In The Name of God

Sharif University of Technology

Electrical Engineering Department

Deep Generative Models

Assignment 3

Fall 2024

Instructor: Dr. S. Amini

Due on Azar 11, 1403 at 23:55

Matin M.babaei 400102114

1 Probability Review

Consider the transformation $f : \mathbb{R}^3 \to \mathbb{R}^3$.

$$X = f(Z) = [z_1 \quad z_2 e^{z_1} \quad (z_3 e^{-z_1} + z_1^2)^{\frac{1}{3}}]$$

Is this an invertible transformation? What's the Jacobian of this transformation? Now assume we have a Multivariate Normal base distribution $p(Z) \sim N_3(0, I)$ on the domain of f. What's $p(x)$ at $x = [0\ 1\ 1/3]^T$. ($p(X)$ and $p(Z)$ are probability density functions)

If invertible, there should be a $Z = f^{-1}(X)$.

Given Z, we have:

$$\begin{cases} x_1 = z_1 & \text{(i)} \\ x_2 = z_2 e^{z_1} & \text{(ii)} \\ x_3 = (z_3 e^{-z_1} + z_1^2)^{\frac{1}{3}} & \text{(iii)} \end{cases}$$

Now, if we are given X, can we obtain Z? We'll see,

$(x_1, x_2, x_3) \xrightarrow{\text{(i)}} z_1 = x_1 \xrightarrow{\text{(ii)}} z_2 = \frac{x_2}{e^{z_1}} = \frac{x_2}{e^{x_1}}$

$\xrightarrow{\text{(iii)}} x_3^3 = z_3 e^{-z_1} + z_1^2 \Rightarrow x_3^3 - x_1^2 = z_3 e^{-x_1} \rightarrow z_3 = \frac{x_3^3 - x_1^2}{e^{-x_1}}$

Since Z was thoroughly determined having X, we can introduce this inverse transformation.

$$Z = f^{-1}(X) = \begin{cases} z_1 = x_1 \\ z_2 = \frac{x_2}{e^{x_1}} = e^{-x_1} x_2 \\ z_3 = \frac{x_3^3 - x_1^2}{e^{-x_1}} = e^{x_1}(x_3^3 - x_1^2) \end{cases}$$

Thus, the transformation is invertible, using the function defined above.

$$J_f(z) = \frac{\partial x}{\partial z} = \begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} & \frac{\partial x_1}{\partial z_\mu} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} & \frac{\partial x_2}{\partial z_\mu} \\ \frac{\partial x_\mu}{\partial z_1} & \frac{\partial x_\mu}{\partial z_2} & \frac{\partial x_\mu}{\partial z_\mu} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ z_2 e^{z_1} & e^{z_1} & 0 \\ \frac{1}{\mu}\left(-z_2 e^{-z_1} + \mu z_1\right)\left(z_2 e^{-z_1} + z_1^2\right)^{-\frac{2}{\mu}} & 0 & \frac{1}{\mu}\left(e^{-z_1}\right)\left(z_2 e^{-z_1} + z_1^2\right)^{-\frac{2}{\mu}} \end{bmatrix}$$

- $x_1 = z_1$ - $x_1 = z_2 e^{z_1}$ - $x_\mu = \left(z_2 e^{-z_1} + z_1^2\right)^{\frac{1}{\mu}}$

$$\frac{\partial x_1}{\partial z_1} = 1 \quad \frac{\partial x_1}{\partial z_2} = \frac{\partial x_1}{\partial z_\mu} = 0 \quad \frac{\partial x_\mu}{\partial z_1} = z_2 e^{z_1} \quad \frac{\partial x_2}{\partial z_2} = e^{z_1} \quad \frac{\partial x_\mu}{\partial z_\mu} = 0$$

$$\frac{\partial x_2}{\partial z_1} = \frac{1}{\mu}\left(-z_2 e^{-z_1} + \mu z_1\right)\left(z_2 e^{-z_1} + z_1^2\right)^{-\frac{2}{\mu}} \quad \frac{\partial x_\mu}{\partial z_2} = 0 \quad \frac{\partial x_\mu}{\partial z_\mu} = \frac{1}{\mu}\left(e^{-z_1}\right)\left(z_2 e^{-z_1} + z_1^2\right)^{-\frac{2}{\mu}}$$

