A study of quality and diversity in K+1 GANs

Ilya Kavalerov
University of Maryland, College Park
ilyak@umd.edu

Wojciech Czaja
University of Maryland, College Park
czaja@umd.edu

Rama Chellappa
Johns Hopkins University, Baltimore
rchella4@jhu.edu

Abstract

We study the K+1 GAN paradigm which generalizes the canonical true/fake GAN by training a generator with a K+1-ary classifier instead of a binary discriminator. We show how the standard formulation of the K+1 GAN does not take advantage of class information fully and show how its learned generative data distribution is no different than the distribution that a traditional binary GAN learns. We then investigate another GAN loss function that dynamically labels its data during training, and show how this leads to learning a generative distribution that emphasizes the target distribution modes. We investigate to what degree our theoretical expectations of these GAN training strategies have impact on the quality and diversity of learned generators on real-world data.

1 Introduction

When GANs were first introduced [1], deep learning had already had striking successes with learning discriminative models while high quality deep generative models had yet to appear. The GAN framework combined these two models in an adversarial setting that allowed the successes of discriminative models to be used to train generative models that produced realistic samples. The first GAN pitted a two class real/fake discriminator against a generator network and real images [1]. Since then a multitude of innovations in the architectures of these networks, and loss functions for how to combine them, have flourished [2], with the chief goals of improving sample diversity and quality [3].

Many works have found new ways to incorporate class information into the GAN training process to improve image generation [4, 5, 6, 7, 8, 9]. In this work we focus on K+1 GANs which combine the task of real/fake discrimination with class discrimination directly [9]. K+1 GANs proved most successful in improving semi-supervised learning classification performance [9, 10, 7], but have more recently been challenged by a different class of methods [11, 12]. They have never been as widely successful as GANs with auxiliary classifiers (ACGAN) [4], or GANs that use class embeddings without the classification task [6, 13, 14]. Here we investigate why the GAN architecture which most closely unifies real sample classification with fake sample classification has been overtaken by other methods.

We study a hypothesis testing inspired extension of the binary GAN into the K+1 setting [9], where K real distributions are classified against each other, and against a fake distribution. We show how the optimal discriminator/classifier in this setting is a straightforward generalization of the optimal binary GAN discriminator, and show that the formulation of the log likelihood loss prevents the generator from fully benefiting from supervision over the classes of real samples. Then we then introduce another generator loss criterion that uses dynamic labeling [7], and describe how it has increased class specificity over the K+1 GAN by emphasizing the modes of the real data. Finally we

1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020), Vancouver, Canada.

demonstrate with experiments on CelebA and CIFAR10 to what degree both GAN formulations have an effect on increasing the quality and diversity of samples over the original binary GAN and more popular ACGAN.

2 GANs and their multi-class generalizations

2.1 Binary GANs and K+1 GANs

The generalization of the binary GAN [1], known as the K+1 GAN (and also as LabelGAN [7]), was originally developed for semi-supervised learning with GANs [9]. We let $M = K + 1$ denote the total number of classes, K real classes plus the fake class, which we refer to as the zeroth class. This generalization echoes the form of the generalization of binary minimax hypothesis testing to be M-ary [15]. We use the notation that X, H, D, Z are respectively the images, hypotheses (fake, real class 1, real class 2, etc), decisions, and random codes. Our discriminator $d(x)$ outputs the vector $d_m(x) = \mathbb{P}[D = m|X = x], m = 0, 1, \ldots, M-1$. The objective function of the binary GAN naturally generalizes to training all outputs of the discriminator:

$$\begin{aligned} \mathrm{L}_{\mathrm{K+1}}(d, g) &= \sum_{m=0}^{M-1} p_m \mathbb{E}[\log \mathbb{P}[D = m \mid X] \mid H = m] \\ &= \sum_{m=1}^{M-1} p_m \int \log d_m(x) f_m(x) dx + p_0 \int \log d_0(g(z)) f_0(g(z)) dF(g(z)) \end{aligned} \tag{1}$$

