§7. Diffusion Models.

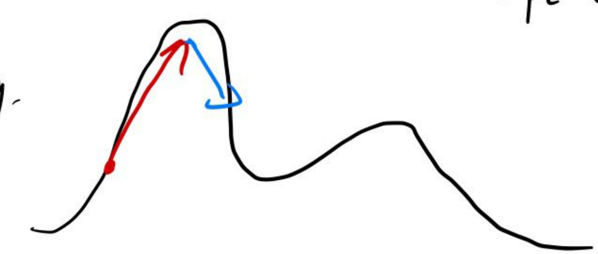
Goal: Given iid samples $\{x_i\}_{i=1}^n$ from the target distribution ρ.

generate more samples from ρ (approximately).

§7.1. Langevin and score-matching.

Suppose $\rho(x) \propto \exp(-V(x))$. $V: \mathbb{R}^d \to \mathbb{R}$

e.g.



Langevin diffusion. $dX_t = -\nabla V(X_t)\,dt + \sqrt{2}\,dB_t$

↳ move toward the modes

↳ some random noise preventing $X_t \to$ (local) max.

<u>Thm.</u> If ρ is <u>log-concave</u> (i.e., V is convex) then $\mathrm{Law}(X_t) \longrightarrow \rho$.

Question. How to estimate the <u>score function</u> $\nabla \log \rho(x) = -\nabla V(x)$?

Let $s(\cdot\,;\theta): \mathbb{R}^d \to \mathbb{R}^d$ be our score estimator. (e.g. some deep network).
$\hookrightarrow$ parameter.

Goal:

$$\underset{\theta}{\text{minimize}} \quad J(\theta) := \mathbb{E}_{x\sim p} \left\| s(x;\theta) - \underbrace{(-\nabla U(x))}_{\text{We don't have access to this.}} \right\|_2^2$$

<u>Lemma.</u> Assume some boundary conditions. We have

$$J(\theta) = \underbrace{\frac{1}{2}\mathbb{E}_{x\sim p}\|s(x;\theta)\|^2 + \mathbb{E}_{x\sim p}\nabla\cdot s(x;\theta)}_{\text{can be estimated using } \{x_i\}_{i=1}^n} + \underbrace{C}_{\text{doesn't depend on } \theta.}$$

Pf. We write

$$J(\theta) = \frac{1}{2}\mathbb{E}_{x\sim p}\|s(x;\theta)\|^2 + \underbrace{\frac{1}{2}\mathbb{E}_{x\sim p}\|\nabla U(x)\|^2}_{=:\,C} + \mathbb{E}_{x\sim p}\langle s(x;\theta), \nabla U(x)\rangle.$$

For the third term, we compute.

$$\mathbb{E}_{\rho}\langle S(x;\theta), \nabla U(x)\rangle = \sum_{k=1}^{d}\int S_k(x;\theta)\, \partial_k U(x)\, \rho(x)\, dx$$

$$= -\sum_{k=1}^{d}\int S_k(x;\theta)\, \partial_k \log \rho(x)\, \rho(x)\, dx$$

$$= -\sum_{k=1}^{d}\int S_k(x;\theta)\, \partial_k \rho(x)\, dx$$

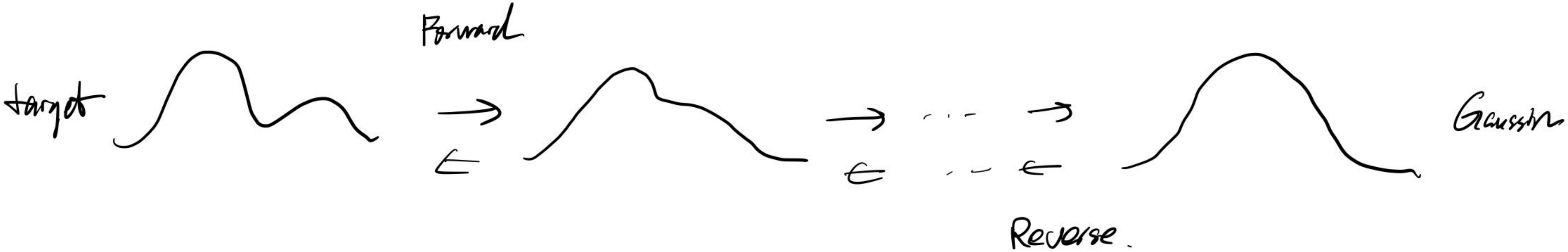
$$= -\sum_{k=1}^{d}\left(\cancelto{0}{S_k(x;\theta)\,\rho(x)\Big|_{\|x\|=\infty}} - \int \partial_k S_k(x;\theta)\, \rho(x)\, dx \right)$$

the boundary condition we need

$$= \int \sum_{k=1}^{d} \partial_k S_k(x;\theta)\, \rho(x)\, dx = \mathbb{E}_{x\sim\rho}\, \nabla \cdot S(x;\theta).$$

$\blacksquare$

§7.2 Diffusion models



Forward process: $dX_t = -X_t\,dt + \sqrt{2}\,dB_t$, $\quad X_0 \sim \rho$.

$$\boxed{q_t := \mathrm{Law}(X_t)}$$

Reverse process: $X_t^{\leftarrow} := X_{T-t}$.

