Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices

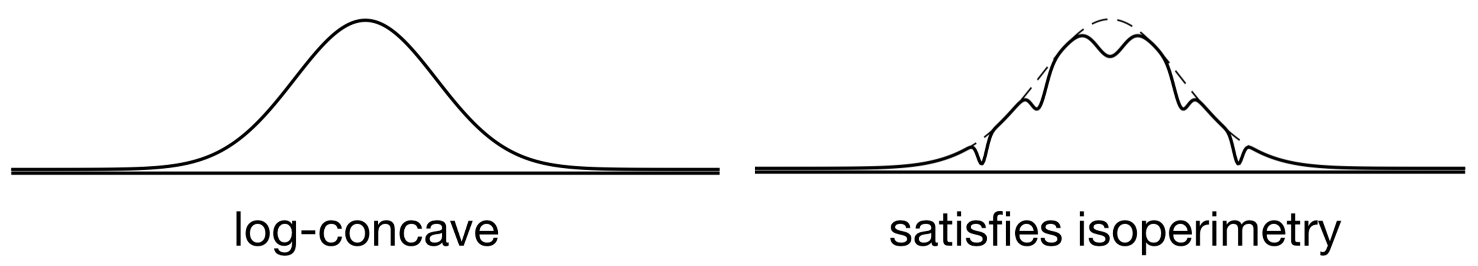
Santosh S. Vempala (Georgia Tech), **Andre Wibisono** (Georgia Tech)

Sampling Problem

Goal: Sample from a distribution $\nu = e^{-f}$ on Euclidean space $\mathbb{R}^n$.

Simple case: when ν is *log-concave* (f is convex), analogous to *convex* optimization. But many distributions in practice are not log-concave (e.g., can be multimodal).

We study sampling under **isoperimetry**: log-Sobolev inequality (LSI) or Poincaré inequality. Isoperimetry is a natural relaxation of log-concavity; it is more stable (preserved under Lipschitz mapping and bounded perturbation), and allows fast sampling in continuous time.



Unadjusted Langevin Algorithm

We study the **Unadjusted Langevin Algorithm (ULA)**:

$$x_{k+1} = x_k - \epsilon \nabla f(x_k) + \sqrt{2\epsilon} z_k$$

where $\epsilon > 0$ is step size and $z_k \sim \mathcal{N}(0, I)$ is independent Gaussian.

ULA is a discretization of **Langevin dynamics** in continuous time:

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

where $(W_t)_{t \ge 0}$ is the standard Brownian motion in $\mathbb{R}^n$.

Log-Sobolev Inequality

We say ν satisfies **log-Sobolev inequality (LSI)** with constant $\alpha > 0$ if for all probability distribution ρ:

$$H_\nu(\rho) \le \tfrac{1}{2\alpha} J_\nu(\rho).$$

Here $H_\nu(\rho)$ is the **KL divergence (relative entropy)**:

$$H_\nu(\rho) = \textstyle\int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)}\, dx$$

and $J_\nu(\rho)$ is the **relative Fisher information**:

$$J_\nu(\rho) = \textstyle\int_{\mathbb{R}^n} \rho(x) \left\| \nabla \log \frac{\rho(x)}{\nu(x)} \right\|^2 dx.$$

If $\nu = e^{-f}$ is α-strongly log-concave (f is α-strongly convex), then ν satisfies α-LSI. But LSI is more general than strong log-concavity.

Convergence of KL Divergence

We recall that when ν satisfies α-LSI, along the Langevin dynamics in continuous time, KL divergence converges exponentially fast: $H_\nu(\rho_t) \le e^{-2\alpha t} H_\nu(\rho_0)$.

We prove a similar convergence guarantee along ULA in discrete time up to the biased limit, when ν satisfies LSI and smoothness.

We say $\nu = e^{-f}$ is **L-smooth** if ∇f is L-Lipschitz ($-LI \preceq \nabla^2 f \preceq LI$). But note we do not assume f is convex.

Theorem: Assume ν satisfies α-LSI and is L-smooth. Then ULA with step size $0 < \epsilon \le \frac{\alpha}{L^2}$ satisfies:

$$H_\nu(\rho_k) \le e^{-\alpha \epsilon k} H_\nu(\rho_0) + \frac{\epsilon n L^2}{\alpha}.$$

Suppose we start from $x_0 \sim \rho_0 = \mathcal{N}(x^*, \frac{1}{L} I)$ where x^* is a stationary point for f ($\nabla f(x^*) = 0$), so $H_\nu(\rho_0) = \tilde{O}(n)$. The theorem above implies the following iteration complexity for ULA.

Corollary: Assume ν satisfies α-LSI and is L-smooth. For $\delta > 0$, to reach $H_\nu(\rho_k) \le \delta$, it suffices to run ULA with step size $\epsilon = \Theta(\frac{\alpha\delta}{nL^2})$ for the following number of iterations:

$$k = \tilde{O}\left(\frac{nL^2}{\alpha^2 \delta}\right).$$

This is the same complexity as previous results for ULA under strong log-concavity, but our result holds under more general condition (LSI).

