Notes on GAN objective functions

Daniel C. Elton*

(Dated: September 25, 2018)

In generative modeling, the task of the generator is to reproduce the training data distribution as closely as possible. Broadly speaking, there are three classes of criteria that are used when comparing the generator's distribution with the training data distribution - log-likelihood or a divergence metric, Parzen windows, and visual inspection (in the case of images, for instance).[1] Interestingly, these three methodologies measure different things, so good performance under one type of criteria does not imply good performance under another.[1] Maximizing log-likelihood is equivalent to minimizing Kullback-Leibler divergence, but often other divergence metrics are used, and finding the most useful (eg. computationally expedient) has been the subject of much research.

The original paper on generative adversarial networks (GANs)[2] introduced the following objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x} \in p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \in p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}))] \tag{1}$$

This form of the objective function has a nice theoretical interpretation as a two person minimax game. The solution to the minimax problem can be interpreted as a Nash equilibrium, a concept from game theory. However, this objective function is rarely used in practice. Firstly, as noted in the original paper, this objective function does not provide a very strong gradient signal when training starts because then $\log(1 - D(G(\boldsymbol{z}))$ saturates (goes to negative infinity) and the numerical derivative becomes impossible to calculate. In practice it is better to maximize $\log(D(G(\boldsymbol{z})))$ when training the generator. Before moving onto to discuss other objective functions, it is worth trying to understand what this objective function does. Equation 1 can simplified as follow:

$$\begin{aligned} V(D,G) &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) d\boldsymbol{x} + \int_z p_{\boldsymbol{z}}(\boldsymbol{z}) \log(1 - D(G(\boldsymbol{z})) d\boldsymbol{z} \\ &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_g(\boldsymbol{x}) \log(1 - D(\boldsymbol{x})) d\boldsymbol{x} \end{aligned} \tag{2}$$

* delton@umd.edu

For a fixed G, the optimal discriminator is :

$$D_G^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \tag{3}$$

If we assume $D = D_G^*$, then the objective function for the generatior can be expressed in terms of the **Jensen-Shannon divergence** $JS(p, q)$:[2]

$$\mathcal{C}(G) = -\log(4) + 2JS(p_{\text{data}}, p_{\theta_G}) \tag{4}$$

Jensen-Shannon divergence is defined as:

$$\delta_{\text{JS}}(p, q) = \frac{1}{2}\left[D_{\text{KL}}\left(p\middle|\middle|\frac{p+q}{2}\right) + D_{\text{KL}}\left(q\middle|\middle|\frac{p+q}{2}\right)\right] \tag{5}$$

where $D_{\text{KL}}(p||q)$ is the famous **Kullback-Leibler (KL) divergence**:

$$\delta_{\text{KL}}(p||q) = \int d\boldsymbol{x} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \tag{6}$$

This expression differs from relatively entropy, which appears widely in information theory and physics, by only a minus sign ($S(p|q) = -\delta_{\text{KL}}(p||q)$).[3]

A. Interpreting KL divergence

Why is KL divergence referenced so much in machine learning? One reason is that it can be shown that maximizing the log-likelihood of data under a model is the same as minimizing the KL divergence between the data distribution and the model distribution (ie. $\min D_{\text{KL}}(p_{\text{data}}||p_\theta)$).[4] To see this, observe that:

$$\begin{aligned} \delta_{\text{KL}}(p_{\text{data}}||p_t heta) &= \int p_{\text{data}}(\boldsymbol{x}) \log_{\text{data}}(\boldsymbol{x}) d\boldsymbol{x} - \int p_{\text{data}}(\boldsymbol{x}) \log_\theta(\boldsymbol{x}) d\boldsymbol{x} \\ &= -S(p_{\text{data}}) - \langle \log p_\theta \rangle_{\text{data}} \end{aligned} \tag{7}$$

rearranging the last line, we see:

$$\langle \log p_\theta \rangle_{\text{data}} = -S(p_{\text{data}}) - \delta_{\text{KL}}(p_{\text{data}}||p_t heta) \tag{8}$$

