Improved generator objectives for GANs

Ben Poole[*]
Stanford University
poole@cs.stanford.edu

Alexander A. Alemi, Jascha Sohl-Dickstein, Anelia Angelova
Google Brain
{alemi, jaschasd, anelia}@google.com

Abstract

We present a framework to understand GAN training as alternating density ratio estimation, and approximate divergence minimization. This provides an interpretation for the mismatched GAN generator and discriminator objectives often used in practice, and explains the problem of poor sample diversity. Further, we derive a family of generator objectives that target arbitrary f-divergences without minimizing a lower bound, and use them to train generative image models that target either improved sample quality or greater sample diversity.

1 Introduction

Generative adversarial networks (GANs) have become a popular method for fitting latent-variable directed generative models to complex datasets [2, 10, 13, 6]. While these models provide compelling visual samples, they are notoriously unstable and difficult to train and evaluate. Many recent papers have focused on new architectures and regularization techniques for improved stability and performance [13, 11, 4], but the objectives they optimize are fundamentally the same as the objectives in the original proposal [2].

The visual quality of samples from generative models trained with GANs often exceeds those of their variationally-trained counterparts [5, 12]. This is often credited to a difference in the divergence between the data and model distribution that each technique optimizes [15]. GAN theory shows that an idealized formulation optimizes Jensen-Shannon divergence, while VAEs optimize a lower bound on log-likelihood, corresponding to a lower bound on the KL divergence. Recent work has generalized the GAN theory to target reverse KL [14] and additional f-divergences (including KL, reverse KL, and JS), allowing GANs to target a diverse set of behaviors [9].

However, these new theoretical advances fail to provide a justification for the GAN objectives that are used in practice. In particular, the generator objective used in practice is different from the one that is theoretically justified [2, 9]. This raises the question as to whether the theory used to motivate GANs applies to these modified objectives, and how the use of mismatched generator and discriminator objectives influences the behavior of GANs in practice.

Here we present a new interpretation of GANs as alternating between steps of density ratio estimation, and divergence minimization. This leads to a new understanding of the GAN generator objective that is used in practice as targeting a mode-seeking divergence that resembles reverse KL, thus providing an explanation for the mode dropping seen in practice. Furthermore, we introduce a set of new objectives for training the generator of a GAN that can trade off between sample quality and sample diversity, and show their effectiveness on CIFAR-10.

---

[*]Work done during an internship at Google Brain.

2 Theory

2.1 Background

Given samples from a data density, $x \sim q(x)$, we would like to learn a generative model with density p that matches the data density q. Often the models we are interested in have intractable likelihoods, so that we can sample x efficiently but cannot evaluate its likelihood. In the GAN framework [2], the intractable likelihood is bypassed by instead training a discriminator to classify between samples from the data and samples from the model. Given this discriminator, the parameters of the generative model are updated to increase the tendency of the discriminator to mis-classify samples from the model as samples from the data. This iterative process pushes the model density towards the data density without ever explicitly computing the likelihood of a sample. More formally, the GAN training process is typically motivated as solving a minimax optimization problem:

$$\underset{p}{\text{minimize}} \ \underset{d}{\max} \ (\mathbb{E}_{x \sim q} [\log d(x)] + \mathbb{E}_{x \sim p} [\log (1 - d(x))]) \tag{1}$$

where p is the generative model distribution, d is the discriminator, and q is the data distribution. Fixing p, the optimal discriminator is $d^*(x) = \frac{q(x)}{q(x)+p(x)}$ [2]. Thus if the inner maximization over the discriminator is performed to completion for each step of p, the GAN objective is equivalent to minimizing:

$$\underset{p}{\text{minimize}} \ \left(\mathbb{E}_{x \sim q} \left[\log \frac{q(x)}{q(x) + p(x)} \right] + \mathbb{E}_{x \sim p} \left[\log \left(1 - \frac{q(x)}{q(x) + p(x)} \right) \right] \right) = 2 \, \text{JS}(q \| p) - \log 4 \tag{2}$$

This has led to the understanding that GANs minimize the Jensen-Shannon divergence between the data density and the model density, and is thought to underlie the difference in sample quality between GANs and VAEs [15]. However, this is not the objective that is used in practice, and we will see below that this alters the analysis.