We denote the fake and real densities as $f_h(x) = \mathbb{P}[X = x \mid H = h],\ h = 0, 1, \ldots, K$ and p_m is the prior on each hypothesis (often taken to be uniform). For every image x the optimal decision vector $d^*(x)$ exists and each of its M elements is given by $d^*_m(x) = p_m f_m(x) / \sum_{i=0}^{M-1} p_i f_i(x)$. This can be proved for discrete densities using the classic optimality criterion for differentiable convex objectives [16]. Plugging in $d^*(x)$ into Equation (1) we see:

$$\begin{aligned} \mathrm{L}_{\mathrm{K+1}}(d^*, g) &= \sum_{m=0}^{M-1} \int \log \frac{p_m f_m(x)}{f_{\mathrm{avg}}(x)} p_m f_m(x) dx \\ &= \sum_{m=0}^{M-1} KL(f(\cdot, m) \mid\mid f_{\mathrm{avg}}(\cdot)) \end{aligned} \tag{2}$$

where $f_{\mathrm{avg}}(x) = \sum_{m=0}^{M-1} p_m f_m(x)$. The task of the generator is $\min_{f_0} \mathrm{L}_{\mathrm{K+1}}(d^*, g)$, which for discrete PMFs f_m is a convex optimization problem, the solution is $p_0 f_0(x) \rightarrow f_{\mathrm{real}}(x) = \sum_{m=1}^{M-1} p_m f_m(x)$. This is consistent with the result from [1] for $M = 2$ and $p_0 = p_1 = 1/2$.

The typical intuition in training neural networks is that learning from more data and sharing parameters across tasks leads to better performance. But here, surprisingly the theory states that by using class information in the $K + 1$ GAN to train a discriminator that can discriminate between classes, and between fake images and each of the classes, is no better than ignoring the class information. That is, if we have labeled classes for data, but choose to throw it away by merging $f_1, \ldots, f_K$ into a single f_{real}, the generator of the binary GAN that we train should have the same distribution f_0 as if we trained the K+1 GAN. This result is also consistent with examining the gradients of the K+1 GAN which reveal a "overlaid-gradient" problem whereby the overall gradient w.r.t. a generated example is the same as that in the binary GAN [7].

The $K + 1$ GAN's original purpose was however to improve the state of the art in semi-supervised classification [9], rather than improve generation of images (though human annotators were said to prefer images from a K+1 GAN). We note that in Equation (1) we omitted the part of the loss for unlabeled real images. Nevertheless we have shown that counter-intuitively, training a fully supervised $K + 1$ discriminator to classify the real examples in parallel with telling fake examples apart from real examples does not change the generator learned.

2.2 Mode Emphasizing GAN via dynamic labeling

One extension that increases the class specificity of the generator learned by a K+1 GAN is to introduce dynamic labeling [7]. Such a choice leads to f_0 converging to emphasize the modes of $f_1, \ldots, f_K$, rather than being the average of them f_{real}.

Looking at the last integral in Equation (1), we notice that the g^* that solves $\min_g \mathbb{E}[\log \mathbb{P}[D = 0 \mid X] \mid H = 0]$ also solves $\min_g -\mathbb{E}[\log \mathbb{P}[D \neq 0 \mid X] \mid H = 0] = \min_g -\mathbb{E}[\log \sum_{m>0} \mathbb{P}[D = m \mid X] \mid H = 0]$. This equivalence provides the intuition for why the density of the generator will be the average of all the real distributions. Let: $\mathrm{L}_{\mathrm{Dyn}}(d, g) = -\mathbb{E}[\log \max_{l>0} \mathbb{P}[D = l \mid X] \mid H = 0]$. That is for each fake $g(z)$ we look at the "most likely mistake" the discriminator makes $d(g(z))$, and train g to emphasize it. For d we keep the K+1 GAN objective unchanged. If we let d^* be optimal in the sense of the previous section, then

$$\min_g \mathrm{L}_{\mathrm{Dyn}}(d^*, g) = \min_g KL(f_0 \mid\mid \text{ptwise } \max_{m>0} f_m) - KL(f_0 \mid\mid f_{\text{avg}}) \tag{3}$$

Comparing this to Equation (2), our intuition expects $f_0(x) \approx \max_{m>0} f_m(x)$ with Dynamic Labeling instead of $f_0(x) = \sum_{m>0} f_m(x)$ in K+1 GAN for the following reason. We note that $\min_{f_0} KL(f_0 \mid\mid \text{ptwise } \max_{m>0} f_m)$ upper bounds Equation (3), for discrete PMFs it is a convex optimization problem, and the solution is $f_0(x) \to \max_{m>0} f_m(x)$.

