Reversible proposal MCMC with heavy-tailed target distributions

Kengo Kamatani

September 2018

Osaka University Supported by CREST JST JPMJCR14D7 with travel support by ISM, NUS

Metrpolis-Hastings (MH): any Markov kernel $Q \rightsquigarrow \Pi$-reversible kernel P.
Q: Proposal kernel, Π: Target probability distribution.

Metrpolis-Hastings (MH): any Markov kernel $Q \rightsquigarrow \Pi$-reversible kernel P.
Q: Proposal kernel, Π: Target probability distribution.

Π-reversible: $\mathcal{L}_\Pi(X_0, X_1, \ldots, X_M) = \mathcal{L}_\Pi(X_M, X_{M-1}, \ldots, X_0)$.

Metrpolis-Hastings (MH): any Markov kernel $Q \rightsquigarrow \Pi$-reversible kernel P.
Q: Proposal kernel, Π: Target probability distribution.

Π-reversible: $\mathcal{L}_{\Pi}(X_0, X_1, \ldots, X_M) = \mathcal{L}_{\Pi}(X_M, X_{M-1}, \ldots, X_0)$.

Reversible proposal MH: $\tilde{\Pi}$-reversible kernel $Q \rightsquigarrow \Pi$-reversible kernel P.

Metrpolis-Hastings (MH): any Markov kernel $Q \rightsquigarrow \Pi$-reversible kernel P.
Q: Proposal kernel, Π: Target probability distribution.

Π-reversible: $\mathcal{L}_{\Pi}(X_0, X_1, \ldots, X_M) = \mathcal{L}_{\Pi}(X_M, X_{M-1}, \ldots, X_0)$.

Reversible proposal MH: $\tilde{\Pi}$-reversible kernel $Q \rightsquigarrow \Pi$-reversible kernel P.

RWM, pCN, Independent-MH $\in$ RMH.

Metrpolis-Hastings (MH): any Markov kernel $Q \rightsquigarrow \Pi$-reversible kernel P.
Q: Proposal kernel, Π: Target probability distribution.

Π-reversible: $\mathcal{L}_\Pi(X_0, X_1, \ldots, X_M) = \mathcal{L}_\Pi(X_M, X_{M-1}, \ldots, X_0)$.

Reversible proposal MH: $\tilde{\Pi}$-reversible kernel $Q \rightsquigarrow \Pi$-reversible kernel P.

RWM, pCN, Independent-MH $\in$ RMH.

We study RMH for heavy-tailed target distributions.

Abstract

In this talk, we will discuss Markov chain Monte Carlo methods for heavy-tailed target probability distributions, based on a reversible proposal transition kernel. We will study the dimensionality effect using the high-dimensional asymptotic analysis of Roberts, Gelman and Gilks. We also study ergodic properties for heavy-tailed target distributions.

Abstract

In this talk, we will discuss Markov chain Monte Carlo methods for heavy-tailed target probability distributions, based on a reversible proposal transition kernel. We will study the dimensionality effect using the high-dimensional asymptotic analysis of Roberts, Gelman and Gilks. We also study ergodic properties for heavy-tailed target distributions.

This talk is mainly from K. JAP 17 and K. Bernoulli 18.

High-dimensional asymptotics

- HDA
- Heavy-tail + HDA
- pCN/MpCN

Ergodicity

Future works

Light-tail + HDA (Roberts, Gelman, Gilks 97)

Random-walk Metropolis (RWM) algorithm

Light-tail + HDA (Roberts, Gelman, Gilks 97)

Random-walk Metropolis (RWM) algorithm

$$X^{d*}(m) = X^d(m-1) \ + \ W^d_m, \ W^d_m \sim N_d(0, \ {\sigma_d}^2 \ I_d)$$

$$X^d(m) = \begin{cases} X^{d*}(m) & \text{if } U^d_m \leq \alpha(X^d(m-1), X^{d*}(m)) \\ X^d(m-1) & \text{if } U^d_m > \alpha(X^d(m-1), X^{d*}(m)) \end{cases}$$

where $\Pi_d(\mathrm{d}x) = \pi_d(x)\mathrm{d}x$ is a prob meas on $\mathbb{R}^d$ and

$$\alpha(x, y) = \min\left\{1, \frac{\pi_d(y)}{\pi_d(x)}\right\}.$$

Light-tail + HDA (Roberts, Gelman, Gilks 97)

Random-walk Metropolis (RWM) algorithm

$$X^{d*}(m) = X^d(m-1) \; + \; W_m^d, \; W_m^d \sim N_d(0, \; {\sigma_d}^2 \; I_d)$$

$$X^d(m) = \begin{cases} X^{d*}(m) & \text{if } U_m^d \le \alpha(X^d(m-1), X^{d*}(m)) \\ X^d(m-1) & \text{if } U_m^d > \alpha(X^d(m-1), X^{d*}(m)) \end{cases}$$

where $\Pi_d(\mathrm{d}x) = \pi_d(x)\mathrm{d}x$ is a prob meas on $\mathbb{R}^d$ and

$$\alpha(x, y) = \min\left\{1, \frac{\pi_d(y)}{\pi_d(x)}\right\}.$$

Product target

$$\Pi_d = F^{\otimes d}$$

where $F(\mathrm{d}x) = f(x)\mathrm{d}x$ is a (regular) prob meas on $\mathbb{R}$.

