COMP 7070 Advanced Topics in AI and ML March 19, 2024

Lecture 10: GAN and WGAN

Instructor: Yifan Chen *Scribes: Riwei Lai* *Proof reader: Zhanke Zhou*

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

10.1 GANs (Generative Adversarial Networks)

10.1.1 Original form of GAN

Starting from Maximum Likelihood Estimation (MLE) and derive the loss function for training Generative Adversarial Networks (GAN) as:

$$\frac{1}{2}\mathbb{E}_{x\sim P(x|Y=1)}[\log D_\varphi(x)] + \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1-D_\varphi(x))].$$

The goal of GAN 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., $P(x|Y=1)$:

$$\max_{D_\varphi,\text{ fix generator } G_\theta} \frac{1}{2}\mathbb{E}_{x\sim P(x|Y=1)}[\log D_\varphi(x)] + \frac{1}{2}\mathbb{E}_{x\sim P_{G_\theta}(x)}[\log(1-D_\varphi(x))].$$

and minimize the probability of identifying generated samples from the generator $G(\theta)$. Since the former part is not related to $G(\theta)$, we only consider the latter part in minimization:

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

This is the original form of GAN. Let us analyze the problems of the original form.

10.1.2 What if the discriminator D_φ is too powerful?

We first state the problem of the minimization part. In the Bayesian Setting, if D_φ is too powerful, it would be:

$$D_\varphi = D^*(x) = \frac{P_{(x|Y=1)}}{\frac{1}{2}P_{(x|Y=1)} + \frac{1}{2}P_{(x|Y=0)}} = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_{G_\theta}(x)}.$$

If conditions apply, the MLE of GAN would be:

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

The first part is actually 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 $\text{KL}(P_{G_\theta}||\frac{1}{2}[P_{\text{data}}(x) + P_{G_\theta}(x)])$, so we can simplify the loss as:

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

Given that JS refers to Jensen-Shannon divergence:

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

We have the MLE of GAN would be:

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

Without the Gaussian assumption, we have JS $\to \log 2$ and the following expressions:

$$2\text{JS}(P_{\text{data}}||P_{G_\theta}) - \log 2 \approx 0.$$

which means that it is hard to train.
We use a log D trick to rewrite the minimization part as:

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

If conditions $D_\varphi = D^*$ apply, we have $\text{KL}(P_{G_\theta}||P_{\text{data}})$:

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

So we have:

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

Since $2\text{JS}(P_{\text{data}}||P_{G_\theta}) - \log 2 \approx 0$, here we try to minimize the $\text{KL}(P_{G_\theta}||P_{\text{data}})$. However, the problem of directly optimizing KL divergence is outlined as follows:

- Minimizing the $\text{KL}(P_{G_\theta}||P_{\text{data}})$ may lead to $P_{G_\theta} \to 0$ and $P_{\text{data}} \to 1$, which means a **bad** generation.
- Minimizing the $\text{KL}(P_{G_\theta}||P_{\text{data}})$ may lead to $P_{G_\theta} \to 1$ and $P_{\text{data}} \to 0$, which means a **unrealistic** generation.

To avoid unrealistic generation, the optimizer will focus on forcing P_{data} to be very large to ensure safe and real generation, which is called the model collapse problem. While in VAE, there is no such problem since the KL divergence it optimized is defined on z, not x.

10.2 WGAN (Wasserstein Generative Adversarial Network)

Recall that traditional GAN uses the JS divergence as a measure to calculate the difference between the data distribution and the model distribution, which can lead to issues like mode collapse.

$$\min_{G_\theta, D_\varphi} \mathrm{JS}(P_{\text{data}} || P_{G_\theta}).$$

WGAN introduces the Wasserstein distance as an alternative to measure the discrepancy between the data and model distributions. The Wasserstein distance provides smoother gradients everywhere, which theoretically allows for the training to progress more steadily and makes it less likely to encounter mode collapse.

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

10.2.1 Wasserstein distance

The Wasserstein distance, also known as the Earth mover's distance (EMD), is a measure of the distance between two probability distributions over a given metric space. It is named after the mathematician Leonid Vaserštejn (Leonid Wasserstein).

