Continuous-Time Diffusion in ML Theory

Schedule

1. Introduction, Brownian motion and Itô's calculus, basic concepts of SDE
2. Fokker-Planck equation, diffusion processes, convergence of diffusion process, coupling and mixing time
3. Functional inequalities and diffusion process convergence
4. Discretization of diffusion process, application to MCMC and stochastic optimization
5. Advanced topics in discretization and sampling, Score-based diffusion generative models
6. Denoising diffusion generative models, discretization and learning of diffusion models
7. RL basics, value functions and value learning
8. Continuous-time control problem, HJB equations, continuous-time RL
9. Advanced topics in continuous-time RL, applications to diffusion model fine-tuning

Lecture 1. Introduction, Brownian motion and Itô's calculus, basic concepts of SDE

Brownian Motion

Motivation: simple random walk
For a discrete-time stochastic process $X_{n+1} = X_n + Z_{n+1}$, where $X_0 = 0$ and
$Z_n = \begin{cases} 1, & \text{w.p. } \frac{1}{2} \\ -1, & \text{w.p. } \frac{1}{2} \end{cases}$
Consider X_{nt}. By CLT, for fixed $t > 0$, $\frac{1}{\sqrt{n}} X_{\lfloor nt \rfloor} \xrightarrow{d} \mathcal{N}(0, t)$
Hope: $\left(\frac{1}{\sqrt{n}} X_{\lfloor nt \rfloor} : 0 \le t \le T\right) \xrightarrow{d} \text{Something}$
Let $(B_t : t \ge 0)$ be the limit. Equivalently, we can define BM as a Gaussian Process.

- For any $t_1, t_2, \cdots, t_n$, $B(t_1), B(t_2), \cdots, B(t_n)$ is a Gaussian r.v.
- Independent increment: $(B(t) - B(s)) \perp B(s)$ for $t > s$.
- $\mathbb{E}\left[B_t B_s\right] = min(t, s)$.
- Regularity: w.p.1, $(B_t : t \ge 0)$ is a continuous function.
  - Everywhere continuous, nowhere differentiable.
  - Graph has fractal dimension 1.5
- BM is a martingale(鞅).
  - Definition of martingale: $\mathbb{E}\left[B_t \mid (B_r : 0 \le r \le s)\right] = B_s$
  - Important theorem of martingale: **Optimal stopping theorem**

- (T is stopping time, satisfying $|B_T| \le C < +\infty$), then $\mathbb{E}[B_T] = \mathbb{E}[B_0] = 0$
  It can be shown that processes satisfying these 5 properties.

Itô's Calculus

> **Problem**
>
> Can we make sense of $\int_0^t Y_s \mathrm{d}B_s$ (for some adapted stochastic process Y), where $Y_t \in \mathcal{F}_t$
>
> - $\mathcal{F}_t$ is filtration (Information about B within time $[0,t]$)

Recall Riemann-Stieltjes integral.
Let $h \in \mathbb{C}^1$, $\int_0^t f(s)\mathrm{d}h(s) := \lim\limits_{n\to+\infty} \sum\limits_{i=0}^{n} f(\xi_i)\left(h(s_{i+1}) - h(s_i)\right)$
For BM, the limit of the sum depends on how we choose $\xi_i \in [s_i, s_{i+1}]$
e.g. For $\int_0^1 B_t \mathrm{d}B_t$

- By selecting left end-point: $I_l(n) = \sum\limits_{i=0}^{n-1} B(s_i)\left(B(s_{i+1}) - B(s_i)\right)$
- By selecting right end-point: $I_r(n) = \sum\limits_{i=0}^{n-1} B(s_{i+1})\left(B(s_{i+1}) - B(s_i)\right)$
  $I_r(n) - I_l(n) = \sum\limits_{i=0}^{n-1} \left(B(s_{i+1}) - B(s_i)\right)^2$. Each term is square of $\mathcal{N}(0, s_{i+1} - s_i)$, so
  $\mathbb{E}[I_r(n)] - \mathbb{E}[I_l(n)] = \sum\limits_{i=0}^{n-1} (s_{i+1} - s_i) = 1$
  In fact, we can compute:
- $I_l(n) \to \frac{1}{2}(B_1^2 - 1)$
- $I_r(n) \to \frac{1}{2}(B_1^2 + 1)$

> **Definition**
>
> **Itô's calculus**. Always take left end-point to calculate the limit.
>
> Itô's calculus define $\int_0^t Y_s \mathrm{d}B_s$ as the limit of $\sum\limits_{i=0}^{n-1} Y_{t_i}(B_{t_{i+1}} - B_{t_i})$ ($\mathbb{L}^2$ convergence)
>
> - $\sum\limits_{i=0}^{n-1} Y_{t_i}(B_{t_{i+1}} - B_{t_i})$ is Cauchy in $\mathbb{L}^2$ metric

Itô's formula

> **Problem**

Is there analogue of Newton-Leibniz for Itô's calculus?

- For Stratonovich integral. ✅
- For Itô integral. Need correction

e.g. $\int_0^1 B_t \mathrm{d}B_t \neq \frac{1}{2}B_1^2 - \frac{1}{2}B_0^2$

Derivation: for $f \in \mathbb{C}^2$, $f(B_t) - f(B_0) = \sum_{i=0}^{n-1}(f(B_{t_{i+1}}) - f(B_{t_i}))$

Taylor expansion:

$$f(B_{t_{i+1}}) - f(B_{t_i}) = (B_{t_{i+1}} - B_{t_i})f'(B_{t_i}) + \frac{1}{2}(B_{t_{i+1}} - B_{t_i})^2 f''(B_{t_i}) + o(|B_{t_{i+1}} - B_{t_i}|^2)$$

Then,

$$f(B_t) - f(B_0) = \sum_{i=0}^{n-1}(B_{t_{i+1}} - B_{t_i})f'(B_{t_i}) + \frac{1}{2}\sum_{i=0}^{n-1}(B_{t_{i+1}} - B_{t_i})^2 f''(B_{t_i}) + o(\sum_{i=0}^{n-1}|B_{t_{i+1}} - B_{t_i}|^2)$$

- Term 1: $\int_0^t f'(B_s)\mathrm{d}B_s$.
- Term 2: $\mathbb{E}\left[(B_{t_{i+1}} - B_{t_i})^2\right] = t_{i+1} - t_i$, so $\sum_{i=0}^{n-1}(B_{t_{i+1}} - B_{t_i})^2 f''(B_{t_i}) \to \int_0^t f''(B_s)\mathrm{d}s$
- Term 3: $(B_{t_{i+1}} - B_{t_i})^2 \sim t_{i+1} - t_i$, so this term converge to zero.

Itô's formulae

Integral form:

$$\boxed{f(B_t) - f(B_0) = \int_0^t f'(B_s)\mathrm{d}B_s + \frac{1}{2}\int_0^t f''(B_s)\mathrm{d}s}$$

Differential form:

$$\boxed{\mathrm{d}f(B_t) = f'(B_t)\mathrm{d}B_t + \frac{1}{2}f''(B_t)\mathrm{d}t}$$

e.g. $f(x) = x^2$

Sol. We have $B_t^2 = B_t^2 - B_0^2 = 2\int_0^t B_s \mathrm{d}B_s + \frac{1}{2}\int_0^t 2\mathrm{d}s$. Then, $\int_0^t B_s \mathrm{d}B_s = \frac{1}{2}(B_t^2 - t)$

Chain rule

Suppose: $\mathrm{d}X_t = Y_t \mathrm{d}t + Z_t \mathrm{d}B_t$

Recall the chain rule in calculus, suppose $\mathrm{d}f(t) = g(t)\mathrm{d}t$, then $\mathrm{d}h\left(f(t)\right) = h'\left(f(t)\right)g(t)\mathrm{d}t$

Calculate $\mathrm{d}f(X_t) = ?$

Idea: telescope sum, for each term:
$f(X_{t_{i+1}}) - f(X_{t_i}) = (X_{t_{i+1}} - X_{t_i})f'(X_{t_i}) + \frac{1}{2}(X_{t_{i+1}} - X_{t_i})^2 f''(X_{t_i}) + o(|X_{t_{i+1}} - X_{t_i}|^2)$

- Term 1 $\approx f'(X_{t_i})[Y_{t_i}(t_{i+1} - t_i) + Z_{t_i}(B_{t_{i+1}} - B_{t_i})]$
- Term 2 $\approx \frac{1}{2}f''(X_{t_i})[Y_{t_i}(t_{i+1} - t_i) + Z_{t_i}(B_{t_{i+1}} - B_{t_i})]^2 \approx \frac{1}{2}f''(X_{t_i})Z_{t_i}^2(B_{t_{i+1}} - B_{t_i})^2$
- Term 3: high order, converge to zero after summation.
  It can be shown:

$$\sum_{i=0}^{n-1} f''(X_{t_i})Z_{t_i}^2(B_{t_{i+1}} - B_{t_i})^2 \to \int_0^t f''(X_s)Z_s^2 \mathrm{d}s$$

  So, we have chain rule in Itô's calculus:

$$f(X_t) - f(X_0) = \int_0^t f'(X_s)\mathrm{d}X_s + \frac{1}{2}\int_0^t f''(X_s)Z_s^2 \mathrm{d}s$$

  And differential form is:

$$\mathrm{d}f(X_t) = f'(X_t)\mathrm{d}X_t + \frac{1}{2}f''(X_t)Z_t^2 \mathrm{d}t$$

Indeed, we define:

$$\begin{aligned}\langle X\rangle_t &:= \lim_{n\to+\infty}\sum_{i=0}^{n-1}(X_{t_{i+1}} - X_{t_i})^2 \\ &= \int_0^t Z_s^2 \mathrm{d}s\end{aligned}$$

Then, $\mathrm{d}\langle X\rangle_t = Z_t^2\mathrm{d}t$.
So,

$$\boxed{\mathrm{d}f(X_t) = f'(X_t)\mathrm{d}X_t + \frac{1}{2}f''(X_t)\mathrm{d}\langle X\rangle_t}$$

Extension
Non-time homogeneous:

$$\mathrm{d}f(t, X_t) = \frac{\partial f}{\partial t}(t, X_t)\mathrm{d}t + \frac{\partial f}{\partial X}(t, X_t)\mathrm{d}X_t + \frac{1}{2}\frac{\partial^2 f}{\partial X^2}(t, X_t)\mathrm{d}\langle X\rangle_t$$

Suppose that $\mathbf{X}_t = \mathbf{Y}_t\mathrm{d}t + \mathbf{Z}_t\mathrm{d}\mathbf{B}_t$, where $\mathbf{X}_t, \mathbf{Y}_t \in \mathbb{R}^d$, $\mathbf{Z}_t \in \mathbb{R}^{d\times d}$ and $(\mathbf{B}_t : t \geq 0)$ is d-dim BM.

$$\begin{aligned}\mathrm{d}f(\mathbf{X}_t) &= \langle \nabla f(\mathbf{X}_t), \mathbf{Y}_t\rangle \mathrm{d}t + \langle \nabla f(\mathbf{X}_t), \mathbf{Z}_t\mathrm{d}\mathbf{B}_t\rangle + \frac{1}{2}\mathrm{Tr}(\mathbf{Z}_t^\top \nabla^2 f(\mathbf{X}_t)\mathbf{Z}_t)\mathrm{d}t \\ &= \langle \nabla f(\mathbf{X}_t), \mathbf{Y}_t\rangle \mathrm{d}t + \langle \nabla f(\mathbf{X}_t), \mathbf{Z}_t\mathrm{d}\mathbf{B}_t\rangle + \frac{1}{2}\mathrm{Tr}(\nabla^2 f(\mathbf{X}_t)\mathbf{Z}_t\mathbf{Z}_t^\top)\mathrm{d}t\end{aligned}$$

> Consider the 2nd order term in Taylor expansion
> $\frac{1}{2}(\mathbf{Z}_{t_i}(\mathbf{B}_{t_{i+1}}-\mathbf{B}_{t_i}))^\top \nabla^2 f(\mathbf{X}_{t_i})(\mathbf{Z}_{t_i}(\mathbf{B}_{t_{i+1}}-\mathbf{B}_{t_i}))$

Stochastic Differential Equation and Diffusion Process

Diffusion process is defined as: $\mathrm{d}\mathbf{X}_t = \mathbf{b}_t(\mathbf{X}_t)\mathrm{d}t + \mathbf{\Sigma}_t(\mathbf{X}_t)^{\frac{1}{2}}\mathrm{d}\mathbf{B}_t$

- Is a Markov process.
- Is a solution to an SDE. (Strong solution exists and is unique when $(\mathbf{b}, \mathbf{\Sigma})$ are Lipschitz)

e.g. Ornstein-Uhlenbeck process. $\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t\mathrm{d}t + \mathrm{d}\mathbf{B}_t$. (In general, $\mathrm{d}\mathbf{X}_t = \mathbf{A}_t\mathbf{X}_t\mathrm{d}t + \mathbf{M}_t\mathrm{d}\mathbf{B}_t$, where $\mathbf{A}_t, \mathbf{M}_t \in \mathbb{R}^{d\times d}$)
Note that $\mathrm{d}(e^t\mathbf{X}_t) = e^t\mathbf{X}_t\mathrm{d}t + e^t\mathrm{d}\mathbf{X}_t = e^t\mathbf{X}_t\mathrm{d}t + e^t(-\mathbf{X}_t\mathrm{d}t + \mathrm{d}\mathbf{B}_t) = e^t\mathrm{d}\mathbf{B}_t$.
Then, $\mathbf{X}_t = e^{-t}(\int_0^t e^s\mathrm{d}\mathbf{B}_s + \mathbf{X}_0)$
Finally, $\mathbf{X}_t \sim \mathcal{N}\left(e^{-t}\mathbf{X}_0, (e^{-2t}\int_0^t e^{2s}\mathrm{d}s)\mathbb{I}_d\right) = \mathcal{N}\left(e^{-t}\mathbf{X}_0, \frac{1-e^{-2t}}{2}\mathbb{I}_d\right)$

e.g. Langevin diffusion. $\mathrm{d}\mathbf{X}_t = -\frac{1}{2}\nabla f(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t$. Convergence: Next lecture

e.g. Black-Scholes model. $\mathrm{d}X_t = \mu X_t\mathrm{d}t + \sigma X_t\mathrm{d}B_t$, $X_0 > 0$
Let $S_t = X_0\exp(at + bB_t)$, calculate that
$\mathrm{d}S_t = aX_0\exp(at+bB_t)\mathrm{d}t + bX_0\exp(at+bB_t)\mathrm{d}B_t + \frac{1}{2}b^2X_0\exp(at+bB_t)\mathrm{d}t$
Then, $\mathrm{d}S_t = (a+\frac{b^2}{2})X_0\exp(at+bB_t)\mathrm{d}t + bX_0\exp(at+bB_t)\mathrm{d}B_t = \left(a+\frac{b^2}{2}\right)S_t\mathrm{d}t + bS_t\mathrm{d}B_t$
Let $a = \mu - \frac{\sigma^2}{2}$, $b = \sigma$, $X_t = X_0\exp((\mu - \frac{\sigma^2}{2})t + \sigma B_t)$ is the solution. "geometric BM"

- e.g. If $\mu < \frac{\sigma^2}{2}$, $t \to 0$, $X_t \to 0$ (exponentially fast). In ML, this corresponds to a stochastic optimization problem.

Lecture 2. Fokker-Planck equation, diffusion processes, convergence of diffusion process, coupling and mixing time

Diffusion process

For Diffusion process(time- homogeneous): $\mathrm{d}\mathbf{X}_t = \mathbf{b}(\mathbf{X}_t)\mathrm{d}t + \mathbf{\Sigma}(\mathbf{X}_t)^{\frac{1}{2}}\mathrm{d}\mathbf{B}_t$
We want to formulate its characterization:
e.g. Analogy to Kolmogorov-Chapman Equation in Markov Chain
We have $X_0, \cdots, X_t$ in Markov Chain with transition matrix $\mathbf{P}$

- If $X_0 \sim \pi_0$, then $X_t \sim \pi_0\mathbf{P}^t$
- For function f on state space, $\mathbb{E}[f(X_t) \mid X_0 = x] = (\mathbf{P}^t f)(x)$
  Important properties: **Stationary**, **Convergence**, **Rate**...

