Investigating Shifts in GAN Output-Distributions

Ricard Durall
Fraunhofer ITWM
IWR, University of Heidelberg
Fraunhofer Center Machine Learning
ricard.durall.lopez@itwm.fraunhofer.de

Janis Keuper
Fraunhofer ITWM
IMLA, Offenburg University
janis.keuper@itwm.fhg.de

Abstract

A fundamental and still largely unsolved question in the context of Generative Adversarial Networks is whether they are truly able to capture the real data distribution and, consequently, to sample from it. In particular, the multidimensional nature of image distributions leads to a complex evaluation of the diversity of GAN distributions. Existing approaches provide only a partial understanding of this issue, leaving the question unanswered. In this work, we introduce a loop-training scheme for the systematic investigation of observable shifts between the distributions of real training data and GAN generated data. Additionally, we introduce several bounded measures for distribution shifts, which are both easy to compute and to interpret. Overall, the combination of these methods allows an explorative investigation of innate limitations of current GAN algorithms. Our experiments on different data-sets and multiple state-of-the-art GAN architectures show large shifts between input and output distributions, showing that existing theoretical guarantees towards the convergence of output distributions appear not to be holding in practice.

1 Introduction

The efficient and accurate modeling of complex multi-modal data distributions is one of the core problem behind many machine learning tasks. While most of the recently very successful classification algorithms can retreat to the simpler sub-problem of finding separation-functions in data distributions, the task of generative algorithms is to provide a model able to reproduce such complex multi-modal data distributions to their full extent. One of the most popular approaches towards generative models are Generative Adversarial Networks (GANs) [1]. GANs have been used for a wide range of use cases and data distributions [2], including numerous applications involving image data generation [3].

In this work, we study the distribution shifts in GANs outputs, and evaluate to what extent GANs can capture the full diversity of the underlying true distribution. To accomplish these tasks, we design a loop-training strategy, where the misalignment between real and generated data distributions is amplified, allowing in this way a painstaking investigation. On top of that, we introduce shift indicators on sub-distribution of the real image space, which allow an easy evaluation of lower bounds for very complex distribution shifts. Finally, we evaluate structural differences between in- and output-distributions for commonly used GAN architectures on several standard benchmarks. Our experiments show that the observable distribution shifts in our indicators are so large, that generated samples can easily be detected by the use of binary classifiers. Taken together, these experimental results provide strong evidence that GANs learn a non-trivial but shifted version of the true distribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

2 Related Work

The original formulation of Generative Adversarial Networks [1] describes GAN training as an optimization problem, minimizing the Jensen-Shannon divergence between generated and real data distributions. Even though there are several theoretical discussions of GAN convergence, like [4, 5, 6], they all come with certain constraints or assume some preconditions – leaving very little guarantees for practical real-world applications. Despite producing visual appealing results on images, [7] showed that this gap does exist and GANs do not produce samples from the real data distribution, but from an approximated low dimensional manifold. These distribution shifts, approximations, lead to limitations in terms of data synthesization and generalization. Recent work [8, 9] also showed that GANs systematically fail to reproduce image distributions in the frequency domain.

3 Methods

Loop-Training. We propose a loop-training scheme (see Figure 1) to highlight the demotion of generated images due to the distribution shifts in GANs models. It consists of (1) training the model and (2) generating a new dataset using the trained generator.

Observable Sub-Distributions of Images. We denote real images $\mathbf{x} \in \mathbb{I}^m \subset \left(\mathbb{R}^{\sqrt{m}} \times \mathbb{R}^{\sqrt{m}} \times 3\right)$ as samples from the space $\mathbb{I}^m$ of all possible color images of a given size m. As real distributions $p_r(\mathbb{I}^m)$ in this image space are typically extremely complex and hard to compare explicitly, we retreat to the analysis of simpler sub-distributions which provide computable but meaningful measures for observable lower bounds of shifts in the full image distributions.

