Wasserstein Generative Adversarial Network (WGAN)

Project Code: C17, Tirath Sojitra (110029415), and Sanketkumar Patel (110023916)

I. Literature Review

THE paper has introduced WGAN, an enhanced version of GAN. It is an alternate way of training the generator model to better approximate the distribution of data obtained from the real data distribution. Instead of using a discriminator model to classify the probability of generated images as being real or fake, WGAN model replaces it with a critic that keeps track of authenticity of generated images [1]. There exists various ways to measure how close the model distribution and the real distribution are, or equivalently, on the various ways to define a distance or divergence. The most common way is to find the distance between their probability distributions [2].

WGAN uses 1-Wasserstein distance to measure the distance between the distributions whereas GAN uses JS-Divergence. The paper shows how Earth Mover (EM) or Wasserstein distance is better than JS-Divergence and KL-Divergence used for VAEs. Although GAN has great ability to generate realistic images, the training is not easy. One of the most common failures of GANs is referred to as Mode Collapse. After training the generator to some extent, it produces the same outputs [3].

This report will provide a comprehensive theoretical analysis of how the Earth Mover (EM) distance behaves in comparison to other popular probability distances and divergences used in the context of learning distributions and will show that the corresponding optimization problem is sound [2].

II. Theory and Concepts

A. *Popular Distances*

- The *Total Variance* (TV) distance.

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A\in\Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)| \ . \quad (1)$$

- The *Kullback-Leibler* (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x) d\mu(x) \quad (2)$$

  Here $P_r(x)$ and $P_g(x)$ is probability distributions of real dataset and generates data, respectively. The KL divergence is asymmetric and results in infinite if the $P_g(x)$ is zero and $P_r(x)$ is non-zero.

- The *Jensen-Shannon* (JS) divergence

$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m) \ , \quad (3)$$

  where P_m is the average of P_r and P_g. This divergence is always symmetric.

- The *Earth-Mover* (EM) distance or Wasserstein-1

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma\in\Pi(\mathbb{P}_r,\mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\, \|x-y\| \,\big] \ , \quad (4)$$

  where $\Pi(P_r,Pg)$ contains all the possible transport plan $\gamma(x,y)$. It indicates how much mass must be transported form x to y to convert the distribution P_r into the distribution P_g. Then the cost to transport the most optimal transport plan is called EM distance.

B. *Learning through an example*

This example will illustrate how probability distributions will converge under EM distance but do not converge under any other distance.

Let $Z \sim U$ [0,1] the uniform distribution on the unit interval. Let P_0 be the distribution of $(0,Z)$ and $g_\theta = (\theta,z)$. with θ a single real parameter. The distances will result as follows,

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$,
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- $KL(\mathbb{P}_\theta \| \mathbb{P}_0) = KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- and $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ . \end{cases}$

The above example shows a case where we can learn a probability distribution in low dimensional using EM distance. This cannot be done with the other distances and divergences because the resulting loss function is not even continuous. It means the gradient of the divergency will eventually diminish. We have close to a zero gradient, i.e. the generator learns nothing from the gradient descent [4]. Its result into mode collapse problem in GANs.

C. *Wasserstein GAN*

Rather than adding noise, Wasserstein GAN proposed a new cost function with more continuous gradient with the help of Wasserstein distance. WGAN learns regardless the performance of generator. The fig 1 below represents the plot of GAN and WGAN. GAN fills the areas with diminishing gradient. Meaning the learning does not takes place after certain distribution. On the other side, WGAN has smoother and continuous gradient irrespective of the quality image obtained by the generator.

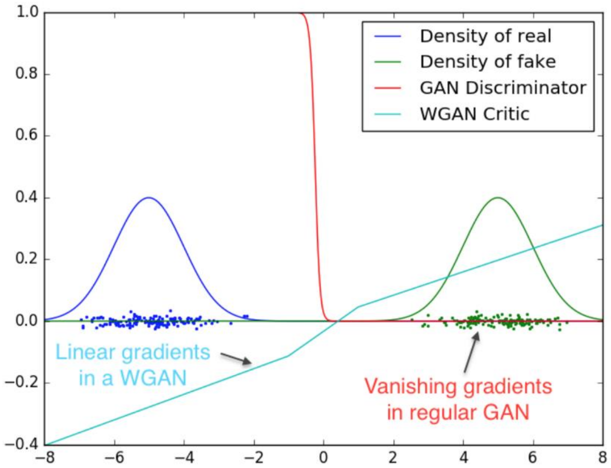

Figure 1: Optimal discriminator and critic when learning to differentiate two Gaussians. As we can see, the traditional GAN discriminator saturates and results in vanishing gradients. Our WGAN critic provides very clean gradients on all parts of the space [2].

