Unbiased Auxiliary Classifier GANs with MINE

Ligong Han
Rutgers University
lh599@cs.rutgers.edu

Anastasis Stathopoulos
Rutgers University
as2947@scarletmail.rutgers.edu

Tao Xue
Rutgers University
tx57@cs.rutgers.edu

Dimitris Metaxas
Rutgers University
dnm@cs.rutgers.edu

Abstract

Auxiliary Classifier GANs (AC-GANs) [14] are widely used conditional generative models and are capable of generating high-quality images. Previous work [17] has pointed out that AC-GAN learns a biased distribution. To remedy this, Twin Auxiliary Classifier GAN (TAC-GAN) [4] introduces a twin classifier to the min-max game. However, it has been reported that using a twin auxiliary classifier may cause instability in training. To this end, we propose an Unbiased Auxiliary GANs (UAC-GAN) that utilizes the Mutual Information Neural Estimator (MINE) [1] to estimate the mutual information between the generated data distribution and labels. To further improve the performance, we also propose a novel projection-based statistics network architecture for MINE. Experimental results on three datasets, including Mixture of Gaussian (MoG), MNIST [11] and CIFAR10 [10] datasets, show that our UAC-GAN performs better than AC-GAN and TAC-GAN.

1. Introduction

Generative Adversarial Networks (GANs) [5] are generative models that can be used to sample from high dimensional non-parametric distributions, such as natural images or videos. Conditional GANs [12] is an extension of GANs that utilize the label information to enable sampling from the class conditional data distribution. Class conditional sampling can be achieved by either (1) conditioning the discriminator directly on labels [12, 8, 13], or by (2) incorporating an additional classification loss in the training objective [14]. The latter approach originates in Auxiliary Classifier GAN (AC-GAN) [14].

Despite its simplicity and popularity, AC-GAN is reported to produce less diverse data samples [17, 13]. This phenomenon is formally discussed in Twin Auxiliary Classifier GAN (TAC-GAN) [4]. The authors of TAC-GAN reveal that due to a missing negative conditional entropy term in the objective of AC-GAN, it does not exactly minimize the divergence between real and fake conditional distributions. TAC-GAN proposes to estimate this missing term by introducing an additional classifier in the min-max game. However, it has also been reported that using such twin auxiliary classifiers might result in unstable training [9].

In this paper, we propose to incorporate the negative conditional entropy in the min-max game by directly estimating the mutual information between generated data and labels. The resulting method enjoys the same theoretical guarantees as that of TAC-GAN and avoids the instability caused by using a twin auxiliary classifier. We term the proposed method UAC-GAN because (1) it learns an Unbiased distribution, and (2) MINE [1] relates to Unnormalized bounds [15]. Finally, our method demonstrates superior performance compared to AC-GAN and TAC-GAN on 1-D mixture of Gaussian synthetic data, MNIST [11], and CIFAR10 [10] dataset.

2. Related Work

Learning unbiased AC-GANs. In CausalGAN [9], the authors incorporate a binary Anti-Labeler in AC-GAN and theoretically show its necessity for the generator to learn the true class conditional data distributions. The Anti-Labeler is similar to the twin auxiliary classifier in TAC-GAN, but it is used only for binary classification. Shu *et al*. [17] formulates the AC-GAN objective as a Lagrangian to a constrained optimization problem and shows that the AC-GAN tends to push the data points away from the decision boundary of the auxiliary classifiers. TAC-GAN [4] builds on the insights of [17] and shows that the bias in AC-GAN is caused by a missing negative conditional entropy term. In addition, [4] proposes to make AC-GAN unbiased by introducing a twin auxiliary classifier that competes in an adversarial game with the generator. The TAC-GAN can be considered as a generalization of CausalGAN's Anti-Labeler to

the multi-class setting.

Mutual information estimation. Learning a twin auxiliary classifier is essentially estimating the mutual information between generated data and labels. We refer readers to [15] for a comprehensive review of variational mutual information estimators. In this paper, we employ the Mutual Information Neural Estimator (MINE) [1].

