COMP 7070 Advanced Topics in AI and ML March 26, 2025

Lecture 10: GAN, WGAN, and Duality

Instructor: Yifan Chen *Scribes: Chenxi Liu, Yuhan Zhao* *Proof reader: Xiong Peng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

10.1 Generative Adversarial Networks (GANs)

10.1.1 Original Formulation of GANs

Beginning with Maximum Likelihood Estimation (MLE), we derive the loss function for training Generative Adversarial Networks (GANs) as follows:

$$\max_{D_\varphi} \mathbb{E}_{X,Y}[D_\varphi(x)]^Y[1 - D_\varphi(x)]^{1-Y}. \tag{10.1}$$

Assuming $P(Y=1) = \frac{1}{2}$, we obtain:

$$\mathbb{E}_{X,Y}\left[\log D_\varphi(x)Y + \log(1 - D_\varphi(x))(1-Y)\right]. \tag{10.2}$$

$$\mathbb{E}_{Y}\mathbb{E}_{X|Y}\left[\log D_\varphi(x)Y + \log(1 - D_\varphi(x))(1-Y)\right]. \tag{10.3}$$

Thus, we have:

$$\frac{1}{2}\mathbb{E}_{x\sim P_{\text{data}}}[\log D_\varphi(x)] + \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D_\varphi(x))]. \tag{10.4}$$

The objective of GANs is to train a discriminator and a generator to maximize the probability of correctly identifying real samples from the data distribution $P_{\text{data}}(x)$, i.e., $P(x|Y=1)$:

$$\max_{D_\varphi,\text{ fix } G_\theta} \frac{1}{2}\mathbb{E}_{x\sim P_{\text{data}}}[\log D_\varphi(x)] + \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D_\varphi(x))]. \tag{10.5}$$

Simultaneously, we aim to minimize the probability of identifying generated samples from the generator G_θ. Since the former part is unrelated to G_θ, we only consider the latter part in minimization:

$$\min_{G_\theta} \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D_\varphi(x))]. \tag{10.6}$$

This is the original form of GANs. Let us analyze the issues inherent in this formulation.

10.1.2 Challenges with a Powerful Discriminator D_φ

We begin by discussing the challenges associated with the minimization component. In a Bayesian context, if D_φ is excessively powerful, it would be:

$$\begin{aligned} D^*(x) &= P(Y=1|X) && (10.7)\\ &= \frac{P(x, Y=1)}{P(x)} && (10.8)\\ &= \frac{P(x|Y=1)}{\frac{1}{2}P(x|Y=1) + \frac{1}{2}P(x|Y=0)} = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_{G_\theta}(x)}. && (10.9) \end{aligned}$$

The Bayes classifier is:

$$h^*(x) = \arg\max_k P(Y = k|X = x) \tag{10.10}$$

h^* minimizes:

$$\begin{aligned} R(h) &= \mathbb{E}_{X,Y}[1_{\{Y \neq h(x)\}}] && (10.11)\\ &= \mathbb{E}_X \mathbb{E}_{Y|X}[1_{\{Y \neq h(x)\}}] && (10.12)\\ &= \mathbb{E}_X \left[P(Y = 1|x)1_{\{h(x)=0\}} + P(Y = 0|x)1_{\{h(x)=1\}}\right] && (10.13) \end{aligned}$$

Thus:

$$R(h) \geq \min\{P(Y = 1|X)P(Y = 0|X)\} = R(h^*) \tag{10.14}$$

Therefore:

$$h^* = \begin{cases} 1, & \text{if } P(Y = 1|X) > P(Y = 0|X) \\ 0, & \text{otherwise} \end{cases} \tag{10.15}$$

Under these conditions, the MLE of GAN becomes:

$$\mathbb{E}_{x \sim P_{\text{data}}} \log \frac{P_{\text{data}}(x)}{\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)]} + \mathbb{E}_{x \sim P_{G_\theta}} \log \frac{P_{G_\theta}(x)}{\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)]} - 2\log 2. \tag{10.16}$$

The first component is the Kullback-Leibler divergence $\text{KL}(P_{\text{data}}||\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)])$, and the second part is $\text{KL}(P_{G_\theta}||\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)])$. Hence, the loss simplifies to:

$$\text{KL}(P_{\text{data}}||\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)]) + \text{KL}(P_{G_\theta}||\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)]) - 2\log 2. \tag{10.17}$$

