CS 59300 – Algorithms for Data Science

Classical and Quantum approaches

Lecture 12 (10/21)

Stochastic Calculus

https://ruizhezhang.com/course_fall_2025.html

Slides are based on Kevin Tian's lecture notes and Sitan Chen's slides

Motivation

- Stochastic calculus is the mathematical tool to study sampling and generative modeling in a continuous space (and continuous time)
  - An analog is the **traditional calculus** for Gradient flow $\longleftrightarrow$ Gradient descent

- It is also a prerequisite for appreciating Eldan's stochastic localization
  - In the mean-field approximation, we use only a single "snapshot" of the stochastic localization at a fixed time

Today's plan

1. Drift-diffusion processes
2. Markov semigroup
3. Optimal transport
4. Functional inequalities

Highly recommend reference: *Log-concave sampling* by Sinho Chewi '25

Today’s plan

1. Drift-diffusion processes

2. Markov semigroup

3. Optimal transport

4. Functional inequalities

Highly recommend reference: *Log-concave sampling* by Sinho Chewi ’25

Brownian motion

$(\boldsymbol{B}_t)_{t\geq 0} \subset \mathbb{R}^d$ is a stochastic process (random sequence of vectors indexed by t) such that:

- **Starts at origin:** $\boldsymbol{B}_0 = 0$ is the origin in $\mathbb{R}^d$
- **Continuous paths:** With prob. 1 over the randomness, $t \mapsto \boldsymbol{B}_t$ is continuous
- **Independent increments:** For all $0 = t_0 < t_1 < \cdots < t_k$, the random vectors $\boldsymbol{B}_{t_{i+1}} - \boldsymbol{B}_{t_i}$ for $0 \leq i < k$ are mutually independent
- **Gaussian increments:** For all $0 \leq s \leq t$, $\boldsymbol{B}_t - \boldsymbol{B}_s \sim N(0, (t-s)I)$

The probability space that Brownian motion is defined on is denoted $\{\mathcal{F}_t\}_{t\geq 0}$, a filtration satisfying $\mathcal{F}_s \subseteq \mathcal{F}_t$ for all $0 \leq s \leq t$. We say that Brownian motion is a stochastic process adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$

"$\mathcal{F}_t$ contains the information of the randomness used up to time t"

Stochastic integrals

Given continuous process $(\mathbf{H}_t)_{t\geq 0} \in \mathbb{R}^{d \times d}$ adapted to the filtration generated by $(\boldsymbol{B}_t)_{t\geq 0}$, the Itô integral

$$\mathbf{x}_t = \int_0^t \mathbf{H}_s \, \mathrm{d}\boldsymbol{B}_s$$

is the stochastic process whose value at time t is the random vector given by taking the probability limit over meshes $P = \{t_k\}_{k\in[|P|]} \subset [0, t]$ such that $0 = t_0 < \cdots < t_{|P|} = t$:

$$\int_0^t \mathbf{H}_s \, \mathrm{d}\mathbf{B}_s = \lim_{\|P\|_{\mathrm{gap}}\to 0} \sum_{j=1}^{k} \mathbf{H}_{t_j} \cdot \left(\boldsymbol{B}_{t_{j+1}} - \boldsymbol{B}_{t_j}\right)$$

where $\|P\|_{\mathrm{gap}} \coloneqq \max_{k\in[|P|]} |t_k - t_{k-1}|$

Stochastic integrals

The Itô integral $\mathbf{x}_t = \int_0^t \mathbf{H}_s \, \mathrm{d}\boldsymbol{B}_s$ is a continuous martingale

- Continuity: With prob. 1 over the randomness, $t \mapsto \mathbf{x}_t$ is continuous
- Martingale: $\mathbb{E}\left[\int_s^t \mathbf{H}_s \, \mathrm{d}\boldsymbol{B}_s \,\middle|\, \mathcal{F}_s\right] = 0$
- Itô isometry:

$$\mathbb{E}\left[\left\|\int_0^t \mathbf{H}_s \, \mathrm{d}\boldsymbol{B}_s\right\|^2\right] = \mathbb{E}\left[\int_0^t \|\mathbf{H}_s\|_F^2 \, \mathrm{d}B_s\right]$$

Intuition: $\mathbb{E}[(a_1 g_1 + a_2 g_2 + \cdots + a_k g_k)^2] = a_1^2 + \cdots + a_k^2$ for $g_1, \ldots, g_k \sim N(0,1)$

"$\mathrm{d}\boldsymbol{B}_t^2 = \mathrm{d}t$" (because $\boldsymbol{B}_{t+\mathrm{d}t} - \boldsymbol{B}_t \sim \mathcal{N}(0, \mathrm{d}t)$)

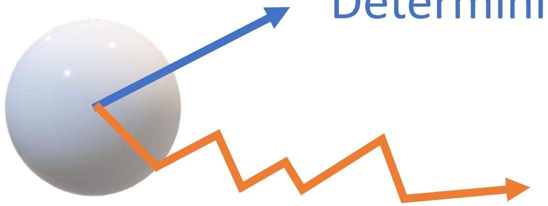
Drift-diffusion processes

A drift-diffusion process $\{\mathbf{x}_t\}_{t\geq 0}$ on $\mathbb{R}^d$ is driven by a vector-valued function $\boldsymbol{\mu}: \mathbb{R}^d \to \mathbb{R}^d$ and a matrix-valued function $\boldsymbol{\sigma}: \mathbb{R}^d \to \mathbb{R}^{d\times d}$, and captured by the stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_t)\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t$$

We can write this stochastic process in integral form:

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t \boldsymbol{\mu}(\mathbf{x}_t)\mathrm{d}t + \int_0^t \boldsymbol{\sigma}(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t \qquad \text{(Itô process)}$$



Drift-diffusion processes

A drift-diffusion process $\{\mathbf{x}_t\}_{t\geq 0}$ on $\mathbb{R}^d$ is driven by a vector-valued function $\boldsymbol{\mu}: \mathbb{R}^d \to \mathbb{R}^d$ and a matrix-valued function $\boldsymbol{\sigma}: \mathbb{R}^d \to \mathbb{R}^{d\times d}$, and captured by the stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_t)\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t$$

