded?
- One can consider the spectral gaps of P and $\tilde{P}$ (remember that $1 - \mathrm{Gap}(\Pi)$ is the second largest eigenvalue of Π).

Theorem (CA and M. Vihola, 2012)

With P the idealised algorithm and $\tilde{P}$ its exact approximation, if the support of the weights is $\mathsf{W} = [0, \bar{w}]$ for some $\bar{w} > 1$ and $\pi(\{x\}) = 0$ for all $x \in \mathsf{X}$ then

$$1 - \mathrm{Gap}(\tilde{P}) \leq 1 - \bar{w}^{-1}\mathrm{Gap}(P) \quad .$$

Remark

Say that we have a sequence $W^N \sim Q_x^N$ and that for all $N \in \mathbb{N} \setminus \{0\}$ and any $x \in \mathsf{X}$, $\epsilon > 0$, $\int_{\bar{w}-\epsilon}^{\bar{w}} Q_x^N(w)\mathrm{d}w > 0$ then it is not possible in general to achieve the rate of convergence of the marginal chain P, even though we may have $\mathrm{var}_{Q_x^N}(W^N) \to 0$ as $N \to \infty$ for all $x \in \mathsf{X}$ (counter-example).

Bounded weights

- One may wonder what happens when the support W of the weights is bounded?
- One can consider the spectral gaps of P and $\tilde{P}$ (remember that $1 - \mathrm{Gap}(\Pi)$ is the second largest eigenvalue of Π).

Theorem (CA and M. Vihola, 2012)

With P the idealised algorithm and $\tilde{P}$ its exact approximation, if the support of the weights is $\mathsf{W} = [0, \bar{w}]$ for some $\bar{w} > 1$ and $\pi(\{x\}) = 0$ for all $x \in \mathsf{X}$ then

$$1 - \mathrm{Gap}(\tilde{P}) \leq 1 - \bar{w}^{-1}\mathrm{Gap}(P) \quad .$$

Remark

Say that we have a sequence $W^N \sim Q_x^N$ and that for all $N \in \mathbb{N} \setminus \{0\}$ and any $x \in \mathsf{X}$, $\epsilon > 0$, $\int_{\bar{w}-\epsilon}^{\bar{w}} Q_x^N(w)\mathrm{d}w > 0$ then it is not possible in general to achieve the rate of convergence of the marginal chain P, even though we may have $\mathrm{var}_{Q_x^N}(W^N) \to 0$ as $N \to \infty$ for all $x \in \mathsf{X}$ (counter-example).

Bounded weights—asymptotic variance

Proposition (CA & Vihola, 2012)

Assume the marginal algorithm is geometrically ergodic, the weights of the pseudo-marginal algorithm are upper-bounded by $\bar{w}$ *and* $\int f^2(x)\pi(x)\mathrm{d}x < \infty$. *Then,*

$$\boxed{\mathrm{var}(f, \tilde{P}) \leq \bar{w}\mathrm{var}(f, P) + (\bar{w} - 1)\mathrm{var}_\pi(f),} \tag{1}$$

Assume $\mathrm{Gap}(P) > 0$ *and*

$$\int_0^{\bar{w}} Q_x(w)\mathrm{d}w = 1 \qquad \text{for } \pi\text{-almost all } x \in \mathsf{X},$$

then (1) holds, where $\mathrm{var}_\pi(f) = \pi\big((f - \pi(f))^2\big)$.

Bounded weights—asymptotic variance

Proposition (CA & Vihola, 2012)

Assume the marginal algorithm is geometrically ergodic, the weights of the pseudo-marginal algorithm are upper-bounded by $\bar{w}$ *and* $\int f^2(x)\pi(x)\mathrm{d}x < \infty$. *Then,*

$$\boxed{\mathrm{var}(f, \tilde{P}) \le \bar{w}\mathrm{var}(f, P) + (\bar{w} - 1)\mathrm{var}_\pi(f),} \tag{1}$$

Assume $\mathrm{Gap}(P) > 0$ *and*

$$\int_0^{\bar{w}} Q_x(w)\mathrm{d}w = 1 \qquad \textit{for } \pi\textit{-almost all } x \in \mathsf{X},$$

then (1) holds, where $\mathrm{var}_\pi(f) = \pi\big((f - \pi(f))^2\big)$.

Ordering of the variances

Theorem (CA & Vihola, 2012)

The pseudo-marginal algorithm is never more efficient than the corresponding marginal algorithm (in terms of the asymptotic variance).

Assume $f : \mathsf{X} \to \mathbb{R}$ satisfies $\pi(f^2) < \infty$. The asymptotic variances of f with respect to the pseudo-marginal algorithm $\tilde{P}$ and the marginal algorithm P always satisfy

$$\boxed{\mathrm{var}(f, P) \leq \mathrm{var}(f, \tilde{P}) \quad .}$$

Remark

The result above is general and does not assume that the weights are bounded.

Remark

Note that although not unexpected, the result requires a non-trivial extension of Peskun's result.

Ordering of the variances

Theorem (CA & Vihola, 2012)

The pseudo-marginal algorithm is never more efficient than the corresponding marginal algorithm (in terms of the asymptotic variance).

Assume $f : \mathsf{X} \to \mathbb{R}$ *satisfies* $\pi(f^2) < \infty$. *The asymptotic variances of* f *with respect to the pseudo-marginal algorithm* $\tilde{P}$ *and the marginal algorithm* P *always satisfy*

$$\boxed{\mathrm{var}(f, P) \leq \mathrm{var}(f, \tilde{P}) \quad .}$$

Remark

The result above is general and does not assume that the weights are bounded.

Remark

Note that although not unexpected, the result requires a non-trivial extension of Peskun's result.

Ordering of the variances

Theorem (CA & Vihola, 2012)

The pseudo-marginal algorithm is never more efficient than the corresponding marginal algorithm (in terms of the asymptotic variance).

Assume $f : \mathsf{X} \to \mathbb{R}$ *satisfies* $\pi(f^2) < \infty$. *The asymptotic variances of* f *with respect to the pseudo-marginal algorithm* $\tilde{P}$ *and the marginal algorithm* P *always satisfy*

$$\boxed{\mathrm{var}(f, P) \leq \mathrm{var}(f, \tilde{P}) \quad .}$$

Remark

The result above is general and does not assume that the weights are bounded.

Remark

Note that although not unexpected, the result requires a non-trivial extension of Peskun's result.

Convergence in terms of variance

- If we combine the last two results, if the weights are upper-bounded by $\bar{w}$ then

$$\boxed{\mathrm{var}(f,P) \leq \mathrm{var}(f,\tilde{P}) \leq \bar{w}\mathrm{var}(f,P) + (\bar{w}-1)\mathrm{var}_{\pi}(f) \quad .}$$

- If we have a sequence $W^N \sim Q_x^N$ and the corresponding supports are $\mathsf{W}_N = [0, \bar{w}^N]$ and $\bar{w}^N \downarrow 1$ then the pseudo-marginal algorithm approaches P in terms of asymptotic variance i.e.

$$\lim_{N\to\infty} \mathrm{var}(f,\tilde{P}_N) = \mathrm{var}(f,P) \quad .$$

- In what follows we show how to extend these results to the (more realistic) case where the weights are unbounded.

