Mixing Time of the Proximal Sampler in Relative Fisher Information via Strong Data Processing Inequality

Andre Wibisono

Yale University

Optimization and Learning: Theory and Applications

CRM, Montreal, May 29, 2025

Based on Joint Work with





1. [Vempala, **W.**, "Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices", NeurIPS 2019]
2. [Chen, Chewi, Salim, **W.**, "Improved Analysis for a Proximal Algorithm for Sampling", COLT 2022]
3. [Mitra, **W.**, "Fast Convergence of Φ-Divergence along the Unadjusted Langevin Algorithm and Proximal Sampler", ALT 2025]
4. [**W.**, "Mixing Time of the Proximal Sampler in Relative Fisher Information via Strong Data Processing Inequality", COLT 2025]

Plan

Sampling in Continuous Time via Langevin Dynamics

Discrete-time Algorithm 1: Unadjusted Langevin Algorithm

Discrete-time Algorithm 2: Proximal Sampler

Proof Technique via Strong Data Processing Inequality

Sampling Problem

Goal: Sample from a probability distribution ν on $\mathbb{R}^d$ with density

$$\nu(x) \propto \exp(-f(x))$$

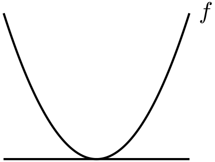
- Assume $f\colon \mathbb{R}^d \to \mathbb{R}$ is twice-differentiable
- Assume we can evaluate score function $\nabla f(x)$, but don't know the normalizing constant $\int_{\mathbb{R}^d} \exp(-f(x))dx < \infty$.
- Useful for Bayesian inference, numerical integration, uncertainty quantification, differential privacy, ...

  e.g.: $p_{\text{posterior}}(x \mid y) \propto p_{\text{prior}}(x) \cdot p_{\text{likelihood}}(y \mid x)$

Optimization and Sampling

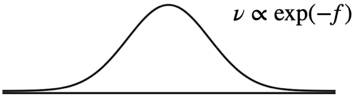
Optimization

$$\min_{x \in \mathbb{R}^d} f(x)$$



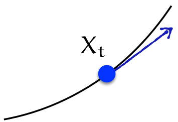
Sampling

$$\nu(x) \propto \exp(-f(x))$$



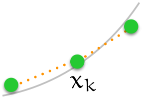
Dynamics and Algorithms for Optimization $\min_{x \in \mathbb{R}^d} f(x)$

Gradient Flow (GF)



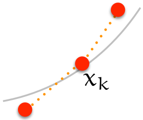
$$\dot{X}_t = -\nabla f(X_t)$$

Gradient Descent (GD)



$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

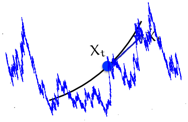
Proximal Point (PP)



$$\begin{aligned} x_{k+1} &= x_k - \eta \nabla f(x_{k+1}) \\ &= \arg\min_{x \in \mathbb{R}^d} \; f(x) + \frac{1}{2\eta} \|x - x_k\|^2 \end{aligned}$$

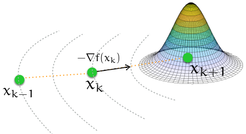
Dynamics and Algorithms for Sampling $\nu \propto \exp(-f)$

Langevin Dynamics (LD)



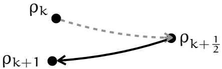
$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$$

Unadjusted Langevin Algorithm (ULA)



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

Proximal Sampler (PS)



$$y_k = x_k + \sqrt{\eta}\,z_k$$

$$x_{k+1} \sim \exp\left(-f(x) - \frac{1}{2\eta}\|x - y_k\|^2\right)$$

Sampling via Langevin Dynamics

To sample from $\nu \propto e^{-f}$, the **Langevin dynamics** is the SDE:

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

where $(W_t)_{t \geq 0}$ is the standard Brownian motion on $\mathbb{R}^d$.

- Target distribution ν is stationary, and $X_t \sim \rho_t \to \nu$ as $t \to \infty$
- Density $\rho_t \colon \mathbb{R}^d \to \mathbb{R}$ evolves via the Fokker-Planck equation:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu} \right)$$

- Optimization meaning: In the space of probability distributions $\mathcal{P}(\mathbb{R}^d)$ with Wasserstein $\mathcal{W}_2$ metric, this is gradient flow for minimizing KL divergence [Jordan, Kinderlehrer, Otto '98]

$$\dot{\rho}_t = -\mathrm{grad}_{\mathcal{W}_2} \mathrm{KL}(\rho_t \,\|\, \nu)$$

KL Divergence, Fisher Information, De Bruijn's Identity

- Kullback-Leibler (KL) Divergence between ρ and ν on $\mathbb{R}^d$ is:

$$\mathsf{KL}(\rho \,\|\, \nu) = \mathbb{E}_\rho \left[\log \frac{\rho}{\nu} \right]$$

  - $\mathsf{KL}(\rho \,\|\, \nu) \geq 0$, and $\mathsf{KL}(\rho \,\|\, \nu) = 0$ iff $\rho = \nu$.

- The Relative Fisher Information between ρ and ν on $\mathbb{R}^d$ is:

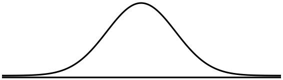
$$\mathsf{FI}(\rho \,\|\, \nu) = \mathbb{E}_\rho \left[\left\| \nabla \log \frac{\rho}{\nu} \right\|^2 \right]$$

- de Bruijn's identity: If ρ_t evolves along Langevin dynamics:

$$\frac{d}{dt} \mathsf{KL}(\rho_t \,\|\, \nu) = -\mathsf{FI}(\rho_t \,\|\, \nu)$$

Definitions: SLC and LSI Distributions

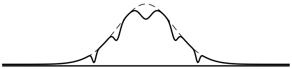
Def: $\nu \propto e^{-f}$ is α-strongly log-concave (SLC)
if f is α-strongly convex ($\nabla^2 f(x) \succeq \alpha I$).



Optimization meaning: $\rho \mapsto \mathsf{KL}(\rho \,\|\, \nu)$ is α-strongly convex on $(\mathcal{P}(\mathbb{R}^d), \mathcal{W}_2)$.