3. Background

3.1. Bias in Auxiliary Classifier GANs

First, we review the AC-GAN [14] and the analysis in [4, 17] to show why AC-GAN learns a biased distribution. The AC-GAN introduces an auxiliary classifier $\mathcal{C}$ and optimizes the following objective

$$\min_{\mathcal{G},\mathcal{C}} \max_{\mathcal{D}} L_{\text{AC}}(\mathcal{G},\mathcal{C},\mathcal{D}) = \tag{1}$$
$$\underbrace{\mathbb{E}_{x\sim P_X}\log \mathcal{D}(x) + \mathbb{E}_{z\sim P_Z, y\sim P_Y}\log(1-\mathcal{D}(\mathcal{G}(z,y)))}_{\text{ⓐ}}$$
$$-\underbrace{\mathbb{E}_{x,y\sim P_{XY}}\log \mathcal{C}(x,y)}_{\text{ⓑ}} - \underbrace{\mathbb{E}_{z\sim P_Z, y\sim P_Y}\log \mathcal{C}(\mathcal{G}(z,y),y)}_{\text{ⓒ}},$$

where ⓐ is the value function of a vanilla GAN, and ⓑ ⓒ correspond to cross-entropy classification error on real and fake data samples, respectively. Let $Q^c_{Y|X}$ denote the conditional distribution induced by $\mathcal{C}$. As pointed out in [4], adding a data-dependent negative conditional entropy $-H_P(Y|X)$ to ⓑ yields the Kullback-Leibler (KL) divergence between $P_{Y|X}$ and $Q^c_{Y|X}$,

$$-H(Y|X) + \text{ⓑ} = \mathbb{E}_{x\sim P_X} D_{\text{KL}}(P_{Y|X}\|Q^c_{Y|X}). \tag{2}$$

Similarly, adding a term $-H_Q(Y|X)$ to ⓒ yields the KL-divergence between $Q_{Y|X}$ and $Q^c_{Y|X}$,

$$-H_Q(Y|X) + \text{ⓒ} = \mathbb{E}_{x\sim Q_X} D_{\text{KL}}(Q_{Y|X}\|Q^c_{Y|X}). \tag{3}$$

As illustrated above, if we were to optimize 2 and 3, the generated data posterior $Q_{Y|X}$ and the real data posterior $P_{Y|X}$ would be effectively chained together by the two KL-divergence terms. However, $H_Q(Y|X)$ cannot be considered as a constant when updating $\mathcal{G}$. Thus, to make the original AC-GAN unbiased, the term $-H_Q(Y|X)$ has to be added in the objective function. Without this term, the generator tends to generate data points that are away from the decision boundary of $\mathcal{C}$, and thus learns a biased (degenerate) distribution. Intuitively, minimizing $-H_Q(Y|X)$ over $\mathcal{G}$ forces the generator to generate diverse samples with high (conditional) entropy.

3.2. Twin Auxiliary Classifier GANs

Twin Auxiliary Classifier GAN (TAC-GAN) [4] tries to estimate $H_Q(Y|X)$ by introducing another auxiliary classifier $\mathcal{C}^{mi}$. First, notice the mutual information can be decomposed in two symmetrical forms,

$$I_Q(X;Y) = H(Y) - H_Q(Y|X) = H_Q(X) - H_Q(X|Y).$$

Herein, the subscript Q denotes the corresponding distribution Q induced by $\mathcal{G}$. Since $H(Y)$ is constant, optimizing $-H_Q(Y|X)$ is equivalent to optimizing $I_Q(X;Y)$. TAC-GAN shows that when Y is *uniform*, the latter form of I_Q can be written as the Jensen-Shannon divergence (JSD) between conditionals $\{Q_{X|Y=1},\dots,Q_{X|Y=K}\}$. Finally, TAC-GAN introduces the following min-max game

$$\min_{\mathcal{G}} \max_{\mathcal{C}^{mi}} V_{\text{TAC}}(\mathcal{G},\mathcal{C}^{mi}) =$$
$$\mathbb{E}_{z\sim P_Z, y\sim P_Y}\log \mathcal{C}^{mi}(\mathcal{G}(z,y),y), \tag{4}$$

to minimize the JSD between multiple distributions. The overall objective is

$$\min_{\mathcal{G},\mathcal{C}} \max_{\mathcal{D},\mathcal{C}^{mi}} L_{\text{TAC}}(\mathcal{G},\mathcal{D},\mathcal{C},\mathcal{C}^{mi}) = L_{\text{AC}} + \underbrace{V_{\text{TAC}}}_{\text{ⓓ}}. \tag{5}$$

