Explicit convergence bounds for Metropolis Markov chains

Isoperimetry, Spectral Gaps and Profiles

Sam Power

University of Bristol

2 December, 2022

Links & Acknowledgements

- Main paper today: arXiv 2211.08959;
- Related: arXiv 2208.05239
- All joint work with
  - Christophe Andrieu (Bristol)
  - Anthony Lee (Bristol)
  - Andi Q. Wang (Bristol $\rightsquigarrow$ Warwick)
- Funded by Bayes4Health EPSRC Grant

Setting: Task

- Task: simulation, integration in complex models
  - posterior inference
  - gradient estimation in intractable models
  - ...
- Approach: MCMC sampling

Random-Walk Metropolis

- Today: target is $\pi(x)$, $x \in E = \mathrm{R}^d$.

1. At x,
   1.1 Propose $x' \sim \mathcal{N}(x, \sigma^2 \cdot I_d)$.
   1.2 Evaluate $r(x, x') = \frac{\pi(x')}{\pi(x)}$.
   1.3 With probability $\min\{1, r(x, x')\}$, move to x'; otherwise, remain at x.

- Leaves π invariant, ergodic under mild conditions.

Some Notation

- $Q\left(x, \mathrm{d}x^{'}\right) = \mathcal{N}\left(\mathrm{d}x^{'}; x, \sigma^2 \cdot I_d\right)$.
- $\alpha(x, x^{'}) := \min\left\{1, r\left(x, x^{'}\right)\right\}$.
- $\alpha(x) = \int Q\left(x, \mathrm{d}x^{'}\right) \alpha(x, x^{'})$.
- The RWM kernel P is given by

$$
\begin{aligned}
P(x, \mathrm{d}x^{'}) = Q\left(x, \mathrm{d}x^{'}\right) \cdot \alpha(x, x^{'}) \\
+ (1 - \alpha(x)) \cdot \delta(x, \mathrm{d}x^{'})
\end{aligned}
$$

Convergence Analysis of RWM

- 'Soft' analysis: Exponential convergence $\overset{\approx}{\Longleftrightarrow}$ Lighter-than-Exponential Tails.
- 'Optimal scaling' analysis: control acceptance rate to optimise efficiency.
- 'Modern' analysis: log-concavity of target, 'optimIsation-style' proofs
  - (and variations on this)
- Today: synthesis of the above.

Main Results

- Suppose that
  - Target is $\pi(x) \propto \exp(-U(x))$,
  - U is m-strongly convex, L-smooth,
  - Write $\kappa = L/m$ (scale-free).
- Run RWM with $\sigma = \upsilon \cdot (L \cdot d)^{-1/2}$.
- Then,
  1. Acceptance rate satisfies $\alpha(x) \geqslant \alpha_0 := \frac{1}{2} \cdot \exp\left(-\frac{1}{2}\upsilon^2\right)$.
  2. Spectral gap satisfies $\gamma_P \geqslant c(\upsilon) \cdot (\kappa \cdot d)^{-1}$.
  3. L^2 mixing time satisfies $T_*(\varepsilon) \lesssim \kappa \cdot d \cdot \log\left(\frac{\kappa \cdot d}{\varepsilon}\right)$
- Paper contains tools which imply simple bounds for much wider class of targets.
- Today: demystify those tools.

Proof Overview

- Roughly:
  1. Large-Scale Properties of Target
  2. + Small-Scale Properties of Sampler
  3. $\rightsquigarrow$ Good Mixing.
- Precisely:
  - 'Isoperimetric' Profile of Target
  - + 'Close Coupling' of Kernels
  - $\rightsquigarrow$ Isoperimetric Profile of *Markov Chain*
    - $\rightsquigarrow$ Good Mixing (in L^2).
- True for fairly general Markov chains on metric spaces.
- For RWM *in particular*:
  - 'Metropolis-type' + Acceptance Control $\rightsquigarrow$ Close Coupling.
- I will explain all of these terms.