Def: ν satisfies α-log-Sobolev inequality (LSI)
if for all probability distributions $\rho \ll \nu$:

$$\mathsf{FI}(\rho \,\|\, \nu) \geq 2\alpha\, \mathsf{KL}(\rho \,\|\, \nu)$$



- **Optimization meaning:** α-Polyak-Łojaciewicz (PL) condition:

$$\|\mathsf{grad}_{\mathcal{W}_2,\rho}\, \mathsf{KL}(\rho \,\|\, \nu)\|_\rho^2 \;\geq\; 2\alpha\, \mathsf{KL}(\rho \,\|\, \nu).$$

- **Lemma:** α-SLC $\Rightarrow$ α-LSI [Bakry-Émery '85]
- LSI is stable under bounded perturbation [Holley-Stroock], Lipschitz mapping

Mixing Time of Langevin Dynamics

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

If ν is α-strongly log-concave (f is α-strongly convex), then:

- Contraction in $\mathcal{W}_2$ distance: If ρ_t, γ_t evolve along Langevin:

$$\mathcal{W}_2(\rho_t, \gamma_t)^2 \le e^{-2\alpha t}\, \mathcal{W}_2(\rho_0, \gamma_0)^2$$

- Convergence in relative Fisher information to $\nu \propto e^{-f}$:

$$\mathsf{FI}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\, \mathsf{FI}(\rho_0 \,\|\, \nu)$$

If ν satisfies α-log-Sobolev inequality (LSI), then:

- Exponential convergence in KL (also Rényi) divergence:

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\, \mathsf{KL}(\rho_0 \,\|\, \nu)$$

Mixing Time of Langevin Dynamics: Optimization View

Langevin Dynamics:

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

- ν is α-strongly log-concave:

$$\mathcal{W}_2(\rho_t, \gamma_t)^2 \le e^{-2\alpha t}\, \mathcal{W}_2(\rho_0, \gamma_0)^2$$

- ν is α-strongly log-concave:

$$\mathsf{FI}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\, \mathsf{FI}(\rho_0 \,\|\, \nu)$$

- ν satisfies α-LSI:

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\, \mathsf{KL}(\rho_0 \,\|\, \nu)$$

Gradient Flow:

$$\dot{X}_t = -\nabla F(X_t)$$

- F is α-strongly convex:

$$\|X_t - Y_t\|^2 \le e^{-2\alpha t}\, \|X_0 - Y_0\|^2$$

- F is α-strongly convex:

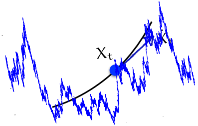
$$\|\nabla F(X_t)\|^2 \le e^{-2\alpha t}\, \|\nabla F(X_0)\|^2$$

- F satisfies α-PL $(\min F = 0)$:

$$F(X_t) \le e^{-2\alpha t}\, F(X_0)$$

Mixing Time of Langevin Dynamics: To Discrete Time?

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



Good mixing time of Langevin dynamics under SLC/LSI

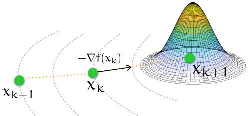
- ($\Leftrightarrow$ Convergence of Gradient flow under strong convexity/PL)
- Langevin also has good convergence in Rényi and Φ-divergence

---

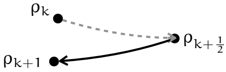
But these are in continuous time! What about in discrete time?

1. Unadjusted Langevin Algorithm, which is explicit but biased.
2. Proximal Sampler, which is implicit but unbiased.

Unadjusted Langevin Algorithm (ULA)



Proximal Sampler (PS)



Plan

Sampling in Continuous Time via Langevin Dynamics

Discrete-time Algorithm 1: Unadjusted Langevin Algorithm

Discrete-time Algorithm 2: Proximal Sampler

Proof Technique via Strong Data Processing Inequality

Optimization & Sampling in Discrete Time

Gradient Flow:

$$\dot{X}_t = -\nabla F(X_t)$$

- F satisfies α-PL $(\min F = 0)$:

$$F(X_t) \le e^{-2\alpha t} F(X_0)$$

Langevin Dynamics:

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

- ν satisfies α-LSI:

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\, \mathsf{KL}(\rho_0 \,\|\, \nu)$$

Gradient Descent:

$$x_{k+1} = x_k - \eta \nabla F(x_k)$$

- F is α-PL & L-smooth, $\eta \le \frac{1}{2L}$:

$$F(x_k) \le (1 - \alpha\eta)^k\, F(x_0)$$

?

Optimization & Sampling in Discrete Time

Gradient Flow:

$$\dot{X}_t = -\nabla F(X_t)$$

- F satisfies α-PL $(\min F = 0)$:

$$F(X_t) \le e^{-2\alpha t}\, F(X_0)$$

Langevin Dynamics:

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

- ν satisfies α-LSI:

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\, \mathsf{KL}(\rho_0 \,\|\, \nu)$$

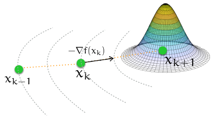
Gradient Descent:

$$x_{k+1} = x_k - \eta \nabla F(x_k)$$

- F is α-PL & L-smooth, $\eta \le \frac{1}{2L}$:

$$F(x_k) \le (1 - \alpha\eta)^k\, F(x_0)$$

Unadjusted Langevin Algorithm?



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

Unadjusted Langevin Algorithm

The **Unadjusted Langevin Algorithm (ULA)** for $\nu \propto e^{-f}$ is:

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

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

- As $\eta \to 0$, ULA recovers the Langevin dynamics.
- For fixed $\eta > 0$, ULA is biased: $x_k \sim \rho_k \xrightarrow{k \to \infty} \nu_\eta \neq \nu$
  - E.g., if $\nu = \mathcal{N}\left(0, \frac{1}{\alpha} I\right)$, then $\nu_\eta = \mathcal{N}\left(0, \frac{1}{\alpha(1 - \frac{1}{2}\eta\alpha)} I\right)$.
  - $\Rightarrow$ Low-accuracy iteration complexity guarantee

Example: Gaussian Target

Suppose $f(x) = \frac{\alpha}{2}\|x\|^2$ so $\nu \propto e^{-f} = \mathcal{N}(0, \alpha^{-1} I)$ on $\mathbb{R}^d$.

Suppose $X_0 \sim \rho_0 = \mathcal{N}(m_0, \sigma_0^2 I)$ for some $m_0 \in \mathbb{R}^d$, $\sigma_0^2 > 0$.

