IFT 6085 - Lecture 13
Wasserstein GANs

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes
Winter 2019: [Jose Gallego, Fabrice Normandin, Guillaume Le Berre]
Winter 2018: [Yuchen Lu, Kyle Kastner]

Instructor: Ioannis Mitliagkas

1 Summary

Previously, we presented the difference between discriminative and generative modelling. We gave a taxonomy of generative models that divides them as *prescribed* or *implicit*. Prescribed generative models provide an explicit description of the distribution of the data, for example, by means of a parametric density function. On the other hand, implicit generative models specify a distribution by providing a sampling procedure that generates the data. A particularly important example of this are Generative Adversarial Networks (GANs). In this lecture we expand on GANs and exhibit certain difficulties regarding the formulation of the problem in terms of the Jensen-Shannon divergence. Ideas from the field of optimal transport will allow us to come up with an alternative divergence minimization problem for training generative models with better theoretical guarantees and improved practical performance.

2 Comparing Probability Distributions

The aim of generative modelling is to be able to produce samples that *resemble* those coming from a "real" distribution $\mathbb{P}_r$ on a space $\mathcal{X}$. Suppose that we are given two generative models which induce the distributions $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$. How can we decide which of these two models to pick? Furthermore, how can we asses how much does our approximation $\mathbb{P}_\theta$ resemble the real data distribution? This leads us to the notion of distances and divergences between probability distributions.

Definition 1 (Metric and Divergence)**.** *Given a set S, a metric on S is a function $d : S \times S \to [0, \infty)$ such that for all $x, y, z \in S$:*

- $d(x, y) \geq 0$ *(non-negativity)*
- $d(x, y) = 0$ *iff* $x = y$ *(identity of indiscernibles)*
- $d(x, y) = d(y, x)$ *(symmetry)*
- $d(x, y) \leq d(x, z) + d(z, y)$ *(triangle inequality)*

A divergence is a weaker notion, for which the symmetry and triangle inequality properties might not be satisfied.

A way to assess "how far is one probability distribution from another" is to construct/define a metric on the space of probability distributions on $\mathcal{X}$, which we will denote $\mathcal{P}_\mathcal{X}$. In this context, the set S in Definition 1, is $\mathcal{P}_\mathcal{X}$ and each of the elements of S corresponds to some probability distribution on $\mathcal{X}$.

Just as there are many ways to define distances between points in $\mathbb{R}^n$ (e.g. Manhattan distance, Euclidean distance, maximum distance), there are are several possible distances and divergences between probability distributions. We now list some of the most common alternatives.

Definition 2 (Total Variation distance)**.** *Let $(\mathcal{X}, \Sigma)$ be a measurable space.*

$$\delta(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \Sigma} |\mathbb{P}(A) - \mathbb{Q}(A)| = \sup_{A \in \Sigma} \left| \int \mathbf{1}_{\{x \in A\}} \, p(x) dx \, - \int \mathbf{1}_{\{x \in A\}} \, q(x) dx \right|$$

In particular, when $\mathcal{X}$ is a finite space, we have that $\delta(\mathbb{P}, \mathbb{Q}) = \max_{x \in \mathcal{X}} |\mathbb{P}(x) - \mathbb{Q}(x)|$.

Definition 3 (Kullback-Leibler divergence)**.**

$$KL(\mathbb{P} \,||\, \mathbb{Q}) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

Note that if p and q are continuous and there is a point $x \in X$ for which $p(x) > 0$ while $q(x) = 0$, the value of the KL divergence is ∞. This is a divergence as it fails to satisfy the symmetry and triangle inequality properties.

Definition 4 (Jensen-Shannon divergence)**.**

$$JS(\mathbb{P} \,||\, \mathbb{Q}) = KL \left(\mathbb{P} \,||\, \frac{\mathbb{P} + \mathbb{Q}}{2} \right) + KL \left(\mathbb{Q} \,||\, \frac{\mathbb{P} + \mathbb{Q}}{2} \right)$$

This is clearly symmetric in $\mathbb{P}$ and $\mathbb{Q}$. However, although the JS divergence does not satisfy the triangle inequality, its square root of the JS divergence does, and thus defines a metric on $\mathcal{P}_{\mathcal{X}}$.

We refer the interested reader to Chapters 2 and 10 of [3], which provide a good overview of the applications of the KL and JS divergences in machine learning.

At this stage we have developed the necessary structure to discuss the convergence of probability distributions. For this we first recall the notion of convergence in a metric space.

