# Zero-Infinity GAN: Stable Dynamics and Implicit Bias of Extragradient

**Kyungjae Lee** NIYANDA01@KAIST.AC.KR
**Donghwan Kim** DONGHWANKIM@KAIST.AC.KR
*KAIST, Daejeon, Republic of Korea*

## Abstract

In supervised learning, gradient descent achieves near-zero empirical risk, while favoring solutions that generalize well—a phenomenon attributed to the *implicit bias* of gradient methods. In stark contrast, in generative models such as generative adversarial networks (GANs), gradient methods typically fail to achieve zero empirical risk, and thus implicit bias is left both empirically elusive and theoretically unexplored. We bridge this gap by developing new perspectives on the loss landscape of GANs together with the gradient dynamics and implicit bias of extragradient. First, regarding the loss, we challenge the prevailing preference for the Wasserstein distance, and instead propose the *zero-infinity* distance—a metric that equals zero when two distributions match exactly and infinity otherwise—as more compatible with gradient-based minimax optimization. On the gradient dynamics side, we prove for the first time in GANs that certain stationary points are *strict non-minimax* points, the minimax analogue of strict saddles in minimization. This enables the two-timescale extragradient method to effectively escape such non-optimal points—similar to gradient descent escaping strict saddles—while being stable at global solutions, in contrast to other existing methods. Lastly, regarding the implicit bias, we show that extragradient favors the minimum-norm generator solution when starting from zero and training only the last layer of neural network.

**Keywords:** Zero-Infinity distance, Two-timescale extragradient, Implicit bias of Extragradient

## 1. Introduction

In supervised learning, the implicit bias of gradient descent—favoring solutions that generalize well—was first observed empirically to play a crucial role in its superior generalization performance [26, 31]. Notably, in overparameterized settings, neural networks trained to near-zero empirical risk without explicit regularization often yielded models that generalized remarkably well among many interpolating solutions—a phenomenon later termed benign overfitting [5]. A theoretical understanding of this bias, especially through the lens of optimization dynamics, subsequently emerged [15, 28, 31], revealing how gradient descent drives models toward particular types of solutions. For instance, in least squares regression and logistic classification, gradient descent has been shown to converge to the minimum-norm and max-margin solutions, respectively [15, 28, 31]. These implicit biases were later shown to underpin the theoretical foundations of benign overfitting [5, 8, 29]. The *minimum-norm* solution is particularly favorable as it is possibly the simplest function consistent with the training data, often leading to strong generalization.

Surprisingly, these insights do not extend to generative models, where overfitting is widely believed to degrade generalization [21, 25, 30]. Moreover, diffusion models [13, 27], the current state-of-the-art, typically parameterize a score function or noise process via neural networks, yet it remains unclear what kinds of solutions should be favored, and whether any implicit bias contributes to generalization. In contrast, generative adversarial networks (GANs) directly parameterize the generator by neural network, making it more transparent which solution types promote generalization. This paper thus studies the implicit bias of gradient methods in GANs, aiming to unify our understanding of generalization in machine learning and to advance it in generative models.

Despite this aim, progress is hindered by two prevailing beliefs: that overfitting harms generalization in generative models, and that GAN training is inherently unstable. The former has recently been challenged by results showing that training solely from presampled latent variables—rather than accessing the full latent distribution—can lead to benign overfitting [7]. The latter remains largely unresolved[1], so our work addresses both the loss landscape and gradient dynamics issues, paving the way toward an implicit bias framework in GANs for the first time.

First, we propose the *zero-infinity* distance for GANs, which is zero when two distributions match exactly, and infinity otherwise. Despite its extreme form, it yields the same minimax loss as the Wasserstein GAN [3], but without requiring the Lipschitz constraint. Consequently, the zero-infinity GAN avoids both the limitations of Lipschitz enforcement and the vanishing gradient issue of the original GAN loss [2, 11, 12]. While the Zero-Infinity GAN loss is favorable, it remains nonconvex–nonconcave. As a next step, we show for the first time in GANs that certain non-optimal stationary points are *strict non-minimax* points [6], the minimax analogue of strict saddles in nonconvex minimization [10]. We demonstrate that the two-timescale extragradient (EG) method [6, 18] escape such strictly non-optimal points—similar to gradient descent escaping strict saddles [20]—while remaining stable at global solutions, in contrast to other existing gradient methods. Finally, we show that if any gradient method converges to a global solution, it will favor the *minimum-norm* solution when starting from zero and training only the last layer of neural network. Our three key contributions are summarized below:

- Section 4 introduces the zero-infinity distance as an alternative that shapes the GAN loss to better align with gradient-based training.
- Section 5 shows that certain stationary points are strict non-minimax points, which the two-timescale extragradient avoids while remaining stable at global solutions.
- Section 6 presents that extragradient favors the minimum-norm solution among many global solutions when starting from zero and training only the last layer of neural network.

