Comparison of Metrics Used in Generative Adversarial Networks

Marco Jiralerspong

December 18, 2019

Abstract

We compare various value functions used in generative adversarial networks (GAN) and the shift from a maximum likelihood approach to metrics on the space of probability measures. Specifically, we study the Cramér GAN and WGAN, examining how they incorporate their associated metrics into the GAN framework, how they deal with the issues caused by maximum likelihood value functions and the assumptions they implicitly make.

1 Introduction

Generative models differ from discriminative models by using a dataset to model the probability distribution of the dataset instead of focusing solely on class prediction. While generative models can be used for classification tasks, they also allow, in certain cases, for the generation of new samples that resemble the provided dataset. Significant improvements have been made to models with the latter goal in recent years. In fact, the faces they are capable of generating have become almost indistinguishable [1] from real faces and are no longer the low resolution, noisy mess they once were. However, for much of the history of machine learning, such generative models were typically overlooked in favor of models designed for classification.

[1] Easily accessible examples are available on `https://thispersondoesnotexist.com/`

2 Original GAN

2.1 Model Structure

Nonetheless, the fate of this subset of machine learning changed drastically with the seminal[2] paper on generative adversarial networks (GAN) [7]. Instead of directly attempting to approximate the underlying probability distribution, they instead model the problem as a game between two agents (a generator and a discriminator). The former takes as input a random vector $\boldsymbol{z}$ and outputs a new data point $G(\boldsymbol{z})$ while the latter takes as input a data point (either $\boldsymbol{x}$ or $G(\boldsymbol{z})$) and outputs its estimate of the probability that its input is an actual data point (either $D(\boldsymbol{x})$ or $D(G(\boldsymbol{z}))$).

The goal of the generator is to generate sample points that are indistinguishable from those of the original dataset. The discriminator, on the other hand, seeks to learn to do the exact opposite (i.e. to distinguish real from generated points). Consider the analogy of an art forger given a series of paintings by a famous artist and an art collector. The art forger is then given a random object, say a chair, and must paint that chair following the style of the famous artist in such a way that the art collector is convinced of its authenticity.

More formally, the goals of the agents are encoded in the game through the value function and its associated payoffs. For a given generator G and discriminator D, the value function is defined as

$$V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_x}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_z}[\log(1 - D(G(\boldsymbol{z})))]$$

where $\boldsymbol{z}$ is some random noise vector (from some predefined distribution p_z) and p_x is the true distribution of the dataset. Both D and G consist of neural networks parametrized by θ_G and θ_D respectively. This value function uses the maximum likelihood approach which is pervasive in machine learning.

2.2 Game Theory Interpretation

As the value function measures how right the discriminator is, it clearly wants to maximize the value function while the generator wants to minimize it. Thus, by setting the payoff of the discriminator to be the output of the value function and the payoff for the generator to be its negation, we encode their

[2] While the originality of the idea of adversarial networks is contested (see [13] and "Predictabality Maximization") its impact on the field is undeniable.

goals and create a 2-player zero-sum game.

For such games, Nash equilibria are known to correspond to each player using their safety strategy [15] (i.e. the strategy that maximizes their payoff given that the other player wants to minimize it) and hence

$$G^* = \operatorname*{argmin}_G \max_D V(D, G)$$

$$D^* = \operatorname*{argmax}_D \min_G V(D, G)$$

corresponds to a mixed Nash equilibrium with value

$$\min_G \max_D V(D, G) = \max_D \min_G V(D, G)$$

Interestingly, Goodfellow et al. never refer to Nash equilibria in the original paper (even though they explicitly model the problem as a minimax game). They even go as far as proving that solving the above yields a value of $\log(\frac{1}{2}) + \log(\frac{1}{2})$ for the generator.

This value is attained precisely when the generator, given the distribution p_z yields a distribution $G(p_z)$ that perfectly matches p_x. Thus the discriminator is forced to assign probability $\frac{1}{2}$ to every sample given to it. However, this value is refered to as a global minimum and it is not until the tutorial [6] that they mention that it corresponds to a Nash equilibrium.