Isoperimetric Profiles of Probability Measures

- For $A \subseteq E$ and $r \geqslant 0$, let $A_r := \{x \in E : \mathrm{d}(x, A) \leqslant r\}$.
- Define the *Minkowski content* of A under π with respect to d by

$$\pi^+(A) = \lim \inf_{r \to 0^+} \frac{\pi(A_r) - \pi(A)}{r}.$$

- The *isoperimetric profile* of π with respect to the metric d is

$$I_\pi(p) := \inf \left\{ \pi^+(A) : A \subseteq E, \pi(A) = p \right\}, \qquad p \in (0, 1).$$

- (usually) increasing on $\left[0, \frac{1}{2}\right]$, symmetric about $1/2$.

Isoperimetric Profiles: Interpretation

- Isoperimetry relates the mass of sets to the mass of their boundaries.
- For Markov chains: isoperimetry captures how difficult it is to escape a given set.
- Escaping small sets ($p \to 0^+$) happens to be the relevant limit.
- If you escape all sets equally easily ($I_\pi(p) \geqslant c \cdot p$),
  - then you mix exponentially quickly.
- If you also escape small sets particularly well ($I_\pi(p) \gg c \cdot p$),
  - then things can be even better at the start.
- If small sets are hard to escape ($I_\pi(p) \ll c \cdot p$),
  - then things can be much worse.

Isoperimetric Profiles: Examples

- $\pi = \mathcal{N}(0, I_d)$ has $I_\pi(p) = \left(\varphi_\gamma \circ \Phi_\gamma^{-1}\right)(p) \sim p \cdot \left(2 \cdot \log \frac{1}{p}\right)^{1/2}$ as $p \to 0^+$.
- $\pi(\mathrm{d}x) \propto \exp(-|x|)\,\mathrm{d}x$ has $I_\pi(p) = \min\{p, 1-p\}$.
- $\pi(\mathrm{d}x) \propto \exp(-|x|^\alpha)\,\mathrm{d}x$ has $I_\pi(p) \geqslant K(\alpha) \cdot p \cdot \left(\log \frac{1}{p}\right)^{1-1/\alpha}$ for $p \in \left[0, \frac{1}{2}\right]$.
- For log-concave measures,
  - $\approx$ preserved under products.
  - functional inequalities (PI, LSI, $\cdots$) imply bounds on I_π.
- Profiles transfer nicely under Lipschitz mappings, bounded change of measure.
- Can be hard to obtain good bounds in some cases.
- Typically very informative.

‘Close Coupling’ of Markov Kernels

- Say that P is $(\mathsf{d}, \delta, \tau)$-close coupling if for some **fixed** $\delta, \tau > 0$, it holds that

$$\mathsf{d}(x, y) \leqslant \delta \implies \mathrm{TV}\left(P_x, P_y\right) \leqslant 1 - \tau.$$

- When two chains get close enough, anywhere in the space,
  - there is a decent chance to make them coalesce.
- In our experience,
  - weaker assumption than global contractivity of dynamics ,
  - typically holds with better constants than minorisation conditions.
- δ is often small (but not tiny)
- τ can be of constant order (e.g. $1/4$).

Isoperimetric Profiles of Markov Chains

- Define

$$I_{\pi,P}(p) := \inf \left\{ \pi \otimes P \left(A \times A^{\complement} \right) : \pi(A) = p \right\}$$

- 'How hard is it for *this chain* to leave sets of a given size?'
- Related to 'conductance', 'conductance profile' of Markov chain.
- Good lower bounds on $I_{\pi,P}$ translate into mixing time bounds for P.

$$T_* (\varepsilon \asymp 1) \lesssim \int_{\chi^2(\mu_0,\pi)^{-1}}^{1/2} \frac{p \, \mathrm{d}p}{I_{\pi,P}(p)^2}.$$

- I will not go into the details of how this is achieved today.