3.3. Insights on Twin Auxiliary Classifier GANs

TAC-GAN from a variational perspective. Training the twin auxiliary classifier minimizes the label reconstruction error on fake data as in InfoGAN [2]. Thus, when optimizing over $\mathcal{G}$, TAC-GAN minimizes a lower bound of the mutual information. To see this,

$$\begin{aligned} V_{\text{TAC}} =& \mathbb{E}_{x,y\sim Q_{XY}}\log \mathcal{C}^{mi}(x,y) \\ =& \mathbb{E}_{x\sim Q_X}\mathbb{E}_{y\sim Q_{Y|X}}\log Q(y|x)\frac{Q^{mi}(y|x)}{Q(y|x)} \\ =& \mathbb{E}_{x\sim Q_X}\mathbb{E}_{y\sim Q_{Y|X}}\log Q(y|x) \\ & - \mathbb{E}_{x\sim Q_X} D_{\text{KL}}(Q_{Y|X}\|Q^{mi}_{Y|X}) \\ \leq& - H_Q(Y|X). \end{aligned} \tag{6}$$

The above shows that ⓓ is a lower bound of $-H_Q(Y|X)$. The bound is tight when classifier $\mathcal{C}^{mi}$ learns the true posterior $Q_{Y|X}$ on fake data. However, minimizing a lower bound might be problematic in practice. Indeed, previous literature [9] has reported unstable training behavior of using an adversarial twin auxiliary classifier in AC-GAN.

TAC-GAN as a generalized CausalGAN. A *binary* version of the twin auxiliary classifier has been introduced as Anti-Labeler in CausalGAN [9] to tackle the issue of *label-conditioned mode collapse*. As pointed out in [9], the use of Anti-Labeler brings practical challenges with gradient-based training. Specifically, (1) in the early stage, the Anti-Labeler quickly minimizes its loss if the generator exhibits label-conditioned mode collapse, and (2) in the later stage, as the generator produces more and more realistic images, Anti-Labeler behaves more like Labeler (the other auxiliary

classifier). Therefore, maximizing Anti-Labeler loss and minimizing Labeler loss become a contradicting task, which ends up with unstable training. To account for this, Causal-GAN adds an exponential decaying weight before the Anti-Labeler loss term (or ⓓ in 5 when optimizing $\mathcal{G}$). In fact, the following theorem shows that TAC-GAN can still induce a degenerate distribution.

Theorem 1. *Given fixed $\mathcal{C}$ and $\mathcal{C}^{mi}$, the optimal $\mathcal{G}^*$ that minimizes ⓒ + ⓓ induces a degenerated conditional $Q^*_{Y|X} = \text{onehot}(\arg\min_k \frac{Q^{mi}(Y=k|x)}{Q^c(Y=k|x)})$, where $Q^{mi}_{Y|X}$ is the distribution specified by $\mathcal{C}^{mi}$.*

Proof. If $\mathcal{G}$ learns the true conditional, and $\mathcal{C}$ and $\mathcal{C}^{mi}$ are both optimally trained so that $Q^c_{Y|X} = Q^{mi}_{Y|X} = P_{Y|X}$, then ⓒ + ⓓ = 0 and the game reaches equilibrium.

If $Q^c_{Y|X}$ and $Q^{mi}_{Y|X}$ are not equal (and $Q^c_{Y|X}$ has non-zero entries),

$$\begin{aligned}
\text{ⓒ} + \text{ⓓ} =& -\mathbb{E}_{x\sim Q_X}\sum_k Q_{Y|X}(Y=k|x)\log Q^c(Y=k|x) \\
&+ \mathbb{E}_{x\sim Q_X}\sum_k Q_{Y|X}(Y=k|x)\log Q^{mi}(Y=k|x) \\
=& \mathbb{E}_{x\sim Q_X}\sum_k Q_{Y|X}(Y=k|x)\log\frac{Q^{mi}(Y=k|x)}{Q^c(Y=k|x)}.
\end{aligned}$$

The minimizing ⓒ + ⓓ is equivalent to minimizing the objective point-wisely for each x,

$$\min_{Q_{Y|X=x}} \sum_k Q_{Y|X}(Y=k|x) r_x(k),$$

where r_x is the log density ratio between Q^{mi} and Q^c. Then the optimized $Q^*_{Y|X}$ is obtained by noticing that

$$\begin{aligned}
\sum_k Q_{Y|X}(Y=k|x) r_x(k) &\geq \sum_k Q_{Y|X}(Y=k|x) r_x(k_m) \\
&= r_x(k_m) \\
&= \sum_k Q^*_{Y|X}(Y=k|x) r_x(k),
\end{aligned}$$

with $k_m = \arg\min_k r_x(k)$ and $Q^*_{Y|X} = \text{onehot}(k_m)$. □

4. Method

To develop a better unbiased AC-GAN while avoiding potential drawbacks by introducing another auxiliary classifier, we resort to directly estimate the mutual information $I_Q(X;Y)$. In this paper, we employ the Mutual Information Neural Estimator (MINE [1]).