The basic idea of the Wasserstein distance is to compute the minimum "work" required to transform one probability distribution into another, where "work" is measured as the amount of distribution weight that must be moved times the distance it has to be moved.

Mathematically, if we have 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:

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

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

$$W_1(P_{\text{data}}, P_{\text{data}}) = \max_{f \in 1-\mathrm{Lip}} \mathbb{E}_{P_{\text{data}}} f(x_1) - \mathbb{E}_{P_{G_\theta}} f(x_2).$$

This equation is called Kantorovich-Rubinstein Duality. In WGAN, the Kantorovich-Rubinstein Duality is utilized to derive a practical training algorithm. The critic (or discriminator) in WGAN is trained to approximate the 1-Lipschitz function that realizes the supremum, which in turn provides the gradient information needed to train the generator so that it can produce samples that minimize the Wasserstein distance to the real data distribution.

10.3 Optimal Transport Problems

10.3.1 Monge problem

Monge's problem is a classical formulation in optimal transportation theory where the goal is to find the most efficient plan to transport a mass distribution, represented by a measure μ, into another mass distribution, represented by a measure ν, with the least possible cost.

Let X and Y be two separate measurable spaces and let μ be a measure on X and ν be a measure on Y. Monge's problem can be formulated as follows:

Find a transport map $T : X \to Y$ that minimizes the total transportation cost:

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

subject to the constraint that T pushes μ forward to ν:

$$T_{\#}\mu = \nu.$$

This means that for every Borel set $B \subseteq Y$,

$$\mu(T^{-1}(B)) = \nu(B).$$

The cost function $c(x, y)$ typically represents the transportation cost for moving a unit mass from x to y. As an example, one might consider $c(x, y) = \|x - y\|^2$ for the squared Euclidean distance, mimicking the "work" or "energy" required to perform the transportation.

In this setup, we often require that μ is absolutely continuous with respect to the Lebesgue measure, ensuring that T is a well-defined map in terms of the Radon-Nikodym derivative.

In calculus of variations, the change of variables trick can be used to convert the problem into a more tractable form by changing the space over which the objective function is optimized:

$$\int_Y f(y)\, d\nu(y) = \int_X f(T(x))\, d\mu(x).$$

Here, f is an arbitrary nonnegative function which could be represented by a combination of indicator functions:

$$f = \sum_i c_i \cdot \mathbf{1}_{A_i}(x).$$

Where $\mathbf{1}_{A_i}(x)$ is the indicator function for the set A, taking on a value of 1 if $x \in A$ and 0 otherwise.

Proof. We start by proving for indicator functions:

For an indicator function $\mathbf{1}_B$ of a measurable set $B \subseteq Y$,

$$\int_Y \mathbf{1}_B(y)\, d\nu(y) = \nu(B).$$

by definition of the integral of an indicator function. Moreover, since ν is the pushforward of μ by T,

$$\nu(B) = \mu(T^{-1}(B)).$$

Now for $x \in X$,

$$\mathbf{1}_B(T(x)) = \begin{cases} 1 & \text{if } T(x) \in B \\ 0 & \text{otherwise} \end{cases}.$$

Therefore,

$$\int_X \mathbf{1}_B(T(x))\, d\mu(x) = \int_X \mathbf{1}_{T^{-1}(B)}(x)\, d\mu(x) = \mu(T^{-1}(B)) = \nu(B).$$

So this proves the desired equality for indicator functions:

$$\int_Y \mathbf{1}_B(y)\, d\nu(y) = \int_X \mathbf{1}_B(T(x))\, d\mu(x).$$

Next, we extend the result to simple functions:

A simple function $s(y)$ is a finite linear combination of indicator functions,

$$s(y) = \sum_{i=1}^{n} a_i \mathbf{1}_{B_i}(y),$$

where a_i are constants and B_i are disjoint measurable sets. Since both sides of the integral are linear, it follows that

$$\int_Y s(y)\,d\nu(y) = \sum_{i=1}^{n} a_i \int_Y \mathbf{1}_{B_i}(y)\,d\nu(y) = \sum_{i=1}^{n} a_i \int_X \mathbf{1}_{B_i}(T(x))\,d\mu(x) = \int_X s(T(x))\,d\mu(x).$$

Finally, we extend to non-negative measurable functions:

A non-negative measurable function $f(y)$ can be approximated from below by an increasing sequence of simple functions $s_n(y)$ that converges to $f(y)$ pointwise. This is due to the Monotone Convergence Theorem.