For the first question

Suppose we have a (**smooth** enough) potential function f

Question: What is $\mathbb{E}[f(\mathbf{X}_t) \mid \mathbf{X}_0 = \mathbf{x}]$ where $\mathbf{X}_t$ follows $\mathrm{d}\mathbf{X}_t = \mathbf{b}(\mathbf{X}_t)\mathrm{d}t + \mathbf{\Sigma}(\mathbf{X}_t)^{\frac{1}{2}}\mathrm{d}\mathbf{B}_t$

$$\begin{aligned}\mathbb{E}[f(\mathbf{X}_T) \mid \mathbf{X}_0 = \mathbf{x}] = f(\mathbf{x}) + \mathbb{E}\bigg[\int_0^T \nabla f(\mathbf{X}_t)^\top \mathbf{b}(\mathbf{X}_t)\mathrm{d}t + \underbrace{\cancel{\int_0^T \nabla f(\mathbf{X}_t)^\top \mathbf{\Sigma}(\mathbf{X}_t)^{\frac{1}{2}}\mathrm{d}\mathbf{B}_t}}_{\text{martingale}} \\ + \frac{1}{2}\int_0^T \mathrm{Tr}(\nabla^2 f(\mathbf{X}_t)\mathbf{\Sigma}(\mathbf{X}_t))\mathrm{d}t\bigg] \\ = f(\mathbf{x}) + \mathbb{E}\left[\int_0^T \left[\nabla f(\mathbf{X}_t)^\top \mathbf{b}(\mathbf{X}_t) + \frac{1}{2}\mathrm{Tr}(\nabla^2 f(\mathbf{X}_t)\mathbf{\Sigma}(\mathbf{X}_t))\right]\mathrm{d}t\right]\end{aligned}$$

Then, calculate that

$$\lim_{T \to 0^+} \frac{\mathbb{E}[f(\mathbf{X}_T) \mid \mathbf{X}_0 = \mathbf{x}] - f(\mathbf{x})}{T} = \nabla f(\mathbf{X}_t)^\top \mathbf{b}(\mathbf{X}_t) + \frac{1}{2}\mathrm{Tr}(\nabla^2 f(\mathbf{X}_t)\mathbf{\Sigma}(\mathbf{X}_t))$$

Analogous to Markov Chains, we define a operator $(\mathcal{P}_t)_{t\geq 0}$ satisfying $(\mathcal{P}_t f)(\mathbf{x}) = \mathbb{E}[f(\mathbf{X}_t) \mid \mathbf{X}_0 = \mathbf{x}]$. We call $(\mathcal{P}_t)_{t\geq 0}$ as semi-group

- Identity element: $\mathcal{P}_0 = \mathbb{I} : f \mapsto f$
- Multiplication: $\mathcal{P}_t \circ \mathcal{P}_s = \mathcal{P}_{t+s}$
- No inverse element.

  Let $\mathcal{A} = \frac{\mathrm{d}\mathcal{P}_t}{\mathrm{d}t}\Big|_{t=0}$, then $\mathcal{A}f(\mathbf{x}) = \nabla f(\mathbf{X}_t)^\top \mathbf{b}(\mathbf{X}_t) + \frac{1}{2}\mathrm{Tr}(\nabla^2 f(\mathbf{X}_t)\mathbf{\Sigma}(\mathbf{X}_t))$

  $\mathcal{A}$ is a linear operator, $\mathcal{A}(\lambda f_1 + \mu f_2) = \lambda\mathcal{A}f_1 + \mu\mathcal{A}f_2$

  e.g. For finite-dim matrix $\mathbf{A}$, if $\frac{\mathrm{d}P_t}{\mathrm{d}t} = \mathbf{A}P_t$, then $P_t = P_0\exp(t\mathbf{A})$

Fokker-Planck Equation

Let π_t be the density of $\mathbf{X}_t$, then,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[f(\mathbf{X}_t)\right] &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} f(\mathbf{x})\pi_t(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} f(\mathbf{x})\frac{\partial \pi_t}{\partial t}(\mathbf{x})\mathrm{d}\mathbf{x}\end{aligned}$$

On the other hand,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[f(\mathbf{X}_t)\right] &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[(\mathcal{P}_t f)(\mathbf{X}_0)\right] \\ &= \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{P}_t f)(\mathbf{X}_0)\right] \\ &= \mathbb{E}\left[(\mathcal{A}\mathcal{P}_t f)(\mathbf{X}_0)\right]\end{aligned}$$

Let $t = 0$, then,

$$\int_{\mathbb{R}^d} f(\mathbf{x}) \frac{\partial \pi_t(\mathbf{x})}{\partial t}\bigg|_{t=0} \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} \mathcal{A} f(\mathbf{x}) \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x}$$

Integration-by-parts formulae

Single variable:

$$\boxed{\int_a^b f'(x) g(x) \mathrm{d}x = f(x) g(x)\Big|_a^b - \int_a^b f(x) g'(x) \mathrm{d}x}$$

Multivariate: For bounded domain Ω (with smooth boundary), if $f(x)g(x)\Big|_{\partial\Omega} = 0$, then

$$\boxed{\int_\Omega \nabla f(x) g(x) \mathrm{d}x = -\int_\Omega f(x) \nabla g(x) \mathrm{d}x}$$

In general, we have

$$\boxed{\int_\Omega \nabla f(x) g(x) \mathrm{d}x = \oint_{\partial\Omega} f(x) g(x) \mathrm{d}\mathbf{n}(x) - \int_\Omega f(x) \nabla g(x) \mathrm{d}x}$$

A non-rigorous proof: Let $\Omega = \mathbb{B}(0, \mathbb{R})$, then

$$\left| \int_\Omega f(\mathbf{x}) \partial_t \pi_0 \mathrm{d}\mathbf{x} - \int_{\mathbb{R}^d} f(\mathbf{x}) \partial_t \pi_0 \mathrm{d}\mathbf{x} \right| \to 0$$

Similarly for $\mathcal{A} f \cdot \pi_0$.

$$\begin{aligned}
\int_\Omega \mathcal{A} f(\mathbf{x}) \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x} &= \int_\Omega \nabla f(\mathbf{x})^\top \mathbf{b}(\mathbf{x}) \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x} + \frac{1}{2} \int_\Omega \mathrm{Tr}(\mathbf{\Sigma}(\mathbf{x}) \nabla^2 f(\mathbf{x})) \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \int_\Omega \left[\nabla f(\mathbf{x})^\top \mathbf{b}(\mathbf{x}) \pi_0(\mathbf{x}) \mathrm{d}\mathbf{x} + \frac{1}{2} \mathrm{Tr}(\mathbf{\Sigma}(\mathbf{x}) \nabla^2 f(\mathbf{x})) \pi_0(\mathbf{x}) \right] \mathrm{d}\mathbf{x}
\end{aligned}$$

For the first term,

$$\text{First term} = \oint_{\partial\Omega} \pi_0(\mathbf{x}) f(\mathbf{x}) \mathbf{b}(\mathbf{x})^\top \mathrm{d}\mathbf{n}(\mathbf{x}) - \int_\Omega f(\mathbf{x}) \nabla (\pi_0 \mathbf{b})(\mathbf{x}) \mathrm{d}\mathbf{x}$$

as $\|\mathbf{x}\| \to +\infty$, $f(\mathbf{x})\mathbf{b}(\mathbf{x})^\top$ grows polynomially; while in $|\partial\Omega| = C_d \mathbb{R}^d$, $\pi_0(\mathbf{x}) \leq C_1 \exp(-C_2 \|\mathbf{x}\|^2)$. Then, $\oint_{\partial\Omega} \pi_0(\mathbf{x}) f(\mathbf{x}) \mathbf{b}(\mathbf{x})^\top \mathrm{d}\mathbf{n}(\mathbf{x}) \to 0$ at $\partial\Omega$.

For the second term,

$$\begin{aligned}\text{Second term} &= \int_{\mathbb{R}^d} \mathrm{Tr}(\nabla^2 f(\mathbf{x})\mathbf{\Sigma}(\mathbf{x}))\pi_0(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= -\int_{\mathbb{R}^d} \nabla f(\mathbf{x}) \cdot (\nabla \cdot \pi\mathbf{\Sigma})(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} f(\mathbf{x})\nabla^2(\pi\mathbf{\Sigma})(\mathbf{x})\mathrm{d}\mathbf{x}\end{aligned}$$

(For simplicity, ignore $\text{boundary term} \to 0$)

Putting them together,

$$\int f(\mathbf{x})\partial_t\pi_0(\mathbf{x})\mathrm{d}\mathbf{x} = \int f(\mathbf{x}) \left[-\nabla \cdot (\pi_0\mathbf{b})(\mathbf{x}) + \frac{1}{2}\nabla^2(\pi_0\mathbf{\Sigma})(\mathbf{x})\right]\mathrm{d}\mathbf{x}$$

so (by time-homogenous), we get **Fokker-Planck equation**:

$$\boxed{\partial_t\pi_t(x) = -\nabla \cdot (\pi_t\mathbf{b})(\mathbf{x}) + \frac{1}{2}\nabla^2(\pi_t\mathbf{\Sigma})(\mathbf{x})}$$

Remark: we got adjoint operator: $\mathcal{A}^\star\pi = -\nabla(\pi_t\mathbf{b})(\mathbf{x}) + \frac{1}{2}\nabla^2(\pi_t\mathbf{\Sigma})(\mathbf{x})$
Let $\langle f, g\rangle = \int_{\mathbb{R}^d} fg\mathrm{d}\mathbf{x}$, then we have $\langle \mathcal{A}f, \pi_t\rangle = \langle f, \partial_t\pi\rangle$, so
$\mathcal{A}^\star\pi = \partial_t\pi = -\nabla \cdot (\pi_t\mathbf{b})(\mathbf{x}) + \frac{1}{2}\nabla^2(\pi_t\mathbf{\Sigma})(\mathbf{x})$

e.g. $\mathrm{d}\mathbf{X}_t = \mathrm{d}\mathbf{B}_t$. Then, we get $\partial_t\pi_t = \frac{1}{2}\Delta\pi_t$. (Fundamental solution: Gaussian convolution)

e.g.(Langevin). $\mathrm{d}\mathbf{X}_t = -\nabla f(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$. Then, we get $\partial_t\pi_t = \nabla \cdot (\pi\nabla f) + \Delta\pi_t$

Sampling Problem and Coupling

Given $f : \mathbb{R}^d \to \mathbb{R}$, generate a sample from $\pi(\mathbf{x}) = \frac{\exp(-f(\mathbf{x}))}{\int_{\mathbb{R}^d}\exp(-f(\mathbf{y}))\mathrm{d}\mathbf{y}}$
Basic idea:

- Importance sampling
- Rejection sampling
  scales poorly with high dim.

MCMC (Markov Chain Monte Carlo): Run a stochastic process with **stationary distribution** π

Stationary distribution of Langevin diffusion
Recall, $\partial_t\pi_t = \nabla \cdot (\pi\nabla f) + \Delta\pi_t$. Its stationary condition is $\nabla \cdot (\pi\nabla f) + \Delta\pi = 0$
Let $\pi = \exp(-f)$, then,

$$\begin{aligned}\nabla \cdot (\pi\nabla f) &= \nabla \cdot (\exp(-f)\nabla f) \\ &= \Delta f \cdot \exp(-f) - |\nabla f|^2\exp(-f)\end{aligned}$$

$$\begin{aligned}\Delta exp(-f) &= \nabla \cdot (\nabla\exp(-f)) \\ &= \nabla \cdot (-\nabla f \cdot \exp(-f)) \\ &= -\Delta f \cdot \exp(-f) + |\nabla f|^2\exp(-f)\end{aligned}$$

so $\pi \propto \exp(-f)$ is a stationary distribution.

- $\exists!$ under weak condition.
- Convergence (rate) in what sense?

Distance (divergence) between probability distribution:
Total variation:

$$d_{\mathrm{TV}}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2} \int |p(x) - q(x)| \mathrm{d}x = \sup_{\|f\|_\infty \leq 1} |\mathbb{E}_\mathcal{P} f(x) - \mathbb{E}_\mathcal{Q} f(x)|$$

Kullback-Leibler:

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{P} \| \mathcal{Q}) = \mathbb{E}_\mathcal{P} \left[\log \frac{p(x)}{q(x)} \right]$$

Pinsker's inequality:

$$d_{\mathrm{TV}}(\mathcal{P}, \mathcal{Q}) \leq \sqrt{\frac{1}{2} \mathcal{D}_{\mathrm{KL}}(\mathcal{P} \| \mathcal{Q})}$$

Wasserstein distance
Let $X \sim \mathcal{P}$, $Y \sim \mathcal{Q}$.
Coupling: Find a joint distribution γ, s.t. $\gamma\Big|_X = \mathcal{P}$, $\gamma\Big|_Y = \mathcal{Q}$.
Then we can define **Wasserstein distance** as follows,

$$\begin{aligned}
\mathcal{W}_1(\mathcal{P}, \mathcal{Q}) &= \inf_{\gamma \in \mathrm{Coupling}(\mathcal{P}, \mathcal{Q})} \mathbb{E}_\gamma [\|X - Y\|] \\
&= \sup_{\|\nabla f\|_2 \leq 1} |\mathbb{E}_\mathcal{P} f(x) - \mathbb{E}_\mathcal{Q} f(x)| \\
\mathcal{W}_2(\mathcal{P}, \mathcal{Q}) &= \sqrt{\inf_{\gamma \in \mathrm{Coupling}(\mathcal{P}, \mathcal{Q})} \mathbb{E}_\gamma [\|X - Y\|^2]}
\end{aligned}$$

By Cauchy-Schwarz inequality, $\mathcal{W}_1(\mathcal{P}, \mathcal{Q}) \leq \mathcal{W}_2(\mathcal{P}, \mathcal{Q})$

Convergence rate of Langevin diffusion

Assumption: $\nabla^2 f \succeq \lambda \mathbb{I}_d$, $\forall \mathbf{x}$, for some $\lambda > 0$
Strong convexity (assumption above) implies

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle \geq \lambda \|\mathbf{x} - \mathbf{y}\|^2$$

Proof

By the following calculation, it is trivial.

$$\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) = \int_0^1 \nabla^2 f\left(t\mathbf{x} + (1-t)\mathbf{y}\right)(\mathbf{x} - \mathbf{y})\mathrm{d}t$$

Goal: construct a coupling between $\mathbf{X}_T$ (time T marginal (π_T) of diffusion) and $\mathbf{X}^\star$(distribution π)

Let $\mathrm{d}\mathbf{X}_t = -\nabla f(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$, $\mathbf{X}_0 = \mathbf{x}_0$ and $\mathrm{d}\mathbf{X}_t^\star = -\nabla f(\mathbf{X}_t^\star)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$, $\mathbf{X}_0^\star \sim \pi$.

We have $\mathbf{X}_T^\star \overset{\mathrm{d}}{=} \mathbf{X}^\star$. We want to construct "Synchronous coupling": same BM.

Let $\mathcal{H}_t := \mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_t^\star\|^2]$

Calculate $\mathrm{d}(\mathbf{X}_t - \mathbf{X}_t^\star) = -(\nabla f(\mathbf{X}_t) - \nabla f(\mathbf{X}_t^\star))\mathrm{d}t$

Then,

$$\begin{aligned}\frac{\mathrm{d}\mathcal{H}_t}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_t^\star\|^2] \\ &= -2\mathbb{E}\left[\langle \mathbf{X}_t - \mathbf{X}_t^\star, \nabla f(\mathbf{X}_t) - \nabla f(\mathbf{X}_t^\star)\rangle\right] \\ &\le -2\lambda\mathbb{E}\left[\|\mathbf{X}_t - \mathbf{X}_t^\star\|^2\right] \\ &= -2\lambda\mathcal{H}_t\end{aligned}$$

Let $\Phi_t = \exp(2\lambda t)\mathcal{H}_t$, then $\frac{\mathrm{d}\Phi}{\mathrm{d}t} \le 0$, so $\Phi_t \le \Phi_0 = \mathcal{H}_0$, so $\mathcal{H}_t \le \exp(-2\lambda t)\mathcal{H}_0$.

Then,

$$\mathcal{W}_2^2(\pi_T, \pi) = \inf \mathcal{H}_T \le \mathcal{H}_T \le \exp(-2\lambda T)\mathbb{E}\left[\|\mathbf{X}_0 - \mathbf{X}_0^\star\|^2\right] \to 0$$

Extension. Underdamped Langevin

[1707.03663] Underdamped Langevin MCMC: A non-asymptotic analysis

$$\begin{cases}\mathrm{d}\mathbf{v}_t = -\gamma\mathbf{v}_t\mathrm{d}t - u\nabla f(\mathbf{x}_t)\mathrm{d}t + \sqrt{2\gamma u}\mathrm{d}\mathbf{B}_t \\ \mathrm{d}\mathbf{x}_t = \mathbf{v}_t\mathrm{d}t\end{cases}$$

Use Fokker-Planck equation, let $\pi_t : \mathbb{R}^{2d} \to \mathbb{R}_+$ be the density function.

