Linking Generative Adversarial Learning and Binary Classification

Akshay Balsubramani
abalsubr@stanford.edu

September 6, 2017

Abstract

In this note, we point out a basic link between generative adversarial (GA) training and binary classification – any powerful discriminator essentially computes an (f-)divergence between real and generated samples. The result, repeatedly re-derived in decision theory, has implications for GA Networks (GANs), providing an alternative perspective on training f-GANs by designing the discriminator loss function.

1 Generating a Distribution

Imagine we are given real data with distribution $P_r(x)$ over a feature space $\mathcal{X}$, and wish to learn a distribution $P_g(x)$ that is as "close" as possible to P_r. Closeness will be measured by some divergence function $D(\cdot, \cdot)$ between probability distributions. So the generator is typically solving

$$P_g^* = \arg\inf_{P_g} D(P_g, P_r) \tag{1}$$

where P_g is in some class of distributions specified by the generator. This manuscript considers $D(\cdot, \cdot)$ to be an f-divergence:

Definition 1. ***f-divergence.*** *For any convex $f(t)$ with $f(1) = 0$, define the f-divergence of P_g to P_r* [1] *as*

$$D_f(P_g, P_r) = \mathbb{E}_{x \sim P_r}\left[f\left(\frac{P_g(x)}{P_r(x)}\right)\right]$$

Here we examine the method of highlighting differences between P_r and P_g by feeding them to a binary classifier (discriminator) with corresponding labels $y = \pm 1$, in equal proportion as for a typical GAN setup, so that the data input to the discriminator are assigned a positive label if they are real data:

$$p := \Pr(y = +1) = \frac{1}{2} \quad , \quad P_r(x) = \Pr(x \mid y = +1) \quad , \quad P_g(x) = \Pr(x \mid y = -1)$$

Recall that any two-class loss function can equivalently be written in terms of *partial losses* $\ell_+(g)$ and $\ell_-(g)$; these are the losses with respect to true labels ± 1 respectively, as a function of the label prediction g.

The discrimination problem is to find a function h in some model class $\mathcal{H}$ that attempts to minimize some loss on average over the data:

$$\inf_{h \in \mathcal{H}} \mathbb{E}_{(x,y)}\left[\ell(y, h(x))\right] \tag{2}$$

The *generative* view of binary classification [RW11] writes this in terms of the class-conditional distributions $\Pr(x \mid y = \pm 1)$:

$$\begin{aligned}
\mathbb{E}_y\left[\ell(y, h(x))\right] &= \Pr(x, y = +1)\,\ell_+(h(x)) + \Pr(x, y = -1)\,\ell_-(h(x)) \\
&= p \Pr(x \mid y = +1)\,\ell_+(h(x)) + (1 - p)\Pr(x \mid y = -1)\,\ell_-(h(x)) \\
&= \frac{1}{2}\left[P_r(x)\ell_+(h(x)) + P_g(x)\ell_-(h(x))\right]
\end{aligned} \tag{3}$$

[1] Note that $D_f(P_g, P_r)$ need not be positive.

The optimization problem (2) is standard in binary classification. Typically, $\mathcal{H}$ is chosen to be a fairly rich class of deep binary classifiers. This means that its performance is close to the Bayes risk, i.e. the minimum risk over measurable functions $\inf_h \mathbb{E}_{(x,y)}[\ell(y, h(x))]$ [RW11]. So the excess risk

$$\epsilon(\mathcal{H}) := \inf_{h\in\mathcal{H}} \mathbb{E}_{(x,y)}[\ell(y, h(x))] - \inf_h \mathbb{E}_{(x,y)}[\ell(y, h(x))]$$

is small.

