Deep Generative Models
Homework Set 3 (Sol)

Mohammad Parsa Dini

13 December, 2024

Problem 1

Inverse Function Calculation We start by computing the inverse function $f^{-1}(x)$. Given x, we find z:

$$z = f^{-1}(x) = \begin{bmatrix} x_1 \\ x_2 e^{-x_1} \\ (x_3^3 - x_1^2)e^{x_1} \end{bmatrix}$$

Substituting the values to find z:

$$z = f^{-1}\left(\begin{bmatrix} 0 \\ 1 \\ \frac{1}{3} \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 1 \\ \frac{1}{27} \end{bmatrix}$$

Jacobian Matrix Calculation The Jacobian matrix J_f is calculated to understand how small changes in z affect $f(z)$:

$$J_f = \begin{bmatrix} 1 & 0 & 0 \\ z_2 e^{z_1} & e^{z_1} & 0 \\ \frac{1}{3}\left(z_3 e^{-z_1} + z_1^2\right)^{-\frac{2}{3}} & 0 & \frac{1}{3}\left(z_3 e^{-z_1} + z_1^2\right)^{-\frac{2}{3}} e^{-z_1} \end{bmatrix}$$

Determinant Relationship The determinant $|J_f|$ and its inverse are crucial in transforming between probability densities. Since $|J_f| = |J_{f^{-1}}|^{-1}$, we have:

Probability Density Calculation Using the determinant relationship and exponential functions, we calculate $p(x)$, the probability density at x:

$$p(x) = (2\pi)^{-\frac{3}{2}} \exp\left(-\frac{1}{2} z^T z\right) \cdot |J_f|^{-1}$$

Substituting the values and simplifying the expression:

$$p(x) = (2\pi)^{-\frac{3}{2}} \exp\left(-\frac{65}{128}\right)\left(\frac{1}{3}(1/27)^{-\frac{2}{3}}\right)^{-1} = \frac{1}{3}(2\pi)^{-\frac{3}{2}} \exp\left(-\frac{65}{128}\right)$$

Problem 2

A. Let $\bar{A}$ be the complement of the set $A = \emptyset$. We can see that:

$$P(\bar{A}) = P(B) = 1 \quad \text{and let} \quad P_\theta(\bar{A}) < 0.9 - \epsilon$$

. We are trying to obtain a lower bound for KL divergence $D_{KL}(p||p_\theta)$ that can be positive for some $\epsilon > 0$ in order to show that there exists an event $E \in \Omega$ such that: $|p_\theta(E) - p(E)| > 0$. Now, we compute the absolute difference:

$$|P(\bar{A}) - P_\theta(\bar{A})| < |0.9 - (0.9 - \epsilon)| = \epsilon$$

Simplifying the expression above leads to:

$$|P(\bar{A}) - P_\theta(\bar{A})| < |1 - 0.9 + \epsilon| = \epsilon$$

Using Pinkster's inequality, we derive a lower bound for the KL divergence:

$$D_{KL}(p_\theta \parallel p) \geq 2 \cdot \delta(p_\theta, p)^2 = 2TV(p_\theta, p)^2$$

So we can also let ϵ such that $\epsilon \leq TV(p_\theta, p)$:

$$D_{KL}(p_\theta \parallel p) \geq 2TV(p_\theta, p)^2 = 2 \cdot \delta(p_\theta, p)^2 \geq 2 \cdot \epsilon^2$$

Hence, for some $\epsilon > 0$, the absolute difference $|p_\theta(E) - p(E)| > 0$, and the KL divergence is bounded below by $2 \cdot \epsilon^2$, confirming the claim.

B. We consider a standard normal random variable $z \sim \mathcal{N}(0, I)$, which has been transformed by a function $x = f(z)$. Given that the probability $p_\theta(\bar{A}) < 0.9 - \epsilon$, we compare this with the probability that the latent variable z lies within a radius r.

Since the function f is applied to z, the probability in the x-domain must be scaled appropriately by dividing by the Jacobian determinant of the transformation. This adjustment accounts for the change of variables from z to x.

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}r^2} > \frac{0.9 - \epsilon}{|\det J_f|},$$

where $r^2 = |z|^2$. Since $\log x \leq x - 1$, we can further derive:

$$r^2 = -2\log\left(\frac{(0.9 - \epsilon)2\pi}{|\det J_f|}\right) = 2\log\left(\frac{|\det J_f|}{(0.9 - \epsilon)2\pi}\right) \leq \frac{|\det J_f| \cdot e^{-1}}{(0.9 - \epsilon)2\pi}.$$

The volume of A is given by:

$$|A| = \frac{1}{|\det J_f|}\pi r^2.$$

Substituting the upper bound for r^2, we have:

$$|A| \leq \frac{1}{(0.9 - \epsilon)^2 \cdot e}.$$

Thus, the volume of the event C in the latent space is bounded above as shown, providing a simple comparison with A.

