CS 236 Autumn 2019/2020 Homework 3

SUNet ID: 06009508

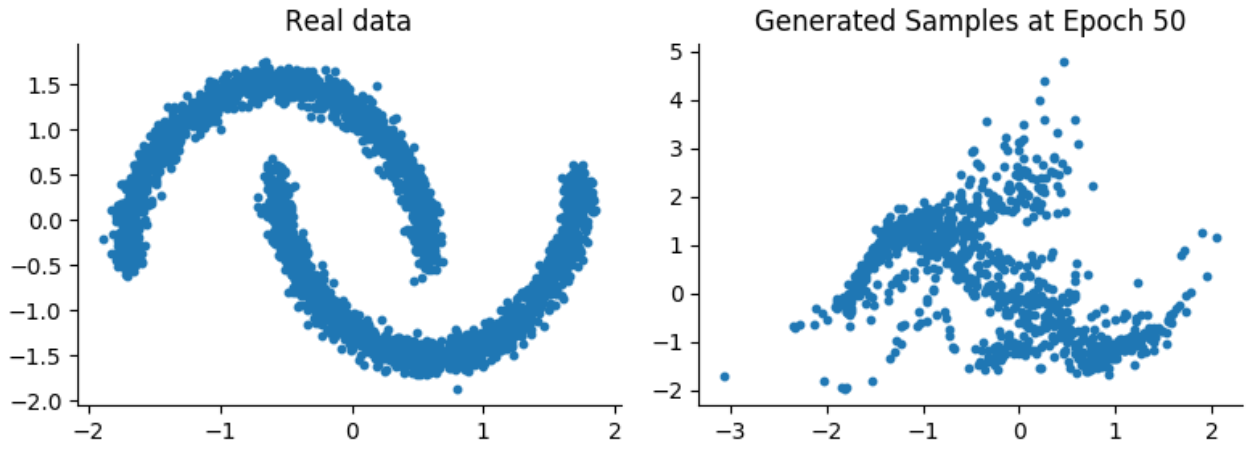
Name: Brandon McKinzie

Collaborators: N/A

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 2: Generative adversarial networks (15 points)

Part 4.



Problem 2: Generative adversarial networks (15 points)

Unfortunately, this form of loss for L_G suffers from a vanishing gradient *problem. In terms of the discriminator's logits, the minimax loss is*

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

Show that the derivative of $L_G^{minimax}$ with respect to θ is approximately 0 if $D(G_\theta(\boldsymbol{z})) \approx 0$, or equivalently, if $h_\phi(G_\theta(\boldsymbol{z})) \ll 0$. You may use the fact that $\sigma'(x) = \sigma(x)(1-\sigma(x))$. Why is this problematic for the training of the generator when the discriminator successfully identifies a fake sample $G_\theta(\boldsymbol{z})$?

Below, I take the derivative with respect to θ and evaluate it for the case of $\sigma(h_\phi(G_\theta(\boldsymbol{z}))) \approx 0$.

$$\frac{\partial L_G^{\text{minimax}}(\theta;\phi)}{\partial\theta} = \frac{\partial}{\partial\theta}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I)}[\log(1-\sigma(h_\phi(G_\theta(\boldsymbol{z}))))] \tag{1}$$

$$= \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I)}[\frac{\partial}{\partial\theta}\log(1-\sigma(h_\phi(G_\theta(\boldsymbol{z}))))] \tag{2}$$

$$\frac{\partial}{\partial\theta}\log(1-\sigma(h_\phi(G_\theta(\boldsymbol{z})))) = \frac{1}{1-\sigma(h_\phi(G_\theta(\boldsymbol{z})))}\frac{\partial}{\partial\theta}(1-\sigma(h_\phi(G_\theta(\boldsymbol{z})))) \tag{3}$$

$$= -\frac{1}{1-\sigma(h_\phi(G_\theta(\boldsymbol{z})))}\frac{\partial}{\partial\theta}\sigma(h_\phi(G_\theta(\boldsymbol{z}))) \tag{4}$$

