Adaptive Langevin Monte Carlo methods for heavy-tailed sampling via weighted functional inequalities

Tyler Farghly

(Joint work with Leo He, Jun Yang & Patrick Rebeschini)

Department of Statistics
University of Oxford

Complexity of sampling

Goal: Approximately sample from a distribution $\pi(dx) \propto \exp(-f(x))dx$, $f : \mathbb{R}^d \to \mathbb{R}$

Applications: Bayesian learning, inverse problems and score-based generative models

Question: For a given algorithm, what is the computational complexity?

Particular interest in **high-dimensional** setting

Langevin Monte Carlo (LMC)

Langevin diffusion:

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}\, dB_t.$$

Stationary with respect to π

Approximate using the **Euler-Maruyama** scheme:

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\, \xi_{k+1}, \quad \xi_k \sim N(0, I_d).$$

Bridge between sampling and optimization

Recent history of LMC

- **Strongly log-concave** setting (TV) [Durmus & Moulines 2016, Dalalyan 2017]
- **Dissipative** setting (Wasserstein) [Cheng et al. 2019]
- **Logarithmic Sobolev inequality** (KL) [Vempala & Wibisono 2019]
- **Poincaré inequality** (Renyi) [Erdogdu et al. 2020, Chewi et al. 2021]

Little known about LMC in **heavy-tailed** settings...

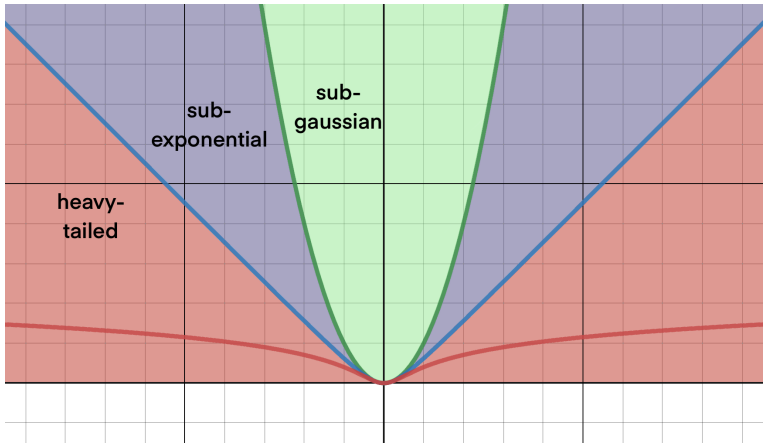
Recent history of LMC

- **Strongly log-concave** setting (TV) [Durmus & Moulines 2016, Dalalyan 2017]
- **Dissipative** setting (Wasserstein) [Cheng et al. 2019]
- **Logarithmic Sobolev inequality** (KL) [Vempala & Wibisono 2019]
- **Poincaré inequality** (Renyi) [Erdogdu et al. 2020, Chewi et al. 2021]

Little known about LMC in **heavy-tailed** settings...

Heavy-tailed measures

$\pi(dx) \propto e^{-f(x)}\,dx$ is **heavy-tailed** if $\|\nabla f(x)\| \to 0$, as $\|x\| \to \infty$



Convergence rates and tail-growth

Logarithmic Sobolev inequality:

$$\mathrm{Ent}_\pi(\varphi) \le 2C_{\mathrm{LSI}} \int \|\nabla\varphi(x)\|^2 \,\pi(dx), \quad \text{for all } \varphi \in \mathcal{C}_c^\infty(\mathbb{R}^d) \tag{LSI}$$

$$\iff \frac{d}{dt} D(\rho_t\|\pi)\Big|_{t=0} \le -\frac{2}{C_{\mathrm{LSI}}} D(\rho_0\|\pi), \quad \text{for all } \rho_0 \in \mathcal{H}(\mathbb{R}^d)$$

Characterises **exponential decay** in KL divergence

But, (LSI) **implies sub-Gaussian concentration** [Bakry et al. 2014]

[Roberts & Tweedie 1996]:

Tails heavier than exponential $\implies$ diffusion not exponentially ergodic

Convergence rates and tail-growth

Logarithmic Sobolev inequality:

$$\mathrm{Ent}_\pi(\varphi) \le 2C_{\mathrm{LSI}} \int \|\nabla\varphi(x)\|^2 \,\pi(dx), \quad \text{for all } \varphi \in \mathcal{C}_c^\infty(\mathbb{R}^d) \tag{LSI}$$

$$\iff \frac{d}{dt} D(\rho_t\|\pi)\Big|_{t=0} \le -\frac{2}{C_{\mathrm{LSI}}} D(\rho_0\|\pi), \quad \text{for all } \rho_0 \in \mathcal{H}(\mathbb{R}^d)$$