Dynamic labeling GAN was found to be experimentally preferable to competing non-dynamically labeled GANs in the sense of Inception Score (IS) [7]. Our analysis here, which looks at the unconditional K+1 GAN with dynamic labeling, explains why but also reveals a potential failure mode. IS is proportional to the KL divergence of $\mathbb{P}(Y|X)$ and $\mathbb{P}(Y)$ for images X and labels Y from a pretrained Inception v3 network [9, 17]. Thus a generator that emphasizes all the modes (classes) of the real data distribution should generate easily recognizable images and have low entropy $\mathbb{P}(Y|X)$ and have a high entropy $\mathbb{P}(Y)$ by displaying all the classes, giving a high IS. However emphasizing the modes of the data could also result in lower observed diversity from finite samplings of g since the density of f_0 is concentrated around the modes. For this reason we also refer to this GAN as "Mode GAN".

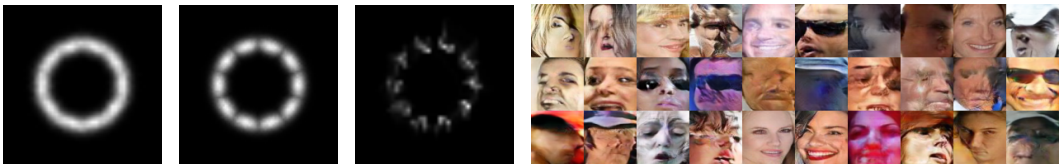
2.3 Other multi-class GANs

In this work we focus on the properties of incorporating class information into GANs via a $K + 1$ architecture. Some other notable GAN architectures that utilize class specific information include Auxiliary Classifier GANs (ACGANs) [4], and GANs that use projection discrimination with spectral normalization (SNGAN) [6]. ACGANs use a standard binary GAN architecture and train an additional fully connected classification layer on the penultimate features of the discriminator. ACGANs don't completely unify the tasks of discrimination and classification like $K + 1$ GANs, but parameters are shared between the two tasks. In projection discrimination GANs, the one hot class label is embedded with a spectrally normalized matrix, and the penultimate features of a binary discriminator network are projected onto this embedding (dot product of two vectors) and added to the binary discriminator output. Projection discrimination has become the state of the art way to incorporate class information in GANs [14, 18]. Interestingly, these learned embeddings are not naturally accurate classifiers by themselves, but adding a training objective to improve their classification ability can improve GAN metrics like IS and FID further [19].

3 Experiments and discussion

3.1 Datasets and methods

We demonstrate our findings on K+1 GANs using three datasets: synthetic 2D Gaussians, CelebA, and CIFAR10. Our Gaussian data is created from 10 highly overlapping 2D Gaussians whose means are distributed uniformly on a circle, samples from different Gaussians are different "classes". Samples from all the classes are shown in Figure 1i. CelebA is a dataset with 40 binary attribute annotations and 5 landmark locations. The images we use are 64×64 and have the background cropped. We split this dataset into 5 classes: Front facing female/male with mouth open/closed are 4, and all faces rotated more than 15 degrees are the 5th. CIFAR10 is a 32×32 dataset with 10 classes. We use all 50,000 images for training, and the 10 classes are the labels for our supervised models.



(i) Data (ii) f_0 Expected (iii) f_0 Learned (iv)

Figure 1: Mode emphasis of Dynamic Labeling GAN. (i) shows 10 overlapping classes of 2D Gaussian data arranged on a circle within $(0, 4) \times (-1, 5)$. (ii) shows the PMF that is the solution of the convex upper-bound of Equation (3), this is the distribution that we expect a GAN trained with $\mathrm{L}_{\mathrm{Dyn}}$ to learn. (iii) shows the histogram of a generator trained with $\mathrm{L}_{\mathrm{Dyn}}$. (iv) shows a random selection 10 images from each GAN that have a MTCNN confidence of 0. The first row from is from a binary GAN, the second row is from ACGAN, and the third row from Dynamic Labeling GAN. This shows that even if modes are emphasized, this does not necessarily mean samples are not realized in what could be considered undesirable low density regions of f_{real}.