Light-tail + HDA (Roberts, Gelman, Gilks 97)

Random-walk Metropolis (RWM) algorithm

$$X^{d*}(m) = X^d(m-1) \ + \ W_m^d, \ W_m^d \sim N_d(0, \ {\sigma_d}^2 \ I_d)$$

$$X^d(m) = \begin{cases} X^{d*}(m) & \text{if } U_m^d \leq \alpha(X^d(m-1), X^{d*}(m)) \\ X^d(m-1) & \text{if } U_m^d > \alpha(X^d(m-1), X^{d*}(m)) \end{cases}$$

where $\Pi_d(\mathrm{d}x) = \pi_d(x)\mathrm{d}x$ is a prob meas on $\mathbb{R}^d$ and

$$\alpha(x,y) = \min\left\{1, \frac{\pi_d(y)}{\pi_d(x)}\right\}.$$

Product target

$$\Pi_d = F^{\otimes d}$$

where $F(\mathrm{d}x) = f(x)\mathrm{d}x$ is a (regular) prob meas on $\mathbb{R}$.

Time scaling / State space scaling / Proposal scaling

$$Y^d(t) = X_1^d([dt]), \ \sigma_d^2 \ = \ l^2/d, \ l \in \mathbb{R}_+.$$

Light-tail + HDA (Roberts, Gelman, Gilks 97)

Random-walk Metropolis (RWM) algorithm

$$X^{d*}(m) = X^d(m-1) \; + \; W_m^d, \; W_m^d \sim N_d(0, \; {\sigma_d}^2 \; I_d)$$

$$X^d(m) = \begin{cases} X^{d*}(m) & \text{if } U_m^d \leq \alpha(X^d(m-1), X^{d*}(m)) \\ X^d(m-1) & \text{if } U_m^d > \alpha(X^d(m-1), X^{d*}(m)) \end{cases}$$

where $\Pi_d(\mathrm{d}x) = \pi_d(x)\mathrm{d}x$ is a prob meas on $\mathbb{R}^d$ and

$$\alpha(x, y) = \min\left\{1, \frac{\pi_d(y)}{\pi_d(x)}\right\}.$$

Product target

$$\Pi_d = F^{\otimes d}$$

where $F(\mathrm{d}x) = f(x)\mathrm{d}x$ is a (regular) prob meas on $\mathbb{R}$.

Time scaling / State space scaling / Proposal scaling

$$Y^d(t) = X_1^d([dt]), \; \sigma_d^2 \; = \; l^2/d, \; l \in \mathbb{R}_+.$$

Why there is a dimensionality effect?

Light-tail + HDA (Roberts, Gelman, Gilks 97)

Random-walk Metropolis (RWM) algorithm

$$X^{d*}(m) = X^d(m-1) \ + \ W_m^d, \ W_m^d \sim N_d(0, \ {\sigma_d}^2 \ I_d)$$

$$X^d(m) = \begin{cases} X^{d*}(m) & \text{if } U_m^d \leq \alpha(X^d(m-1), X^{d*}(m)) \\ X^d(m-1) & \text{if } U_m^d > \alpha(X^d(m-1), X^{d*}(m)) \end{cases}$$

where $\Pi_d(\mathrm{d}x) = \pi_d(x)\mathrm{d}x$ is a prob meas on $\mathbb{R}^d$ and

$$\alpha(x,y) = \min\left\{1, \frac{\pi_d(y)}{\pi_d(x)}\right\}.$$

Product target

$$\Pi_d = F^{\otimes d}$$

where $F(\mathrm{d}x) = f(x)\mathrm{d}x$ is a (regular) prob meas on $\mathbb{R}$.

Time scaling / State space scaling / Proposal scaling

$$Y^d(t) = X_1^d([dt]), \ \sigma_d^2 \ = \ l^2/d, \ l \in \mathbb{R}_+.$$

Why there is a dimensionality effect? *How about* $\sigma_d^2 = l^2/\sqrt{d}$*?*

The dimensionality is due to reversibility.

The dimensionality is due to reversibility.

$\sigma_d^2 = O(d^{-1+\epsilon}) \implies$ bigger jumps + small acceptance probability.

Notice

$$\mathbb{P}(\|X^d(0)\|^2 > \|X^d(1)\|^2) \;=\; \mathbb{P}(\|X^d(0)\|^2 < \|X^d(1)\|^2).$$

The dimensionality is due to reversibility.
$\sigma_d^2 = O(d^{-1+\epsilon}) \implies$ bigger jumps + small acceptance probability.

Notice

$$\mathbb{P}(\|X^d(0)\|^2 > \|X^d(1)\|^2) \;=\; \mathbb{P}(\|X^d(0)\|^2 < \|X^d(1)\|^2).$$

However, when $\sigma_d^2 = O(d^{-1+\epsilon})$,

$$\mathbb{P}(\|X^d(0)\|^2 > \|X^{d*}(1)\|^2) \;\ll\; \mathbb{P}(\|X^d(0)\|^2 < \|X^{d*}(1)\|^2),$$

since

$$\|X^{d*}(1)\|^2 = \|X^d(0) + W_1^d\|^2 = \|X^d(0)\|^2 + \langle X^d(0), W_1^d\rangle + \|W_1^d\|^2.$$

Thus, $X^{d*}(1)$ is likely to be rejected.

The dimensionality is due to reversibility.
$\sigma_d^2 = O(d^{-1+\epsilon}) \implies$ bigger jumps + small acceptance probability.