2 Main Result

Theorem 2. *Take any loss function $\ell_\pm$ and any model class $\mathcal{H}$. Define $f(s) := \sup_\alpha(-\ell_+(\alpha) - s\ell_-(\alpha))$. This is a maximum of linear functions, so it is convex. Then*

$$\inf_{h\in\mathcal{H}} \mathbb{E}_{(x,y)}[\ell(y, h(x))] = -\frac{1}{2}D_f(P_g, P_r) + \epsilon(\mathcal{H})$$

Changing the model class $\mathcal{H}$ only changes the second term of Thm. 2. Therefore, when $\mathcal{H}$ is rich enough that the excess risk $\epsilon(\mathcal{H})$ is small, the loss function ℓ of the discrimination problem corresponds almost exactly to an f-divergence.

2.1 GA Training Solves the Generation Problem with f-Divergences

Revisiting (1), to find P_g^* to be "close" to P_r under some f-divergence D_f, one could solve

$$\begin{aligned} P_g^* &= \arg\inf_{P_g} D_f(P_g, P_r) = \arg\sup_{P_g}[-D_f(P_g, P_r)] = \arg\sup_{\Pr(x|y=-1)}\left[\inf_{h\in\mathcal{H}} \mathbb{E}_{(x,y)}[\ell(y, h(x))] - \epsilon(\mathcal{H})\right] \\ &\approx \arg\sup_{\Pr(x|y=-1)}\left[\inf_{h\in\mathcal{H}} \mathbb{E}_{(x,y)}[\ell(y, h(x))]\right] \end{aligned}$$

So the adversarial game interaction between the generator and discriminator emerges as the solution to the generation problem for powerful enough discriminators, for any $\ell, \mathcal{H}$.

2.2 Examples

Table 1 shows the correspondence between ℓ and f for several common f-divergences. Similar lists can be found in [NWJ09, RW11].

In the GA setup, the variable s is always a function over the data space $\mathcal{X}$. The maximizing α in $\arg\sup_\alpha(-\ell_+(\alpha) - s\ell_-(\alpha))$ is a function of s; as a function of the data $\alpha(x)$, it is the optimal discriminator $h^*(x) = h^*(s(x))$.

3 Related Work

The most related work to this manuscript is the f-GAN approach of [NCT16], to our knowledge. This solves the same problem of minimizing the f-divergence to the true distribution, but by changing the discriminator objective from the binary classification risk, (in contrast to Thm. 2 which just interprets the risk). The key fact is that a convex function f has a well-defined *convex conjugate* function f^* such that $f(u) = \sup_{t\in\mathbb{R}}[tu - f^*(t)]$, so that the following is true[2]:

$$\begin{aligned} D_f(P_r, P_g) &= \mathbb{E}_{x\sim P_g}\left[\sup_t\left(t\frac{P_r(x)}{P_g(x)} - f^*(t)\right)\right] = \sup_t\left(\mathbb{E}_{x\sim P_g}\left[t\frac{P_r(x)}{P_g(x)} - f^*(t)\right]\right) \\ &= \sup_h\left[\mathbb{E}_{x\sim P_r}[h(x)] - \mathbb{E}_{x\sim P_g}[f^*(h(x))]\right] \geq \sup_{h\in\mathcal{H}}\left[\mathbb{E}_{x\sim P_r}[h(x)] - \mathbb{E}_{x\sim P_g}[f^*(h(x))]\right] \end{aligned} \tag{4}$$

[2]Ignoring conjugacy domain issues for simplicity.

Loss ℓ	Partial losses	$f(s)$	$h^*(s)$	f-divergence
0-1	$\ell_\pm(g) = \frac{1}{2}(1 \mp g)$	$\frac{1}{2}\lvert s-1\rvert$	$\operatorname{sgn}(s-1)$	Total variation dist.
Log	$\ell_\pm(g) = \ln\left(\frac{2}{1\pm g}\right)$	$-\ln(1+s) - s\ln\left(\frac{1+s}{s}\right)$	$\frac{1-s}{1+s}$	Jensen-Shannon dist.
Square	$\ell_\pm(g) = (1 \mp g)^2$	$-\frac{s}{1+s} + \frac{1}{2}$	$\frac{1-s}{1+s}$	Triangular discrimination dist.
CW (param. c)	$\ell_\pm(g) = \left(\frac{1}{2} - \left(\frac{1}{2} - c\right)\right)(1 \mp g)$	$\lvert 1-c-cs\rvert - cs+c-\lvert 1-2c\rvert$	$\operatorname{sgn}(1-c-cs)$	–
Exponential	$\ell_\pm(g) = \exp(\mp g)$	$-2\sqrt{s}+2$	$-\frac{1}{2}\ln s$	Hellinger dist.
"Boosting"	$\ell_\pm(g) = \sqrt{\frac{1\mp g}{1\pm g}}$	$-2\sqrt{s}+2$	$\frac{1-s}{1+s}$	Hellinger dist.

