ANNEALED GENERATIVE ADVERSARIAL NETWORKS

Arash Mehrjou
Department of Empirical Inference
Max Planck Institute for Intelligent Systems
arash.mehrjou@tuebingen.mpg.de

Saeed Saremi
Redwood Center for Theoretical Neuroscience
University of California, Berkeley
saeed@berkeley.edu

ABSTRACT

Generative Adversarial Networks (GANs) have recently emerged as powerful generative models. GANs are trained by an adversarial process between a generative network and a discriminative network. It is theoretically guaranteed that, in the nonparametric regime, by arriving at the *unique* saddle point of a minimax objective function, the generative network generates samples from the data distribution. However, in practice, getting close to this saddle point has proven to be difficult, resulting in the ubiquitous problem of "mode collapse". The root of the problems in training GANs lies on the unbalanced nature of the game being played. Here, we propose to level the playing field and make the minimax game balanced by "heating" the data distribution. The empirical distribution is frozen at temperature zero; GANs are instead *initialized* at infinite temperature, where learning is stable. By annealing the heated data distribution, we initialized the network at each temperature with the learnt parameters of the previous higher temperature. We posited a conjecture that learning under continuous annealing in the nonparametric regime is stable, and proposed an algorithm in corollary. In our experiments, the annealed GAN algorithm, dubbed β-GAN, trained with *unmodified* objective function was stable and did not suffer from mode collapse.

1 INTRODUCTION

One of the most fundamental problems in machine learning is the unsupervised learning of high-dimensional data. A class of problems in unsupervised learning is density estimation, where it is assumed that there exist a class of probabilistic models underlying observed data x and the goal of learning is to infer the "right" model(s). The generative adversarial network proposed in Goodfellow et al. (2014) is an elegant framework, which transforms the problem of density estimation to an adversarial process in a minimax game between a generative network $\mathbb{G}$ and a discriminative network $\mathbb{D}$. However, despite their simplicity, GANs have proven to be difficult to train. There are different schools in diagnosing and addressing the problems with training GANs, that have resulted in a variety of algorithms (Denton et al., 2015; Radford et al., 2015; Zhao et al., 2016; Chen et al., 2016; Metz et al., 2016; Tolstikhin et al., 2017; Arjovsky & Bottou, 2017; Arjovsky et al., 2017). Perhaps, the biggest challenge is that the data in the world are highly structured, lying on a very low-dimensional manifold of their ambient space (Goodfellow et al., 2016). At the beginning of training the generative network $\mathbb{G}$ is far off from this low-dimensional manifold and the generated samples get easily rejected by the discriminative network $\mathbb{D}$, causing little room to improve $\mathbb{G}$. The other challenging issue is that GANs optimal point is a *saddle point*. We have good understanding and a variety of optimization methods to find local minima/maxima of objective functions, but *minimax* optimization in high-dimensional spaces have proven to be challenging. Because of these two obstacles, i.e. the nature of high-dimensional data and the nature of the optimization, GANs suffer from stability issues and the ubiquitous problem of *mode collapse*, where the generator completely ignores parts of the low-dimensional data manifold. In this work, we address these two issues at the same time by lifting the minimax game after defining an effective *temperature*[1] for the data distribution in an annealing framework.

Annealing has a rich history in statistical mechanics, with applications to combinatorial optimization and Markov chain Monte Carlo (Kirkpatrick et al., 1983; Marinari & Parisi, 1992; Neal, 2001). In

[1] We will drop the word *effective* throughout the paper.

addition, there are deep connections between the framework here, the nonequilibrium framework for unsupervised learning (Sohl-Dickstein et al., 2015), and recent works on "visible hierarchies" in natural images (Saremi & Sejnowski, 2013; 2015; 2016). We will elaborate on these connections in the extended version of this paper.

2 ANNEALED GAN

In this section, we define the inverse temperature β for the data distribution and provide an algorithm that is built on annealing the data distribution. The convergence of the algorithm is based on a conjecture with stability guarantees in the nonparametric regime and the continuous annealing limit. We will present the conjecture in detail in the extended version of the paper. We named the algorithm β-GAN for the leading role the inverse temperature β plays.

