



Homework 2
 Due 5pm, Friday, May 10, 2024

Problem 1: *Reverse conditional distribution conditioned on X_0 .* Consider the DDPM forward process

$$\mathcal{P}(X_t | X_{t-1}) \sim \mathcal{N}(\sqrt{1 - \beta_t} X_{t-1}, \beta_t I)$$

for $t = 1, 2, \dots$ with $X_0 \sim p_{\text{data}}$. Show that

$$\mathcal{P}(X_{t-1} | X_t, X_0) = \mathcal{N}(\mu_t(X_t | X_0), \tilde{\beta}_t I),$$

$$\mu_t(X_t | X_0) = \frac{1}{\sqrt{1 - \beta_t}} (X_t + \beta_t \nabla_{X_t} \log p_{t|0}(X_t | X_0)), \quad \tilde{\beta}_t = \frac{1 - \prod_{s=1}^{t-1} (1 - \beta_s)}{1 - \prod_{s=1}^t (1 - \beta_s)} \beta_t$$

for $t = 1, 2, \dots$. Do not assume $\beta_t \approx 0$.

Problem 2: *DDIM marginals.* Consider the DDIM “forward” process

$$q(X_1, \dots, X_T | X_0) = q(X_T | X_0) \prod_{t=1}^{T-1} q(X_t | X_{t+1}, X_0)$$

$$q(X_T | X_0) = \mathcal{N}(\sqrt{\alpha_T} X_0, (1 - \alpha_T) I)$$

$$q(X_t | X_{t+1}, X_0) = \mathcal{N}\left(\sqrt{\alpha_t} X_0 + \frac{\sqrt{1 - \alpha_t - \sigma_{t+1}^2}}{\sqrt{1 - \alpha_{t+1}}} (X_{t+1} - \sqrt{\alpha_{t+1}} X_0), \sigma_{t+1}^2 I\right), \quad t = T - 1, \dots, 1,$$

where $\alpha_T, \dots, \alpha_1$ is a sequence in $(0, 1)$ and $\sigma_T, \dots, \sigma_2$ is sequence of positive numbers satisfying $\sigma_{t+1}^2 \leq 1 - \alpha_t$ for all $t = 1, \dots, T - 1$. Show that

$$X_t | X_0 \sim \mathcal{N}(\sqrt{\alpha_t} X_0, (1 - \alpha_t) I), \quad t = 1, \dots, T.$$

Hint. Use the fact that

$$\begin{aligned} X_T &\stackrel{\mathcal{D}}{=} \sqrt{\alpha_T} X_0 + \sqrt{1 - \alpha_T} \varepsilon_T \\ X_t &\stackrel{\mathcal{D}}{=} \sqrt{\alpha_t} X_0 + \frac{\sqrt{1 - \alpha_t - \sigma_{t+1}^2}}{\sqrt{1 - \alpha_{t+1}}} (X_{t+1} - \sqrt{\alpha_{t+1}} X_0) + \sigma_{t+1} \varepsilon_t, \quad t = T - 1, \dots, 1 \end{aligned}$$

for IID $\varepsilon_T, \varepsilon_{T-1}, \dots, \varepsilon_1 \sim \mathcal{N}(0, I)$.

Problem 3: Denoising score matching loss near $t = 0$. Consider the 1-dimensional Ornstein–Uhlenbeck process

$$dX_t = -\frac{1}{2}X_t dt + dW_t$$

for $t \in [0, T]$, where $X_0 \sim p_0$. For simplicity, let $p_0 = \mathcal{N}(0, 1)$. Let

$$\gamma_t = e^{-t/2}, \quad \sigma_t^2 = 1 - e^{-t}.$$

Consider the loss

$$\begin{aligned} \mathcal{L}(\theta) &= \mathbb{E}_{t \sim \text{Uniform}([0, T])} \left[\mathbb{E}_{X_0 \sim p_0} \left[\mathbb{E}_{X_t | X_0} \left[\lambda(t) \left(s_\theta(X_t, t) - \frac{d}{dX_t} \log p_{t|0}(X_t | X_0) \right)^2 \middle| X_0 \right] \right] \right] \\ &= \mathbb{E}_{\substack{t \sim \text{Uniform}([0, T]) \\ X_0 \sim p_0 \\ \varepsilon \sim \mathcal{N}(0, I)}} \left[\frac{\lambda(t)}{\sigma_t^2} (\varepsilon_\theta(\gamma_t X_0 + \sigma_t \varepsilon, t) - \varepsilon)^2 \right], \end{aligned}$$

where $\delta \geq 0$, $\lambda(t) \geq 0$ is a continuous function, s_θ is a score network, and $\varepsilon_\theta(X_t, t) = \sigma_t s_\theta(X_t, t)$. It is customary to use $\delta > 0$ to “avoid numerical instabilities.” In this problem, we explore issues that arise when $\delta = 0$.