4.1. Mutual Information Neural Estimator

The mutual information $I_Q(X;Y)$ is equal to the KL-divergence between the joint Q_{XY} and the product of the marginals $Q_X \otimes Q_Y$ (here we denote $Q_Y = P_Y$ for a consistent and general notation),

$$I_Q(X;Y) = D_{\text{KL}}(Q_{XY} \| Q_X \otimes Q_Y). \tag{7}$$

MINE is built on top of the bound of Donsker and Varadhan [3] (for the KL-divergence between distributions P and Q),

$$D_{\text{KL}}(P\|Q) = \sup_{\mathcal{T}:\Omega\to\mathbb{R}} \mathbb{E}_P[\mathcal{T}] - \log \mathbb{E}_Q[e^{\mathcal{T}}], \tag{8}$$

where $\mathcal{T}$ is a scalar-valued function which takes samples from P or Q as input. Then by replacing P with Q_{XY} and replacing Q with $Q_X \otimes Q_Y$, we get

$$\begin{aligned}
I^{mine}_Q &= \max_{\mathcal{T}} V_{\text{MINE}}(\mathcal{G},\mathcal{T}), \qquad \text{where} \qquad (9)\\
V_{\text{MINE}}(\mathcal{G},\mathcal{T}) =& \mathbb{E}_{z\sim P_Z, y\sim P_y}\mathcal{T}(\mathcal{G}(z,y),y) \\
&- \log \mathbb{E}_{z\sim P_Z, y\sim P_y, \bar{y}\sim P_Y} e^{\mathcal{T}(\mathcal{G}(z,y),\bar{y})}.
\end{aligned}$$

The function $\mathcal{T}: \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is often parameterized by a deep neural network.

4.2. Unbiased AC-GAN with MINE

The overall objective of the proposed unbiased AC-GAN is,

$$\min_{\mathcal{G},\mathcal{C}} \max_{\mathcal{D},\mathcal{T}} L_{\text{UAC}}(\mathcal{G},\mathcal{D},\mathcal{C},\mathcal{T}) = L_{\text{AC}} + V_{\text{MINE}}. \tag{10}$$

Note that when the inner $\mathcal{T}$ is optimal and the bound is tight, $V_{\text{MINE}}(\mathcal{G},\mathcal{T}^*)$ recovers the true mutual information $I_Q(X;Y) = H(Y) - H_Q(Y|X)$. Given that $H(Y)$ is constant, minimizing over the outer $\mathcal{G}$ maximizes the true conditional entropy $H_Q(Y|X)$.

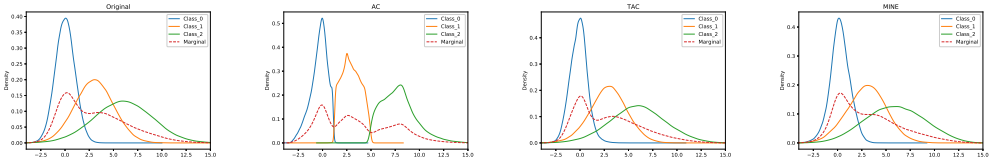
Implementation-wise, a projection-based network $\mathcal{T}$ only adds at most an embedding layer (same as same as a fully connected layer) and a single-class fully connected layer (if replacing the LogSumExp function with a learnable scalar function). Thus, UAC-GAN only adds a negligible computational cost to AC-GANs.

	AC-GAN	TAC-GAN	UAC-GAN
Class_0	0.234 ± 0.054	$\mathbf{0.077 \pm 0.091}$	0.085 ± 0.172
Class_1	4.825 ± 1.883	0.459 ± 0.359	$\mathbf{0.148 \pm 0.274}$
Class_2	527.801 ± 65.174	2.772 ± 2.508	$\mathbf{0.760 \pm 1.474}$
Marginal	52.348 ± 9.660	0.351 ± 0.779	$\mathbf{0.185 \pm 0.494}$

Table 1: MMD distance of 1-D mixture of Gaussian experiment, lower is better. UAC-GAN matches distributions better than TAC-GAN except for `Class_0`.