1. Continuous-time Langevin dynamics:

$$\rho_t = \mathcal{N}\left(e^{-\alpha t} m_0,\ \left(e^{-2\alpha t}\sigma_0^2 + \frac{1 - e^{-2\alpha t}}{\alpha}\right) I\right)$$

2. Discrete-time ULA:

$$\rho_k = \mathcal{N}\left((1-\alpha\eta)^k m_0,\ \left((1-\alpha\eta)^{2k}\sigma_0^2 + \frac{1}{\alpha}\left(\frac{1-(1-\alpha\eta)^{2k}}{(1-\frac{1}{2}\alpha\eta)}\right)\right) I\right)$$

Unadjusted Langevin Algorithm

The **Unadjusted Langevin Algorithm (ULA)** for $\nu \propto e^{-f}$ is:

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

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

- For fixed $\eta > 0$, ULA is biased: $x_k \sim \rho_k \xrightarrow{k \to \infty} \nu_\eta \neq \nu$
  - E.g., if $\nu = \mathcal{N}\left(0, \frac{1}{\alpha} I\right)$, then $\nu_\eta = \mathcal{N}\left(0, \frac{1}{\alpha(1-\frac{1}{2}\eta\alpha)} I\right)$.
  - $\Rightarrow$ Low-accuracy iteration complexity guarantee
- Many biased convergence guarantees for f strongly convex and smooth [Dalalyan '15, Durmus & Moulines '17, Cheng & Bartlett '18, Durmus et al '19 "Analysis of Langevin Monte Carlo via convex optimization"]
- Can remove bias by: ULA + Metropolis filter = MALA
  - High-accuracy, but analysis more complicated, weaker metrics.
  - **Opt meaning:** TV projection to the space of reversible Markov chains [Billera & Diaconis, 2001]

ULA: Biased Convergence Guarantee

Theorem:[1] Assume ν is α-LSI and L-smooth ($\|\nabla^2 f\|_{\mathsf{op}} \le L$). Along ULA $x_k \sim \rho_k$ with step size $\eta \le \frac{\alpha}{L^2}$, for all $k \ge 0$:

$$\mathsf{KL}(\rho_k \,\|\nu) \le e^{-\alpha\eta k}\,\mathsf{KL}(\rho_0 \,\|\, \nu) + \frac{\eta d L^2}{\alpha}$$

$\Rightarrow$ To get $\mathsf{KL}(\rho_k \,\|\, \nu) \le \epsilon$, choose $\eta = \frac{\epsilon\alpha}{dL^2}$, and run ULA from $\rho_0 = \mathcal{N}(x^*, \frac{1}{L}I)$ for number of iterations::

$$k = O\left(\frac{1}{\alpha\eta}\log\frac{\mathsf{KL}(\rho_0 \,\|\, \nu)}{\epsilon}\right) = O\left(\frac{dL^2}{\epsilon\alpha^2}\log\frac{d}{\epsilon}\right)$$

$\therefore$ **Iteration complexity** of ULA for LSI+smooth target: $O(\mathrm{poly}(\frac{1}{\epsilon}))$

- c.f. cts-time Langevin dynamics: $t = O(\frac{1}{\alpha}\log\frac{d}{\epsilon}) = O(\log\frac{1}{\epsilon})$
- c.f. gradient descent: $k = O(\frac{L}{\alpha}\log\frac{d}{\epsilon}) = O(\log\frac{1}{\epsilon})$

[1] [Vempala, **W.**, "Rapid Convergence of ULA: Isoperimetry Suffices", NeurIPS 2019]

Why is ULA Biased?[2]

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

- Sampling is solving a composite optimization problem:

$$\min_{\rho \in \mathcal{P}(\mathbb{R}^d)} \left\{ \mathsf{KL}(\rho \,\|\, \nu) = \mathbb{E}_\rho[f] \quad - H(\rho) \right\}$$

- Langevin dynamics is running the composite gradient flow:

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

- ULA is the "Forward-Flow" discretization:
  1. Run gradient descent for minimizing $\mathbb{E}_\rho[f]$
  2. Run gradient flow for minimizing $-H(\rho)$

  Issue: Forward-Flow is biased for general optimization...

- From **Opt**: Should run "Forward-Backward" $\to$ unbiased
  - But backward method for entropy is not implementable...

[2] [**W.**, "Sampling as Optimization in the Space of Measures: Langevin Dynamics as a Composite Optimization Problem", COLT 2018]

Unbiased Discretizations of Langevin Dynamics

- The backward (proximal) method for KL divergence
  "JKO scheme" [Jordan, Kinderlehrer, Otto, 1998]

$$\rho_{k+1} = \arg\min_{\rho \in \mathcal{P}(\mathbb{R}^d)} \left\{ \mathsf{KL}(\rho \,\|\, \nu) + \frac{1}{2\eta} \mathcal{W}_2(\rho, \rho_k)^2 \right\}$$

- The Forward-Backward algorithm for KL divergence
  [Salim, Korba, Louise, NeurIPS 2020]

$$x_{k+\frac{1}{2}} = x_k - \eta \nabla f(x_k) \;\; \sim \rho_{k+\frac{1}{2}}$$

$$\rho_{k+1} = \arg\min_{\rho \in \mathcal{P}(\mathbb{R}^d)} \left\{ -H(\rho) + \frac{1}{2\eta} \mathcal{W}_2(\rho, \rho_{k+\frac{1}{2}})^2 \right\}$$

Issues: The above are not implementable as an algorithm (that maintains only a sample $x_k \sim \rho_k$), except e.g. for Gaussian target.

Plan

Sampling in Continuous Time via Langevin Dynamics

Discrete-time Algorithm 1: Unadjusted Langevin Algorithm

Discrete-time Algorithm 2: Proximal Sampler

Proof Technique via Strong Data Processing Inequality

Optimization & Sampling in Discrete Time

Gradient Flow:

$$\dot{X}_t = -\nabla F(X_t)$$

- F satisfies α-PL $(\min F = 0)$:

$$F(X_t) \le e^{-2\alpha t}\, F(X_0)$$

Langevin Dynamics:

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

- ν satisfies α-LSI:

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\, \mathsf{KL}(\rho_0 \,\|\, \nu)$$

Proximal Gradient:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} f(x) + \frac{\|x - x_k\|^2}{2\eta}$$

- F satisfies α-PL $(\min F = 0)$:

$$F(x_k) \le \frac{F(x_0)}{(1 + \alpha\eta)^{2k}}$$

?