## 2. Related works

**Statistical distances and their minimax formulations for GANs.** Since direct minimization of statistical distances is typically impractical, GANs optimize the corresponding dual minimax formulation. This has nevertheless been viewed through a minimization lens, partly explaining the popularity of the Wasserstein distance, which yields meaningful non-zero gradients even when minimized directly [3]. Despite this, the minimax optimization of the Wasserstein GAN is infeasible due to the Lipschitz constraint, which is only approximated in practice through regularization, *e.g.,* the gradient penalty [12]. More broadly, existing GANs either rely on such regularization or suffer from vanishing gradients [2, 3]. This motivates our search for a new statistical distance that induces a loss landscape well-suited for gradient-based minimax training.

**Strict non-minimax points and two-timescale EG.** The success of gradient descent in nonconvex minimization relies on two key results: the *strict saddle* property—all (locally) non-optimal

[1] One may argue that popular GAN architectures such as StyleGAN [17] have largely resolved instability in practice. However, as pointed out by [14], they rely on numerous heuristics whose theoretical foundations remain limited.

stationary points are strict saddles—and the result that gradient descent almost surely avoids strict saddle points [10, 20]. Extending them to the minimax and GAN settings, however, has remained elusive. Recently, Chae et al. [6] introduced the notion of *strict non-minimax* points —the minimax analogue of strict saddles in minimization—which can be escaped by two-timescale EG [6, 18]. We show, for the first time, that the GAN loss can also exhibit the analogous *strict non-minimax* property in part, thereby opening the door to more successful minimax training.

## 3. Problem settings

**The original and Wasserstein GANs.** The goal of modern generative models is to train a generator $G : \mathcal{Z} \to \mathcal{X}$ that maps a latent distribution $p_z$ over latent space $\mathcal{Z}$ to a target distribution $p_{\text{data}}$ over data space $\mathcal{X}$. However, directly minimizing a distance between $p_{\text{data}}$ and generated data distribution $p_g := G_\sharp p_z$, the pushforward of $p_z$ by $G$, is challenging. So, the original GAN [11] reformulated minimizing the Jensen-Shannon (JS) divergence as the following minimax problem by introducing a discriminator $D$:[2]

$$\min_G \max_D -\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[l(D(\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{z}\sim p_z}[l(-D(G(\boldsymbol{z})))], \tag{1}$$

where $l(t) = \log(1 + \exp(-t))$ is the logistic loss. Since the original GAN suffers from vanishing gradients [2, 3] due to the logistic loss, the following Wasserstein GAN [3] was introduced:

$$\min_G \max_{D:\|D\|_L \le 1} \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[D(\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{z}\sim p_z}[D(G(\boldsymbol{z}))]. \tag{2}$$

This corresponds to (1) with $l(t) = -t$ but imposes the additional 1-Lipschitz constraint on $D$. Because enforcing the Lipschitz constraint is infeasible in practice, regularization techniques such as the gradient penalty [12] are commonly used. While these approaches improved training stability, the fundamental instability of GAN optimization has remained unresolved.

**Finite sample GAN problems.** In practice, the $p_{\text{data}}$ is unknown, and only a finite set of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ is available. This leads to the empirical problem $\min_G \max_{D\in\mathcal{D}} -\frac{1}{n}\sum_{i=1}^n l(D(\boldsymbol{x}_i)) - \mathbb{E}_{\boldsymbol{z}\sim p_z}[l(-D(G(\boldsymbol{z})))]$, where $\mathcal{D}$ is a constrained set of $D$. However, when this is solved exactly, the generator would merely memorize the training samples $\{\boldsymbol{x}_i\}_{i=1}^n$ [25], rather than generalize. Instead, some consider the following alternative [4]:

$$\min_G \max_{D\in\mathcal{D}} -\frac{1}{n}\sum_{i=1}^n [l(D(\boldsymbol{x}_i)) + l(-D(G(\boldsymbol{z}_i)))], \tag{3}$$

where latent variables $\{\boldsymbol{z}_i\}_{i=1}^n$ are presampled. Recently, Chae et al. [7] showed that solving (3) can exhibit strong generalization, when combined with the implicit bias of gradient methods. Nonetheless, developing a practical training method for (3) remains open. In this work, we aim to improve the loss landscape of (3) and gradient dynamics, and eventually characterize their implicit bias.

## 4. Zero-Infinity GAN: Toward a better loss landscape in GANs

**Motivating example: Dirac GAN.** Mescheder et al. [22] considered a simple yet prototypical example for learning data distributions: the data distribution is a Dirac delta at 0, *i.e.,* $p_{\text{data}} = \delta_0$,

[2] The formulation (1) adopted by [24] is equivalent to the original GAN formulation in [11].

and the latent distribution is a Dirac delta at $1$, *i.e.,* $p_z = \delta_1$. The Dirac GAN is then trained using a linear generator $G(z) = gz$ and a linear discriminator $D(x) = wx$, as $\min_{g\in\mathbb{R}} \max_{w\in\mathcal{W}\subseteq\mathbb{R}} -l(0) - l(-wg)$. This reduces to the original GAN when $l(t)$ is the logistic loss and $\mathcal{W} = \mathbb{R}$, and the Wasserstein GAN when $l(t) = -t$ with a Lipschitz discriminator constraint, *i.e.,* $\mathcal{W} = \{w \in \mathbb{R} : |w| \le 1\}$. In both cases, the unique global solution is $(g_*, w_*) = (0, 0)$.