Given that JS denotes Jensen-Shannon divergence:

$$\text{JS}(P_{\text{data}}||P_{G_\theta}) = \frac{1}{2}\left[\text{KL}(P_{\text{data}}||\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)]) + \text{KL}(P_{G_\theta}||\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)])\right]. \tag{10.18}$$

Consequently, the MLE of GAN is:

$$2\text{JS}(P_{\text{data}}||P_{G_\theta}) - 2\log 2, \text{ JS} \in [0, \log 2]. \tag{10.19}$$

Without the Gaussian assumption, JS can be computed but may be non-informative, approaching JS → $\log 2$.

$$\min_{G_\theta} \frac{1}{2}\mathbb{E}_{x \sim P_{G_\theta}(x)}[\log(1 - D_\varphi(x))], \text{ when } D_\varphi = D^* \tag{10.20}$$

The gradient is:

$$\nabla_\theta \frac{1}{2}\mathbb{E}_{x \sim P_{G_\theta}(x)}[\log(1 - D_\varphi(x))] = \mathbb{E}_{X \sim P_{G_\theta}} \log D^*(x) \nabla_\theta \log P_{G_\theta}(x) \tag{10.21}$$

If $P_{G_\theta}(x) \ll P_{\text{data}}(x)$, $D^*(x) \approx 1$, indicating no alignment to P_{data}. Conversely, if $P_{G_\theta}(x) \gg P_{\text{data}}(x)$, $D^*(x) \approx 0$, leading to gradient vanishing, making training difficult. Thus, we employ the log trick to reformulate the minimization as:

$$\min_{G_\theta} \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D_\varphi(x))] \Rightarrow \min_{G_\theta} -\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log D_\varphi(x)]. \tag{10.22}$$

Under the condition $D_\varphi = D^*$, we have $\mathrm{KL}(P_{G_\theta}||P_{\text{data}})$:

$$\mathrm{KL}(P_{G_\theta}||P_{\text{data}}) = \mathbb{E}_{x\sim P_{G_\theta}} \log \frac{P_{G_\theta}}{P_{\text{data}}} = \mathbb{E}_{x\sim P_{G_\theta}} \log \frac{1 - D^*(x)}{D^*(x)} \tag{10.23}$$

$$= \mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D^*(x))] - \mathbb{E}_{x\sim P_{G_\theta}(x)}[\log D^*(x)] \tag{10.24}$$

$$= 2 \cdot \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D^*(x))] + (-\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log D_\varphi(x)]). \tag{10.25}$$

Thus, we have:

$$\begin{aligned}
-\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log D_\varphi(x)] &= \mathrm{KL}(P_{G_\theta}||P_{\text{data}}) - 2 \cdot \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D^*(x))] && (10.26)\\
&= \mathrm{KL}(P_{G_\theta}||P_{\text{data}}) - 2 \cdot \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D^*(x))] && (10.27)\\
&\quad + \frac{1}{2}\mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D^*(x)] + 2 \cdot \frac{1}{2}\mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D^*(x)] && (10.28)\\
&= \mathrm{KL}(P_{G_\theta}||P_{\text{data}}) - 2\mathrm{JS}(P_{\text{data}}||P_{G_\theta}) + 2\log 2 + \mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D^*(x)]. && (10.29)
\end{aligned}$$

Since $2\mathrm{JS}(P_{\text{data}}||P_{G_\theta}) - \log 2 \approx 0$, our focus is on minimizing $\mathrm{KL}(P_{G_\theta}||P_{\text{data}})$. However, directly optimizing KL divergence presents the following challenges:

- Minimizing $\mathrm{KL}(P_{G_\theta}||P_{\text{data}})$ may result in $P_{G_\theta} \to 0$ and $P_{\text{data}} \to 1$, implying low penalty for **bad generation**.

- Minimizing $\mathrm{KL}(P_{G_\theta}||P_{\text{data}})$ may lead to $P_{G_\theta} \to 1$ and $P_{\text{data}} \to 0$, causing overly large penalty for **unrealistic generation**.

To prevent unrealistic generation, the optimizer focuses on ensuring P_{data} is large to guarantee safe and realistic generation, known as the model collapse problem. In contrast, VAEs do not face this issue as the KL divergence they optimize is defined on latent space z, not x.