Optimization & Sampling in Discrete Time

Gradient Flow:

$$\dot{X}_t = -\nabla F(X_t)$$

- F satisfies α-PL $(\min F = 0)$:

$$F(X_t) \le e^{-2\alpha t}\, F(X_0)$$

Langevin Dynamics:

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

- ν satisfies α-LSI:

$$\mathsf{KL}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\,\mathsf{KL}(\rho_0 \,\|\, \nu)$$

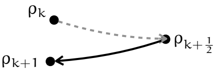
Proximal Gradient:

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^d} f(x) + \frac{\|x - x_k\|^2}{2\eta}$$

- F satisfies α-PL $(\min F = 0)$:

$$F(x_k) \le \frac{F(x_0)}{(1+\alpha\eta)^{2k}}$$

Proximal Sampler:



$$y_k = x_k + \sqrt{\eta}\, z_k$$

$$x_{k+1} \sim \exp\left(-f(x) - \frac{1}{2\eta}\|x - y_k\|^2\right)$$

Proximal Sampler

To sample from $\nu^X(x) \propto e^{-f(x)}$ on $\mathbb{R}^d$, consider joint distribution

$$\nu^{XY}(x,y) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x-y\|^2\right)$$

- Note the x-marginal is ν^X, so it suffices to sample from ν^{XY}.

Algorithm: Run Gibbs sampling on ν^{XY}.

Proximal Sampler: [Titsias, Papaspiliopoulos (2018); Lee, Shen, Tian (2021)]

1. $y_k \mid x_k \sim \nu^{Y|X=x_k} \;= \mathcal{N}(x_k,\, \eta I)$
2. $x_{k+1} \mid y_k \sim \nu^{X|Y=y_k}(x) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x-y_k\|^2\right)$

- Jointly ν^{XY}-reversible $\Rightarrow$ x-marginal is ν^X (unbiased!)
- Second step is called the Restricted Gaussian Oracle (RGO):

$$\nu^{X|Y=y}(x) \propto_x \exp\left(-f(x) - \frac{1}{2\eta}\|x-y\|^2\right)$$

Implementing the RGO

$$\nu^{X|Y=y}(x) \propto_x \exp\left(-f(x) - \frac{1}{2\eta}\|x-y\|^2\right)$$

- Assume f is L-smooth: $-LI \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$.
- If $\eta < \frac{1}{L}$, then $g_y(x) = f(x) + \frac{1}{2\eta}\|x-y\|^2$ is strongly convex and smooth with condition number $\kappa = \frac{1+\eta L}{1-\eta L}$.
- Then we can implement RGO via rejection sampling (with Gaussian proposal) with $\mathbb{E}[\#\text{ queries to } f] \leq \kappa^d$.
- If $\eta = \frac{1}{Ld}$, then $\kappa^d = \left(\frac{1+\frac{1}{d}}{1-\frac{1}{d}}\right)^d \leq O(1)$ is a constant.
- Therefore, can implement the Proximal Sampler with $\eta = \frac{1}{Ld}$.

Example: Gaussian Target

Suppose $f(x) = \frac{\alpha}{2}\|x\|^2$ so $\nu \propto e^{-f} = \mathcal{N}(0, \alpha^{-1} I)$ on $\mathbb{R}^d$.

Suppose $X_0 \sim \rho_0 = \mathcal{N}(m_0, \sigma_0^2 I)$ for some $m_0 \in \mathbb{R}^d$, $\sigma_0^2 > 0$.

1. Continuous-time Langevin dynamics:

$$\rho_t = \mathcal{N}\left(e^{-\alpha t} m_0, \left(e^{-2\alpha t}\sigma_0^2 + \frac{1 - e^{-2\alpha t}}{\alpha}\right) I\right)$$

2. Discrete-time ULA:

$$\rho_k = \mathcal{N}\left((1-\alpha\eta)^k m_0, \left((1-\alpha\eta)^{2k}\sigma_0^2 + \frac{1}{\alpha}\left(\frac{1-(1-\alpha\eta)^{2k}}{(1-\frac{1}{2}\alpha\eta)}\right)\right) I\right)$$

3. Discrete-time Proximal Sampler:

$$\rho_k = \mathcal{N}\left(\frac{m_0}{(1+\alpha\eta)^k}, \left(\frac{\sigma_0^2}{(1+\alpha\eta)^{2k}} + \frac{1}{\alpha}\left(1 - \frac{1}{(1+\alpha\eta)^{2k}}\right)\right) I\right)$$

Proximal Sampler: Unbiased Convergence Guarantees

Theorem:[3] If $\nu^X \propto e^{-f}$ satisfies α-Log Sobolev Inequality (LSI), then along the Proximal Sampler $x_k \sim \rho_k$ with step size $\eta > 0$:

$$\mathsf{KL}(\rho_k \,\|\, \nu^X) \leq \frac{\mathsf{KL}(\rho_0 \,\|\, \nu^X)}{(1+\alpha\eta)^{2k}}$$

- If f is L-smooth, with RGO via rejection sampling with $\eta = \frac{1}{Ld}$:

  To get $\mathsf{KL}(\rho_k \,\|\, \nu^X) \leq \varepsilon$, run Proximal Sampler for # of iterations:

$$k = O\left(\frac{dL}{\alpha} \log \frac{\mathsf{KL}(\rho_0 \,\|\, \nu^X)}{\varepsilon}\right) = O\left(\frac{dL}{\alpha} \log \frac{d}{\epsilon}\right)$$

  - c.f. continuous-time Langevin: $t = O(\frac{1}{\alpha} \log \frac{d}{\varepsilon})$
  - c.f. proximal gradient for optimization: $k = O(\frac{L}{\alpha} \log \frac{d}{\varepsilon})$

[3] [Chen, Chewi, Salim, **W.**, "Improved Analysis for a Proximal Algorithm for Sampling", COLT 2022]