Then,

$$\boxed{\partial_t\pi_t = \gamma\nabla_{\mathbf{v}}\cdot(\pi\mathbf{v}) + u\nabla_{\mathbf{v}}\cdot(\pi\nabla f(\mathbf{x})) - \nabla_{\mathbf{x}}\cdot(\pi\mathbf{v}) + \gamma u\Delta_{\mathbf{v}}\pi}$$

It stationary distribution: $\pi(\mathbf{x}, \mathbf{v}) \propto exp(-(f(\mathbf{x}) + \frac{\|\mathbf{v}\|_2^2}{2u}))$

Convergence: still synchronous coupling (assuming strong convexity)

Construct Lyapunov function:

$$\mathcal{H}_t = \mathbb{E}\left[\left\|\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}\mathbf{x}_t - \mathbf{x}_t^\star \\ \mathbf{v}_t - \mathbf{v}_t^\star\end{pmatrix}\right\|_2^2\right]$$

Let$\mathbf{z}_t = \mathbf{x}_t - \mathbf{x}_t^\star$, $\mathbf{y}_t = \mathbf{v}_t - \mathbf{v}_t^\star$, then,

$$\frac{\mathrm{d}\mathcal{H}_t}{\mathrm{d}t} = 2\mathbb{E}\left[\begin{pmatrix}\mathbf{z}_t + \mathbf{y}_t & \mathbf{y}_t\end{pmatrix}\begin{pmatrix}(1-\gamma)\mathbf{y}_t - u\left(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_t^\star)\right) \\ -\gamma\mathbf{y}_t - u\left(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_t^\star)\right)\end{pmatrix}\right]$$

Detour

Stability of dynamic system $\mathrm{d}\mathbf{x}_t = \mathbf{A}\mathbf{x}_t\mathrm{d}t$
If $\frac{\mathbf{A}+\mathbf{A}^\top}{2}$ has negative eigenvalues, then exp. converge to 0

∇f is nonlinear, but we state that $\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_t^\star) = \int_0^1 \nabla^2 f(\beta\mathbf{x}_t + (1-\beta)\mathbf{x}_t^\star)(\mathbf{x}_t - \mathbf{x}_t^\star)\mathrm{d}\beta$
Assume: $\lambda\mathbb{I} \preceq \nabla^2 f \preceq L\mathbb{I}$
So, $\mathbf{\Lambda}_t = \int_0^1 \nabla^2 f(\beta\mathbf{x}_t + (1-\beta)\mathbf{x}_t^\star)\mathrm{d}\beta \in [\lambda\mathbb{I}, L\mathbb{I}]$
Then,

$$\begin{aligned}\frac{\mathrm{d}\mathcal{H}_t}{\mathrm{d}t} &= -\begin{pmatrix}\mathbf{x}_t - \mathbf{x}_t^\star & \mathbf{v}_t - \mathbf{v}_t^\star\end{pmatrix}\begin{pmatrix}(\gamma-1)\mathbb{I}_d & u\mathbf{\Lambda}_t - (\gamma-1)\mathbb{I}_d \\ -\mathbb{I}_d & \mathbb{I}_d\end{pmatrix}\begin{pmatrix}\mathbf{x}_t - \mathbf{x}_t^\star \\ \mathbf{v}_t - \mathbf{v}_t^\star\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{z}_t + \mathbf{y}_t & \mathbf{y}_t\end{pmatrix}\delta_t\begin{pmatrix}\mathbf{z}_t + \mathbf{y}_t \\ \mathbf{y}_t\end{pmatrix}\end{aligned}$$

Need $\frac{\delta_t+\delta_t^\top}{2}$ negative definite, take $\gamma = 2$, $u = \frac{1}{L} \Rightarrow \lambda_{\min}(-\frac{\delta_t+\delta_t^\top}{2}) \geq \frac{\lambda}{2L}$

Lecture 3. Functional inequalities and diffusion process convergence

For DTMC (Discrete-time Markov Chain): "Mixing time"

- Coupling
- Spectral gap
  Spectral gap:
  Let stationary state $\pi = \pi\mathbf{P}$, where $\pi \in \mathbb{R}^{l\times n}$, $\mathbf{P} \in \mathbf{R}^{n\times n}$. We have $\pi_n = \pi_0\mathbf{P}^n$. Diagonalize $\mathbf{P} = \mathbf{D}\mathbf{\Lambda}\mathbf{D}^{-1}$, then $\pi_n = \pi_0\mathbf{D}\mathbf{\Lambda}^n\mathbf{D}^{-1}$. Its convergence rate depends on $\frac{1}{1-|\lambda_2|}$

Functional Inequalities

"The correct analogue of spectral gap"

log-Sobolev Inequality

Motivation.

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{D}_{\mathrm{KL}}(\pi_t\|\pi) &= \int \partial_t(\pi_t \log\frac{\pi_t}{\pi})\mathrm{d}\mathbf{x} \\
&= \int \partial_t\pi_t \log\frac{\pi_t}{\pi}\mathrm{d}\mathbf{x} + \int \pi_t\partial_t \log\pi_t \mathrm{d}\mathbf{x} \\
&= \int (\nabla\cdot(\pi_t\nabla f) + \Delta\pi_t)\left(\log\frac{\pi_t}{\pi}+1\right)\mathrm{d}\mathbf{x} \\
&= -\int (\pi_t\nabla f + \nabla\pi_t)^\top\left(\nabla\log\frac{\pi_t}{\pi}\right)\mathrm{d}\mathbf{x} \qquad (f=-\log\pi, \nabla\pi_t = \pi_t\nabla\log\pi_t) \\
&= -\int \|\nabla\log\pi_t - \nabla\log\pi\|_2^2\pi_t\mathrm{d}\mathbf{x} \\
&= -\int \pi_t\|\nabla\log\frac{\pi_t}{\pi}\|_2^2\mathrm{d}\mathbf{x}
\end{aligned}
$$

Def. We say π satisfies log-Sobolev inequality with constant C_{LSI} iff for $\forall\mu$ (sufficiently smooth),

$$
\mathcal{D}_{\mathrm{KL}}(\mu\|\pi) \le C_{\mathrm{LSI}} \underbrace{\int \mu(\mathbf{x})\|\nabla\log\frac{\mu}{\pi}\|_2^2\mathrm{d}\mathbf{x}}_{I(\mu\|\pi)\text{: Relative Fisher Information}}
$$

Suppose that LSI holds, we have

$$
\boxed{\mathcal{D}_{\mathrm{KL}}(\pi_t\|\pi) \le \mathcal{D}_{\mathrm{KL}}(\pi_0\|\pi)\exp(-\frac{t}{C_{\mathrm{LSI}}})}
$$

Proof of LSI under strong convexity condition. (This is an instance of "**interpolation method**")

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}I(\pi_t\|\pi) &= \underbrace{-2\mathbb{E}\left[\|\nabla^2\log\frac{\pi_t}{\pi}\|_F^2\right]}_{\le 0}\underbrace{-2\mathbb{E}\left[(\nabla\log\frac{\pi_t}{\pi})^\top\nabla^2 f(\nabla\log\frac{\pi_t}{\pi})\right]}_{\le -2\lambda I(\pi_t\|\pi)} \\
&\le -2\lambda I(\pi_t\|\pi) \qquad \text{(Under strong convexity)} \\
&(\text{Then } I(\pi_t\|\pi) \le I(\pi_0\|\pi)\exp(-2\lambda t))
\end{aligned}
$$

Also, we have

$$
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{D}_{\mathrm{KL}}(\pi_t\|\pi) = -I(\pi_t\|\pi)
$$

So we have

$$
\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(\pi_0\|\pi) &= \int_0^\infty I(\pi_t\|\pi)\mathrm{d}t \\
&\le \int_0^\infty I(\pi_0\|\pi)\exp(-2\lambda t)\mathrm{d}t \\
&= \frac{1}{2\lambda}I(\pi_0\|\pi)
\end{aligned}
$$

Poincare Inequality

Motivation. Let

$$
\begin{aligned}
\chi^2(\mathcal{P}\|\mathcal{Q}) &= \int q(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)^2 \mathrm{d}\mathbf{x} \\
&= \int p(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}-1\right)\mathrm{d}\mathbf{x}
\end{aligned}
$$

Then

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\chi^2(\pi_t\|\pi) &= \frac{\mathrm{d}}{\mathrm{d}t}\int \pi_t\left(\frac{\pi_t}{\pi}-1\right)\mathrm{d}\mathbf{x} \\
&= \int \left(\partial_t\pi_t(\frac{\pi_t}{\pi}-1)+\pi_t\frac{\partial_t\pi_t}{\pi}\right)\mathrm{d}\mathbf{x} \\
&= \int \partial_t\pi_t\left(\frac{2\pi_t}{\pi}-1\right)\mathrm{d}\mathbf{x} \\
&= \int (\nabla\cdot(\pi_t\nabla f)+\Delta\pi_t)\left(\frac{2\pi_t}{\pi}-1\right)\mathrm{d}\mathbf{x} \\
&= -\int (\pi_t\nabla f+\nabla\pi_t)\frac{2\nabla\pi_t}{\pi}\mathrm{d}\mathbf{x} \qquad (f=-\log\pi, \nabla\pi_t=\pi_t\nabla\log\pi_t) \\
&= -2\int \pi_t(\nabla\log\frac{\pi_t}{\pi})\nabla\left(\frac{\pi_t}{\pi}\right)\mathrm{d}\mathbf{x} \quad \left(\nabla\frac{\pi_t}{\pi}=\frac{\pi_t}{\pi}\nabla\log\frac{\pi_t}{\pi}\right) \\
&= -2\int \pi\|\nabla\left(\frac{\pi_t}{\pi}\right)\|_2^2\mathrm{d}\mathbf{x}
\end{aligned}
$$

Def. We say π satisfies Poincare Inequality with constant C_{PI} iff for $\forall h$ (sufficiently smooth function),

$$
\mathrm{Var}_\pi(h(\mathrm{x})) \le C_{\mathrm{PI}}\mathbb{E}\left[\|\nabla h\|^2\right]
$$

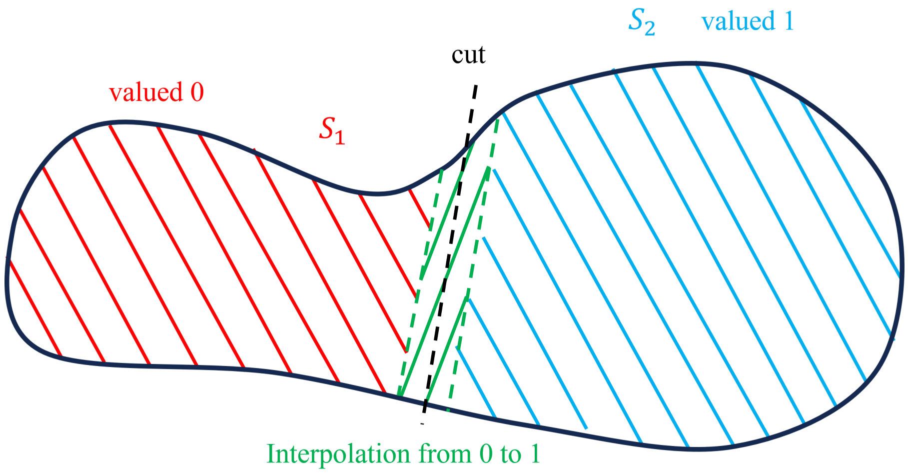
Suppose that PI holds, we have

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

Intuition. Functional inequality $\approx$ Isoperimetry Inequality $\approx$ Difficult to find a cut
Cut $S=S_1\oplus S_2$, a "good" cut is that: $\mathbb{P}(S_1)$ and $\mathbb{P}(S_2)$ big, while $\mathbb{P}(\partial S_1\setminus\partial S)$ small

Isoperimetry.

$$
\min_{S_1}\frac{\mathbb{P}(\partial S_1\setminus\partial S)}{\min(\mathbb{P}(S_1),\mathbb{P}(S_2))}
$$



Suppose we have a good cut, we can construct a continuous function valued 0 in area S_1, valued 1 in area S_2, and interpolation from 0 to 1 near ∂S_1. Then we have $\mathbb{E}\left[\|\nabla f\|^2\right]$ small, while $\mathrm{Var}(f)$ large.

log-Sobolev inequality and Poincare inequality

Compare:

PI: $\mathrm{Var}_\pi(f) \leq C_{\mathrm{PI}} \mathbb{E}_\pi[\|\nabla f\|^2]$
LSI: $\mathcal{D}_{\mathrm{KL}}(\mu\|\pi) \leq C_{\mathrm{LSI}} I(\mu\|\pi)$

Beyond convexity: An equivalent form of log-Sobolev inequality:

Change of variable: Let $f = C_{\mathrm{LSI}}\sqrt{\frac{\mu}{\pi}}$

LSI is equivalent to (for sufficiently smooth function f)

$$\boxed{\mathbb{E}_\pi[f^2 \log f^2] - \mathbb{E}_\pi[f^2]\mathbb{E}_\pi[\log f^2] \leq 4C_{\mathrm{LSI}}\mathbb{E}_\pi[\|\nabla f\|^2]}$$

LSI implies PI:

Apply equivalent form of LSI with $f = 1 + \epsilon h$, for small ϵ

$$\begin{aligned}
\mathrm{LHS} = & \ \mathbb{E}_\pi\left[(1 + 2\epsilon h + \epsilon^2 h^2) \cdot 2\left(\epsilon h - \frac{1}{2}\epsilon^2 h^2 + O(\epsilon^3)\right)\right] \\
& - \mathbb{E}_\pi[(1 + 2\epsilon h + \epsilon^2 h^2)] \cdot \mathbb{E}_\pi[\log\left(1 + 2\epsilon h + \epsilon^2 h^2\right)] \\
& = [2\epsilon\mathbb{E}_\pi[h] + 3\epsilon^2\mathbb{E}_\pi[h^2] + O(\epsilon^3)] \\
& - [1 + 2\epsilon\mathbb{E}_\pi[h] + \epsilon^2\mathbb{E}[h^2]] \cdot [2\epsilon\mathbb{E}_\pi[h] - \epsilon^2\mathbb{E}_\pi[h^2] + O(\epsilon^3)] \\
& = 4\epsilon^2\mathbb{E}_\pi[h^2] - 4\epsilon^2(\mathbb{E}_\pi[h])^2 + O(\epsilon^3) \\
& = 4\epsilon^2\mathrm{Var}_\pi(f) + O(\epsilon^3) \\
\mathrm{RHS} = & \ 4\epsilon^2 C_{\mathrm{LSI}}\mathbb{E}_\pi[\|\nabla h\|^2]
\end{aligned}$$

We get

$$4\epsilon^2\mathrm{Var}_\pi(f)+O(\epsilon^3)\leq 4\epsilon^2 C_{\mathrm{LSI}}\mathbb{E}_\pi[\|\nabla h\|^2]$$

So we get Poincare inequality $\mathbb{E}_\pi[h^2]\leq C_{\mathrm{LSI}}\mathbb{E}_\pi[\|\nabla h\|^2]$, $C_{\mathrm{PI}}\leq C_{\mathrm{LSI}}$

Some distribution and its tail bound

sub-Gaussian:

$$\begin{aligned}&\mathbb{E}[\exp(\lambda(X-\mathbb{E}X))]\leq\exp(\frac{\lambda^2\sigma^2}{2}),\forall\lambda\in\mathbb{R}\\ \iff&\mathbb{P}(|X|\geq t)\leq\exp(-\frac{t^2}{C\sigma^2}),\forall t\end{aligned}$$

sub-Gaussian random vector: $\forall\mathbf{u}$, $\|\mathbf{u}\|_2=1$, $\mathbf{u}^\top\mathbf{X}$ is sub-Gaussian

sub-exponential:

$$\begin{aligned}&\mathbb{E}[\exp(\lambda(X-\mathbb{E}X))]\leq\exp(\frac{\lambda^2\alpha^2}{2}),\text{ when }\lambda\leq\frac{1}{C\alpha}\\ \iff&\mathbb{P}(|X|\geq t)\leq\exp(-\frac{t}{\alpha}),\forall t\end{aligned}$$

Poincare inequality implies concentration of Lipschitz function:
For any 1-Lipschitz function $g:\mathbb{R}^d\to\mathbb{R}$
Let moment generating function $m(\lambda)=\mathbb{E}[\exp(\lambda g(\mathbf{x}))]$
By Poincare inequality,

$$\begin{aligned}\mathbb{E}[\exp(2\lambda g(\mathbf{x}))]-(\mathbb{E}[\exp(\lambda g(\mathbf{x}))])^2&\leq C_{\mathrm{PI}}\lambda^2\mathbb{E}[\|\nabla g\|^2\cdot\exp(2\lambda g(\mathbf{x}))]\quad(g\text{ is 1-Lipschitz})\\&\leq\lambda^2 C_{\mathrm{PI}}\mathbb{E}[\exp(2\lambda g(\mathbf{x}))]\end{aligned}$$

So when $\lambda<\frac{1}{\sqrt{2C_{\mathrm{PI}}}}$,

$$\begin{aligned}m(\lambda)&\leq\frac{1}{1-\lambda^2C_{\mathrm{PI}}}m^2(\frac{\lambda}{2})\\&\leq\frac{1}{1-\lambda^2C_{\mathrm{PI}}}\left(\frac{1}{1-\frac{\lambda^2}{4}C_{\mathrm{PI}}}\right)^2m^4(\frac{\lambda}{4})\\&\leq\prod_{k=0}^{n}\left(\frac{1}{1-\frac{1}{4^k}\lambda^2C_{\mathrm{PI}}}\right)^{2^k}m^{2^n}(\frac{\lambda}{2^{n+1}})\end{aligned}$$

$$
\begin{aligned}
\log \prod_{k=0}^{n}\left(\frac{1}{1-\frac{1}{4^k}\lambda^2 C_{\mathrm{PI}}}\right)^{2^k} &= -\sum_{k=0}^{n} 2^k \log\left(1-\frac{1}{4^k}\lambda^2 C_{\mathrm{PI}}\right) \\
&= \sum_{k=0}^{n} 2^k\left(\frac{1}{4^k}\lambda^2 C_{\mathrm{PI}} + O\left(\frac{\lambda^4}{16^k}\right)\right) \\
&= \sum_{k=0}^{n} \frac{1}{2^k}\lambda^2 C_{\mathrm{PI}} + O(\frac{\lambda^4}{8^k}) \\
&= 2\lambda^2 C_{\mathrm{PI}} - \frac{\lambda^2 C_{\mathrm{PI}}}{2^n} + O(\frac{\lambda^4}{8^k})
\end{aligned}
$$

Then,

$$
\begin{aligned}
m(\lambda) &\leq \prod_{k=0}^{n}\left(\frac{1}{1-\frac{1}{4^k}\lambda^2 C_{\mathrm{PI}}}\right)^{2^k} m^{2^n}\left(\frac{\lambda}{2^{n+1}}\right) \qquad \text{Let n} \to \infty \\
&\leq \exp(2\lambda^2 C_{\mathrm{PI}})
\end{aligned}
$$

e.g. Let $g(\mathbf{x}) = \|\mathbf{x}\|_2$, $\mathbf{x} \in \mathbb{R}^d$. We have, w.h.p. $|\|\mathbf{x}\|_2 - \underbrace{\mathbb{E}[\|\mathbf{x}\|_2]}_{\Theta(\sqrt{d})}| \lesssim \underbrace{\sqrt{C_{\mathrm{PI}}}}_{\text{dim-independent}}$

Nearly all mass are in a thin shell.