Recently, [9] proposed an extension to GANs to target divergences other than Jensen-Shannon. They generalize the set of divergences a GAN can target to the family of f-divergences, where:

$$D_f (q \| p) = \int dx \, p(x) f \left(\frac{q(x)}{p(x)} \right) \tag{3}$$

and $f(u) : \mathbb{R}^+ \to \mathbb{R}$ is a convex function with $f(1) = 0$. The key result they leverage from [8] is that any f-divergence can be lower-bounded by

$$D_f(q \| p) \geq \underset{T \in \mathcal{T}}{\sup} \ (\mathbb{E}_{x \sim q} [T(x)] - \mathbb{E}_{x \sim p} [f^\star(T(x))]) \tag{4}$$

where $f^\star$ is the Fenchel conjugate[2] of f, and T is the variational function also known as the discriminator in the GAN literature[3]. Thus for any T, we have a lower bound on the divergence that recovers exactly the discriminator objective used in the standard GAN when $f(u) = u \log u - (u + 1) \log(u + 1)$. As this is a lower bound on the f-divergence, maximizing it with respect to the discriminator T makes sense, and yields a tighter lower bound on the true divergence.

However, the objective to optimize for the *generative model*, p, remains unclear. In both the original GAN paper [2] and the f-GAN paper [9], two objectives are proposed (denoted as $\mathcal{G}_{\text{LB}}$ and $\mathcal{G}_{\text{ALT}}$):

1. $\mathcal{G}_{\text{LB}}$: Minimize the *lower* bound in Equation 4. For standard GANs, this corresponds to minimizing the probability of the discriminator classifying a sample from the model as fake.

2. $\mathcal{G}_{\text{ALT}}$: Optimize an alternative objective:

$$\underset{p}{\text{minimize}} \ \mathbb{E}_{x \sim p} [-T(x)] \tag{5}$$

   For standard GANs, this corresponds to maximizing the log probability of the discriminator classifying a sample from the model as real.

---

[2]The Fenchel conjugate is defined as $f^\star(t) = \sup_{u \in \text{dom}_f} (ut - f(u))$

[3]We use q as the data distribution and p as the model distribution, which is the opposite of [9].

The first approach minimizes a lower bound, and thus improvements in the objective can correspond to making $D_f(q\|p)$ smaller, or, more problematically, by making the lower bound on $D_f(q\|p)$ looser. In practice this leads to slower convergence, and thus the first objective is not widely used.

The second approach is empirically motivated in [2, 9] as speeding up training, and theoretically motivated by the observation that $p = q$ remains a fixed point of the learning dynamics. However, the behavior of this generator objective when the generative model does not have the capacity to realize the data density remains unclear. This is the regime we care about as most generative models do not have the capacity to exactly model the data.

2.2 Discriminator as a density ratio estimator

To address the theoretical and practical issues we first present a simple relationship between the discriminator and an estimate of the density ratio. Given known data and model densities, the optimal discriminator with respect to an f-divergence, f_D, was derived in [9] as:

$$T^*(x) = f'_D \left(\frac{q(x)}{p(x)} \right) \tag{6}$$

where f'_D is the derivative of f_D. If f'_D is invertible, we can reverse the relationship, and use the discriminator to recover the ratio of the data density to the model density:

$$\frac{q(x)}{p(x)} = (f'_D)^{-1} (T^*(x)) \approx (f'_D)^{-1} (T(x)) \tag{7}$$

In practice we don't have access to the optimal discriminator $T^*(x)$, and instead use the current discriminator $T(x)$ as an approximation.

