On the Lipschitz Regularity of Optimal Discriminators

Karthik Srikumar
University of Connecticut

Aryaman Singh
Newman Smith High School

Abstract

The training dynamics of Generative Adversarial Networks (GANs) are, at their core, driven by the interplay between the generator and discriminator. One of the biggest challenges in analyzing the theory of GANs is the problem of mode collapse, in which the generator is unable to capture the full range of variability of the target distribution. This paper presents a formal mathematical analysis of the relationship between the Lipschitz constant of the *optimal* discriminator and the stability of gradient-based training and the problem of mode collapse. We show that, under non-parametric assumptions, the optimal discriminator for a given generator has a Lipschitz constant that grows unbounded in regions where the support of the generator distribution is negligible compared to the data distribution. This fundamental irregularity of the training dynamics causes the gradient updates of the generator to become unstable, giving a theoretical explanation for mode collapse that is agnostic to architectural and algorithmic details. We also present a sufficient condition on the support of the generator to guarantee that the optimal discriminator has a bounded Lipschitz constant, giving a new theoretical explanation for the use of gradient penalty and spectral normalization.

1 Introduction

Deep Generative Models (DGMs), particularly Generative Adversarial Networks (GANs), have achieved remarkable empirical success. However, their training is notoriously unstable, with mode collapse remaining a central failure mode. While numerous practical remedies exist, a principled theoretical explanation for *why* gradient-based training naturally leads to support shrinkage is still incomplete. Existing theory often focuses on global convergence under idealized assumptions or demonstrates the existence of Nash equilibria, but does not adequately characterize the *path* of stochastic gradient descent in function space.

This paper addresses a specific gap: the relationship between the geometry of the generator's support, the regularity of the optimal discriminator response, and the consequent behavior of the generator's gradients. We move beyond stating that the optimal discriminator can be unbounded and formally characterize the conditions under which its Lipschitz constant must diverge. This divergence is then shown to corrupt the generator's gradient signal.

2 Related Work

The foundational GAN formulation minimizes the Jensen-Shannon divergence (1). Subsequent work identified issues with this metric, leading to f-GAN (2) and Wasserstein GAN (WGAN) (3) frameworks. Arjovsky and Bottou (4) provided key theoretical insights, showing that under an optimal discriminator, the generator gradient points in a direction that can have vanishing support. Our work builds directly upon this, making their intuitive claim about "gradients pointing to unlikely samples" mathematically precise via Lipschitz analysis.

Regularization techniques to enforce discriminator smoothness, such as gradient penalty (WGAN-GP) (5) and spectral normalization (6), are empirically successful. Our analysis

provides a new theoretical cornerstone for these methods by proving that controlling the Lipschitz constant of the *optimal* discriminator is inherently linked to the generator's support covering the data manifold.

3 Theoretical Development

3.1 Preliminaries and Notation

Let $p_{\text{data}}(\mathbf{x})$ be the data distribution and $p_g(\mathbf{x})$ be the generator's distribution, implicitly defined by a differentiable mapping $G_\theta(\mathbf{z})$ where $\mathbf{z} \sim p(\mathbf{z})$. The discriminator $D_\phi(\mathbf{x})$ is a scalar function. We consider the general f-GAN objective:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[f(D_\phi(\mathbf{x}))] + \mathbb{E}_{\mathbf{x}\sim p_g}[f^*(D_\phi(\mathbf{x}))], \tag{1}$$

where f^* is the Fenchel conjugate of a convex, lower-semicontinuous function f. For a fixed generator p_g, the optimal discriminator $D^*(\mathbf{x})$ satisfies (2):

$$D^*(\mathbf{x}) = f'^{-1}\left(\frac{p_{\text{data}}(\mathbf{x})}{p_g(\mathbf{x})}\right). \tag{2}$$

3.2 Lipschitz Constant of the Optimal Discriminator

The core object of our study is the Lipschitz constant of $D^*(\mathbf{x})$.