We can write this stochastic process in integral form:

$$\mathbf{x}_t = \mathbf{x}_0 + \int_0^t \boldsymbol{\mu}(\mathbf{x}_t)\mathrm{d}t + \int_0^t \boldsymbol{\sigma}(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t \qquad \text{(Itô process)}$$

- Technically need to check such a process exists and is uniquely defined, which holds under mild conditions on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ (e.g. Lipschitzness)
- SDEs can be defined w.r.t. more general process, e.g. $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$
- Euler–Maruyama discretization: $\hat{x}_{(k+1)h} \leftarrow \hat{x}_{kh} + h\boldsymbol{\mu}(\hat{x}_{kh}) + \sqrt{h}\,\boldsymbol{\sigma}(\hat{x}_{kh})g \quad g \sim \mathcal{N}(0, I)$

Important example: Langevin diffusion

Given differentiable $V\colon \mathbb{R}^d \to \mathbb{R}$, consider

$$\mathrm{d}\mathbf{x}_t = -\nabla V(\mathbf{x}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t$$

"gradient descent + noise"

i.e. $\boldsymbol{\mu} = -\nabla$ and $\boldsymbol{\sigma} = \sqrt{2}I$

- $\pi^\star \propto e^{-V}$ is a stationary distribution of Langevin diffusion (and is unique under certain assumptions on V)
- By Euler-Maruyama discretization, we get the Unadjusted Langevin Algorithm (ULA):

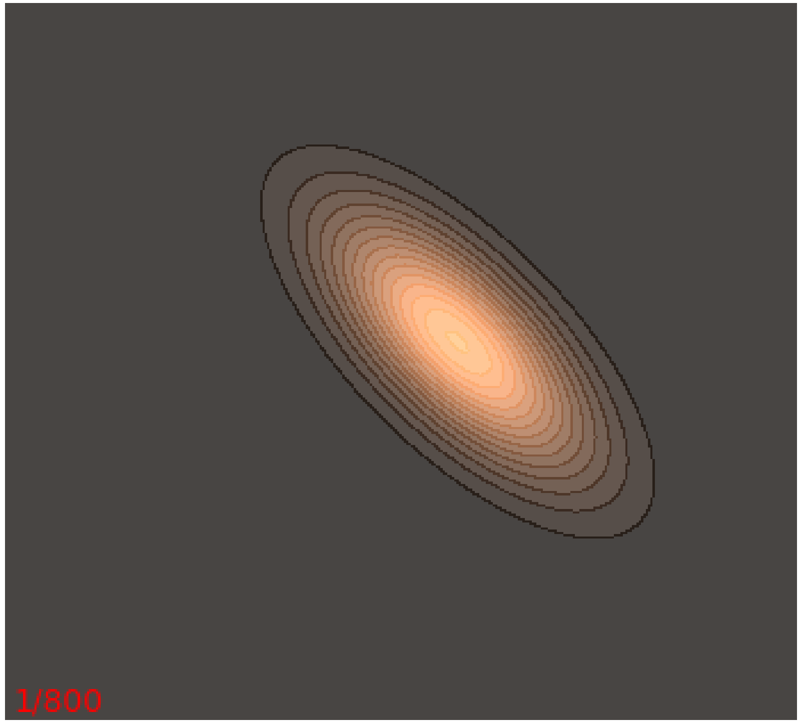
$$\hat{x}_{(k+1)h} \leftarrow \hat{x}_{kh} - h\nabla V(x_t) + \sqrt{2h}\,g \quad \text{for} \quad g \sim \mathcal{N}(0, I)$$

Ornstein-Uhlenbeck process

If we take $V(x) = \|x\|^2/2$, we have

$$\mathrm{d}\mathbf{x}_t = -\mathbf{x}_t \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\boldsymbol{B}_t$$

whose stationary distribution is given by $N(0, \mathrm{Id})$.



https://friedmanroy.github.io/blog/2022/Langevin/

Stochastic chain rule

$$\mathrm{d}f(g) = f'(g) \cdot \mathrm{d}g$$

Itô's lemma. Let $f: \mathbb{R}^d \to \mathbb{R}$ be twice-differentiable, and suppose $\{\mathbf{x}_t\}_{t\geq 0}$ follows the SDE:

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}_t(\mathbf{x}_t)\mathrm{d}t + \boldsymbol{\sigma}_t(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t$$

Then $\{f(\mathbf{x}_t)\}_{t\geq 0}$ is a stochastic process following the SDE:

$$\mathrm{d}f(\mathbf{x}_t) = \left(\langle \nabla f(\mathbf{x}_t), \boldsymbol{\mu}_t(\mathbf{x}_t)\rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}_t), \boldsymbol{\sigma}_t(\mathbf{x}_t)\boldsymbol{\sigma}_t(\mathbf{x}_t)^\top\rangle\right)\mathrm{d}t + \langle \nabla f(\mathbf{x}_t), \boldsymbol{\sigma}_t(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t\rangle$$

Proof sketch:

- $\mathbf{x}_{t+h} \approx \mathbf{x}_t + h\boldsymbol{\mu}(\mathbf{x}_t) + \sqrt{h}\,\boldsymbol{\sigma}(\mathbf{x}_t)g$ for $g \sim \mathcal{N}(0, I)$
- Taylor expand $f(\mathbf{x}_{t+h})$ and only keep first-order (i.e. $\mathcal{O}(h)$) terms:

$$df(\mathbf{x}_t) \approx f(\mathbf{x}_{t+h}) - f(\mathbf{x}_t) \approx \left\langle \nabla f(\mathbf{x}_t), h\boldsymbol{\mu}(\mathbf{x}_t) + \sqrt{h}\boldsymbol{\sigma}(\mathbf{x}_t)g\right\rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}_t), h\boldsymbol{\sigma}(\mathbf{x}_t)\underbrace{gg^\top}_{\approx I}\boldsymbol{\sigma}(\mathbf{x}_t)^\top\rangle$$

$$\approx h(\langle \nabla f(\mathbf{x}_t), \boldsymbol{\mu}(\mathbf{x}_t)\rangle + \langle \nabla^2 f(\mathbf{x}_t), \boldsymbol{\sigma}(\mathbf{x}_t)\boldsymbol{\sigma}(\mathbf{x}_t)^\top\rangle) + \frac{\sqrt{h}}{2}\langle \nabla f(\mathbf{x}_t), \boldsymbol{\sigma}(\mathbf{x}_t)g\rangle$$

