Convergence bounds for the Random Walk Metropolis algorithm

Perspectives from Isoperimetry

Sam Power

University of Bristol

11 May, 2023

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

- Motivating task: making sense of structured probability distributions in high-dimensional spaces
  - posterior inference in Bayesian statistics
  - latent variable models, hidden Markov models
  - generative modeling
  - non-convex optimisation
  - $\ldots$

Markov Chain Monte Carlo (MCMC)

- Task: Generate approximate samples from a probability distribution π to which we have *limited access*.
- MCMC: An iterative approach to this task.
  - Simulate a time-homogeneous Markov chain $(X_n)_{n \geqslant 0}$ such that

$$\mathrm{Law}(X_n) \to \pi \text{ as } n \to \infty.$$

    (and hopefully, quickly)
- Use samples to 'understand' π.

A Glimpse at Modern MCMC

- Current status:
  - Mature algorithmic field, many ‘correct’ solutions are known and practical.
  - Quantitative convergence theory is *challenging*; *important*.
    - ‘Is (this algorithm) $\{$ performant, reliable, preferable, ... $\}$?’
    - ‘Given π, which algorithm do I choose?’
- Many interesting mathematical questions about complexity of sampling
  - Parallels with optimisation
  - Interplay with convex geometry
- Current trend: seek thorough understanding of simple algorithms.
- Burgeoning trend: study approximate inference methodologies with same tools.

Random Walk Metropolis

- Today: Study the *Random Walk Metropolis* (RWM) algorithm
  - Only requires access to density of π, up to a multiplicative constant (typical).
  - Widely-used, simple, 'representative' difficulties

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

- Leaves π invariant; ergodic under mild conditions.

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: convexity assumptions, ‘optimisation-style’ proofs, . . .
- Today: synthesis of the above.

Some Comments on our Results

- Despite ubiquity, sharp complexity analysis of RWM has long been open.
  - Preferable to rejection sampling, quadrature, $\cdots$?
- We obtain a convincing complexity analysis with
  - sharp dependence on the dimension of the problem
  - conjecturally sharp dependence on the conditioning of the problem
- Our proof techniques are remarkably robust, and largely new to this area
- Gives a relatively complete resolution to the question of RWM's mixing

Main Results

- Suppose that
  - Target is $\pi(x) \propto \exp(-U(x))$,
  - U is m-strongly convex, L-smooth,
  - Write $\kappa = L/m$ ('condition number').
- 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.

Isoperimetry, Conductance, and Escapes

- For ‘local’ Markov chains, a powerful tool of analysis is the ‘conductance’ method.
- The core idea is that if a chain cannot get stuck badly in a set of small mass, then the chain must be mixing well.
- Quantitatively: for any (small) set A, [the flow of the chain out of A and into $A^{\complement}$] is comparable to [the mass of A].
- If this condition holds, then the chain is mixing well.
  - Under some conditions, this is a theorem.
  - Weaker and stronger versions of this property are also useful.
  - These each lead to their own theorems.

Conductance Methods for Markov Chains

- Consider for $A \subseteq \mathrm{R}^d$

$$\pi(A) := \int_{x \in A} \pi(x)\,\mathrm{d}x$$

$$\pi \otimes P(A \times A^{\complement}) := \int_{x \in A, y \in A^{\complement}} \pi(x) P(x,y)\,\mathrm{d}x\mathrm{d}y.$$

- If $\pi \otimes P(A \times A^{\complement}) \geqslant c \cdot \pi(A)$, then $\mathrm{P}(X_1 \notin A \mid X_0 \in A) \geqslant c$,
  - so if $c \gg 0$, then the set A is easy for P to escape.
- If every set A is easy for P to escape, then P cannot get stuck ...
  - ... and hence must converge quickly.

Isoperimetric Profiles of Markov Chains

- Define

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

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

$$T_{*}\left(\varepsilon \asymp 1\right) \lesssim \int_{\chi^{2}(\mu_{0},\pi)^{-1}}^{1/2} \frac{p\,\mathrm{d}p}{J_{\pi,P}(p)^{2}}.$$

- I will not go into the technical details of how this is achieved today.
  - (...but $\exists$ bonus slides).

Markov Isoperimetric Profiles: Interpretation

- Classical isoperimetry relates the *mass of sets* to the *mass of their boundaries*.
- For Markov chains, isoperimetry describes 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 ($J_{\pi,P}(p) \geqslant c \cdot p$),
  - then you mix exponentially quickly.
- If you also escape small sets particularly well ($J_{\pi,P}(p) \gg c \cdot p$),
  - then things can be *even better* at the start.
- If small sets are particularly hard to escape ($J_{\pi,P}(p) \ll c \cdot p$),
  - then things *can* be much *worse*.

