Generative Adversarial Networks (Part 4)

Dr. Alireza Aghamohammadi

Overcoming Training Challenges in GANs

Training GANs is challenging due to:

- ❖ Vanishing gradients
- ❖ Mode collapse

A better approach is to use a more effective distance metric:

- ❖ The Wasserstein (Earth Mover's) Distance measures the work required to transport probability mass from one distribution to another.
- ❖ It is well-defined even for disjoint distributions and decreases smoothly as distributions become closer.

Understanding Wasserstein Distance for Discrete Distributions

The Wasserstein distance is easiest to understand for discrete distributions.

- Consider distributions $Pr(x = i)$ and $q(x = j)$ defined over K bins.
- Assume there is a cost associated with moving one unit of mass from bin i in the first distribution to bin j in the second.
- The amounts that are moved form the transport plan and are stored in a matrix P. P is a joint probability distribution.

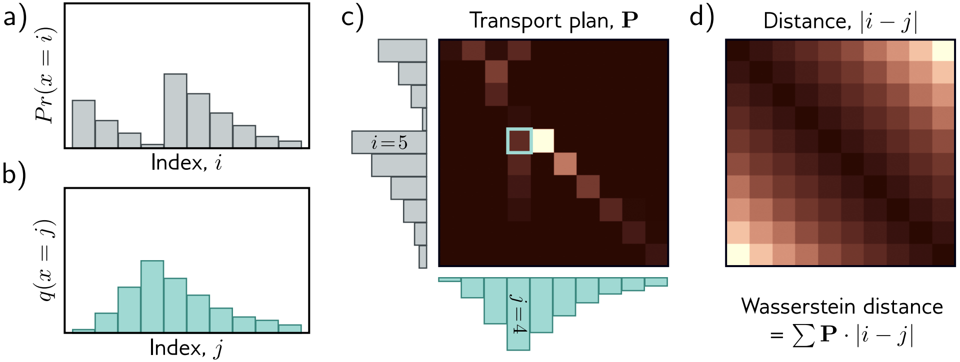

Image credit: Understanding Deep Learning

Defining the Wasserstein Distance

The Wasserstein distance is defined as:

$$D_w\left[Pr(x) \mid\mid q(x)\right] = \min_P \left[\sum_{i,j} P_{i,j} \cdot \mid i - j \mid\right]$$

subject to the constraints:

$$\sum_j P_{i,j} = Pr(x = i) \quad \text{(initial distribution of } Pr(x)\text{)}$$

$$\sum_i P_{i,j} = q(x = j) \quad \text{(initial distribution of } q(x)\text{)}$$

$$P_{i,j} \geq 0 \quad \text{(non-negative masses)}$$

The equivalent dual problem is:

$$D_w\left[Pr(x) \mid\mid q(x)\right] = \max_f \left[\sum_i Pr(x = i) f_i - \sum_j q(x = j) f_j\right]$$

subject to the constraint:

$$\mid f_i - f_j \mid \leq \mid i - j \mid$$

Translating to Continuous Version of Wasserstein Distance

Translating these results to the continuous multi-dimensional domain:

$$D_w\left[Pr(x) \,||\, q(x)\right] = \max_{f[x]} \mathbb{E}_{x\sim Pr(x)}\left[f\left[x\right]\right] - \mathbb{E}_{x\sim q(x)}\left[f\left[x\right]\right]$$

subject to the constraint that the Lipschitz constant of the function $f\left[x\right]$ is less than one

A function $f:\mathbb{R}\to\mathbb{R}$ is called K-Lipschitz continuous if there exists a real constant $K\geq 0$ such that, for all $x_1, x_2 \in \mathbb{R}$:

$$\mid f\left[x_1\right] - f\left[x_2\right] \mid \leq K \mid x_1 - x_2 \mid$$

Defining Loss Functions for Discriminator and Generator

To train GANs, we define the loss functions for the discriminator D and the generator G:

- ❖ Discriminator Loss:

$$\begin{aligned}\mathcal{L}_D = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[D(G(z))\right] - \mathbb{E}_{x\sim\mathsf{Data}}\left[D(x)\right] \\ + \lambda\cdot\left(\|\nabla D(\epsilon\cdot x + (1-\epsilon)\cdot G(z))\|_2 - 1\right)^2\end{aligned}$$

- ❖ Generator Loss:

$$\mathcal{L}_G = -\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[D(G(z))\right]$$