University of Illinois Spring 2018

CS 446: Machine Learning
Homework 11

Due on Tuesday, April 24, 2018, 11:59 a.m. Central Time

1. **[8 points]** Generative Adversarial Network (GAN)

   (a) What is the cost function for classical GANs? Use $D_w(x)$ as the discriminator and $G_\theta(x)$ as the generator, where the generator transforms $z \sim Z$ to $x \in X$.

   Your answer:

$$V(G_\theta, D_w) = \mathbb{E}_{x\sim p_X}\left[\log D_w(x)\right] + \mathbb{E}_{z\sim Z}\left[1 - \log D_w\left(G_\theta(z)\right)\right]$$

   (b) Assume arbitrary capacity for both discriminator and generator. In this case we refer to the discriminator using $D(x)$, and denote the distribution on the data domain induced by the generator via $p_G(x)$. State an equivalent problem to the one asked for in part (a), by using $p_G(x)$ and the ground truth data distribution $p_{data}(x)$.

   Your answer:

$$\begin{aligned}
V(G_\theta, D) &= \int_x p_{data}(x)\log\left(D(x)\right)dx + \mathbb{E}_{z\sim Z}\left[1 - \log D_w\left(G_\theta(z)\right)\right] \\
&= \int_x p_{data}(x)\log\left(D(x)\right)dx + \int_z p_Z(z)\log\left(1 - D\left(G_\theta(z)\right)\right)dz \\
&= \int_x p_{data}(x)\log\left(D(x)\right)dx + \int_x p_G(x)\log\left(1 - D(x)\right)dx \\
&= \int_x p_{data}(x)\log\left(D(x)\right) + p_G(x)\log\left(1 - D(x)\right)dx
\end{aligned}$$

   Where

$$\int_z p_Z(z)\log\left(1 - D\left(G_\theta(z)\right)\right)dz = \int_x p_G(x)\log\left(1 - D(x)\right)dx$$

   follows from applying LOTUS to $\mathbb{E}_{z\sim Z}\left[1 - \log D_w\left(G_\theta(z)\right)\right]$ with the change of variable $X = G_\theta(Z)$.

(c) Assuming arbitrary capacity, derive the optimal discriminator $D^*(x)$ in terms of $p_{data}(x)$ and $p_G(x)$.
You may need the Euler-Lagrange equation:

$$\frac{\partial L(x, D, \dot{D})}{\partial D} - \frac{d}{dx}\frac{\partial L(x, D, \dot{D})}{\partial \dot{D}} = 0$$

where $\dot{D} = \partial D / \partial x$.

Your answer: Following the Euler-Lagrange statement that the stationary point q of

$$S(q) = \int_a^b L(t, q(t), \dot{q}(t))dt$$

is given by the solution of

$$\frac{\partial L(t, q, \dot{q})}{\partial q} - \frac{d}{dx}\frac{\partial L(t, q, \dot{q})}{\partial \dot{q}} = 0$$

In our case, let

$$L(x, D, \dot{D}) = p_{data}(x)\log\left(D(x)\right) + p_G(x)\log\left(1 - D(x)\right)$$

So finding the optimal discriminator is the same thing as finding the stationary point in our Euler-Lagrange equation. Then

$$\begin{aligned}
\frac{\partial L(t, q, \dot{q})}{\partial q} - \frac{d}{dx}\frac{\partial L(t, q, \dot{q})}{\partial \dot{q}} &= \frac{\partial L(t, q, \dot{q})}{\partial q} - \frac{d}{dx}(0) \\
&= \frac{p_{data}}{D^*(x)} - \frac{p_G(x)}{1 - D^*(x)} = 0 \\
&\to p_{data}(x) - D^*(x)\left[p_G(x) + p_{data}(x)\right] = 0 \\
&\to D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)}
\end{aligned}$$

(d) Assume arbitrary capacity and an optimal discriminator $D^*(x)$, show that the optimal generator, $G^*(x)$, generates the distribution $p_G^* = p_{data}$, where $p_{data}(x)$ is the data distribution
You may need the Jensen-Shannon divergence:

$$\text{JSD}(p_{\text{data}}, p_G) = \frac{1}{2}D_{KL}(p_{\text{data}}, M) + \frac{1}{2}D_{KL}(p_G, M) \quad \text{with} \quad M = \frac{1}{2}(p_{\text{data}} + p_G)$$

Your answer:

Let's first look at what value of $V(G, D*)$ we get when $p_G = p_{data}$. Note that a consequence of this is $D*(x) = \frac{1}{2}$. Then

$$\begin{aligned}
V(G, D^*) &= \int_x p_{data}(x) \log \frac{1}{2} + p_G(x) \log \frac{1}{2} dx \\
&= \log \frac{1}{2} \int_x p_{data}(x) + p_G(x) dx \\
&= 2 \log \frac{1}{2} \\
&= -\log 4
\end{aligned}$$

Given the previous result for the optimal D:

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

We can rewrite $V(G, D)$ as

$$\begin{aligned}
V(G, D^*) &= \int_x p_{data}(x) \log (D(x)) + p_G(x) \log (1 - D(x))\, dx \\
&= \int_x p_{data}(x) \log (D(x)) + p_G(x) \log (1 - D(x))\, dx \\
&= \int_x p_{data}(x) \log \left(\frac{p_{data}(x)}{p_{data}(x) + p_G(x)} \right) + p_G(x) \log \left(1 - \frac{p_{data}(x)}{p_{data}(x) + p_G(x)} \right) dx \\
&= \int_x p_{data}(x) \log \left(\frac{p_{data}(x)}{p_{data}(x) + p_G(x)} \right) + p_G(x) \log \left(\frac{p_G(x)}{p_{data}(x) + p_G(x)} \right) dx \\
&= \int_x p_{data}(x) \log \left(\frac{1}{2} \cdot \frac{p_{data}(x)}{\frac{p_{data}(x)+p_G(x)}{2}} \right) + p_G(x) \log \left(\frac{1}{2} \cdot \frac{p_g(x)}{\frac{p_{data}(x)+p_G(x)}{2}} \right) dx \\
&= \int_x p_{data}(x) \left(\log \frac{1}{2} + \log \left(\frac{p_{data}(x)}{\frac{p_{data}(x)+p_G(x)}{2}} \right) \right) + p_G(x) \left(\log \frac{1}{2} + \log \left(\frac{p_G(x)}{\frac{p_{data}(x)+p_g(x)}{2}} \right) \right) dx \\
&= \log \frac{1}{2} \int_x p_{data}(x) + p_G(x) dx + \int_x p_{data}(x) \log \left(\frac{p_{data}(x)}{\frac{p_{data}(x)+p_G(x)}{2}} \right) \\
&\qquad + \int_x p_G(x) \log \left(\frac{p_G(x)}{\frac{p_{data}(x)+p_G(x)}{2}} \right) dx \\
&= -\log 4 + D_{KL}(p_{data} || M) + D_{KL}(p_G || M) \\
&\geq -\log 4
\end{aligned}$$

The inequality follows from the non-negativity of D_{KL} (proven in the previous HW). From this we can conclude that $-\log 4$ is the global minimum and, therefore, the optimal generator $G^*(x)$ generates data with distribution $p_G^* = p_{data}$

(e) More recently, researchers have proposed to use the Wasserstein distance instead of divergences to train the models since the KL divergence often fails to give meaningful information for training. Consider three distributions, $\mathbb{P}_1 \sim U[0, 1]$, $\mathbb{P}_2 \sim U[0.5, 1.5]$, and $\mathbb{P}_3 \sim U[1,2]$. Calculate $D_{KL}(\mathbb{P}_1, \mathbb{P}_2)$, $D_{KL}(\mathbb{P}_1, \mathbb{P}_3)$, $\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_2)$, and $\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_3)$, where $\mathbb{W}_1$ is the Wasserstein-1 distance between distributions.

Your answer:
KL divergence:

$$D_{KL}(\mathbb{P}_1||\mathbb{P}_2) = \infty$$

Since $\mathbb{P}_2$ becomes 0 in a region of $\mathbb{P}_1$ domain and makes the log blow up to ∞. For the same reason:

$$D_{KL}(\mathbb{P}_1||\mathbb{P}_3) = \infty$$

Wasserstein distances:

$$\mathbb{W}_1(U,V) = \int_{\gamma\in\Gamma(U,V)} |F(U) - F(V)| d\gamma$$

where F denotes the CDF of the distribution.
Then:

$$\begin{aligned}
\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_2) &= \int_{\gamma\in\Gamma(\mathbb{P}_1,\mathbb{P}_2)} |F(\mathbb{P}_1) - F(\mathbb{P}_2)| d\gamma \\
&= \int_{\gamma\in\Gamma(\mathbb{P}_1,\mathbb{P}_2)} |F(\mathbb{P}_1) - F(\mathbb{P}_2)| d\gamma \\
&= \int_0^{0.5} F(\mathbb{P}_1) d\gamma + \int_{0.5}^{1} |F(\mathbb{P}_1) - F(\mathbb{P}_2)| d\gamma + \int_1^{1.5} F(\mathbb{P}_2) d\gamma \\
&= \int_0^{0.5} \frac{\gamma - 0}{1-0} d\gamma + \int_{0.5}^{1} |\frac{\gamma - 0}{1-0} - \frac{\gamma - 0.5}{1.5-0.5}| d\gamma + \int_1^{1.5} \frac{\gamma - 0.5}{1.5-0.5} d\gamma \\
&= \frac{\gamma^2}{2}\Big|_0^{0.5} + 0.5\gamma\Big|_{0.5}^{1} + \left(1.5\gamma - \frac{\gamma^2}{2}\right)\Big|_1^{1.5} \\
&= 0.5
\end{aligned}$$

Following the same process (not shown):

$$\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_3) = 1$$