- (a) Show that $p_t = \mathcal{N}(0, 1)$ for all $t > 0$.
- (b) Assume s_θ has been perfectly trained, i.e., $s_\theta(X_t, t) = \frac{d}{dX_t} \log p_t(X_t) = -X_t$. Show that if $\min_{t \in [0, T]} \lambda(t) > 0$, then

$$\begin{aligned} \mathcal{L}(\theta) &\geq \left(\min_{t \in [0, T]} \lambda(t) \right) \mathbb{E}_{t, X_0, X_t} \left[\left(s_\theta(X_t, t) - \frac{d}{dX_t} \log p_{t|0}(X_t | X_0) \right)^2 \right] \\ &= \infty. \end{aligned}$$

- (c) Show that if $\lambda(t) = \sigma_t^2$ (so $\min_{t \in [0, T]} \lambda(t) = 0$) and if $s_\theta(X_t, t) = \frac{d}{dX_t} \log p_t(X_t)$, then

$$\mathcal{L}(\theta) < \infty.$$

- (d) Let $\lambda(t) = \sigma_t^2$. Assume there is a θ^* such that $s_{\theta^*}(X_t, t) = \frac{d}{dX_t} \log p_t(X_t)$. Let θ be such that $s_\theta(X_t, t) = \frac{m}{\sigma_t} X_t + \frac{d}{dX_t} \log p_t(X_t)$ for some small $m > 0$. Show that

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) = m^2.$$

(Conceptually, m^2 is small, so s_θ is nearly optimal with respect to the loss \mathcal{L} .)

- (e) Let $s_\theta(X_t, t) = \frac{m}{\sigma_t} X_t + \frac{d}{dX_t} \log p_t(X_t)$ for some small $m > 0$. Show that the reverse sampling ODE with the trained score s_θ is of the form

$$d\bar{X}_t = F(\bar{X}_t, t) dt,$$

where $F(X_t, t)$ blows up as $t \rightarrow 0$. (Since the ODE is singular, we expect numerical solutions of it via discretizations to be numerically unstable.)

Remark. The ODE

$$d\bar{X}_t = -\frac{1}{\sqrt{t}} \bar{X}_t$$

has a general solution $\bar{X}_t = \exp(-2\sqrt{t})$ for $t \geq 0$, so a singular ODE (an ODE with a RHS that blows up) does not necessarily have a singular solution (a solution that blows up).

Problem 4: Why output projection on MHA? Consider the standard multi-head self-attention (MHA) layer defined by

$$\underbrace{\text{output}}_{L \times d_{\text{out}}} = \underbrace{\text{concat}(\text{head}_1, \dots, \text{head}_H)}_{L \times H d_{\text{head}}} W^O$$

$$\underbrace{\text{head}_h}_{L \times d_{\text{head}}} = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V) \quad \text{for } h = 1, \dots, H,$$

$$\text{Attention}(\tilde{Q}, \tilde{K}, \tilde{V}) = \text{softmax}\left(\frac{\tilde{Q}\tilde{K}^\top}{\sqrt{d_{\text{attn}}}}\right)\tilde{V},$$

where

$$W^O \in \mathbb{R}^{H d_{\text{head}} \times d_{\text{out}}}$$

$$W_h^Q \in \mathbb{R}^{d_Q \times d_{\text{attn}}}, \quad W_h^K \in \mathbb{R}^{d_K \times d_{\text{attn}}}, \quad W_h^V \in \mathbb{R}^{d_V \times d_{\text{head}}}$$

$$Q \in \mathbb{R}^{L \times d_Q}, \quad K \in \mathbb{R}^{L \times d_K}, \quad V \in \mathbb{R}^{L \times d_V}.$$

(Of course, it is often the case that $Q = K = V = X \in \mathbb{R}^{L \times d}$.) Let us call this model MHA1. Next, consider a variant that we call MHA2.

$$\underbrace{\text{output}}_{L \times d_{\text{out}}} = \text{head}_1 + \dots + \text{head}_H$$

$$\underbrace{\text{head}_h}_{L \times d_{\text{head}}} = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V) \quad \text{for } h = 1, \dots, H,$$

$$\text{Attention}(\tilde{Q}, \tilde{K}, \tilde{V}) = \text{softmax}\left(\frac{\tilde{Q}\tilde{K}^\top}{\sqrt{d_{\text{attn}}}}\right)\tilde{V},$$

where

$$W_h^Q \in \mathbb{R}^{d_Q \times d_{\text{attn}}}, \quad W_h^K \in \mathbb{R}^{d_K \times d_{\text{attn}}}, \quad W_h^V \in \mathbb{R}^{d_V \times d_{\text{out}}}$$

$$Q \in \mathbb{R}^{L \times d_Q}, \quad K \in \mathbb{R}^{L \times d_K}, \quad V \in \mathbb{R}^{L \times d_V}.$$

(a) Given an MHA1 model, decompose the rows of W^O as

$$W^O = \begin{bmatrix} W_1^O \\ W_2^O \\ \vdots \\ W_H^O \end{bmatrix} \in \mathbb{R}^{H d_{\text{head}} \times d_{\text{out}}}$$

such that $W_1^O, W_2^O, \dots, W_H^O \in \mathbb{R}^{d_{\text{head}} \times d_{\text{out}}}$. Show that if we set the parameters of an MHA2 model as $W_h^V \leftarrow W_h^V W_h^O$ for $h = 1, \dots, H$ and keep all other parameters the same, then the MHA1 and MHA2 models are equivalent, i.e., $(\text{MHA1}(Q, K, V) = \text{MHA2}(Q, K, V))$ for all inputs Q, K, V .

(b) How many trainable parameters do MHA1 and MHA2 have?

(c) If $d_V = d_{\text{out}} = 512$ and $d_{\text{head}} = 64$, what is the difference in the number of trainable parameters?