Notice

$$\mathbb{P}(\|X^d(0)\|^2 > \|X^d(1)\|^2) \ = \ \mathbb{P}(\|X^d(0)\|^2 < \|X^d(1)\|^2).$$

However, when $\sigma_d^2 = O(d^{-1+\epsilon})$,

$$\mathbb{P}(\|X^d(0)\|^2 > \|X^{d*}(1)\|^2) \ \ll \ \mathbb{P}(\|X^d(0)\|^2 < \|X^{d*}(1)\|^2),$$

since

$$\|X^{d*}(1)\|^2 = \|X^d(0) + W_1^d\|^2 = \|X^d(0)\|^2 + \langle X^d(0), W_1^d \rangle + \|W_1^d\|^2.$$

Thus, $X^{d*}(1)$ is likely to be rejected.
We can show that

$$\mathbb{P}(X^d(0) = X^d(1) = \cdots = X^d(d^k)) = 1 - o(1)$$

for any $k \in \mathbb{N}$.

Theo [Roberts, Gelman, Gilks 97] $Y^d \ \rightarrow \ Y$ where

$$\mathrm{d}Y(t) = a(Y(t))\mathrm{d}t + b\,\mathrm{d}W_t,$$

with

$$a(y) \ = \ \frac{(\log f)'(y)}{2}\ h(l), \ \ b^2 \ = \ h(l)$$
$$h(l) \ = \ 2\ l^2\ \Phi\left(-\frac{l\sqrt{J}}{2}\right), \ J = \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x < \infty.$$

Theo [Roberts, Gelman, Gilks 97] $Y^d \to Y$ where

$$\mathrm{d}Y(t) = a(Y(t))\mathrm{d}t + b\,\mathrm{d}W_t,$$

with

$$a(y) = \frac{(\log f)'(y)}{2}\,h(l),\ b^2 = h(l)$$

$$h(l) = 2\,l^2\,\Phi\left(-\frac{l\sqrt{J}}{2}\right),\ J = \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x < \infty.$$

Convergence rate is d. /

Theo [Roberts, Gelman, Gilks 97] $Y^d \;\to\; Y$ where

$$\mathrm{d}Y(t) = a(Y(t))\mathrm{d}t + b\;\mathrm{d}W_t,$$

with

$$a(y) \;=\; \frac{(\log f)'(y)}{2}\; h(l),\;\; b^2 \;=\; h(l)$$
$$h(l) \;=\; 2\; l^2\; \Phi\left(-\frac{l\sqrt{J}}{2}\right),\;\; J = \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x < \infty.$$

Convergence rate is d. / Efficient when $h(l)$ is large. /

Theo [Roberts, Gelman, Gilks 97] $Y^d \to Y$ where

$$\mathrm{d}Y(t) = a(Y(t))\mathrm{d}t + b\,\mathrm{d}W_t,$$

with

$$a(y) = \frac{(\log f)'(y)}{2}\,h(l),\ b^2 = h(l)$$

$$h(l) = 2\,l^2\,\Phi\left(-\frac{l\sqrt{J}}{2}\right),\ J = \int\left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x < \infty.$$

Convergence rate is d. / Efficient when $h(l)$ is large. / $h(l)$ is maximized when the average acceptance probability is $0.234\cdots$.

Theo [Roberts, Gelman, Gilks 97] $Y^d \to Y$ where

$$\mathrm{d}Y(t) = a(Y(t))\mathrm{d}t + b\,\mathrm{d}W_t,$$

with

$$a(y) = \frac{(\log f)'(y)}{2}\, h(l),\; b^2 = h(l)$$

$$h(l) = 2\, l^2\, \Phi\left(-\frac{l\sqrt{J}}{2}\right),\; J = \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x < \infty.$$

Convergence rate is d. / Efficient when $h(l)$ is large. / $h(l)$ is maximized when the average acceptance probability is $0.234\cdots$.

How about heavy-tail case?

Heavy-tail + HDA

Heavy-tail

$$\forall s > 0 \int_{\mathbb{R}^d} \exp(s\|x\|)\Pi_d(\mathrm{d}x) = +\infty.$$

Heavy-tail + HDA

Heavy-tail

$$\forall s > 0 \ \int_{\mathbb{R}^d} \exp(s\|x\|)\Pi_d(\mathrm{d}x) = +\infty.$$

Mixed normal (MN)

$$\Pi_d(\mathrm{d}x) = \int_0^\infty (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right) Q(\mathrm{d}\sigma^2)\mathrm{d}x.$$

MN $\supset$ Student-t, Stable etc.

Asymptotics: Fix Q, $d \to \infty$.

Asymptotics: Fix Q, $d \to \infty$.

Assumption: $Q(\mathrm{d}y) = q(y)\mathrm{d}y$, $q(y) > 0$ continuously differentiable, q, q' vanishes at $+0$ and $+\infty$.

Time scaling / State space scaling / Proposal scaling

Asymptotics: Fix Q, $d \to \infty$.

Assumption: $Q(\mathrm{d}y) = q(y)\mathrm{d}y$, $q(y) > 0$ continuously differentiable, q, q' vanishes at $+0$ and $+\infty$.

Time scaling / State space scaling / Proposal scaling

$Y^d(t) = X_1^d([d\ t])$, $Z^d(t) = \log\left(\|X^d([d^2\ t])\|/d^{1/2}\right)$, $\sigma_d^2 = l^2/d$, $l \in \mathbb{R}_+$.

Theo K.