Convergence in terms of variance

- If we combine the last two results, if the weights are upper-bounded by $\bar{w}$ then

$$\boxed{\mathrm{var}(f, P) \leq \mathrm{var}(f, \tilde{P}) \leq \bar{w}\mathrm{var}(f, P) + (\bar{w} - 1)\mathrm{var}_\pi(f) \quad .}$$

- If we have a sequence $W^N \sim Q_x^N$ and the corresponding supports are $\mathsf{W}_N = [0, \bar{w}^N]$ and $\bar{w}^N \downarrow 1$ then the pseudo-marginal algorithm approaches P in terms of asymptotic variance i.e.

$$\lim_{N \to \infty} \mathrm{var}(f, \tilde{P}_N) = \mathrm{var}(f, P) \quad .$$

- In what follows we show how to extend these results to the (more realistic) case where the weights are unbounded.

Convergence in terms of variance

- If we combine the last two results, if the weights are upper-bounded by $\bar{w}$ then

$$\boxed{\mathrm{var}(f, P) \leq \mathrm{var}(f, \tilde{P}) \leq \bar{w}\mathrm{var}(f, P) + (\bar{w} - 1)\mathrm{var}_\pi(f) \quad .}$$

- If we have a sequence $W^N \sim Q_x^N$ and the corresponding supports are $\mathsf{W}_N = [0, \bar{w}^N]$ and $\bar{w}^N \downarrow 1$ then the pseudo-marginal algorithm approaches P in terms of asymptotic variance i.e.

$$\lim_{N \to \infty} \mathrm{var}(f, \tilde{P}_N) = \mathrm{var}(f, P) \quad .$$

- In what follows we show how to extend these results to the (more realistic) case where the weights are unbounded.

Rates with w unbounded

- If P is geometric and w unbounded, what rates can one expect for $\tilde{P}$?
- It depends on the tail behaviour of $Q_x(W \geq w)$,
- The "practical" approach developed relies on the drift/minorization approach.
- Establishing these rates of convergence turns out to be essential to characterise the behaviour of $\tilde{P}_N$ as a function of N.

Rates with w unbounded

- If P is geometric and w unbounded, what rates can one expect for $\tilde{P}$?
- It depends on the tail behaviour of $Q_x(W \geq w)$,
- The "practical" approach developed relies on the drift/minorization approach.
- Establishing these rates of convergence turns out to be essential to characterise the behaviour of $\tilde{P}_N$ as a function of N.

Rates with w unbounded

- If P is geometric and w unbounded, what rates can one expect for $\tilde{P}$?
- It depends on the tail behaviour of $Q_x(W \geq w)$,
- The "practical" approach developed relies on the drift/minorization approach.
- Establishing these rates of convergence turns out to be essential to characterise the behaviour of $\tilde{P}_N$ as a function of N.

Rates with w unbounded

- If P is geometric and w unbounded, what rates can one expect for $\tilde{P}$?
- It depends on the tail behaviour of $Q_x (W \geq w)$,
- The "practical" approach developed relies on the drift/minorization approach.
- Establishing these rates of convergence turns out to be essential to characterise the behaviour of $\tilde{P}_N$ as a function of N.

A paedagogical example

- The independent Metropolis-Hastings (IMH) algorithm, albeit of limited practical interest, is relatively easy to analyse.
- If we target $\pi(\mathrm{d}x)$ with a proposal distribution $q(dx)$, the rate of convergence depends on the behaviour of $\mu(x) := \pi(\mathrm{d}x)/q(\mathrm{d}x)$
  1. the IMH is geometric iff. $\sup_{x\in\mathsf{X}} \mu(x) < \infty$ [Mengersen and Tweedie 1996],
  2. if $\int \mu^{\beta}(x)\pi(\mathrm{d}x) < \infty$ then the IMH is polynomially ergodic [Jarner and Roberts 2002],
  3. if $\int \phi\left(\mu(x)\right)\pi(\mathrm{d}x) < \infty$ (e.g. $\phi(x) = \exp(x)$) then the IMH is sub-geometric... [Douc Moulines Soulier 2007].
- We simply exploit that the pseudo-approximation of an IMH is an IMH algorithm (target is $\tilde{\pi}(dx \times dw)$ and the proposal is $q(\mathrm{d}x)Q_x(\mathrm{d}w)$.

A paedagogical example

- The independent Metropolis-Hastings (IMH) algorithm, albeit of limited practical interest, is relatively easy to analyse.
- If we target $\pi(\mathrm{d}x)$ with a proposal distribution $q(dx)$, the rate of convergence depends on the behaviour of $\mu(x) := \pi(\mathrm{d}x)/q(\mathrm{d}x)$
   1. the IMH is geometric iff. $\sup_{x \in \mathsf{X}} \mu(x) < \infty$ [Mengersen and Tweedie 1996],
   2. if $\int \mu^{\beta}(x)\pi(\mathrm{d}x) < \infty$ then the IMH is polynomially ergodic [Jarner and Roberts 2002],
   3. if $\int \phi\,(\mu(x))\,\pi(\mathrm{d}x) < \infty$ (e.g. $\phi(x) = \exp(x)$) then the IMH is sub-geometric... [Douc Moulines Soulier 2007].
- We simply exploit that the pseudo-approximation of an IMH is an IMH algorithm (target is $\tilde{\pi}(dx \times dw)$ and the proposal is $q(\mathrm{d}x)Q_x(\mathrm{d}w)$.

A paedagogical example

- The independent Metropolis-Hastings (IMH) algorithm, albeit of limited practical interest, is relatively easy to analyse.
- If we target $\pi(\mathrm{d}x)$ with a proposal distribution $q(dx)$, the rate of convergence depends on the behaviour of $\mu(x) := \pi(\mathrm{d}x)/q(\mathrm{d}x)$
  1. the IMH is geometric iff. $\sup_{x \in \mathsf{X}} \mu(x) < \infty$ [Mengersen and Tweedie 1996],
  2. if $\int \mu^{\beta}(x)\pi(\mathrm{d}x) < \infty$ then the IMH is polynomially ergodic [Jarner and Roberts 2002],
  3. if $\int \phi\left(\mu(x)\right)\pi(\mathrm{d}x) < \infty$ (e.g. $\phi(x) = \exp(x)$) then the IMH is sub-geometric... [Douc Moulines Soulier 2007].
- We simply exploit that the pseudo-approximation of an IMH is an IMH algorithm (target is $\tilde{\pi}(dx \times dw)$ and the proposal is $q(\mathrm{d}x)Q_x(\mathrm{d}w)$.

A paedagogical example

- The independent Metropolis-Hastings (IMH) algorithm, albeit of limited practical interest, is relatively easy to analyse.
- If we target $\pi(\mathrm{d}x)$ with a proposal distribution $q(dx)$, the rate of convergence depends on the behaviour of $\mu(x) := \pi(\mathrm{d}x)/q(\mathrm{d}x)$
  1. the IMH is geometric iff. $\sup_{x \in \mathsf{X}} \mu(x) < \infty$ [Mengersen and Tweedie 1996],
  2. if $\int \mu^{\beta}(x)\pi(\mathrm{d}x) < \infty$ then the IMH is polynomially ergodic [Jarner and Roberts 2002],
  3. if $\int \phi\left(\mu(x)\right)\pi(\mathrm{d}x) < \infty$ (e.g. $\phi(x) = \exp(x)$) then the IMH is sub-geometric... [Douc Moulines Soulier 2007].
- We simply exploit that the pseudo-approximation of an IMH is an IMH algorithm (target is $\tilde{\pi}(dx \times dw)$ and the proposal is $q(\mathrm{d}x)Q_x(\mathrm{d}w)$.