The equivalence follows from the positively of the KL-divergence and the fact that the entropy of the data $S(p_{\text{data}})$ is constant. A key difference between GAN methods and maximum likelihood methods stems from the fact that $D_{\text{KL}}(p_{\text{data}}||p_\theta)$ and $D_{\text{KL}}(p_{\text{data}}||p_\theta)$ measure different things. A minimization of $D_{\text{KL}}(p_{\text{data}}||p_\theta)$ (ie. max likelihood) prioritizes

models that have high probability where there is lots of data while ignoring the areas of significant probability where there is no data. Minimizing $D_{\mathrm{KL}}(p_{\mathrm{data}}||p_\theta)$, by contrast, focuses on preventing the model from predicting high probabilities where there is no data. In other words, maximum likelihood methods often improperly "fill in" low-probability regions between peaks in the data distribution.

Another appeal to studying KL divergence is that it has information theoretic interpretations. From information theory, the average number of bits needed to compress a data point $x \in p$ is $\log_2 p(x)$.[5] In order to communicate the density p to someone who already knows q we have to communicate information which has a code length that is the difference in code lengths $\lfloor \log_2 p(x) \rfloor - \lfloor \log_2 q(x) \rfloor$ for every data point. On expectation, this is $\mathbb{E}_p \lfloor \log_2 p(x) \rfloor - \lfloor \log_2 q(x) \rfloor \approx \delta_{\mathrm{KL}}(p, q)$. Thus the KL divergence has a clear information-theoretic meaning as the average number of bits required to communicate the density p to someone who already knows q. Further theoretical justification for KL divergence based on a set of three axioms (locality, coordinate invariance, and subsystem independence) has been developed.[3]

KL divergence is always greater than zero and equals zero if and only if $p(\boldsymbol{x}) = q(\boldsymbol{x})$. While KL divergence measures the similarity between two distributions, it violates both symmetry $\delta(p, q) = (q, p)$ and violates the triangle inequality, $\delta(a, b) + \delta(b, c) \geq \delta(a, c)$, so it is not a metric.[4] Jenson-Shannon divergence is a metric and therefore is sometimes called Jenson-Shannon distance.

Objective function 1 runs into issues in high dimensional spaces. Empirically, most high dimensional real world data lies close to a low dimensional manifold. Therefore, when training a GAN it becomes extremely unlikely that the initial generator distribution $G(\boldsymbol{x})$ overlaps with the target distribution – it would be like finding a needle in a haystack. If there is little or no intersection between distributions, KL divergence become infinite and the gradient signal becomes zero. JS divergence is better behaved in the sense that it doesn't become infinite when $p_\theta(\boldsymbol{x}) = 0$ but suffers from the same problem of having zero gradient when there is little or no overlap. This is a well known problem in the context of unsupervised learning of distributions, both for GANs and VAEs. One solution is to add additional noise to model, but in the context of images this makes the generated images blurry.

B. Other divergence measures

In its most general form, **maximum mean discrepancy** (MMD) is defined as:[6]

$$\delta_{\text{MMD}}(p,q) = \sup_{f\in\mathcal{H}} E[f(\boldsymbol{x})] - E[f(\boldsymbol{y})] \tag{9}$$

Here $\boldsymbol{x}$ is an independent sample from p and $\boldsymbol{y}$ is an independent sample from q. In other words, to calculate MMD, one searches the Hilbert space $\mathcal{H}$ for a function which maximizes the difference in expectation of that function over samples from p vs samples from q. In the case where $\mathcal{H}$ is a reproducing kernel Hilbert space, Gretton et al. showed that closed form solutions to the supremum problem can be found.[7] In particular, for a given choice of kernel function $k(\boldsymbol{x},\boldsymbol{y})$ they found that:

$$\text{MMD}(p,q) = (E_{p,q}[k(\boldsymbol{x},\boldsymbol{x}') - 2k(\boldsymbol{x},\boldsymbol{y}) + k(\boldsymbol{y},\boldsymbol{y}')])^{1/2} \tag{10}$$