Faster rate: $Y^d \;\to\; Y$ where

$$\mathrm{d}Y(t) = a(Y(t), Z_0)\mathrm{d}t + b(Z_0)\;\mathrm{d}W_t,$$

with $Z_0 \sim \tilde{Q}$: $\tilde{Q}(\mathrm{d}y) = \tilde{q}(y)\mathrm{d}y$ is the log transform of Q and

$$a(y,z) \;=\; -\,e^{-2z}\mu_0(e^{-z})y/2,\;\; b^2(z) \;=\; \mu_0(e^{-z}).$$

Theo K.

Faster rate: $Y^d \to Y$ where

$$\mathrm{d}Y(t) = a(Y(t), Z_0)\mathrm{d}t + b(Z_0)\,\mathrm{d}W_t,$$

with $Z_0 \sim \tilde{Q}$: $\tilde{Q}(\mathrm{d}y) = \tilde{q}(y)\mathrm{d}y$ is the log transform of Q and

$$a(y,z) \;=\; -e^{-2z}\mu_0(e^{-z})y/2,\;\; b^2(z) \;=\; \mu_0(e^{-z}).$$

Here,

$$\mu_k(\sigma) := \mathbb{E}[|F|^k e^{-F^+}],\;\; F \sim N(\sigma^2/2, \sigma^2)$$

Theo K.

Faster rate: $Y^d \;\to\; Y$ where

$$\mathrm{d}Y(t) = a(Y(t), Z_0)\mathrm{d}t + b(Z_0)\;\mathrm{d}W_t,$$

with $Z_0 \sim \tilde{Q}$: $\tilde{Q}(\mathrm{d}y) = \tilde{q}(y)\mathrm{d}y$ is the log transform of Q and

$$a(y,z) \;=\; -\,e^{-2z}\mu_0(e^{-z})y/2,\;\; b^2(z) \;=\; \mu_0(e^{-z}).$$

Here,

$$\mu_k(\sigma) := \mathbb{E}[|F|^k e^{-F^+}],\;\; F \sim N(\sigma^2/2, \sigma^2)$$

Slower rate: $Z^d \;\to\; Z$ where

$$\mathrm{d}Z(t) = \alpha(Z(t))\mathrm{d}t + \beta(Z(t))\;\mathrm{d}B_t,$$

with

$$\alpha(z) \;=\; (\log\tilde{q})'(z)\mu_2(e^{-z})/2 - e^{-z}\mu_2'(e^{-z})/2,\;\; \beta^2(z) \;=\; \mu_2(e^{-z}).$$

Y is not ergodic, but Z is ergodic. Convergence rate is d^2.

LLN:

$$\frac{M_d}{d^2} \to \infty \implies \frac{1}{M_d} \sum_{m=1}^{M_d} f(X_1^d(m)) - \int f(x_1) \Pi_d(\mathrm{d}x) = o_{\mathbb{P}}(1).$$

Y is not ergodic, but Z is ergodic. Convergence rate is d^2.

LLN:

$$\frac{M_d}{d^2} \to \infty \implies \frac{1}{M_d} \sum_{m=1}^{M_d} f(X_1^d(m)) - \int f(x_1)\Pi_d(\mathrm{d}x) = o_{\mathbb{P}}(1).$$

Key technique: (Classical Stroock-Varadhan's semimartingale characteristics convergence) + Stein's method

$$h(x) \text{ m'ble s.t. } N|h| < \infty \Longrightarrow \exists f \text{ s.t. } h(x) - Nh = f'(x) - xf(x)$$

where $Nh = \int h(x)\phi(x)\mathrm{d}x$.

Y is not ergodic, but Z is ergodic. Convergence rate is d^2.

LLN:

$$\frac{M_d}{d^2} \to \infty \implies \frac{1}{M_d}\sum_{m=1}^{M_d} f(X_1^d(m)) - \int f(x_1)\Pi_d(\mathrm{d}x) = o_{\mathbb{P}}(1).$$

Key technique: (Classical Stroock-Varadhan's semimartingale characteristics convergence) + Stein's method

$$h(x) \text{ m'ble s.t. } N|h| < \infty \implies \exists f \text{ s.t. } h(x) - Nh = f'(x) - xf(x)$$

where $Nh = \int h(x)\phi(x)\mathrm{d}x$. For example, we have a bound

$$\|\mathcal{L}(F) - N(0,1)\|_{\mathrm{TV}} \le \sup_{f\in\mathbf{F}} |\mathbb{E}[f'(F)] - \mathbb{E}[F\ f(F)]|.$$

where $\mathbf{F} = \{f; \|f\|_\infty \le \sqrt{\pi/2},\ \|f'\|_\infty \le 2\}$.

Y is not ergodic, but Z is ergodic. Convergence rate is d^2.

LLN:

$$\frac{M_d}{d^2} \to \infty \implies \frac{1}{M_d} \sum_{m=1}^{M_d} f(X_1^d(m)) - \int f(x_1) \Pi_d(\mathrm{d}x) = o_{\mathbb{P}}(1).$$

Key technique: (Classical Stroock-Varadhan's semimartingale characteristics convergence) + Stein's method

$$h(x) \text{ m'ble s.t. } N|h| < \infty \implies \exists f \text{ s.t. } h(x) - Nh = f'(x) - xf(x)$$

where $Nh = \int h(x)\phi(x)\mathrm{d}x$. For example, we have a bound

$$\|\mathcal{L}(F) - N(0,1)\|_{\mathrm{TV}} \leq \sup_{f \in \mathbf{F}} |\mathbb{E}[f'(F)] - \mathbb{E}[F\ f(F)]|.$$

where $\mathbf{F} = \{f; \|f\|_\infty \leq \sqrt{\pi/2},\ \|f'\|_\infty \leq 2\}$.