A paedagogical example

- The independent Metropolis-Hastings (IMH) algorithm, albeit of limited practical interest, is relatively easy to analyse.
- If we target $\pi(\mathrm{d}x)$ with a proposal distribution $q(dx)$, the rate of convergence depends on the behaviour of $\mu(x) := \pi(\mathrm{d}x)/q(\mathrm{d}x)$
  1. the IMH is geometric iff. $\sup_{x \in \mathsf{X}} \mu(x) < \infty$ [Mengersen and Tweedie 1996],
  2. if $\int \mu^{\beta}(x)\pi(\mathrm{d}x) < \infty$ then the IMH is polynomially ergodic [Jarner and Roberts 2002],
  3. if $\int \phi\left(\mu(x)\right)\pi(\mathrm{d}x) < \infty$ (e.g. $\phi(x) = \exp(x)$) then the IMH is sub-geometric... [Douc Moulines Soulier 2007].
- We simply exploit that the pseudo-approximation of an IMH is an IMH algorithm (target is $\tilde{\pi}(dx \times dw)$ and the proposal is $q(\mathrm{d}x)Q_x(\mathrm{d}w)$.

A paedagogical example

- The independent Metropolis-Hastings (IMH) algorithm, albeit of limited practical interest, is relatively easy to analyse.
- If we target $\pi(\mathrm{d}x)$ with a proposal distribution $q(dx)$, the rate of convergence depends on the behaviour of $\mu(x) := \pi(\mathrm{d}x)/q(\mathrm{d}x)$
    1. the IMH is geometric iff. $\sup_{x \in \mathsf{X}} \mu(x) < \infty$ [Mengersen and Tweedie 1996],
    2. if $\int \mu^{\beta}(x)\pi(\mathrm{d}x) < \infty$ then the IMH is polynomially ergodic [Jarner and Roberts 2002],
    3. if $\int \phi\left(\mu(x)\right)\pi(\mathrm{d}x) < \infty$ (e.g. $\phi(x) = \exp(x)$) then the IMH is sub-geometric... [Douc Moulines Soulier 2007].
- We simply exploit that the pseudo-approximation of an IMH is an IMH algorithm (target is $\tilde{\pi}(dx \times dw)$ and the proposal is $q(\mathrm{d}x)Q_x(\mathrm{d}w)$.

Drift approach

Proposition

Denote $\mu(x) = \pi(\mathrm{d}x)/q(\mathrm{d}x)$. *Suppose that there exists a strictly increasing* $\phi : (0, \infty) \to [1, \infty)$ *with* $\liminf_{t\to\infty} \phi(t)/t > 0$, *such that*

$$\int \tilde{\pi}(\mathrm{d}x, \mathrm{d}w)\phi\big(\mu(x)w\big) < \infty. \tag{2}$$

Then, there exists constants $M, c, \epsilon \in (0, \infty)$ *and a probability measure* ν *on* $(\mathsf{X} \times \mathsf{W}, \mathcal{B}(\mathsf{X}) \times \mathcal{B}(\mathsf{W}))$ *such that for all* $(x, w) \in \mathsf{X} \times \mathsf{W}$,

$$\tilde{P}V(x, w) \le V(x, w) - c\frac{V(x, w)}{\phi^{-1}\big(V(x, w)\big)}, \qquad \mu(x)w > M \tag{3}$$

$$\tilde{P}(x, w; \cdot) \ge \epsilon\nu(\cdot), \qquad \mu(x)w \le M, \tag{4}$$

and $\nu(V) < \infty$, *where* $V(x, w) = \phi\big(\mu(x)w\big)$.

Corollary: polynomial

Corollary

If for some $\beta \geq 1$

$$\int \tilde{\pi}(\mathrm{d}x \times \mathrm{d}w)\big(\mu(x)w\big)^{\beta} < \infty,$$

then there exist constants $M, c, c_V \in (0, \infty)$ *such that for* $\mu(x)w \geq M$*, we have the polynomial drift*

$$\tilde{P}V(x, w) \leq V(x, w) - cV^{\alpha}(x, w),$$

where $V(x, w) = (\mu(x)w)^{\beta} + 1$ *and* $\alpha = 1 - 1/\beta$*. We have for* $\xi \in [0, 1]$

$$\|\mathcal{L}_x(\Phi_n) - \mu\|_{V^{(1-\xi)\alpha}} \leq C_{\xi} V(x) n^{-\frac{\xi\alpha}{1-\alpha}}$$

Corollary: sub-exponential

Corollary

If for some $\gamma > 0$,

$$\int \tilde{\pi}(\mathrm{d}x \times \mathrm{d}w) \exp\left[(\mu(x)w)^{\gamma}\right] < \infty,$$

then there exist constants $M, c, c_V \in (0, \infty)$ *such that for* $\mu(x)w \geq M$, *we have the drift*

$$\tilde{P}V(x,w) \leq V(x,w) - c\kappa\big(V(x,w)\big),$$

where $V(x,w) = \exp\left((\mu(x)w)^{\gamma}\right)$ *and* $\kappa(t) = t(\log t)^{-1/\gamma}$. *We have for* $\xi \in (0,1)$ *and* $b \in \mathbb{R}$

$$\begin{aligned}&\|\mathcal{L}_x(\Phi_n) - \mu\|_{V^{\xi}/(1+\log V)^b} \\ &\qquad \leq C_{\xi} n^{(b+\gamma^{-1})/(1+\gamma^{-1})} \exp\left(-c(1-\xi)\{(1+\gamma^{-1})n^{\gamma/(1+\gamma)}\}\right)\end{aligned}$$

Uniform marginal algorithm

Proposition (CA and Vihola 2012)

Suppose that the one-step expected acceptance probability of the marginal algorithm is bounded away from zero,

$$\alpha_0 := \inf_{x \in \mathsf{X}} \int q(x, \mathrm{d}y) \min\{1, r(x, y)\} > 0,$$

and there exists a non-decreasing convex function $\phi : [0, \infty) \to [1, \infty)$ *satisfying*

$$\liminf_{t \to \infty} \frac{\phi(t)}{t} = \infty \qquad \text{and} \qquad M_W := \sup_{x \in \mathsf{X}} \int \phi(w) Q_x(\mathrm{d}w) < \infty.$$

Then, there exist constants $\delta > 0$ *and* $\bar{w} \in (1, \infty)$ *such that*

$$\tilde{P} V(x, w) \leq V(w) - \delta \frac{V(w)}{w} \mathbb{I}\{w \in [\bar{w}, \infty)\} + M_W \mathbb{I}\{w \in (0, \bar{w})\}.$$

where $V(x, w) = V(w) := \phi(w)$ *(*δ *and* $\bar{w}$ *depend only on* α_0, ϕ *and* M_W*).*

Marginal RWM-uniform moments

- We consider the situation where the marginal algorithm is geometrically convergent Random Walk Metropolis.
- It is known that this is the case when [Jarner & Hansen, 2000] see also [Roberts& Tweedie, 1996].
  1. π has a density which is continuously differentiable and supported on $\mathsf{X} = \mathbb{R}^d$,
  2. the tails of π are super-exponentially decaying and have regular contours, that is,

$$\lim_{|x|\to\infty} \frac{x}{|x|} \cdot \nabla \log \pi(x) = -\infty \qquad \text{and} \qquad \limsup_{|x|\to\infty} \frac{x}{|x|} \cdot \frac{\nabla \pi(x)}{|\nabla \pi(x)|} < 0,$$

  3. the proposal distribution satisfies $q(x, A) = q(A - x) = \int_A q(y - x)\mathrm{d}y$ with a symmetric density q bounded away from zero in some neighbourhood of the origin.
- "Strongly super-exponential condition".