However, the equation for the Wasserstein distance is highly intractable. The cost function of WGAN is denoted by,

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)] \tag{5}$$

where *sup* is the least upper bound, ***f*** is a 1-Lipschitz function, P_r and P_θ is the probability distributions of real dataset and generated dataset, respectively. So, to calculate the Wasserstein distance, we just need to find a 1-Lipschitz function.

The network design shown in fig 2 is almost the same except the critic does not have an output sigmoid function. Another change is that it outputs a scalar score rather than a probability. This score can be interpreted as how real the input images are. In reinforcement learning, we call it the **value function** which measures how good a state (the input) is. We rename the discriminator to **critic** to reflect its new role [4].

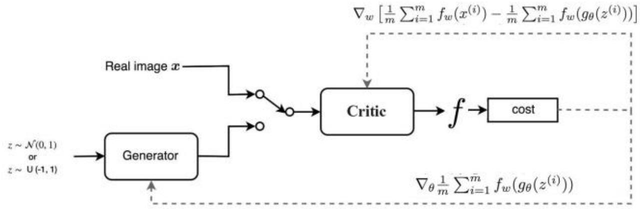
$\nabla_w \left[\frac{1}{m}\sum_{i=1}^m f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))\right]$
Real image x
Critic
f
cost
$z \sim \mathcal{N}(0,1)$ or $z \sim$ U (-1, 1)
Generator
$\nabla_\theta \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))$

Fig 2: Flow Chart of WGANs [4].

III. Algorithm

Now we can put everything together in the pseudo-code below.

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

```
Require: : $\alpha$, the learning rate. $c$, the clipping parameter.
    $m$, the batch size. $n_{\text{critic}}$, the number of iterations of the
    critic per generator iteration.
Require: : $w_0$, initial critic parameters. $\theta_0$, initial generator's parameters.
while $\theta$ has not converged do
    for $t = 0, \ldots, n_{\text{critic}}$ do
        Sample $\{x^{(i)}\}_{i=1}^m \sim \mathbb{P}_r$ a batch from the real data.
        Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of priors.
        $g_w \leftarrow \nabla_w [\frac{1}{m}\sum_{i=1}^m f_w(x^{(i)})$
                 $-\frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))]$
        $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
        $w \leftarrow \text{clip}(w, -c, c)$
    end for
    Sample $\{z^{(i)}\}_{i=1}^m \sim p(z)$ a batch of prior samples.
    $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^m f_w(g_\theta(z^{(i)}))$
    $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while
```

IV. Conclusion

The paper introduced an algorithm named WGAN, an alternative to traditional GAN training. It has overcome the problems like mode collapse that happened in GANs and provide meaningful learning curves useful for debugging and hyperparameter searches. In this new model, researchers has done extensive theoretical work to show how Wasserstein distance is better choice compared to other distances.

References

[1] J. Brownlee, “How to Develop a Wasserstein Generative Adversarial Network (WGAN) From Scratch,” *Machine Learning Mastery*. [Online}Available:https://machinelearningmastery.com/how-to-code-a-wasserstein-generative-adversarial-network-wgan-from-scratch/. [Accessed: 04-Feb-2021].

[2] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” arxiv.org. [Online]. Available: https://arxiv.org/pdf/1701.07875.pdf. [Accessed: 04-Feb-2021].

[3] L. Weng, “From GAN to WGAN,” Lil'Log, 20-Aug-2017. [Online]. Available: https://lilianweng.github.io/lil-log/2017/08/20/from-GAN-to-WGAN.html. [Accessed: 04-Feb-2021].

[4] Hui, J. (2020, March 05). GAN - Wasserstein gan & WGAN-GP. Retrieved February 06, 2021, from https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490