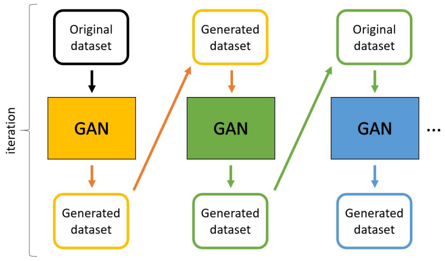

Figure 1: Iterative training scheme. Each GAN block represents one iteration, i.e., a full standard training. In our experiments, we repeat this loop 6 times.

FID: The Standard Measure for GAN Distribution Quality. Fréchet Inception Distance (FID) [10] is the metric commonly used to assess the distribution of generated images. It compares the distribution of generated images with the distribution of those real images used to train the model in the feature-space of a pre-trained Inception-v3 classification model [11] model.

Color Histograms. Besides FID, we propose to use color-distributions as an indicator. Although it neglects all spacial information, its correctness is still a necessary condition for the distribution of generated data $p_g(\mathbb{I}^m)$ and thus defines a suitable bound for the estimation of distribution shifts. We model these color-distributions in form of n averaged samples of color-histograms $\mathcal{H}_b(\mathbf{x}) \in \mathbb{R}^{3b}$ with b bins

$$p_r \approx \frac{1}{n} \sum_{i=0}^{n} \mathcal{H}_b(\mathbf{x_i}), \tag{1}$$

p_g is computed respectively with $\bar{x}_i$.

Class Histograms. We also propose to analyse the shift in class distributions. Here we apply a pre-trained classifier on the generated data and compute the class histograms which are a suitable indicator for partial collapse of output distributions.

Measuring Distribution Similarities. We employ the Kullback-Leibler divergence (KL) to measure the similarities between p_g and p_r (or $\mathcal{N}(0,1)$ respectively) formulated as

$$KL(p_g||p_r) = \sum_i p_g(i) \log \frac{p_g(i)}{p_r(i)}, \tag{2}$$

where $p_r(i)$ denotes the i-th element of the discrete representation (see Equation 1).

4 Empirical Evaluation of GAN Shifted Distributions

For a fair comparison and reproducibility, we conduct the experiments using a standardized GAN library [12]. This allows to train all models under the same conditions using different datasets.

Models. We employ a set of different GANs: Wasserstein GAN with Gradient Penalty (WGAN-GP) [13], Spectral Normalization GAN (SNGAN) [14], Self-supervised GAN (SSGAN) [15], and InfoMax-GAN (InfoGAN) [16].

Datasets. We evaluate results on seven commonly used datasets: CIFAR-10 [17], CIFAR-100 [17], STL-10 [18], CelebA [19] at two different resolutions, ImageNet[1] [20], and LSUN-Bedroom[2] [21].

4.1 Experimental Results

Demotion of Distributions in the training loop. Figure 2 shows the demotion of the FID score over training iterations, using different data-sets and models. We can observe that the FID undergoes an almost linear decline in GANs performances, independent of the scenario. This behavior reveals that the feature space distribution of the generated data is "degenerating" with each iteration, increasing the shift towards the distribution of the real data.

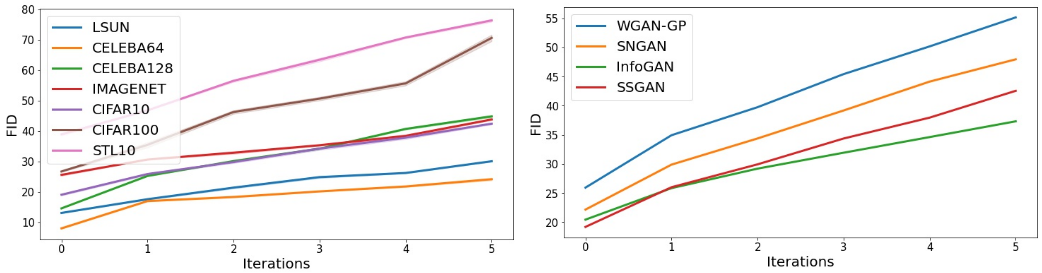

Figure 2: FID curve evolution for loop-training in different scenarios. Independent of the source dataset and the architecture, all experiments show a decrease in FID score over time. (Left) Different datasets on SSGAN. (Right) Different architectures on CIFAR10.