To reach this Nash equilibrium, the GAN trains both models simultaneously using batch gradient descent. m random vectors are sampled from p_z and run through the generator while m samples are chosen randomly from the dataset. The value function is computed for the sample and then both set of parameters are updated using the gradient of the value functions.

3 Kantorovich-Wasserstein metric

Before proceeding further a quick note on nomenclature is necessary. Throughout the papers referenced here, the distance between 2 probability measures P, Q given by

$$KW_p(P, Q) = \inf_{\pi \in \Pi(P,Q)} \left(\int_{X \times X} d(x, y)^p d\pi \right)^{\frac{1}{p}}$$

where $\Pi(P, Q)$ is the set of all couplings has been called the Wasserstein distance, the Mallow's distance, the Earth-Mover's distance, the KW distance and different permutations of the above. As per Prakash Panangaden, an appropriate name would be the Kantorovich-Wasserstein (KW) distance and in this paper it will be henceforth be referred to as such.

3.1 Definition and Model Structure

We now consider the more recent work [1, 2]. While both maintain the overall structure of the GAN some fundamental changes are made to accommodate their updated value function.

Arjovsky et al. begins by giving the usual definition of the KW metric (i.e. as the infimum over couplings), while Bellemare et al. define the class of p-KW metrics KW_p (for 2 distributions P and Q) as:

$$KW_p(P, Q) := \left(\int_0^1 |F_P^{-1}(u) - F_Q^{-1}(u)|^p du \right)^{\frac{1}{p}}$$

where F_P^{-1} and F_Q^{-1} are the inverse CDFs of P and Q respectively. While this definition is perhaps more approachable, it also only applies to univariate probability distributions [10] and as such is not as relevant for many of the applications discussed in the paper (especially since most research around GANs concerns image generation).

Nonetheless, both papers eventually give the equivalent dual definition (one referring to it as Kantorovich-Rubinstein duality and the other as Kantorovich-Monge duality):

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

Arjovsky et al. are quick to note the computational issues associated with maximizing a quantity over the space of 1-Lipschitz functions.

To remedy this in their WGAN, they creatively resort to using a neural network to model the function f since any continuous function can be approximated using an appropriate neural network structure [9]. This neural network, acting as a critic is trained using the gradient of the associated estimated KW distance, its goal being to maximize $\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim Q}[f(x)]$ and approximate the function where the supremum is attained (exists, as shown in class).

The use of a critic marks a departure from the original idea of GANs. Going back to the analogy of the art forger and the art collector, we can now think of the art collector as an art critic. Given the paintings of the forger he gives him advice on how his paintings can be improved to better match the style of the famous painter while simultaneously learning to give more accurate feedback. In fact, the critic is now used to estimate the distance between the generated distribution and the true distribution. In doing so, he provides gradients for the generator to train (the advice in the analogy)

As for the Lipschitz constraint, in the first WGAN [1] it is enforced through weight clipping of the parameters of the generator neural network. In a later paper this is changed to the use of a gradient penalty [8] (the new model is called WGAN-GP).

3.2 Issues with Approach

While the model performs well, an issue that is glossed over is the validity of the estimate of the KW distance. For example, consider an initial generator G and a generator G' obtained from G after one gradient descent step. Seeing that $G(p_z)$ will not be equal to $G'(p_z)$, the function f that maximizes $\mathbb{E}_{x\sim p_x}[f(x)] - \mathbb{E}_{z\sim p_z}[f(G(z))]$ will also conceivably be different from the function that maximizes $\mathbb{E}_{x\sim p_x}[f(x)] - \mathbb{E}_{z\sim p_z}[f(G'(z))]$.

Thus, at each training step, the WGAN critic is trying to estimate a different function which could cause potential issues with its reliability as an estimator. Combined with the fact that this estimate is a sample estimate of the distance potentially explains why Arjovsky et al. train the critic more often than the generator (5 steps for each gradient descent step of the generator). In the same vein, experimental results from Bellemare et al. show that the generator performs much better when the critic is trained 5 times instead of 1 for each step performed on the discriminator. While the experimental results are promising, it would be helpful to have some bound on the accuracy of the estimate of the distance.