Definition 5 (Convergence in a metric space)**.** *A sequence $\{a_n\}_{n \in \mathbb{N}}$ of elements of a metric space (S, d) is said to converge to a limit $L \in S$ if $\lim_{n \to \infty} d(a_n, L) = 0$, i.e., if the real-valued sequence $\{d(a_n, L)\}_{n \in \mathbb{N}}$ converges to 0.*

Notice how the particular metric d in S influences the definition of convergence. In technical lingo, one would say that "d induces a topology on S". This means that if we endow the set S with two different metrics d and d', a fixed sequence $\{a_n\}_{n \in \mathbb{N}}$ might converge with respect to d but not with respect to d'.

In the context of probability distributions, we say that a sequence of distributions $\{\mathbb{P}_n\}_{n \in \mathbb{N}}$ converges to $\mathbb{P}$ with respect to a divergence $\mathcal{D}$ if $\lim_{n \to \infty} \mathcal{D}(\mathbb{P}_n, \mathbb{P}) = 0$. We draw the attention of the reader to the fact that the non-symmetry of divergences requires special care in the order of the arguments, and different choices might lead to different conclusions about the convergence of a sequence.

3 Learning Parallel Lines

The application of generative models to high-dimensional data requires the ability to effectively compare distributions which might have misaligned supports. This is usually formalized under the so-called *manifold hypothesis*, which states that real-world data distributions (e.g. images, text) are supported on low dimensional surfaces in their corresponding spaces.

We already mentioned the fact that the KL divergence can grow arbitrarily large when the supports of the distributions to be compared do not match. Now we show that this issue can appear in a very simple and low-dimensional setting. This is a warning regarding the qualities of these divergences and the need for "weaker" distances [1].

Consider the parametric distribution $\mathbb{P}_\theta$ in $\mathbb{R}^2$ with the following density, illustrated in Figure 1.

$$p_\theta(x, y) = \begin{cases} 1 & x = \theta \text{ and } 0 \leq y \leq 1, \\ 0 & \text{else.} \end{cases}$$

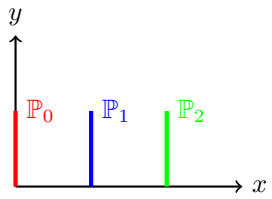

Figure 1: Density "heatmap" of $\mathbb{P}_\theta$ for different values of θ.

Let $\mathbb{P}_0$ be the target distribution we want to learn. Consider the generative model given by let $g_\theta(z) = (\theta, z)$ with $z \sim U[0,1]$ and $\theta \in \mathbb{R}$. Under these conditions, the distribution generated by g_θ is exactly $\mathbb{P}_\theta$. The computation of the aforementioned divergences yields the results shown in Figure 2. By definition, all the divergences are 0 when $\theta = 0$. However, for any $\theta \neq 0$ we obtain the following "saturating" behavior:

- $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = 1$
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \log(2)$
- $KL(\mathbb{P}_0 \,||\, \mathbb{P}_\theta) = KL(\mathbb{P}_\theta \,||\, \mathbb{P}_0) = \infty$.

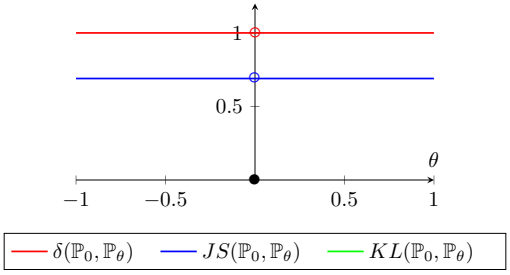

Figure 2: Values of the divergences as functions of θ. KL divergence values are ∞ except at $\theta = 0$.

Therefore (slightly counter-intuitively), even though the sequence $\{\frac{1}{n}\}_{n\in\mathbb{N}}$ converges to 0, the sequence $\{\mathbb{P}_{\frac{1}{n}}\}_{n\in\mathbb{N}}$ does **not** converge with respect to any of the divergences mentioned so far. Moreover, the gradient of these functions with respect to θ is 0 everywhere, except for $\theta = 0$, where it is not even defined. So, even though we can compute all of these divergences analytically, this precludes us from being able to learn even this very simple task with gradient-based methods.

4 Optimal Transport and Wasserstein Distance

One of the shortcomings of the divergences mentioned before is that they disregard extra structure that can be present in $\mathcal{X}$. Notably, in the case where $\mathcal{X}$ is endowed a metric, the three divergences we have mentioned are completely agnostic about the fact that some points of $\mathcal{X}$ are closer to others. This unveils the implicit assumptions we had the previous section when we described the non-convergence of the sequence $\{\mathbb{P}_{\frac{1}{n}}\}_{n\in\mathbb{N}}$ to $\mathbb{P}_0$ as "counter-intuitive". The elements of the sequence are getting closer *geometrically* to $\mathbb{P}_0$, although the divergences seem to be unaware of this.