Following similar arguments, we can prove that LSI $\Rightarrow$ sub-Gaussian concentration.
(and therefore PI $\nRightarrow$ LSI)

Stroock-Holley perturbation principle

Theorem. If π satisfies LSI/PI with constant $C_{\mathrm{LSI}}/C_{\mathrm{PI}}$. For $h(\mathbf{x})$ satisfying $\sup_{\mathbf{x}} \|h(\mathbf{x})\| \leq B$, Let distribution μ satisfies

$$
\mu(\mathbf{x}) = \frac{\pi(\mathbf{x})\exp(h(\mathbf{x}))}{\int \pi(\mathbf{s})\exp(h(\mathbf{s}))\mathrm{d}\mathbf{s}}
$$

Then μ satisfies LSI/PI with constant $e^{2B}C_{\mathrm{LSI}}/e^{2B}C_{\mathrm{PI}}$.

Proof. Note that $e^{-B} \leq \frac{\mu(\mathbf{x})}{\pi(\mathbf{x})} \leq e^B$
For PI,

$$
\begin{aligned}
\mathrm{Var}_\mu(f(\mathbf{x})) &= \inf_{m\in\mathbb{R}} \mathbb{E}_\mu[|f(\mathbf{x}) - m|^2] \\
&\leq \inf_{m\in\mathbb{R}} e^B \mathbb{E}_\pi[|f(\mathbf{x}) - m|^2] \\
&\leq e^B C_{\mathrm{PI}} \mathbb{E}_\pi[\|\nabla f\|^2] \\
&\leq e^{2B} C_{\mathrm{PI}} \mathbb{E}_\mu[\|\nabla f\|^2]
\end{aligned}
$$

For LSI,

$$
\begin{aligned}
\mathbb{E}_\mu[f\log f]-\mathbb{E}_\mu[f]\mathbb{E}_\mu[\log f] &= \inf_{t\geq 0}\mathbb{E}_\mu\left[f\log\frac{f}{t}-f+t\right] \\
&\leq \inf_{t\geq 0} e^B\mathbb{E}_\pi\left[f\log\frac{f}{t}-f+t\right] \\
&\leq 4e^B C_{\mathrm{LSI}}\mathbb{E}_\pi[\|\nabla f\|^2] \\
&\leq 4e^{2B} C_{\mathrm{LSI}}\mathbb{E}_\mu[\|\nabla f\|^2]
\end{aligned}
$$

e.g. For non-convex, L-smooth function f, $|\nabla f(x)-\nabla f(y)|\leq L|x-y|$, for $\forall x,y$. Let $x^\star$ be a stationary point.
Suppose that function is strongly convex ($\nabla^2 f\succeq \lambda\mathbb{I}_d$) for $x\notin\mathbb{B}(x^\star,R)$, only smooth for $x\in\mathbb{B}(x^\star,R)$.
We can construct function $\bar f$ satisfies $\begin{cases}\text{equals to } f, & \text{, outside } \mathbb{B}(x^\star,R)\\ |\bar f - f|\leq LR^2 & \text{, inside } \mathbb{B}(x^\star,R)\end{cases}$
For $\pi\propto\exp(-f)$, $C_{\mathrm{LSI}}(\pi)\leq C_{\mathrm{LSI}}(\exp(-f))\cdot\exp(2LR^2)\leq\frac{\exp(2LR^2)}{\lambda}$

Remark: This bound is near-tight in worst case.
e.g. For double-well potential function f, $\pi\propto\exp(\frac{-f}{\epsilon})$.
Its (rescaled) Langevin diffusion $\mathrm{d}\mathbf{X}_t=-\nabla f(\mathbf{X}_t)\mathrm{d}t+\sqrt{2\epsilon}\mathrm{d}\mathbf{B}_t$
Escape time $\tau\approx\exp(\frac{h}{\epsilon})$, where h represents energy barrier height.

Poincare constant under non-convexity

Theorem.(Bakry et. al.) PI holds true under:
(a). $\langle\mathbf{x},\nabla f(\mathbf{x})\rangle\geq\alpha\|\mathbf{x}\|$, for $x\notin\mathbb{B}(0,R)$
(b). $a\|\nabla f(\mathbf{x})\|^2-\Delta f(\mathbf{x})\geq C$, for $x\notin\mathbb{B}(0,R)$

Proof.
(i). Within $\mathbb{B}(0,R)$, PI is satisfied. (Use Stroock-Holley perturbation principle)

$$
\mu(\mathbf{x})=\frac{\pi(\mathbf{x})\mathbb{1}_{\mathbb{B}(0,R)}(\mathbf{x})}{\int\pi(\mathbf{s})\mathbb{1}_{\mathbb{B}(0,R)}(\mathbf{s})\mathrm{d}\mathbf{s}}
$$

(ii). If Lyapunov function W satisfies $\mathcal{A}W(\mathbf{x})\leq-\theta W(\mathbf{x})+b\mathbb{1}_{\mathbb{B}(0,R)}(\mathbf{x})$, then PI is satisfied. ($\theta<0$, $b<0$).
(iii). Verify the condition, $W(\mathbf{x})=\exp(\gamma\|\mathbf{x}\|_2)$ for γ sufficiently small.

Proof for (ii). for function f satisfying $\mathbb{E}_\pi[f]=0$

$$
\int f^2(\mathbf{x})\mathrm{d}\pi(\mathbf{x})\leq\underbrace{\int\frac{-\mathcal{A}W}{\theta W}f^2\mathrm{d}\pi(\mathbf{x})}_{\leq\int\frac{|\nabla f|^2}{\theta}\mathrm{d}\pi(\mathbf{x})}+\underbrace{\int\frac{b}{\theta W}\mathbb{1}_{\mathbb{B}(0,R)}(\mathbf{x})f^2\mathrm{d}\pi(\mathbf{x})}_{\text{PI within }\mathbb{B}(0,R)}
$$

KLS conjecture. Assume $-\log\pi$ convex, $\mathbb{E}_\pi[\mathbf{X}]=0$, $\mathbb{E}_\pi[\mathbf{X}\mathbf{X}^\top]=\mathbb{I}_d$, then $C_{\mathrm{PI}}(\pi)$ is dim-free.
Chen(2021). $\forall\epsilon>0$, $C_{\mathrm{PI}}(\pi)$ grows slower that $O(d^\epsilon)$

Applications to Non-Convex Optimization

What does stationary distribution give us?

Consider a diffusion process $\mathrm{d}\mathbf{X}_t = -\nabla f(\mathbf{X}_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}\mathbf{B}_t$, $\beta > 0$. Its stationary distribution $\pi(\mathbf{x}) \propto \exp(-\beta f(\mathbf{x}))$
Goal: To analyze $\mathbb{E}_\pi[f(\mathbf{X})] - f_{\min} \leq ?$

$$
\begin{aligned}
\int \pi(\mathbf{x}) f(\mathbf{x})\mathrm{d}\mathbf{x} &= \int \pi(\mathbf{x}) \left(-\frac{1}{\beta}\log\pi(\mathbf{x}) - \frac{1}{\beta}\mathbf{z}_\beta\right)\mathrm{d}\mathbf{x} \\
&= \frac{1}{\beta} \underbrace{\int \pi(\mathbf{x}) \log\frac{1}{\pi(\mathbf{x})}\mathrm{d}\mathbf{x}}_{\text{differential entropy} \leq \text{D.E. of Gaussian}} - \frac{1}{\beta}\log\mathbf{z}_\beta \\
&\leq \frac{1}{\beta}\frac{d}{2}\log(\mathbb{E}_\pi[\|\mathbf{X}\|^2]) - \frac{1}{\beta}\log\int\exp(-\beta f(\mathbf{z}))\mathrm{d}\mathbf{z} \\
\text{Second term} &= f_{\min} - \frac{1}{\beta}\log\int\exp(-\beta(f(\mathbf{z}) - f_{\min}))\mathrm{d}\mathbf{z} \\
&\leq f_{\min} - \frac{1}{\beta}\log\int \underbrace{\exp\left(-\frac{\beta L\|\mathbf{x}-\mathbf{x}^\star\|^2}{2}\right)}_{\text{Gaussian p.d.f.}}\mathrm{d}\mathbf{x} \\
&\leq f_{\min} + \frac{d}{\beta}\log(\cdots)
\end{aligned}
$$

Finally,

$$
\mathbb{E}_\pi[f(\mathbf{X})] - f_{\min} \leq \frac{d}{\beta}\log(\cdots)
$$

Noise helps!

Lecture 4. Discretization of diffusion process, application to MCMC and stochastic optimization

Discretization

Forward-Euler method. Take Langevin diffusion $\mathrm{d}\mathbf{X}_t = -\nabla f(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$ with step size η.
Then, $\tilde{\mathbf{X}}_{(k+1)\eta} = \tilde{\mathbf{X}}_{k\eta} - \eta\nabla f(\tilde{\mathbf{X}}_{k\eta}) + \sqrt{2\eta}\xi_k$, where $(\xi_0, \xi_1, \xi_2, \cdots) \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathbb{I}_d)$
Define a stochastic process $\mathrm{d}\tilde{\mathbf{X}}_t = -\nabla f(\tilde{\mathbf{X}}_{k\eta})\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$, for $t \in [\tilde{\mathbf{X}}_{k\eta}, \tilde{\mathbf{X}}_{(k+1)\eta})$
Synchronous coupling: Drive $(\mathbf{X}_t)_{t\geq 0}$ and $(\tilde{\mathbf{X}}_t)_{t\geq 0}$ using the same BM.
Then, we get $\mathrm{d}(\mathbf{X}_t - \tilde{\mathbf{X}}_t) = -(\nabla f(\tilde{\mathbf{X}}_{k\eta}) - \nabla f(\mathbf{X}_t))\mathrm{d}t$

Note

In ML applications. Dependence on:
(a). Total time
(b). Problem dimension

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\|\tilde{\mathbf{X}}_t - \mathbf{X}_t\|_2^2] &= -2\mathbb{E}[\langle \tilde{\mathbf{X}}_t - \mathbf{X}_t, \nabla f(\tilde{\mathbf{X}}_{k\eta}) - \nabla f(\mathbf{X}_t)\rangle] \\ &\leq -2\mathbb{E}[\langle \tilde{\mathbf{X}}_t - \mathbf{X}_t, \nabla f(\tilde{\mathbf{X}}_t) - \nabla f(\mathbf{X}_t)\rangle] - 2\mathbb{E}[\langle \tilde{\mathbf{X}}_t - \mathbf{X}_t, \nabla f(\tilde{\mathbf{X}}_{k\eta}) - \nabla f(\tilde{\mathbf{X}}_t)\rangle] \\ &\leq -2\lambda\mathbb{E}[\|\tilde{\mathbf{X}}_t - \mathbf{X}_t\|_2^2] + 2\sqrt{\mathbb{E}[\|\tilde{\mathbf{X}}_t - \mathbf{X}_t\|_2^2]} \cdot \sqrt{\mathbb{E}[\|\nabla f(\tilde{\mathbf{X}}_{k\eta}) - \nabla f(\tilde{\mathbf{X}}_t)\|_2^2]}\end{aligned}$$

Under smoothness,

$$\begin{aligned}\mathbb{E}[\|\nabla f(\tilde{\mathbf{X}}_{k\eta}) - \nabla f(\tilde{\mathbf{X}}_t)\|_2^2] &\leq L^2\mathbb{E}[\|\tilde{\mathbf{X}}_{k\eta} - \tilde{\mathbf{X}}_t\|_2^2] \\ \tilde{\mathbf{X}}_t - \tilde{\mathbf{X}}_{k\eta} &= \underbrace{\int_{k\eta}^t \nabla f(\tilde{\mathbf{X}}_s)\mathrm{d}s}_{O(\eta\sqrt{d})} + \underbrace{\sqrt{2}\int_{k\eta}^t \sqrt{2}\mathrm{d}\mathbf{B}_s}_{O(\sqrt{\eta d})}\end{aligned}$$

Due to,

$$\begin{aligned}\mathbb{E}[\|\text{First term}\|_2^2] &\leq (t-k\eta)\int_{k\eta}^t \mathbb{E}[\|\nabla f(\tilde{\mathbf{X}}_s)\|_2^2\mathrm{d}s] \\ &\leq (t-k\eta)\int_{k\eta}^t L^2\mathbb{E}[\|\tilde{\mathbf{X}}_s - \mathbf{X}^\star\|_2^2]\mathrm{d}s \\ &= O(\eta^2 d) \\ \mathbb{E}[\|\text{Second term}\|_2^2] &= 2d(t-k\eta) \\ &= O(\eta d)\end{aligned}$$

Substituting back, $\Phi_t = \mathbb{E}[\|\tilde{\mathbf{X}}_t - \mathbf{X}_t\|^2]$

$$\begin{aligned}\frac{\mathrm{d}\Phi_t}{\mathrm{d}t} &\leq -\lambda\Phi_t + \sqrt{\Phi_t \cdot C \cdot d\eta} \\ &\leq -\lambda\Phi_t + \frac{\lambda}{2}\Phi_t + \frac{2C}{\lambda}d\eta \\ &= -\frac{\lambda}{2}\Phi_t + C'\eta d\end{aligned}$$

Solve for recursion,

$$\Phi_t \leq \exp\left(-\frac{\lambda t}{2}\right)\Phi_0 + \frac{2C'}{\lambda}\eta d \leq \epsilon^2$$

We need $\eta \leq \Theta(\frac{\epsilon^2}{d})$, $t \geq \frac{4}{\lambda}\log(\frac{1}{\epsilon})$. Number of steps: $O(\frac{d}{\epsilon^2}\log(\frac{1}{\epsilon}))$

Theorem. Under above parameter setup, we have

$$\mathcal{W}_2(\tilde{\pi}_n, \pi^\star) \le \epsilon$$

Remark.

