Dauphine | PSL
UNIVERSITÉ PARIS

f-gan

Professors:

Alexandre Verine
Benjamin Negrevergne

Group Name:

Brasil AI

Students:

Artur Dandolini Pescador
Caio Azevedo

November
2023

Contents

1 Introduction

The main objective of this project is to implement and to understand the capabilities of Generative Adversarial Networks (GANs). It uses two neural networks that are trained simultaneously in a competitive setting. The objective is to generate of highly realistic synthetic data.

Initially, our objective was set on implementing the Wasserstein GAN (W-GAN) [1]. W-GAN introduces a new way to measure the distance between two distributions (the one from the data generated by the GAN and the distribution of the real data) using the Wasserstein distance.

After understanding and implementing the W-GAN, it was decided to follow on another way towards f-GAN [2]. The general idea of the f-GAN framework is to generalize the GAN objective to a broader class of divergences, called f-divergences.

This report contains some explanations of our initial implementation of W-GAN and also of f-GAN. The results will be shown throughout this report to see how each approach dealt with the MNIST dataset.

2 Dataset

The dataset that was used for training our GAN model and subsequently for generating new images was the MNIST dataset.

The MNIST dataset is a large database of handwritten digits. It contains 70,000 images of handwritten digits divided into a training set of 60,000 examples and a test set of 10,000 examples. Each image is a grayscale image, size normalized and centered in a fixed-size (28x28 pixels) frame.

The simplicity and size (28 x 28 pixels) of the MNIST dataset make it ideal for training our GAN model.

3 Architecture

Originally, the model given for training consisted in identical (except for the last activation) multi-layer perceptrons, with four layers of increasing (in the generator) or decreasing (in the discriminator) number of neurons.

By tuning the learning rate and trying different initialization schemes one could have found a setting in which this network produces high quality samples. Since in all settings we tried, we ran into the problem of vanishing gradients or severe instabilities, we decided to simplify the network and change final activations to arrive at reasonable generations.

4 W-GAN

The Wasserstein GAN [1] proposes that the Earth-Mover distance W between distributions, shown in Equation 1, has desirable properties which make it a better objective function to be optimized than usual ones such as Kullback-Leibler or Jensen-Shannon divergences. We consider

$$W(P||Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma} \left[\|x - y\|\right] \tag{1}$$

where $\Pi(P, Q)$ is the set of all joint distributions with marginals P and Q. Through what is called the Kantorovich-Rubinstein equation, the authors show that, if we let g_θ be the generator and f_w the discriminator, and add some kind of constraint to keep f_w in the family of K-Lipschitz functions for some constant K, we can approximately minimize the EM-distance by solving the optimization problem

$$\max_{w \in \mathcal{W}} \mathbb{E}_{x \sim P_r} \left[f_w(x) \right] - \mathbb{E}_{z \sim p(z)} \left[f_w(g_\theta(z)) \right]. \tag{2}$$

We implement this method and are able to obtain, as expected, a much more stable training procedure and high quality samples, although very low variability in the modes generated (precision 0.60, recall 0.08). Generations can be seen in Figure 1.

Figure 1: Example generations from our W-GAN implementation.

Since the W-GAN is a very well-known upgrade to the original algorithm, we decide to shift focus and explore another research direction, that of the f-GAN, which allows for a more general understanding of the training procedure for GANs.

5 f-GAN

In recent advancements in generative modeling, f-GANs (f-divergence Generative Adversarial Networks) introduce a more generalized approach by utilizing f-divergences to measure and minimize the discrepancy between the model's generated distributions and the real data distribution.

The f-GAN framework extends the standard GAN by the possibility of using different divergence functions, each one tailored to different aspects of distribution similarity. This aspect gives more contrel over the training process and outcomes making it suitable for different applications in generative models.

The f-divergence is a way to quantify the difference between two probability distributions. It is a class of functions that measure the divergence between two probability distributions P and Q.

These are defined as:

$$D_f(P||Q) = \int_X q(x) f\left(\frac{p(x)}{q(x)}\right) dx \tag{3}$$

where P and Q are two probability distributions with density functions p and q, respectively and f is a convex function with the property that $f(1) = 0$, ensuring that the divergence is zero if and only if $P = Q$.

Some of the most known divergences include the Total Variation, Kullback-Leibler divergence, and Jensen Shannon, among others. Each of them captures different aspects of the distributions, making them suitable for a bunch of different tasks.

The image below illustrates a range of f-divergences with their respective generating functions and transformations.