A similar issue is found in [2] where they use a critic (i.e. get the function f represented by the trained neural network) from WGAN-GP trained on the same dataset. Once again, the accuracy of this estimate is not discussed and is potentially more problematic than the previous case.

In fact, in the previous example, we at least had that G' is obtained from G through one gradient descent step and thus could be assumed to be reasonably close. Such an argument wouldn't apply for the case of the Cramér GAN. If such GANs are substantially different from WGANs, then it is even more likely that the f that maximizes the estimate of KW distance for one of the generators is different from the one that does the same for the other generator.

Essentially, two results are needed/missing to guarantee the soundness of the training procedure:

- Some bound of $|\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)] - W(p_{G'}, p_x)|$ given that f is a function that attains the supremum of $\mathbb{E}_{x\sim P}[f(x)]-\mathbb{E}_{x\sim Q^*}[f(x)]$ for some other distribution Q^*. With such a bound, we could get some estimate of the validity of training a critic function f on a dataset P and generator G to then reuse this f to estimate the KW distance between some other generator G' and the same dataset P.

- Some result on the convergence of the critic to the true KW distance. There is no result in either paper that indicates that performing gradient descent with respect to their estimate of the distance would converge to a global optimum (the true distance).

Nonetheless, the empirical results of Arjovsky et al. seem to indicate that, at least in the setting it was used, their estimate of the KW distance is sound. Indeed, one of the advantages they mention is the usefulness of their approximate metric as a gauge of sample quality. They observe that the approximate KW distance correlates well with visual quality of generated images, a property not observed for other GAN losses based on maximum likelihood.

4 Cramér and Energy Distances

4.1 Definition

While WGAN performs well, Bellemare et al.'s propose the Cramér distance (and its multi-dimensional analog, the energy distance) as an alternative to the KW distance that doesn't suffer from biased sample gradients. Simply put, the Cramér distance is just the l_2^2 distance between the respective cumulative distributions of P and Q [2]

$$l_2^2(P, Q) := \int_{\mathbb{R}} (F_P(x) - F_Q(x))^2 dx$$

where F_P is the CDF of P and F_Q is the CDF of Q.

As for the energy distance, let $X, X' \sim P$ and $Y, Y' \sim Q$ be independent. The energy distance between P and Q is then [2]

$$\epsilon(P, Q) := \epsilon(X, Y) := 2\mathbb{E}||X - Y||_2 - \mathbb{E}||X - X'||_2 - \mathbb{E}||Y - Y'||_2$$

with $l_2^2(P, Q) = \frac{1}{2}\epsilon(P, Q)$ when P, Q are on $\mathbb{R}$ [14].

Interestingly, they comment on the simililarity of l_p metrics and the KW metric, specifically through the dual of the l_p metrics, given by

$$l_p(P, Q) = \sup_{f \in F_q} |\mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{x \sim Q} f(x)|$$

Since F_q is the set of absolutely continuous functions where $||\frac{df}{dx}||_q \leq 1$ such that $\frac{1}{p} = \frac{1}{q} = 1$, they state that for $p = 1$ (and thus $q = \infty$) we have equivalence of the l_p metric and the KW metric.

4.2 Model Structure

Seeing that the sample energy function is tractable, the need for the use of the GAN framework is less clear. It would be possible to dispense of the critic and simply use the sample gradient of the energy function to update the parameters of the generator. Doing so would resemble a more traditional machine learning model where the objective is simply

$$\min_{G_\theta} \epsilon(G_\theta(z), P)$$

with P representing the underlying distribution of the dataset.

This possibility is not mentioned by Bellemare et al. and it would be useful to see how well it would work. Instead, they use as critic a function (implemented as a neural network) $h(x) : \mathbb{R}^d \to R^k$ (with $k < d$) which performs dimensionality reduction. The sample energy distance is computed using the values that h maps the samples to, with h being trained to maximize the energy distance.

Experimentally, the more they train h, the better the results, indicating that maintaining the GAN framework is beneficial. However, this observation is never justified mathematically.

