Provable Benefit of Annealed Langevin Monte Carlo for Non-log-concave Sampling

Georgia Institute of Technology

Wei Guo Molei Tao Yongxin Chen

Challenges for Non-log-concave Sampling

Aim: we study **sampling** from a probability distribution $\pi \propto \mathrm{e}^{-V}$ on $\mathbb{R}^d$, an important task in computational statistics, Bayesian inference, statistical physics, etc.

The **Langevin diffusion (LD)** is the SDE $\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ t \in [0, \infty)$. Its Euler-Maruyama discretization is known as the **Langevin Monte Carlo (LMC)** algorithm:

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}\mathcal{N}(0, hI),\ k = 0, 1, \ldots.$$

When π has good isoperimetry conditions (e.g., being log-concave or satisfying Poincaré or log-Sobolev inequalities (PI/LSI)), LD converges exponentially fast in KL; furthermore, when V is β-smooth, LMC also converges exponentially with a bias that vanishes when $h \to 0$.

However, the effectiveness of LMC diminishes when dealing with target distributions that are **multimodal** (such as mixtures of Gaussians): the sampler often becomes confined to a single mode.

Annealing to Address Multimodality

Annealing: construct a sequence of distributions $\pi_0, \pi_1, \ldots, \pi_M$ that interpolates between an easily samplable distribution π_0 (e.g., $\mathcal{N}(0, I)$) and the target distribution $\pi_M = \pi$. Start with samples from π_0 and progressively sample from each π_i until π_M is reached.

Our contribution: we propose a novel strategy to analyze the non-asymptotic complexity bounds of annealed LMC algorithm, bypassing the need for assumptions such as log-concavity or isoperimetry.

Wasserstein Distance, Metric Derivative, and Action

For probability measures μ, ν on $\mathbb{R}^d$, the **Wasserstein-2 distance** is defined as $W_2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int \|x - y\|^2 \gamma(\mathrm{d}x, \mathrm{d}y)\right)^{\frac{1}{2}}$, where $\Pi(\mu, \nu)$ is the set of all couplings of (μ, ν).

A vector field $v = (v_t : \mathbb{R}^d \to \mathbb{R}^d)_{t\in[a,b]}$ on $\mathbb{R}^d$ **generates** a curve of probability measures $\rho = (\rho_t)_{t\in[a,b]}$ if the **continuity equation** $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0,\ t \in [a, b]$ holds.

The **metric derivative** of ρ at $t \in [a, b]$ is defined as $|\dot\rho|_t := \lim_{\delta\to0} \frac{W_2(\rho_{t+\delta}, \rho_t)}{|\delta|}$, which can be interpreted as the "speed" of this curve. If $|\dot\rho|_t$ exists and is finite for a.e. $t \in [a, b]$, we say that ρ is **absolutely continuous (AC)**. Its **action** is defined as $\int_a^b |\dot\rho|_t^2 \mathrm{d}t$, which is a key property characterizing the effectiveness of a curve in annealed sampling.

Lemma (Relationship between Metric Derivative and Continuity Equation [AGS08])

For an AC curve of probability measures $(\rho_t)_{t\in[a,b]}$, any vector field $(v_t)_{t\in[a,b]}$ that generates $(\rho_t)_{t\in[a,b]}$ satisfies $|\dot\rho|_t \le \|v_t\|_{L^2(\rho_t)}$ for a.e. $t \in [a, b]$. Moreover, there exists a unique vector field $(v_t^*)_{t\in[a,b]}$ generating $(\rho_t)_{t\in[a,b]}$ that satisfies $|\dot\rho|_t = \|v_t^*\|_{L^2(\rho_t)}$ for a.e. $t \in [a, b]$.

Properties of the Action

Given an AC curve of probability measures $(\rho_t)_{t\in[0,1]}$, and let $\mathcal{A}$ be its action. Then

- *$\mathcal{A} \ge W_2^2(\rho_0, \rho_1)$. The equality is attained when $(\rho_t)_{t\in[0,1]}$ is a constant-speed Wasserstein geodesic, i.e., let (X_0, X_1) follow the optimal coupling of (ρ_0, ρ_1) and define $\rho_t = \mathrm{Law}((1-t)X_0 + tX_1)$.*
- *If ρ_t satisfies $C_{\mathrm{LSI}}(\rho_t)$-LSI for all t, then $\mathcal{A} \le \int_0^1 C_{\mathrm{LSI}}(\rho_t)^2 \|\partial_t \nabla \log \rho_t\|^2_{L^2(\rho_t)} \mathrm{d}t$.*
- *If ρ_t satisfies $C_{\mathrm{PI}}(\rho_t)$-PI for all t, then $\mathcal{A} \le \int_0^1 2C_{\mathrm{PI}}(\rho_t) \|\partial_t \log \rho_t\|^2_{L^2(\rho_t)} \mathrm{d}t$.*

Problem Setting

We consider a curve of probability measures $(\pi_\theta)_{\theta\in[0,1]}$ from prior to target distribution.