Marginal RWM-uniform moments

- We consider the situation where the marginal algorithm is geometrically convergent Random Walk Metropolis.
- It is known that this is the case when [Jarner & Hansen, 2000] see also [Roberts& Tweedie, 1996].
  1. π has a density which is continuously differentiable and supported on $\mathsf{X} = \mathbb{R}^d$,
  2. the tails of π are super-exponentially decaying and have regular contours, that is,

$$\lim_{|x|\to\infty} \frac{x}{|x|} \cdot \nabla \log \pi(x) = -\infty \qquad \text{and} \qquad \limsup_{|x|\to\infty} \frac{x}{|x|} \cdot \frac{\nabla \pi(x)}{|\nabla \pi(x)|} < 0,$$

  3. the proposal distribution satisfies $q(x, A) = q(A - x) = \int_A q(y - x)\mathrm{d}y$ with a symmetric density q bounded away from zero in some neighbourhood of the origin.

- "Strongly super-exponential condition".

Marginal RWM-uniform moments

- We consider the situation where the marginal algorithm is geometrically convergent Random Walk Metropolis.
- It is known that this is the case when [Jarner & Hansen, 2000] see also [Roberts& Tweedie, 1996].
  1. π has a density which is continuously differentiable and supported on $\mathsf{X} = \mathbb{R}^d$,
  2. the tails of π are super-exponentially decaying and have regular contours, that is,

$$\lim_{|x|\to\infty} \frac{x}{|x|} \cdot \nabla \log \pi(x) = -\infty \qquad \text{and} \qquad \limsup_{|x|\to\infty} \frac{x}{|x|} \cdot \frac{\nabla\pi(x)}{|\nabla\pi(x)|} < 0,$$

  3. the proposal distribution satisfies $q(x, A) = q(A - x) = \int_A q(y - x)\mathrm{d}y$ with a symmetric density q bounded away from zero in some neighbourhood of the origin.
- "Strongly super-exponential condition".

Marginal RWM-uniform moments

- If in addition to the condition on the marginal algorithm we have a uniform moment condition on the distributions $\{Q_x\}_{x\in\mathsf{X}}$: there exist constants $\alpha' > 0$ and $\beta' > 1$ such that

$$M_W := \mathrm{esssup}_{x\in\mathsf{X}} \int \max\{w^{-\alpha'} \vee w^{\beta'}\} Q_x(\mathrm{d}w) < \infty, \tag{5}$$

  (the essential supremum is taken with respect to the Lebesgue measure).
- Then one can establish polynomial drift condition and conclude about the polynomial convergence of the pseudo-marginal algorithm,
- In fact one can replace the condition with more general moments and obtain other sub-geometric rates.
- What about non-uniform moments...?

Marginal RWM-uniform moments

- If in addition to the condition on the marginal algorithm we have a uniform moment condition on the distributions $\{Q_x\}_{x \in \mathsf{X}}$: there exist constants $\alpha' > 0$ and $\beta' > 1$ such that

$$M_W := \operatorname{esssup}_{x \in \mathsf{X}} \int \max\{w^{-\alpha'} \vee w^{\beta'}\} Q_x(\mathrm{d}w) < \infty, \tag{5}$$

  (the essential supremum is taken with respect to the Lebesgue measure).

- Then one can establish polynomial drift condition and conclude about the polynomial convergence of the pseudo-marginal algorithm,
- In fact one can replace the condition with more general moments and obtain other sub-geometric rates.
- What about non-uniform moments...?

Marginal RWM-uniform moments

- If in addition to the condition on the marginal algorithm we have a uniform moment condition on the distributions $\{Q_x\}_{x \in \mathsf{X}}$: there exist constants $\alpha' > 0$ and $\beta' > 1$ such that

$$M_W := \operatorname{esssup}_{x \in \mathsf{X}} \int \max\{w^{-\alpha'} \vee w^{\beta'}\} Q_x(\mathrm{d}w) < \infty, \qquad (5)$$

  (the essential supremum is taken with respect to the Lebesgue measure).
- Then one can establish polynomial drift condition and conclude about the polynomial convergence of the pseudo-marginal algorithm,
- In fact one can replace the condition with more general moments and obtain other sub-geometric rates.
- What about non-uniform moments...?

Marginal RWM-uniform moments

- If in addition to the condition on the marginal algorithm we have a uniform moment condition on the distributions $\{Q_x\}_{x \in \mathsf{X}}$: there exist constants $\alpha' > 0$ and $\beta' > 1$ such that

$$M_W := \mathrm{esssup}_{x \in \mathsf{X}} \int \max\{w^{-\alpha'} \vee w^{\beta'}\} Q_x(\mathrm{d}w) < \infty, \tag{5}$$

  (the essential supremum is taken with respect to the Lebesgue measure).
- Then one can establish polynomial drift condition and conclude about the polynomial convergence of the pseudo-marginal algorithm,
- In fact one can replace the condition with more general moments and obtain other sub-geometric rates.
- What about non-uniform moments...?

Ajelehtia Rambo pohjoisesta “drift Rambo from the North”.

Let $\hat{w} : \mathsf{X} \to [1, \infty)$ be a function bounded on compact sets and tending to infinity as $|x| \to \infty$. Let $\psi : (0, \infty) \to [1, \infty)$ be a non-increasing function such that $\psi(t) \to \infty$ as $t \to 0$, and define $g(x) := \psi(\pi(x))$.

1. There exist constants $\alpha' > 0$ and $\beta' > 1$ such that

$$\mathrm{esssup}_{x \in \mathsf{X}} g^{-1}(x) \int u^{-\alpha'} \vee u^{\beta'} Q_x(\mathrm{d}u) \leq 1,$$

2. There exist constants $\xi_w \in (0, \beta' - 1)$ and $\xi_\pi \in (0, \beta' - 1 - \xi_w)$,

$$\sup_{x \in \mathsf{X}} \frac{g(x)}{\hat{w}^{\xi_\pi}(x)} \sup_{z \in R_x} \left[\left(\frac{\pi(x+z)}{\pi(x)} \right)^{\xi_\pi} \frac{g(x+z)}{g(x)} \right] < \infty, \tag{6}$$

where $R_x := \{z : \frac{\pi(x+z)}{\pi(x)} < 1\}$ is the set of possible rejection for the marginal random-walk Metropolis algorithm.

3. For any $b > 1$, one must have $\sup_{x \in \mathsf{X}} M_W\big(b(|x| \vee 1)\big) / \hat{w}^{\xi_w}(x) < \infty$

$$M_W(r) := \mathrm{esssup}_{|x| \leq r} \int u^{-\alpha'} \vee u^{\beta'} Q_x(\mathrm{d}u) \leq \mathrm{esssup}_{|x| \leq r} g(x) \quad .$$

More

Surprisingly these conditions are implied by the simpler conditions...