Characterises **exponential decay** in KL divergence

But, (LSI) **implies sub-Gaussian concentration** [Bakry et al. 2014]

[Roberts & Tweedie 1996]:

Tails heavier than exponential $\implies$ diffusion not exponentially ergodic

Slow start behaviour

Towards a Complete Analysis of Langevin Monte Carlo: Beyond Poincaré Inequality

TF, A Mousavi-Hosseini, Y He, K Balasubramanian, MA Erdogdu (2023)

Result (informal)

Suppose there exists $\alpha \in [0, 2]$ *such that* $\|\nabla f(x)\| = \mathcal{O}(\|x\|^{\alpha-1})$ *and there exists* $k \in \mathbb{N}$, $q, q', \Delta_0 \in (1, \infty)$ *such that for all* $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$,

$$R_{q'}(\rho_0 \| \pi) \leq \Delta_0 \implies R_q(\rho_k \| \pi) \leq 1.$$

Then, when η *is sufficiently small, it must hold that*

$$\begin{aligned} \alpha = 2: &\quad k\eta \gtrsim \ln(\Delta_0), \\ \alpha \in (0, 2): &\quad k\eta \gtrsim d^{1-\alpha/2} \Delta_0^{\frac{(2-\alpha)^2}{2\alpha}}, \\ \alpha = 0: &\quad k\eta \gtrsim d \exp(\Delta_0/\nu), \end{aligned}$$

New diffusion

Summary: the Langevin diffusion is slow on heavy-tailed targets

Idea: Consider a different diffusion

Weighted LSI

Generalisation of LSI that uses a **weighting function** $\kappa : \mathbb{R}^d \to \mathbb{R}_+$:

$$\mathrm{Ent}_\pi(\varphi) \leq 2C_{\mathrm{LSI}} \int \kappa(x) \|\nabla \varphi(x)\|^2 \, \pi(dx), \quad \text{for all } \varphi \in \mathcal{C}_c^\infty(\mathbb{R}^d).$$

Characterises **exponential decay** for the **weighted Langevin diffusion**,

$$dX_t = -\kappa(X_t)\nabla f(X_t)dt + \nabla \kappa(X_t)dt + \sqrt{2\kappa(X_t)}dB_t. \tag{1}$$

Satisfied by a variety of heavy-tailed measures:
E.g. generalized Cauchy, s-concave and subexponential

P Cattiaux et al. (2010); P Cattiaux, A Guillin, LM Wu (2011); SG Bobkov, M Ledoux (2009)

Discretisation

Difficult to approximate this diffusion

- Coefficients $\kappa\nabla f$, $\nabla\kappa$ and κ are non-globally Lipschitz
- Euler-Maruyama scheme blows up in finite time as $\eta \to 0$
- Non-constant diffusion coefficient

Idea: Use an adaptive Euler-Maruyama scheme

Adaptive step-size and time-change

Can simulate using a time-change/adaptive step-size:

$$dX_t = -\kappa(X_t)\nabla f(X_t)dt + \nabla\kappa(X_t)dt + \sqrt{2\kappa(X_t)}dB_t$$

$$dY_t = \Big(-\nabla f(Y_t) + \nabla \ln \kappa(Y_t)\Big)dt + \sqrt{2}\, dB_t$$

$$d\phi_t = \kappa(Y_t)^{-1}dt$$

Due to B Öksendal (1990): $X_t \simeq Y_{\phi^{-1}(t)}$

Langevin Monte Carlo

Input: potential f, initial $y_0 \sim \rho_0$

For each iteration, $k \leq k_{\max}$:

(i) Update iterate: $y_k = y_{k-1} - \eta \nabla f(y_{k-1}) + \sqrt{2\eta}\, \xi_k$

(ii) Collect the sample y_k

Our algorithm

Input: potential f, initial $y_0 \sim \rho_0$, weighting function κ

For each iteration, $k \leq k_{\max}$:

(i) Update iterate: $y_k = y_{k-1} - \eta \nabla f(y_{k-1}) + \eta \nabla \ln \kappa(y_{k-1}) + \sqrt{2\eta}\, \xi_k$