Better MCMC?

pCN/MpCN

Pre-conditioned Crank-Nicolson (pCN) Neal 99, Beskos et al 08

$$X^{d*}(m) = \sqrt{\rho}\, X^d(m-1) \; + \; \sqrt{1-\rho}\, \sigma \; W_m^d, \;\; W_m^d \sim N_d(0, I_d)$$

Acceptance probability:

$$\alpha(x,y) = \min\left\{1, \frac{\pi_d(y)\exp(-\|x\|^2/2\sigma^2)}{\pi_d(x)\exp(-\|y\|^2/2\sigma^2)}\right\}.$$

pCN/MpCN

Pre-conditioned Crank-Nicolson (pCN) Neal 99, Beskos et al 08

$$X^{d*}(m) = \sqrt{\rho}\, X^d(m-1) \; + \; \sqrt{1-\rho}\, \sigma \; W^d_m, \;\; W^d_m \sim N_d(0, I_d)$$

Acceptance probability:

$$\alpha(x, y) = \min\left\{1, \frac{\pi_d(y)\exp(-\|x\|^2/2\sigma^2)}{\pi_d(x)\exp(-\|y\|^2/2\sigma^2)}\right\}.$$

Works well only for light-tail target since the proposal is reversible to $\tilde{\Pi} = N_d(0, \sigma^2 I_d)$.

pCN/MpCN

Pre-conditioned Crank-Nicolson (pCN) Neal 99, Beskos et al 08

$$X^{d*}(m) = \sqrt{\rho}\, X^d(m-1) \;+\; \sqrt{1-\rho}\,\sigma\; W_m^d,\;\; W_m^d \sim N_d(0, I_d)$$

Acceptance probability:

$$\alpha(x,y) = \min\left\{1, \frac{\pi_d(y)\exp(-\|x\|^2/2\sigma^2)}{\pi_d(x)\exp(-\|y\|^2/2\sigma^2)}\right\}.$$

Works well only for light-tail target since the proposal is reversible to $\tilde{\Pi} = N_d(0, \sigma^2 I_d)$.

$\rightsquigarrow$ We can put prior on σ^2

pCN/MpCN

Pre-conditioned Crank-Nicolson (pCN) Neal 99, Beskos et al 08

$$X^{d*}(m) = \sqrt{\rho}\, X^d(m-1) \;+\; \sqrt{1-\rho}\,\sigma\, W_m^d,\;\; W_m^d \sim N_d(0, I_d)$$

Acceptance probability:

$$\alpha(x, y) = \min\left\{1, \frac{\pi_d(y)\exp(-\|x\|^2/2\sigma^2)}{\pi_d(x)\exp(-\|y\|^2/2\sigma^2)}\right\}.$$

Works well only for light-tail target since the proposal is reversible to $\tilde{\Pi} = N_d(0, \sigma^2 I_d)$.

$\rightsquigarrow$ We can put prior on σ^2 $\rightsquigarrow$ How about Jeffreys/Haar prior?

pCN/MpCN

Pre-conditioned Crank-Nicolson (pCN) Neal 99, Beskos et al 08

$$X^{d*}(m) = \sqrt{\rho}\, X^d(m-1) \; + \; \sqrt{1-\rho}\, \sigma \; W_m^d, \;\; W_m^d \sim N_d(0, I_d)$$

Acceptance probability:

$$\alpha(x, y) = \min\left\{1, \frac{\pi_d(y) \exp(-\|x\|^2/2\sigma^2)}{\pi_d(x) \exp(-\|y\|^2/2\sigma^2)}\right\}.$$

Works well only for light-tail target since the proposal is reversible to $\tilde{\Pi} = N_d(0, \sigma^2 I_d)$.

$\rightsquigarrow$ We can put prior on σ^2 $\rightsquigarrow$ How about Jeffreys/Haar prior?

$\rightsquigarrow$ MpCN algorithm.

pCN/MpCN

Pre-conditioned Crank-Nicolson (pCN) Neal 99, Beskos et al 08

$$X^{d*}(m) = \sqrt{\rho}\, X^d(m-1) \;+\; \sqrt{1-\rho}\,\sigma\; W_m^d,\;\; W_m^d \sim N_d(0, I_d)$$

Acceptance probability:

$$\alpha(x,y) = \min\left\{1, \frac{\pi_d(y)\exp(-\|x\|^2/2\sigma^2)}{\pi_d(x)\exp(-\|y\|^2/2\sigma^2)}\right\}.$$

Works well only for light-tail target since the proposal is reversible to $\tilde{\Pi} = N_d(0, \sigma^2 I_d)$.

$\rightsquigarrow$ We can put prior on σ^2 $\rightsquigarrow$ How about Jeffreys/Haar prior?

$\rightsquigarrow$ MpCN algorithm. $\rightsquigarrow$ $\tilde{\Pi} = \|x\|^{-d}\mathrm{d}x$.