We have $dX_t^{\leftarrow} = \left(X_t^{\leftarrow} + 2\nabla \log q_{T-t}(X_t^{\leftarrow})\right) dt + \sqrt{2}\,dB_t$, $\quad X_0^{\leftarrow} \sim q_T$.

↳ score fn. can be learned in the forward process.

Facts. 1) $\mathrm{Law}(X_t) \to \gamma^d$ (d-dim std. Gaussian).

2) $\mathrm{Law}(X_t^{\leftarrow}) = q_{T-t}$.

Discretized reverse process. (step size $h > 0$, score estimator $s_t \approx \nabla \log q_t$).

$$dX_t^{\leftarrow} = \left(X_t^{\leftarrow} + 2 s_{T-kh}(X_{T-kh}^{\leftarrow}) \right) dt + \sqrt{2}\, dB_t, \qquad t \in [kh, (k+1)h].$$

$X_0^{\leftarrow} \sim \gamma^d$

not q_T since we can't sample from q_T.

(No need to replace all $X_t^{\leftarrow}$ with $X_{kh}^{\leftarrow}$. The above (1-step) linear SDE can be solved in closed form.)

$P_t := \mathrm{Law}(X_t^{\leftarrow})$.

Goal: $P_T \approx q_0 = \rho$.

☆ No log-concavity conditions or isoperimetric inequalities.

Assumptions.

a). $\forall t \geq 0$. $\mathrm{supp}\, q_t = \mathbb{R}^d$, $\nabla \log q_t$ is L-Lipschitz.

b). For some $\eta > 0$. $\mathbb{E}_\rho \|X\|^{2+\eta} < \infty$. Define $m_2^2 := \mathbb{E}_\rho \|X\|^2$

c). $\forall k \in [N]$. $\mathbb{E}_{q_{kh}} \| s_{kh} - \nabla \log q_{kh} \|^2 \leq \varepsilon_{score}^2$.

Error sources: 1) $X_0^{\leftarrow} \sim \gamma^d$ instead of q_T / forward process not converging

2) Using the time-discretized SDE.

Øksendal.

3) non-exact score estimator s.

Thm. 2 of (Chen et al. 2023). Under the previous assumptions. choose $h = T/N \le \min\{1/L, 1\}$.

$$\Rightarrow \quad TV(P_T, \rho) \le \sqrt{KL(\rho \| \gamma^d)} \exp(-T) + (2\sqrt{dh} + L m_2 h)\sqrt{T} + \varepsilon_{score}\sqrt{T}.$$

error 1) 2) 3).

1). $\tilde{P}_T$: distribution of $X_T^{\leftarrow}$ if we run the continuous reverse SDE with init $= \gamma^d$.

$$TV(\tilde{P}_T, \rho) \le TV(\gamma^d, q_T) \le \sqrt{KL(q_T \| \gamma^d)} \le \exp(-T)\sqrt{KL(\rho \| \gamma^d)}$$

Data processing inequality along the reverse process

Pinsker's inequality

exp. convergence of the forward process.

To handle errors 2) and 3), we need Girsanov's theorem.

<u>Thm 1 (Girsanov diffusion)</u>. Consider $dX_t = b(X_t)\,dt + \sigma(X_t)\,dB_t$, $\quad t \le T$.

and $dY_t = [\,b(Y_t) + \gamma(t,\omega)\,]\,dt + \sigma(Y_t)\,dB_t$, $\quad t \le T$.

Assume some regularity conditions (and Novikov's condition). Define

$$u(t,\omega) = \gamma(t,\omega) / \sigma(Y_t) \longrightarrow \text{normalized error}$$

$$M_t(\omega) = \exp\left(-\int_0^t u(s,\omega)\,dB_s - \frac{1}{2}\int_0^t u^2(s,\omega)\,ds\right).$$

Then, we have $\mathrm{Law}(Y_t) / \mathrm{Law}(X_t) = M_t$ The statement here is wrong. TODO: fix this

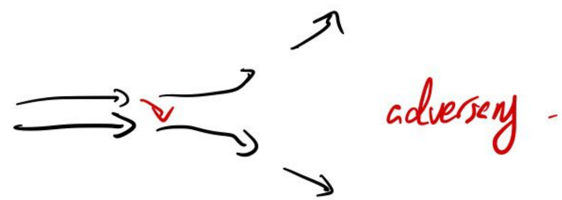
<u>Corollary</u>. Put $P = \mathrm{Law}(X_T)$, $Q = \mathrm{Law}(Y_T)$. Then we have

$$KL(P\|Q) = \mathbb{E}_P \log\frac{P}{Q} = -\mathbb{E}_P \log M_t = \overset{0}{\cancel{\mathbb{E}_P \int_0^T u(s,\omega)\,dB_s}} + \underbrace{\frac{1}{2}\,\mathbb{E}_P \int_0^T u^2(s,\omega)\,ds}_{L^2 \text{ error of the drift term.}}$$

Remark. Novikov's condition doesn't hold in our setting, and we actually need some limiting argument.

