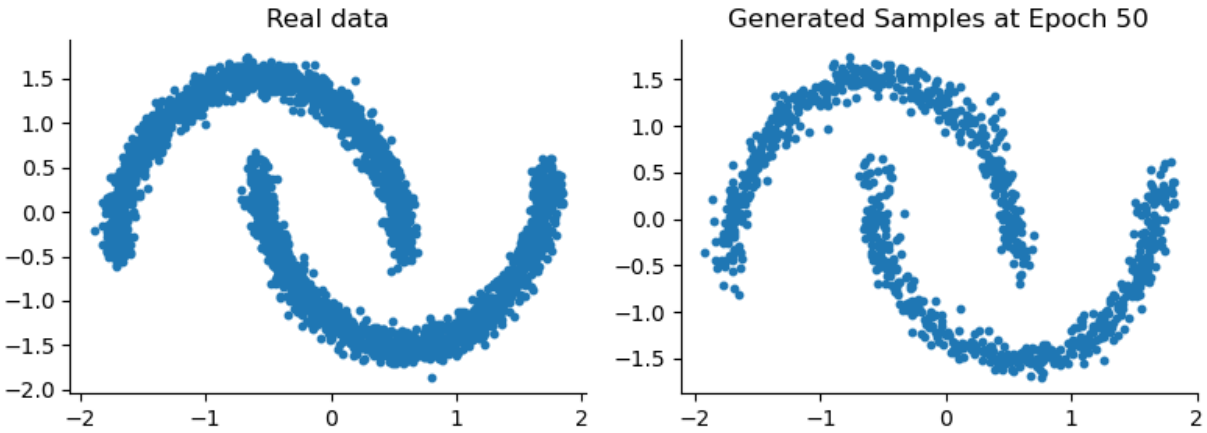
CS236: Deep Generative Models Homework #3 Solution

(Fall 2021)

Shubham Shrivastava email: shubhams@stanford.edu

Problem 1: Flow models

4. Visualize 1000 samples drawn the model after is has been trained.



Problem 2: Generative adversarial networks

Q2) 1) The minmax loss is given as:

$$L_G^{minmax}(\theta;\phi) = E_{z\sim N(0,I)}[\log(1 - D_\phi(G_\theta(z)))]$$

where, $G_\phi(x) = \sigma(h_\phi(x))$

and, $\sigma'(x) = \sigma(x)(1-\sigma(x))$

Let's compute the derivative of minmax loss w.r.t. θ

$$\frac{\partial}{\partial\theta} L_G^{minmax}(\theta;\phi) = E_{z\sim N(0,I)}\left[\frac{\partial}{\partial\theta}\log(1-\sigma(h_\phi(G_\theta(z))))\right]$$

$$= E_{z\sim N(0,I)}\left[\frac{1}{(1-\sigma(h_\phi(G_\theta(z))))}\frac{\partial}{\partial\theta}(1-\sigma(h_\phi(G_\theta(z))))\right]$$

$$= E_{z\sim N(0,I)}\left[\frac{-\sigma(h_\phi(G_\theta(z)))(1-\sigma(h_\phi(G_\theta(z))))}{(1-\sigma(h_\phi(G_\theta(z))))}\frac{\partial}{\partial\theta}h_\phi(G_\theta(z))\right]$$

$$= E_{z\sim N(0,I)}\left[-\sigma(h_\phi(G_\theta(z)))\frac{\partial}{\partial\theta}h_\phi(G_\theta(z))\right]$$

When discriminator successfully identifies a fake sample $G_\theta(z)$, its output $\sigma(h_\phi(G_\theta(z))) \approx 0$, which causes the derivative of the loss function to be very close to

zero (i.e. vanishing gradient). This stops the network from learning anything.

2. Implement and train a non-saturating GAN on Fashion MNIST for one epoch.

Problem 3: Divergence minimization

Q3) 1) From the given hint:

$$f(t) = -p_{data}(x)\log t - p_\theta(x)\log(1-t) \quad \text{(i)}$$

In order to find the t which minimizes $f(t)$, we differentiate $f(t)$ with respect to 't' and then set it to zero. i.e.