Unlike gradient descent ascent (GDA) diverging for both cases, the extragradient (EG) [19]—a two-step variant of GDA—can successfully converge to the global solution without requiring the Lipschitz constraint. Figure 1 illustrates the EG trajectories for the first 100 iterations. Notably, in the logistic loss case, the EG trajectory stalls in a flat region, where the gradient nearly vanishes, leading to a slow convergence.[3] This does not happen for the linear loss.

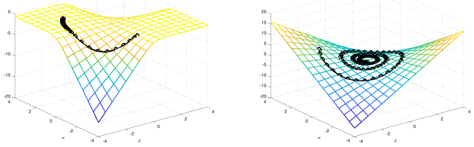


Figure 1: EG trajectories for Dirac GAN: (L) logistic loss and (R) linear loss

**Zero-Infinity distance and GAN.** Motivated by the Dirac GAN, we consider the linear loss $l(t) = -t$ without the Lipschitz constraint. This yields arguably the simplest GAN loss known to date, while still sharing the same global solutions as other GANs. In particular, solving GAN (1) with the linear loss is equivalent to minimizing what we refer to as the zero-infinity distance:[4]

$$\mathfrak{D}_{0/\infty}(p_{\text{data}}, p_g) := \begin{cases} 0, & p_{\text{data}} = p_g, \\ \infty, & \text{otherwise.} \end{cases}$$

For a realistic setting, we consider a neural network generator $G(\boldsymbol{z}) = \boldsymbol{G}\varphi_{\boldsymbol{\theta}}(\boldsymbol{z})$ with a linear last layer and a two-layer[5] neural network discriminator $D(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{x})$, where $\varphi_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathbb{R}^{d_2}$ is a $\boldsymbol{\theta}$-parameterized neural network, $\boldsymbol{\sigma}$ is a sigmoid function, $\boldsymbol{w} \in \mathbb{R}^r$ is a vector, $\boldsymbol{A} \in \mathbb{R}^{r\times d_1}$ and $\boldsymbol{G} \in \mathbb{R}^{d_1\times d_2}$ are matrices. For notational simplicity, let $\boldsymbol{g} := \operatorname{vec}(\boldsymbol{G}) \in \mathbb{R}^{d_1 d_2}$ and $\boldsymbol{\Phi}_i(\boldsymbol{\theta}) := \boldsymbol{I}_{d_1} \otimes \varphi_{\boldsymbol{\theta}}(\boldsymbol{z}_i)^\top \in \mathbb{R}^{d_1\times d_1 d_2}$, so that $G(\boldsymbol{z}_i) = \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}$. The resulting finite-sample GAN is given by:

$$\min_{\boldsymbol{g},\boldsymbol{\theta}} \max_{\boldsymbol{w},\boldsymbol{A}} \left\{ f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) := \frac{1}{n}\boldsymbol{w}^\top \sum_{i=1}^{n} [\boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{x}_i) - \boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})] \right\}, \tag{4}$$

We are now left to discuss a method that has potential to find a global solution of the nonconvex-nonconcave problem (4), and favors a minimum-norm generator.

## 5. Stable dynamics of extragradient in Zero-Infinity GAN

### 5.1. Strict non-minimax points and two-timescale methods

Escaping strict saddle via gradient descent is a cornerstone of success in nonconvex minimization [10, 20]. However, comparable results in minimax optimization have only emerged recently in [6, 18]. In particular, Chae et al. [6] introduced the following notion of a strict non-minimax point, the minimax analogue of a strict saddle, which can be escaped by two-timescale methods.

[3] Vanishing gradients in GANs are often attributed to the discriminator approaching optimality [2, 3], but they are in fact an inherent property of the loss landscape.

[4] Although not a proper distance due to its infinite values, we use an extended definition to accommodate this. Then, the zero-infinity distance qualifies as both an $f$-divergence [1] and an integral probability metric (IPM) [23].

[5] We study the two-layer case for simplicity, but believe the extension to deeper networks is straightforward.

For notational simplicity, consider the minimax problem $\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$, whose associated saddle gradient operator is defined as $\boldsymbol{F} := (\nabla_{\boldsymbol{x}} f, -\nabla_{\boldsymbol{y}} f)$.