Remark. For ODEs, small L^2 error $\not\Rightarrow$ small worst case error

e.g.



For SDEs, we don't have "worst case"!

Bounding the discretization error. Under the assumptions of Thm 1, assuming Novikov's condition, let $Q_T^{\leftarrow}$ and $P_T^{q_T}$ denote the distribution of the continuous and discretized reverse process initialized at q_T, respectively.

Goal: Bound $TV(P_T^{q_T}, Q_T^{\leftarrow})$.

Pf. Girsanov $\Rightarrow$ $KL(Q_T^{\leftarrow} \| P_T^{q_T}) \leq \sum_{k=0}^{N-1} \mathbb{E}_{Q_T^{\leftarrow}} \int_{kh}^{(k+1)h} \| s_{T-kh}(X_{kh}) - \nabla \log q_{T-t}(X_t) \|^2 dt$.

$$\mathbb{E}_{Q_T^\epsilon} \| s_{T-kh}(X_{kh}) - \nabla \log q_{T-t}(X_t) \|^2$$

$$\lesssim \mathbb{E}_{Q_T^\epsilon} \| s_{T-kh}(X_{kh}) - \nabla \log q_{T-kh}(X_{kh}) \|^2 + \mathbb{E}_{Q_T^\epsilon} \| \nabla \log q_{T-kh}(X_{kh}) - \nabla \log q_{T-t}(X_{kh}) \|^2$$

$$+ \mathbb{E}_{Q_T^\epsilon} \| \nabla \log q_{T-t}(X_{kh}) - \nabla \log q_{T-t}(X_t) \|^2$$

$$\lesssim \varepsilon_{score}^2 + \underbrace{\mathbb{E}_{Q_T^\epsilon} \| \nabla \log \frac{q_{T-kh}}{q_{T-t}}(X_{kh}) \|^2}_{\text{change of the score function within each step along the forward process}} + L^2 \underbrace{\mathbb{E}_{Q_T^\epsilon} \| X_{kh} - X_t \|^2}_{\text{movement of } X}.$$

Define $S(x) = \exp(-(t-kh))x$. We have $q_{T-kh} = S\# q_{T-t} * N(0, 1-\exp(-2(t-kh)))$.

<u>Lemma.</u> (Score perturbation lemma). $M_0, M_1 \in \mathbb{R}^{d\times d}$, $\|M_0 - I_d\|_{op} \le \zeta < 1$, M_1 symmetric.

$q = \exp(-H) \in P(\mathbb{R}^d)$. ∇H L-Lipschitz. $L \le \frac{1}{4\|M\|_{op}}$.

$$\Rightarrow \| \nabla \log \frac{M_0 \# q * N(0, M_1)}{q}(\theta) \| \lesssim L\sqrt{\|M_1\|_{op}\, d} + L\zeta \|\theta\| + (\zeta + L\|M_1\|_{op}) \|\nabla H(\theta)\|.$$

$$\Rightarrow \mathbb{E}_{\mathbb{Q}_T^\epsilon} \left\| \nabla \log \frac{q_{T-kh}}{q_{T-t}}(X_{kh}) \right\|^2 = \mathbb{E}_{\mathbb{Q}_T^\epsilon} \left\| \nabla \log \frac{S\# q_{T-t} * N(0, 1-\exp(-2(t-kh)))}{q_{T-t}}(X_{kh}) \right\|^2$$

$$\lesssim L^2 d h + L^2 h^2 \mathbb{E}_{\mathbb{Q}_T^\epsilon} \|X_{kh}\|^2 + L^2 h^2 \mathbb{E}_{\mathbb{Q}_T^\epsilon} \underbrace{\| \nabla \log q_{T-t}(X_{kh}) \|^2}$$

$$\hookrightarrow \lesssim \| \nabla \log q_{T-t}(X_t) \|^2 + L^2 \|X_{kh} - X_t\|^2$$

Claim. $\mathbb{E}\|X_t\|^2 \le d \vee m_2^2$. $\mathbb{E}\|\nabla \log q_t(X_t)\|^2 \le Ld$. $\mathbb{E}\|X_t - X_s\|^2 \lesssim (t-s)^2 m_2^2 + (t-s)d$.

Thus.

$$\mathbb{E}_{\mathbb{Q}_T^\epsilon} \| s_{T-kh}(X_{kh}) - \nabla \log q_{T-t}(X_t) \|^2 \le \varepsilon_{score}^2 + L^2 dh + L^2 h^2 (d + m_2^2) + L^3 h^2 d + L^4 h^2 (h^2 m_2^2 + dh)$$

$$\lesssim \varepsilon_{score}^2 + L^2 dh + L^2 m_2^2 h^2 .$$

$\blacksquare$