Stochastic chain rule

$$\mathrm{d}f(g) = f'(g) \cdot \mathrm{d}g$$

Itô's lemma. Let $f: \mathbb{R}^d \to \mathbb{R}$ be twice-differentiable, and suppose $\{\mathbf{x}_t\}_{t\geq 0}$ follows the SDE:

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}_t(\mathbf{x}_t)\mathrm{d}t + \boldsymbol{\sigma}_t(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t$$

Then $\{f(\mathbf{x}_t)\}_{t\geq 0}$ is a stochastic process following the SDE:

$$\mathrm{d}f(\mathbf{x}_t) = \left(\langle \nabla f(\mathbf{x}_t), \boldsymbol{\mu}_t(\mathbf{x}_t)\rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}_t), \boldsymbol{\sigma}_t(\mathbf{x}_t)\boldsymbol{\sigma}_t(\mathbf{x}_t)^\top\rangle\right)\mathrm{d}t + \langle \nabla f(\mathbf{x}_t), \boldsymbol{\sigma}_t(\mathbf{x}_t)\mathrm{d}\boldsymbol{B}_t\rangle$$

Example:

- Consider the Langevin diffusion: $\mathrm{d}x_t = -\nabla V(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t$
- By Itô's lemma,

$$\mathrm{d}f(\mathbf{x}_t) = (-\langle \nabla f, \nabla V\rangle + \Delta f)\mathrm{d}t + \sqrt{2}\langle \nabla f, \mathrm{d}\boldsymbol{B}_t\rangle$$

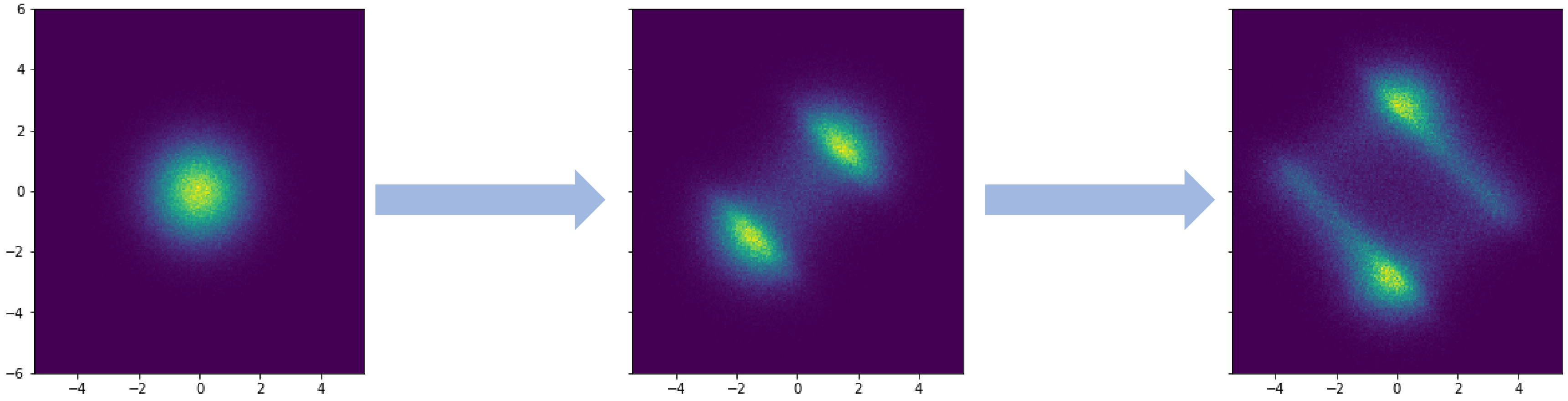
$$\Delta f \coloneqq \mathrm{tr}[\nabla^2 f]$$

Today's plan

1. Drift-diffusion processes
2. Markov semigroup
3. Optimal transport
4. Functional inequalities

Highly recommend reference: *Log-concave sampling* by Sinho Chewi '25

Dual view on SDEs



- Move to the density space $\mathcal{P}(\mathbb{R}^d)$ (i.e. the set of continuous prob. densities on $\mathbb{R}^d$)
- In discrete setting, the evolution of prob. density is characterized by the Markov chain transition matrix $P \in \mathbb{R}^{D \times D}$: $\pi_{t+1} = P\pi_t$
- For a drift-diffusion process $\{\mathbf{x}_t\}_{t\geq 0}$, we define the Markov semigroup $\{P_t\}_{t\geq 0}$:

$$(P_t f)(\mathbf{x}) \coloneqq \mathbb{E}[f(\mathbf{x}_t) \mid \mathbf{x}_0 = \mathbf{x}] \qquad \text{for} \quad f: \mathbb{R}^d \to \mathbb{R}$$

Markov semigroup

- For a drift-diffusion process $\{\mathbf{x}_t\}_{t\geq 0}$, we define the Markov semigroup $\{P_t\}_{t\geq 0}$:

$$(P_t f)(\mathbf{x}) \coloneqq \mathbb{E}[f(\mathbf{x}_t) \mid \mathbf{x}_0 = \mathbf{x}] \qquad \text{for} \quad f: \mathbb{R}^d \to \mathbb{R}$$

If $f = \mathbf{1}_S$ for a subset S, then $(P_t f)(\mathbf{x}) = \Pr[\mathbf{x}_t \in S \mid \mathbf{x}_0 = \mathbf{x}]$

- Markov property:

$$P_{t+s} f = P_t P_s f = P_s P_t f \qquad \forall\, f: \mathbb{R}^d \to \mathbb{R}, \forall\, s, t \geq 0$$