$$\frac{\partial}{\partial\theta}\sigma(h_\phi(G_\theta(\boldsymbol{z}))) = \sigma(h_\phi(G_\theta(\boldsymbol{z})))(1-\sigma(h_\phi(G_\theta(\boldsymbol{z}))))\frac{\partial}{\partial\theta}h_\phi(G_\theta(\boldsymbol{z})) \tag{5}$$

$$\therefore \left.\frac{\partial L_G^{\text{minimax}}(\theta;\phi)}{\partial\theta}\right|_{h_\phi<<0} = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I)}[-\frac{\sigma(h_\phi(G_\theta(\boldsymbol{z})))(1-\sigma(h_\phi(G_\theta(\boldsymbol{z}))))\frac{\partial}{\partial\theta}h_\phi(G_\theta(\boldsymbol{z}))}{1-\sigma(h_\phi(G_\theta(\boldsymbol{z})))}] \tag{6}$$

$$= \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,I)}[-\sigma(h_\phi(G_\theta(\boldsymbol{z})))\frac{\partial}{\partial\theta}h_\phi(G_\theta(\boldsymbol{z}))] \tag{7}$$

$$\approx 0 \tag{8}$$

This is problematic since it suggests that if our generator G_θ is horrible – that is, it produces obviously fake samples from the perspective of our discriminator (or, equivalently, if our discriminator is perfect) – then its gradients will be (nearly) zero and thus its parameters won't get updated.

Problem 3: Divergence minimization (25 points)

Part 1. *Show that L_D is minimized when $D_\phi = D^*$, where*

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

(Hint: for a fixed $\boldsymbol{x}$, what t minimizes $f(t) = -p_{data}(\boldsymbol{x}) \log t - p_\theta(\boldsymbol{x}) \log(1-t)$?)

First we expand out the equation for L_D as an integral:

$$L_D(\phi; \theta) = -\int \left[p_{data}(\boldsymbol{x}) \log D_\phi(\boldsymbol{x}) + p_\theta(\boldsymbol{x}) \log(1 - D_\phi(\boldsymbol{x}))\right] \mathrm{d}\boldsymbol{x} \tag{9}$$

The derivative of the integrand with respect to D is

$$p_{data}(\boldsymbol{x}) \frac{1}{D_\phi(\boldsymbol{x})} - p_\theta(\boldsymbol{x}) \frac{1}{1 - D_\phi(\boldsymbol{x})} \tag{10}$$

We can set this to zero and solve for D to obtain D^*.

$$p_{data}(\boldsymbol{x}) \frac{1}{D_\phi(\boldsymbol{x})} = p_\theta(\boldsymbol{x}) \frac{1}{1 - D_\phi(\boldsymbol{x})} \tag{11}$$

$$p_{data}(\boldsymbol{x})(1 - D_\phi(\boldsymbol{x})) = p_\theta(\boldsymbol{x}) D_\phi(\boldsymbol{x}) \tag{12}$$

$$p_{data}(\boldsymbol{x}) = D_\phi(\boldsymbol{x})(p_\theta(\boldsymbol{x}) + p_{data}(\boldsymbol{x})) \tag{13}$$

$$D_\phi(\boldsymbol{x}) = \frac{p_{data}(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p_{data}(\boldsymbol{x})} \tag{14}$$

Part 2. *Recall that $D_\phi(\boldsymbol{x}) = \sigma(h_\phi(\boldsymbol{x}))$. Show that the logits $h_\phi(\boldsymbol{x})$ of the discriminator estimate the log of the likelihood ratio of $\boldsymbol{x}$ under the true distribution compared to the model's distribution; that is, show that if $D_\phi = D^*$, then*