- We did not optimize dependence or condition number $\kappa = \frac{l}{\lambda}$. (may be improved)
- "Unadjusted Langevin Algorithm"

Faster convergence possible under Metropolis adjustment.
Brief description of Metropolis-Hastings:

- ("Reversible") Fact, if an MC satisfies $p(y|x)\pi(x) = p(x|y)\pi(y)$. Then π is a stationary distribution.
- MH: Accept or reject proposal.
  Let p be the proposal transition probability. For each step, generate $Y_k \sim p(\cdot|X_k)$.
  Then, accept Y_k to proceed with probability $min(1, \frac{p(X_k|Y_k)\pi(Y_k)}{p(Y_k|X_k)\pi(X_k)})$.
  We get adjusted Markov Chain, which satisfies

$$q(y|x)\pi(x) = q(x|y)\pi(y)$$

  $\Rightarrow$ MALA(Metropoli-adjusted Langevin Algorithm)
- With $O(d\log(\frac{1}{\epsilon}))$ iterations, we get $d_{\mathrm{TV}}(\tilde{\pi}_n^{\mathrm{MALA}}, \pi^\star) \le \epsilon$. (may be sensitive to η e.t.c.)

Underdamped Langevin

$$\begin{cases} \mathrm{d}\mathbf{v}_t = -\gamma \mathbf{v}_t \mathrm{d}t - u\nabla f(\mathbf{x}_t)\mathrm{d}t + \sqrt{2\gamma u}\mathrm{d}\mathbf{B}_t \\ \mathrm{d}\mathbf{x}_t = \mathbf{v}_t \mathrm{d}t \end{cases}$$

Discretization, for $t \in [k\eta, (k+1)\eta)$

$$\begin{cases} \mathrm{d}\tilde{\mathbf{v}}_t = -\gamma \tilde{\mathbf{v}}_t \mathrm{d}t - u\nabla f(\tilde{\mathbf{x}}_{k\eta})\mathrm{d}t + \sqrt{2\gamma u}\mathrm{d}\mathbf{B}_t \\ \mathrm{d}\tilde{\mathbf{x}}_t = \tilde{\mathbf{v}}_t \mathrm{d}t \end{cases}$$

Ornstein-Uhlenbeck process:
Given $(\tilde{\mathbf{x}}_{k\eta}, \tilde{\mathbf{v}}_{k\eta})$, the conditional distribution of $(\tilde{\mathbf{x}}_{k\eta}, \tilde{\mathbf{v}}_{k\eta})$ is normal, with closed-form formulae for mean and variance.
Use this at update algo.

Let Lyapunov function

$$\mathbf{\Phi}_t = \mathbb{E}\left[\left\| \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{x}}_t - \mathbf{x}_t \\ \tilde{\mathbf{v}}_t - \mathbf{v}_t \end{pmatrix} \right\|_2^2\right]$$

Let $\mathbf{z}_t = \tilde{\mathbf{x}}_t - \mathbf{x}_t$, $\mathbf{y}_t = \tilde{\mathbf{v}}_t - \mathbf{v}_t$

$$
\begin{aligned}
\frac{\mathrm{d}\Phi_t}{\mathrm{d}t} &= 2\mathbb{E}\left[\begin{pmatrix}\mathbf{z}_t+\mathbf{y}_t & \mathbf{y}_t\end{pmatrix}\begin{pmatrix}(1-\gamma)\mathbf{y}_t - u\left(\nabla f(\tilde{\mathbf{x}}_{k\eta})-\nabla f(\mathbf{x}_t)\right)\\ -\gamma\mathbf{y}_t - u\left(\nabla f(\tilde{\mathbf{x}}_{k\eta})-\nabla f(\mathbf{x}_t)\right)\end{pmatrix}\right]\\
&= -(\text{something nice}) + u\mathbb{E}\left[\begin{pmatrix}\mathbf{z}_t+\mathbf{y}_t & \mathbf{y}_t\end{pmatrix}\begin{pmatrix}-\left(\nabla f(\tilde{\mathbf{x}}_{k\eta})-\nabla f(\tilde{\mathbf{x}}_t)\right)\\ -\left(\nabla f(\tilde{\mathbf{x}}_{k\eta})-\nabla f(\tilde{\mathbf{x}}_t)\right)\end{pmatrix}\right]\\
&\le -\frac{\lambda}{2L}\Phi_t + u\sqrt{\Phi_t\cdot\mathbb{E}[\|\nabla f(\tilde{\mathbf{x}}_{k\eta})-\nabla f(\tilde{\mathbf{x}}_t)\|^2]}
\end{aligned}
$$

Due to

$$
\begin{aligned}
\mathbb{E}[\|\nabla f(\tilde{\mathbf{x}}_{k\eta})-\nabla f(\tilde{\mathbf{x}}_t)\|^2] &\le L^2\mathbb{E}[\|\tilde{\mathbf{x}}_{k\eta}-\tilde{\mathbf{x}}_t\|^2]\\
\tilde{\mathbf{x}}_t &= \tilde{\mathbf{x}}_{k\eta} + \int_{\eta k}^{t}\tilde{\mathbf{v}}_t\mathrm{d}t
\end{aligned}
$$

So

$$
\mathbb{E}[\|\tilde{\mathbf{x}}_{k\eta}-\tilde{\mathbf{x}}_t\|^2] \le \eta^2\max_{k\eta\le t\le(k+1)\eta}\mathbb{E}[\|\tilde{\mathbf{v}}_t\|^2]
$$

Similar to Langevin Algorithm, $\mathbb{E}[\|\tilde{\mathbf{v}}_t\|^2]\le O(d)$
Finally, we get

$$
\Phi_t \le \exp\left(-\frac{\lambda}{2L}t\right)\Phi_0 + O(\eta^2 d)
$$

If we want $\mathcal{W}_2(\tilde{\pi}_t,\pi^\star)\le\epsilon$, we need $\eta\le O(\frac{\epsilon}{\sqrt{d}})$, $t\ge\Omega(\frac{L}{\lambda}\log(\frac{1}{\epsilon}))$
Iteration complexity $n\asymp\frac{\sqrt{d}}{\epsilon}\log(\frac{1}{\epsilon})$

Remark. some further improvements

- (Shen & Lee, 2019) Randomized midpoint, $O(d^{\frac{1}{3}})$
- (Mou et. al., 2019) High-order Langevin (Using integration orade, $O(d^{\frac{1}{4}})$
- (Song, Lee & Venpala, 2019) With "nice data" (Specialized to Bayesian logistic regression), $O(\mathrm{polylog}(d))$
- (Zhang, 2015) Randomized midpoint, $O(d^{\frac{1}{4}})$
  Still open: $\mathrm{polylog}(1/\epsilon)$, sub-linear in d for cold start.

How about non-convex?
For diffusion process

$$
\begin{aligned}
\mathrm{d}\mathbf{x}_t &= -\nabla f(\mathbf{x}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t\\
\mathrm{d}\tilde{\mathbf{x}}_t &= -\nabla f(\tilde{\mathbf{x}}_{k\eta})\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t
\end{aligned}
$$

Idea 1: Still use synchronous coupling to estimate

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\|^2] &\le -2\mathbb{E}[\langle \mathbf{x}_t - \tilde{\mathbf{x}}_t, \nabla f(\mathbf{x}_t) - \nabla f(\tilde{\mathbf{x}}_{k\eta})\rangle] \\
&\le 2L \cdot \mathbb{E}[\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\| \cdot \|\mathbf{x}_t - \tilde{\mathbf{x}}_{k\eta}\|] \\
&\le 2L \cdot \mathbb{E}[\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2] + 2L\mathbb{E}[\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|\|\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{k\eta}\|] \\
&\le 2L \cdot \mathbb{E}[\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2] + O(\eta d)
\end{aligned}
$$

Use Gronwall inequality,

$$
\mathbb{E}[\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\|^2] \le O(\eta d) \cdot \exp(2Lt)
$$

This is unavailable for ODEs, but the noise may help!

Solution 1. pathwise KL divergence

Radon-Nikodym derivative. Let $\mathbb{P}^{(1)}$, $\mathbb{P}^{(2)}$ be two probability distribution on $\mathbf{X}$. $\frac{\mathrm{d}\mathbb{P}^{(1)}}{\mathrm{d}\mathbb{P}^{(2)}}$ as a random variable, we have

$$
\mathbb{E}\left[f(\mathbf{X}^{(2)})\frac{\mathrm{d}\mathbb{P}^{(1)}}{\mathrm{d}\mathbb{P}^{(2)}}\right] = \mathbb{E}[f(\mathbf{X}^{(1)})]
$$

Underlying idea: $p^{(1)}$, $p^{(2)}$ are densities,

$$
\frac{\mathrm{d}\mathbb{P}^{(1)}}{\mathrm{d}\mathbb{P}^{(2)}} = \frac{p^{(1)}}{p^{(2)}}(\mathbf{X}^{(2)})
$$

where $X^{(2)} \sim \mathbb{P}^{(2)}$.

$$
\mathcal{D}_{\mathrm{KL}}(\mathbb{P}^{(1)}\|\mathbb{P}^{(2)}) = -\mathbb{E}\left[\log\frac{\mathrm{d}\mathbb{P}^{(2)}}{\mathrm{d}\mathbb{P}^{(1)}}\right]
$$

Definition.

$\tilde{\mathbb{P}}_{[0,T]}$: Law of $(\tilde{\mathbf{X}}_t : t \in [0,T])$

$\mathbb{P}_{[0,T]}$: Law of $(\mathbf{X}_t : t \in [0,T])$

Analogy: discrete-time version.

Let

$$
\begin{aligned}
\mathbf{X}_{t+1}^{(1)} &= \mathbf{h}^{(1)}(\mathbf{X}_1^{(1)}, \cdots, \mathbf{X}_t^{(1)}) + \xi_{t+1} \\
\mathbf{X}_{t+1}^{(2)} &= \mathbf{h}^{(2)}(\mathbf{X}_1^{(2)}, \cdots, \mathbf{X}_t^{(2)}) + \xi_{t+1}
\end{aligned}
$$

where $\xi_t \sim \mathcal{N}(0, \mathbb{I}_d)$.

$$
\begin{aligned}
P_{\mathbf{X}_{t+1}^{(1)}}(\mathbf{X}|\mathbf{X}_1^{(1)}, \cdots, \mathbf{X}_t^{(1)}) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\|\mathbf{X} - \mathbf{h}^{(1)}(\mathbf{X}_1^{(1)}, \cdots, \mathbf{X}_t^{(1)})\|^2\right) \\
P_{\mathbf{X}_{t+1}^{(2)}}(\mathbf{X}|\mathbf{X}_1^{(2)}, \cdots, \mathbf{X}_t^{(2)}) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\|\mathbf{X} - \mathbf{h}^{(1)}(\mathbf{X}_1^{(2)}, \cdots, \mathbf{X}_t^{(2)})\|^2\right)
\end{aligned}
$$

Then,

$$\frac{P_{\mathbf{X}_{t+1}^{(2)}}}{P_{\mathbf{X}_{t+1}^{(1)}}} = \exp\left(\langle \mathbf{X}, \mathbf{h}^{(2)} - \mathbf{h}^{(1)}\rangle - \frac{1}{2}\|\mathbf{h}^{(2)}\|^2 + \frac{1}{2}\|\mathbf{h}^{(1)}\|^2\right)$$
$$= \exp\left(\langle \mathbf{X} - \mathbf{h}^{(1)}, \mathbf{h}^{(2)} - \mathbf{h}^{(1)}\rangle - \frac{1}{2}\|\mathbf{h}^{(1)} - \mathbf{h}^{(2)}\|^2\right)$$

Finally,

$$\frac{\mathrm{d}\mathbb{P}^{(2)}}{\mathrm{d}\mathbb{P}^{(1)}} = \exp(\sum_{i=1}^{n}\langle \xi_i, \mathbf{h}^{(2)}(\mathbf{X}_1^{(1)}, \cdots, \mathbf{X}_{i-1}^{(1)}) - \mathbf{h}^{(1)}(\mathbf{X}_1^{(1)}, \cdots, \mathbf{X}_{i-1}^{(1)})\rangle$$
$$-\frac{1}{2}\|\mathbf{h}^{(1)}(\mathbf{X}_1^{(1)}, \cdots, \mathbf{X}_{i-1}^{(1)}) - \mathbf{h}^{(2)}(\mathbf{X}_1^{(1)}, \cdots, \mathbf{X}_{i-1}^{(1)})\|^2)$$

Theorem. (Girsanov) Suppose $\mathrm{d}\mathbf{X}_t^{(i)} = \mathbf{g}^{(i)}(\mathbf{X}_{[0,T]}^{(i)})\mathrm{d}t + \mathrm{d}\mathbf{B}_t$, $i \in \{1, 2\}$
Calculate,

$$\frac{\mathrm{d}\mathbb{P}_{[0,T]}^{(1)}}{\mathrm{d}\mathbb{P}_{[0,T]}^{(2)}} = \exp\left(\int_0^T (\mathbf{g}^{(1)}(\mathbf{X}_{[0,T]}^{(2)}) - \mathbf{g}^{(2)}(\mathbf{X}_{[0,T]}^{(2)}))^\top \mathrm{d}\mathbf{B}_t - \frac{1}{2}\int_0^T \|\mathbf{g}^{(1)}(\mathbf{X}_{[0,T]}^{(2)}) - \mathbf{g}^{(2)}(\mathbf{X}_{[0,T]}^{(2)})\|^2 \mathrm{d}t\right)$$

Then, (under weak assumptions)

$$\mathcal{D}_{\mathrm{KL}}(\tilde{\mathbb{P}}_{[0,T]}\|\mathbb{P}_{[0,T]}) = -\mathbb{E}\left[\log\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_2}\right] - \int_0^T \mathbb{E}[\|\nabla f(\mathbf{X}_{k\eta}) - \nabla f(\mathbf{X}_t)\|^2]\mathrm{d}t$$
$$\leq L^2 \int_0^T \underbrace{\mathbb{E}[\|\mathbf{X}_t - \mathbf{X}_{k\eta}\|^2]}_{\leq O(d\eta)}\mathrm{d}t$$

$$\mathcal{D}_{\mathrm{KL}}(\hat{\pi}_T\|\pi_T) \leq \mathcal{D}_{\mathrm{KL}}(\tilde{\mathbb{P}}_{[0,T]}\|\mathbb{P}_{[0,T]}) \leq O(T\eta d)$$

Pinsker+triangle inequality,

$$d_{\mathrm{TV}}(\hat{\pi}_T, \pi^\star) \leq \exp\left(-\frac{T}{2C_{\mathrm{LSI}}}\right) + O(\sqrt{\eta T d})$$

For ϵ-close in TV, we need $\eta \lesssim \frac{\epsilon^2}{Td}$, $T \gtrsim C_{\mathrm{LSI}}\log(\frac{1}{\epsilon})$, Its complexity $n \asymp \frac{C_{\mathrm{LSI}}^2 d}{\epsilon^2}\log(\frac{1}{\epsilon})$
We get $O(\sqrt{\eta})$ numerical order, not tight! (compared to Euler method in ODE)

Fokker-Planck equation method. (Mou et. al., 2020)

$$\partial_t \pi_t = \nabla\cdot(\pi_t \nabla f) + \Delta\pi_t$$
$$\partial_t \tilde{\pi}_t = \nabla\cdot(\tilde{\pi}_t \mathbb{E}[\nabla f(\tilde{\mathbf{x}}_{k\eta})|\tilde{\mathbf{x}}_t = \mathbf{x}]) + \Delta\hat{\pi}_t$$

Lecture 5. Advanced topics in discretization and sampling, Score-based diffusion generative models

From last lecture, we have

$$\mathcal{D}_{\mathrm{KL}}(\hat{\mathbb{P}}_{[0,T]}\|\mathbb{P}_{[0,T]}) \leq O(\eta d T)$$

Numerical order not tight enough.