Since the sequence $\{s_n\}$ converges pointwise to f and $s_n \leq f$ for all n,

$$\lim_{n\to\infty} \int_Y s_n(y)\,d\nu(y) = \int_Y f(y)\,d\nu(y),$$

and

$$\lim_{n\to\infty} \int_X s_n(T(x))\,d\mu(x) = \int_X f(T(x))\,d\mu(x).$$

So,

$$\int_Y f(y)\,d\nu(y) = \int_X f(T(x))\,d\mu(x).$$

This completes the proof for non-negative measurable functions f.

For real-valued f which can be negative, you would apply the above argument to the positive part f^+ and the negative part f^- of f separately, and combine the results, noting that

$$\int_Y f(y)\,d\nu(y) = \int_Y f^+(y)\,d\nu(y) - \int_Y f^-(y)\,d\nu(y).$$

and similar for the integral over X.

Remember, the key assumption here that makes this proof work is that the map T "pushes forward" the measure μ to ν in the sense that $\nu(B) = \mu(T^{-1}(B))$ for every measurable set B. If your function f, spaces X, Y, map T, or measures μ, ν don't satisfy the requirements needed for this proof, modifications would have to be made.

□

10.3.2 Kantorovich problem: a relaxation

In its simplest form, the Kantorovich problem can be described as follows: Imagine you have a certain amount of goods at several locations (e.g., warehouses) and a number of destinations where the goods need to be delivered (e.g., stores). You also have a cost function that describes the expense of transporting the goods from each origin to each destination.

The goal of the Kantorovich problem is to find the most efficient way to transport goods from the origins to the destinations, minimizing the total transportation cost, subject to the constraints of the supply at each origin and demand at each destination.

Mathematically, the Kantorovich problem can be formulated as a linear program where you optimize over a transport plan, which is a matrix that specifies how much of the goods to transport from each origin to each destination. The problem can be expressed as follows:

Given a cost function $c(x, y)$, where x and y are points in two separate spaces (representing the locations of supply and demand), the Kantorovich problem seeks a transport plan π, which is a probability measure on the product space of the two, that minimizes the total cost:

$$\min \int_{X\times Y} c(x, y)\,d\pi(x, y).$$

subject to constraints that ensure the marginal distribution of π matches the given supply and demand distributions μ (on X) and ν (on Y):

$$\pi(A \times Y) = \mu(A) \quad \text{for all measurable sets } A \subseteq X,$$
$$\pi(X \times B) = \nu(B) \quad \text{for all measurable sets } B \subseteq Y.$$

Assuming that T is the optimal transport map pushing forward μ into ν, $T_{\#}\mu = \nu$, meaning that T pushes μ to ν. The cost of transport for mass "moving" from position x to position $T(x)$ with respect to the measure μ is given by the left-hand side of the equation:

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

This integral calculates the total cost associated with transporting the distribution μ to ν using the transport map T, where each point x from X is sent to the point $T(x)$ in Y.

In practice, we often treat the transport problem as a form of linear programming:

$$\min\langle T, C\rangle, \text{s.t. constraints defined by distribution vectors } a \text{ and } b$$

The transport matrix T, where $T \cdot \mathbf{1}_m = \mathbf{a}$ and $T^\top \cdot \mathbf{1}_n = \mathbf{b}$, specifies the distribution from n supply points to m demand points with n parameters and m constraints.

The expected cost in the transport problem can be modeled by an integral or expectation:

$$\mathbb{E}[c(X, T(X))] = \int c(x, T(x)) d\mu(x).$$

We consider a specific case with quadratic costs:

$$\mathbb{E}[\|X - Y\|^2] = \mathbb{E}[\|X\|^2] + \mathbb{E}[\|Y\|^2] - 2\mathbb{E}[X^\top Y].$$

10.4 Duality in Optimization

10.4.1 General Duality Concept

For a set A and a point x, the duality concept can be expressed as follows:

$$d(A, x) = \min_{a \in A} d(x, a) = \max_{s \text{ separates } A, x} d(x, s).$$

where s is the boundary separating point x from set A.