$$\frac{df(t)}{dt} = 0$$

$$\Rightarrow \frac{-p_{data}(x)}{t} - \frac{p_\theta(x)}{(1-t)}(-1) = 0$$

$$\Rightarrow \frac{-(1-t)p_{data}(x) + t\,p_\theta(x)}{t(1-t)} = 0$$

$$\Rightarrow -p_{data}(x) + t\left(p_{data}(x) + p_\theta(x)\right) = 0$$

$$\Rightarrow t = \frac{p_{data}(x)}{p_{data}(x) + p_\theta(x)}$$

Hence, for $f = D_\phi(x)$, the function L_D is minimized when $D_\phi = D^*$

where, $$D^* = \frac{p_{data}(x)}{p_\theta(x) + p_{data}(x)} \quad \text{— (i)}$$

2) $$D_\phi(x) = \sigma(h_\phi(x))$$

where, $$\sigma(h_\phi(x)) = \frac{1}{1 + e^{-h_\phi(x)}}$$

$$\Rightarrow D_\phi(x) = \frac{1}{1 + e^{-h_\phi(x)}} \quad \text{— (ii)}$$

Using equation (i), when $D_\phi(x) = D^*(x)$, eqn (ii) becomes

$$\frac{p_{data}(x)}{p_\theta(x) + p_{data}(x)} = \frac{1}{1 + e^{-h_\phi(x)}}$$

$$\Rightarrow e^{-h_\phi(x)} = \frac{p_\theta(x) + p_{data}(x)}{p_{data}(x)} - 1$$

$$\Rightarrow \quad e^{-h_\phi(x)} = \frac{p_\theta(x)}{p_{data}(x)} + \frac{p_{data}(x)}{p_{data}(x)} - 1 = \frac{p_\theta(x)}{p_{data}(x)}$$

Taking log on both sides of eq^{n} (iii):

$$-h_\phi(x) = \log \frac{p_\theta(x)}{p_{data}(x)} = -\log \frac{p_{data}(x)}{p_\theta(x)}$$

$$\Rightarrow \quad h_\phi(x) = \log \frac{p_{data}(x)}{p_\theta(x)}$$

3)

Generator Loss:

$$L_G(\theta;\phi) = E_{x\sim p_\theta(x)}\left[\log\left(1 - D_\phi(x)\right)\right]$$

$$- E_{x\sim p_\theta(x)}\left[\log D_\phi(x)\right]$$

$$= E_{x\sim p_\theta(x)}\left[\log \frac{\left(1 - D_\phi(x)\right)}{D_\phi(x)}\right] \quad \text{— (i)}$$

when, $D_\phi(x) = D^*(x)$, $D_\phi(x) = \dfrac{p_{data}(x)}{p_\theta(x) + p_{data}(x)}$ — (ii)

Substituting eqn (ii) in (i),

$$L_G(\theta;\phi) = E_{x\sim p_\theta(x)}\left[\log \frac{1 - \left(\frac{p_{data}(x)}{p_\theta(x)+p_{data}(x)}\right)}{\left(\frac{p_{data}(x)}{p_\theta(x)+p_{data}(x)}\right)}\right]$$

$$= E_{x\sim p_\theta(x)}\left[\log \frac{\left(\frac{p_\theta(x)}{p_\theta(x)+p_{data}(x)}\right)}{\left(\frac{p_{data}(x)}{p_\theta(x)+p_{data}(x)}\right)}\right]$$

$$= E_{x\sim p_\theta(x)}\left[\log \frac{p_\theta(x)}{p_{data}(x)}\right]$$

$$\Rightarrow L_G(\theta;\phi) = KL\left(p_\theta(x) \,||\, p_{data}(x)\right)$$

4) The negative log-likelihood can be rewritten as

$$-E_{x\sim p_{data}(x)}\left[\log p_\theta(x)\right] = -E_{x\sim p_{data}(x)}\left[\log p_\theta(x)\right] + E_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right] - E_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right]$$

$$= E_{x\sim p_{data}(x)}\left[\log p_{data}(x) - \log p_\theta(x)\right] - E_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right]$$

$$= E_{x\sim p_{data}}\left[\log \frac{p_{data}(x)}{p_\theta(x)}\right] - E_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right]$$

$$= KL\left(p_{data}(x) \,\|\, p_\theta(x)\right) - \underbrace{E_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right]}_{\text{Constant with respect to } \theta}$$

The GAN generator objective in the previous part is $L_G(\theta;\phi) = KL(p_\theta(x) \,||\, p_{data}(x))$ and we know that the KL-divergence in general is non-symmetric.

i.e. $KL(p_\theta(x) \,||\, p_{data}(x)) \neq KL(p_{data}(x) \,||\, p_\theta(x))$

Hence, the VAE decoder trained with negative ELBO and a GAN generator trained with L_G are not learning the same objective.