Name	$D_f(P\|Q)$	Generator $f(u)$	$T^*(x)$
Total variation	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \,\mathrm{d}x$	$\frac{1}{2}\lvert u-1\rvert$	$\frac{1}{2}\operatorname{sign}(\frac{p(x)}{q(x)}-1)$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\,\mathrm{d}x$	$u\log u$	$1+\log\frac{p(x)}{q(x)}$
Reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\,\mathrm{d}x$	$-\log u$	$-\frac{q(x)}{p(x)}$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)}\,\mathrm{d}x$	$(u-1)^2$	$2(\frac{p(x)}{q(x)}-1)$
Neyman χ^2	$\int \frac{(p(x)-q(x))^2}{q(x)}\,\mathrm{d}x$	$\frac{(1-u)^2}{u}$	$1-\left[\frac{q(x)}{p(x)}\right]^2$
Squared Hellinger	$\int \left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\,\mathrm{d}x$	$(\sqrt{u}-1)^2$	$(\sqrt{\frac{p(x)}{q(x)}}-1)\cdot\sqrt{\frac{q(x)}{p(x)}}$
Jeffrey	$\int (p(x)-q(x))\log\left(\frac{p(x)}{q(x)}\right)\,\mathrm{d}x$	$(u-1)\log u$	$1+\log\frac{p(x)}{q(x)}-\frac{q(x)}{p(x)}$
Jensen-Shannon	$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$	$-(u+1)\log\frac{1+u}{2}+u\log u$	$\log\frac{2p(x)}{p(x)+q(x)}$
Jensen-Shannon-weighted	$\int p(x)\pi\log\frac{p(x)}{\pi p(x)+(1-\pi)q(x)}+(1-\pi)q(x)\log\frac{q(x)}{\pi p(x)+(1-\pi)q(x)}\,\mathrm{d}x$	$\pi u\log u-(1-\pi+\pi u)\log(1-\pi+\pi u)$	$\pi\log\frac{p(x)}{(1-\pi)q(x)+\pi p(x)}$
GAN	$\int p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x-\log(4)$	$u\log u-(u+1)\log(u+1)$	$\log\frac{p(x)}{p(x)+q(x)}$
α-divergence ($\alpha\notin\{0,1\}$)	$\frac{1}{\alpha(\alpha-1)}\int\left(p(x)\left[\left(\frac{q(x)}{p(x)}\right)^{\alpha}-1\right]-\alpha(q(x)-p(x))\right)\,\mathrm{d}x$	$\frac{1}{\alpha(\alpha-1)}\left(u^{\alpha}-1-\alpha(u-1)\right)$	$\frac{1}{\alpha-1}\left[\left[\frac{p(x)}{q(x)}\right]^{\alpha-1}-1\right]$

Figure 2: f-divergence functions

The convex conjugate of f, denoted as f^*, allows us to rewrite D_f as:

$$D_f(P \parallel Q) = \sup_{T\in\mathcal{T}} \left(\mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))]\right), \tag{4}$$

where $\mathcal{T}$ is the set of all measurable functions for which the expectations are defined.

From this formulation, we can derive the objective function for the f-GAN training process:

$$F(\theta,\omega) = \mathbb{E}_{x\sim P}[T_\omega(x)] - \mathbb{E}_{x\sim Q_\theta}[f^*(T_\omega(x))]. \tag{5}$$

We train the discriminator T_ω to maximize the objective function, while the generator Q_θ is trained to minimize it.

The generative capabilities under each divergence are illustrated through a series of generated images and the accumulated loss over epochs below.