2.3 A new set of generator objectives

Given access to an approximate density ratio $q(x)/p(x)$, we can now optimize any objective that depends only on samples from q or p and the value of the density ratio. Conveniently, f-divergences are a family of divergences that depend only on samples from one distribution and the density ratio! Given samples from p and an estimate of the density ratio at each point, we can compute an estimate of the f-divergence, f_G between p and q:

$$D_{f_G} (p\|q) = \mathbb{E}_{x \sim p} \left[f_G \left(\frac{q(x)}{p(x)} \right) \right] \approx \mathbb{E}_{x \sim p} \left[f_G \left((f'_D)^{-1} (T(x)) \right) \right] \equiv \mathcal{G}_{f_D, f_G} \tag{8}$$

where $\mathcal{G}_{f_D, f_G}$ is the generator objective, f_G is the f-divergence targeted for the generator, and f_D the f-divergence targeted for the discriminator. f_G and f_D need not be the same f-divergence. For non-optimal discriminators, this objective will be a biased approximation of the f-divergence, but is not guaranteed to be either an upper or lower bound on f_G.

Our new algorithm for GAN training iterates the following steps:

1. Optimize the discriminator, T, to maximize a lower-bound on $D_{f_D} (q\|p)$ using Equation 4.

2. Optimize the generator, p, to minimize $\mathcal{G}_{f_D, f_G}$, using the estimate of the density ratio from the current discriminator, T, in Equation 8.

While the first step is identical to the standard f-GAN training algorithm, the second step comprises a new generator update that can be used to fit a generative model to the data while targeting any f-divergence. In practice, we alternate single steps of optimization on each minibatch of data.

2.4 Related work

Several recent papers have identified novel objectives for GAN generators. In [14], they propose a generator objective corresponding to f_G being reverse KL, and show that it improves performance on image super-resolution. [3] identifies the generator objective that corresponds to minimizing the KL divergence, but does not empirically evaluate this objective.

Concurrent with our work, two papers propose closely related GAN training algorithms. In [16], they directly estimate the density ratio by optimizing a different discriminator objective that corresponds to rewriting the discriminator in terms of the density ratio:

$$D_f(q\|p) \geq \sup_r \left(\mathbb{E}_{x \sim q} \left[f'(r(x)) \right] - \mathbb{E}_{x \sim p} \left[f' \left(f^\star(T(x)) \right) \right] \right) \tag{9}$$

This approach requires learning a network that directly outputs the density ratio, which can be very small or very large and in practice the networks that parameterize the density ratio must be clipped [16]. We found estimating a function of the density ratio to be more stable, in particular using the GAN discriminator objective the discriminator $T(x)$ estimates $\log \frac{q(x)}{q(x)+p(x)}$. However, there are likely ways of combining these approaches in the future to directly estimate stable functions of the density ratio independent of the discriminator divergence.

More generically, the training process can be thought of as two interacting systems: one that identifies a statistic of the model and data, and another that uses that statistic to make the model closer to the data. [7] discusses many approaches similar to the one presented here, but do not present experimental results.

3 Interpreting the GAN generator objective used in practice, $\mathcal{G}_{\textbf{ALT}}$

We can use our new family of generator objectives to better understand $\mathcal{G}_{\text{ALT}}$, the objective that is used in practice (Eq. 5). Given that f_D is the standard GAN divergence, we can solve for the generator divergence, f_G, such that $\mathcal{G}_{\text{ALT}} = \mathcal{G}_{f_D, f_G}$, yielding:

$$f_G(u) = \log \left(1 + \frac{1}{u} \right) \tag{10}$$

Thus minimizing $\mathcal{G}_{\text{ALT}}$ corresponds to minimizing an approximation of the f_G divergence between the data density and the model density, not minimizing the Jensen-Shannon divergence.

To better understand the behavior of this divergence, we fit a single Gaussian to a mixture of two Gaussians in one dimension (Figure 1). We find that the GAN divergence optimized in practice is even more mode-seeking than JS and reverse KL. This behavior is likely the cause of many problems experienced with GANs in practice: samples often fail to cover the diversity of the dataset.
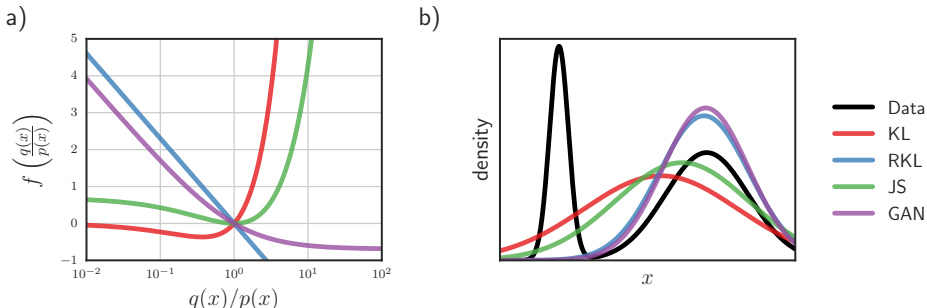