Problem 4: Conditional GAN with projection discriminator

Q4) 1) From problem 3.2, we get

$$h^*_\phi(x,y) = \log \frac{p_{data}(x,y)}{p_\theta(x,y)}$$

$$= \log \frac{p_{data}(x|y)\, p_{data}(y)}{p_\theta(x|y)\, p_\theta(y)}$$

$$= \log \frac{p_{data}(x|y)}{p_\theta(x|y)} + \log \frac{p_{data}(y)}{p_\theta(y)} \quad \text{— (i)}$$

For simplicity, we assume that $p_{data}(x) = p_\theta(x) = \frac{1}{m}$

$$\Rightarrow \log \frac{p_{data}(y)}{p_\theta(y)} = 0$$

Equation (i) then becomes:

$$h_\theta(x, y) = \log \frac{p_{data}(x|y)}{p_\theta(x|y)}$$

$$= \log \frac{\mathcal{N}(\varphi(x) | \mu_y, I)}{\mathcal{N}(\varphi(x) | \hat{\mu}_y, I)} \quad \text{— (ii)}$$

We know that,

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

for $\Sigma = I$, we get,

$$\mathcal{N}(\varphi(x) | \mu_y, I) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} \|\varphi(x) - \mu_y\|^2\right) \quad \text{— (iii)}$$

and,

$$\mathcal{N}(\varphi(x) | \hat{\mu}_y, I) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} \|\varphi(x) - \hat{\mu}_y\|^2\right) \quad \text{— (iv)}$$

Substituting equations (iii) and (iv) in (ii)

$$h^*_\phi(x,y) = \log \frac{\exp\left(-\frac{1}{2}\|\varphi(x) - \mu_y\|^2\right)}{\exp\left(-\frac{1}{2}\|\varphi(x) - \hat{\mu}_y\|^2\right)}$$

$$= -\frac{1}{2}\|\varphi(x) - \mu_y\|^2 + \frac{1}{2}\|\varphi(x) - \hat{\mu}_y\|$$

$$= -\frac{1}{2}\|\varphi(x)\|^2 - \frac{1}{2}\|\mu_y\|^2 - \mu_y^T\varphi(x)$$

$$+ \frac{1}{2}\|\varphi(x)\|^2 + \frac{1}{2}\|\hat{\mu}_y\|^2 + \hat{\mu}_y^T\varphi(x)$$

$$= \frac{1}{2}\left(\|\hat{\mu}_y\|^2 - \|\mu_y\|^2\right) + \left(\hat{\mu}_y^T - \mu_y^T\right)\varphi(x) \quad \text{(v)}$$

Let y^T be a one-hot vector denoting the class y, then $(\hat{\mu}_y^T - \mu_y^T)$ can be rewritten as:

$$(\hat{\mu}_y^T - \mu_y^T) = y^T \begin{bmatrix} \hat{\mu}_1^T - \mu_1^T \\ \hat{\mu}_2^T - \mu_1^T \\ \vdots \\ \hat{\mu}_y^T - \mu_y^T \\ \vdots \\ \hat{\mu}_m - \mu_m^T \end{bmatrix}$$

$\|\hat{\mu}_y\|^2 - \|\mu_y\|^2$ can also be rewritten as:

$$\|\hat{\mu}_y\|^2 - \|\mu_y\|^2 = y^T \begin{bmatrix} \|\hat{\mu}_1\|^2 - \|\mu_1\|^2 \\ \|\hat{\mu}_2\|^2 - \|\mu_2\|^2 \\ \vdots \\ \|\hat{\mu}_y\|^2 - \|\mu_y\|^2 \\ \vdots \\ \|\hat{\mu}_m\|^2 - \|\mu_m\|^2 \end{bmatrix}$$

Hence, equation (i) can be rewritten as

$$h_\phi^*(x,y) = y^T \left(\underbrace{\begin{bmatrix} \hat{\mu}_1^T - \mu_1^T \\ \hat{\mu}_2^T - \mu_2^T \\ \vdots \\ \hat{\mu}_y^T - \mu_y^T \\ \vdots \\ \hat{\mu}_m - \mu_m^T \end{bmatrix}}_{A} \varphi(x) + \underbrace{\begin{bmatrix} \|\hat{\mu}_1\|^2 - \|\mu_1\|^2 \\ \|\hat{\mu}_2\|^2 - \|\mu_2\|^2 \\ \vdots \\ \|\hat{\mu}_y\|^2 - \|\mu_y\|^2 \\ \vdots \\ \|\hat{\mu}_m\|^2 - \|\mu_m\|^2 \end{bmatrix}}_{b} \right)$$