(ii) Collect the sample y_k

Our algorithm

Input: potential f, initial $y_0 \sim \rho_0$, weighting function κ

For each iteration, $k \leq k_{\max}$, such that $\phi_{k-1} \leq \phi_{\max}$:

(i) Update iterate: $y_k = y_{k-1} - \eta \nabla f(y_{k-1}) + \eta \nabla \ln \kappa(y_{k-1}) + \sqrt{2\eta}\, \xi_k$

(ii) Update clock: $\phi_k = \phi_{k-1} + \eta \kappa(y_{k-1})^{-1}$

(iii) If $\lfloor \phi_k/\eta \rfloor > \lfloor \phi_{k-1}/\eta \rfloor$, collect the sample y_k

Our algorithm

Input: potential f, initial $y_0 \sim \rho_0$, weighting function κ

For each iteration, $k \leq k_{\max}$, such that $\phi_{k-1} \leq \phi_{\max}$:

(i) Update iterate: $y_k = y_{k-1} - \eta \nabla f(y_{k-1}) + \eta \nabla \ln \kappa(y_{k-1}) + \sqrt{2\eta}\, \xi_k$

(ii) Update clock: $\phi_k = \phi_{k-1} + \kappa(y_{k-1})^{-1}$

(iii) If $\lfloor \phi_k \rfloor > \lfloor \phi_{k-1} \rfloor$, collect the sample y_k

- Discretisation of the SDE $dX_t = -\kappa(X_t)\nabla f(X_t)dt + \nabla \kappa(X_t)dt + \sqrt{2\kappa(X_t)}dB_t$
- Adaptive EM discretisation with step-size $\propto \kappa(x)^{-1}$.

Our algorithm

Input: potential f, initial $y_0 \sim \rho_0$, weighting function κ

For each iteration, such that $\phi_{k-1} \leq \phi_{\max}$:

(i) Update iterate: $y_k = y_{k-1} - \eta \nabla f(y_{k-1}) + \eta \nabla \ln \kappa(y_{k-1}) + \sqrt{2\eta}\, \xi_k$

(ii) Update clock: $\phi_k = \phi_{k-1} + \kappa(y_{k-1})^{-1}$

(iii) If $\lfloor \phi_k \rfloor > \lfloor \phi_{k-1} \rfloor$, collect the sample y_k

- Generalisation of LMC: $\kappa \equiv 1$ recovers LMC
- κ is chosen according to the tail-growth of the target

How to choose κ?

Using the framework of functional inequalities

E.g. Bakry calculus, Lyapunov conditions, bounded perturbations, ...

Target density	$\kappa(x)$
$\exp(-\|x\|^\alpha)$, $\alpha \in (0,2]$	$\|x\|^{2-\alpha}$
$(1+\|x\|^2)^{-\frac{d+\nu}{2}}$	$(1+\|x\|^2)^2$
$V(x)^{-(d+\alpha)}$, V convex, $\alpha > 0$	$(1+\|x\|^2)\log(e+\|x\|^2)$

Analysis: assumptions

(A1) Weighted LSI with weighting $\kappa : \mathbb{R}^d \to \mathbb{R}_{\geq 1}$ satisfying $\kappa(x) \leq c(1 + \|x\|^2)^{r/2}$ for some $c, r \geq 0$

(A2) (Moments) There exists $p \in (r, \infty)$ such that $\sigma_p = \pi(\|\cdot\|^p)^{1/p} < \infty$

(A3) $\nabla \ln \kappa$ is L-Lipschitz and $D^2 f \leq M\kappa^{1/2}$ holds a.e. for some $c \in \mathbb{R}_+$

Analysis: continuous-time algorithm

Result (simplified)

Suppose (A1)–(A3) hold, $1 < \alpha^* < p/r$ *and*

$$\phi_{\max} \geq \ln\left(\frac{4R_\alpha(\rho_0\|\pi)}{\varepsilon}\right)\frac{\alpha C_{LSI}}{2}, \quad t_{\max} \geq 4\phi_{\max}\varepsilon^{-1}\Big(8(\alpha-1) \vee \exp((\alpha-1)L\phi_{\max})\Big)$$

for some $\varepsilon > 0$*, then* $R_\alpha(\rho_{\phi_{max}}^{t_{\max}}\|\pi) \leq \varepsilon$.