Theorem

Suppose π *is strongly super-exponential and* q *regular, and that there exist* $\alpha' > 0$, $\beta' > 1$, $c < \infty$ *and* $\rho' \in [0, \rho - 1)$ *such that*

$$\int \max\left\{w^{-\alpha'}, w^{\beta'}\right\} Q_x(w)\mathrm{d}w \leq c \max\left\{1, |x|^{\rho'}\right\},$$

Then, defining $V(x, w) := \|\pi\|_\infty^\eta \pi^{-\eta}(x) \max\{w^{-\alpha}, w^\beta\}$ *for any*

$$\eta \in \left(0, \alpha' \wedge (\beta' - 1) \wedge 1\right), \quad \alpha \in (\eta, \alpha'], \quad \beta \in (1 - \eta, \beta' - \eta),$$

then there exist $\bar{w}, M, b \in [1, \infty)$, $\underline{w} \in (0, 1]$ *and* $\delta_V > 0$ *such that*

$$\boxed{\tilde{P}V(x, w) \leq \begin{cases} V(x, w) - \delta_V V^{\frac{\beta - 1}{\beta}}(x, w), & \text{for all } (x, w) \notin \mathsf{C}, \\ b, & \text{for all } (x, w) \in \mathsf{C}, \end{cases}}$$

where $\mathsf{C} := \{(x, w) : |x| \leq M,\ w \in [\underline{w}, \bar{w}]\}$.

Uniform vanishing of the IA's tails

- Showing that $\lim_{N\to\infty} \mathrm{var}(f, \tilde{P}_N) = \mathrm{var}(f, P)$ seem to require a fundamental property.
- Denote by $\tilde{X}_n^N$ the stationary pseudo-marginal chain with weight distribution Q_x^N. We require that for $f : \mathsf{X} \to \mathbb{R}$, denoting $\bar{f} = f - \pi(f)$,

$$\lim_{n\to\infty} \sup_{N\in\mathbb{N}} \left| \sum_{k=n}^{\infty} \mathbb{E}[\bar{f}(\tilde{X}_0^N)\bar{f}(\tilde{X}_k^N)] \right| = 0.$$

- The drift conditions established earlier allow us to verify these conditions, and in fact one can even obtain quantitative bounds.

Uniform vanishing of the IA's tails

- Showing that $\lim_{N\to\infty} \mathrm{var}(f, \tilde{P}_N) = \mathrm{var}(f, P)$ seem to require a fundamental property.
- Denote by $\tilde{X}_n^N$ the stationary pseudo-marginal chain with weight distribution Q_x^N. We require that for $f : \mathsf{X} \to \mathbb{R}$, denoting $\bar{f} = f - \pi(f)$,

$$\lim_{n\to\infty} \sup_{N\in\mathbb{N}} \left| \sum_{k=n}^{\infty} \mathbb{E}[\bar{f}(\tilde{X}_0^N)\bar{f}(\tilde{X}_k^N)] \right| = 0.$$

- The drift conditions established earlier allow us to verify these conditions, and in fact one can even obtain quantitative bounds.

Uniform vanishing of the IA's tails

- Showing that $\lim_{N\to\infty} \mathrm{var}(f, \tilde{P}_N) = \mathrm{var}(f, P)$ seem to require a fundamental property.
- Denote by $\tilde{X}_n^N$ the stationary pseudo-marginal chain with weight distribution Q_x^N. We require that for $f : \mathsf{X} \to \mathbb{R}$, denoting $\bar{f} = f - \pi(f)$,

$$\lim_{n\to\infty} \sup_{N\in\mathbb{N}} \left| \sum_{k=n}^{\infty} \mathbb{E}[\bar{f}(\tilde{X}_0^N)\bar{f}(\tilde{X}_k^N)] \right| = 0.$$

- The drift conditions established earlier allow us to verify these conditions, and in fact one can even obtain quantitative bounds.

Convergence of the variance

Theorem (CA & Vihola, 2012)

Under general technical conditions, the asymptotic variance of the pseudo-marginal algorithm converges to the asymptotic variance of the marginal algorithm.

Assume that $\int |f(x)|^{2+\delta}\pi(x)\mathrm{d}x < \infty$ *for some* $\delta > 0$, $\sum_{k=1}^{\infty} \mathbb{E}[\bar{f}(X_0)\bar{f}(X_k)] = c \in \mathbb{R}$ *and the Uniform IA vanishing assumption. Suppose also that,*

$$\lim_{N\to\infty} \int Q_x^N(w)|1-w|\mathrm{d}w = 0 \qquad \textit{for all } x \in \mathsf{X}.$$

Then,

$$\boxed{\lim_{N\to\infty} \mathrm{var}(f, \tilde{P}_N) = \mathrm{var}(f, P)\ .}$$

Convergence of the variance

Theorem (CA & Vihola, 2012)

Under general technical conditions, the asymptotic variance of the pseudo-marginal algorithm converges to the asymptotic variance of the marginal algorithm.

Assume that $\int |f(x)|^{2+\delta}\pi(x)\mathrm{d}x < \infty$ *for some* $\delta > 0$, $\sum_{k=1}^{\infty} \mathbb{E}[\bar{f}(X_0)\bar{f}(X_k)] = c \in \mathbb{R}$ *and the Uniform IA vanishing assumption. Suppose also that,*

$$\lim_{N\to\infty} \int Q_x^N(w)|1-w|\mathrm{d}w = 0 \qquad \textit{for all } x \in \mathsf{X}.$$

Then,

$$\boxed{\lim_{N\to\infty} \mathrm{var}(f, \tilde{P}_N) = \mathrm{var}(f, P)\ .}$$

Explicit bounds

- As a by-product of the proof one can get an "explicit" upper bound

$$\mathrm{var}(\tilde{P}_N) - \mathrm{var}(P) \leq C\left(S_N^{1/q} + r^{-1}[n_0(N)]\right)$$

- where (here for simplicity in the "marginal uniform" case)

$$S_N = n_0(N)\left[\sup_{x\in X} Q_x^N(|U-1| > \check{\epsilon}(N)) + \check{\epsilon}(N) + 2\sup_{x\in X}\int_1^{\infty} Q_x^N(U > t)dt\right]$$

for an adequate choice $n_0(N) \to \infty$ and $\check{\epsilon}(N) \downarrow 0$

- The bound depends explicitly on the distribution of the weights, which we can again characterise in terms of moments.

Explicit bounds

- As a by-product of the proof one can get an "explicit" upper bound

$$\mathrm{var}(\tilde{P}_N) - \mathrm{var}(P) \leq C\left(S_N^{1/q} + r^{-1}[n_0(N)]\right)$$

- where (here for simplicity in the "marginal uniform" case)

$$S_N = n_0(N)\left[\sup_{x\in\mathsf{X}} Q_x^N(|U-1| > \check{\epsilon}(N)) + \check{\epsilon}(N) + 2\sup_{x\in\mathsf{X}}\int_1^\infty Q_x^N(U > t)dt\right]$$

  for an adequate choice $n_0(N) \to \infty$ and $\check{\epsilon}(N) \downarrow 0$

- The bound depends explicitly on the distribution of the weights, which we can again characterise in terms of moments.