Idea: compare $\hat{\pi}_t$ and π_t
By Fokker-Planck equation:

$$\frac{\partial \pi_t}{\partial t} = \nabla \cdot (\pi_t \nabla f) + \Delta \pi_t$$

And we have,

$$\mathrm{d}\hat{\mathbf{x}}_t = -\nabla f(\hat{\mathbf{x}}_{k\eta})\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$$

Then,

$$\frac{\partial \hat{\pi}_t|_{k\eta}}{\partial t} = \nabla \cdot (\hat{\pi}_t|_{k\eta} \cdot \underbrace{\nabla f(\hat{\mathbf{x}}_{k\eta})}_{\substack{\text{seen as deterministic} \\ \text{since we condition on } \mathcal{F}_{k\eta}}}) + \Delta \hat{\pi}_t|_{k\eta}$$

From $\hat{\pi}_t|_{k\eta}$ to $\hat{\pi}_t$: integrate out $\hat{\mathbf{x}}_{k\eta}$,

$$\frac{\partial \hat{\pi}_t(\mathbf{x})}{\partial t} = \nabla \cdot (\hat{\pi}_t(\mathbf{x}) \underbrace{\mathbb{E}\left[\nabla f(\hat{\mathbf{x}}_{k\eta}) | \hat{\mathbf{x}}_t = \mathbf{x}\right]}_{=:\hat{\mathbf{b}}_t(\mathbf{x})}) + \Delta \hat{\pi}_t$$

Let $\mathbf{b}(\mathbf{x}) := \nabla f(\mathbf{x})$ for simplicity,

$$\partial_t \hat{\pi}_t = \nabla \cdot (\hat{\pi}_t \hat{\mathbf{b}}_t) + \Delta \hat{\pi}_t$$
$$\partial_t \pi_t = \nabla \cdot (\pi_t \mathbf{b}) + \Delta \pi_t$$

Then, we calculate,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{D}_{\mathrm{KL}}(\hat{\pi}_t \| \pi_t) &= \int \frac{\partial}{\partial t} \left(\hat{\pi}_t \log \frac{\hat{\pi}_t}{\pi_t} \right) \mathrm{d}\mathbf{x} \\
&= \int (\nabla \cdot (\hat{\pi}_t \hat{\mathbf{b}}_t) + \Delta \hat{\pi}_t) \log \frac{\hat{\pi}_t}{\pi_t} \mathrm{d}\mathbf{x} + \int \pi_t \frac{\partial}{\partial t} \left(\frac{\hat{\pi}_t}{\pi_t} \right) \mathrm{d}\mathbf{x} \\
&= \int (\nabla \cdot (\hat{\pi}_t \hat{\mathbf{b}}_t) + \Delta \hat{\pi}_t) \log \frac{\hat{\pi}_t}{\pi_t} \mathrm{d}\mathbf{x} + \int \pi_t \frac{\partial_t \hat{\pi}_t \pi_t - \hat{\pi}_t \partial_t \pi_t}{\pi_t^2} \mathrm{d}\mathbf{x} \\
&= \int \langle \mathbf{b} - \hat{\mathbf{b}}_t, \nabla \log \frac{\hat{\pi}_t}{\pi_t} \rangle \hat{\pi}_t \mathrm{d}\mathbf{x} - \int \|\nabla \log \frac{\hat{\pi}_t}{\pi_t}\|^2 \hat{\pi}_t \mathrm{d}\mathbf{x} \\
&\leq \int \|\mathbf{b}_t(\mathbf{x}) - \hat{\mathbf{b}}_t(\mathbf{x})\|^2 \hat{\pi}_t(\mathbf{x}) \mathrm{d}\mathbf{x}
\end{aligned}$$

Previously, we have, $\mathbb{E}[\|\mathbf{b}(\hat{\mathbf{x}}_t) - \hat{\mathbf{b}}(\hat{\mathbf{x}}_{k\eta})\|^2]$
Now we get,

$$\mathbb{E}[\|\mathbf{b}(\hat{\mathbf{x}}_t) - \mathbb{E}[\hat{\mathbf{b}}(\hat{\mathbf{x}}_{k\eta}) | \hat{\mathbf{x}}_t]\|^2]$$

Is it better?

Conceptual example: two extremes

- If $\hat{\mathbf{x}}_{k\eta}$ is deterministic, $\mathbb{E}[\|\mathbf{b} - \hat{\mathbf{b}}_t\|^2] = \Theta(\eta)$.
- If the density of $\hat{\mathbf{x}}_{k\eta}$ is "uniform", $\mathbb{E}[\|\mathbf{b} - \hat{\mathbf{b}}_t\|^2] = \Theta(\eta) \leq O(\eta^2)$ in the "best case".

Analysis based on this intuition:
Step 1. linear approximate to b.

$$\mathbf{b}(\hat{\mathbf{x}}_{k\eta}) = \mathbf{b}(\hat{\mathbf{x}}_t) + \nabla \cdot \mathbf{b}(\hat{\mathbf{x}}_t)(\hat{\mathbf{x}}_{k\eta} - \hat{\mathbf{x}}_t) + O(\eta^2)$$

Only need to analyze (Condition on $\hat{\mathbf{x}}_t = \mathbf{x}$)

$$\mathbb{E}[\hat{\mathbf{x}}_{k\eta}|\hat{\mathbf{x}}_t] - \hat{\mathbf{x}}_t = \int \frac{\hat{\pi}_{k\eta}(\mathbf{y}) \cdot \hat{p}_{t|k\eta}(\mathbf{x}|\mathbf{y})}{\hat{\pi}_t(\mathbf{x})}(\mathbf{y} - \mathbf{x})\mathrm{d}\mathbf{y}$$

where $\hat{p}_{t|k\eta} \sim \mathcal{N}(\mathbf{y} - \eta b(\mathbf{y}), 2\eta\mathbb{I}_d)$ density evaluated at $\mathbf{x}$.
Then,

$$\mathbb{E}[\hat{\mathbf{x}}_{k\eta}|\hat{\mathbf{x}}_t] - \hat{\mathbf{x}}_t \approx \int \frac{\hat{\pi}_{k\eta}(\mathbf{y}) \cdot \nabla_y \hat{p}_{t|k\eta}(\mathbf{x}|\mathbf{y})}{\hat{\pi}_t(\mathbf{x})}\mathrm{d}\mathbf{y}$$

Then,

$$\mathbb{E}[\|\mathbb{E}[\hat{\mathbf{x}}_{k\eta}|\hat{\mathbf{x}}_t] - \hat{\mathbf{x}}_t\|^2] \leq \eta^2 \int \hat{\pi}_{k\eta}\|\nabla \log \hat{\pi}_{k\eta}\|^2 \mathrm{d}\mathbf{x} + O(\eta^2)$$

Intuition:

$$\hat{\pi}_{k\eta} \xrightarrow{\mathbf{x} \mapsto \mathbf{x} - \eta \mathbf{b}(\mathbf{x})} \text{increase FI} \xrightarrow{\text{add } \mathcal{N}(0, \eta\mathbb{I}_d)} \text{decrease FI} \to \hat{\pi}_{(k+1)\eta}$$

Then we conclude,

$$\mathcal{D}_{\mathrm{KL}}(\hat{\pi}_t\|\pi_t) \leq O(\eta^2 T)$$

Wang and Li (2024). Extension to general diffusion

Score-based generative models

Goal: Learn $p^\star$ from $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \sim p^\star$

Idea 1: Learn $\hat{\mathbf{s}}(\mathbf{x}) \approx \nabla \log p^\star(\mathbf{x})$. Run diffusion process $\mathrm{d}\hat{\mathbf{x}}_t = -\hat{\mathbf{s}}(\mathbf{x}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$ (and its numerical simulation)
Problem: How to learn $\nabla \log p^\star(\mathbf{x})$?
In regression problems, we learn $f^\star$ from $\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i) + \epsilon_i$, $\mathbb{E}[\mathbf{y}|\mathbf{x}] = \mathbf{f}(\mathbf{x})$, $\text{loss} = \mathbb{E}[\|\mathbf{Y} - \mathbf{f}(\mathbf{X})\|^2]$.
Empirically, $\text{loss} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{f}(\mathbf{x}_i))^2$
In score-based generative models, if we define

$$\begin{aligned}\text{loss}(\mathbf{s}) &= \mathbb{E}[\|\nabla \log p^\star(\mathbf{X}) - \mathbf{s}(\mathbf{X})\|^2] \\ &= \underbrace{\mathbb{E}[\|\mathbf{s}(\mathbf{X})\|^2]}_{\text{easy to compute}} - 2\mathbb{E}[\langle \mathbf{s}(\mathbf{X}), \nabla \log p^\star(\mathbf{X})\rangle] + \underbrace{\mathbb{E}[\|\nabla \log p^\star(\mathbf{X})\|^2]}_{\text{independend of } \mathbf{s}}\end{aligned}$$

For the second term,

$$\begin{aligned}2\mathbb{E}[\langle \mathbf{s}(\mathbf{X}), \nabla \log p^\star(\mathbf{X})\rangle] &= \int \mathbf{s}(\mathbf{x})^\top \nabla \log p^\star(\mathbf{x}) p^\star(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int \mathbf{s}(\mathbf{x})^\top \nabla p^\star(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= -\int (\nabla \cdot \mathbf{s}(\mathbf{x})) p^\star(\mathbf{x}) \mathrm{d}\mathbf{x}\end{aligned}$$

So we have,

$$\text{loss}(\mathbf{s}) = \mathbf{E}[\|\mathbf{s}(\mathbf{X})\|^2 + 2(\nabla \cdot \mathbf{s}(\mathbf{x}))] + \text{something independent of } \mathbf{s}$$

Then,

$$\text{empirical loss} = \frac{1}{n}\sum_{i=1}^{n}(\|\mathbf{s}(\mathbf{x}_i)\|^2 + 2\nabla \cdot \mathbf{s}(\mathbf{x}_i))$$

e.g. If we know $s^\star = \nabla \log p^\star \in \mathcal{F}$, $\mathcal{F}$ is convex. We can solve $\hat{\mathbf{s}}$ using convex optimization.

Performance guarantees?
e.g. $\mathbb{E}[\|\nabla \log p^\star(\mathbf{X}) - \mathbf{s}(\mathbf{X})\|^2] \le \epsilon_{\text{score}}^2$

$$\mathcal{D}_{\text{KL}}(\mathbb{P}^\star_{[0,T]} \| \hat{\mathbb{P}}_{[0,T]}) \le T\epsilon_{\text{score}}^2 + \text{discretization error} \le O(\eta T d)$$

e.g. $\mathbf{s}(\mathbf{x}) = \nabla \log q(\mathbf{x})$, $q \in \mathcal{Q}$

$$I(p^\star \| \hat{q}) = \mathbb{E}_{p^\star}[\|\mathbf{s}(\mathbf{X}) - \nabla \log p^\star(\mathbf{X})\|^2]$$

Learning score function of data distribution

Goal: Fit $\nabla \log p^\star(\mathbf{x})$ using function set $\mathcal{F}$

$$\mathbf{s} = \arg\min_{\mathbf{s} \in \mathcal{F}} \mathbb{E}[\|\nabla \log p^\star(\mathbf{X}) - \mathbf{s}(\mathbf{X})\|^2]$$

Empirically,

$$\mathbf{s} = \arg\min_{\mathbf{s} \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(\|\mathbf{s}(\mathbf{x}_i)\|^2 + 2\nabla \cdot \mathbf{s}(\mathbf{x}_i))$$

Suppose $\mathcal{F} = \{\nabla \log q : q \in \mathcal{Q}\}$,

$$L(\nabla \log q) = \mathbb{E}_{p^\star}[\|\nabla \log p^\star - \nabla \log q\|^2]$$

If $\forall q \in \mathcal{Q}$, q satisfies LSI (C_0), then

$$\mathcal{D}_{\mathrm{KL}}(p^\star \| q) \le C_0 L(\nabla \log q)$$

Q: How to control $L(\nabla \log q)$?

ML Theory basics

In general, we want to minimize $L(f) = \mathbb{E}[l(f; \mathbf{X})]$
Empirically, we calculate

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} l(f; \mathbf{X}_i)}_{L_n(f)}$$

How about $L(\hat{f}_n)$?

- $\forall f$, $\mathbb{E}[L_n(f)] = L(f)$,
- $\hat{f}_n$ minimizes L_n,
- $\mathbb{E}[L_n(\hat{f}_n)] \neq L(f^\star)$

Let $f^\star = \arg\min_{f \in \mathcal{F}} L(f)$,

$$\begin{aligned} L(\hat{f}_n) - L(f^\star) &= L(\hat{f}_n) - L_n(\hat{f}_n) + \underbrace{L_n(\hat{f}_n) - L_n(f^\star)}_{\le 0} + \underbrace{L_n(f^\star) - L(f^\star)}_{\substack{\text{easily obtained} \\ \text{from concentration}}} \\ &\le \sup_{f \in \mathcal{F}} |L(f) - L_n(f)| \end{aligned}$$

Need uniform convergence. Depending on complexity of $\mathcal{F}$
e.g. Suppose $|\mathcal{F}| < +\infty$

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \mathcal{F}} |L_n(f) - L(f)| \ge \epsilon \right) &\le |\mathcal{F}| \sup_{f \in \mathcal{F}} \mathbb{P}(|L_n(f) - L(f)| \ge \epsilon) \\ &\le |\mathcal{F}| \exp(-Cn\epsilon^2) \end{aligned}$$

If we want $\mathbb{P}\left(\sup_{f \in \mathcal{F}} |L_n(f) - L(f)| \ge \epsilon\right) \le \delta$, we can achieve

$$\epsilon_n = C \sqrt{\frac{\log\left(\frac{|\mathcal{F}|}{\delta}\right)}{n}}$$

e.g. For infinite $\mathcal{F}$, use discrete approximation

However, relax to $\sup_{f \in \mathcal{F}} |\cdots|$ can be overly conservative.

Using **localization**: consider $\sup$ over $\mathcal{F} \cap \mathbb{B}(f^\star, r_n)$

Application to generative models
Suppose $q^\star \in \mathcal{Q}$, any distribution in $\mathcal{Q}$ satisfies LSI (C_0), then

$$\mathcal{D}_{\mathrm{KL}}(p^\star \| \hat{q}_n) \le C_0 \cdot \mathrm{complexity}_n(\mathcal{F})$$

where $\mathcal{F} = \{\|\nabla \log q\|^2 + 2\Delta \log q : q \in \mathcal{Q}\}$

- Isoperimetry is unavoidable for this model

Lecture 6. Denoising diffusion generative models, discretization and learning of diffusion models

Denoising diffusion model

Forward process (add noise to image)
Let $\mathbf{X}_0 \sim p^\star$. Run OU process $\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$, $t \in [0, T]$.
Converging to $\mathcal{N}(0, \mathbb{I}_d)$ exponentially fast, independent of isoperimetry of $p^\star$.

We only need marginal distribution to be close.
Use Fokker-Planck equation:

$$\begin{aligned} \pi_0 &= p^\star \\ \partial_t \pi_t &= \nabla \cdot (\mathbf{X}_t \pi_t) + \Delta \pi_t \end{aligned}$$

We want to reverse in time, let $\overleftarrow{\pi}_t := \pi_{T-t}$, for $t \in [0, T]$, satisfying

$$\begin{aligned} \partial_t \overleftarrow{\pi}_t &= -\nabla \cdot (\mathbf{X} \overleftarrow{\pi}_t) - \Delta \overleftarrow{\pi}_t \\ &= \Delta \overleftarrow{\pi}_t + (-\nabla \cdot (\mathbf{X} \overleftarrow{\pi}_t) - 2\Delta \overleftarrow{\pi}_t) \\ &= \Delta \overleftarrow{\pi}_t + (-\nabla \cdot [(\mathbf{X} + 2\nabla \log \overleftarrow{\pi}_t) \overleftarrow{\pi}_t]) \end{aligned}$$

This is Fokker-Planck of another diffusion,

$$\mathrm{d}\overleftarrow{\mathbf{X}}_t = (\overleftarrow{\mathbf{X}}_t + 2\nabla \log \pi_{T-t}(\overleftarrow{\mathbf{X}}_t))\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$$

- Exactly the same marginal $\overleftarrow{\pi}_T = \pi_0 = p^\star$
- Suppose:
    - (i). We know π_{T-t}
    - (ii). $\overleftarrow{\mathbf{X}}_0 \sim \pi_T$
    - (iii). Exact simulation

Discussion.
For (i). Goal: find $\mathbf{s} \in \mathcal{F}$ to minimize

$$
\begin{aligned}
L_t(\mathbf{s}) &= \mathbb{E}[\|\nabla \log \pi_{T-t}(\overleftarrow{\mathbf{X}}_t) - \mathbf{s}_t(\overleftarrow{\mathbf{X}}_t)\|^2] \\
&= \mathbb{E}_{\pi_{T-t}}[\|\nabla \log \pi_{T-t}(\mathbf{X}) - \mathbf{s}_t(\mathbf{X})\|] \\
&= \mathbb{E}_{\pi_{T-t}}[\|\mathbf{s}_t(\mathbf{X})\|^2 + 2\nabla \cdot \mathbf{s}_t(\mathbf{X})]
\end{aligned}
$$

Let global loss function $L(\mathbf{s}) = \int_0^T w(t) L_t(\mathbf{s}) \mathrm{d}t$
Training of diffusion generative model:

- Observed data $\{X^{(i)}\}_{i=1}^n$
- For each i, randomly pick $t_i \sim w$
- Simulate $\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2} \mathrm{d}\mathbf{B}_t$ from time 0 to $T - t_i$ with $\mathbf{X}_0^{(i)} = \mathbf{X}^{(i)}$ (Can be simulated exactly). Then we get $\mathbf{X}_{t_i}^{(i)}$
- Solve $\min\limits_{\mathbf{s} \in \mathcal{F}} \frac{1}{n} \sum\limits_{i=1}^n \|\mathbf{s}(\mathbf{X}_{t_i}^{(i)})\|^2 + 2\nabla \cdot \mathbf{s}(\mathbf{X}_{t_i}^{(i)})$. (Can also use multiple time points)
  Statistical error analysis using aforementioned framework.
  e.g. For Holder-smooth $p^\star$. With careful choice of $\mathcal{F}$, we can achieve minimax (optimal) rate of convergence.