$p(x)$ @ $x = [0, 1, \frac{1}{\mu}]^T$:

$$p(x) = p(z) \cdot \left|\det\left(J_f(z)\right)\right|^{-1}$$

$$z_1 = x_1 = 0, \quad z_2 = \frac{x_2}{e^{z_1}} = 1, \quad z_\mu = e^{z_1}\left(x_\mu^\mu - z_1^2\right) = \left(\frac{1}{\mu}\right)^\mu = \frac{1}{2\mu}$$

$$p(z_\mu) = (2\pi)^{-\frac{d}{2}} \left|\Sigma\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(z-\mu)^T \Sigma^{-1} (z-\mu)\right)$$

$$d = 2, \ \Sigma = I, \ \mu = 0 : \quad \|z\|_\mu^2 = z_1^2 + z_2^2 + \dots + z_\mu^2 = 0 + 1 + \left(\frac{1}{2\mu}\right)^2$$

$$p(z_\mu) = \frac{1}{(2\pi)^{\frac{2}{2}} (1)^{\frac{1}{2}}} \exp\left(-\frac{1}{2} z^T z\right) = (2\pi)^{-\frac{2}{2}} \exp\left(-\frac{1}{2}\|z\|_\mu^2\right)$$

$$\rightarrow p(z) = (2\pi)^{-\frac{\mu}{2}} \exp\left(-\frac{1}{2}\left(\frac{2\mu}{2\mu}\right)\right) \qquad J_f(z)\Big|_{z} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -\frac{1}{9} & 0 & \mu \end{bmatrix}$$

$$\left|\det\left(J_f(z)\right)\right| = (1)(\mu - 0) = \mu$$

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

2 VP-NFs are not universal

Remember a Normalizing Flow formulated as:

$$p_\theta(\mathbf{x}) = p(z = f_\theta^{-1}(\mathbf{x})) \left| \det \frac{\partial f_\theta^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right|$$

In a Volume-Preserving Normalizing Flow (VP-NF), the determinant of the Jacobian matrix of each transformation is constant, specifically:

$$\left| \det \frac{\partial f_\theta(\mathbf{z})}{\partial \mathbf{z}} \right| = 1$$

This implies that the transformation preserves volume in the latent space, meaning the model can rearrange the probability mass but cannot scale it.

A set of probability distributions $\mathcal{P}$ is called a distributional universal approximator if for every possible target distribution $p(x)$ there is a sequence of distributions $p_n(x) \in \mathcal{P}$ such that $\lim_{n\to\infty} p_n(x) \sim p(x)$ with a chosen type of convergence.

We want to prove: The family of Normalizing Flows with constant Jacobian determinant $\left|\det \frac{\partial f_\theta(\mathbf{z})}{\partial \mathbf{z}}\right| = const$ with latent $z \sim \mathcal{N}(0,1)$ is not a universal distribution approximator under KL divergence. We do this by finding a lower bound for the KL objective. Use the Pinsker's inequality as below and the counter example given:

$$\delta(p_\theta, q) = \sup_{E\text{: a measurable event}} \{|p_\theta(E) - q(E)|\} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(p\|q)}$$

counter example:

$$p(x,y) = \begin{cases} 0.9 & \text{if } (x,y) \in [-0.5, 0.5] \times [-0.5, 0.5], \\ 0.9 - k \cdot (|x| - 0.5) & \text{if } |x| \in [0.5, 0.9k + 0.5] \text{ and } |y| \in [0, |x|], \\ 0.9 - k \cdot (|y| - 0.5) & \text{if } |y| \in [0.5, 0.9k + 0.5] \text{ and } |x| \in [0, |y|], \\ 0 & \text{otherwise.} \end{cases}$$

$$A = \{(x,y) \in \mathbb{R}^2 : p_\theta(x,y) \geq 0.9 - \epsilon\}$$
$$B = [-0.5, 0.5] \times [-0.5, 0.5]$$
$$\bar{A} = B \setminus A$$