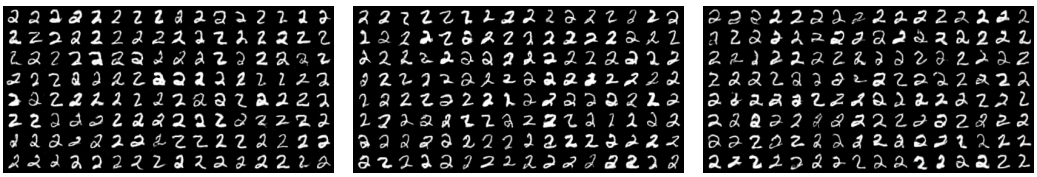
5. Experiments

We borrow the evaluation protocol in [4] to compare the distribution matching ability of AC-GAN, TAC-GAN, and our UAC-GAN on (1-D) mixture of Gaussian synthetic data. Then, we evaluate the image generation performance of UAC-GAN on MNIST [11] and CIFAR10 [10] dataset.

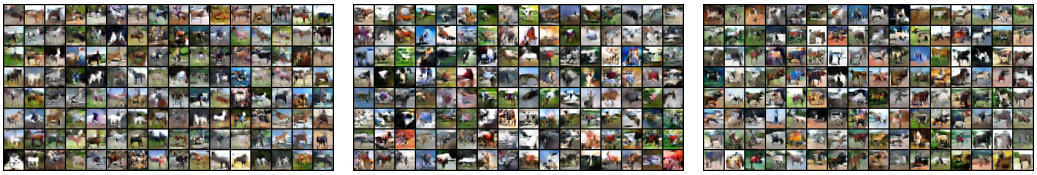


(a) Target Distribution (b) AC-GAN (c) TAC-GAN (d) UAC-GAN

Figure 1: Results on 1-D mixture of Gaussian dataset. The generated data points in (b) are well-separated, which clearly illustrates how AC-GAN learns a biased conditional distribution.



(a) AC-GAN (b) TAC-GAN (c) UAC-GAN



(d) AC-GAN (e) TAC-GAN (f) UAC-GAN

Figure 2: Results on MNIST (a-c) and CIFAR10 (d-f) dataset. Samples are drawn from a single class "2" (a-c) and "horse" (d-f) to illustrate the label-conditioned diversity.

	MNIST		CIFAR10	
Method	IS ↑	FID ↓	IS ↑	FID ↓
AC-GAN	2.52	4.17	4.71	47.75
TAC-GAN	2.60	3.70	4.17	54.91
UAC-GAN (ours)	**2.68**	**3.68**	**4.92**	**43.04**

Table 2: Inception Scores (IS) and Fréchet Inception Distances (FID) on MNIST and CIFAR10 dataset.

5.1. Mixture of Gaussian

The 1-D mixture of Gaussian (MoG) experiment is shown in Figure 1. The MoG data is sampled from three Gaussian components, $\mathcal{N}(0,1)$, $\mathcal{N}(3,2)$, and $\mathcal{N}(6,3)$, labeled as `Class_0`, `Class_1`, and `Class_2`, respectively. The estimated density is obtained by applying kernel density estimation as used in [4], and the maximum mean discrepancy (MMD) [6] distances are reported in Table 1. As shown, in most cases (except for `Class_0`), UAC-GAN outperforms TAC-GAN and is generally more stable across different runs.

5.2. MNIST and CIFAR10

Table 2 reports the Inception Scores (IS) [16] and Fréchet Inception Distances (FID) [7] on the MNIST and CIFAR10 datasets. To visually inspect whether the model exhibits label-conditioned mode collapse, we condition the generator on a single class. Samples are shown in Figure 2. It is obvious to conclude from the image samples that the proposed UAC-GAN generates more diverse images; moreover, as demonstrated in quantitative evaluations, UAC-GAN outperforms AC-GAN and TAC-GAN.

6. Conclusion

In this paper, we reviewed the low intra-class diversity problem of the AC-GAN model. We analyzed the TAC-GAN model and showed in theory why introducing a twin auxiliary classifier may cause unstable training. To address this, we proposed to directly estimate the mutual information using MINE. The effectiveness of the proposed method is demonstrated by a distribution matching experiment and image generation experiments on MNIST and CIFAR10.

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 1, 2, 3

[2] 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. 2

[3] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983. 3

[4] Mingming Gong, Yanwu Xu, Chunyuan Li, Kun Zhang, and Kayhan Batmanghelich. Twin auxilary classifiers gan. In *Advances in Neural Information Processing Systems*, pages 1328–1337, 2019. 1, 2, 3, 4

[5] 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. 1

[6] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. 4

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 4

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[9] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017. 1, 2

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 3

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1, 3

[12] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[13] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 1

[14] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017. 1, 2

[15] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019. 1, 2

[16] 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. 4

[17] Rui Shu, Hung Bui, and Stefano Ermon. Ac-gan learns a biased distribution. In *NIPS Workshop on Bayesian Deep Learning*, 2017. 1, 2