An Informal Introduction to the Maths Behind Wasserstein GAN

Qinxun Bai

February, 2017

Briefly speaking, [1] explains from a theoretical perspective why training GAN is hard, and [2] suggests a solution from the perspective of geometric foundation of generative modeling, which is more fundamental than GAN itself. The algorithm suggested by [2] ends up being another variant of GAN, purely due to optimization reasons, rather than being inherently designed to follow any form of adversarial training.

The problem with GANs

Given a generator $g : \mathcal{Z} \to \mathcal{X}$, and a distribution $p(z)$ on $\mathcal{Z}$, the generator's data distribution p_g is determined by the process of first sampling from $p(z)$ and then applying g to the samples. Denote the real data distribution as p_r, then explicit maximum likelihood methods are equivalent to minimizing $KL(p_r \| p_g)$. GANs, instead, minimize the following symmetric Jensen-Shanon divergence,

$$JSD(p_r \| p_g) = \frac{1}{2} KL(p_r \| p_a) + \frac{1}{2} KL(p_g \| p_a),$$

where $p_a = \frac{p_r + p_g}{2}$. The original minimax objective for GANs,

$$\min_g \max_D \{\mathbb{E}_{x \sim p_r}[\log D(x)] + \mathbb{E}_{x \sim p_g}[\log(1 - D(x))]\}, \tag{1}$$

where $D : \mathcal{X} \to \{0, 1\}$ is the discriminator. A revised objective function for optimizing g,

$$\max_g \mathbb{E}_{x \sim p_g}[\log(D(x))]. \tag{2}$$

While achieving intriguing success in generating realistic and sharp images, some mysterious problems during GAN training are widely observed:

1. Updates of the generator gets worse as D gets better, for both loss (1) and (2).
2. GAN training is massively unstable.

A recent work [1] studies these problems and makes important steps towards a better theoretical understanding of the training of GANs.

The support of p_r and p_g

In my opinion, the key observation of [1] is to focus on the supports of p_r and p_g, and relates them with a classical result in differential topology. Let $\mathcal{M} = \text{supp}(p_r), \mathcal{P} = \text{supp}(p_g)$. If $\mathcal{M} \cap \mathcal{P} = \emptyset$, it is straightforward that there exists a perfect discriminator that is constant on both $\mathcal{M}$ and $\mathcal{P}$ ($D^*|_{\mathcal{M}} = 1, D^*|_{\mathcal{P}} = 0$). If $\mathcal{M} \cap \mathcal{P} = \mathcal{L} \neq \emptyset$, the General Position Lemma gives the following marvelous result,

Theorem (General Position Lemma)**.** *If $\mathcal{M}$ and $\mathcal{P}$ are two submanifolds of $\mathbb{R}^d$ that do not have full-dimension, for almost every $a \in \mathbb{R}^d$, the perturbed submanifold $\mathcal{M} + a$ intersects transversely with $\mathcal{P}$, i.e., for every $x \in \mathcal{L}, T_x\mathcal{M} + T_x\mathcal{P} = T_x\mathbb{R}^d$.*

This lemma leads to horrible situations for GAN training. Since it is widely believed that real image data lies in low dimensional submanifold, and $\mathcal{P} \subset g(\mathcal{Z})$ also lies in a submanifold that does not have full dimension, as a result, transversality holds almost surely for $\mathcal{M}$ and $\mathcal{P}$. Then it is straightford to see that $\mathcal{L}$ has dimensionality strictly lower than that of both $\mathcal{M}$ and $\mathcal{P}$, i.e., $\mathcal{L}$ has measure 0 on both $\mathcal{M}$ and $\mathcal{P}$. Therefore, as in the disjoint case, there exists perfect discriminator that is constant on both $\mathcal{M}$ and $\mathcal{P}$.

Then rest of the reslults in [1] are not surprising. Such as KL-divergence will be infinity, JSD will be maxed out, either the updates to D is inaccurate or the gradients for training g will vanish. Note that (2) might not be a good choice either, causing unstable updates and mode dropping.

