# On the Convergence of the Random Walk Metropolis Algorithm

Sam Power

University of Bristol

30 March, 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

# Talk Goals

- In recent work, we resolve (to a large extent) the convergence behaviour of the Random Walk Metropolis MCMC algorithm.
- In this talk, I hope to convey conceptual messages rather than technical details.
- I want to enable your understanding of how our results can be used to deduce convergence estimates for concrete sampling problems.

# 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
  - ...

# Markov Chain Monte Carlo (MCMC)

- Task: Generate approximate samples from a probability distribution $\pi$ 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’ $\pi$.

# Random Walk Metropolis

- Today: Study the *Random Walk Metropolis* (RWM) algorithm
  - Only requires access to density of $\pi$, 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 $\pi$ 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 $\left\{1, r\left(x, x^{'}\right)\right\}$.
- $\alpha(x) := \int Q\left(x, \mathrm{d}x^{'}\right) \alpha(x, x^{'})$.
- $\alpha_0 :=$ inf $\left\{\alpha(x) : x \in \mathrm{R}^d\right\}$
- The Random Walk Metropolis kernel $P$ is given by

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

# Headlines

- Our main results take the (stylised) form
  - If $\alpha_0 \gtrsim 1$, then

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

  where $\mathrm{OLD}$ is the continuous-time 'Overdamped Langevin Diffusion' process.
- Can all be made precise via '$\mathrm{L}^2$ convergence', 'mixing times', etc.
- Boils down to
  1. Control worst-case acceptance rates (i.e. bound $\alpha_0$ from below).
  2. Understand how 'nice' the target $\pi$ is (i.e. understand mixing of $\mathrm{OLD}\left(\pi\right)$).

# The Overdamped Langevin Diffusion

- Let $\pi \propto \exp(-U)$ be our target; call $U$ the 'potential'.
- $\mathrm{OLD}(\pi)$ is the SDE given by

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

- 'The canonical $\pi$-reversible diffusion'.
- Fundamental tool for analysing $\{$ geometry, concentration, $\cdots$ $\}$ of $\pi$.
- Many aspects are well-understood by now.

# Crash-Course in Convergence of Langevin Diffusions

1. If $U$ is convex,
   - then $\mathrm{OLD}\left(\pi\right)$ converges at some exponential rate.
2. If $U$ is uniformly quadratically convex,
   - then $\mathrm{OLD}\left(\pi\right)$ initially converges at a *doubly-exponential* rate.
     - (intuition: ‘burn-in’ phase is very quick)
3. If $U$ has slower-than linear growth in the tails,
   - then $\mathrm{OLD}\left(\pi\right)$ can only converge at slower-than-exponential rates.
     - (intuition: time-consuming to get in and out of the tails)

# Crash-Course in Convergence of Langevin Diffusions

1. KLS *Conjecture*: if $U$ is convex,
   - then the exponential rate satisfies

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

     *independently of the dimension*.
     - (intuition: only bottleneck is the 'worst' one-dimensional marginal of $\pi$)
2. Various transfer principles (change of measure, transport, . . .).
   - (intuition: convergence behaviour is robust to various classes of perturbation to $\pi$)

# Back to the Random Walk Metropolis

- Rough intuition: when the dimension is high (small moves) and the target is nice (not too rough), the RWM 'looks like' $\mathrm{OLD}$.
- We can show that (again, glossing over details)

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

  - (Actually, this is an instance of a surprisingly general result; can discuss offline.)
- For RWM, we know that it is not $\{$ interesting, relevant $\}$ to consider $\alpha_0 \to 0$.
- So, how do we control $\alpha_0 \gg 0$?

# Regularity Assumptions on Potential

- Recall that $\alpha(x, x') = \min\left\{1, \frac{\pi(x')}{\pi(x)}\right\}$: natural to control $U = -\log\pi$.
- Smoothness assumption: for some $p \in [1, 2]$, $\psi : \mathbb{R}_+ \to \mathbb{R}$, it holds that

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

- e.g. $\nabla U$ is $\alpha$-Hölder $\rightsquigarrow$ can take $p = 2$, $\psi(r) \sim r^{1+\alpha}$.
- $p = 2$ is typically the easiest case to handle.
- Other $p$ correspond to forms of heterogeneity, roughness.

# Acceptance Rate Control for RWM

- Lemma: The acceptance rate satisfies

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

  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 on step-size.
  - n.b. here, we treat *worst-case* rather than *average-case*.
- For $p < 2$, smaller step-sizes are needed to stabilise $\alpha_0$.

# Taking Stock

- So, to ‘know’ the convergence of Random Walk Metropolis on a given target . . .,
  1. Examine regularity of $U$, extract a step-size $\sigma$ which will control $\alpha_0$.
  2. Examine global structure of $\pi$, characterise the convergence of $\mathrm{OLD}(\pi)$.
  3. ‘Multiply these two things together’.
- This is the punchline, particularly if you don’t want to get your hands too dirty.

# ‘Similarity’ of RWM, OLD

- Morally, the key principle is that $\mathrm{RWM} \approx \mathrm{OLD}$ somehow.
- In what sense is this true?
- Not pathwise, nor uniformly (tails).
- In terms of *exit, boundary* behaviour:
  - If $A \subseteq \mathrm{R}^d$, and $\pi(A)$ is not too tiny, then $\mathrm{RWM}$ and $\mathrm{OLD}$ both require similar amounts of effort to exit $A$; cross the boundary between $A$, $A^{\complement}$.
  - Mathematically: isoperimetry, conductance, . . ..
- For $\mathrm{L}^2$ convergence, this is a sufficient ‘resemblance’ between the processes.

# Some Examples

- Throughout,
  - Assume that initial distance to stationarity is $\exp\Omega(d)$ (typical; achievable).
  - Drop log factors (forgive me).
- Model problems:
  - $U(x) = \|x\|_2^2$: take $\sigma \sim d^{-1/2}$, gives $T_{\text{mix}} \lesssim d$
  - $U(x) = \|x\|_2^\alpha$: take $\sigma \sim d^{-1/2}$, gives $T_{\text{mix}} \lesssim d^{2/\alpha}$ for $\alpha \in [1, 2)$.
  - $U(x) = \|x\|_p^\alpha$: take $\sigma \sim d^{-1/p}$, gives $T_{\text{mix}} \lesssim d^{2/p+2/\alpha-1}$ for $\alpha \in [1, 2)$, $p \in [1, 2]$.
    - One factor for roughness, one factor for tails.
    - One factor for step-size, one factor for diffusion.
- Main example in paper:
  - Suppose that $U$ is convex, with $\text{eigs}\left(U''\right) \in [m, L]$ ('well-conditioned').
  - Take $\sigma \sim (L \cdot d)^{-1/2}$.
  - This gives $T_{\text{mix}} \lesssim \kappa \cdot d$, where $\kappa = L/m$.

# Recap

- Random Walk Metropolis for MCMC sampling.
- Analysis reduces to:
  - Am I accepting my proposed moves? (roughness)
  - How would the corresponding Langevin diffusion mix? (concentration)
- Relatively easy to use, gives sharp results in many cases.