Homework 4

Deep Learning 2025 Spring

Due on 2025/3/31

1 True or False

Problem 1. In practice, we expect a higher FID score for better generators.

Problem 2. To train a GAN, generally, it is good to first update the discriminator to optimality before each generator update.

2 Q&A

Problem 3. (Jensen-Shannon Divergence) Jensen-Shannon Divergence (JSD) between two distributions p and q is defined as

$$\mathrm{JSD}(p\|q) = \frac{1}{2}\left(\mathrm{KL}\left(p\middle\|\frac{p+q}{2}\right) + \mathrm{KL}\left(q\middle\|\frac{p+q}{2}\right)\right)$$

where $\mathrm{KL}(\cdot\|\cdot)$ represents KL divergence. We only consider discrete distributions in this problem.

1. Prove

   $$\mathrm{JSD}(p\|q) = H\left(\frac{p+q}{2}\right) - \frac{1}{2}\left(H(p) + H(q)\right)$$

   where $H(\cdot)$ is Shannon entropy.

2. Prove that JSD is bounded: $0 \leq \mathrm{JSD}(p\|q) \leq \log 2$.

3. Prove the Pythagorean theorem of JSD, i.e., for any distributions p_1, p_2, p_3, the following inequality holds:

   $$\sqrt{\mathrm{JSD}(p_1\|p_2)} + \sqrt{\mathrm{JSD}(p_2\|p_3)} \geq \sqrt{\mathrm{JSD}(p_1\|p_3)}.$$

Problem 4. (Wasserstein Distance) We recall the definition of L_2 Wasserstein distance $W(p,q)$, where p and q are two discrete distributions over a finite alphabet $\mathcal{X}$. Namely,

$$W(p,q) = \inf_{\gamma\in\Gamma} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|_2 \,\mathrm{d}\gamma(x,y) = \inf_{\gamma\in\Gamma} \sum_{x,y} \gamma(x,y)\|x-y\|_2$$

where Γ is the set of all possible joint distributions on $\mathcal{X} \times \mathcal{X}$ that have marginals p and q. We denote p_{x} as the probability distribution of random variable x. Let $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ be an isotropic Gaussian random variable. We define $p_{\mathsf{x}+\epsilon}$ as the density of random variable $\mathsf{y} = \mathsf{x} + \epsilon$, i.e.,

$$p_{\mathsf{x}+\epsilon}(y) = \int p_{\mathsf{x}+\epsilon|\mathsf{x}}(y|x) p_{\mathsf{x}}(x) \, \mathrm{d}x = \mathbb{E}_{x \sim p_{\mathsf{x}}} \left[p_\epsilon(y - x) \right].$$

1. Apply Kantorovich-Rubinstein duality to prove the Pythagorean theorem of Wasserstein distance, i.e., for any discrete distributions p, q, r over a finite alphabet $\mathcal{X}$, we have

$$W(p, q) \leq W(p, r) + W(r, q)$$

2. Prove $W(p_{\mathsf{x}}, p_{\mathsf{x}+\epsilon}) \leq V^{\frac{1}{2}}$, where $V = \mathbb{E}\left[\|\epsilon\|_2^2\right]$ is the variance of ϵ.

3. Denote $r \sim p_{\mathsf{r}}$ and $g \sim p_{\mathsf{g}}$ as random variables of ground truth and generated images respectively. If $p_{\mathsf{r}+\epsilon}$ and $p_{\mathsf{g}+\epsilon}$ have support contained on a ball of diameter C, prove

$$W(p_{\mathsf{r}}, p_{\mathsf{g}}) \leq 2V^{\frac{1}{2}} + 2C\sqrt{\mathrm{JSD}(p_{\mathsf{r}+\epsilon} \| p_{\mathsf{g}+\epsilon})}.$$

   Hint 1: If p_{x} and p_{y} have support contained on a ball of diameter C, we have

$$W(p_{\mathsf{x}}, p_{\mathsf{y}}) \leq C\delta(p_{\mathsf{x}}, p_{\mathsf{y}}) \leq C\sqrt{\frac{1}{2} D_{\mathrm{KL}}(p_{\mathsf{x}}, p_{\mathsf{y}})}$$

   where $\delta(p_{\mathsf{x}}, p_{\mathsf{y}}) = \sup |p_{\mathsf{x}} - p_{\mathsf{y}}|$ is the total variation distance between p_{x} and p_{y}.

   Hint 2: Try to apply triangular inequalities for Wasserstein distance and total variation.

4. If we train GANs using the original JSD objective, what tricks can we apply to implicitly minimize the Wasserstein distance according to the above result? What's the potential issue of this trick?