Estimating the profile $J_{\pi,P}$

- Directly computing $J_{\pi,P}$ involves a difficult infimum over measurable sets.
- Our route will be to show that under verifiable conditions, we can estimate $J_{\pi,P}$.
- These conditions are nicely decoupled as $\approx$:
  1. A global condition about the target measure π.
  2. A local condition about the kernel P.
- Remark: This part of the analysis should work for ~generic problems.
  - When guided by local information (e.g. gradients) to solve global problems (e.g. sampling), conditions of this form are relevant.
- I will explain the conditions, and then explain how they fit together.

Isoperimetric Profiles of Probability Measures

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

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

  - $\approx$ 'boundary mass' of A under π
- 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 $1/2$.
- For experts: This is (basically) $J_{\pi,P}$ for the Langevin diffusion.

Isoperimetric Profiles: Examples (1)

- $\pi(\mathrm{d}x) \propto \exp(-|x|)\,\mathrm{d}x$ has $I_\pi(p) = p$.
- $\pi = \mathcal{N}(0, I_d)$ 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 $p \to 0^+$.
- For $\alpha \in (1, 2)$, $\pi(\mathrm{d}x) \propto \exp(-|x|^\alpha)\,\mathrm{d}x$ has $I_\pi(p) \geqslant K(\alpha) \cdot p \cdot \left(\log \frac{1}{p}\right)^{1-1/\alpha}$.
- (many other explicit examples in one dimension)

Isoperimetric Profiles: Examples (2)

- For log-concave measures,
  - Essentially preserved under products.
  - Functional inequalities (PI, LSI, $\cdots$) directly imply bounds on I_π.
  - Heat flow improves things.
  - Implied by certain concentration inequalities.
- General transfer principles:
  - Pushforward by Lipschitz transport map.
  - Change of measure by log-bounded weight.
- In specific cases, it can be hard to obtain good bounds.
- Once you have a bound, it is typically very informative about $\{$ mixing, concentration, $\cdots$ $\}$.

‘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

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

- If two chains get close enough, anywhere in the space,
  - then there is a decent chance to make them coalesce.
- In our experience,
  - weaker assumption than e.g. global contractivity of dynamics, and
  - typically holds with better constants than minorisation conditions.
- δ is often small ($\approx$ step-size).
- τ can be of constant order (e.g. $1/8$).
- Remark: this condition can hold, with good (δ, τ), for chains which mix **badly**.

Obtaining $J_{\pi,P}$

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

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

- Interpretation:
  - If P is 'nice' at small scales,
  - and if π is 'nice' at large scales,
  - then P will mix well!
- Alternatively: P mixes 'as well as' the Langevin diffusion, slowed down by (τ, δ).
- For algorithms: no point in making τ too big; think of it as a constant.
  - $\rightsquigarrow$ Tune algorithm to find a good δ which gives a desired τ.

Isoperimetry: from π to P, to mixing

- Corollary 1: L^2 mixing time satisfies

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

  (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

- Recall: want to show that

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

- For MH algorithms, natural to try using Δ inequality:

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

  This appears to have some limitations.
  - Roughly: tail behaviour ruins two of the three terms.

Close Coupling for RWM (2)

- We will see that being 'Metropolis-type' (not just 'Metropolis-Hastings-type')

$$\alpha(x, x') = \mathrm{Monotone}(f(x')/f(x))$$

  lets us do better.
  - Key: no 'cross terms', as in general MH.
- We will see that if we can control the **marginal** acceptance rates of the chain, then we can guarantee the close coupling condition.
  - $\rightsquigarrow$ 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: Explicitly construct a coupling (next slide).

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 almost surely,

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

- 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 $U = -\log \pi$.
- Smoothness 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

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

  - (compute KL between Gaussians; apply Csiszar-Kullback-Pinsker)
- Using the coupling result,

$$\begin{aligned} \mathrm{TV}\left(P_x, P_y\right) &\leqslant \mathrm{TV}\left(Q_x, Q_y\right) + (1 - \alpha_0) \\ &\leqslant 1 - \frac{1}{2}\alpha_0, \end{aligned}$$

  i.e. one may take $\tau = \frac{1}{2} \cdot \alpha_0$.

Isoperimetric Profile and Mixing of RWM

- Recalling that

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

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

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

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

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

- Still *very* general at this stage.

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.
    - Remark: Failure of these conditions corresponds to known failure modes for RWM.