10.4.2 Dual Form Derivation

Consider the primal optimization problem:

$$\min_{x \in D} f(x), \text{ subject to some constraints.}$$

Here, $f(x)$ is a function to be minimized over the domain D.

Transforming it into the Lagrangian form with the use of Lagrange multipliers λ and y for the inequality constraints $g(x) \leq 0$ and the equality constraint $h(x) = 0$, respectively, we obtain the Lagrangian:

$$L(x, \lambda, y) = f(x) + \lambda^\top g(x) + y^\top h(x).$$

The dual problem involves flipping the min and max, expressed as:

$$\max_{\lambda \geq 0, y} \min_{x \in D} L(x, \lambda, y).$$

This forms the dual optimization problem, where λ is the vector of Lagrange multipliers for the inequality constraints, and $g(x) \geq 0$ ensures feasibility.

10.4.3 Wasserstein Case

Considering the Wasserstein distance, a special case in optimal transport:

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

where $\pi \in \Pi(P_X, P_Y)$ is a transport plan within the set of all possible plans between the probability measures P_X and P_Y, and $c(x, y)$ is the cost function, often chosen to be a distance measure between x and y.

Consider any transport map $\pi(x, y)$ without constraints. The optimality condition can be given as:

$$\mu(x) - \int_Y \pi(x,y)\,dy = 0.$$

Additionally, for π in some set of transport plans UV, we have the following conditions:

$$\begin{aligned}\nu(y) - \int_X \pi(x,y)\,dx &= 0,\\ -\pi(x,y) &\le 0.\end{aligned}$$

The Lagrangian $L(\Pi, \phi, \psi, \tau)$ can be expressed as:

$$\begin{aligned}L(\Pi,\phi,\psi,\tau) =& C_k(\pi) + \int \phi(x)\left[\mu(x) - \int_Y \pi(x,y)\,dy\right]\,dx\\ &+ \int \psi(y)\left[v(y) - \int_X \pi(x,y)\,dx\right]\,dy\\ &+ \int\int \tau(x,y)\left[-\pi(x,y)\right]\,dxdy.\end{aligned}$$

The primal problem can be stated as finding the infimum of the Lagrangian form:

$$\inf_{\pi\in\Pi}\sup_{\phi,\psi,\tau\ge 0} L(\Pi,\phi,\psi,\tau).$$

Moreover, we derive the inner integration in the Lagrangian formulation concerning measures and potential functions: [Not sure]

$$\int L(\Pi,\phi,\psi,\tau) = \inf_{\Pi}\int \pi(x,y)\left[c(x,y) - \tau(x,y) - \phi(x), -\psi(y)\right]\,dy.$$

And we have:

$$c(x,y) - \tau(x,y) - \phi(x) - \psi(y) = 0, \quad \text{when } \Pi(x,y) > 0.$$

Upon eliminating terms, we have:

$$c(x,y) - \phi(x) - \psi(y) \ge 0.$$

Thus we have the dual formulation:

$$\sup_{\phi,\psi}\left(\int \phi(x)\,du(x) + \int \psi(y)\,dv(y)\right), \text{subject to } \phi(x) + \psi(y) \le c(x,y).$$

10.4.4 Economic Explanation[Not sure]

Here we introduce an optimization problem, characterized by the following expressions:

$$\inf \left\{ \int \int c(x,y) \cdot \pi(x,y)\, dx\, dy \right\},$$

where we minimize the cost function subject to certain constraints:

$$u(x) + v(y) = c(x,y), \tag{1}$$
$$u(x) \geq v(y). \tag{2}$$

We interpret (2) as the utility gained by the payor as greater than or equal to the utility of the vendor, suggesting an economic transaction.

The collection fee ($\phi(x)$) and delivery fee ($\psi(y)$) are presumably related to the utility functions described above.