For (ii). We know that π_T is close enough to $\mathcal{N}(0, \mathbb{I}_d)$,

$$
d_{\mathrm{TV}}(\pi_T, \mathcal{N}(0, \mathbb{I}_d)) \lesssim e^{-T}
$$

For (iii). We have discussed it.

Alternative ways of investing the process.

$$
\begin{aligned}
\partial_t \overleftarrow{\pi}_t &= -\nabla \cdot (\mathbf{X} \overleftarrow{\pi}_t) - \Delta \overleftarrow{\pi}_t \\
&= -\nabla \cdot (\overleftarrow{\pi}_t (\mathbf{X} + \nabla \log \overleftarrow{\pi}_t))
\end{aligned}
$$

Corresponding flow in original space

$$
\mathrm{d}\overleftarrow{\mathbf{X}}_t = (\overleftarrow{\mathbf{X}}_t + \nabla \log \pi_{T-t}(\overleftarrow{\mathbf{X}}_t)) \mathrm{d}t
$$

We only need to simulate ODE. (Only randomly from $\overleftarrow{\mathbf{X}}_0 \sim \mathcal{N}(0, \mathbb{I}_d)$)

ODE might not have enough randomness, but can help with faster generation. (Sometimes it's easier to learn solution mapping directly.)

Examples of extension.
If we want to sample from $p^\star(\mathbf{x}|\mathbf{c})$ (e.g. $\mathbf{c}$ is text description of image)
(a). learn $\nabla \log \pi_t(\mathbf{x}|\mathbf{c})$ directly. ($\mathbf{s}_t(\mathbf{x}; \mathbf{c})$)
(b). Classifier guidance (w is training parameter)

$$
\nabla_{\mathbf{x}} \log \pi_t(\mathbf{x}|\mathbf{c}) = \nabla_{\mathbf{x}} \log \pi_t(\mathbf{x}) + (w) \nabla_{\mathbf{x}} \log p(\mathbf{c}|\mathbf{X} = \mathbf{x})
$$

(c). Classifier-free guidance

$$\nabla_{\mathbf{x}} \log \pi_t(\mathbf{x}) + w(\nabla_{\mathbf{x}} \log \pi_t(\mathbf{x}|\mathbf{c}) - \nabla_{\mathbf{x}} \log \pi_t(\mathbf{x}))$$

e.g. from 2D to 3D (Dreamfusion)
Use a pretrained 2D diffusion model.

$$\theta \in \mathbb{R}^d \to \text{3D shape} \xrightarrow{\textit{projection}} \text{2D image}$$

Then,

$$\underset{\theta}{\text{minimize}} \mathbb{E}[\|\substack{\text{projection of shape} \\ \text{generated by } \theta} - \substack{\text{image from} \\ \text{diffusion model}}\|^2]$$

Fine-tuning.

$$\mathrm{d}\overleftarrow{\mathbf{X}}_t = (\overleftarrow{\mathbf{X}}_t + 2\nabla \log \pi_{T-t}(\overleftarrow{\mathbf{X}}_t))\mathrm{d}t + \mathbf{a}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$$

where $\mathbf{a}_t$ is adopted process.

Goal:

$$\text{minimize}\, \mathbb{E}[\text{loss}(\overleftarrow{\mathbf{X}}_T)] + \alpha \mathcal{D}_{\text{KL}}(\mathbb{P}_{[0,T]} \| \mathbb{P}^{\star}_{[0,T]})$$

- Improve images' visual quality. (e.g. reward model learned from human performance.)
- From experimental data in AI for Science.
- (mostly in LLM) Verifiable reward in reasoning.

Recall score matching objective

$$\min_{\mathbf{s}} \mathbb{E}[\|\nabla \log \pi_t(\mathbf{X}_t) - \mathbf{s}_t(\mathbf{X}_t)\|^2]$$

(a). Implies score matching $\min_{\mathbf{s}} \mathbb{E}[\|\mathbf{s}_t(\mathbf{X})\|^2 + 2\nabla \cdot \mathbf{s}_t(\mathbf{X})]$
(b). Sliced score matching

$$\min_{\mathbf{s}} \mathbb{E}[\|\mathbf{s}_t(\mathbf{X}_t)\|^2 + 2\nu^\top \nabla_{\mathbf{x}}(\nu^\top \mathbf{s}_t(\mathbf{X}_t))]$$

where $\nu \sim \mathcal{N}(0, \mathbb{I}_d)$
(c). Denoising score matching

$$\min_{\mathbf{s}} \mathbb{E}[\|\mathbf{s}_t(\mathbf{X}_t) - \nabla_{\mathbf{y}} \log p_t(\mathbf{X}_t|\mathbf{X}_0)\|^2]$$

(We denote $\nabla_{\mathbf{y}} p_t(\mathbf{y}|\mathbf{x})$ as derivative w.r.t. the first variable)

$$\text{DSM objective} = \mathbb{E}[\|\mathbf{s}_t(\mathbf{X}_t)\|^2] - 2\mathbb{E}[\langle \mathbf{s}_t(\mathbf{X}_t), \nabla_{\mathbf{y}} \log p_t(\mathbf{X}_t|\mathbf{X}_0)\rangle] + \text{sth indp of } \mathbf{s}$$

The second term is

$$
\begin{aligned}
\mathbb{E}[\langle \mathbf{s}_t(\mathbf{X}_t), \nabla_{\mathbf{y}} \log p_t(\mathbf{X}_t|\mathbf{X}_0)\rangle] &= \iint \mathbf{s}_t(\mathbf{y})^\top \nabla_{\mathbf{y}} \log p_t(\mathbf{y}|\mathbf{x}) \pi_0(\mathbf{x}) p_t(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&= \int \mathbf{s}_t(\mathbf{y})^\top \nabla_{\mathbf{y}} \left(\int \pi_0(\mathbf{x}) p_t(\mathbf{y}|\mathbf{x}) \mathrm{d}\mathbf{x} \right) \mathrm{d}\mathbf{y} \\
&= \int \mathbf{s}_t(\mathbf{y})^\top \nabla \pi_t(\mathbf{y}) \mathrm{d}\mathbf{y} \\
&= \mathbb{E}[\langle \mathbf{s}_t(\mathbf{X}_t), \nabla \log \pi_t(\mathbf{X}_t)\rangle]
\end{aligned}
$$

Computing DSM objective

$$
\mathrm{loss}_t(\mathbf{s}_t) = \mathbb{E}[\|\mathbf{s}_t(\mathbf{X}_t) - \nabla_{\mathbf{y}} \log p_t(\mathbf{X}_t|\mathbf{X}_0)\|^2]
$$

where $p_t(\cdot|\mathbf{X}_0) \sim \mathcal{N}(\mu_t(\mathbf{X}_0), \sigma_t \mathbb{I}_d)$. (For OU process, $\mu_t(\mathbf{X}_0) = e^{-t}\mathbf{X}_0$, $\sigma_t = 1 - e^{-2t}$)
Then,

$$
\min_{\mathbf{s}} \mathbb{E}\left[\left\| \mathbf{s}_t(\mathbf{X}_t) + \frac{\mathbf{X}_t - \mu_t(\mathbf{X}_0)}{\sigma_t^2} \right\|^2\right]
$$

Lecture 7. RL basics, value functions and value learning

Discrete-Time RL basics

Notations.

- $\mathcal{S}$: State space.
- $\mathcal{A}$: Action space.
- r: Reward function.
- P: Transition dynamics.

For MDP (Markov Decision Process) $\cdots \to s_{t-1} \xrightarrow{P(\cdot|s_{t-1}, A_{t-1})} s_t \xrightarrow{P(\cdot|s_t, A_t)} s_{t+1} \to \cdots$

Value functions.

- Finite horizon, time: $1, \cdots, T$

$$
\begin{aligned}
v_t^\pi(x) &= \mathbb{E}_\pi\left[\sum_{s=t}^{T} R_s \middle| X_t = x\right] \\
v_t^\star(x) &= \max_\pi v_t^\pi(x)
\end{aligned}
$$

- Discounted, time $0, 1, 2, \cdots$. $\gamma \in (0, 1)$ is called discounted factor.

$$
\begin{aligned}
v^\pi(x) &= \mathbb{E}_\pi\left[\sum_{t=0}^{+\infty} \gamma^t R_t \middle| X_0 = x\right] \\
v^\star(x) &= \max_\pi v^\pi(x)
\end{aligned}
$$

- Infinite horizon, undiscounted. "average reward RL" (Often approximated well by discounted problems)

$$\max_{\pi} \lim_{T\to+\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_\pi[R_t]$$

  The discussion mainly focuses on discounted case for simplicity.

Value function satisfies,

$$\begin{aligned} v^\pi(x) &= \mathbb{E}_\pi[r(x,A) + \gamma v^\pi(X^+)] = J^\pi(v^\pi) \\ v^\star(x) &= \max_{a\in\mathcal{A}} \mathbb{E}_{X^+\sim p(\cdot|x,a)}[r(x,a) + \gamma v^\star(X^+)] = J^\star(v^\star) \end{aligned}$$

Define q-function ,

$$\begin{aligned} Q^\pi(x,a) &= \mathbb{E}_\pi\left[\sum_{t=0}^{T} \gamma^t R_t \mid X_0 = x, A_0 = a\right] \\ Q^\star(x,a) &= \max_\pi Q^\pi(x,a) \end{aligned}$$

Similarly, Q^π, $Q^\star$ also satisfy Bellman equations.

$$\begin{aligned} Q^\pi(x,a) &= \mathbb{E}_\pi[r(x,a) + \gamma Q^\pi(X^+, A^+)] = J_Q^\pi(Q^\pi) \\ Q^\star(x,a) &= \mathbb{E}_\pi[r(x,a)] + \gamma \mathbb{E}\left[\max_{a^+\in\mathcal{A}} Q^\star(x^+, a^+)\right] = J_Q^\star(Q^\star) \end{aligned}$$

Fact. J^π, $J^\star$, J_Q^π, and $J_Q^\star$ are γ-contractions under $\|\cdot\|_\infty$ norm. (i.e. $\|Jv_1 - Jv_2\|_\infty \le \gamma\|v_1 - v_2\|_\infty$)

Fixed-point algorithms: Value Iteration and Policy Iteration.

Value Iteration: $Q^{(t+1)} = J_Q^\star(Q^{(t)})$ $(t = 0, 1, 2, \cdots)$ Until $\|Q^{(t)} - Q^\star\|_\infty < \gamma^t \|Q^{(0)} - Q^\star\|_\infty$

Policy Iteration: $Q^{(t+1)} = (\mathbb{I} - \gamma \mathbf{P}_\pi(t))^{-1} r_{\pi^{(t)}}$, $\pi^{(t+1)} = \mathrm{greedy} - \mathrm{optimal}(Q^{(t+1)})$ Until $\|Q^{(t)} - Q^\star\|_\infty < \gamma^t \|Q^{(0)} - Q^\star\|_\infty$ (The convergence can actually be faster like a Newton method)

Data-driven solutions: RL

Naive idea: plug-in solution.
e.g. for value iteration.

$$\begin{aligned} Q^{(t+1)} &= J_Q^\star(Q^{(t)}) \\ \hat{Q}_n^{(t+1)} &= \hat{J}_n(\hat{Q}_n^{(t)}) \end{aligned}$$

where $\hat{J}_n$ is empirical average estimation for Bellman equation

Alternatively, stochastic approximation:

$$\hat{Q}^{(t+1)} = (1-\beta_t)\hat{Q}^{(t)} + \beta_t J_{t+1}(\hat{Q}^{(t)})$$

It's Q-learning algorithm

From fixed points to projected fixed points

Function approximation: Use function class $\mathcal{F}$ to approximate $Q^\star$

$$\bar{Q} = \Pi_{\mathcal{F}} \circ J^\star(\bar{Q})$$

Ideally, we want to find $\Pi_{\mathcal{F}}(Q^\star)$

Q-learning with function approximation

$$\hat{Q}_n = \hat{\Pi}_n \circ \hat{J}_n(\hat{Q}_n)$$

In practice, update the operator each time using new data point.

Similarly, for policy iteration: "**Actor-Critic algorithm**"

Questions:

- Easy to solve projected fixed point?
- Does projected fixed point always lead to good solution?

In general negative, with some known hardness results.

- High-level reason: L^2-L^∞ mismatch.
- Some special cases are solvable
  - For policy evaluation $\|J^\pi v_1 - J^\pi v_2\|_{L^2} \le \gamma \|v_1 - v_2\|_{L^2}$

$$\mathbb{E}[\|\hat{v}_n^\pi - v^\pi\|_{L^2}^2] \le \frac{1}{1-\gamma} \inf_{v \in \mathcal{F}} \|v - v^\pi\|_{L^2}^2 + \epsilon_\pi^2$$

  - For optimal Q-function
    e.g. Optimal stopping. L^2 contraction
    e.g. Bellman closure / linear MDP

Lecture 8. Continuous-time control problem, HJB equations, continuous-time RL

Continuous-Time RL

For diffusion process with control:

$$\mathrm{d}\mathbf{X}_t = \mathbf{b}_t(\mathbf{X}_t, A_t)\mathrm{d}t + \Sigma_t^{\frac{1}{2}}(\mathbf{X}_t, A_t)\mathrm{d}\mathbf{B}_t$$

Action $A_t = \pi_t(\mathbf{X}_t)$
Define value function:

$$v^\pi(\mathbf{x}) = \begin{cases} \mathbb{E}_\pi\left[\int_t^T r(\mathbf{X}_s)\mathrm{d}s|\mathbf{X}_t = \mathbf{x}\right] & \text{, finite-horizon} \\ \mathbb{E}_\pi\left[\int_0^{+\infty} e^{-\beta t} r(\mathbf{X}_t)\mathrm{d}t|\mathbf{X}_0 = \mathbf{x}\right] & \text{, discounted} \end{cases}$$

$$v^\star(x) = \max_\pi v^\pi(x)$$

Observation/action application every Δt time

- This is just a discrete-time MDP
- We can run any algorithm

Underlying differential equation.

Consider finite-horizon, controlled Markov diffusion, for $[t, t+\Delta t]$

$$v_t^\star(\mathbf{x}) \approx \max_{a\in\mathcal{A}} \mathbb{E}\left[\underbrace{\int_0^{\Delta t} r_{t+s}(\mathbf{X}_{t+s}, a)\mathrm{d}s}_{\approx r_{t+s}(\mathbf{X}_{t+s}, a)\Delta t} + v_{t+\Delta t}^\star(\mathbf{X}_{t+\Delta t})\middle|\mathbf{X}_t = \mathbf{x}\right]$$

By Ito's formulae,

$$\begin{aligned} v_{t+\Delta t}^\star(\mathbf{X}_{t+\Delta t}) = v_t^\star(\mathbf{x}) &+ \int_0^{\Delta t} \frac{\partial v_{t+s}}{\partial s}(\mathbf{X}_{t+s})\mathrm{d}s + \int_0^{\Delta t} \langle \nabla v_{t+s}^\star(\mathbf{X}_{t+s}), \mathbf{b}_{t+s}(\mathbf{X}_{t+s})\rangle \mathrm{d}s \\ &+ \underbrace{\int_0^{\Delta t} \nabla v_{t+s}^\star(\mathbf{X}_{t+s})^\top \Sigma_{t+s}^{\frac{1}{2}}(\mathbf{X}_{t+s})\mathrm{d}\mathbf{B}_s}_{\text{martingale}} + \frac{1}{2}\int_0^{\Delta t} \mathrm{Tr}(\Sigma_{t+s}(\mathbf{X}_{t+s})\nabla^2 v_{t+s}^\star(\mathbf{X}_{t+s}))\mathrm{d}s \end{aligned}$$

Then, we get HJB (Hamilton-Jacobi-Bellman) equation:

$$\boxed{\partial_t v_t^\star(\mathbf{x}) + \max_{a\in\mathcal{A}} r_t(\mathbf{x}, a) + \langle \nabla v_t^\star(\mathbf{x}), \mathbf{b}_t(\mathbf{x}, a)\rangle + \frac{1}{2}\mathrm{Tr}(\nabla^2 v_t^\star(\mathbf{x})\Sigma_t(\mathbf{x}, a)) = 0}$$

Similarly, for discounted case, we have

$$\boxed{\beta v^\star(\mathbf{x}) = \max_{a\in\mathcal{A}} r(\mathbf{x}, a) + \langle \nabla v^\star(\mathbf{x}), \mathbf{b}(\mathbf{x}, a)\rangle + \frac{1}{2}\mathrm{Tr}(\nabla^2 v^\star(\mathbf{x})\Sigma(\mathbf{x}, a)) = 0}$$