E.g. for $R_\alpha(\rho_0\|\pi) \geq 1$, it is sufficient to have

$$t_{\max} = \tilde{\Theta}(\varepsilon^{-(\alpha C_{\mathrm{LSI}}/2+1)}R_\alpha(\rho_0\|\pi)^{\alpha C_{\mathrm{LSI}}/2})$$

Analysis: discrete-time algorithm

Result (simplified)

Let $q \leq 2p$ and consider the same setting with $\alpha = 2$, and assume further that

$$\eta \leq (M+L)^{-1} d^{-1} t_{\max}^{-2} \varepsilon^2,$$

and $k_{\max} = \lfloor t_{\max}/\eta \rfloor$, $\mathrm{W}_q(\hat{\rho}_{\phi_{max}}^{k_{\max}}, \pi) \leq \sigma_p \varepsilon$.

E.g. for $R_2(\rho_0 \| \pi) \geq 1$, it is sufficient to have

$$k_{\max} = \tilde{\Theta}(\varepsilon^{-(3\alpha C_{\mathrm{LSI}}/2+5)} R_2(\rho_0 \| \pi)^{3C_{\mathrm{LSI}}} d(M+L))$$

Example: generalized Cauchy

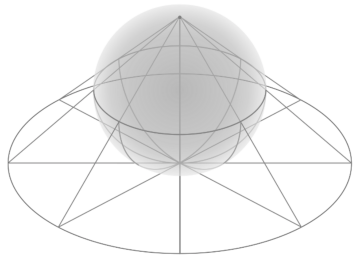
$$\pi(dx) \propto (1 + \|x\|^2)^{-\frac{d+\nu}{2}} dx, \quad \nu > 2$$

- LMC: $\mathcal{O}\left(d^5 \varepsilon^{-4/\nu+1} \exp\left[4(2q-1)\nu^{-1} R_\infty(\rho_0 \| \pi)\right]\right)$ in Renyi
- TULA: $\widetilde{\mathcal{O}}(e^{2d} d^{d+1} \varepsilon^{-1} \ln(D(\rho_0 \| \pi)\varepsilon^{-1}))$ in KL
- **Our work:** $\widetilde{\mathcal{O}}((d+\nu)^6 d^4 \varepsilon^{-17/4} R_\alpha(\rho_0 \| \pi))$ for $\nu > 6$ using $\kappa(x) = (1 + \|x\|^2)^2$

Geometry of the diffusion

$$dX_t = \Big(-\kappa(X_t)\nabla f(X_t) + \nabla\kappa(X_t)\Big)dt + \sqrt{2\kappa(X_t)}dB_t$$

$$\implies \text{Riemannian Langevin diffusion on } (\mathbb{R}^d, \delta_{ij}/\kappa)$$



$\kappa(x) = (1 + \|x\|^2)^2 \implies$ metric induced by stereographic projection

Geometry of the algorithm

Sampling methods that induce a Riemannian structure:

- Standard LMC: $G = \delta_{ij}$
- **Our algorithm**: $G = (\kappa)^{-1}\delta_{ij}$
- Mirror LMC: $G = (D^2\Phi)^{-1}\delta_{ij}$
- Variable transformation with $h : (M, g) \to \mathbb{R}^d$: $G = (h^{-1}) * g$

Variable transformation geometry $\subseteq$ weighted geometry $G = (\kappa)^{-1}\delta_{ij}$

Geometry of the algorithm

Sampling methods that induce a Riemannian structure:

- Standard LMC: $G = \delta_{ij}$
- **Our algorithm**: $G = (\kappa)^{-1}\delta_{ij}$
- Mirror LMC: $G = (D^2\Phi)^{-1}\delta_{ij}$
- Variable transformation with $h : (M, g) \to \mathbb{R}^d$: $G = (h^{-1}) * g$

Variable transformation geometry $\subseteq$ weighted geometry $G = (\kappa)^{-1}\delta_{ij}$

Summary

- We propose a generalisation of LMC that can adapt to different tail-growths
- Based on weighted functional inequalities and adaptive EM
- Analysis gives polynomial dependence on d and ε^{-1} and initial Renyi in heavy-tailed settings

Future directions Riemmanian LMC based on Adaptive EM schemes:

- Simulating Riemannian Langevin diffusions based on adaptive EM schemes

  Super light-tailed targets

  Non-smooth potentials

  Sampling from Riemannian manifolds
- Design accept/reject mechanism