We hypothesize that this behavior is caused by larger distributions shifts, which is also supported by the strong degradation of class distributions shown in Figure 3. As one might expect, those models that undergo several iterations will inevitably start to suffer drastically from mode collapse. Moreover, we observe empirically, that this misalignment between the real and the synthetic distributions cannot not be prevented by alternative loss formulations such as Wasserstein [13], nor by advanced regularizes such as spectrum normalization [14] nor even by semi-supervised approaches such as SSGAN [15].

Observation: Strong Shift Towards Gaussian Sub-Distributions. Our empirical analysis of the color-histogram sub-distributions of the same experiments show another surprising effect: the distributions appear to converge towards a Gaussian over the iterations of the loop-training. Figure 4 shows a typical example of this observation. Measuring the KL between color-distributions of generated and real data, we see a strong correlation to the increase in the FID score in Figure 5.

Impact of Distribution Shifts on Individual Samples. The previous results show strong shifts in the distributions from real input images to GAN generated images. However, these distributions have been estimated from a larger number of samples. Hence, it is also interesting to assess if these shifts are systematic enough to be detectable in single samples. We therefore train simple classifier models on the color-histograms of single samples, labelled as real or generated. Table 1 shows that it is, in fact possible, to predict generated data with up to 85% accuracy for a wide range of investigated datasets.

dataset	LSUN	CelebA128	CelebA64	ImageNet	CIFAR10	CIFAR100	STL10
SVM	75.37%	74.81%	74.62%	81.00%	76.25%	77.12%	82.87%
Random Forest	75.12%	71.25%	76.50%	82.75%	80.37%	79.12%	85.00%

Table 1: Test accuracy on classification of samples: real or generated image (from iteration 0). For each dataset, we use 10K samples split into 80% training and 20% testing set.

[1] For practical reasons, we employ tiny-ImageNet.

[2] For practical reasons, we train with 300K images instead of the whole dataset.

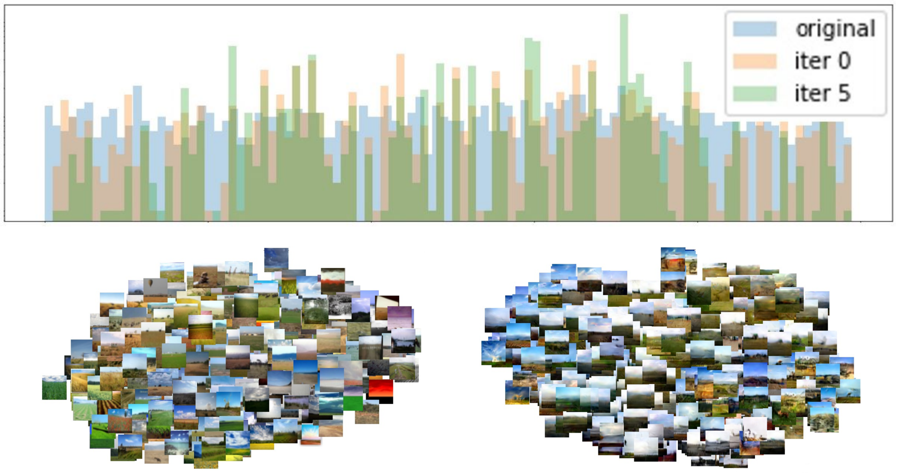

Figure 3: (Top) Histograms from original, iteration 0 and iteration 5 datastes on CIFAR100, showing the evolution of the amount of classes and the number of class members. Iteration 5 contains roughly only half of the classes, and additionally, they are quite unbalanced. (Bottom) Decline on intra-class variance. Real data on the left, generated data from iteration 5 on the right. The latter shows clear signs of mode collapse.