Figure 1: The GAN generator objective used in practice ($\mathcal{G}_{\text{ALT}}$) is mode-seeking when fit to a mixture of two Gaussians in one dimension. **(a)** Value of the divergence function, f, as a function of the density ratio. The behavior of the GAN objective used in practice ($\mathcal{G}_{\text{ALT}}$) resembles reverse KL when the model density is greater than the data density. **(b)** Learned densities when fitting a single Gaussian generative model to a mixture of two Gaussians (data, black). KL and JS are more mode-covering learning a generative model with larger variance that covers both modes of the data density, while reverse KL (RKL) and the GAN generator used in practice (GAN) are more mode-seeking, with smaller variance that covers only the higher density mode.

4 Experiments

We evaluate our proposed generator objectives at improving the sample quality and diversity on CIFAR-10. All models were trained using identical architectures and hyperparameters (see Appendix

(a) $\alpha = -3$

(b) $\alpha = -1$

(c) $\mathcal{G}_{\text{ALT}}$ (typical objective)

(d) Reverse KL ($\alpha \to 0$)

(e) squared Hellinger ($\alpha = 0.5$)

(f) KL ($\alpha \to 1$)

Figure 2: Different generator objectives yield different degrees of sample diversity. As we move from mode seeking α-divergences with low α to mode covering divergences with $\alpha > 0$ we see visual evidence of the increase in sample diversity, without a noticeable decrease in sample quality. In particular, note the overabundance of green and brown tones in the most mode seeking objectives. Sub-captions give the targeted generator divergence and are ordered from the most mode seeking to most mode covering. In all cases, the discriminator was trained using the standard GAN objective.

B). The discriminator in all models was trained to optimize the normal GAN objective, corresponding to maximizing Equation 4 with $f_D(u) = u \log u - (u+1) \log(u+1)$, and using $T(x) = g_f(V(x))$ with $V(x) \in \mathbb{R}$ being the output of a neural network and $g_f(v) = -\log(1 + \exp(v))$ being used to constrain the range of T as in [9]. For each model, we optimized a different generator objective by using different values for f_G in Equation 8. The generator objectives are derived and listed in Appendix A.

In order to highlight the effect the generator objective can have on the generated samples, we targeted several objectives at various α divergences, as well as the traditional generator objective $\mathcal{G}_{\text{ALT}}$. In Figure 2, we see that the generator objective has a large impact on sample diversity. In particular, for very mode-seeking divergences ($\alpha = -3$ and $\alpha = -1$), the samples fail to capture the diversity of class labels in the dataset, as is immediately visually obvious from over-representation of greens and browns in the generated samples. For more mode-covering divergences ($\alpha = 0.5$ (squared Hellinger), KL) we see much better diversity in colors and sampled classes, without any noticeably degradation in sample quality.

5   Discussion

Our work presents a new interpretation of GAN training, and a new set of generator objectives for GANs that can be used to target any f-divergence. We demonstrate that targeting JS for the discriminator and targeting other objectives for the generator yields qualitatively different samples, with mode-seeking objectives producing less diverse samples, and mode-covering objectives producing more diverse samples. However, training with very mode-seeking objectives does not yield extremely high-quality samples. Similarly, targeting mode-covering objectives like KL improves sample diversity, but the quality of samples does not visibly worsen. Visual evaluation of sample quality is a potentially fraught measure of quality however. Future work will be needed to investigate the impact of alternate generator objectives and provide better quantitative metrics and understanding of what factors drive sample quality and diversity in GANs.

References

[1] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua 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.

[3] Ian J Goodfellow. On distinguishability criteria for estimating generative models. *arXiv preprint arXiv:1412.6515*, 2014.

