AC-GAN Learns a Biased Distribution

Rui Shu
Stanford University

Hung Bui
Adobe Research

Stefano Ermon
Stanford University

Abstract

The Auxiliary Classifier GAN (AC-GAN) was proposed in [1] and was able to yield high-quality images and state-of-the-art Inception Score. However, it is not immediately apparent why exactly AC-GAN improves upon GAN with respect to visual quality and the Inception Score. In this paper, we show that AC-GAN is a Lagrangian to a constrained primal objective function that explicitly pushes the density of the generator distribution away from the classifier's decision boundary. We verify empirically on MNIST-based experiments that AC-GAN indeed learns a biased distribution that down-samples points near the decision boundary. Our analysis suggests that AC-GAN's bias is a contributing factor for AC-GAN's performance on the visual quality and Inception Score metrics.

1 Introduction

The development of generative adversarial networks (GAN) has enabled rapid advancements in the learning of natural image distributions. However, GANs often have difficulty generating globally coherent, high-resolution samples. To address this issue, the GAN literature has proposed a range of solutions, including optimizing the architecture [2], improving the objective function [3], and improving the optimization procedure [4].

In this paper, we focus on [1]'s Auxiliary Classifier GAN (AC-GAN), a variant of InfoGAN [5] that incorporates a supervised learning signal. Let $p^*(x, y)$ be the true joint distribution and $p_\theta(x, y)$ be our generator, where x is an image and y is the class label. We further denote $d(p^*(x), p_\theta(x))$ as the Jensen-Shannon divergence and $H_\theta(Y|X)$ as the conditional entropy under the distribution $p_\theta(x, y)$. Given an auxiliary classifier $q_\phi(y \mid x)$, AC-GAN minimizes the following objective

$$\min_{\theta,\phi}. d(p^*(x), p_\theta(x)) + \lambda_m \mathcal{L}_m(\theta, \phi) + \lambda_c \mathcal{L}_c(\phi), \tag{1}$$

where λ_m, λ_c are weighting factors and the last two terms are

$$\mathcal{L}_m(\theta, \phi) = -\mathbb{E}_{(x,y)\sim p_\theta} \ln q_\phi(y \mid x) \tag{2}$$

$$\mathcal{L}_c(\phi) = \mathbb{E}_{x\sim p^*} \mathrm{D}_{\mathrm{KL}}(p^*(y \mid x) \| q_\phi(y \mid x)), \tag{3}$$

where D_{KL} is the Kullback-Leibler divergence. In this paper, we make the following contributions.

1. We show that the AC-GAN objective is a Lagrangian to a constrained optimization problem that rejects the sampling of points near the classifier decision boundary.
2. We verify empirically that if the true joint distribution $p^*(x, y)$ contains points near the decision boundary, AC-GAN will learn a biased distribution that down-samples those points.
3. We relate our analysis to why AC-GAN yields more visually appealing images and better Inception Score.

2 AC-GAN: A Lagrangian Perspective

In this section, we shall demonstrate that the AC-GAN objective can be interpreted as a Lagrangian to a constrained optimization problem that imposes restrictions on the density of $p_\theta(x)$. To simplify

Second workshop on Bayesian Deep Learning (NIPS 2017), Long Beach, CA, USA.

our analysis, we shall assume that the support of p_θ (which we shall denote $\mathcal{X}_\theta$) is contained within the support of p^* (denoted as $\mathcal{X}^*$). Formally, we assume that for all $\theta \in \Theta$, $\mathcal{X}_\theta \subseteq \mathcal{X}^*$. Additionally, we set $p_\theta(y)$ to be a fixed prior distribution (i.e. uniform distribution). We now begin by considering the following constrained optimization problem, which we denote as the primal objective,

$$\begin{aligned} \min_{\theta,\phi}.\ & d(p^*(x), p_\theta(x)) \\ \text{s.t. } & H_\theta(Y|X) \le \epsilon \\ & \mathbb{E}_{x\sim p_\theta} \mathrm{D}_{\mathrm{KL}}(p_\theta(y \mid x) \| q_\phi(y \mid x)) = 0 \\ & \mathbb{E}_{x\sim p^*} \mathrm{D}_{\mathrm{KL}}(p^*(y \mid x) \| q_\phi(y \mid x)) = 0. \end{aligned} \tag{4}$$