Recipe in practice

The key idea for a possible recipe is, either to break the basic assumptions of the General Position Lemma, i.e., to ensure the input to D has full dimension, or to use other distance measures that do not suffer from the General Position Lemma. Regarding the first direction, the paper [1] suggests adding continuous noise to the inputs of D, for both real data and generated data, which is also a typical recipe in machine learning. However, adding noise, in essence, alters the original problem and might degrade the quality of generated results. The second direction leads to paper [2], which in my opinion, is seminal and has vital importance beyond GANs.

Maximum Likelihood for generative modeling

Solving the typical MLE problem $\max_g \mathbb{E}_{x \in p_r}[p_g(x)]$ amounts to minimizing the Kullback-Leibler divergence $KL(p_r \| p_g)$ on a subspace (depending on $p(z)$ and

g) of the space of probability measures (denoted by $\text{Prob}(\mathcal{X})$). MLE has been used by default as the "golden standard" in machine learning for many years, yet whether KL-divergence is the most proper distance measure in $\text{Prob}(\mathcal{X})$ has not been studied carefully. The WGAN paper [2] analyzes the topology of $\text{Prob}(\mathcal{X})$ in the light of pursuing better continuity/convergence properties for doing optimization on $\text{Prob}(\mathcal{X})$. This is indeed the first step one should do for a given space before any geometry (such as distances, angles, gradients, etc) can be discussed. We summarize briefly in the last section the rigorous mathematical workflow for an arbitrary space on which we would like to study any geometry.

Topology, Convergence, and Continiuity

The notion of convergence and continuity can be well-defined without having a "distance/metric", but solely based on the topology. A **topology** associated with a space S is the collection of all "open" subsets of S. For a given S, it is possible to put different topologies on it by introducing different senses of "open". If $\mathcal{T}_1$ and $\mathcal{T}_2$ are two topologies on S, $\mathcal{T}_1$ is smaller (weaker in the language of [2]) than $\mathcal{T}_2$ if $\mathcal{T}_1 \subseteq \mathcal{T}_2$.

A point sequence $\{x_i\}$ **converges** in S, if there exists some $x_\infty \in S$, such that for any open set A (an element of the topology) containing x_∞, there always exists some N, such that for all $n > N, x_n \in A$. Then it is clear that the smaller the topology is, the easier it is for $\{x_i\}$ to converge under that topology, since there are "fewer" open sets to worry about.

Continuity is a property of a function, i.e., two (topological) spaces are involved. Say $f : M \to N$, where M has topology $\mathcal{T}_M$ and N has topology $\mathcal{T}_N$,

Definition 1. *f is* **continuous** *if for any open set $A(\in \mathcal{T}_N)$ on N, $f^{-1}(A)$ is open $(\in \mathcal{T}_M)$ in M.*

It can be easily shown that this definition of continuity leads to the following definition used in [2], if the notion of convergence is defined as above.

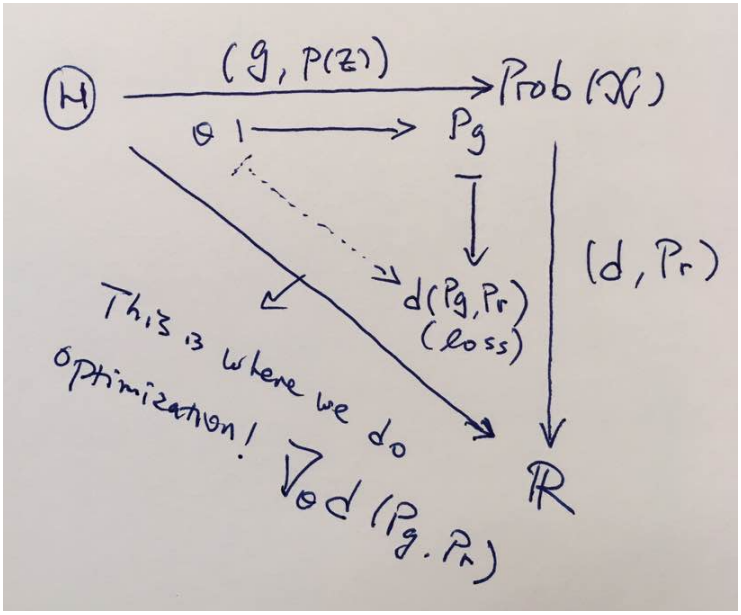
Definition 2. *f is* **continuous** *if for any convergent (under $\mathcal{T}_M$) sequence $\{x_i\}$ in M, the sequence $\{f(x_i)\}$ converges (under $\mathcal{T}_N$) in N.*

From both definitions of continuity, it is quite clear that the smaller $\mathcal{T}_N$ is, the easier it is for f to be continuous. This lies the foundation for the arguments of [2].