2.1 Part (a)

Prove there exist an event E such that $|p_\theta(E) - p(E)| > 0$ when $A = \emptyset$, and show the lower bound for KL divergence can be positive for some $\epsilon > 0$, using Pinkster's inequality. (hint: prove $\bar{A}$ can be such E)

2.2 Part(b)

Using the change of variables formula, create the event C in latent space resulting from A. Compare its volume with A. Then compute its volume and provide a simple upper bound for it. (hint: volume here is a 2D circle because of the normal distribution)

2.3 Part(c)

Now show that there is an event E such that $|p_\theta(E) - p(E)| > 0$ when $A \neq \emptyset$, and show the lower bound for KL divergence can be positive for some $\epsilon > 0$, using Pinkster's inequality

2.4 Part(d)

Now briefly explain why this means VP-NFs cant approximate every distribution.

2.1)

Note that $A = \emptyset$, this means:

i) $\forall (x,y) \in B$: $P_\theta(x,y) < 0.9 - \epsilon$

ii) $\bar{A} = B \setminus A \rightarrow \bar{A} = B \setminus \emptyset = B$

So, using (i) $\bar{A} = [-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}] \rightarrow P(\bar{A}) = 0.9$ (Counter example)

and using (ii), $\forall (x,y) \in \mathbb{R}^2: P_\theta(x,y) < 0.9 - \epsilon \Rightarrow \forall (x,y) \in \bar{A}: P_\theta(\bar{A}) < 0.9 - \epsilon$

So taking $E := \bar{A}$, we would have:

$|P_\theta(E) - P(E)| > |0.9 - (0.9 - \epsilon)| = \epsilon \rightarrow$ So for $E := \bar{A}$, there is

an event which $|P_\theta(E) - P(E)| > 0$ when $A = \phi$ (Eq. $E = A$)

Now, let's see.

we know: $|P_\theta(E) - P(E)| > 0$

Also: $|P_\theta(E) - P(E)| < \delta(P_\theta, P) = \sup_E \{|P_\theta(E) - P(E)|\} \leq \sqrt{\frac{1}{2} D_{KL}(P_\theta \| P)}$

So we infer: $\delta(P_\theta, P) > \epsilon > 0$, $D_{KL} \geq 2\delta^2(P_\theta, P)$

2.2) $z \xrightarrow{f_\theta} x$: $P_\theta(x) = P_z(f_\theta^{-1}(x)) \cdot \frac{1}{|\det(J_f)|}$ (J_f symmetric)

Note that: $z \sim N(0, I)$

$\forall x \in A,\ P_\theta(x) \geq 0.9 - \epsilon \Rightarrow P_z(z) \frac{1}{|\det(J_f)|} \geq 0.9 - \epsilon$

$\rightarrow P_z(z) \geq |\det(J_f)|(0.9 - \epsilon) \rightarrow \log P_z(z) \geq \log|\det(J_f)| + \log(0.9 - \epsilon)$

$\rightarrow \log(\frac{1}{2\pi} \exp(-\frac{1}{2} z^T z)) = -\log(2\pi) - \frac{1}{2}\|z\|_2^2 \geq \log|\det J_f| + \log(0.9 - \epsilon)$

$\rightarrow \|z\|_2^2 \leq -2(\log(2\pi) + \log|\det J_f| + \log(0.9 - \epsilon)) = -2m$

بنابراین C شامل مجموعه نقاطی خواهد بود که شرط بالا در آن ها برقرار باشد. یعنی $m < 0$