Note that the second and third constraints serve as posterior regularization, effectively chaining $p^*(y \mid x)$, $q_\phi(y \mid x)$, and $p_\theta(y \mid x)$ together. Since $\mathcal{X}_\theta \subseteq \mathcal{X}^*$ and by the properties of the Kullback-Leibler divergence (non-negativity and identity of indiscernibles), it follows that for all $x \in \mathcal{X}_\theta$, $p_\theta(y \mid x) = q_\phi(y \mid x) = p^*(y \mid x)$. Thus, the primal objective is equivalently expressed as,

$$\begin{aligned} \min_{\theta,\phi}.\ & d(p^*(x), p_\theta(x)) \\ \text{s.t. } & \mathbb{E}_{x\sim p_\theta} H(p^*(y \mid x)) \le \epsilon \\ & \mathbb{E}_{x\sim p_\theta} \mathrm{D}_{\mathrm{KL}}(p_\theta(y \mid x) \| q_\phi(y \mid x)) = 0 \\ & \mathbb{E}_{x\sim p^*} \mathrm{D}_{\mathrm{KL}}(p^*(y \mid x) \| q_\phi(y \mid x)) = 0. \end{aligned} \tag{5}$$

Note that this objective minimizes the divergence between $p^*(x)$ and $p_\theta(x)$, subject to the constraint that $p_\theta(x)$, on expectation, cannot sample points for which $p^*(y \mid x)$ has high uncertainty. In other words, p_θ is not allowed to sample near the decision boundary of $p^*(y \mid x)$.

The Lagrangian for Eq. (4) suggests the following optimization problem

$$\min_{\theta,\phi}.\ d(p^*(x), p_\theta(x)) - \lambda_m H_\theta(Y|X) + \lambda_p \mathcal{L}_p(\theta, \phi) + \lambda_c \mathcal{L}_c(\phi), \tag{6}$$

where $\mathcal{L}_p(\theta, \phi), \mathcal{L}_c(\phi)$ are the LHS terms of the second and third constraints in Eq. (4). When $\lambda_m = \lambda_p$, Eq. (6) reduces to

$$\min_{\theta,\phi}.\ d(p^*(x), p_\theta(x)) + \lambda_m \mathcal{L}_m(\theta, \phi) + \lambda_c \mathcal{L}_c(\phi), \tag{7}$$

where $\mathcal{L}_m(\theta, \phi) = -\mathbb{E}_{x,y\sim p_\theta} \ln q_\phi(y \mid x)$. Note that this objective in Eq. (7) is exactly the AC-GAN objective in Eq. (1).

Our analysis exposes the AC-GAN objective as a Lagrangian to a constrained objective which rejects samples near the decision boundary of the classification problem. We therefore predict that AC-GAN is biased toward down-sampling points that are close to the decision boundary of the classifier, even if the true density $p^*(x)$ does contain points near the decision boundary.

3 Verifying AC-GAN's Bias

We verify the down-sampling behavior of AC-GAN in a simple density estimation experiment constructed using the MNIST dataset. Let y be one of two labels in $\{A, B\}$. We consider the following generative model for p^*: sample y from $\{A, B\}$ uniformly. If $y = A$, sample x uniformly from MNIST 0's and 1's. If $y = B$, sample x uniformly from MNIST 0's and 2's. Assuming equal numbers of 0's, 1's, and 2's digits in the MNIST dataset, then the digit distribution in our constructed $p^*(x)$ should be $[1/2, 1/4, 1/4]$ for 0's, 1's, and 2's respectively. Furthermore, the Bayes-optimal A/B classifier is maximally uncertain for 0's, meaning that the 0's digits lie exactly in the decision boundary. Our analysis predicts that, as a result, 0's will be down-sampled. We show in Fig. 1 and Fig. 2 that AC-GAN indeed down-samples images of 0's.