Let us assume networks $\mathbb{G}$ and $\mathbb{D}$ have enough capacity, parameterized by deep neural networks $G(z;\theta_G)$ and $D(x;\theta_D)$. The heated data distribution at inverse temperature β is given by:

$$p_{\text{data}}(x;\beta) = \frac{1}{N}\left(\frac{\beta}{2\pi}\right)^{D/2}\sum_i \exp\left(-\frac{\beta(x-x^{(i)})^2}{2}\right),$$

where the empirical distribution for N observations $\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\} \in \mathcal{R}^D$ is recovered in the limit $\beta \to \infty$. The minimax optimization task at each β is:

$$\begin{aligned}\theta_G^*(\beta) &= \arg\min_{\theta_G}\max_{\theta_D} f(\theta_D,\theta_G;\beta),\\ f(\theta_D,\theta_G;\beta) &= \mathop{\mathbb{E}}_{x\sim p_{\text{data}}(x;\beta)} \log\left(D(x;\theta_D)\right) + \mathop{\mathbb{E}}_{z\sim p(z)} \log(1-D(G(z;\theta_G);\theta_D)).\end{aligned}$$

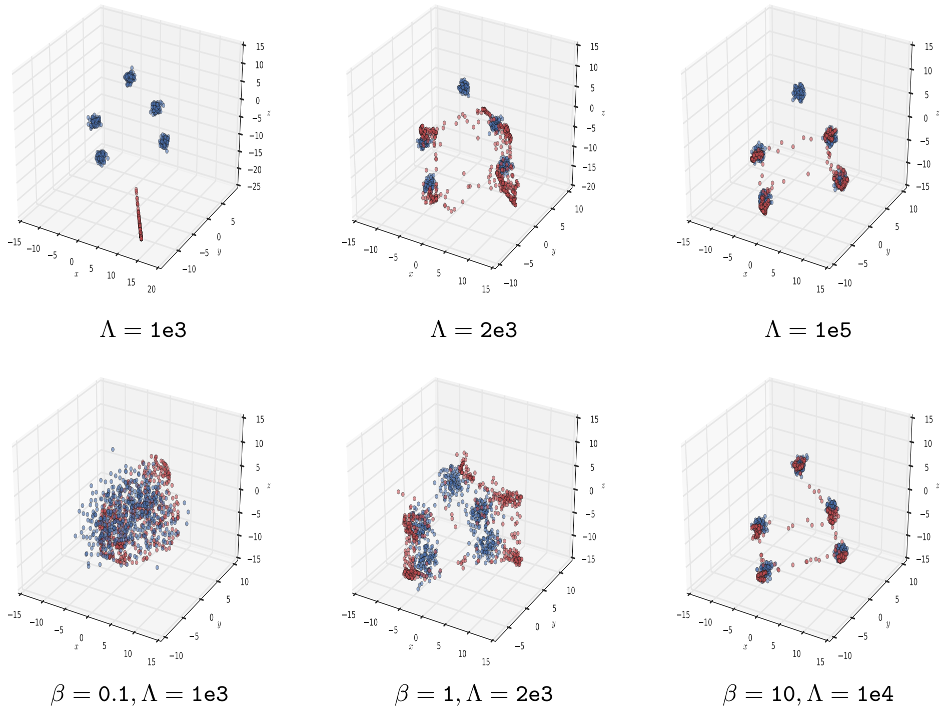
Note that the optimal parameters θ_G^* and θ_D^* are both functions of β. With this setup, annealed GAN algorithm is given below:

Algorithm 1 Minibatch stochastic gradient descent training of annealed generative adversarial networks.

```
• Receive $\beta_0$, $\beta_\infty$, and $K$, which correspond to inverse infinite temperature, inverse zero temperature, and the number of cooling steps respectively.
• Compute $\alpha > 1$ as the geometric cooling factor:
      $$\alpha = \left(\frac{\beta_\infty}{\beta_0}\right)^{\frac{1}{K}} = \left(\frac{\text{Temperature}_\infty}{\text{Temperature}_0}\right)^{\frac{1}{K}}.$$
• Initialize $\beta$: $\beta \leftarrow \beta_0$
for number of cooling steps $(K)$ do
    for number of training steps $(n)$ do
        • Sample minibatch of $m$ noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p(z)$.
        • Sample minibatch of $m$ examples $\{x^{(1)}, \ldots, x^{(m)}\}$ from data generating distribution $p_{\text{data}}(x;\beta)$.
        • Update the discriminator by ascending its stochastic gradient:
              $$\nabla_{\theta_{d,\beta}} \frac{1}{m}\sum_{i=1}^{m}\left[\log D\left(x^{(i)};\theta_{d,\beta}\right) + \log\left(1 - D\left(G\left(z^{(i)};\theta_{g,\beta}\right);\theta_{d,\beta}\right)\right)\right].$$
        • Sample minibatch of $m$ noise samples $\{z^{(1)}, \ldots, z^{(m)}\}$ from noise prior $p(z)$.
        • Update the generator by descending its stochastic gradient:
              $$\nabla_{\theta_{g,\beta}} \frac{1}{m}\sum_{i=1}^{m}\log\left(1 - D\left(G\left(z^{(i)};\theta_{g,\beta}\right);\theta_{d,\beta}\right)\right).$$
    end for
    • Increase $\beta$ geometrically: $\beta \leftarrow \beta * \alpha$
end for
```