Explicit bounds

- As a by-product of the proof one can get an "explicit" upper bound

$$\mathrm{var}(\tilde{P}_N) - \mathrm{var}(P) \leq C\left(S_N^{1/q} + r^{-1}[n_0(N)]\right)$$

- where (here for simplicity in the "marginal uniform" case)

$$S_N = n_0(N)\left[\sup_{x\in\mathsf{X}} Q_x^N(|U-1| > \check{\epsilon}(N)) + \check{\epsilon}(N) + 2\sup_{x\in\mathsf{X}}\int_1^\infty Q_x^N(U > t)dt\right]$$

for an adequate choice $n_0(N) \to \infty$ and $\check{\epsilon}(N) \downarrow 0$

- The bound depends explicitly on the distribution of the weights, which we can again characterise in terms of moments.

Exponential moments

- We drop the dependence on x here and assume $\mathbb{E}\left[\exp(t(W-1))\right] < \infty \text{ for } |t| < H$ and we simply average N iid realisations
- Then by optimising $n_0(N) \to \infty$ and $\check{\epsilon}(N) \downarrow 0$

$$\mathrm{var}(P) - \mathrm{var}(\tilde{P}_N) \leq C \left(\log(N) \left[N^{-1/2} + g \log^{1/2}(N)/\sqrt{N} + \sqrt{2\pi g/N} \right] \right.$$
$$\left. + 2(NT)^{-1} \exp\left(-gT(N^2)/2\right) + \exp(-(\log(N))^{\gamma}) \right)$$

Polynomial moments

- Here we assume $\mathbb{E}\left[W^\beta\right] < \infty$ for $\beta \geq 2$
- And finds

$$\mathrm{var}(P) - \mathrm{var}(\tilde{P}_N) \leq \left(A + B/N^{\frac{1}{2}\frac{\beta}{1+\beta}}\right) N^{-\frac{1}{2}(\beta-1)/(\beta+1)}$$

Polynomial moments

- Here we assume $\mathbb{E}\left[W^{\beta}\right] < \infty$ for $\beta \geq 2$
- And finds

$$\mathrm{var}(P) - \mathrm{var}(\tilde{P}_N) \leq \left(A + B/N^{\frac{1}{2}\frac{\beta}{1+\beta}}\right) N^{-\frac{1}{2}(\beta-1)/(\beta+1)}$$

Sub-polynomial moments

- Just kidding...

Sub-polynomial moments

- Just kidding...

Conclusions

- Many recently proposed algorithms share the underlying noisy structures considered here,
- We have some understanding and characterisation of the properties of these algorithms in terms of moments of the "noise",
- In some recent work we show the monotonicity of $\mathrm{var}(\tilde{P}_N)$ and other quantities $\Rightarrow$ adaptive algorithms.

Conclusions

- Many recently proposed algorithms share the underlying noisy structures considered here,
- We have some understanding and characterisation of the properties of these algorithms in terms of moments of the "noise",
- In some recent work we show the monotonicity of $\mathrm{var}(\tilde{P}_N)$ and other quantities $\Rightarrow$ adaptive algorithms.

Conclusions

- Many recently proposed algorithms share the underlying noisy structures considered here,
- We have some understanding and characterisation of the properties of these algorithms in terms of moments of the "noise",
- In some recent work we show the monotonicity of $\mathrm{var}(\tilde{P}_N)$ and other quantities $\Rightarrow$ adaptive algorithms.

Thanks.

Thanks for your attention!

Counter-example

- Consider the independent MH algorithm, in the discrete case. It is possible to characterise exactly the second largest eigenvalue of the transition probability.
  - For P it takes the form $1-\left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}$
  - For $\tilde{P}$ it takes the form $1-\left(\sup_{(\theta,w)\in\Theta\times\mathsf{W}}\frac{\pi(\theta)}{q(\theta)}w\right)^{-1}$.
- If $\sup_{w\in\mathsf{W}} w$ is independent of θ, the second largest eigenvalue is exactly $1-\left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}\left(\sup_{w\in\mathsf{W}} w\right)^{-1}$ which is larger than $1-\left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}$ - **even for an arbitrarily small variance!**

Counter-example

- Consider the independent MH algorithm, in the discrete case. It is possible to characterise exactly the second largest eigenvalue of the transition probability.
  - For P it takes the form $1 - \left(\sup_{\theta\in\Theta} \frac{\pi(\theta)}{q(\theta)}\right)^{-1}$
  - For $\tilde{P}$ it takes the form $1 - \left(\sup_{(\theta,w)\in\Theta\times\mathsf{W}} \frac{\pi(\theta)}{q(\theta)} w\right)^{-1}$.
- If $\sup_{w\in\mathsf{W}} w$ is independent of θ, the second largest eigenvalue is exactly $1 - \left(\sup_{\theta\in\Theta} \frac{\pi(\theta)}{q(\theta)}\right)^{-1} \left(\sup_{w\in\mathsf{W}} w\right)^{-1}$ which is larger than $1 - \left(\sup_{\theta\in\Theta} \frac{\pi(\theta)}{q(\theta)}\right)^{-1}$ - **even for an arbitrarily small variance!**

Counter-example

- Consider the independent MH algorithm, in the discrete case. It is possible to characterise exactly the second largest eigenvalue of the transition probability.
  - For P it takes the form $1 - \left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}$
  - For $\tilde{P}$ it takes the form $1 - \left(\sup_{(\theta,w)\in\Theta\times\mathsf{W}}\frac{\pi(\theta)}{q(\theta)}w\right)^{-1}$.
- If $\sup_{w\in\mathsf{W}} w$ is independent of θ, the second largest eigenvalue is exactly $1 - \left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}\left(\sup_{w\in\mathsf{W}} w\right)^{-1}$ which is larger than $1 - \left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}$ - **even for an arbitrarily small variance!**

Counter-example

- Consider the independent MH algorithm, in the discrete case. It is possible to characterise exactly the second largest eigenvalue of the transition probability.
  - For P it takes the form $1-\left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}$
  - For $\tilde{P}$ it takes the form $1-\left(\sup_{(\theta,w)\in\Theta\times\mathsf{W}}\frac{\pi(\theta)}{q(\theta)}w\right)^{-1}$.
- If $\sup_{w\in\mathsf{W}} w$ is independent of θ, the second largest eigenvalue is exactly $1-\left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}(\sup_{w\in\mathsf{W}} w)^{-1}$ which is larger than $1-\left(\sup_{\theta\in\Theta}\frac{\pi(\theta)}{q(\theta)}\right)^{-1}$ - **even for an arbitrarily small variance!**

Counter-example

- Consider the independent MH algorithm, in the discrete case. It is possible to characterise exactly the second largest eigenvalue of the transition probability.
  - For P it takes the form $1 - \left(\sup_{\theta\in\Theta} \frac{\pi(\theta)}{q(\theta)}\right)^{-1}$
  - For $\tilde{P}$ it takes the form $1 - \left(\sup_{(\theta,w)\in\Theta\times\mathsf{W}} \frac{\pi(\theta)}{q(\theta)} w\right)^{-1}$.
- If $\sup_{w\in\mathsf{W}} w$ is independent of θ, the second largest eigenvalue is exactly $1 - \left(\sup_{\theta\in\Theta} \frac{\pi(\theta)}{q(\theta)}\right)^{-1} (\sup_{w\in\mathsf{W}} w)^{-1}$ which is larger than $1 - \left(\sup_{\theta\in\Theta} \frac{\pi(\theta)}{q(\theta)}\right)^{-1}$ - **even for an arbitrarily small variance!**

Un petit détour (I)

- Before turning to the study of pseudo-marginal algorithms, we show on one of their cousins why the convex order may be useful.
- Consider the following algorithm with transition

$$\mathring{P}(x;\mathrm{d}y) = q(x,\mathrm{d}y)\int_{\mathsf{W}} Q_{xy}(\mathrm{d}\varpi)\min\left\{1, r(x,y)\varpi\right\} + \delta_x(\mathrm{d}y)\mathring{\rho}(x)$$

  where $r(x,y)$ is the acceptance ratio of P.