Remark. *Though the notion of topology does not depend on a pre-defined distance, if given a distance d on S, it naturally induces a topology, i.e., the set of all unions of open balls (an open ball is $B(x, r) = \{y \in S : d(x, y) < r\}$).*

Diagram of functional spaces that generative modeling algorithms deal with

Let Θ denote the parameter space of the parameterized generator (g_θ), which is a subspace of some Euclidean space and therefore comes with the standard topology, metric, and distance of Euclidean space. The following diagram commutes, so the continuity of function $(\theta \to d(p_g, p_r))$ is equivalent to the continuity of function $(\theta \to p_g)$.



Wasserstein distance for continuity concerns

[2] starts with GAN's problem raised by [1]: if $d(p_g, p_r) = KL(p_r \| p_g)$ (or its symmetric form, the Jensen-Shanon divergence) as is by default for MLE, due to the General Position Lemma, the function $(\theta \to d(p_g, p_r))$ on which our learning algorithm performs SGD, is unfortnately discontinuous, which makes gradient-based training very hard. Given the above relation between the topology on $\text{Prob}(\mathcal{X})$ and the difficulty of $(\theta \to p_g)$ being continuous, it is natural to consider a smaller topology on $\text{Prob}(\mathcal{X})$ which is more likely to enable a continuous $(\theta \to p_g)$, and therefore a continuous $(\theta \to d(p_g, p_r))$ for gradient-based training of θ.

[2] suggests putting the topology induced by the Wasserstein distance on $\text{Prob}(\mathcal{X})$, which is shown (Thm 2 of [2]) to be smaller than topologies induced by KL-divergence, Jensen-Shanon divergence, and Total Variation divergence. While this is just a qualitative result pointing to a more hopeful direction, [2]

further shows in Thm 1 that picking the Wasserstein distance as the loss function and putting its induced topology on $\text{Prob}(\mathcal{X})$ is indeed sufficient to enable a continuous $(\theta \to d(p_g, p_r))$ for training θ. Rest of the work is to design a computationally tractable algorithm under this setup.

Optimization with Wasserstein distance

While the Wasserstein distance and its induced topology have appealing theoretical properties, it is harder to compute/optimize than the Jensen-Shanon divergence. The K-R duality is then introduced to enable a tractable approximation of the Wasserstein distance, which in the end leads to a minmax optimization problem. This is exactly why the WGAN algorithm ends up being in the form of another GAN. From a higher level perspective, however, WGAN is quite different from the original GAN, with adversarial training being a computational compromise rather than the motivating idea.

Regarding the learning algorithm, WGAN makes the following modifications compared with the origninal GAN:

1. Remove the final sigmoid layer of the discriminator
2. No log for loss computation
3. Clamp the updated weights to a fixed closed interval
4. Avoid momentum based SGD methods

While the last modification depends largely on empirical observations, 1-3 are simple tricks directly inspired by the theoretical analysis. It is hardly possible for someone to simultaneously fix 1-3 in practice without the theoretical guidance.

A rigorous workflow to study the geometry of an arbitrary space

Given an arbitrary space S, we first need to define a topology on it, based on which, we can further study some fundamental "qualitative" properties, such as convergence, compactness, connectedness, and continuity. The crucial step from "qualitative" to "quantitative" study is the introduction of some Riemannian metric, which defines a smoothly varying (from point to point) inner product on the tangent space at each point of S. With a Riemannian metric specified, it is then possible (not always though) to define a series of "quantitative" geometric notions, such as angles, curvatures, distances/geodesics, areas/volumes, and gradients of functions on S.

References

[1] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.