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

We just plug in $D_\phi(\boldsymbol{x}) = \sigma(h_\phi(\boldsymbol{x}))$ and solve for $h_\phi(\boldsymbol{x})$ using basic arithmetic:

$$\sigma(h_\phi(\boldsymbol{x})) = \frac{p_{data}(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p_{data}(\boldsymbol{x})} \tag{15}$$

$$\text{Let } F := \frac{p_{data}(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p_{data}(\boldsymbol{x})} \tag{16}$$

$$\frac{1}{1 + e^{-h_\phi(\boldsymbol{x})}} = F \tag{17}$$

$$1 = (1 + e^{-h_\phi(\boldsymbol{x})})F \tag{18}$$

$$= F + Fe^{-h_\phi(\boldsymbol{x})} \tag{19}$$

$$\frac{1 - F}{F} = e^{-h_\phi(\boldsymbol{x})} \tag{20}$$

$$\log \frac{1 - F}{F} = -h_\phi(\boldsymbol{x}) \tag{21}$$

$$\therefore h_\phi(\boldsymbol{x}) = \log \frac{F}{1 - F} \tag{22}$$

$$= \log \frac{p_{\text{data}}(\boldsymbol{x})}{p_\theta(\boldsymbol{x})} \tag{23}$$

Part 3. *Consider a generator loss defined by the sum of the minimax loss and the non-saturating loss,*

$$L_G(\theta;\phi) = \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}[\log(1 - D_\phi(\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}[\log D_\phi(\boldsymbol{x})].$$

Show that if $D_\phi = D^*$*, then*

$$L_G(\theta;\phi) = KL(p_\theta(\boldsymbol{x})||p_{data}(\boldsymbol{x})).$$

Again, we just plug in for D^* and shuffle some terms around to arrive at the result.

$$L_G(\theta;\phi) = \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}[\log(\frac{p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p_{data}(\boldsymbol{x})})] - \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}[\log \frac{p_{data}(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p_{data}(\boldsymbol{x})}] \tag{24}$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}[\log \frac{p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p_{data}(\boldsymbol{x})} / \frac{p_{data}(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p_{data}(\boldsymbol{x})}] \tag{25}$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}[\log \frac{p_\theta(\boldsymbol{x})}{p_{data}(\boldsymbol{x})}] \tag{26}$$

$$= \mathrm{KL}(p_\theta(\boldsymbol{x})||p_{\mathrm{data}}(\boldsymbol{x})) \tag{27}$$

Part 4. *Recall that when training VAEs, we minimize the negative ELBO, an upper bound to the negative log likelihood. Show that the negative log likelihood, $-\mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$, can be written as a KL divergence plus an additional term that is constant with respect to θ.*

We can use the definition of $KL(p_{data}(x)||p_\theta(x))$ and rearrange some terms:

$$KL(p_{data}(x)||p_\theta(x)) \triangleq \mathbb{E}_{x\sim p_{data}(x)}\left[\log \frac{p_{data}(x)}{p_\theta(x)}\right] \tag{28}$$

$$= \mathbb{E}_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right] - \mathbb{E}_{x\sim p_{data}(x)}\left[\log p_\theta(x)\right] \tag{29}$$

$$\therefore -\mathbb{E}_{x\sim p_{data}(x)}\left[\log p_\theta(x)\right] = KL(p_{data}(x)||p_\theta(x)) - \mathbb{E}_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right] \tag{30}$$

and we see that $-\mathbb{E}_{x\sim p_{data}(x)}\left[\log p_\theta(x)\right]$ can be written as $KL(p_{data}(x)||p_\theta(x))$ plus a θ-independent term, $-\mathbb{E}_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right]$. This of course implies that

$$\min_\theta -\mathbb{E}_{x\sim p_{data}(x)}\left[\log p_\theta(x)\right] = \min_\theta \Big[KL(p_{data}(x)||p_\theta(x)) - \mathbb{E}_{x\sim p_{data}(x)}\left[\log p_{data}(x)\right]\Big] \tag{31}$$

$$= \min_\theta KL(p_{data}(x)||p_\theta(x)) \tag{32}$$

Does this mean that a VAE decoder trained with ELBO and a GAN generator trained with the L_G defined in the previous part are implicitly learning the same objective? Explain.

No, they are not the same, since the GAN generator from the previous part minimizes

$$L_G(\theta;\phi) = KL(p_\theta(x)||p_{data}(x)) \tag{33}$$

and $KL(p_\theta(x)||p_{data}(x)) \neq KL(p_{data}(x)||p_\theta(x))$ since the KL divergence is not symmetric.