We train three SN GAN networks [13] (Binary GAN, ACGAN, Mode GAN) with default hyperparameters and only vary the loss function used. We also train two additional SN GAN networks (Mode GAN and K+1 GAN) that also used projection discrimination [20]. All networks were trained for 200k iters for CelebA and 500k iters for CIFAR10.

Quantitative evaluation of GANs is itself an active area of research and many information measures for GANs exist. We focus on the two most popular, Inception Score (IS) and Frechet Inception Distance (FID) [9, 21]. The Inception Score calculates $\mathrm{KL}(\mathbb{P}(Y|X) \mid\mid \mathbb{P}(Y))$, where $\mathbb{P}$ is a reference pretrained Inception v3 network. It evaluates whether generated images can be classified with confidence by the reference network. Frechet Inception distance calculates $\mid\mid \mu_r - \mu_g \mid\mid + \mathrm{Tr}(\mathrm{Cov}_r + \mathrm{Cov}_g - 2\sqrt{\mathrm{Cov}_r \mathrm{Cov}_g})$ from the penultimate features of the same reference network. That is μ_r and μ_g are the mean feature vectors of the real images and generated images respectively, and Cov_r and Cov_g are the covariance matrices of the features of real and generated images respectively.

3.2 Discussion

When the form of the true densities $f_m(x)$ are known, we can observe the learned density of the generator f_0 via the histogram of its samples. For the Dynamically Labeled GAN we see the PMF that minimizes the convex upper-bound of Equation (3) in Figure 1ii emphasizes the modes of the data. In Figure 1iii we see that the learned g from minimizing Equation (3) while training a GAN indeed emphasizes the modes of the data: the samples are in high density around the means of each class, and are low density in the overlap between classes. However, this simple example does not generalize neatly to real data. We may expect an analogous situation in the domain of faces to be that few samples exist between the "means" of the 5 classes we chose: i.e. no interpolations between males facing left and females facing right. But we see in Figure 1iv that Dynamically Labeled GAN generates unwanted low quality samples just like the baseline Binary GAN and ACGAN networks.

To measure the quality of our CelebA GAN's output, we use the confidence output of MTCNN, a publicly available pretrained face detector, to measure the frequency of bad generator samples. In Figure 1iv are examples of images that MTCNN assigns 0 confidence output by all 3 GANs. For images that are far away from the modes of the CelebA data density, for example images that blur together multiple face orientations into 1 image, MTCNN will assign a low confidence. The right hand side of Figure 2 shows out of 100,000 samples generated by each GAN, the percentage of images with MTCNN confidence greater than $c = 0, 1 - 10^{-1}, \ldots, 1 - 10^{-4}$ is always largest in Mode GAN, but only by a small margin. And as shown in the figure, all the networks produce a score 0 face 4-6 times every 10000 images.

The left hand side of Figure 2 shows that performance was tied for the three networks on CelebA and CIFAR10. Adding projection discrimination (PJ) to the networks yeilded a much greater difference in IS performance than the changes in loss function. We show some samples from the CIFAR10 GANs in Figure 3, which are hard to qualitatively rank.

Method	**FID** CelebA	**IS** CIFAR-10
SN GAN	4.13	8.03
SN ACGAN	5.07	8.02
SN Mode GAN	3.65	8.13
PJ SN Mode GAN		8.50
PJ SN K+1 GAN		8.43

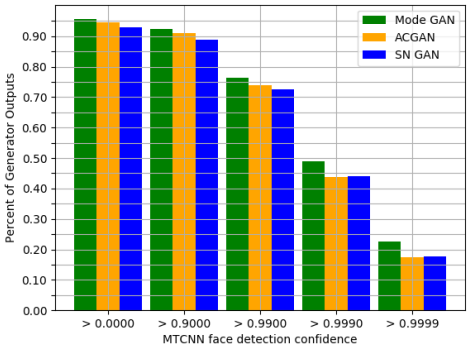


Figure 2: On the left are FIDs image generation on CelebA and Inception scores for CIFAR-10. On the right is the percentage of 100,000 samples generated by each GAN for which MTCNN has greater than c confidence that a face is present, where $c = 0, 1-10^{-1}, 1-10^{-2}, 1-10^{-3}, 1-10^{-4}$. We use the publicly available pretrained MTCNN face detector network to compute the confidence scores.