- Generator:

$$\mathcal{L} f \coloneqq \lim_{\eta \to 0} \frac{P_\eta f - f}{\eta}$$

Kolmogorov's backward equation. For all $t \geq 0$ and $f: \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\frac{\partial}{\partial t} P_t f = \mathcal{L} P_t f = P_f \mathcal{L} f$$

Kolmogorov's equations

Kolmogorov's backward equation. For all $t \geq 0$ and $f: \mathbb{R}^d \to \mathbb{R}$, it holds that

$$\frac{\partial}{\partial t} P_t f = \mathcal{L} P_t f = P_f \mathcal{L} f$$

Proof.

$$\mathcal{L} P_t f = \lim_{\eta \to 0} \frac{P_\eta - \overbrace{P_0}^{\equiv \mathrm{Id}}}{\eta} P_t f = \lim_{\eta \to 0} \frac{P_{t+\eta} - P_t}{\eta} f = \lim_{\eta \to 0} P_t \frac{P_\eta - P_0}{\eta} f = P_t \mathcal{L} f$$

∎

Kolmogorov's forward equation. Let P_t^* be the adjoint of P_t:

$$\mathbb{E}[f(\mathbf{x}_t)] = \int \mathbb{E}[f(\mathbf{x}_t) \mid \mathbf{x}_0 = \mathbf{x}] \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x} = \int P_t f(\mathbf{x}) \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x} = \int f(\mathbf{x}) P_t^* \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Then,

$$\langle P_t f, \pi_0 \rangle = \langle f, P_t^* \pi_0 \rangle$$

$$\frac{\partial}{\partial t} \underbrace{P_t^* \pi_0}_{\pi_t} = \mathcal{L}^* P_t^* \pi_0 = P_t^* \mathcal{L}^* \pi_0$$

Kolmogorov's forward equation for drift-diffusion process

$$\frac{\partial}{\partial t}\pi_t = \mathcal{L}^*\pi_t$$

Fokker-Planck equation. Let $\{x_t\}_{t\geq 0}$ follows $\mathrm{d}\mathbf{x}_t = \boldsymbol{\mu}_t(\boldsymbol{x}_t)\mathrm{d}t + \boldsymbol{\sigma}_t \mathrm{d}\boldsymbol{B}_t$ and $\mathbf{x}_0 \sim \pi_0$. Then for all $t \geq 0$, denoting the law of x_t by π_t, we have

$$\frac{\partial}{\partial t}\pi_t(\mathbf{x}) = -\nabla \cdot \big(\boldsymbol{\mu}_t(\mathbf{x})\pi_t(\mathbf{x})\big) + \frac{1}{2}\sum_{i,j\in[d]} \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\big(\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^\top(\mathbf{x})_{ij}\pi_t(\mathbf{x})\big)$$

Proof of Fokker-Planck equation

- Recall that $\mathrm{d}f(\mathbf{x}) = \left(\langle \nabla f(\mathbf{x}), \boldsymbol{\mu}_t(\mathbf{x})\rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}), \boldsymbol{\sigma}_t(\mathbf{x})\boldsymbol{\sigma}_t(\mathbf{x})^\top\rangle\right)\mathrm{d}t + \underbrace{\langle \nabla f(\mathbf{x}), \boldsymbol{\sigma}_t(\mathbf{x})\mathrm{d}\boldsymbol{B}_t\rangle}_{\text{martingale}}$
- Then $\mathrm{d}\underbrace{\mathbb{E}[f(\mathbf{x})]}_{P_t f(\mathbf{x})} = \left(\underbrace{\langle \nabla f(\mathbf{x}), \boldsymbol{\mu}_t(\mathbf{x})\rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}), \boldsymbol{\sigma}_t(\mathbf{x})\boldsymbol{\sigma}_t(\mathbf{x})^\top\rangle}_{\mathcal{L}f(\mathbf{x})}\right)\mathrm{d}t$
- Thus, for all prob. densities π,

$$\int f(\mathbf{x})\mathcal{L}^*\pi(\mathbf{x})\mathrm{d}\mathbf{x} = \int \mathcal{L}f(\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x} = \int \left(\langle \nabla f(\mathbf{x}), \boldsymbol{\mu}_t(\mathbf{x})\rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}), \boldsymbol{\sigma}_t(\mathbf{x})\boldsymbol{\sigma}_t(\mathbf{x})^\top\rangle\right)\pi(\mathbf{x})\mathrm{d}\mathbf{x}$$

(Integral by parts) $$= \int \left(-f(\mathbf{x})\nabla\cdot\big(\boldsymbol{\mu}_t(\mathbf{x})\pi(\mathbf{x})\big) - \frac{1}{2}\left\langle \nabla f(\mathbf{x}), \nabla\cdot\big(\boldsymbol{\sigma}_t(\mathbf{x})\boldsymbol{\sigma}_t(\mathbf{x})^\top\pi(\mathbf{x})\big)\right\rangle\right)\mathrm{d}\mathbf{x}$$

(Integral by parts) $$= \int f(\mathbf{x})\left(-\nabla\cdot\big(\boldsymbol{\mu}_t(\mathbf{x})\pi(\mathbf{x})\big) + \frac{1}{2}\sum_{i,j\in[d]}\frac{\partial^2}{\partial\mathbf{x}_i\partial\mathbf{x}_j}\big(\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^\top(\mathbf{x})_{ij}\pi(\mathbf{x})\big)\right)\mathrm{d}\mathbf{x}$$