5 Issues of the Original Value Function

While GANs can achieve impressive results, they are notorious for being incredibly hard to train (i.e. to get them to converge to some sensible generator). The two main issues plaguing GANs are as follows:

- Mode collapse, which corresponds to cases when the generator maps most random vectors $z \sim p_z$ to a single or a small amount of output points. Such a generator is only capable of generating a few distinct samples.
- Poor gradients impeding the training of the generator.

Arjovsky et al. argue that the choice of value function is to blame for these issues which forms the basis of their use of the KW distance.

[7] doesn't explicitly make the link between maximizing log-likelihood and the Kullback-Leibler divergence (defined as $KL(P||Q) := \int_{\mathbb{R}} \log \frac{P(x)}{Q(x)} P(x) d\mu(x))$. However, as minimizing the KL divergence is equivalent to maximizing the log-likelihood [5] (the value function used in [7]), both [1] and [2] compare their metrics to the KL divergence.

For the issue of mode collapse, Arjovsky et al. do not observe it occuring in any of their experiments. While they do not provide an explicit justification, the intuition is fairly clear. When using the log-likelihood as a value function, for a given discriminator, the generator is incentivized to always output the point where D is maximized and never the other points in order to minimize $\mathbb{E}_{z \sim p_z} \log(1 - D(G(z)))$.

However, when using the KW distance and considering the intuition of optimal transport, it is clear that a generator that minimizes the KW distance will not yield a distribution where all the probability mass is concentrated on a single point or a few points (unless that is truly the underlying distribution) since doing so does not minimize the optimal transport distance. Thus, it is conceivable that using a WGAN would generally solve the issue of mode collapse.

As for the issue of gradients, problems arise when the discriminator is too well trained. In such cases, the loss is close to 0 and has a very small gradient, hindering the training of the generator. On the other hand, with a WGAN, we have that training the critic only improves the estimate of the KW distance. Hence, "excessive" training simply improves the estimate of the gradient and the only reason the critic is not trained to optimality is due to computational constraints.

Essentially, when using log-likelihood, training is a balancing act. If the discriminator is not trained often enough, it remains relatively static in which case the generator is prone to mode collapse. If it is trained too often, it will become too accurate and no longer provide useful gradients.

Finally, in any case where $P(x) = 0$ while $Q(x) > 0$, the KL divergence is infinite on top of being asymmetric. On the other hand, in cases of non-overlapping supports, the KW metric is still defined and is symmetric as a metric. As such, the use of the KW remedies many of the issues associated with a maximum likelihood approach to loss.

6 Geometry of Outcomes

Embedded in the ability of the KW metric to solve these issues is the intuitive advantage of the KW metric (as well as the Cramér/energy distance) over the KL divergence. While the KL divergence is independent of the metric on the space, the KW distance takes into account the "geometry of outcomes".

For example, given 2 Dirac random variables δ_1 and δ_{1+x} defined on $\mathbb{R}$, we have that $KL(\delta_1, \delta_{1+x}) = \infty$ for any $x > 0$. A similar issue exists for the Jensen-Shannon divergence (another divergence mentionned in [1]). Using the same example, we have that:

$$JS(\delta_1, \delta_{1+x}) = \frac{1}{2}[\log(2) + \log(2)] = \log 2$$

for any 2 Dirac random variables with $x > 0$. However, intuitively, we would want δ_1 to be closer to δ_{1+x} for x small (which is the case for the KW-distance as it is proportional to the value of x).

Bellemare et al. formalize this notion with a property they coin as scale sensitivity (S) where a divergence d is considered scale sensitive if, $\exists \beta > 0$ such that for all random variables X, Y and all $c \in \mathbb{R}^+$:

$$d(cX, cY) \leq |c|^\beta d(X, Y)$$

While they cite [3] instead of proving the result ([3] only states the result, presumably since it follows directly from the linearity of integration), it can be seen that for $\mathbb{R}^d$ with the Euclidean metric and $\beta = 1$ the property holds:

$$W_p(cP, cQ) = \inf_{\pi \in \Pi(cP, cQ)} \left(\int_{X \times X} d(x, y)^p d\pi \right)^{\frac{1}{p}}$$

$$= \inf_{\pi \in \Pi(P, Q)} \left(\int_{X \times X} d(cx, cy)^p d\pi \right)^{\frac{1}{p}}$$

$$= \inf_{\pi \in \Pi(P, Q)} \left(\int_{X \times X} |c|^p d(x, y)^p d\pi \right)^{\frac{1}{p}}$$

$$= |c| \inf_{\pi \in \Pi(P, Q)} \left(\int_{X \times X} d(x, y)^p d\pi \right)^{\frac{1}{p}} = |c| W_p(P, Q)$$