**Definition 1 ([6], Definition 5)** *A stationary point $\boldsymbol{z}_* := (\boldsymbol{x}_*, \boldsymbol{y}_*)$ is said to be a strict non-minimax point of $f(\boldsymbol{x}, \boldsymbol{y})$ if $\lambda_{\min}(\boldsymbol{S}_{res}(D\boldsymbol{F}(\boldsymbol{z}_*))) < 0$ or $\lambda_{\min}(-\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}_*)) < 0$, where $\lambda_{\min}(\boldsymbol{A})$ denotes the smallest eigenvalue of $\boldsymbol{A}$, and $\boldsymbol{S}_{\mathrm{res}}(D\boldsymbol{F}(\boldsymbol{z}_*))$ denotes the restricted Schur complement [6] of $D\boldsymbol{F}$.*

Unlike gradient descent in minimization, plain GDA cannot distinguish between optimal and non-optimal stationary points [16]. Consequently, two-timescale methods that use a smaller step size for $\boldsymbol{x}$ have been found to be effective [6, 9, 16, 18]. We consider two-timescale GDA $\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k)$, where $\boldsymbol{z}_k := (\boldsymbol{x}_k, \boldsymbol{y}_k)$ and $\boldsymbol{\Lambda}_\tau := \mathrm{diag}\{(1/\tau)\boldsymbol{I}, \boldsymbol{I}\}$, and the two-timescale EG:

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k)).$$

For sufficiently large $\tau$, these methods can escape strict non-minimax points, while two-timescale GDA can even escape certain (local) optima [6, 9, 16, 18].

### 5.2. Stict non-minimax points in Zero-Infinity GAN

The following characterizes certain strict non-minimax points of (4), and implies that they exist in the Zero-Infinity GAN. For example, when $\boldsymbol{w}_l = \boldsymbol{0}$ and $\boldsymbol{a}_l = \boldsymbol{0}$, it is straightforward to construct a $\boldsymbol{g}, \boldsymbol{\theta}$ such that $\sum_{i=1}^n [\boldsymbol{x}_i - \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}] \neq \boldsymbol{0}$, which results in a strict non-minimax point.

**Theorem 2** *A stationary point of the Zero-Infinity GAN is a strict non-minimax point if there exists an index $l$ such that $w_l = 0$ and $\sum_{i=1}^n [\sigma'(\boldsymbol{a}_l^\top \boldsymbol{x}_i)\boldsymbol{x}_i - \sigma'(\boldsymbol{a}_l^\top \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}] \neq \boldsymbol{0}$.*

A more interesting question is whether all non-optimal stationary points are strict non-minimax points. However, this stronger statement does not hold. Consider a non-optimal stationary point $(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A})$ satisfying $\sum_{i=1}^n [\boldsymbol{x}_i - \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}] = \boldsymbol{0}$, $\boldsymbol{w} = \boldsymbol{0}$ and $\boldsymbol{A} = \boldsymbol{0}$. Since the Hessian $\nabla^2 f$ is a zero matrix, it is not a strict non-minimax point, despite being non-optimal.

### 5.3. Two-timescale EG (locally) converges to global solution

We now examine whether any two-timescale methods locally converge to global solutions. At any global solution of the Zero-Infinity GAN, where $\boldsymbol{w}$ is zero, all Hessians are zero except $\nabla^2_{\boldsymbol{gw}} f$ and $\nabla^2_{\boldsymbol{\theta w}} f$, so the discriminator Hessian is singular. This singularity affects the stability of two-timescale methods, where two-timescale EG outperforms two-timescale GDA.

**Theorem 3** *For sufficiently large $\tau$, the two-timescale GDA almost surely escapes global solutions of the Zero-Infinity GAN, where $\boldsymbol{w}$ is zero, whereas the two-timescale EG converges to such points.*

## 6. Implicit bias of extragradient in Zero-Infinity GAN

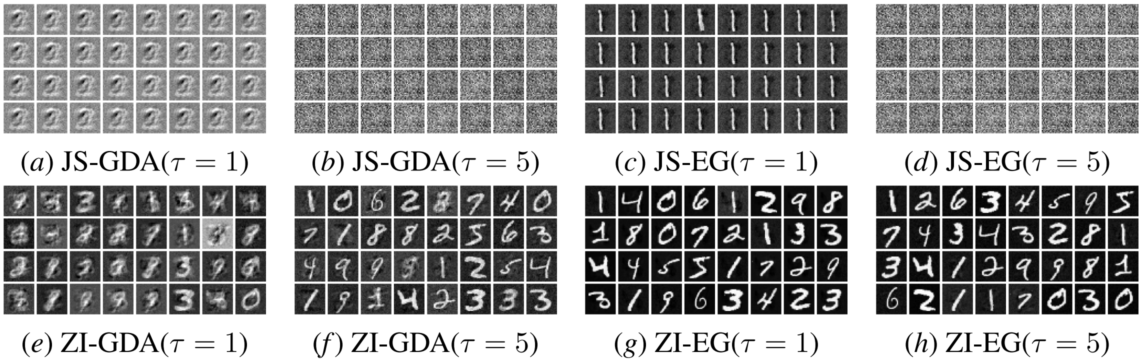
As a first step toward analyzing implicit bias, we study the regime where only the final layer $\boldsymbol{g}$ of generator is updated, while other layers ($\boldsymbol{\theta}$) are fixed; this mirrors the initial study of implicit bias in least squares regression [31]. Extension to realistic neural network settings is left for future work.

[6] The restricted Schur complement of $D\boldsymbol{F}$ reduces to the standard Schur complement when $\nabla^2_{\boldsymbol{yy}} f$ is invertible. We defer the detailed definition to the Appendix B.2.