Proof of Fokker-Planck equation

- Recall that $\mathrm{d}f(\mathbf{x}) = \left(\langle \nabla f(\mathbf{x}), \boldsymbol{\mu}_t(\mathbf{x})\rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}), \boldsymbol{\sigma}_t(\mathbf{x})\boldsymbol{\sigma}_t(\mathbf{x})^\top\rangle\right)\mathrm{d}t + \underbrace{\langle \nabla f(\mathbf{x}), \boldsymbol{\sigma}_t(\mathbf{x})\mathrm{d}\boldsymbol{B}_t\rangle}_{\text{martingale}}$
- Then $\mathrm{d}\underbrace{\mathbb{E}[f(\mathbf{x})]}_{P_t f(\mathbf{x})} = \left(\underbrace{\langle \nabla f(\mathbf{x}), \boldsymbol{\mu}_t(\mathbf{x})\rangle + \frac{1}{2}\langle \nabla^2 f(\mathbf{x}), \boldsymbol{\sigma}_t(\mathbf{x})\boldsymbol{\sigma}_t(\mathbf{x})^\top\rangle}_{\mathcal{L}f(\mathbf{x})}\right)\mathrm{d}t$
- Thus, for all prob. densities π,

$$\int f(\mathbf{x})\mathcal{L}^*\pi(\mathbf{x})\mathrm{d}\mathbf{x} = \int f(\mathbf{x})\underbrace{\left(-\nabla\cdot\left(\boldsymbol{\mu}_t(\mathbf{x})\pi(\mathbf{x})\right) + \frac{1}{2}\sum_{i,j\in[d]}\frac{\partial^2}{\partial \mathbf{x}_i\partial \mathbf{x}_j}\left(\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^\top(\mathbf{x})_{ij}\pi(\mathbf{x})\right)\right)}_{\mathcal{L}^*\pi(\mathbf{x})}\mathrm{d}\mathbf{x}$$

- Then it follows from the Kolmogorov's forward equation $\frac{\partial}{\partial t}\pi_t = \mathcal{L}^*\pi_t$

$\blacksquare$

Fokker-Planck equation for the Langevin diffusion

$$\frac{\partial}{\partial t}\pi_t(\mathbf{x}) = -\nabla \cdot \big(\boldsymbol{\mu}_t(\mathbf{x})\pi_t(\mathbf{x})\big) + \frac{1}{2}\sum_{i,j\in[d]} \frac{\partial^2}{\partial \mathbf{x}_i \partial \mathbf{x}_j}\Big(\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^\top(\mathbf{x})_{ij}\pi(\mathbf{x})\Big)$$

- When $\boldsymbol{\mu} = -\nabla V$ and $\boldsymbol{\sigma} = \sqrt{2}I$,

$$\frac{\partial}{\partial t}\pi_t(\mathbf{x}) = \underbrace{\nabla \cdot \big(\nabla V(\mathbf{x})\pi_t(\mathbf{x})\big)}_{\text{drift term}} + \underbrace{\Delta\pi_t}_{\text{heat equation term}}$$

Theorem. The stationary distribution for the Langevin dynamics is $\pi^\star \propto \exp(-V)$

Proof.

- Let $U \coloneqq -\log \pi^\star$
- Stationary $\Leftrightarrow \frac{\partial}{\partial t}\pi_t(\mathbf{x}) = 0$
- $0 = \nabla \cdot \big(\nabla V(\mathbf{x})\pi^\star(\mathbf{x})\big) + \Delta\pi^\star = \nabla \cdot \big(\nabla V(\mathbf{x})\pi^\star(\mathbf{x}) + \nabla\pi^\star(\mathbf{x})\big) = \nabla \cdot \Big(\underbrace{\big(\nabla V(\mathbf{x}) - \nabla U(\mathbf{x})\big)}_{\text{Solved by } U = V + c}\pi^\star(\mathbf{x})\Big)$

Solved by $U = V + c$

$\blacksquare$

Spectral gap and the Dirichlet form

- **Motivation:** we want to understand the spectrum of $\mathcal{L}^*$

$$\frac{\partial}{\partial t}\pi_t = \mathcal{L}^*\pi_t \quad \Longrightarrow \quad \pi_t = \exp(t\mathcal{L}^*)\pi_0$$

If $\pi^\star$ is an eigenfunction of $\mathcal{L}^*$ with eigenvalue 0 and all other eigenvalues are negative

Then $\pi_t \longrightarrow \pi^\star$ as $t \longrightarrow \infty$

- To quantitively analyze the convergence rate, we define the Dirichlet form:

$$\mathcal{E}(f,g) \coloneqq -\int f(\mathbf{x})\mathcal{L}g(\mathbf{x})\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} \qquad \forall\, f,g\colon \mathbb{R}^d \to \mathbb{R}$$

$$-\langle f, \mathcal{L}g\rangle_{\pi^\star}$$

- For the Langevin diffusion,

$$\mathcal{E}(f,g) \coloneqq \int \langle \nabla f(\mathbf{x}), \nabla g(\mathbf{x})\rangle \pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}$$

Courant-Fischer:

$$\frac{v^\top \boldsymbol{A} v}{v^\top v}$$

Dirichlet form

$$\mathcal{E}(f,g) \coloneqq \int \langle \nabla f(\mathbf{x}), \nabla g(\mathbf{x}) \rangle \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}$$

- Langevin is reversible since $\mathcal{E}(f,g) = \mathcal{E}(g,f)$ $\langle \mathcal{L}f, g \rangle_{\pi^\star} = \langle f, \mathcal{L}g \rangle_{\pi^\star}$
- Constant function is an eigenfunction of $\mathcal{E}$ with eigenvalue 0
- The spectral gap is captured by all eigenfunctions of $\mathcal{E}$ orthogonal to the constant function

$$\mathrm{gap} \coloneqq \min_{f \perp 1} \frac{\mathcal{E}(f,f)}{\langle f, f \rangle_{\pi^\star}}$$

$$\mathcal{E}(f,g) = -\int \mathcal{L}f(\mathbf{x}) g(\mathbf{x}) \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \langle \nabla f, \nabla V \rangle g \pi^\star \mathrm{d}\mathbf{x} - \int \Delta f g \pi^\star \mathrm{d}\mathbf{x}$$

(I.B.P.)

$$= \int \langle \nabla f, \nabla V \rangle g \pi^\star \mathrm{d}\mathbf{x} + \left(\int \langle \nabla f, \nabla g \rangle \pi^\star \mathrm{d}\mathbf{x} + \int \langle \nabla f, \nabla \pi^\star \rangle g \mathrm{d}\mathbf{x} \right)$$