Mixed pre-conditioned Crank-Nicolson (MpCN) K. 18 Bernoulli

$$X^{d*}(m) = \sqrt{\rho}\, X^d(m-1) \;+\; \sqrt{1-\rho} \;\; \frac{\|X^d(m-1)\|}{\|\tilde{W}^d_m\|} W^d_m$$

Acceptance probability:

$$\alpha(x,y) = \min\left\{1, \frac{\pi_d(y)\|x\|^{-d}}{\pi_d(x)\|y\|^{-d}}\right\}.$$

Mixed pre-conditioned Crank-Nicolson (MpCN) K. 18 Bernoulli

$$X^{d*}(m) = \sqrt{\rho}\, X^d(m-1) \; + \; \sqrt{1-\rho} \;\; \frac{\|X^d(m-1)\|}{\|\tilde{W}^d_m\|} W^d_m$$

Acceptance probability:

$$\alpha(x,y) = \min\left\{1, \frac{\pi_d(y)\|x\|^{-d}}{\pi_d(x)\|y\|^{-d}}\right\}.$$

Time scaling / State space scaling / Proposal scaling

$$Y^d(t) = X^d_1([t]),\; Z^d(t) = \log\left(\|X^d([d\;\,t])\|/d^{1/2}\right).$$

$Y^d \to Y$ and $Z^d \to Z$ for some diffusions Y and Z.

The convergence rate is d.

Key technique : Malliavin calculus and Stein's method such as

$$|\mathbb{E}[f'(F)] - \mathbb{E}[F\ f(F)]| \leq \|f'\|_\infty\ \mathbb{E}[|1 - \langle DF, -DL^{-1}F\rangle_H|].$$

The convergence rate is d.

Key technique : Malliavin calculus and Stein's method such as

$$|\mathbb{E}[f'(F)] - \mathbb{E}[F\ f(F)]| \leq \|f'\|_\infty\ \mathbb{E}[|1 - \langle DF, -DL^{-1}F\rangle_H|].$$

Recall that Hermite Polynomials are

$$H_0(x) = 1,\ H_1(x) = x,\ H_2(x) = \frac{x^2-1}{2}, \ldots.$$

Then the Malliavin derivative D operates

$$DH_n(W(e)) = H_{n-1}(W(e))e,\ e \in H,\ \|e\|_H = 1$$

and

$$L^{-1}H_n(W(e)) = -\frac{H_n(W(e))}{n}.$$

The convergence rate is d.

Key technique : Malliavin calculus and Stein's method such as

$$|\mathbb{E}[f'(F)] - \mathbb{E}[F\ f(F)]| \leq \|f'\|_\infty\ \mathbb{E}[|1 - \langle DF, -DL^{-1}F\rangle_H|].$$

Recall that Hermite Polynomials are

$$H_0(x) = 1,\ H_1(x) = x,\ H_2(x) = \frac{x^2 - 1}{2}, \dots.$$

Then the Malliavin derivative D operates

$$DH_n(W(e)) = H_{n-1}(W(e))e,\ e \in H,\ \|e\|_H = 1$$

and

$$L^{-1}H_n(W(e)) = -\frac{H_n(W(e))}{n}.$$

How about light-tail case?

The convergence rate is d.

Key technique : Malliavin calculus and Stein's method such as

$$|\mathbb{E}[f'(F)] - \mathbb{E}[F\ f(F)]| \le \|f'\|_\infty\ \mathbb{E}[|1 - \langle DF, -DL^{-1}F\rangle_H|].$$

Recall that Hermite Polynomials are

$$H_0(x) = 1,\ H_1(x) = x,\ H_2(x) = \frac{x^2-1}{2}, \dots.$$

Then the Malliavin derivative D operates

$$DH_n(W(e)) = H_{n-1}(W(e))e,\ e \in H,\ \|e\|_H = 1$$

and

$$L^{-1}H_n(W(e)) = -\frac{H_n(W(e))}{n}.$$

How about light-tail case? For light-tail case, the convergence rate is 1 (no time scaling).

Summary

Method	Light-tail	Heavy-tail
RW	d	d^2
pCN	1	$+\infty$
MpCN	1	d

High-dimensional asymptotics

Ergodicity

- Regular variation
- Ergodicity + Heavy-tail
- Ergodicity + Light-tail

Future works

Regular variation

Def $h : \mathbb{R}^d \to \mathbb{R}_+$ is regularly varying if

$$\frac{h(\ r\ x\ )}{h(\ r\ \mathbf{1}\ )} \longrightarrow_{r\to\infty} \Lambda(x)$$

locally uniformly on $\mathbb{R}^d \backslash \{0\}$, where $\Lambda(x) > 0$ is a continuous function and $\mathbf{1} = (1, 0, 0, \dots, 0)$.

Regular variation

Def $h : \mathbb{R}^d \to \mathbb{R}_+$ is regularly varying if

$$\frac{h(\ r\ x\ )}{h(\ r\ \mathbf{1}\ )} \longrightarrow_{r\to\infty} \Lambda(x)$$

locally uniformly on $\mathbb{R}^d \backslash \{0\}$, where $\Lambda(x) > 0$ is a continuous function and $\mathbf{1} = (1, 0, 0, \ldots, 0)$.

h is RV $\implies \exists \alpha \in \mathbb{R}$ (exponent of variation), $\Lambda(x) = \|x\|^{-\alpha}\Lambda(x/\|x\|)$.

π pdf is RV $\implies \alpha \geq d$, Π is heavy-tail (Karamata).

See Monographs such as Resnick 08 Springer, Bingham et al 89 Cambridge U. press.

Examples/Counter examples

RV $\supset$ Student-t, Stable, polynomial target [Jarner and Roberts 07]

RV $\not\ni \|x\|^{-d-2-\sin(\|x\|)}$.