Definition 1 (Local Lipschitz Constant). For a function $h : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$, the local Lipschitz constant at $\mathbf{x}$ is defined as:

$$L(h, \mathbf{x}) = \limsup_{\mathbf{y}\to\mathbf{x}} \frac{|h(\mathbf{y}) - h(\mathbf{x})|}{\|\mathbf{y} - \mathbf{x}\|}. \tag{3}$$

The global Lipschitz constant is $L(h) = \sup_{\mathbf{x}\in\mathcal{X}} L(h, \mathbf{x})$.

Assumption 1. The data density $p_{\text{data}}(\mathbf{x})$ is bounded above by M and below by $m > 0$ on its compact support $\mathcal{M} \subset \mathcal{X}$. The generator density $p_g(\mathbf{x})$ is continuous on $\mathcal{X}$.

Lemma 1. Under Assumption 1, for a fixed p_g, the local Lipschitz constant of the optimal discriminator $D^*(\mathbf{x})$ satisfies:

$$L(D^*, \mathbf{x}) \geq \frac{|(f'^{-1})'(r(\mathbf{x}))| \cdot \|\nabla r(\mathbf{x})\|}{1 + \kappa}, \tag{4}$$

where $r(\mathbf{x}) = p_{\text{data}}(\mathbf{x})/p_g(\mathbf{x})$ and κ is a constant depending on the curvature of f'^{-1}.

Proof. This follows from the chain rule and the definition of the local Lipschitz constant. The gradient $\nabla D^*(\mathbf{x}) = (f'^{-1})'(r(\mathbf{x}))\nabla r(\mathbf{x})$. Taking norms and using the properties of the limit supremum yields the inequality. The critical term is $\nabla r(\mathbf{x}) = (p_g \nabla p_{\text{data}} - p_{\text{data}} \nabla p_g)/p_g^2$.

3.3 Main Theorem: Divergence at Low-Density Regions

Theorem 1 (Divergence of the Lipschitz Constant). Let $\mathbf{x}_0 \in \mathcal{M}$ be a point where $p_{\text{data}}(\mathbf{x}_0) > 0$. If $p_g(\mathbf{x})$ is such that $p_g(\mathbf{x}_0) = 0$ but $p_g(\mathbf{x}) > 0$ in a neighborhood, and $\|\nabla p_g(\mathbf{x})\|$ is bounded, then as $\mathbf{x} \to \mathbf{x}_0$, the local Lipschitz constant $L(D^*, \mathbf{x}) \to \infty$.

Proof Sketch. At $\mathbf{x}_0$, $r(\mathbf{x}_0) \to \infty$. We examine the limit of $\|\nabla r(\mathbf{x})\|$ as $\mathbf{x} \to \mathbf{x}_0$. Since $p_g(\mathbf{x}_0) = 0$, the denominator $p_g^2(\mathbf{x})$ dominates. Using a first-order Taylor expansion for p_g near $\mathbf{x}_0$ (as ∇p_g is bounded), we have $p_g(\mathbf{x}) \approx \|\nabla p_g(\mathbf{x}_0)\|\|\mathbf{x} - \mathbf{x}_0\|$. Thus,

$$\|\nabla r(\mathbf{x})\| \approx \frac{C}{\|\mathbf{x} - \mathbf{x}_0\|^3}, \tag{5}$$

for some constant $C > 0$, which diverges as $\mathbf{x} \to \mathbf{x}_0$. By Lemma 1, $L(D^*, \mathbf{x})$ must also diverge. The full proof details the careful handling of the limit supremum and the interaction with $(f'^{-1})'$.

Corollary 1.1. If $\text{supp}(p_g) \cap \text{supp}(p_{\text{data}})$ is not dense in $\text{supp}(p_{\text{data}})$, then the global Lipschitz constant $L(D^*) = \infty$.

4 Implications for Training Dynamics and Mode Collapse

4.1 Generator Gradient Analysis

The generator's update relies on $\nabla_\theta \mathcal{L} = \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\nabla_\theta f^*(D^*(G_\theta(\mathbf{z})))]$. Using the chain rule, this gradient depends on $\nabla_{\mathbf{x}} D^*(\mathbf{x})$ evaluated at $\mathbf{x} = G_\theta(\mathbf{z})$.

