Problem 1: Optimal GAN discriminator

Assume that the generator is defined as $G_\theta : \mathbb{R}^{d_{\text{latent}}} \to \mathbb{R}^{d_{\text{data}}}$, and the discriminator $D : \mathbb{R}^{d_{\text{data}}} \to [0, 1]$ outputs a probability $p(\text{real} \mid x)$. Let p_{data} be the density of the real data distribution and $p_\theta = (G_\theta)_{\#} p_{\text{latent}}$ be the density of the distribution of generated samples.
Show that for a fixed generator G_θ the optimal discriminator D^* for the following optimisation problem

$$\max_D \left\{ \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_{\text{latent}}} [\log(1 - D(G_\theta(z)))] \right\}$$

has the following form: $D^*(\text{real} \mid x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)}$ (hint: for $a, b > 0$ find the maximiser of $a \log(x) + b \log(1 - x)$ with respect to x, then use this result to find the optimal D^*).

Problem 2: Alternative GAN losses

1. Alternatively, losses for training discriminator and generator can be written is follows:

$$\mathcal{L}_{\text{MSE}}(\varphi) = \mathbb{E}_{x \sim p_{\text{data}}} \left[(1 - D_\varphi(x))^2 \right] + \mathbb{E}_{z \sim p_{\text{latent}}} \left[(D_\varphi(G_\theta(z)))^2 \right],$$
$$\mathcal{L}_{\text{MSE}}(\theta) = \mathbb{E}_{z \sim p_{\text{latent}}} \left[(1 - D_\varphi(G_\theta(z)))^2 \right],$$

   which are minimised with respect to φ and θ respectively. This contrasts with the vanilla objective

$$\mathcal{L}_{\text{vanilla}}(\theta, \varphi) = \mathbb{E}_{x \sim p_{\text{data}}} \left[\log D_\varphi(x) \right] + \mathbb{E}_{z \sim p_{\text{latent}}} \left[\log(1 - D_\varphi(G_\theta(z))) \right]$$

   Show that the gradients $\nabla_\varphi \mathcal{L}_{\text{MSE}}(\varphi)$ and $\nabla_\theta \mathcal{L}_{\text{MSE}}(\theta)$ are proportional to $\nabla_\varphi \mathcal{L}_{\text{vanilla}}(\varphi, \theta)$ and $\nabla_\theta \mathcal{L}_{\text{vanilla}}(\varphi, \theta)$ respectively.

2. In order to improve the training stability one can use 'non-saturating' losses for training the generator: $-\log D(G_\theta(z))$ in place of $\log(1 - D(G_\theta(z)))$:

$$\mathcal{L}_{\text{non-sat}}(\theta) = \mathbb{E}_{z \sim p_{\text{latent}}} \left[-\log D(G_\theta(z)) \right].$$

   Are the non-saturating GAN gradients proportional to the vanilla ones?

Problem 3: Wasserstein GAN

Read **Wasserstein GAN paper**. Let $\mathbb{P}_r$ represent the distribution of real data and $\mathbb{P}_g$ the distribution of generated data. Answer the following:

1. Explain (informally) how using $\text{KL}(\mathbb{P}_r \,\|\, \mathbb{P}_g)$ as a measure of distance between two distributions differs from $\mathbb{W}_1(\mathbb{P}_r, \mathbb{P}_g)$ and why the latter is more sensible.

2. (**Hard**) Explain (informally) why:

$$\inf_{\pi \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \pi} \left[\|x - y\| \right] = \sup_{\|f\|_L \leq 1} \left\{ \mathbb{E}_{x \sim \mathbb{P}_r} [f(x)] - \mathbb{E}_{y \sim \mathbb{P}_g} [f(y)] \right\}$$

   Answer the following questions:

   (a) Give the definition of a 1-Lipschitz function and explain (informally) why f should be 1-Lipschitz.

   (b) If f is parameterised by a neural network, what are the possible ways to make f 1-Lipschitz?