- **Assump. 1**: each π_θ has a finite second-order moment, and the curve $(\pi_\theta)_{\theta\in[0,1]}$ is AC with finite action $\mathcal{A} = \int_0^1 |\dot\pi|_\theta^2 \mathrm{d}\theta$.
- **Assump. 2**: V is β-smooth, and there exists a global minimizer x_* of V such that $\|x_*\| \le R$. Moreover, π has finite second-order moment.

Analysis of Annealed Langevin Dynamics (ALD)

ALD: with reparametrized curve $(\widetilde\pi_t := \pi_{t/T})_{t\in[0,T]}$ for some duration T, run the following SDE:

$$\mathrm{d}X_t = \nabla \log \widetilde\pi_t(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ t \in [0, T];\ X_0 \sim \widetilde\pi_0 \implies X_T \sim \nu^{\mathrm{ALD}}.$$

Theorem (Convergence Guarantee of ALD)

Under assump. 1, when choosing $T = \frac{\mathcal{A}}{4\varepsilon^2}$, it follows that $\mathrm{KL}(\pi\|\nu^{\mathrm{ALD}}) \le \varepsilon^2$.

Sketch of Proof: Girsanov Theorem + Metric Derivative

Let $\mathbb{Q}$ be the path measure of ALD, and define $\mathbb{P}$ as the path measure of the reference SDE

$$\mathrm{d}X_t = (\nabla \log \widetilde\pi_t + v_t)(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ X_0 \sim \widetilde\pi_0,\ t \in [0, T].$$

The vector field v is designed such that $X_t \sim \widetilde\pi_t$ for all t, which happens if.f. v generates $\widetilde\pi$. By Girsanov theorem, $\mathrm{KL}(\pi\|\nu^{\mathrm{ALD}}) \le \mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = \frac{1}{4}\mathbb{E}_{\mathbb{P}} \int_0^T \|v_t(X_t)\|^2 \mathrm{d}t = \frac{1}{4}\int_0^T \|v_t\|^2_{L^2(\widetilde\pi_t)} \mathrm{d}t$. Choosing v_t that minimizes the $L^2(\widetilde\pi_t)$-norm yields $\frac{\mathcal{A}}{4T}$.

Analysis of Annealed Langevin Monte Carlo

Theorem (Convergence Guarantee of ALMC)

Under Assumps. 1 and 2, consider the geometric interpolation $\pi_\theta \propto \exp\left(-\eta(\theta)V - \frac{\lambda(\theta)}{2}\|\cdot\|^2\right)$, where the annealing schedules $\eta(\cdot)$ and $\lambda(\cdot)$ satisfy $\eta_0 = \eta(0) \nearrow \eta(1) = 1$ and $\lambda_0 = \lambda(0) \searrow \lambda(1) = 0$. Then, ALMC generates a distribution ν^{ALMC} satisfying $\mathrm{KL}(\pi\|\nu^{\mathrm{ALMC}}) \le \varepsilon^2$ within $\widetilde O\left(\frac{d\beta^2\mathcal{A}^2}{\varepsilon^6}\right)$ calls to the oracle of V and ∇V in expectation.

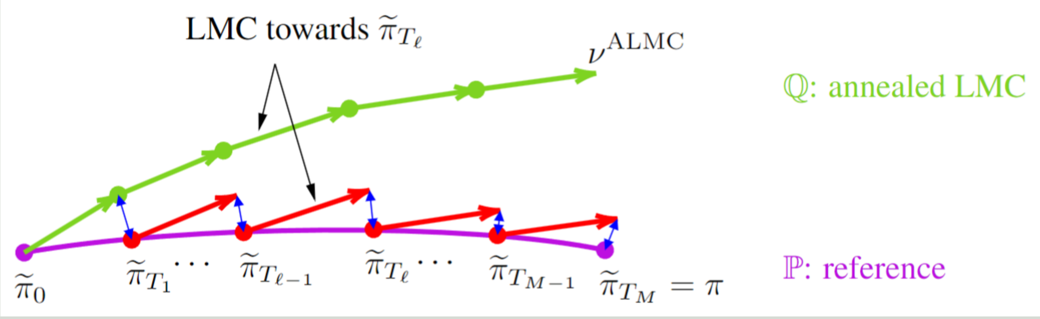

Figure 1. Illustration of ALMC. The ℓ-th green arrow represents one step of LMC towards $\widetilde\pi_{T_\ell}$, while each red arrow corresponds to the application of the same transition kernel, initialized at $\widetilde\pi_{T_{\ell-1}}$ on the reference trajectory $\mathbb{P}$ (in purple). To evaluate $\mathrm{KL}(\mathbb{P}\|\mathbb{Q})$, we only need to bound the aggregate KL divergence across each small interval (i.e., the sum of the blue "distances").