Optimization & Sampling in Discrete Time

Gradient Flow:

$$\dot{X}_t = -\nabla F(X_t)$$

∘ F satisfies α-PL $(\min F = 0)$:

$$F(X_t) \leq e^{-2\alpha t}\, F(X_0)$$

Langevin Dynamics:

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

∘ ν satisfies α-LSI:

$$\mathsf{KL}(\rho_t \,\|\, \nu) \leq e^{-2\alpha t}\, \mathsf{KL}(\rho_0 \,\|\, \nu)$$

Proximal Gradient:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} f(x) + \frac{\|x - x_k\|^2}{2\eta}$$

∘ F satisfies α-PL $(\min F = 0)$:

$$F(x_k) \leq \frac{F(x_0)}{(1+\alpha\eta)^{2k}}$$

Proximal Sampler:

$$x_{k+1} \sim \exp\left(-f(x) - \frac{\|x - x_k + \sqrt{\eta} z_k\|^2}{2\eta}\right)$$

∘ ν satisfies α-LSI: [CCSW. '22]

$$\mathsf{KL}(\rho_k \,\|\, \nu) \leq \frac{\mathsf{KL}(\rho_0 \,\|\, \nu)}{(1+\alpha\eta)^{2k}}$$

Review: Mixing Time of Proximal Sampler

1. ν strongly log-concave $\Rightarrow$ exponential contraction in $\mathcal{W}_2$ distance
   [Lee, Shen, Tian, "Structured Logconcave Sampling with a Restricted Gaussian Oracle", COLT 2021]

2. Log-Sobolev inequality $\Rightarrow$ exp. convergence in KL, Rényi divergence
   Poincaré inequality $\Rightarrow$ exp. convergence in χ^2-divergence
   [Chen, Chewi, Salim, **W.**, "Improved Analysis for a Proximal Algorithm for Sampling", COLT 2022]

3. Φ-Sobolev inequality $\Rightarrow$ exponential convergence in Φ-divergence
   [Mitra, **W.**, "Fast Convergence of Φ-Divergence along the Unadjusted Langevin Algorithm and Proximal Sampler", ALT 2025]

4. Strongly log-concave $\Rightarrow$ exp. decay of mutual information (x_0, x_k)
   [Liang, Mitra, **W.**, "Characterizing Dependence of Samples along the Langevin Dynamics & Algorithms via Contraction of Φ-Mutual Information", COLT 2025]

5. Strongly log-concave $\Rightarrow$ exp. convergence in Fisher information
   [**W.**, "Mixing Time of Proximal Sampler in Relative Fisher Information via Strong Data Processing Inequality", COLT 2025]

Relative Fisher Information

Recall the Relative Fisher Information of ρ with respect to ν is:

$$\mathsf{FI}(\rho \,\|\, \nu) = \mathbb{E}_\rho\left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right]$$

- This is the "non-parametric" relative Fisher information (gradient ∇ is in the state variable x, not in the parameter)
- Optimization meaning: In $(\mathcal{P}(\mathbb{R}^d), \mathcal{W}_2)$:

$$\|\mathsf{grad}_{\mathcal{W}_2}\mathsf{KL}(\rho \,\|\, \nu))\|_\rho^2 = \mathsf{FI}(\rho \,\|\, \nu)$$

- ν satisfies α-LSI $\Leftrightarrow$ $\mathsf{FI}(\rho \,\|\, \nu) \geq 2\alpha\, \mathsf{KL}(\rho \,\|\, \nu)$
- ν is α-Poincaré ineq. $\Rightarrow$ $\mathsf{FI}(\rho \,\|\, \nu) \geq 4\alpha\, \mathsf{TV}(\rho \,\|\, \nu)^2$
- Can construct ρ, $\nu = \mathcal{N}(0,1)$ s.t. $\mathsf{KL}(\rho \,\|\, \nu) \leq \epsilon$, $\mathsf{FI}(\rho \,\|\, \nu) \geq \frac{1}{\epsilon}$
$\therefore$ guarantees in FI is strictly stronger than KL

Optimization & Sampling in Discrete Time

Gradient Flow:

$$\dot{X}_t = -\nabla F(X_t)$$

- F is α-strongly convex:

$$\|\nabla F(X_t)\|^2 \le e^{-2\alpha t} \|\nabla F(X_0)\|^2$$

Langevin Dynamics:

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

- ν is α-SLC:

$$\mathsf{FI}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\, \mathsf{FI}(\rho_0 \,\|\, \nu)$$

Proximal Gradient:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} f(x) + \frac{\|x - x_k\|^2}{2\eta}$$

- F is α-strongly convex:

$$\|\nabla F(x_k)\|^2 \le \frac{\|\nabla F(x_0)\|^2}{(1 + \alpha\eta)^{2k}}$$

Proximal Sampler:

$$x_{k+1} \sim \exp\left(-f(x) - \frac{\|x - x_k + \sqrt{\eta} z_k\|^2}{2\eta}\right)$$

- ν is α-SLC:

?

Mixing Time of Proximal Sampler in Fisher Information

Theorem:[4] Assume $\nu^X \propto e^{-f}$ is α-strongly log-concave. Along the discrete-time Proximal Sampler $x_k \sim \nu^X$ with step size $\eta > 0$:

$$\mathsf{FI}(\rho_k \,\|\, \nu^X) \le \frac{\mathsf{FI}(\rho_0 \,\|\, \nu^X)}{(1+\alpha\eta)^{2k}}$$

- If f is L-smooth, with RGO via rejection sampling with $\eta = \frac{1}{Ld}$:

  To get $\mathsf{FI}(\rho_k \,\|\, \nu^X) \le \varepsilon$, run Proximal Sampler for # of iterations:

$$k = O\left(\frac{dL}{\alpha} \log \frac{\mathsf{FI}(\rho_0 \,\|\, \nu^X)}{\varepsilon}\right) = O\left(\frac{dL}{\alpha} \log \frac{d}{\varepsilon}\right)$$

  - c.f. cts-time Langevin: $t = O(\frac{1}{\alpha} \log \frac{d}{\varepsilon})$
  - c.f. proximal gradient for optimization: $k = O(\frac{L}{\alpha} \log \frac{d}{\varepsilon})$

[4] [**W.**, "Mixing Time of the Proximal Sampler in Relative Fisher Information via Strong Data Processing Inequality", COLT 2025]