- It can be shown that the condition $Q_{xy}(\mathrm{d}\varpi)\times\varpi = Q_{yx}(\mathrm{d}(\varpi^{-1}))$ for any $x,y\in\mathsf{X}$ ensures that it is reversible with respect to π.
- For example, for any $a>0$ the distribution $Q(\mathrm{d}w) = \left[\delta_a\left(\mathrm{d}w\right) + a\,\delta_{a^{-1}}\left(\mathrm{d}w\right)\right]/(1+a)$ satisfies this condition, but this is also the case for the log-normal distribution...
- These algorithms are exact approximations of MCMC, but here it is the acceptance probability which is directly approximated.

Un petit détour (I)

- Before turning to the study of pseudo-marginal algorithms, we show on one of their cousins why the convex order may be useful.
- Consider the following algorithm with transition

$$\mathring{P}(x; \mathrm{d}y) = q(x, \mathrm{d}y) \int_{\mathsf{W}} Q_{xy}(\mathrm{d}\varpi) \min\{1, r(x,y)\varpi\} + \delta_x(\mathrm{d}y)\mathring{\rho}(x)$$

where $r(x, y)$ is the acceptance ratio of P.

- It can be shown that the condition $Q_{xy}(\mathrm{d}\varpi) \times \varpi = Q_{yx}(\mathrm{d}(\varpi^{-1}))$ for any $x, y \in \mathsf{X}$ ensures that it is reversible with respect to π.
- For example, for any $a > 0$ the distribution $Q(\mathrm{d}w) = [\delta_a(\mathrm{d}w) + a\,\delta_{a^{-1}}(\mathrm{d}w)]/(1+a)$ satisfies this condition, but this is also the case for the log-normal distribution...
- These algorithms are exact approximations of MCMC, but here it is the acceptance probability which is directly approximated.

Un petit détour (I)

- Before turning to the study of pseudo-marginal algorithms, we show on one of their cousins why the convex order may be useful.
- Consider the following algorithm with transition

$$\mathring{P}(x; \mathrm{d}y) = q(x, \mathrm{d}y) \int_{\mathsf{W}} Q_{xy}(\mathrm{d}\varpi) \min\{1, r(x,y)\varpi\} + \delta_x(\mathrm{d}y)\mathring{\rho}(x)$$

where $r(x, y)$ is the acceptance ratio of P.

- It can be shown that the condition $Q_{xy}(\mathrm{d}\varpi) \times \varpi = Q_{yx}(\mathrm{d}(\varpi^{-1}))$ for any $x, y \in \mathsf{X}$ ensures that it is reversible with respect to π.
- For example, for any $a > 0$ the distribution $Q(\mathrm{d}w) = [\delta_a(\mathrm{d}w) + a\,\delta_{a^{-1}}(\mathrm{d}w)]/(1 + a)$ satisfies this condition, but this is also the case for the log-normal distribution...
- These algorithms are exact approximations of MCMC, but here it is the acceptance probability which is directly approximated.

Un petit détour (I)

- Before turning to the study of pseudo-marginal algorithms, we show on one of their cousins why the convex order may be useful.
- Consider the following algorithm with transition

$$\mathring{P}(x; \mathrm{d}y) = q(x, \mathrm{d}y) \int_{\mathsf{W}} Q_{xy}(\mathrm{d}\varpi) \min\{1, r(x,y)\varpi\} + \delta_x(\mathrm{d}y)\mathring{\rho}(x)$$

  where $r(x, y)$ is the acceptance ratio of P.

- It can be shown that the condition $Q_{xy}(\mathrm{d}\varpi) \times \varpi = Q_{yx}(\mathrm{d}(\varpi^{-1}))$ for any $x, y \in \mathsf{X}$ ensures that it is reversible with respect to π.
- For example, for any $a > 0$ the distribution $Q(\mathrm{d}w) = [\delta_a(\mathrm{d}w) + a\,\delta_{a^{-1}}(\mathrm{d}w)]/(1 + a)$ satisfies this condition, but this is also the case for the log-normal distribution...
- These algorithms are exact approximations of MCMC, but here it is the acceptance probability which is directly approximated.

Un petit détour (I)

- Before turning to the study of pseudo-marginal algorithms, we show on one of their cousins why the convex order may be useful.
- Consider the following algorithm with transition

$$\mathring{P}(x; \mathrm{d}y) = q(x, \mathrm{d}y) \int_{\mathsf{W}} Q_{xy}(\mathrm{d}\varpi) \min\left\{1, r(x,y)\varpi\right\} + \delta_x(\mathrm{d}y)\mathring{\rho}(x)$$

  where $r(x, y)$ is the acceptance ratio of P.

- It can be shown that the condition $Q_{xy}(\mathrm{d}\varpi) \times \varpi = Q_{yx}(\mathrm{d}(\varpi^{-1}))$ for any $x, y \in \mathsf{X}$ ensures that it is reversible with respect to π.
- For example, for any $a > 0$ the distribution $Q(\mathrm{d}w) = \left[\delta_a\left(\mathrm{d}w\right) + a\,\delta_{a^{-1}}\left(\mathrm{d}w\right)\right]/(1+a)$ satisfies this condition, but this is also the case for the log-normal distribution...
- These algorithms are exact approximations of MCMC, but here it is the acceptance probability which is directly approximated.

A small detour (II)

- Now compare

$$\mathring{P}^{(i)}(x; \mathrm{d}y) = q(x, \mathrm{d}y) \int_{\mathsf{W}} Q_{xy}^{(i)}(\mathrm{d}\varpi) \min\{1, r(x,y)\varpi\} + \delta_x(\mathrm{d}y)\mathring{\rho}^{(i)}(x)$$

- These define Markov chains $\{\mathring{X}^{(1)}\}$ and $\{\mathring{X}^{(2)}\}$ with common invariant distribution (Peskun!).
- In contrast with pseudo-marginal algorithms for which the Markov chain involves the weight sequence, i.e. $\{X^{(1)}, W^{(1)}\}$.
- If we have for any $x, y \in \mathsf{X}^2$ that $\bar{W}_{xy}^{(1)} \leq_{cx} \bar{W}_{xy}^{(2)}$ then, noting that $u \mapsto -\min\{1, u\}$ is convex,

$$\int_{\mathsf{W}} Q_{xy}^{(2)}(\mathrm{d}\varpi_2) \min\{1, r(x,y)\varpi_2\} \leq \int_{\mathsf{W}} Q_{xy}^{(1)}(\mathrm{d}\varpi_1) \min\{1, r(x,y)\varpi_1\}.$$

- This therefore allows us to apply Peskun's result directly and conclude that $\mathrm{var}(f, \mathring{P}_2) \geq \mathrm{var}(f, \mathring{P}_1)$.