10.2 Wasserstein Generative Adversarial Networks (WGANs)

Traditional GANs employ JS divergence to measure the difference between data and model distributions, potentially leading to issues like mode collapse:

$$\min_{G_\theta} \max_{D_\varphi} \frac{1}{2}\mathbb{E}_{x\sim P_{\text{data}}}[\log D_\varphi(x)] + \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1 - D_\varphi(x))] \iff \min_{G_\theta, D_\varphi} \mathrm{JS}(P_{\text{data}}||P_{G_\theta}). \tag{10.30}$$

WGANs introduce the Wasserstein distance as an alternative to quantify the discrepancy between data and model distributions. The Wasserstein distance offers smoother gradients throughout, theoretically facilitating more stable training and reducing the likelihood of mode collapse:

$$\min_{G_\theta} W_1(P_{\text{data}}, P_{G_\theta}). \tag{10.31}$$

10.2.1 Wasserstein Distance

The Wasserstein distance, also known as the Earth Mover's Distance (EMD), measures the distance between two probability distributions over a given metric space. Named after mathematician Leonid Vaserštejn (Leonid Wasserstein), it quantifies the minimum "work" required to transform one probability distribution into another, where "work" is defined as the amount of distribution weight moved times the distance it is moved.

Mathematically, given two probability measures P_{data} and P_{G_θ} on a metric space (M, d), the p-Wasserstein distance between P_{data} and P_{G_θ} is defined as:

$$W_p(P_{\text{data}}, P_{G_\theta}) = \min_{\gamma \in \Gamma(P_{\text{data}}, P_{G_\theta})} \mathbb{E}_{(x_1, x_2) \sim \gamma} ||x_1 - x_2||^p. \tag{10.32}$$

Here, $\Gamma(P_{\text{data}}, P_{G_\theta})$ is the set of all couplings of P_{data} and P_{G_θ}, which are measures with marginals P_{data} and P_{G_θ} on the first and second factors, respectively. The case where $p = 1$ is often utilized in the context of GANs:

$$W_1(P_{\text{data}}, P_{G_\theta}) = \max_{f \in \text{1-Lip}} \mathbb{E}_{x \sim P_{\text{data}}}[f(x)] - \mathbb{E}_{x \sim P_{G_\theta}}[f(x)]. \tag{10.33}$$

This equation is known as the Kantorovich-Rubinstein Duality. In WGANs, the Kantorovich-Rubinstein Duality is employed to derive a practical training algorithm. The critic (or discriminator) in WGANs is trained to approximate the 1-Lipschitz function that realizes the supremum, thereby providing gradient information essential for training the generator to produce samples that minimize the Wasserstein distance to the real data distribution.

10.3 Optimal Transport Problems

10.3.1 Monge Problems

Motivation How to transport mass from a source distribution to another target one efficiently?

Mathematical Formulation Given two probability measures μ (source) and ν (target) on spaces X and Y, and also a cost function $c(x, y)$ (e.g., Euclidean distance $\|x - y\|$), we seek an optimal transport map $T : X \to Y$ s.t. $\nu(B) = \mu(T^{-1}(B))$ for all measurable B, and

$$\inf_T \int_X c(x, T(x))\, d\mu(x)$$

10.3.2 Kantorovich Problems (Relaxation)

Leonid Kantorovich (1942) proposed a **relaxation** of the Monge problem. It relaxes the original problem to a convex one:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times Y} c(x, y)\, d\pi(x, y)$$

where:

- $\Pi(\mu, \nu) = \{\pi \text{ on } X \times Y \mid \pi(A \times Y) = \mu(A), \pi(X \times B) = \nu(B)\}$.
- π represents a **transport plan**, where $\pi(x, y)$ specifies how much mass moves from x to y.

Example Let μ be uniform on $[0, 1]$ and ν uniform on $[1, 2]$, with quadratic cost $c(x, y) = (x - y)^2$.

The optimal transport map is $T(x) = x + 1$:

$$\int_0^1 (x - T(x))^2\, dx = \int_0^1 (x - (x + 1))^2\, dx = \int_0^1 1\, dx = 1.$$

10.4 Duality

10.4.1 General Derivation of Dual Form via Lagrangian Multipliers

Original problem Consider:

$$\min_x f(x) \quad \text{s.t.} \quad g_i(x) \leq 0, \quad h_j(x) = 0$$

where:

- $f(x)$ is the objective function,
- $g_i(x) \leq 0$ are inequality constraints,
- $h_j(x) = 0$ are equality constraints.

