Name:

University of Illinois Spring 2020

CS 446/ECE 449 Machine Learning
Homework 9: Generative Adversarial Nets (GANs)

Due on Tuesday April 28 2020, noon Central Time

1. **[24 points]** Generative Adversarial Nets (GANs) and Duality

   Consider the following program for a dataset $\mathcal{D} = \{(x)\}$ of points:

$$\max_{\theta} \min_{w} -\sum_{x\in\mathcal{D}} \log p_w(y=1|x) - \sum_{z\in\mathcal{Z}} \log(1 - p_w(y=1|G_\theta(z))) + \frac{C}{2}\|w\|_2^2. \quad (1)$$

   Hereby θ denotes the parameters of the generator $G_\theta(z)$, which transforms 'perturbations' $z \in \mathcal{Z}$ into artificial data, w refers to the parameters of the discriminator model $p_w(y|x)$, $y \in \{0,1\}$ denotes artificial or real data, and $C \geq 0$ is a fixed hyper-parameter.

   (a) (1 point) What is the original motivation (the one used in Goodfellow *et al*. (NIPS'14)) underlying generative adversarial nets (GANs)?

   Your answer:

   GANs help to prevent the need for any Markov chains or unrolled approximate inference networks during training and sample generation. The whole network can simply be trained with back propagation. The motivation comes from framing the problem as a game between generator and discriminator. In the paper, the authors liken the generator as a team of counterfeiters trying to produce fake currency without getting caught (realistic output). The discriminators can be thought of as the police trying to catch the counterfeiters. Competition between the two should result in the GAN network producing counterfeits that are practically indistinguishable from real data.

   (b) (1 point) Without restrictions on the generator model G_θ and the discriminator model p_w, what are challenges in solving the program given in Eq. (1)?

   Your answer:

   Since we are solving an optimization problem, we want the generator model to be concave since we are maximizing it over theta and we want the discriminator to be convex since we are minimizing it over w.

   (c) (2 points) We now restrict the discriminator as follows:

$$p_w(y=1|x) = \frac{1}{1+\exp w^\top x}.$$

   Using this discriminator, write down the resulting cost function for the program given in Eq. (1).

   Your answer:

$$\max_{\theta}\min_{w} -\sum_{x\in D} \log \frac{1}{1+\exp w^T x} - \sum_{z\in Z} \log\left(\frac{\exp w^T x}{1+\exp w^T x}\right) + \frac{C}{2}\|w\|_2^2$$

$$\max_{\theta}\min_{w} -\sum_{x\in D}\left[\log 1 - \log\left[1+\exp w^T x\right]\right] - \sum_{z\in Z}\log\left(\frac{\exp w^T x}{1+\exp w^T x}\right) + \frac{C}{2}\|w\|_2^2$$

$$\max_{\theta}\min_{w} \sum_{x\in D} \log\left[1+\exp w^T x\right] - \sum_{z\in Z}\log\left(\frac{\exp w^T x}{1+\exp w^T x}\right) + \frac{C}{2}\|w\|_2^2$$

Name:

(d) (2 points) When is the function $\frac{C}{2}\|a\|_2^2 - a^\top b$ convex in a? Why?

Your answer:

$$\frac{\partial}{\partial a} = Ca - b$$

$$\frac{\partial^2}{\partial^2 a} = C$$

Convex when $C \geq 0$

(e) (2 points) When is the function $\log(1 + \exp a^\top b)$ convex in a? Why?

Your answer:

$$\frac{\partial}{\partial a} = \frac{b\exp(a^\top b)}{1+\exp(a^\top b)}$$

$$\frac{\partial^2}{\partial a^2} = \frac{(1+\exp(a^\top b))\cdot[b^\top b\exp(a^\top b)] - b^\top b\exp(a^\top b)\cdot\exp(a^\top b)}{(1+\exp(a^\top b))^2}$$

$$= \frac{b^\top b\exp(a^\top b) + \cancel{b^\top b\exp(2a^\top b) - b^\top b\exp(2a^\top b)}}{(1+\exp(a^\top b))^2}$$

$$= \frac{b^\top b\exp(a^\top b)}{(1+\exp(a^\top b))^2} \geq 0$$, always b/c output of $b^\top b$, $\exp(a^\top b)$ and $(1+\exp(a^\top b))^2$ will always be positive.