A small detour (II)

- Now compare

$$\mathring{P}^{(i)}(x;\mathrm{d}y) = q(x,\mathrm{d}y)\int_{\mathsf{W}} Q_{xy}^{(i)}(\mathrm{d}\varpi)\min\{1, r(x,y)\varpi\} + \delta_x(\mathrm{d}y)\mathring{\rho}^{(i)}(x)$$

- These define Markov chains $\{\mathring{X}^{(1)}\}$ and $\{\mathring{X}^{(2)}\}$ with common invariant distribution (Peskun!).
- In contrast with pseudo-marginal algorithms for which the Markov chain involves the weight sequence, i.e. $\{X^{(1)}, W^{(1)}\}$.
- If we have for any $x, y \in \mathsf{X}^2$ that $\bar{W}_{xy}^{(1)} \leq_{cx} \bar{W}_{xy}^{(2)}$ then, noting that $u \mapsto -\min\{1, u\}$ is convex,

$$\int_{\mathsf{W}} Q_{xy}^{(2)}(\mathrm{d}\varpi_2)\min\{1, r(x,y)\varpi_2\} \leq \int_{\mathsf{W}} Q_{xy}^{(1)}(\mathrm{d}\varpi_1)\min\{1, r(x,y)\varpi_1\}.$$

- This therefore allows us to apply Peskun's result directly and conclude that $\mathrm{var}(f, \mathring{P}_2) \geq \mathrm{var}(f, \mathring{P}_1)$.

A small detour (II)

- Now compare

$$\mathring{P}^{(i)}(x; \mathrm{d}y) = q(x, \mathrm{d}y) \int_{\mathsf{W}} Q_{xy}^{(i)}(\mathrm{d}\varpi) \min\{1, r(x, y)\varpi\} + \delta_x(\mathrm{d}y)\mathring{\rho}^{(i)}(x)$$

- These define Markov chains $\{\mathring{X}^{(1)}\}$ and $\{\mathring{X}^{(2)}\}$ with common invariant distribution (Peskun!).
- In contrast with pseudo-marginal algorithms for which the Markov chain involves the weight sequence, i.e. $\{X^{(1)}, W^{(1)}\}$.
- If we have for any $x, y \in \mathsf{X}^2$ that $\bar{W}_{xy}^{(1)} \leq_{cx} \bar{W}_{xy}^{(2)}$ then, noting that $u \mapsto -\min\{1, u\}$ is convex,

$$\int_{\mathsf{W}} Q_{xy}^{(2)}(\mathrm{d}\varpi_2) \min\{1, r(x, y)\varpi_2\} \leq \int_{\mathsf{W}} Q_{xy}^{(1)}(\mathrm{d}\varpi_1) \min\{1, r(x, y)\varpi_1\}.$$

- This therefore allows us to apply Peskun's result directly and conclude that $\mathrm{var}(f, \mathring{P}_2) \geq \mathrm{var}(f, \mathring{P}_1)$.

A small detour (II)

- Now compare

$$\mathring{P}^{(i)}(x;\mathrm{d}y) = q(x,\mathrm{d}y)\int_{\mathsf{W}} Q_{xy}^{(i)}(\mathrm{d}\varpi)\min\left\{1, r(x,y)\varpi\right\} + \delta_x(\mathrm{d}y)\mathring{\rho}^{(i)}(x)$$

- These define Markov chains $\{\mathring{X}^{(1)}\}$ and $\{\mathring{X}^{(2)}\}$ with common invariant distribution (Peskun!).
- In contrast with pseudo-marginal algorithms for which the Markov chain involves the weight sequence, i.e. $\{X^{(1)}, W^{(1)}\}$.
- If we have for any $x, y \in \mathsf{X}^2$ that $\bar{W}_{xy}^{(1)} \leq_{cx} \bar{W}_{xy}^{(2)}$ then, noting that $u \mapsto -\min\{1, u\}$ is convex,

$$\int_{\mathsf{W}} Q_{xy}^{(2)}(\mathrm{d}\varpi_2)\min\left\{1, r(x,y)\varpi_2\right\} \leq \int_{\mathsf{W}} Q_{xy}^{(1)}(\mathrm{d}\varpi_1)\min\left\{1, r(x,y)\varpi_1\right\}.$$

- This therefore allows us to apply Peskun's result directly and conclude that $\mathrm{var}(f, \mathring{P}_2) \geq \mathrm{var}(f, \mathring{P}_1)$.

A small detour (II)

- Now compare

$$\mathring{P}^{(i)}(x; \mathrm{d}y) = q(x, \mathrm{d}y) \int_{\mathsf{W}} Q_{xy}^{(i)}(\mathrm{d}\varpi) \min\{1, r(x, y)\varpi\} + \delta_x(\mathrm{d}y)\mathring{\rho}^{(i)}(x)$$

- These define Markov chains $\{\mathring{X}^{(1)}\}$ and $\{\mathring{X}^{(2)}\}$ with common invariant distribution (Peskun!).
- In contrast with pseudo-marginal algorithms for which the Markov chain involves the weight sequence, i.e. $\{X^{(1)}, W^{(1)}\}$.
- If we have for any $x, y \in \mathsf{X}^2$ that $\bar{W}_{xy}^{(1)} \leq_{cx} \bar{W}_{xy}^{(2)}$ then, noting that $u \mapsto -\min\{1, u\}$ is convex,

$$\int_{\mathsf{W}} Q_{xy}^{(2)}(\mathrm{d}\varpi_2) \min\{1, r(x, y)\varpi_2\} \leq \int_{\mathsf{W}} Q_{xy}^{(1)}(\mathrm{d}\varpi_1) \min\{1, r(x, y)\varpi_1\}.$$

- This therefore allows us to apply Peskun's result directly and conclude that $\mathrm{var}(f, \mathring{P}_2) \geq \mathrm{var}(f, \mathring{P}_1)$.

Extremal distributions (III)

When the interval has infinite support, one can constrain the problem by e.g. imposing a variance on the class of distributions, $\mathscr{P}(\mu, \sigma^2, [0, \infty))$ for $\sigma^2 < \infty$

Theorem

Let $\sigma_x^2 : \mathsf{X} \to [0, \infty)$. Consider the class of pseudo marginal algorithms $\tilde{P}$ such that for any $x \in \mathsf{X}$ the weight distribution Q_x is such that $\mathscr{P}(1, \sigma_x^2, [0, \infty))$. Then for any $f \in L^2(\mathsf{X}, \pi)$,

$$\mathrm{var}\left(P, f\right) \leq \mathrm{var}\left(\tilde{P}, f\right) \leq \mathrm{var}\left(\tilde{P}_{\max}, f\right) \quad ,$$

where for any $x \in \mathsf{X}$

$$Q_x^{\max}\left(W \leq t; \sigma_x^2\right) := \begin{cases} 0 & \text{for } t \leq 0 \\ \frac{\sigma_x^2}{1+\sigma_x^2} & \text{for } 0 \leq t \leq (\sigma_x^2 + 1)/2 \\ \frac{1}{2} + \frac{1}{2}\frac{t-1}{\sqrt{\sigma_x^2 + (t-1)^2}} & \text{for } t \geq (\sigma_x^2 + 1)/2 \end{cases}$$