An Example of Mixture of Gaussian

Consider a d-dimensional mixture of Gaussian defined by $\pi = \sum_{i=1}^N p_i \mathcal{N}(y_i, \beta^{-1}I)$, where $\|y_i\| = r$ for all i. The potential $V = -\log\pi$ is B-smooth, where $B = \beta(4r^2\beta + 1)$. With an annealing schedule defined by $\eta(\cdot) \equiv 1$ and $\lambda(\theta) = dB(1-\theta)^\gamma$ for some $1 \le \gamma = O(1)$, we have $\mathcal{A} = O\left(d(r^2\beta + 1)\left(r^2 + \frac{d}{\beta}\right)\right)$.

In the special case $N = 2$, $y_1 = -y_2$, and $r^2 \gg \beta^{-1}$, the complexity to obtain an ε-accurate sample in TV distance is $\widetilde O(d^3\beta^2 r^4(r^4\beta^2 \vee d^2)\varepsilon^{-6})$; in contrast, as the LSI constant of π is $\Omega(\mathrm{e}^{\Theta(\beta r^2)})$, existing analysis of LMC can only provide an exponential complexity $\widetilde O(\mathrm{e}^{\Theta(\beta r^2)} d\varepsilon^{-2})$.

Comparison of Complexity Bounds

Table 1. Comparison of oracle complexities in terms of d, ε, and the LSI constant for sampling from $\pi \propto \mathrm{e}^{-V}$.

Algorithm	Isoperimetric Assumptions	Other Assumptions	Criterion	Complexity
LMC [VW19]	C-LSI	Potential smooth	ε^2, $\mathrm{KL}(\cdot\|\pi)$	$\widetilde O(C^2 d\varepsilon^{-2})$
PS [FYC23]	C-LSI	Potential smooth	ε, TV	$\widetilde O(Cd^{1/2}\log\varepsilon^{-1})$
STLMC [GLR18]	/	Translated mixture of a well-conditioned distribution	ε, TV	$O(\mathrm{poly}(d, \varepsilon^{-1}))$
RDMC [HDH+24]	/	Potential smooth, nearly convex at ∞	ε, TV	$O(\mathrm{poly}(d)\mathrm{e}^{\mathrm{poly}(\varepsilon^{-1})})$
RS-DMC [HZD+24]	/	Potential smooth	ε^2, $\mathrm{KL}(\pi\|\cdot)$	$\exp(O(\log^3 d\varepsilon^{-2}))$
ZOD-MC [HRT24]	/	Potential growing at most quadratically	ε, $\mathrm{TV} + \mathrm{W}_2$	$\exp(\widetilde O(d)O(\log\varepsilon^{-1}))$
ALMC (ours)	/	Potential smooth	ε^2, $\mathrm{KL}(\pi\|\cdot)$	$\widetilde O(d\mathcal{A}(d)^2\varepsilon^{-6})$

Conclusion and Future Work

Our framework can also be used to analyze the statistical efficiency of normalizing constant (free energy) estimation using Jarzynski equality and annealed importance sampling, see [GTC25].

References

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2 edition, 2008.

[FYC23] Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1473–1521. PMLR, 12–15 Jul 2023.

[GLR18] Rong Ge, Holden Lee, and Andrej Risteski. Simulated tempering Langevin Monte Carlo II: An improved proof using soft Markov chain decomposition. *arXiv preprint arXiv:1812.00793*, 2018.

[GTC25] Wei Guo, Molei Tao, and Yongxin Chen. Complexity analysis of normalizing constant estimation: from jarzynski equality to annealed importance sampling and beyond. *arXiv preprint arXiv:2502.04575*, 2025.

[HDH+24] Xunpeng Huang, Hanze Dong, Yifan Hao, Yian Ma, and Tong Zhang. Reverse diffusion Monte Carlo. In *The Twelfth International Conference on Learning Representations*, 2024.

[HRT24] Ye He, Kevin Rojas, and Molei Tao. Zeroth-order sampling methods for non-log-concave distributions: Alleviating metastability by denoising diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[HZD+24] Xunpeng Huang, Difan Zou, Hanze Dong, Yi-An Ma, and Tong Zhang. Faster sampling without isoperimetry via diffusion-based Monte Carlo. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2438–2493. PMLR, 30 Jun–03 Jul 2024.

[VW19] Santosh S. Vempala and Andre Wibisono. *Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices*, volume 32. Curran Associates, Inc., 2019.

Paper link on ArXiv: https://arxiv.org/abs/2407.16936

Accepted at the Thirteenth International Conference on Learning Representations (ICLR 2025), Singapore

Contact: {wei.guo,mtao,yongchen}@gatech.edu