Isoperimetry: from π to P, to mixing

- Suppose that π has profile I_π, and P is $(\mathrm{d}, \delta, \tau)$-close coupling. Then

$$I_{\pi,P}(p) \gtrsim \tau \cdot \min\{p, \delta \cdot I_\pi(p)\}$$

- Corollary 1: L^2 mixing time satisfies

$$T_*(\varepsilon \asymp 1) \lesssim \tau^{-2} \cdot \delta^{-2} \cdot \int_{\chi^2(\mu_0,\pi)^{-1}}^{1/2} \frac{p\,\mathrm{d}p}{I_\pi(p)^2}.$$

(overlooking an additional annoying term related to the min)

- Corollary 2: for log-concave π, it holds that

$$\gamma_P \gtrsim \tau^2 \cdot \delta^2 \cdot I_\pi\left(\frac{1}{2}\right)^2.$$

- Our target is fixed, now: look at the kernel P, and control (τ, δ).

Close Coupling for RWM

- For MH algorithms, natural to try

$$\mathrm{TV}\left(P_x, P_y\right) \leqslant \mathrm{TV}\left(P_x, Q_x\right) + \mathrm{TV}\left(Q_x, Q_y\right) + \mathrm{TV}\left(Q_y, P_y\right).$$

  This appears to have some limitations.
- Being ‘Metropolis-type’ (not just ‘Metropolis-Hastings-type’) lets us do better.
  - $\alpha(x, x') = \mathtt{Monotone}(f(x')/f(x))$.
  - No ‘cross terms’, as in general MH.
- We will see that it suffices to control the acceptance rates.
  - $\rightsquigarrow$ need to control the regularity of π.

Total Variation Bound between Metropolis Kernels

- Lemma: Let P be a Metropolis kernel, and suppose that $\inf_{x \in E} \alpha(x) \geqslant \alpha_0 > 0$. Then for any $x, y \in E$, it holds that

$$\mathrm{TV}\left(P_x, P_y\right) \leqslant \mathrm{TV}\left(Q_x, Q_y\right) + (1 - \alpha_0).$$

- Proof: WLOG, assume that $\pi(x) \geqslant \pi(y)$. If both chains propose moving to z, then it is possible to couple the acceptance steps so that whenever x accepts the move, so does y. Use $P(A \cap B) \geqslant P(A) + P(B) - 1$ to see that chains meet w.p. $\geqslant (1 - \mathrm{TV}\left(Q_x, Q_y\right)) + \alpha_0 - 1 = \alpha_0 - \mathrm{TV}\left(Q_x, Q_y\right)$. Conclude by coupling inequality.

Proof Sketch

- WLOG, assume that $\pi(x) \geqslant \pi(y)$.
- If both chains propose moving to z, then $\alpha(x, z) \leqslant \alpha(y, z)$.
- Thus, can couple the acceptance steps so that

$$x \text{ accepts move} \implies y \text{ accepts move}$$

- Use $P(A \cap B) \geqslant P(A) + P(B) - 1$ to see that

$$\begin{aligned} P\left(X' = Y'\right) &\geqslant P\left(\tilde{X} = \tilde{Y}\right) + P\left(X' = \tilde{X}\right) - 1 \\ &\geqslant \left(1 - \mathrm{TV}\left(Q_x, Q_y\right)\right) + \alpha_0 - 1 \\ &= \alpha_0 - \mathrm{TV}\left(Q_x, Q_y\right). \end{aligned}$$

- Conclude by coupling inequality.