(f) (2 points) Assume we restrict ourselves to the domain (if any) where $\frac{C}{2}\|a\|_2^2 - a^\top b$ and $\log(1 + \exp a^\top b)$ are convex in a, what can we conclude about convexity of the function

$$\sum_{x \in \mathcal{D}} \log(1 + \exp w^\top x) + \sum_{z \in \mathcal{Z}} \log(1 + \exp(w^\top G_\theta(z))) - \sum_{z \in \mathcal{Z}} w^\top G_\theta(z) + \frac{C}{2}\|w\|_2^2$$

in w and why?

Your answer:

Because both functions in part d and e have positive 2nd derivatives and since the function in question in f is just the sum of these functions, we can say the sum of the second derivative of these functions from d and e are also positive. Therefore the function in question in part f is also convex.

Name:

(g) (2 points) Let us introduce variables $\xi_x = w^\top x$ and $\xi_z = w^\top G_\theta(z)$ and let us consider the following program:

$$\min_w \quad \sum_{x\in\mathcal{D}} \log(1+\exp\xi_x) + \sum_{z\in\mathcal{Z}} \log(1+\exp\xi_z) - \sum_{z\in\mathcal{Z}} w^\top G_\theta(z) + \frac{C}{2}\|w\|_2^2 \qquad (2)$$

$$\text{s.t.} \quad \begin{cases} \xi_x = w^\top x & \forall x\in\mathcal{D} \quad (C1) \\ \xi_z = w^\top G_\theta(z) & \forall z\in\mathcal{Z} \quad (C2) \end{cases}.$$

What is the Lagrangian for this program? Use the Lagrange multipliers λ_x and λ_z for the constraints (C1) and (C2) respectively.

Your answer:

$L() = \sum_{x\in D} \log(1+\exp\xi_x) + \sum_{z\in Z}\log(1+\exp\xi_z) - \sum_{z\in Z}\xi_z + \frac{C}{2}\|w\|_2^2 + \sum_{x\in D}\lambda_x(\xi_x - w^T x) + \sum_{z\in Z}\lambda_z(\xi_z - w^T G_\theta(z))$

(h) (2 points) What is the value of

$$\min_w \frac{C}{2}\|w\|_2^2 - w^\top b$$

in terms of b and C?

Your answer:

$\frac{\partial}{\partial w}: Cw - b = 0$

$w = \frac{b}{C}$

$\boxed{\frac{C}{2}\left\|\frac{b}{C}\right\|_2^2 - \frac{b^T b}{C}}$

(i) (2 points) What is the value of

$$\min_\xi \lambda\xi + \log(1+\exp\xi)$$

in terms of λ? What is the valid domain for λ?

Your answer:

$\frac{\partial}{\partial \xi}: \lambda + \frac{\exp\xi}{1+\exp\xi} = 0$

$\exp\xi = -\lambda - \lambda\exp\xi$

$\exp\xi = \frac{-\lambda}{1+\lambda}$

$\xi = \log\left(\frac{-\lambda}{1+\lambda}\right)$

$\lambda\log\left(\frac{-\lambda}{1+\lambda}\right) + \log\left(1 - \frac{\lambda}{1+\lambda}\right)$

$\boxed{\lambda\log\left(\frac{-\lambda}{1+\lambda}\right) + \log\left(\frac{1}{1+\lambda}\right)}$

DOMAIN:

$\frac{-\lambda}{1+\lambda} > 0 \Rightarrow \boxed{-1 < \lambda < 0}$

Name: ________________________

(j) (6 points) Combine your results from the previous two sub-problems to derive the dual function of the program given in Eq. (2). Also state the dual program and clearly differentiate it from the dual function. State how this dual program can help to address a challenge in GAN training.