3 Experiments

To check the stability of β-GAN, we ran experiments on mixtures of 1D, 2D, and 3D Gaussians. The results reported here for vanilla GAN was the best among many runs; in most of them it only captured one mode or failed to capture any mode. However, β-GAN consistently produced similar results. Vanilla GAN requires the modification of the generator loss to $\log(D(G(z;\theta_G)))$ to avoid saturation of discriminator (Goodfellow et al., 2014), while in β-GAN we did not make any modification, staying with the generator loss $\log(1 - D(G(z;\theta_G);\theta_D))$. In the experiments, the total number of training iterations in β-GAN was the same as vanilla GAN, but distributed over many intermediate temperatures, thus curbing the computational cost. Despite its low computational cost, the algorithm was stable during training and did not suffer from mode collapse (see Fig. 1). We should also emphasize that other GAN architectures can be easily augmented with β-GAN as the outer loop. In this paper, we chose the original generative adversarial network of Goodfellow et al. (2014) as the inner loop (see Algorithm 1).



$\Lambda = 1e3$

$\Lambda = 2e3$

$\Lambda = 1e5$

$\beta = 0.1, \Lambda = 1e3$

$\beta = 1, \Lambda = 2e3$

$\beta = 10, \Lambda = 1e4$

Figure 1: Top row: performance of vanilla GAN on a mixture of 4 Gaussian components in 3 dimensions. Bottom row: performance of β-GAN on the same dataset. Blue/red dots are real/generated data. To compare the computational cost, we report Λ, which is the total number of gradient evaluations from the start.

4 Discussions

This work fits in a larger picture of approaches that have recently emerged in stabilizing the training of GANs by either "better" distance measures (Nowozin et al., 2016; Arjovsky et al., 2017) or adding noise during training, including "label noise" (Salimans et al., 2016) and "instance noise" (Kaae Sønderby et al., 2016). Instead of adding noise, we defined a more general concept called *inverse temperature*, which we annealed geometrically (Neal, 2001). We thus avoided adding any extra noise to the generator or the discriminator. Geometric annealing also gave us the flexibility to quickly zoom in on the data distribution at small temperatures. Our approach is especially appealing because we start the training at *infinite* temperature (high-entropy/uniform distribution), where the data distribution *fills* the ambient space, guaranteeing that the generator output always remains a subspace of the heated data manifold. This work opens up a new space for exploring GAN stability both theoretically and for rich datasets occupying arbitrarily complex manifolds.

ACKNOWLEDGMENTS

SS acknowledges the support by The Canadian Institute for Advanced Research.

REFERENCES

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

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*, pp. 2172–2180, 2016.

Emily L Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 1486–1494, 2015.

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*, pp. 2672–2680, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19(6):451, 1992.

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

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

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

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*, pp. 2226–2234, 2016.

Saeed Saremi and Terrence J Sejnowski. Hierarchical model of natural images and the origin of scale invariance. *Proceedings of the National Academy of Sciences*, 110(8):3071–3076, 2013.

Saeed Saremi and Terrence J Sejnowski. The Wilson machine for image modeling. *arXiv preprint arXiv:1510.07740*, 2015.

Saeed Saremi and Terrence J Sejnowski. Correlated percolation, fractal structures, and scale-invariant distribution of clusters in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):1016–1020, 2016.

Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

Ilya Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. Adagan: Boosting generative models. *arXiv preprint arXiv:1701.02386*, 2017.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.