For policy evaluation,

$$\boxed{\begin{cases} \partial_t v_t^\pi(\mathbf{x}) + r_t(\mathbf{x}) + \mathcal{L}_\pi v_t^\pi(\mathbf{x}) = 0 & \text{, finite-horizon} \\ \beta v^\pi(\mathbf{x}) = r^\pi(\mathbf{x}) + \mathcal{L}_\pi v^\pi(\mathbf{x}) & \text{, discounted} \end{cases}}$$

where $\mathcal{L}_\pi$ is the diffusion generator

For discounted policy evaluation.
Discrete RL guarantee gives,

$$\mathbb{E}[\|\hat{v} - v^\pi\|_{L^2}^2] \le \underbrace{\frac{1}{1-\gamma}}_{\approx \frac{1}{\beta \Delta t}} \inf_{v \in \mathcal{F}} \|v - v^\pi\|_{L^2}^2 + \epsilon_{\text{statistics}}^2$$

Lecture 9. Advanced topics in continuous-time RL, applications to diffusion model fine-tuning

Recall. policy evaluation

$$\mathrm{d}\mathbf{X}_t^\pi = \mathbf{b}^\pi(\mathbf{X}_t^\pi)\mathrm{d}t + \Sigma^\pi(\mathbf{X}_t^\pi)\mathrm{d}\mathbf{B}_t$$
$$v^\pi(\mathbf{x}) = \int_0^{+\infty} \exp(-\beta t)\mathbb{E}_\pi[r(\mathbf{X}_t)|\mathbf{X}_0 = \mathbf{x}]\mathrm{d}t$$

Observe $\mathbf{X}_0$, $\mathbf{X}_{\Delta t}$, $\mathbf{X}_{2\Delta t}$, $\cdots$, where $\mathbf{X}_0 \sim \text{stationary(denoted as } \mu)$

Projected fixed point

$$\bar{v} = \Pi_{\mathcal{F}, L^2(\mu)} \circ J(\bar{v})$$

Solvable using data (e.g. $\mathcal{F} = \mathbb{S} = \text{span}(\mathbf{\Phi}_1, \cdots, \mathbf{\Phi}_m)$ is linear subspace)

Its empirical version is,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{\Phi}(\mathbf{X}_{i\Delta t})(\mathbf{\Phi}(\mathbf{X}_{i\Delta t}) - \exp(-\beta\Delta t)\mathbf{\Phi}(\mathbf{X}_{(i+1)\Delta t}))^\top \hat{\theta} = \frac{1}{n}\sum_{i=1}^n r_{i\Delta t}\mathbf{\Phi}(\mathbf{X}_{i\Delta t})\Delta t$$

Then,

$$\|\bar{v} - v^\pi\|_{L^2(\mu)} \le \underbrace{\frac{1}{\sqrt{1 - e^{-\beta\Delta t}}}}_{O\left(\frac{1}{\sqrt{\Delta t}}\right)} \inf_{v \in \mathcal{S}} \|\bar{v} - v^\pi\|_{L^2(\mu)} \left(+\sqrt{\Delta t}\right)$$

On the other hand, v^π satisfies $\beta v^\pi - \mathcal{L}v^\pi = r$, where $\mathcal{L}$ is diffusion generator.
If coefficients were known, Galerkin method guarantees good approximations.

Galerkin method: find function $\bar{v} \in \mathbb{S}$, s.t.

$$\langle(\beta - \mathcal{L})\bar{v} - r, f\rangle_{L^2(\mu)} = 0, \forall f \in \mathbb{S}$$

How well does Galerkin method work?

$$\langle(\beta - \mathcal{L})\bar{v} - (\beta - \mathcal{L})v^\pi, f\rangle_{L^2(\mu)} = 0, \forall f \in \mathbb{S}$$

Let $\tilde{v} = \arg\min_{v\in\mathbb{S}}\|v^\pi - v\|$, $f = \bar{v} - \tilde{v}$, then,

$$\langle(\beta - \mathcal{L})(\bar{v} - v^\pi), \bar{v} - \tilde{v}\rangle = 0$$

Hope:

$$\begin{aligned}\lambda\|\bar{v} - \tilde{v}\|_{H^1} &\leq \langle(\beta - \mathcal{L})(\bar{v} - \tilde{v}), \bar{v} - \tilde{v}\rangle_{L^2(\mu)} \\ &= \langle(\beta - \mathcal{L})(\tilde{v} - v^\pi), \bar{v} - \tilde{v}\rangle_{L^2(\mu)} \\ &\leq L\|\tilde{v} - \bar{v}\|_{H^1}\|\tilde{v} - v^\pi\|_{H^1}\end{aligned}$$

where,

$$\|f\|_{H^1(\mu)} = \mathbb{E}_\mu|f(X)|^2 + \mathbb{E}_\mu|\nabla f(X)|^2$$

Lax-Milgram theorem: ∃! PDE solution under these estimates.
Proof.
Lower bound:

$$\begin{aligned}\langle(\beta - \mathcal{L})f, f\rangle_{L^2(\mu)} &= \int\left(\beta f(x) - b^\top\nabla f(x) - \frac{1}{2}\mathrm{Tr}(\Sigma\cdot\nabla^2 f)(x)\right)f(x)\mu(x)\mathrm{d}x \\ &= \beta\|f\|_{L^2(\mu)} + \frac{1}{2}\int\nabla f(x)^\top\nabla(\Sigma f\mu)\mathrm{d}x + \int f(x)\nabla(fb\mu)\mathrm{d}x \\ &\left(\text{using the fact that } -\nabla(b\mu) + \frac{1}{2}\nabla^2(\Sigma\mu) = 0\right) \\ &= \beta\|f\|_{L^2(\mu)} + \frac{1}{2}\int\nabla f(x)^\top\Sigma(x)\nabla f(x)\mu(x)\mathrm{d}x\end{aligned}$$

Assuming $\Sigma(x) \succeq \lambda_{\min}\mathbb{I}_d$, (uniform ellipticity) for $\lambda_{\min} > 0$, $\forall x \in \mathbb{R}^d$

$$\langle(\beta - \mathcal{L})f, f\rangle_{L^2(\mu)} \geq \min\left(\beta, \frac{\lambda_{\min}}{2}\right)\|f\|_{H^1}^2$$

Similarly,

$$|\langle(\beta - \mathcal{L})f, g\rangle_{L^2(\mu)}| \leq L\|f\|_{H^1}^2\|g\|_{H^1}^2$$

So we get (Cea, 1964)

$$\|\bar{v} - v^\pi\|_{H^1(\mu)} \leq \frac{L}{\min\left(\beta, \frac{\lambda_{\min}}{2}\right)}\inf_{v\in\mathbb{S}}\|v - v^\pi\|_{H^1(\mu)}$$

(Possible extension to "hypo-ellipticity", e.g. underdamped Langevin)

For RL (temporal difference) method,

$$\bar{v} = \Pi_{\mathbb{S}, L^2(\mu)}(\exp(-\beta\Delta t)\mathcal{P}_{\Delta t}\bar{v} + r_{\Delta t})$$

i.e.

$$\left\langle \underbrace{\frac{(\mathbb{I}-\exp(-\beta\Delta t)\mathcal{P}_{\Delta t})\bar{v}}{\Delta t}}_{\lim_{\Delta t\to 0}(\cdots)=\beta-\mathcal{L}} - \underbrace{\frac{r_{\Delta t}}{\Delta t}}_{\approx r}, f \right\rangle_{L^2(\mu)} = 0, f \in \mathbb{S}$$

Ideally, we want upper & lower bounds for

$$\frac{\mathbb{I}-\exp(-\beta\Delta t)\mathcal{P}_{\Delta t}}{\Delta t}$$

for finite Δt.
i.e.

$$\left\langle \frac{\mathbb{I}-\exp(-\beta\Delta t)\mathcal{P}_{\Delta t}}{\Delta t} f, f \right\rangle_{L^2(\mu)} \underbrace{\gtrsim}_{?} \|f\|^2_{H^1(\mu)}$$

- impossible in general
- but under some regularity conditions on basis functions of $\mathbb{S}$, true for $\Delta t \le \Delta t_{\text{thres}}(\mathbb{S})$, the lower bound holds for $\forall f \in \mathbb{S}$
- upper bounds still hold for finite Δt

Conclusion:

$$\|\bar{v} - v^\pi\|_{H^1} \le C \cdot \inf_{v\in\mathbb{S}} \|v - v^\pi\|_{H^1}$$

$\bar{v}$ is target of TD algo, its statistic error is:

$$\|\bar{v} - \hat{v}_T\|_{H^1(\mu)} \lesssim \frac{m \cdot t_{\text{mix}}}{T} \inf_{v\in\mathbb{S}} \|v - v^\pi\|_{W^{1,p}} + \underbrace{\frac{g(m)}{T}}_{g(m)=o(m)}$$

- Nonstandard trade-off, policy evaluation is easier than regression.
- Extension to general Q-learning: Possible under certain cases. (ongoing work)

Diffusion fine-tuning:

$$\min \mathbb{E}\left[C(T) + \int_0^T y(X_t)\mathrm{d}t\right] + \alpha\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_{[0,T]} \| \mathbb{P}^{\text{pretrained}}_{[0,T]})$$

s.t. $\mathrm{d}\mathbf{X}_t = \mathbf{b}(\mathbf{X}_t)\mathrm{d}t + \mathbf{A}_t\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$

HJB eq:

$$\frac{\partial v_t^\star(x)}{\partial t} + \min_{a\in\mathbb{R}^d}[\langle b(x)+a, \nabla v_t^\star(x)\rangle + \Delta v_t^\star(x) + y(x) + \alpha|a|^2] = 0$$

then,

$$\partial_t v_t^\star(x) + \langle b(x), \nabla v_t^\star(x)\rangle + \Delta v_t^\star(x) - \frac{1}{4\alpha}|\nabla v_t^\star(x)|^2 + y_t(x) = 0$$

Exponential transformation, let $f_t^\star = \exp\left(\frac{v_t^\star}{\alpha}\right)$, then,

$$(\partial_t + \alpha y_t + \mathcal{L}_t) f_t^\star = 0$$

boundary condition: $f_T^\star = \exp(\frac{C}{\alpha})$

RL-fine-tuning is as easy as regression for diffusion models (and even easier)

Talk is cheap, show me the code/implementations

Some key points to bridge the gap between theoretical analysis and empirical implementation.

Denoising diffusion probabilistic model

We have data points $x_i \sim p_{\text{data}}$. Given a (invertible) mapping that add noise to images $f : x_{t-1} \mapsto x_t$ such that $\lim_{T\to\infty} f^T(x_0) \approx z \sim \mathcal{N}(0, \mathbf{I})$, if we have learned $\mu \approx f^{-1}$, we can generate any image distributed in p_{data} by $\hat{x} = \mu^T(z)$, where $z \sim \mathcal{N}(0, \mathbf{I})$.

In DDPM, we formulate $f : x_{t-1} \mapsto x_t = \alpha_t x_{t-1} + \beta_t \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. (Markov? Maybe helpful) Then we have $x_T = (\prod_{i=1}^T \alpha_i)x_0 + \sum_{i=1}^T (\prod_{j=i+1}^T \alpha_j)\beta_i\epsilon_i$. If we have $\alpha_i^2 + \beta_i^2 = 1$, then the noisy term $\sum_{i=1}^T(\prod_{j=i+1}^T \alpha_j)\beta_i\epsilon_i \sim \mathcal{N}(0, 1 - (\prod_{i=1}^T \alpha_i)^2)$. Denote $\bar{\alpha}_T = \prod_{i=1}^T \alpha_i$, $\bar{\beta}_T = \sqrt{1 - \bar{\alpha}_T^2}$, then we can add noise by one step sampling:

$$x_T = \bar{\alpha}_T x_0 + \bar{\beta}_T \bar{\epsilon}_T, \quad \bar{\epsilon}_T \sim \mathcal{N}(0, \mathbf{I})$$

How to train DDPM empirically?
We have training dataset $\mathcal{D} = \{(x_{t-1}, x_t)\}$ consisting of data pairs.
Simple idea: Since we get x_t by adding noise $x_t = \alpha_t x_{t-1} + \beta_t\epsilon_t$, we can denoise it by $x_{t-1} = \frac{1}{\alpha_t}(x_t - \beta_t\epsilon_t)$. Then we can:

- Learn a model $\mu_\theta : x_t \mapsto x_{t-1}$, then $\theta = \arg\min_\theta \mathcal{L} = \arg\min_\theta \|x_{t-1} - \mu_\theta(x_t)\|_2^2$
- Formulate $\mu_\theta(x_t) = \frac{1}{\alpha_t}(x_t - \beta_t\epsilon_\theta(x_t, t))$
- Then we have the new objective $\mathcal{L} = \|x_{t-1} - \mu_\theta(x_t)\|_2^2 = (\frac{\beta_t}{\alpha_t})^2\|\epsilon_t - \epsilon_\theta(x_t, t)\|_2^2$

We can sample $x_t = \bar{\alpha}_t x_0 + \bar{\beta}_t\bar{\epsilon}_t$. However, note that ϵ_t and $\bar{\epsilon}_t$ are not independent, so we sample $x_t = \alpha_t x_{t-1} + \beta_t\epsilon_t = \alpha_t(\bar{\alpha}_{t-1}x_0 + \bar{\beta}_{t-1}\bar{\epsilon}_{t-1}) + \beta_t\epsilon_t$
Then the objective is (remove weight term):

$$\begin{aligned}\mathcal{L} &= \|\epsilon_t - \epsilon_\theta(x_t, t)\|_2^2 \\ &= \|\epsilon_t - \epsilon_\theta(\bar{\alpha}_t x_0 + \alpha_t\bar{\beta}_{t-1}\bar{\epsilon}_{t-1} + \beta_t\epsilon_t, t)\|_2^2\end{aligned}$$

During training phase, we should sample $x_0 \sim p_{\text{data}}$, $t \sim \mathcal{U}(1, T)$ and $\epsilon_t, \bar{\epsilon}_{t-1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$ and minimize $\mathbb{E}(\mathcal{L})$. But sampling 4 r.v. makes $\text{Var}(\mathcal{L})$ too big.

Note that $(\alpha_t\bar{\beta}_{t-1})^2+\beta_t^2=\bar{\beta}_t^2$, so we have,

$$
\begin{aligned}
\alpha_t\bar{\beta}_{t-1}\bar{\epsilon}_{t-1}+\beta_t\epsilon_t &\sim \bar{\beta}_t\epsilon \\
\beta_t\bar{\epsilon}_{t-1}-\alpha_t\bar{\beta}_{t-1}\epsilon_t &\sim \bar{\beta}_t\omega
\end{aligned}
$$

where $\epsilon,\omega{\sim}\mathcal{N}(0,\mathbf{I})$. Then we could calculate,

$$
\epsilon\omega^\top=\frac{1}{\bar{\beta}_t^2}[\alpha_t\bar{\beta}_{t-1}\beta_t(\bar{\epsilon}_{t-1}\bar{\epsilon}_{t-1}^\top-\epsilon_t\epsilon_t^\top)+(\beta_t^2-\alpha_t^2\bar{\beta}_{t-1}^2)\bar{\epsilon}_{t-1}\epsilon_t^\top]
$$

So we have $\mathbb{E}(\epsilon\omega^\top)=0$, i.e. $\epsilon\perp\omega$. So the final objective is

$$
\begin{aligned}
\mathbb{E}(\mathcal{L}) &= \mathbb{E}_{\bar{\epsilon}_{t-1},\epsilon_t\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\mathbf{I})}[\|\epsilon_t-\epsilon_\theta(\bar{\alpha}_t x_0+\alpha_t\bar{\beta}_{t-1}\bar{\epsilon}_{t-1}+\beta_t\epsilon_t,t)\|_2^2] \\
&= \mathbb{E}_{\epsilon,\omega\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\mathbf{I})}\left[\left\|\frac{1}{\bar{\beta}_t}(\beta_t\epsilon-\alpha_t\bar{\beta}_{t-1}\omega)-\epsilon_\theta(\bar{\alpha}_t x_0+\bar{\beta}_t\epsilon,t)\right\|_2^2\right] \\
&= \frac{\beta_t^2}{\bar{\beta}_t^2}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\left\|\epsilon-\frac{\bar{\beta}_t^2}{\beta_t^2}\epsilon_\theta(\bar{\alpha}_t x_0+\bar{\beta}_t\epsilon,t)\right\|_2^2\right]+\text{some constant}
\end{aligned}
$$

And the empirical loss is

$$
\mathcal{L}_{\text{empirical}}=\frac{1}{n}\sum\left\|\epsilon-\frac{\bar{\beta}_t^2}{\beta_t^2}\epsilon_\theta(\bar{\alpha}_t x_0+\bar{\beta}_t\epsilon,t)\right\|_2^2
$$

where $x_0\sim p_{\text{data}},\epsilon\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}(1,T)$.