We reach the conclusion by characterizing the globally optimal generator $(\boldsymbol{g}^*, \boldsymbol{\theta}^*)$. For some permutation $\pi$ over $n$ elements, the solution satisfies the linear system $\boldsymbol{x}_{\pi(i)} = \boldsymbol{\Phi}_i(\boldsymbol{\theta}^*)\boldsymbol{g}^*$ for $i = 1, \ldots, n$. Stacking the equations gives $\boldsymbol{x} = \boldsymbol{Q}_\pi \boldsymbol{\Phi}(\boldsymbol{\theta}^*)\boldsymbol{g}^*$, where $\boldsymbol{x} \in \mathbb{R}^{nd_1}$, $\boldsymbol{\Phi}(\boldsymbol{\theta}^*) \in \mathbb{R}^{nd_1 \times d_1 d_2}$ are formed by concatenating $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{\Phi}_i(\boldsymbol{\theta}^*)\}$, respectively, and $\boldsymbol{Q}_\pi$ is the permutation matrix associated with $\pi$. Hereafter we omit $\boldsymbol{\theta}^*$ for simplicity. For each $\pi$, the solution set is $\{\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{Q}_\pi^\top \boldsymbol{x} + \boldsymbol{y} : \boldsymbol{y} \in \mathcal{N}(\boldsymbol{\Phi})\}$. Each such set admits a minimum-norm solution $\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{Q}_\pi^\top \boldsymbol{x}$, whose norm can vary across permutations. Nevertheless, all minimum-norm solutions lie in $\mathcal{R}(\boldsymbol{\Phi}^\top)$, while other solutions do not. Gradient-based methods inherently favor those within $\mathcal{R}(\boldsymbol{\Phi}^\top)$, since $\mathcal{R}(\boldsymbol{\Phi}_i^\top) \subseteq \mathcal{R}(\boldsymbol{\Phi}^\top)$, and the gradient with respect to $\boldsymbol{g}$ consists of components that lie in $\mathcal{R}(\boldsymbol{\Phi}_i^\top)$. Therefore, starting from zero, any gradient method that converges to a global solution will favor the minimum-norm solution.

## 7. Experiment

Our preliminary experiments use only 32 MNIST samples, so we focus on verifying whether the algorithms reach global solutions; studying implicit bias are left for future work. We compare training with the original GAN (JSGAN) and Zero-Infinity GAN (ZIGAN). Figure 2 shows that ZIGAN outperforms JSGAN for all training methods. Moreover, in the ZIGAN case, (two-timescale) EG successfully reaches a global solution, memorizing and reproducing the training data, whereas (two-timescale) GDA approaches the vicinity of the global solution but fails to remain stable. See Appendix C for additional experiments under different initializations, showing similar results.


($a$) JS-GDA($\tau = 1$) ($b$) JS-GDA($\tau = 5$) ($c$) JS-EG($\tau = 1$) ($d$) JS-EG($\tau = 5$)

($e$) ZI-GDA($\tau = 1$) ($f$) ZI-GDA($\tau = 5$) ($g$) ZI-EG($\tau = 1$) ($h$) ZI-EG($\tau = 5$)

Figure 2: Generated MNIST samples from JSGAN and ZIGAN, trained with GDA and EG

## 8. Conclusion

This work advances the theoretical foundations of generative adversarial networks (GAN)—despite their waning popularity due to unstable training—by improving three key aspects: the loss landscape, gradient dynamics, and implicit bias, mirroring the role of gradient descent in supervised learning. We introduced the Zero-Infinity GAN, a formulation more compatible with gradient-based minimax optimization, and showed that certain non-optimal stationary points are strict and thus escapable via two-timescale extragradient method. We also uncover a novel implicit bias: extragradient dynamics initialized at zero converge to a minimum-norm generator solution. These findings provide a new lens on GAN training and lay theoretical groundwork for designing stable and generalizable generative models. While this is an initial step, understanding global convergence and scaling to more practical settings remain important directions for future research.

## Acknowledgement

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (Nos. RS-2019-NR040050, RS-2022-NR071715 and RS-2025-00515159).

## References

[1] Syed M. Ali and Samuel D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

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

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 214–223. PMLR, 2017.

[4] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International conference on machine learning*, pages 224–232. PMLR, 2017.

[5] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[6] Jiseok Chae, Kyuwon Kim, and Donghwan Kim. Two-timescale extragradient for finding local minimax points. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Jiseok Chae, Kyuwon Kim, and Donghwan Kim. Rethinking memorization–generalization trade-off in generative models. In *International Conference on Machine Learning Workshop on High-dimensional Learning Dynamics*, 2025.

[8] Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.

[9] Tanner Fiez and Lillian J Ratliff. Local convergence analysis of gradient descent ascent with finite timescale separation. In *International Conference on Learning Representations*, 2021.

[10] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[11] 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 (NeurIPS)*, 2014.

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, 2017.

[13] Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[14] Nick Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The GAN is dead; long live the GAN! a modern GAN baseline. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[15] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019.

