On the Convergence of the Random Walk Metropolis Algorithm

Sam Power, 9 Oct 2023, Probability Seminar, King’s College London

Links and Acknowledgements

- Main paper today: arXiv 2211.08958
- Related technical report: arXiv 2208.05239
- All joint work with
  - Christophe Andrieu (Bristol)
  - Anthony Lee (Bristol)
  - Andi Q. Wang (Bristol ⇝ Warwick)
- Funded by Bayes4Health EPSRC Grant

Talk Overview

- Markov chains are a useful tool for exploring probability distributions.
- The Random Walk Metropolis (RWM) is such an algorithm.
- In recent work, we study the quantitative convergence of the RWM.

Talk Goals

- In this talk, my focus is conceptual rather than technical.
- I hope to enable your intuition for

  1. which factors influence the convergence behaviour of RWM, and
  2. which properties must be verified to prove so.

Quantitative Convergence of the Random Walk Metropolis

Quantitative Convergence of the RWM

- What is the Random Walk Metropolis?
- To what is it trying to ‘converge’?
- How do we quantify its success in doing so?

Motivating Task

- “making sense of structured probability distributions in complex spaces”
  - posterior inference in Bayesian statistics
  - latent variable models, hidden Markov models
  - statistical physics
  - generative modeling
  - non-convex optimisation
  - …

Markov Chain Monte Carlo (MCMC)

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

$$\text{Law}\left(X_n\right) \to \pi \quad \text{as}\, n \to \infty$$

      (and hopefully, quickly)

    - Use $\left(X_n : n \in \mathbf{N}\right)$ to 'understand' π.

Describing Random Walk Metropolis

- Today: Study the *Random Walk Metropolis* (RWM) algorithm
  - Only requires pointwise access to the density of π
    - (up to a multiplicative constant - typical in applications)
  - ‘fancy guess-and-check’
  - Widely-used, simple; ‘representative’ difficulties

Defining Random Walk Metropolis

RWM $\left(\pi, \sigma^2\right)$

1. At x,

    1.1. Propose $x' \sim \mathcal{N}\left(x' \mid x, \sigma^2 \cdot I_d\right)$.

    1.2. Evaluate $r(x, x') = \pi(x')/\pi(x)$.

    1.3. With probability $\alpha(x, x') = \min\{1, r(x, x')\}$, move to x'.

    - Otherwise, remain at x.

- Leaves π invariant; ergodic under mild conditions.

Some Relevant Objects

- Worthwhile to bear these in mind going forward:
    - $Q\left(x, \mathrm{d}x'\right) := \mathcal{N}\left(\mathrm{d}x'; x, \sigma^2 \cdot I_d\right)$
    - $r(x, x') = \pi(x')/\pi(x), \quad \alpha(x, x') := \min\{1, r(x, x')\}$
    - $\alpha(x) := \int Q\left(x, \mathrm{d}x'\right) \alpha(x, x'), \quad \alpha_0 := \inf\left\{\alpha(x) : x \in \mathbf{R}^d\right\}$
    - The Markov kernel P corresponding to RWM $\left(\pi, \sigma^2\right)$ writes as

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

Quantifying Convergence of the RWM

- We want to quantify statements of the form "Law $\left(X_n\right) \to \pi$ as $n \to \infty$".
- Many possible metrics, divergences, etc.
- We work with "convergence in $\mathrm{L}^2(\pi)$".
    - Strong; implies other forms of convergence (TV, $\mathscr{T}_p$, KL, $\cdots$).
    - Details will be suppressed in the talk; can elaborate afterwards.

Prior Work

- Early work: qualitative (exponential or not?), quantitative rates left implicit
- Optimal Scaling: dimension-dependence for a product-form model problem
- Modern era: focus on ‘convex’ regime; non-asymptotic convergence
- Goal for our work: retain all lessons learned from prior studies

Presenting Convergence of the RWM

- In the paper, we provide non-asymptotic estimates on this convergence,
  - holding at all times n, for any step-size σ, and
  - with explicit dependence on the details of the target π.
- In this talk, I will instead present more digestible ‘mixing time’ estimates.
  - Interpret as “how large must n be to get within $\mathcal{O}(1)$ of π?”.
  - Provides a simple complexity analysis.

The Ingredients of Convergence for Random Walk Metropolis

The Ingredients of Convergence for RWM

- Convergence of $\mathsf{RWM}\left(\pi, \sigma^2\right)$ is largely dictated by two features:
    1. Convergence of a related diffusion process, $\mathsf{OLD}(\pi)$, and
    2. Control of the worst-case acceptance rate, $\alpha_0 = \alpha_0(\sigma)$.
- In general, it holds that

$$\mathsf{Mixing}\left(\mathsf{RWM}\left(\pi, \sigma^2\right)\right) \gtrsim \alpha_0^4 \cdot \sigma^2 \cdot \mathsf{Mixing}\left(\mathsf{OLD}(\pi)\right)$$

- Simple to discern that $\alpha_0 \to 0^+$ is generally un-interesting; focus on $\alpha_0 \gtrsim 1$.