$\pi^\star \propto \exp(-V)$

$$= \int \langle \nabla f, \nabla V \rangle g \pi^\star \mathrm{d}\mathbf{x} + \int \langle \nabla f, \nabla g \rangle \pi^\star \mathrm{d}\mathbf{x} - \int \langle \nabla f, \nabla V \rangle g \pi^\star \mathrm{d}\mathbf{x} = \int \langle \nabla f, \nabla g \rangle \pi^\star \mathrm{d}\mathbf{x}$$

Today’s plan

1. Drift-diffusion processes
2. Markov semigroup
3. Optimal transport
4. Functional inequalities

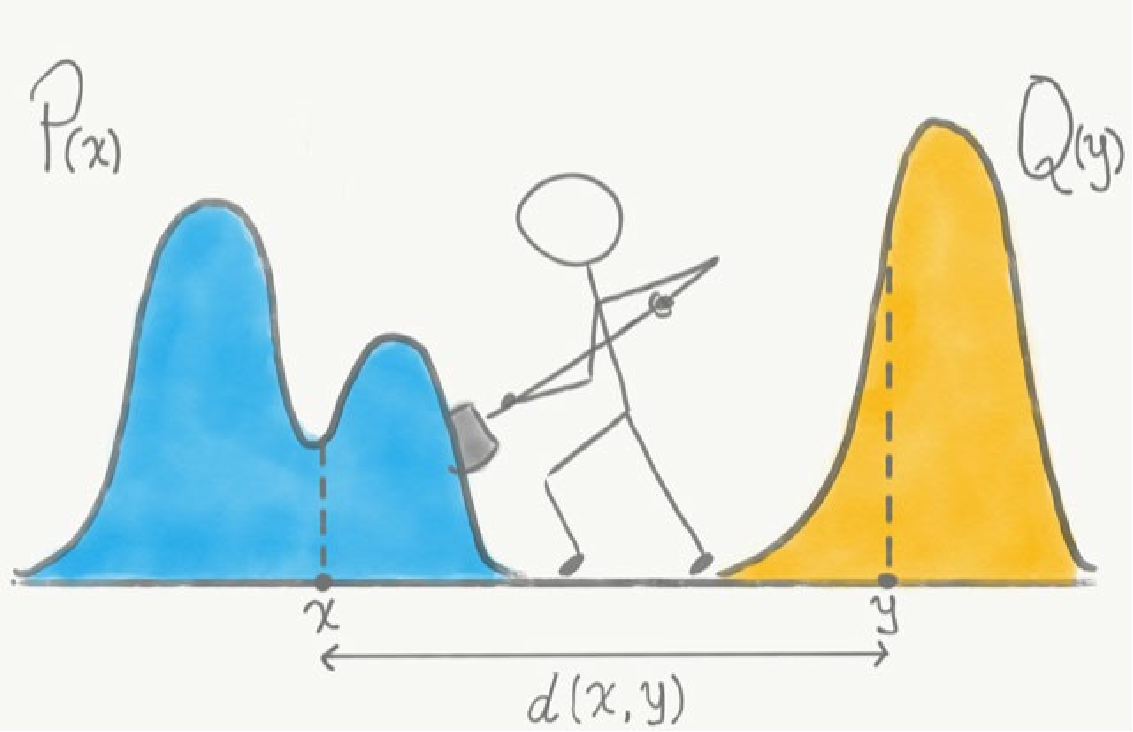
Highly recommend reference: *Log-concave sampling* by Sinho Chewi ’25

Optimal transport

Wasserstein distance:

$$W_2(\mu,\pi) := \inf_{\gamma\in\mathcal{C}(\mu,\pi)} \sqrt{\int \|x-y\|^2 \gamma(x,y)\mathrm{d}x\mathrm{d}y}$$

γ is a coupling of μ and ν



Optimal transport

Wasserstein distance:

$$W_2(\mu, \pi) := \inf_{\gamma \in \mathcal{C}(\mu,\pi)} \sqrt{\int \|x-y\|^2 \gamma(x,y) \mathrm{d}x \mathrm{d}y}$$

Kantorovich: By the strong duality of LP,

$$W_2(\mu, \pi) := \sup_{(f,g) \in \mathcal{D}(\mu,\pi)} \int f(x)\mu(x)\mathrm{d}x + \int g(y)\nu(y)\mathrm{d}y$$

$$\mathcal{D}(\mu, \nu) := \left\{ (f,g) \in L^1 \,\middle|\, f(x) + g(y) \le \frac{1}{2}\|x-y\|^2 \;\; \forall\, x, y \in \mathbb{R}^d \right\}$$

Brenier: the optimal coupling is $x \sim \mu$ and $y \leftarrow \nabla\phi(x)$ for some convex potential ϕ

Wasserstein convergence of Langevin dynamics

- $\pi \propto \exp(-V)$ is μ-strongly logconcave if V is μ-strongly convex (i.e., $\nabla^2 V \succcurlyeq \mu I$)

Theorem. Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the Langevin diffusion with stationary distribution $\pi^\star$ being μ-strongly logconcave. Then for all $t \geq 0$,

$$W_2^2(\pi_t, \pi) \leq \exp(-2\mu t)\, W_2^2(\pi_0, \pi^\star)$$

Proof.

- We use the "coupling method":
  - $\mathrm{d}\mathbf{x}_t = -\nabla V(\mathbf{x}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t$

    same $\qquad (\mathbf{x}_0, \mathbf{x}_0^\star) \sim \gamma_0$ the optimal coupling for $W_2(\pi_0, \pi^\star)$
  - $\mathrm{d}\mathbf{x}_t^\star = -\nabla V(\mathbf{x}_t^\star)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t$
- We get a coupling γ_t for $(\pi_t, \pi^\star)$ for every $t \geq 0$

Wasserstein convergence of Langevin dynamics

- $\pi \propto \exp(-V)$ is μ-strongly logconcave if V is μ-strongly convex (i.e., $\nabla^2 V \succcurlyeq \mu I$)

Theorem. Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the Langevin diffusion with stationary distribution $\pi^\star$ being μ-strongly logconcave. Then for all $t \geq 0$,

$$W_2^2(\pi_t, \pi) \leq \exp(-2\mu t)\, W_2^2(\pi_0, \pi^\star)$$

Proof.