Acceptance Rate Bounds for RWM

- Recall that $\alpha(x, x') = \min\left\{1, \frac{\pi(x')}{\pi(x)}\right\}$.
- Natural to control growth of $U = -\log\pi$.
- Assumption: for some ψ, it holds that

$$U(x+h) - U(x) - \langle \nabla U(x), h \rangle \leqslant \psi(|h|).$$

- Lemma: The acceptance rate satisfies

$$\alpha(x) \geqslant \frac{1}{2} \cdot \exp\left(-\int \mathcal{N}(\mathrm{d}z; 0, I_d) \cdot \psi(\sigma \cdot |z|)\right),$$

  and taking $\sigma = \upsilon \cdot d^{-1/2}$ gives that

$$\alpha(x) \geqslant \frac{1}{2} \cdot \exp\left(-\psi(\upsilon) + \mathcal{O}\left(d^{-1}\right)\right).$$

Close Coupling for RWM

- Taking $\sigma = \upsilon \cdot d^{-1/2}$ allows for $\alpha_0 \geqslant \frac{1}{2} \cdot \exp\left(-\psi(\upsilon) + \mathcal{O}\left(d^{-1}\right)\right)$.
- Taking $\delta = \sigma \cdot \alpha_0$ allows for

$$\mathsf{d}(x, y) \leqslant \delta \implies \mathrm{TV}\left(Q_x, Q_y\right) \leqslant \frac{1}{2} \cdot \alpha_0.$$

- Using the coupling result, one may then take $\tau = \frac{1}{2} \cdot \alpha_0$.

Isoperimetric Profile and Mixing of RWM

- Recalling that

$$I_{\pi,P}(p) \gtrsim \tau \cdot \min\{p, \delta \cdot I_\pi(p)\}$$

and taking υ so that $\alpha_0 \asymp 1$, obtain that

$$I_{\pi,P}(p) \gtrsim \min\{p, \sigma \cdot I_\pi(p)\},$$

$$\gamma_P \gtrsim \sigma^2 \cdot I_\pi\left(\frac{1}{2}\right)^2$$

$$T_*(\varepsilon \asymp 1) \lesssim \sigma^{-2} \cdot \int_{\chi^2(\mu_0,\pi)^{-1}}^{1/2} \frac{p \, \mathrm{d}p}{I_\pi(p)^2}.$$

Deducing main results (1)

- Under m-strong log-concavity, can bound isoperimetric profile as

$$I_\pi(p) \geqslant c \cdot m^{1/2} \cdot p \cdot \left(\log \frac{1}{p}\right)^{1/2}$$

- Under L-smoothness, take $\sigma = \upsilon \cdot (L \cdot d)^{-1/2}$ and control acceptance ratio as

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

- Good isoperimetry, good acceptance rates $\rightsquigarrow$ Good mixing.

Deducing main results (2)

- Combining earlier results, obtain

$$\gamma_P \gtrsim 1/(\kappa \cdot d)$$

$$T_* (\varepsilon \asymp 1) \lesssim \sigma^{-2} \cdot m^{-1} \int_{\chi^2(\mu_0,\pi)^{-1}}^{1/2} \frac{\mathrm{d}p}{p \cdot \log\left(\frac{1}{p}\right)}$$

$$\lesssim \kappa \cdot d \cdot \log\log \chi^2(\mu_0, \pi).$$

- Same strategy works well for other targets:
  - Characterise the isoperimetric profile (out of your hands).
  - Control the acceptance rates.

Not discussed in detail

- Sharpness of bounds w.r.t. d.
- Implications for asymptotic variance.
- ‘Multi-phase convergence’, initialisation.
- RWM on targets ‘between exponential and Gaussian’.
- RWM on rougher targets.
- pCN for Gaussian prior, ‘centered’ log-concave likelihood.

Ongoing and future work

- RWM on Heavy-tailed targets.
- Other Metropolis algorithms.
- Other non-Metropolis algorithms.

Recap

- RWM for MCMC sampling.
- MCMC Convergence analysis via:
  - Isoperimetry (of target).
  - Close Coupling (of kernels).
- Explicit control of RWM acceptance rates.
- Estimates of spectral gap, L^2 mixing times, asymptotic variance, etc.