The field of Optimal Transport studies the optimal transportation and allocation of resources. In the context of probability theory, one can think of defining a notion of distance between probability distributions by considering what would be the minimum possible cost for transporting all of the mass from one distribution to another.

The Wasserstein [\vasershtān\] distance W is a metric between probability distributions defined on a given metric space. Therefore, in some sense, the Wasserstein distance construction is more demanding since it requires a metric structure on $\mathcal{X}$. Note that in this setting we are dealing with two metric spaces simultaneously: $(\mathcal{X}, d)$ is the metric *data* space (e.g. images with a pixel-wise squared distance) and $(\mathcal{P}_\mathcal{X}, W)$ is the metric space of distributions over $\mathcal{X}$.

Let us now introduce the fundamental notion of couplings and show how they relate to transportation problems.

Definition 6 (Coupling)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on the spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ be the set of all joint distributions on $\mathcal{X}\times\mathcal{Y}$. An element $\pi \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ is called a coupling of $\mathbb{P}$ and $\mathbb{Q}$ if its marginals coincide with $\mathbb{P}$ and $\mathbb{Q}$, respectively. The set of all couplings between $\mathbb{P}$ and $\mathbb{Q}$ is denoted by $\Pi(\mathbb{P},\mathbb{Q})$.*

Equivalently, $\pi \in \Pi(\mathbb{P},\mathbb{Q})$ if and only if:

$$\int_{\mathcal{X}\times\mathcal{Y}} \pi(\cdot, y)dy = p(\cdot) \quad \text{and} \quad \int_{\mathcal{X}\times\mathcal{Y}} \pi(x,\cdot)dx = q(\cdot)$$

When $\mathcal{X}$ and $\mathcal{Y}$ are finite spaces of sizes n and m, the couplings correspond to matrices $\pi \in [0,1]^{n\times m}$ such that $\pi\mathbf{1}_m = \mathbb{P}$ and $\mathbf{1}_n\pi = \mathbb{Q}$. Thus, $\Pi(\mathbb{P},\mathbb{Q})$ is a polytope in $\mathbb{R}^{n\times m}$.

Example 7. *Let X and Y be two Bernoulli(0.5) random variables. Tables 1 and 2 show two possible couplings of X and Y. Note that $\pi(x,y) = \mathbb{P}(x)\mathbb{Q}(y)$ is always a coupling, and therefore $\Pi(\mathbb{P},\mathbb{Q})$ is always non-empty.*

x \ y	0	1
0	0.25	0.25
1	0.25	0.25

Table 1: Independent coupling $\pi(x,y) = \mathbb{P}(x)\mathbb{Q}(y)$

x \ y	0	1
0	0.5	0
1	0	0.5

Table 2: Correlated coupling $X = Y$

Example 8. *Suppose that the production (P) of avocados in Canada is split 70% in Montréal, and 30% in Québec City. However, 60% of the avocados are consumed (C) by the people in Québec City, while the remaining 40% are eaten by the Montrealers. Suppose that the cost of transporting one avocado is $0.001\$$ per mile, and the distance between Montréal and Québec City is 150 miles. What is the best transportation plan to fulfill the demand in both cities at the minimum cost?*

*Tables 3 and 4 show the independent coupling (always possible to construct) and the optimal coupling (check this!). The transportation cost under the independent coupling is $0.081 = 0.54 * 0.001 * 150\$$ per avocado, while the optimal transportation cost is $0.045 = 0.3 * 0.001 * 150\$$. Note the role the distance between the cities plays here!*

One can interpret the entries of π as a transportation plan as follows: $\frac{\pi(p,c)}{P(p)}$ represents the proportion of avocados produced in location p that will be shipped to location c. For example, in the optimal coupling, approximately $57\% \approx \frac{0.4}{0.7}$ of the avocados produced in Montréal remain there.

P \ C	M	Q
M	0.28	0.42
Q	0.12	0.09

Table 3: Independent coupling

P \ C	M	Q
M	0.4	0.3
Q	0	0.3

Table 4: Optimal coupling

Definition 9 (Wasserstein-1 distance)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions over a metric space $(\mathcal{X}, d)$.*

$$W(\mathbb{P},\mathbb{Q}) = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(x,y)\sim\pi}[d(x,y)] = \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \langle \pi, d\rangle,$$

where the $\langle \pi, d\rangle$ is the Frobenius product between the coupling π and a matrix representation of the distance d.