They also describe the sum invariance property (I) which states that for A independent of X and Y:

$$d(A + X, A + Y) \leq d(X, Y)$$

Together, (S) and (I) yield what they call an "ideal divergence". Finally, the last property they mention is unbiased sample gradients (U) with the main contention being that the Cramér distance should be selected over the KW distance since it has all three properties while the latter lacks (U). Notice that the biased sample gradients are referring to an issue with the generator and not the potential bias of the critic brought up earlier.

Specifically, they prove the bias exists for 2 Bernouilli distributions (P the true distribution and Q_θ, a Bernouilli distribution with parameter θ trained using gradient descent) by finding a lower bound for the bias and showing that Q_θ doesn't converge to the distribution that minimizes the KW distance. It remains to see if the result holds generally for most distributions or if the bias is restricted to cases of simple distributions (and if there are different optimization techniques that avoid this issue).

Finally, the 3 properties are shown for the Cramér (l_2) and energy distances (interestingly they show that (U) holds **only** for $p = 2$ and not the other l_p metrics) and are used to justify the statement that the Cramér/energy distance is "strictly superior" to the KW distance "for machine learning applications" [2]. While the statement is true with respect to the 3 properties, the Cramér distance lacks other benefits of the KW distance (for example the intuitive interpretation) and as such it would be unfair to describe it as strictly better.

7 Conclusion

Ultimately, both the KW and Cramér distance provide improvements in terms of stability and convergence to GANs relative to the original maximum likelihood value function. In doing so, they change the paradigm to one where the second agent's goal is to give an estimate of the distance between the generator's distribution and the true distribution (that takes into account the underlying metric) while providing useful gradients for the training of the generator.

It would be relevant to see how training a generator using just the gradient of one of these loss functions (outside of the framework of a GAN) would fare. Additionally, further research is needed to give theoretical validity (not just experimental validity) to some of the estimation techniques used by both papers.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, 2017.

[2] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The cramér distance as a solution to biased wasserstein gradients, 2017.

[3] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

[4] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

[5] N. Defreitas. Machine learning, December 2019.

[6] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2016. cite arxiv:1701.00160Comment: v2-v4 are all typo fixes. No substantive changes relative to v1.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017.

[9] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861 – 867, 1993.

[10] P. Major. On the invariance principle for sums of independent identically distributed random variables. *Journal of Multivariate Analysis*, 8(4):487 – 517, 1978.

[11] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving gans using optimal transport, 2018.

[13] J. Schmidhuber. Learning factorial codes by predictability minimization. Technical Report CU-CS-565-91, Dept. of Comp. Sci., University of Colorado at Boulder, Dec. 1991.

[14] G. Szekely. E-statistics: The energy of statistical samples. 01 2003.

[15] A. Vetta. Lecture notes in algorithmic game theory, September 2019.

Appendix

7.1 Related Work

Included below are some other interesting papers related to [1, 2]:

Cuturi's work [4] on Sinkhorn distances yields a much a quicker computation of an approximation of the optimal transport distance. The associated algorithm could be used to quickly evaluate generative models at different stages (seeing that the KW distance is a good indication of visual quality sample) or potentially even as critic.

Mao et al.'s work [11] on the Least Squares GAN (LSGAN) demonstrates another attempt at incorporating a loss function that takes into account the underlying geometry (albeit in a simpler manner). The resulting GAN shows similar improvements as the WGAN (higher stability, better gradients and no mode collapse).

Finally, Salimans et al.'s work [12] is particularly relevant as it combines elements of both papers by using a new metric that corresponds to a combination of the KW metric (in primal form) with a learned energy distance with the goal of "improving GANs using optimal transport".