Ergodicity + Heavy-tail

Heavy-tail 1 [Jarner, Tweedie 03]

π RV $\Longrightarrow$ RWM is not geometrically ergodic (GE).

Heavy-tail 2 [Fort, Moulines 03, Jarner, Roberts 07]

π RV + some additional conditions $\Longrightarrow$ RWM is polynomially ergodic.

Note [Johnson and Geyer 12]:
Heavy-tail $\Longrightarrow$ RWM + variable transform is GE.

Theo [K 17 JAP] π RV. Then

$$\text{MpCN is GE} \iff \exists s > 0,\ \int_{\mathbb{R}^d} \|x\|^s \Pi(\mathrm{d}x) < \infty.$$

Theo [K 17 JAP] π RV. Then

$$\text{MpCN is GE} \iff \exists s > 0, \int_{\mathbb{R}^d} \|x\|^s \Pi(\mathrm{d}x) < \infty.$$

Proof is standard: Drift condition $V(x) = (\pi(x)\|x\|^d)^{-1/2}$ for symmetric Λ + Compare spectral gaps by Dirichlet form $\mathcal{E}(f,g) = (f,(I-P)g)$.

Key property: The proposal kernel is a (logarithmic squared root) random-walk kernel under

$$x \mapsto \log \|x\|^2.$$

Theo [K 17 JAP] π RV. Then

$$\text{MpCN is GE} \iff \exists s > 0, \int_{\mathbb{R}^d} \|x\|^s \Pi(\mathrm{d}x) < \infty.$$

Proof is standard: Drift condition $V(x) = (\pi(x)\|x\|^d)^{-1/2}$ for symmetric Λ + Compare spectral gaps by Dirichlet form $\mathcal{E}(f, g) = (f, (I - P)g)$.

Key property: The proposal kernel is a (logarithmic squared root) random-walk kernel under

$$x \mapsto \log \|x\|^2.$$

How about light-tail case?

Ergodicity + Light-tail

Def $h : \mathbb{R}^d \to \mathbb{R}_+$ is rapidly varying if

$$\frac{h(\ r\ s\ x\ )}{h(\ r\ x\ )} \longrightarrow_{r\to+\infty} \begin{cases} 0 & \text{if } s>1 \\ +\infty & \text{if } s<1. \end{cases}$$

Ergodicity + Light-tail

Def $h : \mathbb{R}^d \to \mathbb{R}_+$ is rapidly varying if

$$\frac{h(\ r\ s\ x\ )}{h(\ r\ x\ )} \longrightarrow_{r \to +\infty} \begin{cases} 0 & \text{if } s > 1 \\ +\infty & \text{if } s < 1. \end{cases}$$

Rapidly varying $\supset$ Super exponential [Jarner, Hansen 00], exponential [Fort, Moulines 00].

Ergodicity + Light-tail

Def $h : \mathbb{R}^d \to \mathbb{R}_+$ is rapidly varying if

$$\frac{h(\ r\ s\ x\ )}{h(\ r\ x\ )} \longrightarrow_{r \to +\infty} \begin{cases} 0 & \text{if } s > 1 \\ +\infty & \text{if } s < 1. \end{cases}$$

Rapidly varying $\supset$ Super exponential [Jarner, Hansen 00], exponential [Fort, Moulines 00].

Note: (Heavy-tail) $\bigcap$ (Rapidly varying) $\neq \emptyset$. Ex. $\exp(-\|x\|^{\beta})$, $\beta < 1$.

Ergodicity + Light-tail

Def $h : \mathbb{R}^d \to \mathbb{R}_+$ is rapidly varying if

$$\frac{h(\ r\ s\ x\ )}{h(\ r\ x\ )} \longrightarrow_{r \to +\infty} \begin{cases} 0 & \text{if } s > 1 \\ +\infty & \text{if } s < 1. \end{cases}$$

Rapidly varying $\supset$ Super exponential [Jarner, Hansen 00], exponential [Fort, Moulines 00].

Note: (Heavy-tail) $\bigcap$ (Rapidly varying) $\neq\ \emptyset$. Ex. $\exp(-\|x\|^{\beta}),\ \beta < 1$.

Def π satisfies the curvature condition [CC] if

$$\limsup_{\|x\| \to \infty} \left\langle \frac{x}{\|x\|}, \frac{\nabla \log \pi(x)}{\|\nabla \log \pi(x)\|} \right\rangle < 0.$$

Theo [K 17 JAP] π rapidly varying + CC $\implies$ MpCN is GE.

Note: If π is rapidly varying, then MpCN is $GE \iff \sup P(x, \{x\}) = 1$.

Theo [K 17 JAP] π rapidly varying + CC $\Longrightarrow$ MpCN is GE.

Note: If π is rapidly varying, then MpCN is $GE \iff \sup P(x, \{x\}) = 1$.

As a linear operator in L^2

The Markov kernel P is a self-adjoint positive operator. Thus,

$$\text{GE} \iff \text{Variance bounding.}$$

Summary

Method	E-Rapid $e^{-\|x\|^\alpha}, \alpha > 1$	Rapid $e^{-\|x\|^\alpha}, \alpha \in (0,1)$	Regular $\|x\|^{-d-\delta}, \delta > 0$
RWM	OK	NO	NO
pCN	Conditional	NO	NO
MpCN	OK	OK	OK

Application of MpCN: K. Uchida 2016, K. Nogita and Uchida 2016.
Implemented in Yuima package.