Deducing main results (2)

- Combining earlier results, obtain

$$
\begin{aligned}
\gamma_P &\gtrsim 1/(\kappa \cdot d) \\
T_*(\varepsilon \asymp 1) &\lesssim \sigma^{-2} \cdot m^{-1} \cdot \int_{\chi^2(\mu_0,\pi)^{-1}}^{1/2} \frac{\mathrm{d}p}{p \cdot \log\left(\frac{1}{p}\right)} \\
&\lesssim \kappa \cdot d \cdot \log\log \chi^2(\mu_0, \pi).
\end{aligned}
$$

- 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.
- New algorithms inspired by proof techniques.

Beyond MCMC (1)

- Convexity of potentials is not essential to our results.
- The key ingredient is really **isoperimetry**, which is a more robust notion.
- For $\sim$general approximate inference schemes, we expect similar.
  - A caveat: need to assume that we work in a fixed coordinate system.
  - This is because changing coordinates changes the isoperimetry.
  - For e.g. ‘dense’ VGA, might need a more refined analysis.

Beyond MCMC (2)

- We expect isoperimetric ideas to be insightful quite generally.
- For fixed-covariance VGA, isoperimetry of π ...
  - $\rightsquigarrow$ regularity of KL $(\cdot, \pi)$
  - $\rightsquigarrow$ bounds on quality of approximation in e.g. $\mathcal{T}_2$
- For normalising flows, isoperimetry of π ...
  - $\rightsquigarrow$ (im)possibility of transport maps with good regularity
  - $\rightsquigarrow$ practical need for stable non-Lipschitz transport maps
- For denoising diffusions, isoperimetry of π ...
  - $\rightsquigarrow$ how large T should be so that $\pi P_T \approx \mathcal{N}(0, I_d)$
  - $\rightsquigarrow$ regularity of backwards diffusion

Beyond MCMC (3)

- A general message: when processing probability measures to be 'nice',
  1. Prioritise good isoperimetry, and
  2. Prioritise good smoothness.
- $\{$ Preconditioning, Flow Transport, $\cdots$ $\}$ are particular instances of this idea.
- It is difficult to have much better isoperimetry than a Gaussian measure.
  - This arguably justifies the routine use of Gaussian measures as references.
- Many algorithms will implicitly require either these properties.
  - When not, they tend to require something **very** different (e.g. Gibbs sampling, CAVI).

Recap

- RWM for MCMC sampling.
- MCMC Convergence analysis via:
  - Isoperimetry (of target), and
  - Close Coupling (of kernels).
- Explicit bounds with interpretable dependence on problem parameters.

Bonus Slides 1 : Technical Details

- Isoperimetric Profile to Conductance Profile:

$$\Phi_P(v) := \inf\left\{\frac{J_{\pi,P}(v)}{v} : 0 < v \leqslant \frac{1}{2}\right\}. \tag{1}$$

- Conductance Profile to Spectral Profile:

$$\Lambda_P(v) \geqslant \frac{1}{2}\Phi_P(v)^2. \tag{2}$$

- Spectral Profile to Functional Inequality: for $f \geqslant 0$,

$$\frac{\mathcal{E}(P, f)}{\mathrm{Var}_\pi(f)} \geqslant \frac{1}{2} \cdot \Lambda_P\left(4 \cdot \frac{\pi(f)^2}{\mathrm{Var}_\pi(f)}\right). \tag{3}$$

- Functional Inequality to Mixing Time: consider $\|P^n f\|_2^2$ with $f = \frac{\mathrm{d}\mu}{\mathrm{d}\pi}$.

Bonus Slides 2 : Super-Poincaré Inequalities

- Spectral Profile to Super-Poincaré Inequality

$$\mathrm{Var}_{\pi}(f) \leqslant s \cdot \mathcal{E}(P, f) + \beta_P(s) \cdot \pi(|f|)^2 ; \tag{4}$$

can express β_P in terms of Λ_P.

Some Pointers to the Literature

1. <u>MCMC Overview</u>: Roberts-Rosenthal (General MCMC, 'Qualitative' Analysis, Optimal Scaling), Vempala, Chewi, Lee-Vempala ('Convex' Perspectives)
   - (+ my thesis)
2. <u>Conductance</u>: Lawler-Sokal, Jerrum-Sinclair (original papers), Douc-Moulines-Priouret-Soulier (Chapter on Spectral Theory); C. Sherlock (notes)
3. <u>Isoperimetric Profiles</u>: Bobkov-Houdré (1997 book), E. Milman (papers 2007-2010)
4. <u>Functional Inequalities</u>: Montenegro-Tetali (monograph), Diaconis-Saloff-Coste (papers)
   - (+ our tech report)