Your answer:

$$L() = \sum_{x\in D} \log(1-\exp\xi_x) + \sum_{z\in Z} \log(1+\exp\xi_z) - \sum_{z\in Z} w^T G_\theta(z) + \frac{C}{2}\|w\|_2^2 + \sum_{x\in D} \lambda_x (\xi_x - w^T x) + \sum_{z\in Z} \lambda_z (\xi_z - w^T G_\theta(z))$$

$$\frac{\partial L}{\partial w}: -\sum_{z\in Z} G_\theta(z) - \sum_{x\in D} \lambda_x x - \sum_{z\in Z} \lambda_z G_\theta(z) + Cw = 0$$

$$w = \frac{1}{C}\left[\sum_{x\in D} \lambda_x x + \sum_{z\in Z} G_\theta(z)(1+\lambda_z)\right]$$

$$\frac{\partial L}{\partial \xi_x}: \sum_{x\in D} \frac{\exp(\xi_x)}{1+\exp(\xi_x)} + \sum_{x\in D}\lambda_x = 0 \xrightarrow[\text{From (i)}]{} \xi_x = \log\left(\frac{-\lambda_x}{1+\lambda_x}\right)$$

$$\frac{\partial L}{\partial \xi_z}: \sum_{z\in Z} \frac{\exp(\xi_z)}{1+\exp(\xi_z)} + \sum_{z\in Z}\lambda_z = 0 \xrightarrow[\text{From (i)}]{} \xi_z = \log\left(\frac{-\lambda_z}{1+\lambda_z}\right)$$

DUAL PROGRAM:

$$\max_{\lambda_x,\lambda_z} g(\lambda_x,\lambda_z)$$
$$\text{S.T. } -1<\lambda_x<0$$
$$-1<\lambda_z<0$$

DUAL FUNCTION:

$$g(\lambda_x,\lambda_z) = \sum_{x\in D}\left[\log\left(1-\frac{\lambda_x}{1+\lambda_x}\right) + \lambda_x\left(\log\left(\frac{-\lambda_x}{1+\lambda_x}\right) - \frac{1}{C}\left(\sum_{x\in D}\lambda_x x + \sum_{z\in Z} G_\theta(z)(1+\lambda_z)\right)\right)\right]$$
$$+ \sum_{z\in Z}\left[\log\left(1-\frac{\lambda_z}{1+\lambda_z}\right) - \frac{G_\theta(z)}{C}\left(\sum_{x\in D}\lambda_x x + \sum_{z\in Z} G_\theta(z)(1+\lambda_z)\right) + \lambda_z\left(\log\left(\frac{-\lambda_z}{1+\lambda_z}\right) - \frac{G_\theta(z)}{C}\left(\sum_{x\in D}\lambda_x x + \sum_{z\in Z} G_\theta(z)(1+\lambda_z)\right)\right)\right]$$
$$+ \frac{1}{2C}\left\|\sum_{x\in D}\lambda_x x + \sum_{z\in Z} G_\theta(z)(1+\lambda_z)\right\|_2^2$$

A challenge in GAN training is in the beginning the discriminator is artificially really good because the initial outputs from the generator are really bad. This causes a problem because the discriminator will have very low gradient and our training will get stuck because we are trying to minimize the loss of our discriminator. If instead we maximize the dual function, we wont get stuck.

(k) (2 points) Implement and provide the loss for the discriminator and the generator when using the '-log-D' trick in `A9_GAN.py`.

Your answer:

Discriminator:

loss = criterion(logit, target1)

loss after 250 epochs = 0.595237

Generator:

loss = criterion(logit, target2).

loss after 250 epochs = 0.896365