[4] Ferenc Huszar. An alternative update rule for generative adversarial networks. http://www.inference.vc/an-alternative-update-rule-for-generative-adversarial-networks/, 2015.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[6] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2016.

[7] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[8] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NIPS*, pages 1089–1096, 2007.

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

[10] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

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

[12] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*. Citeseer, 2014.

[13] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.

[14] Casper Kaae Sonderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszar. Amortised map inference for image super-resolution, 2016.

[15] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, Apr 2016.

[16] Masatoshi Uehara, Issei Sato, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.

A Deriving the generator objectives

Here we derive the generator objectives when the discriminator divergence is $f_D(u) = u \log u - (u + 1) \log(u+1)$, corresponding to the standard GAN discriminator objective. As in [9], we parameterize the discriminator as $T(x) = g_f(V(x))$ where g_f has the same range as f'_D. For the GAN case, this corresponds to $g_f(v) = -\log(1 + \exp(-v))$.

First, we can compute the inverse of the gradient of f_D which is used to estimate the density ratio:

$$(f'_D)^{-1}(t) = -\frac{e^t}{e^t - 1}$$

For GANs, the discriminator is parameterized as $T(x) = -\log(1 + \exp(-V(x))$, so we can compute the density ratio as:

$$\frac{q(x)}{p(x)} \approx -\frac{e^{T(x)}}{e^{T(x)} - 1} = e^{V(x)} \tag{11}$$

Given this estimate of the density ratio, we can then compute the generator objective as $f_G(e^{V(x)})$. The table below contains the generator objectives for many different f_G given $f_D(u) = u \log u - (u+1) \log(u+1)$:

Name	Generator f-divergence (f_G)	Generator objective (minimized)
GAN-standard	$\log(1 + \frac{1}{u})$	$\log\left(1 + e^{-V(x)}\right) = -T(x)$
GAN-RKL	$-\log u$	$-V(x)$
GAN-KL	$u \log u$	$V(x)e^{V(x)}$
GAN-α	$\frac{1}{\alpha(\alpha-1)}\left(u^\alpha - 1 - \alpha(u-1)\right)$	$\frac{1}{\alpha(\alpha-1)}\left(e^{\alpha V(x)} - 1 - \alpha(e^{V(x)} - 1)\right)$

$$\tag{12}$$

B CIFAR-10 architecture details

This is a slightly modified version of the architecture from [1]. Input images were scaled from $[0, 255]$ to $[0, 1]$.

Operation	Kernel	Strides	Feature maps	BN?	Dropout	Nonlinearity
$G_x(z) - 64 \times 1 \times 1$ input						
Transposed convolution	4×4	1×1	256	$\checkmark$	0.0	Leaky ReLU
Transposed convolution	4×4	2×2	128	$\checkmark$	0.0	Leaky ReLU
Transposed convolution	4×4	1×1	64	$\checkmark$	0.0	Leaky ReLU
Transposed convolution	4×4	2×2	32	$\checkmark$	0.0	Leaky ReLU
Transposed convolution	5×5	1×1	32	$\checkmark$	0.0	Leaky ReLU
Convolution	1×1	1×1	32	$\checkmark$	0.0	Leaky ReLU
Convolution	1×1	1×1	3	$\times$	0.0	Sigmoid
$V(x) - 3 \times 32 \times 32$ input						
Convolution	5×5	1×1	32	$\times$	0.2	Maxout
Convolution	4×4	2×2	64	$\times$	0.5	Maxout
Convolution	4×4	1×1	128	$\times$	0.5	Maxout
Convolution	4×4	2×2	256	$\times$	0.5	Maxout
Convolution	4×4	1×1	512	$\times$	0.5	Maxout
Convolution	1×1	1×1	1024	$\times$	0.5	Maxout
Convolution	1×1	1×1	128	$\times$	0.5	Maxout
Convolution	1×1	1×1	1	$\times$	0.5	Linear

Optimizer	Adam ($\alpha = 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$)
Batch size	128
Leaky ReLU slope, maxout pieces	0.1, 2
Weight, bias initialization	Isotropic gaussian ($\mu = 0$, $\sigma = 0.01$), Constant(0)

Table 1: CIFAR10 model hyperparameters. Maxout layers are used in the discriminator.