(i) SNGAN (ii) ACGAN (iii) Mode GAN (iv) CIFAR10

Figure 3: Generated images on CIFAR10 for the three methods we compare. Images are randomly selected from each generator and sorted by class by an auxiliary DenseNet121 classifier with 95% test accuracy into the 10 classes of CIFAR10: plane, car, bird, cat, deer, dog, frog, horse, ship, truck. In iv we plot the mean pairwise distance of the outputs of the three GANs. That is for each image in a batch of size in the range $n * 128, n = 1, 2, \ldots, 79$ an embedding was generated by an auxiliary network, and then the pairwise Euclidean distances were computed. For each of these 79 batches of increasing size the mean embedding distance per generator was recorded. In each plot the average of 10 trials is shown with the standard deviation shaded. The embedding network was the 1,024 dimension penultimate layer of a DenseNet-121.

To measure diversity within a batch of generated images, we embed the images in a feature space with an auxiliary network, calculate all the pairwise distances, and record the mean. In Figure 3iv we see Mode GAN always has a higher mean pairwise distance only within a margin of error. Thus we don't observe that fitting the modes of real data yields a big impact on quality or diversity of samples.

4 Conclusion

This work provided a hypothesis testing perspective on the canonical binary GAN, its K+1 GAN generalization, and an extension meant to increase the class specificity during training, Dynamically Labeled GAN. We showed that the GAN's generator is not benefited by the classification of real vs real samples in K+1 GANs. And we demonstrated that Dynamically Labeled GAN emphasizes the modes of the real data distributions, but this has a tenuous link to increased quality and diversity of generated samples.

References

[1] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 2672–2680. MIT Press, 2014.

[2] Hao-Wen Dong and Yi-Hsuan Yang. Towards a Deeper Understanding of Adversarial Losses. *arXiv:1901.08753 [cs, stat]*, January 2019. arXiv: 1901.08753.

[3] Martin Arjovsky and Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. *arXiv:1701.04862 [cs, stat]*, January 2017. arXiv: 1701.04862.

[4] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[5] Jost Tobias Springenberg. Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks. *arXiv:1511.06390 [cs, stat]*, November 2015. arXiv: 1511.06390.

[6] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018.

[7] Zhiming Zhou, Han Cai, Shu Rong, Yuxuan Song, Kan Ren, Weinan Zhang, Jun Wang, and Yong Yu. Activation maximization generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[8] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[9] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. *arXiv:1606.03498 [cs]*, June 2016. arXiv: 1606.03498.

[10] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan R Salakhutdinov. Good Semi-supervised Learning That Requires a Bad GAN. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6510–6520. Curran Associates, Inc., 2017.

[11] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5050–5060. Curran Associates, Inc., 2019.

[12] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv:2001.07685 [cs, stat]*, January 2020. arXiv: 2001.07685.

[13] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. *arXiv:1802.05957 [cs, stat]*, February 2018. arXiv: 1802.05957.

[14] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[15] H Vincent Poor. *An introduction to signal detection and estimation*. Springer Science & Business Media, 2013.

[16] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[17] Shane Barratt and Rishi Sharma. A Note on the Inception Score. *arXiv:1801.01973 [cs, stat]*, June 2018. arXiv: 1801.01973.

[18] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[19] Ilya Kavalerov, Wojciech Czaja, and Rama Chellappa. cGANs with Multi-Hinge Loss. *arXiv:1912.04216 [cs, stat]*, November 2020. arXiv: 1912.04216.

[20] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[21] 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. *arXiv:1706.08500 [cs, stat]*, January 2018. arXiv: 1706.08500.