Table 1: Some discriminator losses, with corresponding f-divergences.

[NCT16] use this bound from [NWJ10]. It is quite tight when $\mathcal{H}$ is rich, exactly when Thm. 2 is strong, though the order of the arguments is switched. [3]

More broadly, since the original GAN paper [GPAM$^+$14], the GA approach has enjoyed a string of recent empirical successes with very rich model classes $\mathcal{H}$ [RMC15, BSM17], in accordance with Thm. 2.

4 Summary

The correspondences here fundamentally link generative adversarial training and the generation problem, and most are well known in decision theory. However, within the GAN literature they do not appear well known and lack references, which we address in this note.

References

[BSM17] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[GPAM$^+$14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[LV06] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[NCT16] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[NWJ09] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. On surrogate loss functions and f-divergences. *The Annals of Statistics*, pages 876–904, 2009.

[3] Note that (4) puts the modeled (generated) distribution in the second argument rather than the first; the two orderings are related by Csiszár duality [RW11]. Our proof of Theorem 2 can be followed to prove an exact analogue of Theorem 2 viewing the discriminator risk as $D_f(P_r, P_g)$ with the arguments interchanged as in (4). The analogue result only differs in the definition of the convex function generating the divergence, which is $\sup_\alpha(-\ell_-(\alpha) - s\ell_+(\alpha))$ instead of f as defined in Thm. 2. In this manuscript, we follow the convention of taking the real distribution to be the second argument with respect to which we measure the "excess description length" of using P_g.

[NWJ10] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[RW11] Mark D Reid and Robert C Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(Mar):731–817, 2011.

5 Proofs

This section relates binary classification loss functions to f-divergences, recapitulating [LV06, NWJ09].

Proof of Theorem 2. From (3), if μ is the base measure over $\mathcal{X}$,

$$\begin{aligned}
&\inf_h \mathbb{E}_{(x,y)}\left[\ell(y,h(x))\right] = \inf_h \mathbb{E}_x\left[\mathbb{E}_y\left[\ell(y,h(x))\right]\right] = \frac{1}{2}\inf_h \mathbb{E}_{x\sim\mu}\left[P_r(x)\ell_+(h(x)) + P_g(x)\ell_-(h(x))\right] \\
&\quad = \frac{1}{2}\inf_h \mathbb{E}_{x\sim\mu}\left[P_r(x)\left(\ell_+(h(x)) + \frac{P_g(x)}{P_r(x)}\ell_-(h(x))\right)\right] = \frac{1}{2}\inf_h \mathbb{E}_{x\sim P_r}\left[\ell_+(h(x)) + \frac{P_g(x)}{P_r(x)}\ell_-(h(x))\right] \\
&\quad = \frac{1}{2}\mathbb{E}_{x\sim P_r}\left[\inf_\alpha\left(\ell_+(\alpha) + \frac{P_g(x)}{P_r(x)}\ell_-(\alpha)\right)\right] = -\frac{1}{2}\mathbb{E}_{x\sim P_r}\left[\sup_\alpha\left(-\ell_+(\alpha) - \frac{P_g(x)}{P_r(x)}\ell_-(\alpha)\right)\right] \\
&\quad = -\frac{1}{2}D_f(P_g,P_r)
\end{aligned}$$

Adding $\epsilon(\mathcal{H})$ to both sides proves the result. □