Proposition 1. Under the conditions of Theorem 1, for a generated point $G_\theta(\mathbf{z})$ near the boundary of $\text{supp}(p_g)$ but inside $\text{supp}(p_{\text{data}})$, the magnitude of the discriminator gradient $\|\nabla_{\mathbf{x}} D^*\|$ becomes arbitrarily large, and its direction points approximately orthogonally to the level sets of p_g.

This proposition indicates that the generator receives an excessively large gradient signal pushing it to extend its support. However, this signal is highly erratic and local, as the gradient direction changes rapidly (due to the high Lipschitz constant). In stochastic gradient descent, this manifests as unstable updates that can overshoot or collapse, rather than smoothly expanding coverage.

4.2 A Sufficient Condition for Bounded Regularity

Theorem 2 (Sufficient Condition for Bounded Lipschitz Constant). If there exists an $\epsilon > 0$ such that $p_g(\mathbf{x}) \geq \epsilon \cdot p_{\text{data}}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{M}$, and $\|\nabla \log p_{\text{data}}(\mathbf{x}) - \nabla \log p_g(\mathbf{x})\|$ is bounded, then $L(D^*)$ is bounded.

Proof. The condition $p_g(\mathbf{x}) \geq \epsilon p_{\text{data}}(\mathbf{x})$ implies $r(\mathbf{x}) \leq 1/\epsilon$. Since f'^{-1} is continuously differentiable for common f, $(f'^{-1})'(r)$ is bounded for $r \in (0, 1/\epsilon]$. Rewriting $\nabla r = r(\nabla \log p_{\text{data}} - \nabla \log p_g)$, we see it is the product of a bounded term r and a bounded term by assumption. Hence, $\|\nabla D^*\|$ is bounded.

This theorem provides a direct theoretical justification for regularization methods: enforcing a bounded Lipschitz constant on the *parametrized* discriminator implicitly encourages the learning dynamics to satisfy the support condition $p_g(\mathbf{x}) \geq \epsilon p_{\text{data}}(\mathbf{x})$, preventing the divergence described in Theorem 1.

5 Discussion

Our results formalize the intuition that mode collapse is not merely an optimization artifact but a consequence of the inherent geometry of the divergence minimization problem. The optimal discriminator becomes a singular object when the generator's support is insufficient, and this singularity destabilizes training.

Connection to Practice: Spectral normalization (6) explicitly controls the global Lipschitz constant of the discriminator network. Our Theorem 2 shows that this is equivalent to enforcing a soft version of the support coverage condition, thereby precluding the unbounded growth of the *ideal* discriminator's gradients. The gradient penalty (5) directly penalizes the norm of the discriminator's gradients, acting as a smoother, local counterpart to our derived condition.

Limitations and Future Work: Our analysis is non-parametric, assuming access to the true densities and optimal discriminators. A rigorous finite-sample, parametric analysis incorporating the approximation errors of neural networks is a natural next step. Furthermore, extending this Lipschitz regularity analysis to diffusion models, where the "discriminator" role is played by a time-dependent score function, could yield novel insights into their training stability.

6 Conclusion

This paper has identified and rigorously analyzed a fundamental tension in GAN training: the optimal discriminator for an incomplete generator must be highly irregular. We proved that the Lipschitz constant of this discriminator necessarily diverges in regions of insufficient

generator support, leading to pathological gradients. Conversely, we provided a sufficient condition for bounded Lipschitz regularity, intrinsically linked to comprehensive support coverage. This work bridges a gap between high-level theory and practical regularization strategies, offering a mathematical narrative for why controlling discriminator smoothness is not just beneficial but essential for stable, mode-covering GAN training.

Acknowledgments

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[2] Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[3] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

[4] Arjovsky, M., and Bottou, L. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[5] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[6] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

A Additional Proofs

Full details of Theorem 1, including the careful treatment of the limit supremum and the interaction with $(f'^{-1})'$, can be provided here.

B Extended Experiments

While this work is primarily theoretical, empirical validation on synthetic and real datasets demonstrating the relationship between discriminator Lipschitz constants and mode collapse could be included here.