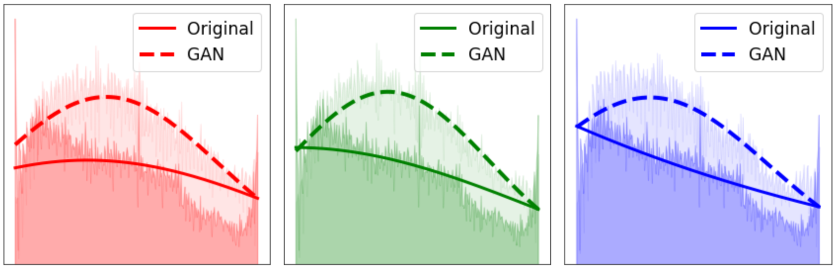

Figure 4: Accumulated color histograms of all real CIFAR10 train images of the class "Car" (strong color) with its best Gaussian fit (solid line) and the respective generated samples for the same class (light color) and their Gaussian fit (dashed line)

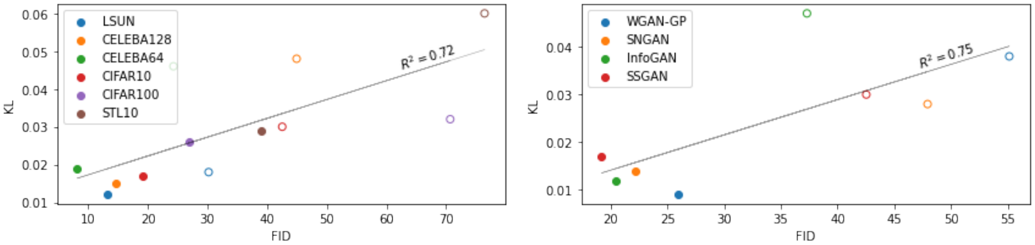

Figure 5: Correlation (R^2) between FID and KL, where the latter is computed between the real and the generated color histogram distributions. Filled markers represent results on iteration 0, and the empty ones iteration 5. (Left) Different datasets on SSGAN*. (Right) Different architectures on CIFAR10. *ImageNet follows the same tendency, but with a much pronounced slope.

5 Discussion and outlook

We present clear experimental results, investigating the distribution shifts, from which GANs architectures suffer. Furthermore, we introduce a sub-distribution based on color-image space, which can be utilized to gain valuable insights of the generator capacity. We also find a strong correlation between sub-distribution shifts and the FID scores. Taken together, these results show that GANs learn a shifted version of the true distribution, and immediately raise the question, why discriminators appear to be unable to detect evident differences between input and output distributions of a GAN.

References

[1] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[2] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*, 2020.

[3] Yongjun Hong, Uiwon Hwang, Jaeyoon Yoo, and Sungroh Yoon. How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys (CSUR)*, 52(1):1–43, 2019.

[4] Gérard Biau, Benoît Cadre, Maxime Sangnier, Ugo Tanielian, et al. Some theoretical properties of gans. *Annals of Statistics*, 48(3):1539–1566, 2020.

[5] Jerry Li, Aleksander Madry, John Peebles, and Ludwig Schmidt. On the limitations of first-order approximation in gan dynamics. In *International Conference on Machine Learning*, pages 3005–3013. PMLR, 2018.

[6] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

[7] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[8] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7890–7899, 2020.

[9] Steffen Jung and Margret Keuper. Spectral distribution aware image generation. *arXiv preprint arXiv:2012.03110*, 2020.

[10] 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 preprint arXiv:1706.08500*, 2017.

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[12] Kwot Sin Lee and Christopher Town. Mimicry: Towards the reproducibility of gan research. *arXiv preprint arXiv:2005.02494*, 2020.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[14] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[15] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019.