- We use the "coupling method":
  - $\mathrm{d}\mathbf{x}_t = -\nabla V(\mathbf{x}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t$

    same

  - $\mathrm{d}\mathbf{x}_t^\star = -\nabla V(\mathbf{x}_t^\star)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t$
- We get a coupling γ_t for $(\pi_t, \pi^\star)$ for every $t \geq 0$

$$\frac{\mathrm{d}}{\mathrm{d}s}(\mathbf{x}_s - \mathbf{x}_s^\star) = \nabla V(\mathbf{x}_s^\star) - \nabla V(\mathbf{x}_s)$$

$$\Longrightarrow \frac{\mathrm{d}}{\mathrm{d}s}\|\mathbf{x}_s - \mathbf{x}_s^\star\|^2 = -2\langle \nabla V(\mathbf{x}_s) - \nabla V(\mathbf{x}_s^\star), \mathbf{x}_s - \mathbf{x}_s^\star\rangle$$
$$\leq -2\mu\|\mathbf{x}_s - \mathbf{x}_s^\star\|^2$$

by the strong convexity of V

Wasserstein convergence of Langevin dynamics

- $\pi \propto \exp(-V)$ is μ-strongly logconcave if V is μ-strongly convex (i.e., $\nabla^2 V \succcurlyeq \mu I$)

Theorem. Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the Langevin diffusion with stationary distribution $\pi^\star$ being μ-strongly logconcave. Then for all $t \geq 0$,

$$W_2^2(\pi_t, \pi) \leq \exp(-2\mu t)\, W_2^2(\pi_0, \pi^\star)$$

Proof.

- We use the "coupling method":
  - $\mathrm{d}\mathbf{x}_t = -\nabla V(\mathbf{x}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t$

    same

  - $\mathrm{d}\mathbf{x}_t^\star = -\nabla V(\mathbf{x}_t^\star)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t$
- We get a coupling γ_t for $(\pi_t, \pi^\star)$ for every $t \geq 0$

$$\frac{\mathrm{d}}{\mathrm{d}s}(\mathbf{x}_s - \mathbf{x}_s^\star) = \nabla V(\mathbf{x}_s^\star) - \nabla V(\mathbf{x}_s)$$

$$\Longrightarrow \frac{\mathrm{d}}{\mathrm{d}s}\|\mathbf{x}_s - \mathbf{x}_s^\star\|^2 = -2\langle \nabla V(\mathbf{x}_s) - \nabla V(\mathbf{x}_s^\star), \mathbf{x}_s - \mathbf{x}_s^\star\rangle$$

$$\leq -2\mu\|\mathbf{x}_s - \mathbf{x}_s^\star\|^2$$

$$\Longrightarrow \|\mathbf{x}_t - \mathbf{x}_t^\star\|^2 \leq \exp(-2\mu t)\,\|\mathbf{x}_0 - \mathbf{x}_0^\star\|^2$$

by Gronwall's inequality

Wasserstein convergence of Langevin dynamics

- $\pi \propto \exp(-V)$ is μ-strongly logconcave if V is μ-strongly convex (i.e., $\nabla^2 V \succcurlyeq \mu I$)

Theorem. Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the Langevin diffusion with stationary distribution $\pi^\star$ being μ-strongly logconcave. Then for all $t \geq 0$,

$$W_2^2(\pi_t, \pi) \leq \exp(-2\mu t)\, W_2^2(\pi_0, \pi^\star)$$

Proof.

- Since W_2 minimizes over all couplings, we have

$$\begin{aligned} W_2^2(\pi_t, \pi^\star) \leq \mathbb{E}_{(\mathbf{x}_t, \mathbf{x}_t^\star)\sim\gamma_t}[\|\mathbf{x}_t - \mathbf{x}_t^\star\|^2] &\leq \exp(-2\mu t)\, \mathbb{E}_{(\mathbf{x}_0, \mathbf{x}_0^\star)\sim\gamma_0}[\|\mathbf{x}_0 - \mathbf{x}_0^\star\|^2] \\ &= \exp(-2\mu t)\, W_2(\pi_0, \pi^\star)^2 \end{aligned}$$

■

Today’s plan

1. Drift-diffusion processes
2. Markov semigroup
3. Optimal transport
4. Functional inequalities

Highly recommend reference: *Log-concave sampling* by Sinho Chewi ’25

Poincaré inequality

We say that $\pi^\star$ satisfies a Poincaré inequality with constant C_{PI} if for all differentiable f,

$$\underset{\langle f, f \rangle_{\pi^\star}}{\mathrm{Var}_{\pi^\star}[f]} \leq C_{\mathrm{PI}} \underset{\mathcal{E}(f, f)}{\int \|\nabla f(\mathbf{x})\|^2 \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}}$$

- Poincaré inequality is equivalent to the statement that

$$\mathrm{gap} \coloneqq \min_{f \perp 1} \frac{\mathcal{E}(f, f)}{\langle f, f \rangle_{\pi^\star}} \geq \frac{1}{C_{\mathrm{PI}}}$$

- The χ^2 divergence between μ and ν is defined as:

$$\chi^2(\mu \| \nu) \coloneqq \int \left(\frac{\mu(x)}{\nu(x)} \right)^2 \nu(x) \mathrm{d}x - 1$$

- PI $\Longrightarrow$ convergence in χ^2 divergence

Convergence from Poincaré inequality

Theorem. Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the Langevin diffusion with stationary distribution $\pi^\star$ satisfying a Poincaré inequality with constant C_{PI}. Then for all $t \geq 0$,

$$\chi^2(\pi_t \| \pi^\star) \leq \exp\left(-\frac{2t}{C_{\mathrm{PI}}}\right)\chi^2(\pi_0 \| \pi^\star)$$