[16] Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Kyuwon Kim and Donghwan Kim. Double-step alternating extragradient with increasing timescale separation for finding local minimax points: Provable improvements. In *Forty-first International Conference on Machine Learning*, 2024.

[19] G. M. Korpelevich. An extragradient method for finding saddle points and other problems. *Ekonomika i Mateaticheskie Metody*, 12(4):747–56, 1976.

[20] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, 2016.

[21] Lorenzo Luzi, Yehuda Dar, and Richard Baraniuk. Double descent and other interpolation phenomena in GANs. *arXiv preprint arXiv:2106.04003*, 2021.

[22] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[23] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[24] Vaishnavh Nagarajan and J. Zico Kolter. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 5585–5595. Curran Associates, Inc., 2017.

[25] Vaishnavh Nagarajan, Colin Raffel, and Ian J Goodfellow. Theoretical insights into memorization in GANs. In *Integration of Deep Learning Theories Workshop, NeurIPS*, 2018.

[26] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning,. In *International Conference on Learning Representations workshop track*, 2015.

[27] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[28] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[29] Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, 4(1):260–284, 2022.

[30] TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *International Conference on Machine Learning Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

[31] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

## Contents

**1 Introduction** 1

**2 Related works** 2

**3 Problem settings** 3

**4 Zero-Infinity GAN: Toward a better loss landscape in GANs** 3

**5 Stable dynamics of extragradient in Zero-Infinity GAN** 4

5.1 Strict non-minimax points and two-timescale methods . . . . . 4

5.2 Stict non-minimax points in Zero-Infinity GAN . . . . . 5

5.3 Two-timescale EG (locally) converges to global solution . . . . . 5

**6 Implicit bias of extragradient in Zero-Infinity GAN** 5

**7 Experiment** 6

**8 Conclusion** 6

**A Proofs and missing details in Section 4** 11

A.1 Preliminaries . . . . . 11

A.2 Linear loss: Zero-infinity distance . . . . . 11

A.3 Zero-infinity distance is an instance of $f$-divergence and integral probability metric 11

**B Proofs and missing details for Section 5** 12

B.1 Convergence behavior of two-timescale methods . . . . . 12

B.2 Definition of the restricted Schur complement . . . . . 12

B.3 Gradients and Hessians of the Zero-Infinity GAN loss . . . . . 13

B.4 Proof of Theorem 2 . . . . . 14

B.5 Proof of Theorem 3 . . . . . 15

**C Experiment detail and additional experiments** 15

## Appendix A. Proofs and missing details in Section 4

### A.1. Preliminaries

Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on a measurable space $(\Omega, \mathcal{F})$, such that both are absolutely continuous with respect to a measure $\mu$. This guarantees the existence of Radon-Nikodym derivatives $p$ and $q$ defined $\mu$-almost everywhere by $p := \frac{d\mathbb{P}}{d\mu}$ and $q := \frac{d\mathbb{Q}}{d\mu}$.

In addition, let $\mathbb{Q}$ be the pushforward of a latent variable $z \sim p_z$ through the generator $G$, *i.e.,* $\mathbb{Q} = G_\sharp p_z$. Then, for a given generator $G$, it is known that the optimal value of the inner maximization problem of GAN (1):

$$\begin{aligned}\mathfrak{D}(\mathbb{P}, \mathbb{Q}) &:= \max_D -\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}[l(D(\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[l(-D(\boldsymbol{x}))] \\ &= \max_D - \int_\Omega [l(D(\boldsymbol{x}))p(\boldsymbol{x}) + l(-D(\boldsymbol{x}))q(\boldsymbol{x})] d\mu(\boldsymbol{x})\end{aligned}$$

can measure a certain distance or divergence between the two probability measures $\mathbb{P}$ and $\mathbb{Q}$.

### A.2. Linear loss: Zero-infinity distance

Consider the linear loss $l(t) = -t$. Then we have

$$\mathfrak{D}(\mathbb{P}, \mathbb{Q}) = \max_D \int_\Omega D(\boldsymbol{x})[p(\boldsymbol{x}) - q(\boldsymbol{x})] d\mu(\boldsymbol{x}),$$

which can be maximized by

$$D^*(\boldsymbol{x}) = \begin{cases} \infty, & p(\boldsymbol{x}) > q(\boldsymbol{x}), \\ -\infty, & p(\boldsymbol{x}) < q(\boldsymbol{x}), \\ 0, & \text{otherwise.} \end{cases}$$

This leads to the zero-infinity distance.

$$\mathfrak{D}(\mathbb{P}, \mathbb{Q}) = \mathfrak{D}_{0/\infty}(\mathbb{P}, \mathbb{Q}).$$

### A.3. Zero-infinity distance is an instance of $f$-divergence and integral probability metric

The $f$-divergence [1] and integral probability metric [23] are defined as follows.

- $f$-divergence[7]:

$$\begin{aligned}\mathfrak{D}_f(\mathbb{P}, \mathbb{Q}) &:= \int_\Omega p(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\mu(\boldsymbol{x}) \\ &\geq \sup_{D\in\mathcal{D}} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[f^*(D(\boldsymbol{x}))],\end{aligned}$$

  where $f^*(t) = \sup_{s\in\mathrm{dom} f}\{st - f(s)\}$.