Optimization & Sampling in Discrete Time

Gradient Flow:

$$\dot{X}_t = -\nabla F(X_t)$$

- F is α-strongly convex:

$$\|\nabla F(X_t)\|^2 \le e^{-2\alpha t}\,\|\nabla F(X_0)\|^2$$

Langevin Dynamics:

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

- ν is α-SLC:

$$\mathsf{FI}(\rho_t \,\|\, \nu) \le e^{-2\alpha t}\,\mathsf{FI}(\rho_0 \,\|\, \nu)$$

Proximal Gradient:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^d} f(x) + \frac{\|x - x_k\|^2}{2\eta}$$

- F is α-strongly convex:

$$\|\nabla F(x_k)\|^2 \le \frac{\|\nabla F(x_0)\|^2}{(1+\alpha\eta)^{2k}}$$

Proximal Sampler:

$$x_{k+1} \sim \exp\left(-f(x) - \frac{\|x - x_k + \sqrt{\eta} z_k\|^2}{2\eta}\right)$$

- ν is α-SLC: [**W.** '25]

$$\mathsf{FI}(\rho_k \,\|\, \nu) \le \frac{\mathsf{FI}(\rho_0 \,\|\, \nu)}{(1+\alpha\eta)^{2k}}$$

Plan

Sampling in Continuous Time via Langevin Dynamics

Discrete-time Algorithm 1: Unadjusted Langevin Algorithm

Discrete-time Algorithm 2: Proximal Sampler

Proof Technique via Strong Data Processing Inequality

Proximal Sampler Decomposition

Each iteration of Proximal Sampler is a composition of two steps:

1. **Forward step:**

$$y_k \mid x_k \sim \mathcal{N}(x_k, \eta I)$$

   - **Key:** Interpret as application of Gaussian channel.

2. **Backward step:**

$$x_{k+1} \mid y_k \sim \exp\left(-f(x) - \frac{1}{2\eta}\|x - y_k\|^2\right)$$

   - **Key:** Interpret as application of reverse Gaussian channel.

To prove mixing time, we show strong data processing inequality (SDPI) for each channel.

Proximal Sampler: Forward Step

$$y_k \mid x_k \sim \mathcal{N}(x_k, \eta I)$$

- **Interpretation:** Gaussian channel
  - Run from ρ_k^X, to get $\rho_k^Y = \rho_k^X * \mathcal{N}(0, \eta I)$.
  - Run from ν^X, to get $\nu^Y = \nu^X * \mathcal{N}(0, \eta I)$.
- SDPI for Gaussian channel under LSI:

Lemma [CCSW.'22]**:** If ν^X satisfies α-LSI, then

$$\mathsf{KL}(\rho_k^Y \,\|\, \nu^Y) \le \frac{\mathsf{KL}(\rho_k^X \,\|\, \nu^X)}{1 + \alpha\eta}$$

(SDPI also holds in Rényi divergence and in all Φ-divergence.)

Proximal Sampler: Backward Step

$$x_{k+1} \mid y_k \;\sim\; \nu^{X|Y=y_k}(x) \;\propto\; \exp\left(-f(x) - \frac{1}{2\eta}\|x - y_k\|^2\right)$$

Interpretation: The distribution $\nu^{X|Y=y}$ is the output of the reverse Gaussian channel at time $t = \eta$ from $X_0 = y$:

$$dX_t = \nabla \log \nu_{\eta - t}(X_t)\, dt + dW_t$$

where $\nu_t = \nu^X * \mathcal{N}(0, tI)$ and $(W_t)_{t \geq 0}$ is Brownian motion.

- This is the same principle as Diffusion Model (DM).
- But we run for short time $\eta \sim \frac{1}{Ld}$ (vs. long time $\eta \to \infty$ for DM). We implement via rejection sampling (vs. score estimation in DM).

Proximal Sampler: Backward Step

$$x_{k+1} \mid y_k \;\sim\; \nu^{X|\{Y=y_k\}}(x) \;\propto\; \exp\left(-f(x) - \frac{1}{2\eta}\|x - y_k\|^2\right)$$

- **Interpretation:** Output of the reverse Gaussian channel

$$dX_t = \nabla \log \nu_{\eta-t}(X_t)\, dt + dW_t$$

  - Run from $X_0 = y_k \sim \rho_k^Y$ to get $X_\eta \overset{d}{=} x_{k+1} \sim \rho_{k+1}^X$.
  - Also run from $X_0^* \sim \nu^Y$ to get back $X_\eta^* \sim \nu^X$.

- (Restricted) SDPI for reverse Gaussian channel under LSI

Lemma [CCSW.'22]: If ν^X satisfies α-LSI, then

$$\mathsf{KL}(\rho_{k+1}^X \,\|\, \nu^X) \le \frac{\mathsf{KL}(\rho_k^Y \,\|\, \nu^Y)}{1 + \alpha\eta}$$

(SDPI also holds in Rényi divergence and in all Φ-divergence.)

Review: Proximal Sampler in KL/Rényi/Φ Divergence

Theorem:[5] Assume ν^X satisfies α-LSI. Then for each $k \geq 0$:

1. Forward step: From $x_k \sim \rho_k^X$ to $y_k \sim \rho_k^Y$,

$$\mathsf{KL}(\rho_k^Y \,\|\, \nu^Y) \leq \frac{\mathsf{KL}(\rho_k^X \,\|\, \nu^X)}{1 + \alpha\eta}$$

2. Backward step: From $y_k \sim \rho_k^Y$ to $x_{k+1} \sim \rho_{k+1}^X$,

$$\mathsf{KL}(\rho_{k+1}^X \,\|\, \nu^X) \leq \frac{\mathsf{KL}(\rho_k^Y \,\|\, \nu^Y)}{1 + \alpha\eta}$$

Therefore,

$$\mathsf{KL}(\rho_k^X \,\|\, \nu^X) \leq \frac{\mathsf{KL}(\rho_0^X \,\|\, \nu^X)}{(1 + \alpha\eta)^{2k}}$$

(Same analysis for Rényi divergence and Φ-divergence [Mitra, **W.** '25].)

[5] [Chen, Chewi, Salim, **W.**, "Improved Analysis for a Proximal Algorithm for Sampling", COLT 2022]

Data Processing Inequality in Relative Fisher Information?

