MULTIMODAL NOISE AND COVERING INITIALIZATIONS FOR GANS

David Lopez-Paz
Facebook AI Research
Paris, France
dlp@fb.com

Maxime Oquab
Inria - WILLOW Team
Paris, France
maxime.oquab@inria.fr

ABSTRACT

This note describes two simple techniques to stabilize the training of Generative Adversarial Networks (GANs) on multimodal data. First, we propose a *covering initialization* for the generator. This initialization pre-trains the generator to match the empirical mean and covariance of its samples with those of the real training data. Second, we propose using *multimodal input noise* distributions. Our experiments reveal that the joint use of these two simple techniques stabilizes GAN training, and produces generators with a richer diversity of samples. Our code is available at http://pastebin.com/GmHxL0e8.

1 INTRODUCTION

In GANs (Goodfellow et al., 2014), two machines learn about some data $x \sim P$ by competing against each other. On the one hand, the *generator* $g : \mathbb{R}^d \to \mathbb{R}^D$ competes to transform *noise vectors* $z \sim Q$ into *fake samples* $g(z)$ that resemble *real samples* $x \sim P$. On the other hand, the *discriminator* competes to distinguish between real samples $x \sim P$ and fake samples $g(z)$ synthesized by the generator. Mathematically, GAN training approximates the optimization problem

$$\min_g \max_d \mathop{\mathbb{E}}_{x \sim P} \log(d(x)) + \mathop{\mathbb{E}}_{z \sim Q} \log(1 - d(g(z))).$$

Although GANs show a remarkable performance in a variety of generative modeling tasks, their training is notoriously difficult (Goodfellow, 2017). One common failure mode of GANs is known as *mode collapsing* (Metz et al., 2017; Che et al., 2017; Tolstikhin et al., 2017). This happens when the generator is only able to produce fake samples resembling a small subset of the modes of P.

To illustrate mode collapsing, consider training a single GAN to reproduce the mixture of Gaussians data drawn in blue in Figure 1a. In this example, the probability density between modes is near zero. Therefore, a generator g able to reproduce such multimodal data, given unimodal noise vectors $z \sim Q$, must be a discontinuous function. However, in most situations the generator g is a deep neural network, which is a differentiable —and therefore continuous— function. In these situations, multimodality is pursued by saturating sigmoidal activation functions, which approximate discontinuous step functions at the cost of larger weights and vanishing gradients.

Even in the case where these do not cause issues in training, the resulting fake samples $g(z)$ are not truly multimodal. This can be easily revealed by interpolating between two noise vectors mapped to different modes, and observing a continuous deformation between the two samples. However, samples living on a "bridge" between two modes make little sense with regards to the real distribution. In a nutshell, continuous generators are constrained to place mass *between modes* in an attempt to mimick multimodality.

Another major cause for mode collapsing are bad initializations: when the fake samples are initially near a mode of the real samples, the generator often collapses to that mode, ignoring all others.

In the following, we defend two simple strategies to stabilize the training of GANs on multimodal data: *multimodal noise distributions* and *covering initializations*. First, *multimodal noise distributions* are a natural way to provide continuous generators with multimodal behaviour. Since the noise distribution can be understood as our *prior distribution*, we believe it should carry our domain

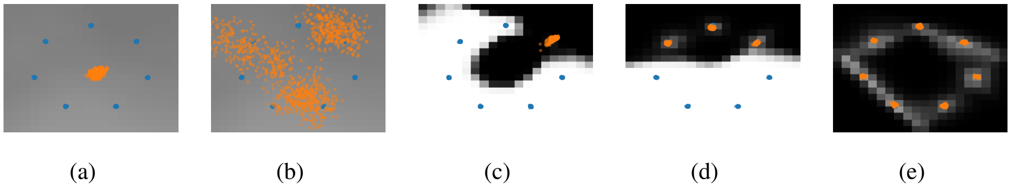

Figure 1: Experiments on the mixture of Gaussians dataset, showing (a) normal initialization, (b) covering initialization, (c) converged GAN with covering initialization and unimodal noise, (d) converged GAN with normal initialzation and multimodal noise, and (e) converged GAN with covering initialization and multimodal noise. Real data is depicted in blue, fake data is depicted in orange, and the discriminator scores are depicted in a heatmap ranging from black (fake) to white (real).

knowledge (including multimodality) about the real samples. Second, *covering initialization* is a technique to pre-train the generator to match the empirical mean and covariance of the fake and real data. This promotes overlap between the fake and real distributions, which protects our GANs from converging to the degenerate solutions described above.

2 Multimodal Noise

Given a noise distribution Q, we construct multimodal noise vectors $z \in \mathbb{R}^{d+c}$ by

i) sampling $a \sim Q$, where $a \in \mathbb{R}^d$,
ii) constructing a one-hot vector $b \in \mathbb{R}^c$, where $b_i = 1$ for $i \sim$ Multinomial $[1, c]$,
iii) constructing $z =$ Concatenate(a, b).

For instance, if $Q = U[0, 1]$ and $d = c = 3$, we may sample $z = [0.247, 0.877, 0.324, 0, 1, 0]$.

Choosing the number of categorical noise dimensions c is an analogous problem to choosing the number of continuous noise dimensions d. As such, c can be either cross-validated (Lopez-Paz & Oquab, 2017) or estimated by means of a clustering algorithm.

3 Covering Initialization

Before training, we pre-train the generator g to match the empirical mean and covariance of large batches of fake and real data. This *covering initialization* promotes the intersection between the supports of the real data distribution and the synthesized data distribution, in the same spirit as *instance noise* (Sønderby et al., 2017; Arjovsky & Bottou, 2017). A more sophisticated alternative would be to pre-train the generator g to match the MMD statistic (Dziugaite et al., 2015).

4 Numerical simulations

We train a GAN to reproduce the mixture of Gaussians benchmark introduced in (Metz et al., 2017). The generator and discriminator are two-layer neural networks with 128 hidden ReLU neurons. Both the generator and the discriminator are trained using the Adam optimizer with default parameters (Kingma & Ba, 2015). As recommended by Arjovsky et al. (2017); Chintala et al. (2016), we train the discriminator close to optimality, by taking 10 Adam steps per iteration.

Figure 1 summarizes our results. First, we show the real samples (blue) together with the fake samples (orange) synthesized by the generator, when initialized normally (Figure 1a) or using the *covering initalization* (Figure 1b) described in Section 3. Second, we also show the fake samples (orange) synthesized by the generator after the GAN training has converged, when using a covering initialization and unimodal noise (Figure 1c), a regular initialization and multimodal noise (Figure 1d) or, as proposed, a covering initialization and a multimodal noise (Figure 1e). All figures illustrate the discriminator decision surface as a heatmap, ranging from black (fake) to white (real).

The joint use of covering initialization and multimodal noise solves to perfection the task at hand. Furthermore, our model places no mass between the modes, and is robust with respect to all hyperparameters and random initializations, as long as the discriminator is trained close to optimality.

REFERENCES

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *ICLR*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv*, 2017.

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *ICLR*, 2017.

Soumith Chintala, Emily Denton, Martin Arjovsky, and Michael Mathieu. How to train a GAN? Tips and tricks to make GANs work. `https://github.com/soumith/ganhacks`, 2016.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv*, 2015.

Ian Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 2014.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

David Lopez-Paz and Maxime Oquab. Revisiting Classifier Two-Sample Tests. *ICLR*, 2017.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *ICLR*, 2017.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *ICLR*, 2017.

Ilya Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. AdaGAN: Boosting Generative Models. *arXiv*, 2017.