- Integral probability metrics:

$$\mathfrak{D}_{\mathrm{IPM}}(\mathbb{P}, \mathbb{Q}) := \max_{D\in\mathcal{D}} \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\mathbb{Q}}[D(\boldsymbol{x})].$$

[7] Here, we assume that $\mathbb{P}$ is absolutely continuous with respect to a measure $\mathbb{Q}$.

First, for the $f$-divergence, choosing

$$f(s) = \begin{cases} 0, & s = 1, \\ \infty, & \text{otherwise.} \end{cases}$$

yields the conjugate $f^*(t) = t$, and placing no constraint on $D$ leads to the zero-infinity distance. Similarly, taking $\mathcal{D}$ to be the full domain also results in the zero-infinity distance.

## Appendix B. Proofs and missing details for Section 5

### B.1. Convergence behavior of two-timescale methods

As explained in main text, two-timescale methods can escape strict non-minimax points when $\tau$ is sufficiently large. Moreover, two-timescale gradient descent-ascent can even escape certain local optima, as informally summarized below.

(P1) Both two-timescale GDA and EG methods almost surely escape strict non-minimax points [6, Theorem 5.7]. [8]

(P2) Two-timescale GDA can even escape certain (local) optimal points where $\nabla^2_{\boldsymbol{yy}} f$ is singular, whereas two-timescale EG converge to such points [6, Theorem 4.4 and 5.6].

### B.2. Definition of the restricted Schur complement

For convenience, we denote the Hessian of $f$ by $\boldsymbol{A} \coloneqq \nabla^2_{\boldsymbol{xx}} f$, $\boldsymbol{B} \coloneqq \nabla^2_{\boldsymbol{yy}} f$, and $\boldsymbol{C} \coloneqq \nabla^2_{\boldsymbol{xy}} f$. The definition of the restricted Schur complement is given below; it reduces to the standard Schur complement $\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top$ when $\boldsymbol{B}$ is invertible.

**Definition 4 (6, Definition 4)** *For $f \in C^2$, the restricted Schur complement is defined as $\boldsymbol{S}_{\mathrm{res}}(D\boldsymbol{F}) \coloneqq \boldsymbol{U}^\top(\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^\dagger\boldsymbol{C}^\top)\boldsymbol{U}$, with the matrix $\boldsymbol{U}$ defined below.*

Since $\boldsymbol{B}$ is symmetric, it is orthogonally diagonalizable as $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}^\top$, where $r = \mathrm{rank}(\boldsymbol{B})$, $\boldsymbol{\Delta} = \mathrm{diag}\{\delta_1, \dots, \delta_r, 0, \dots, 0\}$, and $\boldsymbol{P}$ is an orthogonal matrix. Let $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ denote the submatrices of $\boldsymbol{C}\boldsymbol{P}$, consisting of the first $r$ columns and the remaining $d_2 - r$, respectively. Then, $\boldsymbol{U}$ is defined as a matrix whose columns form an orthonormal basis for $\mathcal{R}(\boldsymbol{C}_2)^\perp$. The matrix $\boldsymbol{U}$ is not unique in general, but since only the spectrum of $\boldsymbol{S}_{\mathrm{res}}(D\boldsymbol{F})$ matters in the definition of a strict non-minimax point in Definition 1, any such choice of $\boldsymbol{U}$ suffices.

[8] Although Theorem 5.7 in [6] does not discuss the behavior of two-timescale GDA, this directly applies to two-timescale GDA, based on the fact that it is unstable than two-timescale EG from the dynamical system view.

### B.3. Gradients and Hessians of the Zero-Infinity GAN loss

The gradients of the Zero-Infinity GAN loss (4) are

$$\nabla_{\boldsymbol{g}} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) = -\frac{1}{n}\sum_{i=1}^{n}\left[(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta}))^\top \mathrm{diag}(\boldsymbol{\sigma}'(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}))\right]\boldsymbol{w}, \tag{5}$$

$$\nabla_{\theta_j} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) = -\frac{1}{n}\sum_{i=1}^{n}\left[\left(\boldsymbol{A}\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_j}\boldsymbol{g}\right)^\top \mathrm{diag}(\boldsymbol{\sigma}'(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}))\right]\boldsymbol{w}, \quad j = 1, \ldots, m, \tag{6}$$

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) = \frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{x}_i) - \boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})], \quad \text{and} \tag{7}$$

$$\nabla_{\boldsymbol{a}_l} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) = \frac{1}{n} w_l \sum_{i=1}^{n}[\sigma'(\boldsymbol{a}_l^\top \boldsymbol{x}_i)\boldsymbol{x}_i - \sigma'(\boldsymbol{a}_l^\top \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}], \quad l = 1, \ldots, r. \tag{8}$$