Problem 4: Conditional GAN with projection discriminator (20 points)

Part 1. *Suppose that when $(\boldsymbol{x}, y) \sim p_{data}(\boldsymbol{x}, y)$, there exists a feature mapping φ under which $\varphi(\boldsymbol{x})$ becomes a mixture of m unit Gaussians, with one Gaussian per class label y. Assume that when $(\boldsymbol{x}, y) \sim p_\theta(\boldsymbol{x}, y)$, $\varphi(\boldsymbol{x})$ also becomes a mixture of m unit Gaussians, again with one Gaussian per class label y. Concretely, we assume that the ratio of the conditional probabilities can be written as*

$$\frac{p_{data}(\boldsymbol{x}|y)}{p_\theta(\boldsymbol{x}|y)} = \frac{\mathcal{N}(\varphi(\boldsymbol{x})|\boldsymbol{\mu}_y, I)}{\mathcal{N}(\varphi(\boldsymbol{x})|\hat{\boldsymbol{\mu}}_y, I)},$$

where $\boldsymbol{\mu}_y$ and $\hat{\boldsymbol{\mu}}_y$ are the means of the Gaussians for p_{data} and p_θ respectively.

Show that under this simplifying assumption, the optimal discriminator's logits $h^*(\boldsymbol{x}, y)$ can be written in the form

$$h^*(\boldsymbol{x}, y) = \boldsymbol{y}^T(A\varphi(\boldsymbol{x}) + \boldsymbol{b})$$

for some matrix A and vector $\boldsymbol{b}$, where $\boldsymbol{y}$ is a one-hot vector denoting the class y. In this problem, the discriminator's output and logits are related by $D_\phi(\boldsymbol{x}, y) = \sigma(h_\phi(\boldsymbol{x}, y))$. (Hint: use the result from problem 2.2.)

To rephrase the question a bit for clarity: there exists some mapping $\varphi : (\boldsymbol{x} \in \mathbb{R}^d, y) \mapsto \boldsymbol{r} \in \mathbb{R}^r$ where $\boldsymbol{r}$ is modeled by the distribution $\mathcal{N}(\boldsymbol{r}; \boldsymbol{\mu}_y, I)$. The result from problem 3.2 states that the optimal unconditional discriminator's logits $h^*(\boldsymbol{x})$ satisfy

$$h^*(\boldsymbol{x}) = \log \frac{p_{data}(\boldsymbol{x})}{p_\theta(\boldsymbol{x})} \tag{34}$$

For the class-conditional setting, our optimal discriminator logits $h^*(\boldsymbol{x}, y)$ will instead follow

$$\begin{aligned}
h^*(\boldsymbol{x}, y) &= \log \frac{p_{data}(\boldsymbol{x} \mid y)}{p_\theta(\boldsymbol{x} \mid y)} && (35)\\
&= \log \frac{\mathcal{N}(\varphi(\boldsymbol{x})|\boldsymbol{\mu}_y, I)}{\mathcal{N}(\varphi(\boldsymbol{x})|\hat{\boldsymbol{\mu}}_y, I)} && (36)\\
&= \log \frac{\exp\left(-\frac{1}{2}\left(\varphi(\boldsymbol{x}) - \boldsymbol{\mu}_y\right)^T\left(\varphi(\boldsymbol{x}) - \boldsymbol{\mu}_y\right)\right)}{\exp\left(-\frac{1}{2}\left(\varphi(\boldsymbol{x}) - \hat{\boldsymbol{\mu}}_y\right)^T\left(\varphi(\boldsymbol{x}) - \hat{\boldsymbol{\mu}}_y\right)\right)} && (37)\\
&= -\frac{1}{2}\left(\varphi(\boldsymbol{x}) - \boldsymbol{\mu}_y\right)^T\left(\varphi(\boldsymbol{x}) - \boldsymbol{\mu}_y\right) + \frac{1}{2}\left(\varphi(\boldsymbol{x}) - \hat{\boldsymbol{\mu}}_y\right)^T\left(\varphi(\boldsymbol{x}) - \hat{\boldsymbol{\mu}}_y\right) && (38)\\
&= \varphi(\boldsymbol{x})^T\left[\boldsymbol{\mu}_y - \hat{\boldsymbol{\mu}}_y\right] + \frac{1}{2}\left(\hat{\boldsymbol{\mu}}_y^2 - \boldsymbol{\mu}_y^2\right) && (39)
\end{aligned}$$

