Report project of Optimal Transport: Does Wasserstein-GAN approximate Wasserstein distances?

Mohamed Fyras Telmini

January 28, 2024

Abstract

This report serves to explain the different results of paper[1] and summazize them. Then, a comprehensive experiment is conducted with synthetic data in order to see if the paper's critic network is in fact approximating Wasserstein distances or not.

1 Paper summary: Wasserstein GAN

The paper is concerned with estimating probability densities of inputs, in the context of generative networks. The authors begin by explaining how instead of estimating the density of the real data directly, which may not exist, they suggest defining a random variable with a fixed distribution and passing it through a parametric function that directly generates samples. This approach is akin to the Generative-Adversarial-Approaches which has a generator component that is tasked with correctly generating samples from the desired distribution. This approach is well-known and commonly used. The main aim of the paper is the focus on the critic part of the GAN network, as it directly relates to the notion of distance between 'real' samples and 'fake' samples. First, they provide a theoretical analysis on how the Wasserstein distance is a suitable candidate for the critic's distance, and compare it to other more common distances. Then they formally define the Wasserstein-GAN and argue on its usefulness. The authors also then discuss the empirical benefits of Wasserstein GANs (WGANs) over traditional GANs. Since a full summary of the paper is not part of the task, i will only summerize the part where they formulate the estimation for the Wasserstein distance.

1.1 Wasserstein distance as a GAN loss function

After theoretically justifying their choice of the Wasserstein distance as a suitable loss function for the GAN problem, the authors then introduce the main idea that allows for the estimation of the Wasserstein distance. First, Wasserstein distance is defined as:

$$W(Pr, Pg) = \inf_{\gamma \in \Pi(Pr, Pg)} E_{(x,y)\sim\gamma}[\|x - y\|] \tag{1}$$

The authors use the Kantorovich-Rubinstein duality:

$$W(Pr, P) = \sup_{\|f\|_L \leq 1} E_{x\sim Pr}[f(x)] - E_{x\sim P}[f(x)] \tag{2}$$

Knowing this formulation, the problem of estimating 1 becomes the problem of finding the correct function f that maximizes 2. The authors suggest training a neural network parameterized with weights w lying in a compact space W and then backpropagating through the expectation of the gradient of f. This is done within the context of GANs and allows the critic to be trained in order to simulate the Wasserstein distance correctly.

2 Empirical verification of the validity of the critic network as a Wasserstein distance estimator

By exploiting the results explained above, the main component of interest in the network is the critic network. The idea is that the critic is trained on a set of 'real' images and 'fake' images, and is able to determine the correct Wasserstein distance between the two. 'fake' images are usually samples of the 'real' images which have additive noise. A main issue with this is that the critic cannot generalize to be a good estimator of the Wasserstein distance between two distributions if its not properly trained with both of them. Which means that recovering an already trained checkpoint from a GAN serves no real purpose since the images on which it had been trained on don't have analytically determined ground truth distances. And the task of determining the consistency of the critic's estimation would only be done versus other estimators (in our case we will use the sinkhorn estimator) and hold little to know value since we don't know for sure if either of them is close to the real, extremely computationally expensive in this case, optimal transport.
A more sound approach is to then train a critic network on two distributions, of which we already know the Wasserstein distance. In this case, 3-D multivariate gaussian distributions are used. This is because they can be mapped to RGB images with 3 color channels, each channel corresponding to a coordinate in the distributions. In particular, the following two distributions are used:

$$X \sim \mathcal{N}(\mu_1, \Sigma_1)\mu_1 = \begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} \Sigma_1 = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 2.0 & 0.0 \\ 0.0 & 0.0 & 3.0 \end{pmatrix} \tag{3}$$

$$Y \sim \mathcal{N}(\mu_2, \Sigma_2)\mu_2 = \begin{pmatrix} 3.0 \\ 3.0 \\ 3.0 \end{pmatrix} \Sigma_2 = \begin{pmatrix} 2.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix} \tag{4}$$

Various images are then constructed iteratively by sampling from X and Y and then mapping the coordinates of each samples to its corresponding color channels. This seemed to be the most intuitive way to represent a desired distribution as an image. Roughly 3000 images from each distribution were generated for the training of the critic network, with 300 left for testing it. Samples were generated through JAX. The Wasserstein distance

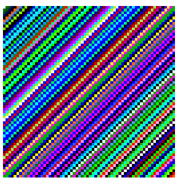

Figure 1: Example of a generated image

between the distributions from which X and Y are sampled is known analytically, and represents a valid baseline for the experiment, additionally I reported on the sinkhorn estimation in order to be certain of the validity of the analytical solution (both should be quite close). This provides an additional valid estimator to support the validity of the approach.
Since the way the critic network works is by maximizing the output for the 'real' sample and minimizing the

output for the 'fake' sample. The proxy for the distance is then equivalent to the solution of 2 but is always in [-1,1]. So if the network is effectively able to estimate the Wasserstein distance, it would be through a coefficient λ. With

$$W(Pr, P) = \lambda \| critic(Pr) - critic(P) \| \tag{5}$$

We can obtain λ through the real known value of $W(X, Y)$ such that $\lambda = \frac{W_{real}(X,Y)}{\|critic(x) - critic(y)\|}$ with x,y samples from X and Y. We can then test the critic on a third new random variable, and see if it can correctly estimate the distance.

3 Experiments and results

The experiments described above are executed in the colab notebook: OTT-Work.ipynb. The notebook contains the most major part of the conducted work, from generating samples to training and testing the critic network. The results of the experiments are specified in it. The third variable Z was chosen such as:

$$Z \sim \mathcal{N}(\mu_3, \Sigma_3)\mu_3 = \begin{pmatrix} 1.0 \\ 1.0 \\ 1.0 \end{pmatrix} \Sigma_3 = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 2.0 & 0.0 \\ 0.0 & 0.0 & 3.0 \end{pmatrix} \tag{6}$$

Bibliography

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. 2017. arXiv: `1701.07875 [stat.ML]`.