Volume of $C := V(C)$

$\rightarrow m < 0: \quad V(C) = \pi \|z\|_2^2 \leq -2m\pi$

$V(A) = |\det J_f|\, V(C) \rightarrow V(A) \leq |\det J_f|(-2\pi m)$

$\rightarrow U(A) \leq |\det J_f|(-r_{min}\pi) = r_\pi |\det J_f| \left(\log \frac{1}{r_\pi |\det J_f| (1/9 - \epsilon)} \right)$

$\log(x) \leq \frac{x}{e} \;:\; \frac{1}{x} \log(x_{min}) \leq \frac{1}{e}$

put $x \leftarrow \frac{1}{r_\pi |\det J_f| (1/9 - \epsilon)}$ then:

$$r_\pi |\det J_f| (1/9 - \epsilon) \left(\log \frac{1}{r_\pi |\det J_f| (1/9 - \epsilon)} \right) \leq \frac{1}{e}$$

So:

$$U(A) \leq r_\pi |\det J_f| \left(\log \frac{1}{r_\pi |\det J_f| (1/9 - \epsilon)} \right) \leq \frac{1}{e} \frac{1}{1/9 - \epsilon}$$

$$\Rightarrow U(C) \leq \frac{1}{|\det J_f|} \frac{1}{e} \frac{1}{1/9 - \epsilon}$$

2.3)

$A \neq \emptyset$:

i) $P_\theta(A) = U(A)(1/9 - \epsilon) \overset{(2.2)}{\leq} (1/9 - \epsilon) \frac{1}{(1/9 - \epsilon)\epsilon} = \frac{1}{e} \simeq 1/3$

ii) $P(A) \geq P(B) = 1/9$

So $|P_\theta(A) - P(A)| > 1/9 - \frac{1}{e} \simeq 1/20$

Assume $E = A$:

$$\delta(P_\theta, P) = \sup_E \{|P_\theta(A) - P(A)|\} = 1/9 - \frac{1}{e} \simeq 1/20$$

Using Pinsker's Inequality: So $|P_\theta(A) - P(A)| > \epsilon > 0$

$$\delta(P_\theta, P) \leq \sqrt{\frac{1}{2} D_{KL}(P, P_\theta)} \rightarrow D_{KL}(P, P_\theta) \geq 2\delta^2(P, P_\theta) = 2(1/9 - \frac{1}{e})^2$$

2.4) we saw either $A = \emptyset$ || $A = \Omega$, $D_{KL}(P, P_\theta) \geq 2\delta^2(P, P_\theta) > 0$

And $\delta(P, P_\theta) > 0 \longrightarrow$ UP-NFs can not approximate every distribution.

3 Optimality of GAN Framework

In class, we covered that the optimal discriminator satisfies the following condition

$$D_G^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \tag{1}$$

for arbitrary values of $\boldsymbol{x}$. In this problem, you will prove the optimality of this discriminator. To be more specific, for a fixed generator θ, show that $D_\phi^*(\boldsymbol{x})$ minimizes the following loss function:

$$\mathcal{L}(\phi;\theta) = -\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log D_\phi(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}\left[\log\left(1 - D_\phi(\boldsymbol{x})\right)\right] \tag{2}$$

---

we only need to consider what happens at each input x pointwise, since our discriminator is perfectly expressive.

The relevant part of the objective for the discriminator value at a particular x is:

$$\mathcal{L}(\phi;\theta) = \min_{D(x)} -\left(p_{data}(x)\log D_\phi(x) + p_\theta(x)\log(1 - D_\phi(x))\right)$$