Analysis

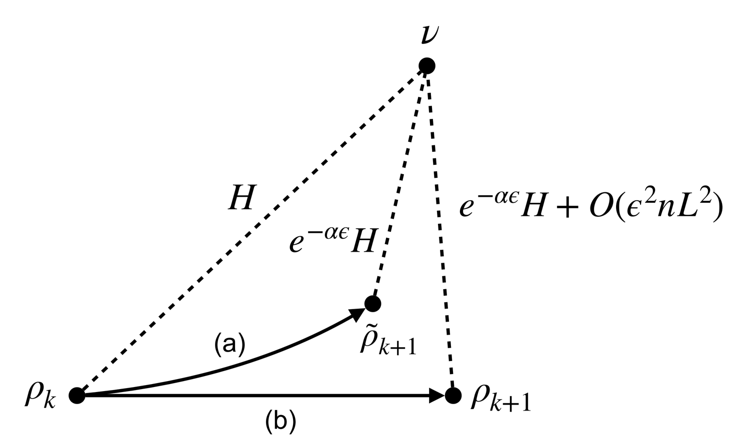
We show KL divergence decreases by a constant factor in each step of ULA, with an additional $O(\epsilon^2)$ error term. Iterating this bound yields the result above with $O(\epsilon)$ bias.

Lemma: Assume ν satisfies α-LSI and is L-smooth. Then ULA with step size $0 < \epsilon \le \frac{\alpha}{L^2}$ satisfies

$$H_\nu(\rho_{k+1}) \le e^{-\alpha\epsilon} H_\nu(\rho_k) + \epsilon^2 n L^2.$$

Proof idea:

1. We compare one step of ULA with the Langevin dynamics.
2. We use Talagrand's inequality to bound the difference.



Rényi Divergence

Rényi divergence of order $q > 0$ ($q \neq 1$) of ρ with respect to ν is:

$$R_{q,\nu}(\rho) = \frac{1}{q-1} \log \int_{\mathbb{R}^n} \frac{\rho(x)^q}{\nu(x)^{q-1}}\, dx$$

The case $q = 1$ recovers KL divergence: $\lim_{q \to 1} R_{q,\nu}(\rho) = H_\nu(\rho)$.

Rényi divergence is a family of generalization of KL divergence which is stronger ($q \mapsto R_{q,\nu}$ is increasing). It has fundamental applications in statistics, physics, computer science (e.g., for differential privacy).

Convergence of Rényi Divergence

We can show when ν satisfies α-LSI, Rényi divergence converges exponentially fast along the Langevin dynamics: $R_{q,\nu}(\rho_t) \le e^{-\frac{2\alpha}{q}t} R_{q,\nu}(\rho_0)$.

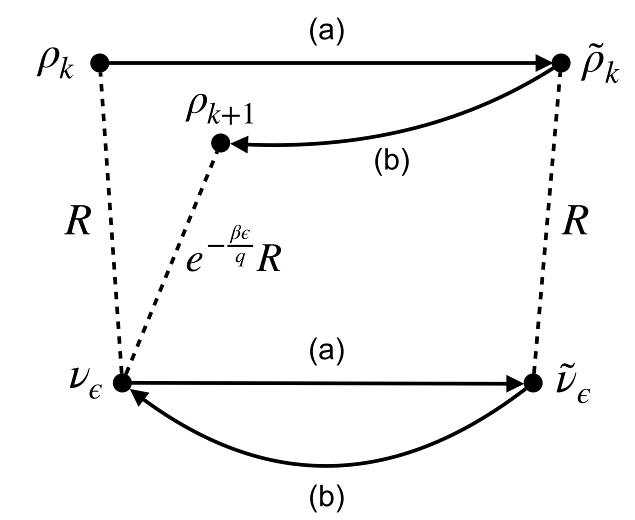
Theorem: Assume ν is L-smooth, the biased limit ν_ϵ of ULA satisfies β-LSI, and $0 < \epsilon \le \min\{\frac{1}{L}, \frac{1}{\beta}\}$. For $q > 1$, ULA satisfies:

$$R_{q,\nu}(\rho_k) \le \left(\frac{q - \frac{1}{2}}{q-1}\right) R_{2q,\nu_\epsilon}(\rho_0) e^{-\frac{\beta\epsilon k}{2q}} + R_{2q-1,\nu}(\nu_\epsilon).$$

Iteration complexity is determined by the bias $R_{2q-1,\nu}(\nu_\epsilon)$.

Proof idea:

1. We show Rényi divergence converges exponentially fast along ULA to the biased limit ν_ϵ.
2. We use a decomposition result for Rényi divergence (triangle inequality with increasing q).



Poincaré inequality

We prove similar convergence guarantee for Rényi divergence under Poincaré inequality (weaker than LSI). The convergence is initially linear before becoming exponential. The iteration complexity for ULA under Poincaré is a factor of n larger than the complexity under LSI.

Presented at the 33rd Conference on Neural Information Processing Systems (NeurIPS), December 8–14, 2019, Vancouver, Canada.