In other words, the Wasserstein distance formalizes the idea of "looking over all possible transportation plans and selecting the best one" by means of the infimum over the space of couplings between both distributions.

Note that the final formulation and the fact that $\Pi(\mathbb{P},\mathbb{Q})$ is a polytope imply that the computation of the Wasserstein distance is equivalent to solving a linear optimization problem. In finite dimensions, it is possible to solve LPs remarkably efficiently in practice (with a worst-case exponential complexity). However, the infinite dimensional case is in general intractable.

5 Back to Parallel Lines

Let us consider again the problem of learning the parameter θ for a generator $g_\theta(z) = (\theta, z)$, with $z \sim U[0, 1]$. With our insights on the relation between the Wasserstein distance and transportation problems, it is easy to see that the optimal coupling between $\mathbb{P}_0$ and $\mathbb{P}_\theta$ is given by:

$$\pi((x, y), (x', y')) = \begin{cases} 1 & x = 0,\ x' = \theta \text{ and } y = y', \\ 0 & \text{else.} \end{cases}$$

Therefore, $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$ (check this!). So, the Wasserstein distance allows us to assess effectively how far the generated distribution is from the target distribution, and provides a good training signal with strong gradients, suitable for standard machine learning optimization techniques.

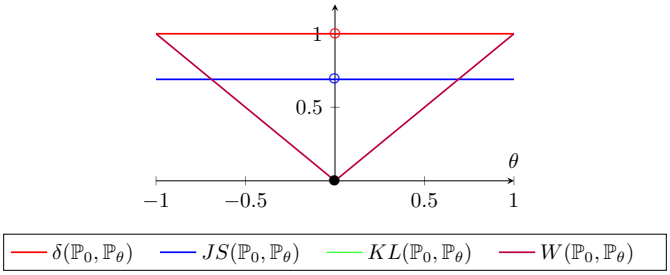

Figure 3: Behavior of the Wasserstein distance for the parallel lines problem.

6 Wasserstein GANs

The previous example gives the hope that the Wasserstein distance can be a useful tool when trying to learn distributions with misaligned supports. However, in order to make this practical in high-dimensional settings we need to somehow overcome the optimization challenges posed by the computation of the Wasserstein distance. Let us evoke the celebrated Kantorovich-Rubinstein duality:

Theorem 10 (Kantorovich-Rubinstein duality)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions over a metric space $(\mathcal{X}, d)$.*

$$W(\mathbb{P}, \mathbb{Q}) = \sup_{||f||_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}}[f(x)] - \mathbb{E}_{x\sim\mathbb{Q}}[f(x)],$$

where the $||f||_L$ denotes the Lipschitz constant of a function $f : \mathcal{X} \to \mathbb{R}$.

Let us ponder what the implications of this theorem are. We have translated the problem of looking over the space of couplings and find the best transportation plan to a problem of finding a good *witness* function in a particular function space that highlights as much as possible the differences between $\mathbb{P}$ and $\mathbb{Q}$.

Now, how does this help at all? Think about how you would parameterize an infinite-dimensional coupling in order to perform numerical optimization on it. The KR duality allows us to avoid working with the couplings directly, at the expense of having to search over a space of functions.

However we do know how to do this (c.f. universal approximation theorems for neural networks)! Select a parametric model for your witness function f_ϕ for $\phi \in \Phi$ and optimize the parameters ϕ so as to maximize the objective in Theorem 10. Moreover, since we have assumed a parametric family for the generator, the approach of [2] recovers a GAN-like zero sum game where f takes the role of the "discriminator".

$$\min_{\theta\in\Theta} \max_{\phi\in\Phi} \mathbb{E}_{x\sim\mathbb{P}}[f_\phi(x)] - \mathbb{E}_{z\sim\mathbb{P}_z}[f_\phi(g_\theta(z))]$$

Furthermore, there is a strong theoretical incentive regarding optimizing f_ϕ to convergence as achieving a tighter and tighter lower bound on the Wasserstein distance. Note that the fact that the optimization has to be performed over the space of 1-Lipchitz functions implies that proper care needs to be taken into account with respect to ensuring this

requirement is fulfilled. In the case of a neural network, a naive way to achieve this is by clipping the weights to a box $[0, 1]^p$, where p is the number of parameters of the network. However, finding better ways to do this is an active area of research.

References

[1] M. Arjovsky and L. Bottou. Towards Principled Methods for Training Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1701.04862, Jan 2017.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[3] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.