- Data Processing Inequality (DPI) along any noisy channel:

$$D_\Phi(\rho^Y \,\|\, \nu^Y) \le D_\Phi(\rho^X \,\|\, \nu^X)$$

  - For any $\rho^Y = P^{Y|X} \circ \rho^X$ and $\nu^Y = P^{Y|X} \circ \nu^X$
  - For any Φ-divergence $D_\Phi(\rho \,\|\, \nu) = \mathbb{E}_\nu[\Phi(\frac{\rho}{\nu})]$, $\Phi \ge 0$ convex
  - Strong DPI: Strict contraction rate < 1

- Question: Do we have DPI in relative Fisher information?

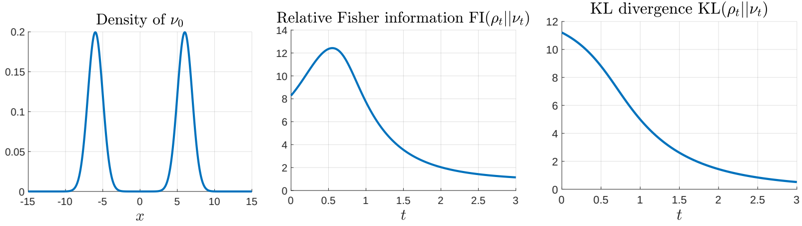
$$\mathsf{FI}(\rho^Y \,\|\, \nu^Y) \overset{?}{\le} \mathsf{FI}(\rho^X \,\|\, \nu^X)$$

  - $\mathsf{FI}(\rho \,\|\, \nu) = \mathbb{E}_\rho[\|\nabla \log \frac{\rho}{\nu}\|^2]$ is not a Φ-divergence
  - $\mathsf{FI}(\rho \,\|\, \nu)$ is convex in ρ, but *not* convex in ν
    So proof technique via Jensen's inequality fails.

Failure of DPI in Relative Fisher Information

Gaussian channel in $d = 1$ dimension: $\rho_t = \rho_0 * \mathcal{N}(0, t)$

- Let $\rho_0 = \mathcal{N}(0, 1)$
- Can construct ν_0 so that DPI in FI initially does not hold (see paper[6] for explicit expression)



- Note $\frac{d}{dt}\mathsf{KL}(\rho_t \,\|\, \nu_t) = -\frac{1}{2}\mathsf{FI}(\rho_t \,\|\, \nu_t)$

  So initially $t \mapsto \mathsf{KL}(\rho_t \,\|\, \nu_t)$ is decreasing in a concave way, then eventually in a convex way.

[6] [**W.**, "Mixing Time of the Proximal Sampler in Relative Fisher Information via SDPI", COLT 2025]

(S)DPI in FI along Gaussian Channel under SLC

Theorem:[7] If $\rho_t = \rho_0 * \mathcal{N}(0, tI)$ and $\nu_t = \nu_0 * \mathcal{N}(0, tI)$, then:

(i) If ν_0 is log-concave, then we have DPI:

$$\mathsf{FI}(\rho_t \,\|\, \nu_t) \le \mathsf{FI}(\rho_0 \,\|\, \nu_0).$$

(ii) If ν_0 is α-strongly log-concave (SLC), then we have SDPI:

$$\mathsf{FI}(\rho_t \,\|\, \nu_t) \le \frac{\mathsf{FI}(\rho_0 \,\|\, \nu_0)}{(1 + \alpha t)^2}.$$

Also have (see paper):

- Improved SDPI rate if ρ_0 satisfies Poincaré and symmetry
- Eventual SDPI if ν_0 is a log-Lipschitz perturbation of SLC.

[7] [**W.**, "Mixing Time of the Proximal Sampler in Relative Fisher Information via Strong Data Processing Inequality", 2025]

Proof of (S)DPI in FI along Gaussian Channel

Analysis via time differentiation along simultaneous heat flows.

Lemma: If $(\rho_t)_{t\geq 0}$, $(\nu_t)_{t\geq 0}$ evolve following the heat equation:

$$\partial_t \rho_t = \frac{1}{2}\Delta \rho_t \qquad \partial_t \nu_t = \frac{1}{2}\Delta \nu_t$$

then for any $t \geq 0$:

$$\frac{d}{dt}\mathsf{FI}(\rho_t \,\|\, \nu_t) = -\mathbb{E}_{\rho_t}\left[\left\|\nabla^2 \log \frac{\rho_t}{\nu_t}\right\|_{\mathrm{HS}}^2\right] - 2\mathbb{E}_{\rho_t}\left[\left\|\nabla \log \frac{\rho_t}{\nu_t}\right\|_{(-\nabla^2 \log \nu_t)}^2\right].$$

- c.f. for KL divergence: $\frac{d}{dt}\mathsf{KL}(\rho_t \,\|\, \nu_t) = -\frac{1}{2}\,\mathsf{FI}(\rho_t \,\|\, \nu_t)$
- (S)DPI follows by evolution of SLC constant along heat flow:

  If $-\nabla^2 \log \nu_0(x) \succeq \alpha I$, then $-\nabla^2 \log \nu_t(x) \succeq \frac{\alpha}{1+\alpha t} I$

Evolution of FI along General Fokker-Planck Channel

Lemma: If $(\rho_t)_{t\geq 0}$, $(\nu_t)_{t\geq 0}$ evolve following Fokker-Planck equations:

$$\partial_t \rho_t = -\nabla \cdot (\rho_t b_t) + \frac{c}{2}\Delta \rho_t \ ,$$
$$\partial_t \nu_t = -\nabla \cdot (\nu_t b_t) + \frac{c}{2}\Delta \nu_t$$

for any smooth vector field $b_t \colon \mathbb{R}^d \to \mathbb{R}^d$ and $c \geq 0$. Then:

$$\frac{d}{dt}\mathsf{FI}(\rho_t \,\|\, \nu_t) = -c\,\mathbb{E}_{\rho_t}\left[\left\|\nabla^2 \log \frac{\rho_t}{\nu_t}\right\|_{\mathrm{HS}}^2\right]$$