The unbiased estimator of this is:

$$\text{MMD}(p,q) = \left(\frac{1}{N(N-1)}\sum_{n\neq n'} k(\boldsymbol{x}_n,\boldsymbol{x}_{n'}) - \frac{2}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M} k(\boldsymbol{x}_n,\boldsymbol{y}_m) + \frac{1}{M(M-1)}\sum_{m\neq m'} k(\boldsymbol{y}_m,\boldsymbol{y}_{m'})]\right)^{1/2} \tag{11}$$

where as before $\boldsymbol{x},\boldsymbol{x}'$ are independent samples from p and $\boldsymbol{y},\boldsymbol{y}'$ are independent samples from q. As should be obvious from inspection of the equation, the construction of the kernel function $k(\boldsymbol{x},\boldsymbol{y})$ is critical to this method. Typically a Gaussian kernel is used, with the width of the kernel becoming a hyperparameter. The MMD metric was applied to GANs by Dziugaite, et al.[6] They view their "MMD-Nets" as GANs where the discriminator is replaced with an MMD metric, which has a learned test function.

Arjovsky and Bottou (2017) introduced the **Wasserstein distance**, which is now one of the most popular metrics.[8] The Wasserstein distance (also called the "Earth mover's distance") can be informally understood by imagining the probability distribution to be a pile of dirt, and the distance to a different distribution to be the number of buckets of dirt that need to be moved, times the sum of the distances each bucket must be moved in order to transform the first pile to the second. Mathematically this is expressed as :

$$W(p,q) = \inf_{\gamma\in\Pi(p,q)} \mathbb{E}_{(x,y)\in\gamma}||x-y||] \tag{12}$$

$\Pi(x,y)$ can be understood to be the "transport plan" explaining how much probability mass we move from from x to y. To be a valid transport plan the constraints $\sum_x \Pi(x,y) = q(y)$

and $\sum_y \Pi(x,y) = q(x)$ must also hold. In other words, p and q must be marginals of $\Pi(x,y)$. The optimization problem to find the optimal transport plan can be solved with linear programming. Unfortunately linear programming is intractable for everything but small discrete-space problems, so some sophisticated math (the Kantorovich-Rubinstein duality and other tricks) is required to transform to a dual problem which is tractable.

Other metrics include ***f*-divergence**[9] and **total variation**:

$$\delta(p,q) = \sup_{\boldsymbol{x}} \in |p(\boldsymbol{x}) - q(\boldsymbol{x})| \tag{13}$$

Total variation is used in the energy based GAN (EBGAN) of Zhao, Mathieu, and LeCun.

---

[1] L. Theis, A. van den Oord, and M. Bethge, in *International Conference on Learning Representations* (2016), URL `http://arxiv.org/abs/1511.01844`.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014), pp. 2672–2680, URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[3] A. Caticha, in *AIP Conference Proceedings* (AIP, 2004), URL `https://doi.org/10.1063/1.1751358`.

[4] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, ArXiv e-prints (2018), 1803.08823.

[5] C. E. Shannon, Bell System Technical Journal **27**, 379 (1948).

[6] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence* (AUAI Press, Arlington, Virginia, United States, 2015), UAI'15, pp. 258–267, ISBN 978-0-9966431-0-8, URL `http://dl.acm.org/citation.cfm?id=3020847.3020875`.

[7] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, in *Advances in Neural Information Processing Systems 19*, edited by B. Schölkopf, J. C. Platt, and T. Hoffman (MIT Press, 2007), pp. 513–520, URL `http://papers.nips.cc/paper/3110-a-kernel-method-for-the-two-sample-problem.pdf`.

[8] M. Arjovsky, S. Chintala, and L. Bottou, ArXiv e-prints (2017), 1701.07875.

[9] S. Nowozin, B. Cseke, and R. Tomioka, ArXiv e-prints (2016), 1606.00709.