High-dimensional asymptotics

Ergodicity

Future works

Future works

Adaptive MpCN [K. + Chimisov, Łatuszyński]

General version of MpCN

$$x^* \leftarrow \mu + \sqrt{\rho}\,(x-\mu) \;+\; \sqrt{1-\rho}\,\|\Sigma^{-1/2}(x-\mu)\|\;\frac{\Sigma^{1/2}w}{\|\tilde{w}\|}$$

$w, \tilde{w} \sim N_d(0, I_d)$.

$$\text{Parameter } \theta \;=\; (\rho, \mu, \Sigma).$$

Future works

Adaptive MpCN [K. + Chimisov, Łatuszyński]

General version of MpCN

$$x^* \leftarrow \mu + \sqrt{\rho}\,(x-\mu) \;+\; \sqrt{1-\rho}\,\|\Sigma^{-1/2}(x-\mu)\|\;\frac{\Sigma^{1/2}w}{\|\tilde{w}\|}$$

$w, \tilde{w} \sim N_d(0, I_d)$.

$$\text{Parameter } \theta \;=\; (\rho, \mu, \Sigma).$$

Choose θ (apart from ρ) so that

$$\|\Sigma^{-1/2}(x-\mu)\|^{-d}(\det \Sigma)^{-1/2}\mathrm{d}x \;\approx\; \Pi_d.$$

Future works

Adaptive MpCN [K. + Chimisov, Łatuszyński]

General version of MpCN

$$x^* \leftarrow \mu + \sqrt{\rho}\,(x-\mu) \;+\; \sqrt{1-\rho}\,\|\Sigma^{-1/2}(x-\mu)\|\;\frac{\Sigma^{1/2}w}{\|\tilde{w}\|}$$

$w, \tilde{w} \sim N_d(0, I_d)$.

$$\text{Parameter } \theta \;=\; (\rho, \mu, \Sigma).$$

Choose θ (apart from ρ) so that

$$\|\Sigma^{-1/2}(x-\mu)\|^{-d}(\det\Sigma)^{-1/2}\mathrm{d}x \;\approx\; \Pi_d.$$

Matrix MpCN [K. + Beskos]

Future works

Adaptive MpCN [K. + Chimisov, Łatuszyński]

General version of MpCN

$$x^* \leftarrow \mu + \sqrt{\rho}\,(x-\mu) \;+\; \sqrt{1-\rho}\,\|\Sigma^{-1/2}(x-\mu)\|\,\frac{\Sigma^{1/2}w}{\|\tilde{w}\|}$$

$w, \tilde{w} \sim N_d(0, I_d)$.

$$\text{Parameter } \theta \;=\; (\rho, \mu, \Sigma).$$

Choose θ (apart from ρ) so that

$$\|\Sigma^{-1/2}(x-\mu)\|^{-d}(\det\Sigma)^{-1/2}\mathrm{d}x \;\approx\; \Pi_d.$$

Matrix MpCN [K. + Beskos]

Gradient MpCN

Future works

Adaptive MpCN [K. + Chimisov, Łatuszyński]

General version of MpCN

$$x^* \leftarrow \mu + \sqrt{\rho}\,(x-\mu) \;+\; \sqrt{1-\rho}\,\|\Sigma^{-1/2}(x-\mu)\|\,\frac{\Sigma^{1/2}w}{\|\tilde{w}\|}$$

$w, \tilde{w} \sim N_d(0, I_d)$.

$$\text{Parameter } \theta \;=\; (\rho, \mu, \Sigma).$$

Choose θ (apart from ρ) so that

$$\|\Sigma^{-1/2}(x-\mu)\|^{-d}(\det\Sigma)^{-1/2}\mathrm{d}x \;\approx\; \Pi_d.$$

Matrix MpCN [K. + Beskos]

Gradient MpCN

PDMP + MpCN

Future works

Adaptive MpCN [K. + Chimisov, Łatuszyński]

General version of MpCN

$$x^* \leftarrow \mu + \sqrt{\rho}\,(x-\mu) \;+\; \sqrt{1-\rho}\,\|\Sigma^{-1/2}(x-\mu)\|\,\frac{\Sigma^{1/2}w}{\|\tilde{w}\|}$$

$w, \tilde{w} \sim N_d(0, I_d)$.

$$\text{Parameter } \theta \;=\; (\rho, \mu, \Sigma).$$

Choose θ (apart from ρ) so that

$$\|\Sigma^{-1/2}(x-\mu)\|^{-d}(\det\Sigma)^{-1/2}\mathrm{d}x \;\approx\; \Pi_d.$$

Matrix MpCN [K. + Beskos]

Gradient MpCN

PDMP + **MpCN**

Analysis of MpCN: Non-asymptotic / Sub-geometric ergodicity

Summary

In high-dimension, an MCMC may have several convergence rates.

Summary

In high-dimension, an MCMC may have several convergence rates.

The (slower) convergence rate of RWM is d for a light-tail target and d^2 for a heavy-tail target.

Summary

In high-dimension, an MCMC may have several convergence rates.

The (slower) convergence rate of RWM is d for a light-tail target and d^2 for a heavy-tail target.

On the other hand, the convergence rate of MpCN is 1 for a light-tail target and d for a heavy-tail target.

Summary

In high-dimension, an MCMC may have several convergence rates.

The (slower) convergence rate of RWM is d for a light-tail target and d^2 for a heavy-tail target.

On the other hand, the convergence rate of MpCN is 1 for a light-tail target and d for a heavy-tail target.

MpCN has a good ergodic property.