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\pi_t \| \pi^\star) = \frac{\mathrm{d}}{\mathrm{d}t}\int\left(\frac{\pi_t(\mathbf{x})^2}{\pi^\star(\mathbf{x})^2} - 1\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} = 2\int\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\left(\frac{\partial}{\partial t}\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x}$$

$$= 2\int\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\left(\frac{\mathcal{L}^*\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\pi^\star(\mathbf{x})\mathrm{d}\mathbf{x} = -2\mathcal{E}\left(\frac{\pi_t}{\pi^\star}, \frac{\pi_t}{\pi^\star}\right)$$

$$\leq -\frac{2}{C_{\mathrm{PI}}}\mathrm{Var}_{\pi^\star}\left[\frac{\pi_t}{\pi^\star}\right] = -\frac{2}{C_{\mathrm{PI}}}\chi^2(\pi_t \| \pi^\star)$$

■

Log-Sobolev inequality

We say that $\pi^\star$ satisfies a log-Sobolev inequality with constant C_{LSI} if for all differentiable f,

$$\mathrm{Ent}_{\pi^\star}[f^2] \leq 2C_{\mathrm{LSI}} \int \|\nabla f(\mathbf{x})\|^2 \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}$$

where $\mathrm{Ent}_{\pi^\star}[f] \coloneqq \mathbb{E}_{\pi^\star}[f \log f] - \mathbb{E}_{\pi^\star}[f] \log \mathbb{E}_{\pi^\star}[f]$.

Equivalently,

$$D_{\mathrm{KL}}(\pi \| \pi^\star) \leq \frac{C_{\mathrm{LSI}}}{2} \int \left\| \nabla \log \frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})} \right\|^2 \pi(\mathbf{x}) \mathrm{d}\mathbf{x}$$

where $D_{\mathrm{KL}}(\pi \| \pi^\star) \coloneqq \int \pi(\mathbf{x}) \log\left(\frac{\pi(\mathbf{x})}{\pi^\star(\mathbf{x})}\right) \mathrm{d}\mathbf{x}$

Fisher information

- LSI $\Longrightarrow$ convergence in KL divergence

Convergence from log-Sobolev inequality

Theorem. Let $\{\mathbf{x}_t\}_{t\geq 0}$ follow the Langevin diffusion with stationary distribution $\pi^\star$ satisfying a log-Sobolev inequality with constant C_{LSI}. Then for all $t \geq 0$,

$$D_{\mathrm{KL}}(\pi_t \| \pi^\star) \leq \exp\left(-\frac{2t}{C_{\mathrm{LSI}}}\right) D_{\mathrm{KL}}(\pi_0 \| \pi^\star)$$

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}t} D_{\mathrm{KL}}(\pi_t \| \pi^\star) = \frac{\mathrm{d}}{\mathrm{d}t} \int \pi_t(\mathbf{x}) \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right) \mathrm{d}\mathbf{x} = \int \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\left(\frac{\mathcal{L}^* \pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right) \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$= -\mathcal{E}\left(\frac{\pi_t}{\pi^\star}, \log\left(\frac{\pi_t}{\pi^\star}\right)\right) = -\int \left\langle \nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right), \nabla\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right\rangle \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x}$$

$$= -\int \left\| \nabla \log\left(\frac{\pi_t(\mathbf{x})}{\pi^\star(\mathbf{x})}\right)\right\|^2 \pi^\star(\mathbf{x}) \mathrm{d}\mathbf{x} \leq -\frac{2}{C_{\mathrm{LSI}}} D_{\mathrm{KL}}(\pi_t \| \pi^\star)$$

■

LSI vs PI convergence

$$\chi^2(\pi_t \| \pi^\star) \le \exp\left(-\frac{2t}{C_{\mathrm{PI}}}\right) \chi^2(\pi_0 \| \pi^\star)$$

$$D_{\mathrm{KL}}(\pi_t \| \pi^\star) \le \exp\left(-\frac{2t}{C_{\mathrm{LSI}}}\right) D_{\mathrm{KL}}(\pi_0 \| \pi^\star)$$

LSI provides stronger convergence guarantee than PI since KL divergence can be exponentially smaller than χ^2 divergence

- π_0 is a β-warm start w.r.t. $\pi^\star$ if $\frac{\pi_0(\mathbf{x})}{\pi^\star(\mathbf{x})} \le \beta$ for all $\mathbf{x} \in \mathbb{R}^d$
- $\chi^2(\pi_0 \| \pi^\star) \le \beta^2$
- $D_{\mathrm{KL}}(\pi_0 \| \pi^\star) = \mathbb{E}_{\pi_0}[\log(\pi_0/\pi^\star)] \le \log \beta$

LSI and PI constants

Lemma. If π satisfies a log-Sobolev inequality with constant C, it also satisfies a Poincaré inequality with constant C.

How to determine the LSI or PI constant?

$$\mu\text{-strongly logconcave} \implies C_{\mathrm{PI}} = \frac{1}{\mu} \quad \text{(Brascamp-Lieb inequlity)}$$

$$\implies C_{\mathrm{LSI}} = \frac{1}{\mu} \quad \text{(Bakry–Émery thoerem)}$$

LSI / PI implies concentration

Lemma. Let $\pi^\star$ satisfies a log-Sobolev inequality with constant C_{LSI}. Then for any 1-Lipschitz function $f: \mathbb{R}^d \to \mathbb{R}$, we have that

$$\Pr_{\mathbf{x}\sim\pi^\star}[f(\mathbf{x}) \geq \mathbb{E}_{\pi^\star}[f] + \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2C_{\mathrm{LSI}}}\right) \quad \forall\, \epsilon > 0$$

sub-Gaussian

Lemma. Let $\pi^\star$ satisfies a Poincaré inequality with constant C_{PI}. Then for any 1-Lipschitz function $f: \mathbb{R}^d \to \mathbb{R}$, we have that

$$\Pr_{\mathbf{x}\sim\pi^\star}[f(\mathbf{x}) \geq \mathbb{E}_{\pi^\star}[f] + \epsilon] \leq 3\exp\left(-\frac{\epsilon}{\sqrt{C_{\mathrm{PI}}}}\right) \quad \forall\, \epsilon > 0$$

sub-exponential