- c.f. for KL divergence: $\frac{d}{dt}\mathsf{KL}(\rho_t \,\|\, \nu_t) = -\frac{c}{2}\,\mathsf{FI}(\rho_t \,\|\, \nu_t)$
- Heat flow: $b_t = 0$, $c = 1$
- Ornstein-Uhlenbeck (Langevin for Gaussian): $b_t(x) = -\gamma x$, $c = 2$
- Reverse Gaussian channel: $b_t(x) = \nabla \log(\nu * \mathcal{N}(0, tI))$, $c = 1$

Application: Mixing Time of Proximal Sampler in FI

Theorem:[8] Assume ν^X is α-SLC. Then for each $k \geq 0$:

1. Forward step: From $x_k \sim \rho_k^X$ to $y_k \sim \rho_k^Y$,

$$\mathsf{FI}(\rho_k^Y \,\|\, \nu^Y) \leq \frac{\mathsf{FI}(\rho_k^X \,\|\, \nu^X)}{(1+\alpha\eta)^2}$$

2. Backward step: From $y_k \sim \rho_k^Y$ to $x_{k+1} \sim \rho_{k+1}^X$,

$$\mathsf{FI}(\rho_{k+1}^X \,\|\, \nu^X) \leq \mathsf{FI}(\rho_k^Y \,\|\, \nu^Y)$$

Therefore,

$$\mathsf{FI}(\rho_k^X \,\|\, \nu^X) \leq \frac{\mathsf{FI}(\rho_0^X \,\|\, \nu^X)}{(1+\alpha\eta)^{2k}}$$

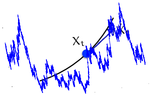
- Recall for KL/Rényi, have SDPI for both forward and backward steps.
- For FI, have SDPI in forward, and only weak DPI in backward step.

[8] [**W.**, "Mixing Time of the Proximal Sampler in Relative Fisher Information via Strong Data Processing Inequality", 2025]

Summary

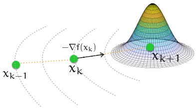
- Sampling as Optimization in the space of distributions:
  - Cts. time: Langevin dynamics $\Leftrightarrow$ Gradient flow
  - Disc. time: Proximal Sampler $\approx$ Proximal gradient method

Langevin Dynamics (LD)



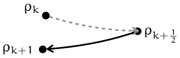
$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2}\, dW_t$$

Unadjusted Langevin Algorithm (ULA)



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

Proximal Sampler (PS)



$$y_k = x_k + \sqrt{\eta}\, z_k$$

$$x_{k+1} \sim \exp\left(-f(x) - \frac{1}{2\eta}\|x - y_k\|^2\right)$$

Summary

- Sampling as Optimization in the space of distributions:
  - Cts. time: Langevin dynamics $\Leftrightarrow$ Gradient flow
  - Disc. time: Proximal Sampler $\approx$ Proximal gradient method
- Proximal Sampler has unbiased convergence guarantees, matching Langevin dynamics and Proximal gradient:
  - In KL divergence/Rényi divergence under LSI
  - In Φ-divergence under SLC ($\Rightarrow$ Φ-Sobolev)
  - In relative FI under SLC
- **Technique:** SDPI along Fokker-Planck channels
  - SDPI in KL/Rényi always holds under LSI
  - DPI in FI does *not* always hold, even for Gaussian channel
  - (S)DPI in FI holds under (strong) log-concavity

Questions

- SDPI in FI for other channels, under LSI or weaker conditions?
- Mixing time in relative FI for other sampling algorithms?
- Acceleration in Sampling ($\Leftrightarrow$ matching rates with Opt)?
    - Want $\tilde{O}(\sqrt{\frac{L}{\alpha}})$ iteration complexity in discrete time

      (c.f. Proximal Sampler needs $\tilde{O}(\frac{dL}{\alpha})$ iterations)

Thank you!

[**W.**, "Mixing Time of Proximal Sampler in Relative Fisher Information via Strong Data Processing Inequality", COLT 2025]

Key: SDPI in KL along Fokker-Planck Channel under LSI

Lemma: Suppose $(\rho_t)_{t\geq 0}$ and $(\nu_t)_{t\geq 0}$ evolve following the PDE:

$$\partial_t \rho_t = -\nabla \cdot (\rho_t b_t) + \frac{c}{2}\Delta \rho_t$$
$$\partial_t \nu_t = -\nabla \cdot (\nu_t b_t) + \frac{c}{2}\Delta \nu_t$$

for any smooth vector field $b_t \colon \mathbb{R}^d \to \mathbb{R}^d$ and constant $c \geq 0$.
Then for any $t \geq 0$:

$$\frac{d}{dt}\mathsf{KL}(\rho_t \,\|\, \nu_t) = -\frac{c}{2}\,\mathsf{FI}(\rho_t \,\|\, \nu_t).$$

Therefore, if we know that ν_t satisfies α_t-LSI for all $t \geq 0$, then:

$$\mathsf{KL}(\rho_t \,\|\, \nu_t) \leq \exp\left(-c\int_0^t \alpha_s \, ds\right)\mathsf{KL}(\rho_0 \,\|\, \nu_0).$$

- Identity also holds for Rényi and all Φ-divergence.
- To apply, key is to control evolution of LSI constant along PDE.

Eventual SDPI in FI along Ornstein-Uhlenbeck Channel

Theorem: Along the **OU channel** (Langevin to $\mathcal{N}(0, \gamma^{-1} I)$):

$$X_t = e^{-\gamma t} X_0 + \sqrt{\frac{1 - e^{-2\gamma t}}{\gamma}} Z, \quad Z \sim \mathcal{N}(0, I)$$

If ν_0 is α-strongly log-concave, then we have (eventual) SDPI:

$$\mathsf{FI}(\rho_t \,\|\, \nu_t) \le \frac{\gamma^2 \, \mathsf{FI}(\rho_0 \,\|\, \nu_0)}{(\alpha + e^{-2\gamma t}(\gamma - \alpha))^2} \, e^{-2\gamma t}$$

- Improved rate if ρ_0 satisfies Poincaré and symmetry
- If $\gamma \to 0$, this recovers the Gaussian channel result.

Eventual SDPI in FI along Ornstein-Uhlenbeck Channel

Example: Along the **OU channel** (targeting $\mathcal{N}(0,1)$):

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}}\, Z, \quad Z \sim \mathcal{N}(0,1)$$

- Let $\rho_0 = \mathcal{N}(0, 0.01)$, and $\nu_0 = \mathcal{N}(0, 10)$