We can equivalently rewrite this to use the one-hot vector $\boldsymbol{y}$ by defining an m-element vector $\boldsymbol{b}$ whose ith entry equals $\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_i^2 - \boldsymbol{\mu}_i^2\right)$ (corresponding to class i). Similarly, we define the matrix $\boldsymbol{A}$ with m rows and r columns (dimensionality of $\varphi(\boldsymbol{x})$), with $A_{i,:} := \boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i$. This results in

$$\begin{aligned}
h^*(\boldsymbol{x}, y) &= \left(\boldsymbol{A}\varphi(\boldsymbol{x}) + \boldsymbol{b}\right)_y && (40)\\
&= \boldsymbol{y}^T\left(\boldsymbol{A}\varphi(\boldsymbol{x}) + \boldsymbol{b}\right) && (41)
\end{aligned}$$

Problem 5: Wasserstein GAN (35 points)

Part 1. *Let $p_\theta(x) = \mathcal{N}(x|\theta, \epsilon^2)$ and $p_{data}(x) = \mathcal{N}(x|\theta_0, \epsilon^2)$ be normal distributions with standard deviation ϵ centered at $\theta \in \mathbb{R}$ and $\theta_0 \in \mathbb{R}$ respectively. Show that*

$$\mathrm{KL}(p_\theta(x)||p_{data}(x)) = \frac{(\theta - \theta_0)^2}{2\epsilon^2}.$$

$$\begin{aligned}
\mathrm{KL}(p_\theta(x)||p_{\mathrm{data}}(x)) &= \mathbb{E}_{x\sim p_\theta}\left[\log \frac{p_\theta(x)}{p_{data}(x)}\right] && (42)\\
&= \int_{\mathbb{R}} p_\theta(x) \log \frac{p_\theta(x)}{p_{data}(x)} \mathrm{d}x && (43)\\
&= \int_{\mathbb{R}} \mathcal{N}(x \mid \theta, \epsilon^2) \log \frac{\exp\left(-\frac{(x-\theta)^2}{2\epsilon^2}\right)}{\exp\left(-\frac{(x-\theta_0)^2}{2\epsilon^2}\right)} \mathrm{d}x && (44)\\
&= \int_{\mathbb{R}} \mathcal{N}(x \mid \theta, \epsilon^2) \left(-\frac{1}{2\epsilon^2}\left((x-\theta)^2 - (x-\theta_0)^2\right)\right) \mathrm{d}x && (45)\\
&= \int_{\mathbb{R}} \mathcal{N}(x \mid \theta, \epsilon^2) \left(-\frac{1}{2\epsilon^2}\left(-2x\theta + \theta^2 + 2x\theta_0 - \theta_0^2\right)\right) \mathrm{d}x && (46)\\
&= \frac{1}{2\epsilon^2} \int_{\mathbb{R}} \mathcal{N}(x \mid \theta, \epsilon^2) \left(2x(\theta - \theta_0) + \theta_0^2 - \theta^2\right) \mathrm{d}x && (47)\\
&= \frac{1}{2\epsilon^2} \mathbb{E}_{x\sim p_\theta}\left[2x(\theta - \theta_0) + \theta_0^2 - \theta^2\right] && (48)\\
&= \frac{1}{2\epsilon^2}\left(2\theta(\theta - \theta_0) + \theta_0^2 - \theta^2\right) && (49)\\
&= \frac{1}{2\epsilon^2}\left(2\theta^2 - 2\theta\theta_0 + \theta_0^2 - \theta^2\right) && (50)\\
&= \frac{1}{2\epsilon^2}\left(\theta^2 - 2\theta\theta_0 + \theta_0^2\right) && (51)\\
&= \frac{(\theta - \theta_0)^2}{2\epsilon^2} && (52)
\end{aligned}$$