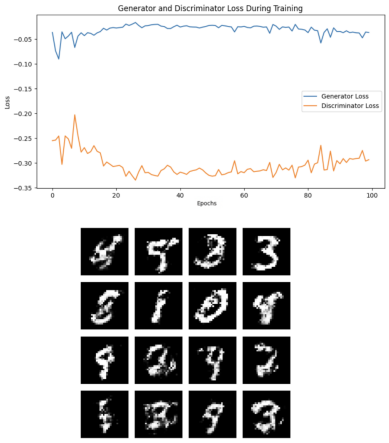

Figure 3: Total Variational

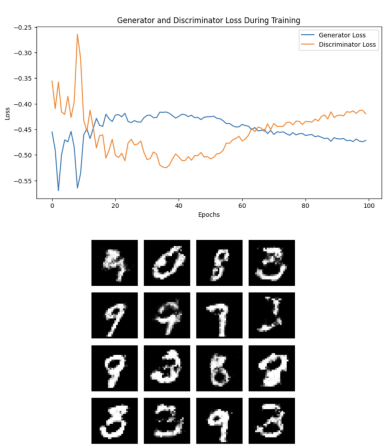

Figure 4: Forward KL

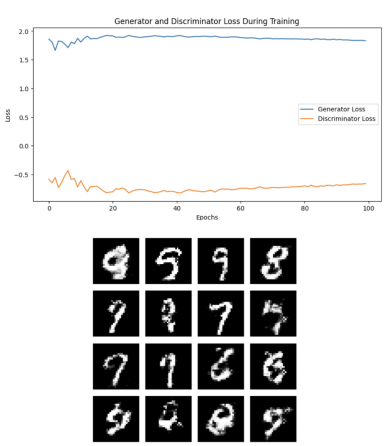

Figure 5: Reverse KL

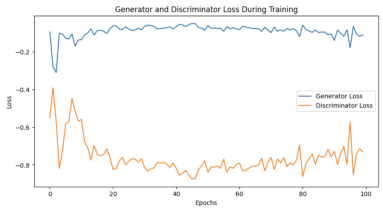

Figure 6: Pearson

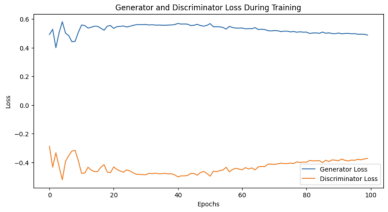

Figure 7: Hellinger

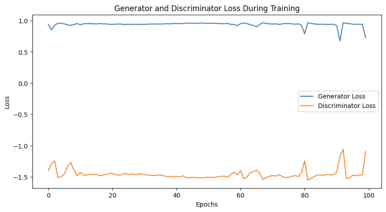

Figure 8: Jensen-Shannon

Based on the generated images above, it is possible to observe that each of divergence functions produce the most varied (high recall) and accurate (high precision) images. The plots show the loss over epochs, which gives some insights about the stability and convergence speed of the model under each divergence.

5.1 Precision-Recall Trade-offs in f-GANs

Achieving a balance between precision and recall is really important for generative models. Precision measures the quality of the generated samples, while recall assesses the diversity. A novel approach to optimize this balance is through the concept of PR-divergence, as introduced by Verine et al. [3].

PR-divergence is introduced to explicitly control the trade-off between precision and recall during the training of generative models by a new parameter λ that is introduced into the f-divergence function. It is defined as follows:

Given a trade-off parameter $\lambda \in [0, +\infty]$, the PR-divergence ($D_{\lambda-PR}$) is defined for a function $f_\lambda : \mathbb{R}_+ \to \mathbb{R}$ given by:

$$f_\lambda(u) = \max(\lambda u, 1) - \max(\lambda, 1) \tag{6}$$

for $\lambda \in \mathbb{R}_+$ and $f_\lambda(u) = 0$ for $\lambda = +\infty$.

It is possible to adjusts the emphasis on precision or recall by the choice of λ.

The results of MNIST generations with $\lambda = 1$ and $\lambda = 10$ are as follows (9 and 10 respectively).

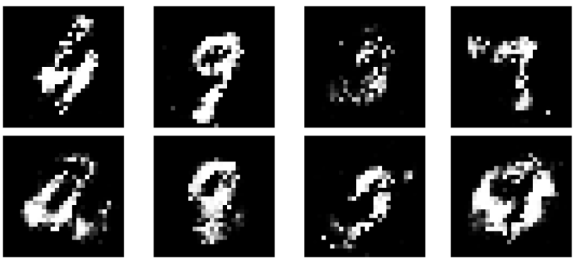

Figure 9: MNIST generations with $\lambda = 1$

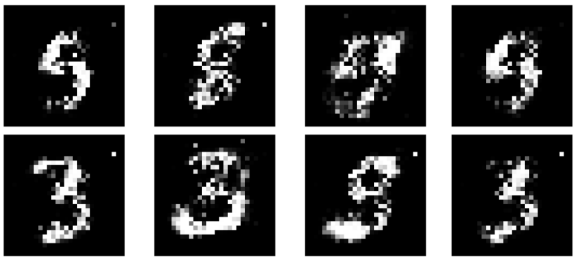

Figure 10: MNIST generations with $\lambda = 10$

As it is possible to see in the figures 9 and 10, increasing λ from 1 to 10 results in samples with higher precision (clearer and more defined shapes of the digits) but with lower recall

indicated by the reduced variability among the samples. This is consistent with the theoretical framework of PR-divergence [3], where a higher λ value emphasizes precision over recall.

6 Conclusion

In conclusion, f-GANs (f-divergence Generative Adversarial Networks) is a good way to have a more general approach in generative modeling. By employing f-divergences to measure the similarity between generated and real data distributions, f-GANs provide a flexible framework that can adapt to various types of data and applications.

The introduction of PR-divergence, as detailed in the work of Verine et al. [3], further enhances the capabilities of f-GANs. By incorporating a parameter λ into the f-divergence function to control the trade-off between precision and recall during the training of generative models.

This adaptability and control make f-GANs a valuable addition to the general GAN to have more flexibility on the training process.

References

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

[2] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. arXiv preprint arXiv:1606.00709, 2016.

[3] Alexandre Verine, Benjamin Negrevergne, Muni Sreenivas Pydi, and Yann Chevaleyre. Precision-recall divergence optimization for generative modeling with gans and normalizing flows. arXiv preprint arXiv:2305.18910, 2023. NeurIPS 2023.

[1] [2] [3]