$$\Rightarrow h_\phi^*(x,y) = y^T \left(A\,\varphi(x) + b \right)$$

2. Implement and train a conditional GAN on Fashion MNIST for one epoch.

Problem 5: Wasserstein GAN

Q5) 1) $p_\theta(x) = \mathcal{N}(x \mid \theta, \varepsilon^2)$

$p_{data}(x) = \mathcal{N}(x \mid \theta_0, \varepsilon^2)$

$$\Rightarrow KL(p_\theta(x) \,\|\, p_{data}(x)) = E_{x \sim p_\theta(x)}\left[\log \frac{p_\theta(x)}{p_{data}(x)}\right]$$

$$= E_{x \sim \mathcal{N}(\theta, \varepsilon^2)}\left[\log \frac{\mathcal{N}(x \mid \theta, \varepsilon^2)}{\mathcal{N}(x \mid \theta_0, \varepsilon^2)}\right]$$

we know that,

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$\Rightarrow KL(p_\theta(x) \,\|\, p_{data}(x)) =$$

$$E_{x \sim \mathcal{N}(\theta, \varepsilon^2)}\left[\log \frac{\frac{1}{(2\pi)^{m/2} \varepsilon^{1/2}} \exp\left(-\frac{1}{2\varepsilon^2}(x-\theta)^2\right)}{\frac{1}{(2\pi)^{m/2} \varepsilon^{1/2}} \exp\left(-\frac{1}{2\varepsilon^2}(x-\theta_0)^2\right)}\right]$$

$$= E_{x \sim \mathcal{N}(\theta, \varepsilon^2)}\left[\log \frac{\exp\left(-\frac{1}{2\varepsilon^2}(x-\theta)^2\right)}{\exp\left(-\frac{1}{2\varepsilon^2}(x-\theta_0)^2\right)}\right]$$

$$= E_{x\sim N(\theta,\varepsilon^2)}\left[-\frac{1}{2\varepsilon^2}(x-\theta)^2+\frac{1}{2\varepsilon^2}(x-\theta_0)^2\right]$$

$$= E_{x\sim N(\theta,\varepsilon^2)}\left[\frac{1}{2\varepsilon^2}\left(-x^2-\theta^2+2x\theta+x^2+\theta_0^2-2x\theta_0\right)\right]$$

$$= E_{x\sim N(\theta,\varepsilon^2)}\left[\frac{1}{2\varepsilon^2}\left(\theta_0^2-\theta^2+2x(\theta-\theta_0)\right)\right]$$

$$= \frac{1}{2\varepsilon^2}(\theta_0^2-\theta^2)+\frac{1}{2\varepsilon^2}E_{x\sim N(\theta,\varepsilon^2)}\left[2x(\theta-\theta_0)\right]$$

$$= \frac{1}{2\varepsilon^2}\left(\theta_0^2-\theta^2+2\theta(\theta-\theta_0)\right)$$

$$= \frac{1}{2\varepsilon^2}\left(\theta_0^2-\theta^2+2\theta^2-2\theta\theta_0\right)$$

$$= \frac{1}{2\varepsilon^2}(\theta-\theta_0)^2$$

$$\Rightarrow KL(p_\theta(x)\,\|\,p_{data}(x)) = \frac{(\theta-\theta_0)^2}{2\varepsilon^2}$$

2) If $\theta \neq \theta_0$ and $\varepsilon \to 0$, the KL-divergence and its derivative will go to infinity. This will result in a very unstable generator training since the gradients obtained by the optimizer will be large.

3) The loss function, $L_D(\phi; \theta)$, is unbounded.
As $\varepsilon \to 0$, if $\theta \neq \theta_0$, there will not be an overlap between the distributions.
In the process of minimizing L_D, the optimizer can set $D_\phi(\theta) = -\infty$ and $D_\phi(\theta_0) = \infty$, thereby L_D approaching $-\infty$. Hence, there is no discriminator D_ϕ that minimizes this new objective.

4) $L_D = D_\phi(\theta) - D_\phi(\theta_0)$
derivative of D_ϕ (same as the slope) is restricted to be between -1 and $+1$
When D_ϕ is set to be either -1 or $+1$, that is when we achieve the smallest $D_\phi(\theta)$
This can be done by connecting two points at (θ_0, k) & $(\theta, k - |\theta - \theta_0|)$

5. Implement and train WGAN-GP for one epoch on Fashion MNIST.