Let's optimize this:

$$a = \frac{\partial \mathcal{L}}{\partial D_\phi(x)} = -p_{data}(x)\frac{1}{D_\phi(x)} + p_\theta(x)\cdot\frac{1}{1 - D_\phi(x)}$$

$$\Rightarrow \quad p_\theta(x)\, D_\phi^*(x) = p_{data}(x)\left(1 - D_\phi^*(x)\right)$$

$$\Rightarrow \quad \boxed{D_\phi^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_\theta(x)}}$$

$$\frac{\partial}{\partial D_\phi(x)}\mathcal{L}(\phi,\theta) = -\mathop{E}_{x\sim p_{data}}\left[\frac{1}{D_\phi(x)}\right] - \mathop{E}_{x\sim p_\theta(x)}\left[\frac{1}{1 - D_\phi(x)}\right]$$

$$\frac{\partial^2}{\partial D_\phi(x)^2}\mathcal{L}(\phi,\theta) = -\mathop{E}_{x\sim p_{data}}\left[\frac{1}{D_\phi'(x)}\right] - \mathop{E}_{x\sim p_\theta}\left[-\frac{1}{(1 - D_\phi(x))'}\right]$$

$$= \mathop{E}_{x\sim p_{data}}\left[\frac{1}{D_\phi'(x)}\right] + \mathop{E}_{x\sim p_\theta}\left[\frac{1}{(1 - D_\phi(x))'}\right] > 0$$

$\longrightarrow$ $\mathcal{L}(\phi,\theta)$ is convex w.r.t D_ϕ, thus D_ϕ^* is the global minima.

$$\mathcal{L}(D_\phi^*,\theta) < \mathcal{L}(D_\phi,\theta)$$

4 Perfect Discriminator, Perfect Generation

Recall that when training GAN, the gradient signals for θ come from the term

$$\mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}\left[\log\left(1-D_\phi(\boldsymbol{x})\right)\right], \tag{3}$$

which can also be written as

$$L_G(\theta;\phi)=\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\log\left(1-\sigma\left(h_\phi\left(G_\theta(\boldsymbol{z})\right)\right)\right)\right], \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $h_\phi(\cdot)$ denotes the logits, and G_θ is the generator. Now assume that our discriminator is perfect, i.e. $\nabla_\theta L_G(\theta;\phi)\to 0$, show that the gradient of θ would vanish to zero in this case, i.e. $\nabla_\theta L_G(\theta;\phi)\to 0$.

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

Let's compute gradient:

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

we shall use Chain Rule:

$\nabla_\theta \log\left(1-\sigma(h_\phi(G_\theta(z)))\right) = \dfrac{\nabla_\theta\left(1-\sigma(h_\phi(G_\theta(z)))\right)}{1-\sigma(h_\phi(G_\theta(z)))}$

$\nabla_\theta\left(1-\sigma(h_\phi(G_\theta(z)))\right) = -\nabla_\theta\,\sigma(h_\phi(G_\theta(z)))$

For Sigmoid, Gradient is: $\nabla\sigma = \sigma(1-\sigma)$

$\nabla_\theta\,\sigma(h_\phi(G_\theta(z))) = \sigma(h_\phi(G_\theta(z)))\left(1-\sigma(h_\phi(G_\theta(z)))\right)\nabla_\theta h_\phi(G_\theta(z))$

So we have: $\nabla_\theta \log\left(1-\sigma(h_\phi(G_\theta(z)))\right) = -\dfrac{\sigma(h_\phi(G_\theta(z)))\left(1-\sigma(h_\phi(G_\theta(z)))\right)}{1-\sigma(h_\phi(G_\theta(z)))}\nabla_\theta h_\phi(G_\theta(z))$ $(<\infty)$

So: $\nabla_\theta L_G(\theta;\phi) = -\mathbb{E}_{z\sim(\cdot,I)}\left[\sigma(h_\phi(G_\theta(z)))\,\nabla_\theta h_\phi(G_\theta(z))\right]$

If the discriminator is perfect, we are told that: $\sigma(h_\phi(G_\theta(z)))\to 0$

So, in spite of $\nabla_\theta h_\phi(G_\theta(z))$, when $\sigma(h_\phi(G_\theta(z)))\to 0$, the entire expression vanishes: $\nabla_\theta L_G(\theta;\phi)\to 0$