4 Inception Score and the Pretty Image Bias

Our analysis shows that AC-GAN down-samples difficult-to-classify images. From this perspective, it is easy to see why blurry or globally incoherent images, which are probably more difficult to classify, will be down-sampled. Furthermore, if humans consider easier-to-classify images to be prettier, then

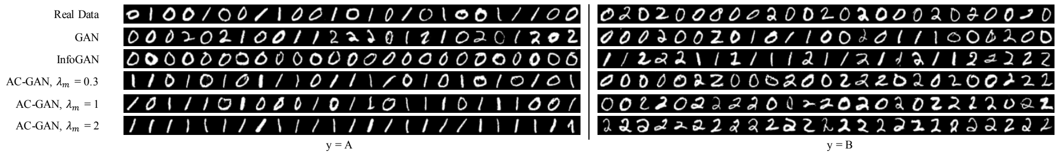


Figure 1: Visual comparison of how GAN, InfoGAN, and AC-GAN learns to use the discrete latent variable y and demonstration of the down-sampling of zero's in AC-GAN. In contrast, GAN ignores the latent code and learns the correct distribution. InfoGAN, despite also containing a mutual information maximizing term, simply re-assigns the latent code ($A \to \{0\}, B \to \{1, 2\}$) so that the mutual information objective does not conflict with the density estimation objective.

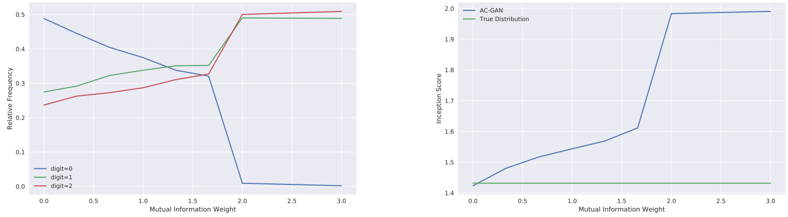
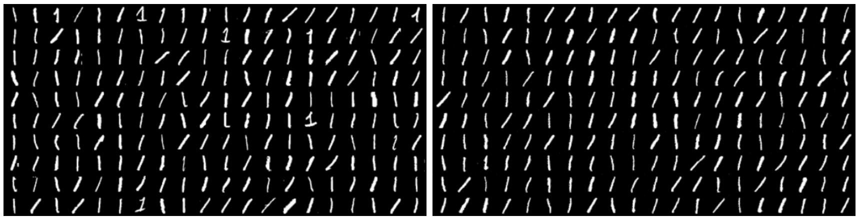


Figure 2: Evaluation of the relative frequency of each digit (0's, 1's, 2's) as we vary λ_m in the range of $[0, 2]$. Note that, for $p^*(x)$, the digit distribution should be $[1/2, 1/4, 1/4]$. Since 0's are down-sampled, the Inception Score improves too.

through AC-GAN's biased sampling, it is, in theory, possible for the sampled images to be more visually pleasing than the original distribution $p^*(x)$. In Fig. 3, we show that phenomenon holds in practice. We trained AC-GAN on the original MNIST dataset and digit labels. Since the inclusion of serifs may cause confusion between 1's and 2's, AC-GAN biases sampling toward sans-serif 1's.



(a) Real MNIST

(b) AC-GAN $\lambda_m = 2$

Figure 3: Comparison of 1's digits from the real MNIST dataset versus AC-GAN. Note that AC-GAN does not sample 1's with serifs, as this may cause confusion with 2's. The Inception Score for real MNIST is 9.80, while that for AC-GAN is 9.94—a higher score.

Our analysis also explains why AC-GAN performs well on the Inception Score [6]. If the Inception Score is computed using the Bayes-optimal classifier, the log of the Inception Score corresponds exactly to the mutual information objective $I(X, Y)$ under the joint distribution $p_\theta(x)p^*(y \mid x)$. If $p_\theta(y \mid x) = p^*(y \mid x)$ and $p_\theta(y)$ is set to be uniform, then AC-GAN is explicitly designed to maximize the mutual information term underlying the Inception Score. In cases where the true distribution $p^*(x)$ contains point near the decision boundary of $p^*(y \mid x)$, it is in fact possible for AC-GAN to out-perform the real distribution on the Inception Score metric. We verify this in our toy example in Fig. 2 and on the real MNIST dataset too in Fig. 3.

Conclusion. In this paper, we showed that the AC-GAN objective is a Lagrangian to a constrained optimization problem that reveals AC-GAN's bias against points near the decision boundary of the auxiliary classifier. This bias is most apparent when the true distribution assigns density to points near the decision boundary and causes AC-GAN to learn an incorrect distribution. We showed how the incorrect distribution can in fact achieve a better Inception Score and recommend practitioners to take this bias into account when deciding whether to use the AC-GAN objective.

References

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

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

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

[4] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[6] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.