Part 2. *Suppose $p_\theta(x)$ and $p_{data}(x)$ both place probability mass in only a very small part of the domain; that is, consider the limit $\epsilon \to 0$. What happens to $\mathrm{KL}(p_\theta(x)||p_{data}(x))$ and its derivative with respect to θ, assuming that $\theta \neq \theta_0$? Why is this problematic for a GAN trained with the loss function L_G defined in problem 2.3?*

As $\epsilon \to 0$, $KL \to \infty$. It's derivative,

$$\frac{\partial}{\partial \theta} \frac{(\theta - \theta_0)^2}{2\epsilon^2} = \frac{2(\theta - \theta_0)}{2\epsilon^2} = \frac{\theta - \theta_0}{\epsilon^2} \tag{53}$$

also goes to ∞ as $\epsilon \to 0$. We saw in problem 3.3 that if $D_\phi = D^*$, then $L_G(\theta, \phi) = KL(p_\theta(x)||p_{data}(x))$. Therefore, if we have a sufficiently good discriminator $\approx D^*$, small $\epsilon \to 0$, and $\theta \neq \theta_0$, our losses and gradients will explode and prevent our model from training successfully.

Part 3. *To avoid this problem, we'll propose an alternative objective for the discriminator and generator. Consider the following alternative objectives:*

$$L_D(\phi;\theta) = \mathbb{E}_{x\sim p_\theta(x)}[D_\phi(x)] - \mathbb{E}_{x\sim p_{data}(x)}[D_\phi(x)]$$
$$L_G(\theta;\phi) = -\mathbb{E}_{x\sim p_\theta(x)}[D_\phi(x)],$$

where D_ϕ is no longer constrained to functions that output a probability; instead D_ϕ can be a function that outputs any real number. As defined, however, these losses are still problematic. Again consider the limit $\epsilon \to 0$; that is, let $p_\theta(x)$ be the distribution that outputs $\theta \in \mathbb{R}$ with probability 1, and let $p_{data}(x)$ be the distribution that outputs $\theta_0 \in \mathbb{R}$ with probability 1. Why is there no discriminator D_ϕ that minimizes this new objective L_D?

In this case, L_D simply becomes

$$L_D(\phi,\theta) = D_\phi(\theta) - D_\phi(\theta_0) \tag{54}$$

Proceed via proof by contradiction. Assume that there exists some D^* that minimizes L_D. Since D can be any function that outputs a real number, let another function D^{**} be defined as

$$D^{**} = \begin{cases} D^*(\theta) - 1 & \text{if } x = \theta \\ D^*(\theta_0) & \text{if } x = \theta_0 \\ 0 & \text{otherwise} \end{cases} \tag{55}$$

Plugging this into L_D reveals

$$L_{D^{**}}(\phi,\theta) = D^{**}(\theta) - D^{**}(\theta_0) \tag{56}$$
$$= D^*(\theta) - 1 - D^*(\theta_0) \tag{57}$$
$$= L_{D^*}(\phi,\theta) - 1 \tag{58}$$

and we see that $L_{D^{**}} < L_{D^*}$, which contradicts our premise that $D^* = \arg\min_D L_D(\phi,\theta)$. Therefore, there does not exist any function D that minimizes this form of L_D (under the assumptions of p_θ, p_{data}, etc).

Part 4. *Let's tweak the alternate objective so that an optimal discriminator exists. Consider the same objective L_D and the same limit $\epsilon \to 0$. Now, suppose that D_ϕ is restricted to differentiable functions whose derivative is always between -1 and 1. It can still output any real number. Is there now a discriminator D_ϕ out of this class of functions that minimizes L_D? Briefly describe what the optimal D_ϕ looks like as a function of x.*

If we add the constraint that $-1 < \frac{\partial D_\phi(x)}{\partial x} < 1$, then

$$-|\theta - \theta_0| < \ L_D(\phi, \theta) = D_\phi(\theta) - D_\phi(\theta_0) \ < |\theta - \theta_0| \tag{59}$$

$$\inf L_D(\phi, \theta) = -|\theta - \theta_0| \tag{60}$$

If, for example, we had $\theta \in \mathbb{R}^+$ only, then $D_\phi(x) = -x$ would satisfy this.