Lagrangian Introduce Lagrange multipliers $\lambda_i \geq 0$ (for inequalities) and ν_j (for equalities):

$$\mathcal{L}(x, \lambda, \nu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \nu_j h_j(x)$$

Dual Function is:

$$g(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu)$$

Dual Problem is:

$$\max_{\lambda \geq 0, \nu} g(\lambda, \nu)$$

10.4.2 Example: Dual Form of Wasserstein Distance

Original problem Consider:

$$C(\pi) := \inf_{\pi \in \Pi} \int_{X \times Y} c(x, y) \cdot \pi(x, y) \, dx \, dy,$$

with constraints:

$$\begin{aligned} \mu(x) - \int_Y \pi(x, y) \, dy &= 0 \\ \nu(y) - \int_X \pi(x, y) \, dx &= 0 \\ -\pi(x, y) &\leq 0 \end{aligned}$$

Lagrangian Introduce Lagrange multipliers $\tau \geq 0$ (for inequalities) and ϕ, ψ (for equalities):

$$\begin{aligned} \mathcal{L}(\pi, \tau, \phi, \psi) =& C(\pi) + \int \phi(x) \left[\mu(x) - \int_Y \pi(x, y) \, dy \right] \, dx \\ &+ \int \psi(y) \left[\nu(x) - \int_X \pi(x, y) \, dx \right] \, dy \\ &+ \int \int \tau(x, y) \left[-\pi(x, y) \right] \, dx \, dy \\ =& \int_{X \times Y} \left[c(x, y) - \phi(x) - \psi(y) - \tau(x, y) \right] \pi(x, y) \, dx \, dy \\ &+ \int_X \phi(x) \mu(x) \, dx + \int_Y \psi(y) \nu(y) \, dy. \end{aligned}$$

Dual Function is:

$$g(\phi, \psi, \tau) = \inf_{\pi} \mathcal{L}(\pi, \tau, \phi, \psi)$$

First, if $c(x, y) - \phi(x) - \psi(y) - \tau(x, y) < 0$ at some point, then g must be $-\infty$. Now, we assume $c(x, y) - \phi(x) - \psi(y) - \tau(x, y) \geq 0$, thus:

$$g(\phi, \psi, \tau) = \int_X \phi(x)\mu(x)\, dx + \int_Y \psi(y)\nu(y)\, dy.$$

Dual Problem Since $\tau \geq 0$, we have $c(x, y) \geq \phi(x) + \psi(y)$, the dual problem is

$$\sup_{\phi,\psi} \int_X \phi(x)\mu(x)\, dx + \int_Y \psi(y)\nu(y)\, dy \quad \text{s.t.} \quad \phi(x) + \psi(y) \leq c(x, y)$$

Wasserstein Distance Set $c(x, y) = |x - y|$, we have

$$\sup_{\phi,\psi} \int_X \phi(x)\mu(x)\, dx + \int_Y \psi(y)\nu(y)\, dy \quad \text{s.t.} \quad \phi(x) + \psi(y) \leq |x - y|$$

10.5 First Variation and Functional Derivative

10.5.1 First Variation

Given a functional:

$$J[y] = \int_a^b F(x, y, y')\, dx,$$

its **first variation** under a perturbation $y \mapsto y + \epsilon\eta$ (where $\eta(a) = \eta(b) = 0$) is:

$$\delta J = \frac{d}{d\epsilon} J[y + \epsilon\eta] \bigg|_{\epsilon=0}.$$

1. Expand $J[y + \epsilon\eta]$ to first order in ϵ:

$$J[y + \epsilon\eta] = \int_a^b F(x, y + \epsilon\eta, y' + \epsilon\eta')\, dx.$$

2. Differentiate under the integral and set $\epsilon = 0$:

$$\delta J = \int_a^b \left(\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta' \right) dx.$$

3. Integrate by parts to eliminate η':

$$\delta J = \int_a^b \left(\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'} \right) \eta\, dx + \left[\frac{\partial F}{\partial y'}\eta \right]_a^b = \int_a^b \left(\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'} \right) \eta\, dx.$$

The boundary term vanishes due to $\eta(a) = \eta(b) = 0$.

Euler-Lagrange Equation For $\delta J = 0$ (stationarity), the integrand must vanish:

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'} = 0.$$

10.5.2 Functional Derivative

The functional derivative $\frac{\delta J}{\delta y}$ is the function satisfying:

$$\delta J = \int_a^b \frac{\delta J}{\delta y}(x)\,\eta(x)\,dx.$$

From the first variation, we identify:

$$\frac{\delta J}{\delta y} = \frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'}.$$

Examples

- **Dirichlet Energy** $J[y] = \frac{1}{2}\int_a^b y'(x)^2\,dx$:

$$\frac{\delta J}{\delta y} = -y''(x).$$

- **Potential Energy** $J[y] = \int_a^b V(y(x))\,dx$:

$$\frac{\delta J}{\delta y} = V'(y(x)).$$