[16] Kwot Sin Lee, Ngoc-Trung Tran, and Ngai-Man Cheung. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. *arXiv preprint arXiv:2007.04589*, 2020.

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

[18] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[21] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

6 Supplementary Material

This supplementary section contains an extended version of the implementation details that we have presented in the main paper. Additionally, we provide more results, discussing the direct effect that distribution shifts have on our sub-distribution.

6.1 Details on Loop-Training

Figure 1 shows the training pipeline of the loop-training experiment, introduced in the main paper. We conduct further evaluations using the loop-training setup. In particular, we compare FID results between loop-and long-training pipelines. The difference between these experiments is that despite both trainings train the same amount of epochs, the latter trains with real data as reference during the whole training. In Figure 6, we can observe that the distribution shifts of the long-training, i.e., the capacity of the GAN model to generate images covering the real distribution, remains stable. On the other hand, when using generated data to train the next model (the next iteration), the results start to worsen, offering every iteration a more skewed and shifted distribution of the output distributions.

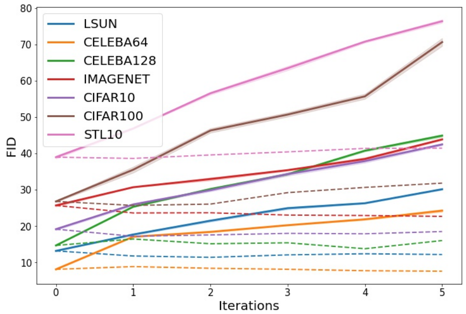

Figure 6: FID curve evolution on different datasets on SSGAN. Solid lines indicate the score for loop-trainings, and dashed lines for long-trainings.

6.2 Details on Observable Sub-Distributions

Figure 7 illustrates how is measured the distribution similarity. As we explained before, the idea is to compute the best Gaussian fit and the proportion that this fit represents for the whole Gauss curve to weight the final KL scores. In this way, we can penalize those fits that barely have a Gaussian shape but still have a good fitting, e.g., those fits that just cover a small part of a tail.

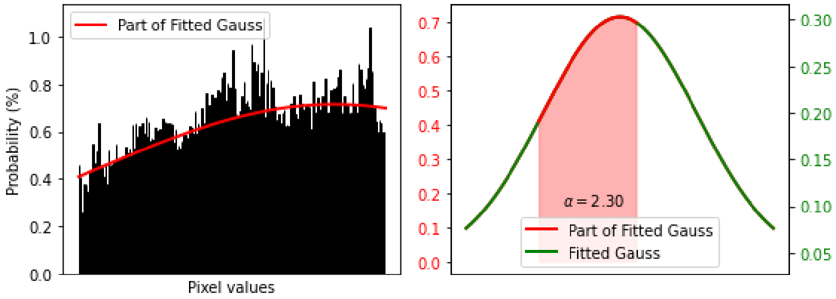

Figure 7: (Left) Example of a histogram and its Gaussian fit. (Right) Visualization of the fit w.r.t. the whole Gaussian curve. Factor $\alpha := \frac{1}{Area\ fit}$ weights the final KL scores.

6.3 Color-Distribution Cluster Analysis

Following the classification analysis from the main paper, we propose this time an unsupervised clustering approach for the color-distribution space. Using this rather simple method, one can easily visualize the mode-differences between in- and output distributions, even allowing to pinpoint single samples from the training data which can not be reconstructed by the GAN. Showing again the notorious capacity of GANs to model entirely the real distribution.