The Overdamped Langevin Diffusion

The Small-Step-Size Limit of RWM
Taylor Heuristics

- For $\sigma \to 0^+$, Taylor expansions yield that

$$\mathbf{E}_P\left[Y \mid X = x\right] \approx x + \frac{\sigma^2}{2}\nabla \log \pi(x)$$

$$\mathbf{Var}_P\left[Y \mid X = x\right] \approx \sigma^2 \cdot I_d$$

- ⇝ Study some ‘simpler’ Markov process with these characteristics

The Overdamped Langevin Diffusion

Definitions

- Write our target as $\pi \propto \exp(-U)$; call U the 'potential'.
- The Overdamped Langevin Diffusion with target π is

$$\mathrm{d}X_t = -\nabla U\left(X_t\right)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$$

  - Write $\mathsf{OLD}(\pi)$ for this process.
- Straightforward to check that this process is π-reversible
  - (hence invariant)

The Overdamped Langevin Diffusion

‘just another Markov process?’

- Far from it!

- OLD (π) is somehow a ‘canonical’ object.

- Fundamental tool for analysing { geometry, concentration, $\cdots$ } of π

- Many aspects are very well-understood by now.

Crash-Course in the Convergence of OLD

Examples (1)

1. If U is convex,
   - then OLD converges at some exponential rate.
2. If U is uniformly convex,
   - then OLD also initially converges at a *faster-than-exponential* rate.
3. If U grows sublinearly in the tails,
   - then OLD can only ever converge at a *slower-than-exponential* rate.

Crash-Course in the Convergence of OLD
Examples (2)

4. If U is convex,
    - then *conjecturally* (KLS), the exponential rate satisfies

$$\gamma_\pi \gtrsim \left\| \mathrm{Cov}_\pi\left(\mathrm{id}\right) \right\|_{\mathrm{op}}^{-1}$$

    independently of the dimension.

5. Various transfer principles:
    - Change of measure, Lipschitz transport, …

In what sense does OLD **'resemble'** RWM?

Nature of Approximation

- In what sense are the processes close?
  - Not pathwise, nor uniformly (tails)
  - In terms of { *exit, boundary* } behaviour!
- If $A \subseteq \mathbf{R}^d$ (and $\pi(A)$ is not too small),
  - then both RWM, OLD require similar amounts of effort to exit A
  - Mathematically: isoperimetry, conductance, $\ldots$
- For convergence in $\mathrm{L}^2(\pi)$, this 'resemblance' is sufficient.

Controlling the Acceptance Rate

Controlling the Acceptance Rate

Regularity of the Potential

- Recall that $\alpha(x, x') = \min\{1, \pi(x')/\pi(x)\}$
  - Natural to study regularity of $U = -\log \pi$.
- Smoothness assumption: for some $p \in [1,2]$, $\psi : \mathbf{R}_+ \to \mathbf{R}$,

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

- e.g. if ∇U is α-Hölder, then one can take $p = 2$, $\psi(r) \sim r^{1+\alpha}$.
- $p = 2$ usually easiest; other p reflect heterogeneity, roughness, $\ldots$

Controlling the Acceptance Rate

Explicit Bounds

- Lemma: The acceptance rate satisfies

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

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

$$\alpha(x) \geqslant \frac{1}{2} \cdot \exp\left(-\psi\left(c_p \cdot \upsilon\right) + o\left(1\right)\right)$$

- For $p = 2$, consistent with usual ‘optimal scaling’ results.
- For $p \in [1,2)$, rougher target ⇝ smaller step-sizes are required to stabilise α_0.

Partial Recap

- Convergence is dictated by

$$\text{Mixing}\left(\text{RWM}\left(\pi,\sigma^2\right)\right) \gtrsim \alpha_0^4 \cdot \sigma^2 \cdot \text{Mixing}\left(\text{OLD}\left(\pi\right)\right)$$

- Understand ‘global’ picture by studying $\text{OLD}\left(\pi\right)$
- Understand ‘local’ picture by examining the regularity of π, bounding α_0.
- (division of labour)

Explicit Examples

Explicit Examples
(norm-based model problems)

- Throughout, will assume ‘realistic’ initialisation and omit log factors
    - (forgive me; minor points + happy to elaborate later)
- Some (nested) model problems:

1. $U(x) = \|x\|_2^2$: take $\sigma \sim d^{-1/2}$, gives $T_{\text{mix}} \lesssim d$.
2. $U(x) = \|x\|_2^\alpha$: take $\sigma \sim d^{-1/2}$, gives $T_{\text{mix}} \lesssim d^{2/\alpha}$ for $\alpha \in [1,2]$.
3. $U(x) = \|x\|_p^\alpha$: take $\sigma \sim d^{-1/p}$, gives $T_{\text{mix}} \lesssim d^{2/p+2/\alpha-1}$ for $\alpha, p \in [1,2]$.

Examining the model problems

- $\alpha, p \in [1,2], U(x) = \|x\|_p^\alpha \rightsquigarrow T_{\mathsf{mix}} \lesssim d^{2/p} \cdot d^{2/\alpha - 1}$
  - One factor for roughness, one factor for tails.
  - One factor for step-size, one factor for diffusion.
  - (division of labour)

Another Explicit Example

(well-conditioned convex potentials)

- Suppose that U is convex, with $\mathrm{eigs}\left(U''\right) \in [m, L]$ ('well-conditioned').
- Scale $\sigma \sim \left(L \cdot d\right)^{-1/2}$ to stabilise α_0.
- This gives $T_{\text{mix}} \lesssim \kappa \cdot d$, where $\kappa = L/m$.
    - One factor for roughness, one factor for tails.

Recap

- Random Walk Metropolis (RWM) is an algorithm for approximate sampling.
- We conduct an analysis of its rate of convergence to equilibrium.
- Reduces to two (largely decoupled) questions:
    1. Am I accepting my proposed moves? (local regularity)
    2. Would the Overdamped Langevin diffusion mix well? (global regularity)
- Very general, easy to apply, and often gives sharp results.