C. Define the difference in probabilities for event $\bar{A}$:

$$\delta(p_\theta, p)^2 = |P(\bar{A}) - P_\theta(\bar{A})|.$$

Substituting the probabilities, we have:

$$\delta(p_\theta, p)^2 = |A| \cdot 0.9 - P_\theta(\bar{A}) \geq |A| \cdot 0.9 - |A| \cdot (0.9 - \epsilon) = |A| \cdot \epsilon.$$

Since $|A| \leq 1$, the following inequality holds:

$$\delta(p_\theta, p)^2 \geq (1 - |A|) \cdot \epsilon.$$

Using an upper bound for $|A|$:

$$|A| \leq \frac{1}{(0.9 - \epsilon)^2},$$

we can further bound $\delta(p_\theta, p)^2$:

$$\delta(p_\theta, p)^2 \geq \left(1 - \frac{1}{(0.9 - \epsilon)^2}\right) \cdot \epsilon.$$

This expression is positive for some $\epsilon > 0$. Therefore:

$$D_{KL}(p \parallel q) \geq 2 \cdot \delta(p_\theta, p)^2 > 0 \quad \text{for some} \quad \epsilon > 0.$$

Thus, the KL divergence lower bound is positive under these conditions.

D. We found a positive lower bound for the KL divergence between p and any p_θ modeled by a VP-NF:

$$D_{KL}(p \parallel p_\theta) \geq 2 \cdot \delta(p_\theta, p)^2 > 0 \quad \text{for some} \quad \epsilon > 0.$$

Given that:

$$D_{KL}(p \parallel q) = 0 \iff p = q,$$

this implies that the model's distribution p_θ and the target distribution p can never converge to be the same.

This result arises because the transformation $x = f(z)$, dictated by the flow, imposes geometric constraints

Problem 3

Assuming that $D_\phi(x)$ has infinite capacity, we can use the calculus of variations to find the optimal function for $D_\phi(x)$.

The objective function is defined as:

$$L(\phi; \theta) = -\mathbb{E}_{x\sim p_{\text{data}}}[\log(D_\phi(x))] - \mathbb{E}_{x\sim p_\theta}[\log(1 - D_\phi(x))]$$

Simplifying this for a generic discriminator $D(x)$, we have:

$$L(D; \theta) = -\mathbb{E}_{x\sim p_{\text{data}}}[\log(D(x))] - \mathbb{E}_{x\sim p_\theta}[\log(1 - D(x))]$$

Rewriting as an integral form:

$$L(D; \theta) = -\int_x p_{\text{data}}(x) \log(D(x)) + p_\theta(x) \log(1 - D(x))\, dx$$

In order to find the optimal $D(x)$, we take the functional derivative of $L(D; \theta)$ with respect to $D(x)$:

$$-\int_x \delta(x) \left(\frac{p_{\text{data}}(x)}{D(x)} - \frac{p_\theta(x)}{1 - D(x)} \right) dx = 0 \quad \forall \delta(x)$$

This implies:

$$\frac{p_{\text{data}}(x)}{D^*(x)} - \frac{p_\theta(x)}{1 - D^*(x)} = 0$$

Rearranging, we find the optimal discriminator $D^*(x)$:

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)}$$

The optimal discriminator $D^*(x)$ represents the ratio of the true data distribution $p_{\text{data}}(x)$ to the total distribution, which is the sum of $p_{\text{data}}(x)$ and $p_\theta(x)$. This result forms the basis of generative adversarial networks (GANs) where the discriminator learns to distinguish between true data samples and generated samples.

Problem 4

The gradient of θ with respect to $L_G(\theta;\phi)$ can be derived using the chain rule, which is commonly referred to as backpropagation in machine learning:

$$\nabla_\theta L_G(\theta;\phi) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[\nabla_\theta \log(1-\sigma(h_\phi(G_\theta(z))))\right]$$

Expanding the gradient term:

$$= \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[\frac{1}{(1-\sigma(h_\phi(G_\theta(z))))}\cdot\sigma(h_\phi(G_\theta(z)))(1-\sigma(h_\phi(G_\theta(z))))\cdot\nabla_\theta h_\phi(G_\theta(z))\right]$$

Simplifying:

$$= \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[\sigma(h_\phi(G_\theta(z)))\cdot J_{h_\phi}\nabla_\theta G_\theta(z)\right]$$

A perfect discriminator can effectively distinguish between real samples and generated samples. For $x\sim p_\theta$, the discriminator output satisfies $D(x)\approx 0$. Equivalently, for z:

$$D(G(z)) = \sigma(h_\phi(G(z)))\approx 0 \quad\Rightarrow\quad h_\phi(G(z))\gg 0$$

Under these conditions, the gradient term for $L_G(\theta;\phi)$ becomes:

$$\nabla_\theta L_G(\theta;\phi) = \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[\sigma(h_\phi(G_\theta(z)))\cdot J_{h_\phi}\nabla_\theta G_\theta(z)\right]$$

Since $\sigma(h_\phi(G(z)))\approx 0$, the gradient vanishes:

$$\nabla_\theta L_G(\theta;\phi)\approx 0$$

When the discriminator is perfect, the generator's gradient $\nabla_\theta L_G(\theta;\phi)$ approaches zero. This highlights the vanishing gradient problem in adversarial training, where the generator struggles to improve when the discriminator becomes too effective.