We run clustering experiments based on k-means, where we set different sizes of clusters (k). In this way, we can observe which images are plausible to be reproduced by the generative model, and which ones are not. Figure 8 depicts two different experiments on CIFAR10. We can see how the real data has a more diverse distribution, while the generated samples reduce to a few clusters. As a result,

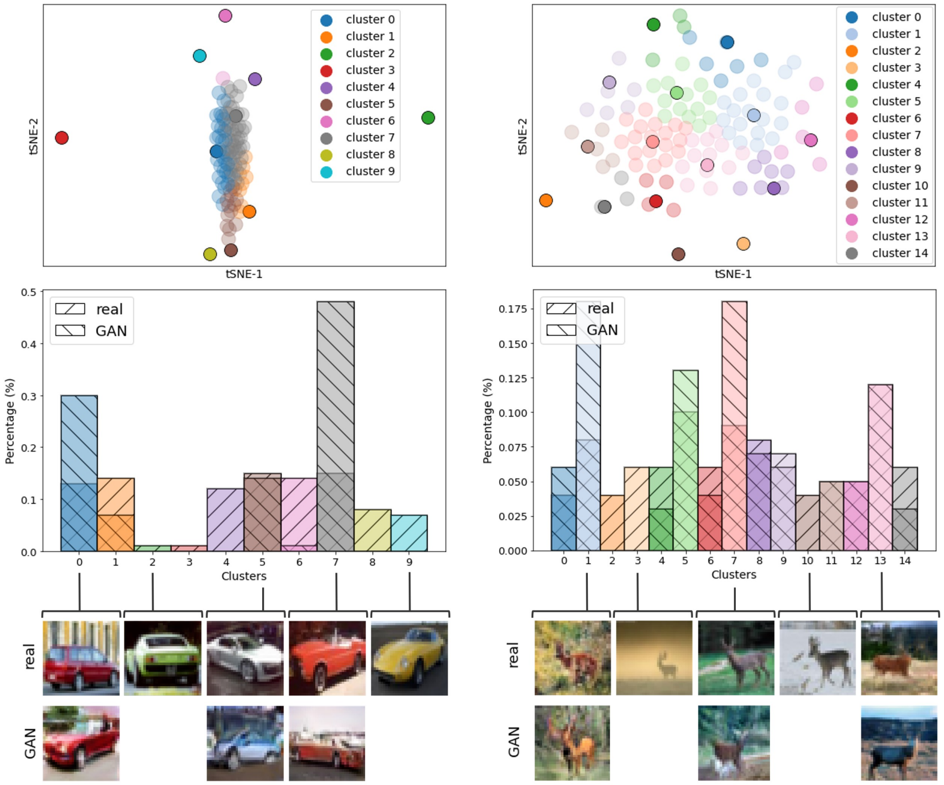

Figure 8: Clustering results for $k = 10$ on class "Car" and for $k = 15$ on class "Deer" of CIFAR10. (Top) tSNE visualization with the centroids from real data in strong markers, and the generated samples in light markers. (Bottom) Histogram of clusters frequency for real data and for generated.

it is likely that generative model starts to produce samples very similar to each other and eventually this might lead to mode collapse.

6.4 Gaussian-like Distribution Analysis

Finally, we conduct an experimental study to visualize the Gaussian tendency that output distributions of GANs seem to be shifted to. To accomplish this task, we approximate the color-distribution of generated data and real data to their best Gaussian fit, and then, we compute the KL distance between these results. Figure 9 give a good impression of the typical distribution shift that generated data undergoes, resulting in a dominant Gaussian shape of GAN outputs distributions.

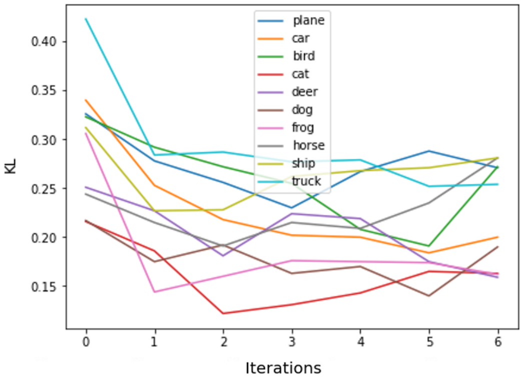

Figure 9: KL curve evolution on CIFAR10 classes. All the cases display a tendency towards Gaussian-like distributions.