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.

   Your answer:
   Original GAN minimizes a divergence/distance between probability distributions.

   $$\min_G \max_D V(D_w(x), G_\theta(x)) = \mathbb{E}_{x\sim p_{data}}[\log D_w(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D_w(G_\theta(x)))]$$

   (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:

   $$\begin{aligned} \max_D V(G, D) &= \mathbb{E}_{x\sim p_{data}}[\log D_w(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D_w(G_\theta(x)))] && (1) \\ &= \mathbb{E}_{x\sim p_{data}}\left[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}\right] + \mathbb{E}_{x\sim p_g}\left[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}\right] && (2) \end{aligned}$$

(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:
The optimal discriminator $D^*(x)$ is given by fixed G (generator). Given any G, maximize $V(G, D)$

$$\begin{aligned} V(G, D) &= \int_x p_{data}(x)\log(D(x))dx + \int_z p_z(z)\log(1 - D(G(z)))dz && (3) \\ &= \int_x p_{data}(x)\log(D(x)) + p_G(x)\log(1 - D(x))dx && (4) \end{aligned}$$

The reason why transforming from Eq. (3) to Eq. (4) is since $x = G(z)$, we can replace $G(z)$ with variable x. Also in this case, p_g is the distribution of x (in generator).

For any $(a, b) \in \mathbb{R}^2 \backslash \{0, 0\}$, the function $y \rightarrow a\log(y) + b\log(1 - y)$ achieves its maximum in $[0, 1]$ at $\frac{a}{a+b}$. So the optimal discriminator $D^*(x)$ will be given by:

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

(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:

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

Your answer:
For $p^*_G = p_{data}$, we can get

$$D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_{data}(x)} = \frac{1}{2},$$

so that $\max\limits_D V(G, D) = \log \frac{1}{2} + \log \frac{1}{2} = -\log 4$

From Eq. (2), we can get:

$$\begin{aligned}
\max_D V(G, D) &= \mathbb{E}_{x \sim p_{data}}[\log D^*(x)] + \mathbb{E}_{x \sim p_G}[\log(1 - D^*(x)] \qquad (5)\\
&= \mathbb{E}_{x \sim p_{data}}[-\log 2] + \mathbb{E}_{x \sim p_G}[-\log 2] \qquad (6)\\
&= -\log 4 \qquad (7)
\end{aligned}$$

Meanwhile, if we simply Eq. (2):

$$\begin{aligned}
C(G) &= \max_D V(G, D) \qquad (8)\\
&= \mathbb{E}_{x \sim p_{data}}\left[\log \frac{p_{data}(x)}{p_{data}(x) + p^*_G(x)}\right] + \mathbb{E}_{x \sim p^*_G}\left[\log \frac{p^*_G(x)}{p_{data}(x) + p^*_G(x)}\right] \qquad (9)\\
&= \int_x \left(p_{data}(x) \log\left(\frac{p_{data}(x)}{p_{data}(x) + p^*_G(x)}\right) + p^*_G(x) \log\left(\frac{p^*_G(x)}{p_{data}(x) + p^*_G(x)}\right)\right) dx \qquad (10)\\
&= \int_x \left(p_{data}(x) \log\left(\frac{2p_{data}(x)}{p_{data}(x) + p^*_G(x)}\right) + p^*_G(x) \log\left(\frac{2p^*_G(x)}{p_{data}(x) + p^*_G(x)}\right)\right) dx - \log 4 \qquad (11)\\
&= -\log 4 + KL\left(p_{data} \| \frac{p_{data}(x) + p^*_G(x)}{2}\right) + KL\left(p^*_G \| \frac{p_{data}(x) + p^*_G(x)}{2}\right) \qquad (12)\\
&= -\log 4 + 2 \cdot JSD(p_{data}, p^*_G) \qquad (13)
\end{aligned}$$

Since the Jensen-Shannon divergence between two distributions is always non-negative and equals to zero if and only if the two distributions are equal. So that if $C(G)$ wants to be the optimal value (global optimum), the only solution is:

$$p^*_G = p_{data}$$

(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:

$$\begin{aligned} D_{KL}(\mathbb{P}_1, \mathbb{P}_2) &= \int \mathbb{P}_1 \log \frac{\mathbb{P}_1}{\mathbb{P}_2} \\ &= \inf \end{aligned}$$

$$\begin{aligned} D_{KL}(\mathbb{P}_1, \mathbb{P}_3) &= \int \mathbb{P}_1 \log \frac{\mathbb{P}_1}{\mathbb{P}_3} \\ &= \inf \end{aligned}$$

$$\begin{aligned} \mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_2) &= \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x,y) d\gamma(x,y) \\ &= \sup_{f \in \mathcal{F}_L} |\mathbb{E}_{x \sim \mathbb{P}_1}[f(x)] - \mathbb{E}_{y \sim \mathbb{P}_2}[f(y)]| \\ &= 0.5 \end{aligned}$$

$$\begin{aligned} \mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_3) &= \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x,y) d\gamma(x,y) \\ &= \sup_{f \in \mathcal{F}_L} |\mathbb{E}_{x \sim \mathbb{P}_1}[f(x)] - \mathbb{E}_{y \sim \mathbb{P}_3}[f(y)]| \\ &= 1 \end{aligned}$$

$\Gamma(\mu, \nu)$ denotes the collection of all measures on $M \times M$ with marginals μ and ν on the first and second factors respectively. (The set $\Gamma(\mu, \nu)$ is also called the set of all couplings of μ and ν.)
$f \in \mathcal{F}_L$, a natural class of smooth functions is the class of 1-Lipschitz functions, i.e.

$$\mathcal{F}_L = \{f : f \texttt{continuous}, |f(x) - f(y)| \leq \|x - y\|\}$$