Moreover, the Hessians of the Zero-Infinity GAN loss (4) are

$$\begin{aligned}
\nabla^2_{\boldsymbol{g}\boldsymbol{g}} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) &= -\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta}))^\top \mathrm{diag}(\boldsymbol{\sigma}''(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}))\mathrm{diag}(\boldsymbol{w})\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta}),\\
\nabla^2_{\boldsymbol{g}\boldsymbol{w}} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) &= -\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta}))^\top \mathrm{diag}(\boldsymbol{\sigma}'(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})),\\
\nabla^2_{\boldsymbol{w}\boldsymbol{w}} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) &= \boldsymbol{0},\\
\nabla^2_{\boldsymbol{a}_l\boldsymbol{a}_j} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) &= \begin{cases} \frac{1}{n} w_l \sum_{i=1}^{n}[\sigma''(\boldsymbol{a}_l^\top \boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top - \sigma''(\boldsymbol{a}_l^\top \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}\boldsymbol{g}^\top \boldsymbol{\Phi}_i(\boldsymbol{\theta})^\top], & j = l,\\ \boldsymbol{0}, & \text{otherwise,}\end{cases}\\
\nabla^2_{\boldsymbol{g}\boldsymbol{a}_l} f(\boldsymbol{g}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{A}) &= -\frac{1}{n} w_l \sum_{i=1}^{n}[\sigma''(\boldsymbol{a}_l^\top \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})\boldsymbol{\Phi}_i(\boldsymbol{\theta})^\top \boldsymbol{a}_l \boldsymbol{g}^\top \boldsymbol{\Phi}_i(\boldsymbol{\theta})^\top + \sigma'(\boldsymbol{a}_l^\top \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})\boldsymbol{\Phi}_i(\boldsymbol{\theta})^\top], \quad \text{and}\\
\nabla^2_{\boldsymbol{w}\boldsymbol{a}_l} f(\boldsymbol{g}, \boldsymbol{w}, \boldsymbol{A}) &= \begin{bmatrix} \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \\ \frac{1}{n}\sum_{i=1}^{n}[\sigma'(\boldsymbol{a}_l^\top \boldsymbol{x}_i)\boldsymbol{x}_i^\top - \sigma'(\boldsymbol{a}_l^\top \boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})(\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})^\top] \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix} \begin{matrix} \\ \\ \\ \leftarrow l\text{th row} \\ \\ \\ \\ \end{matrix}
\end{aligned}$$

$$\begin{aligned}
\nabla^2_{\boldsymbol{g}\theta_j} f(\boldsymbol{g},\boldsymbol{\theta},\boldsymbol{w},\boldsymbol{A}) &= -\frac{1}{n}\sum_{i=1}^{n}\Bigg[\left(\boldsymbol{A}\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_j}\right)^{\top}\mathrm{diag}(\boldsymbol{\sigma}'(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}))\boldsymbol{w} \\
&\qquad + (\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta}))^{\top}\,\mathrm{diag}(\boldsymbol{\sigma}''(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}))\mathrm{diag}(\boldsymbol{w})\boldsymbol{A}\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_j}\boldsymbol{g}\Bigg] \\
\nabla^2_{\theta_j\theta_k} f(\boldsymbol{g},\boldsymbol{\theta},\boldsymbol{w},\boldsymbol{A}) &= -\frac{1}{n}\sum_{i=1}^{n}\Bigg[\left(\boldsymbol{A}\frac{\partial^2 \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_k\partial \theta_j}\boldsymbol{g}\right)^{\top}\mathrm{diag}(\boldsymbol{\sigma}'(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}))\boldsymbol{w} \\
&\qquad + \left(\boldsymbol{A}\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_j}\boldsymbol{g}\right)^{\top}\mathrm{diag}(\boldsymbol{\sigma}''(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}))\mathrm{diag}(\boldsymbol{w})\left(\boldsymbol{A}\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_k}\boldsymbol{g}\right)\Bigg] \\
\nabla^2_{\boldsymbol{w}\theta_j} f(\boldsymbol{g},\boldsymbol{\theta},\boldsymbol{w},\boldsymbol{A}) &= -\frac{1}{n}\sum_{i=1}^{n}\Bigg[\left(\boldsymbol{A}\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_j}\boldsymbol{g}\right)^{\top}\mathrm{diag}(\boldsymbol{\sigma}'(\boldsymbol{A}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}))\Bigg] \\
\nabla^2_{\boldsymbol{a}_l\theta_j} f(\boldsymbol{g},\boldsymbol{\theta},\boldsymbol{w},\boldsymbol{A}) &= -\frac{1}{n}\sum_{i=1}^{n}\Bigg[\sigma''(\boldsymbol{a}_l^{\top}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})\left(\boldsymbol{a}_l^{\top}\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_j}\boldsymbol{g}\right)\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g} + \sigma'(\boldsymbol{a}_l^{\top}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{\theta})}{\partial \theta_j}\boldsymbol{g}\Bigg] w_l
\end{aligned}$$

for $l = 1, \ldots, r$ and $j, k = 1, \ldots, m$.

### B.4. Proof of Theorem 2

**Proof** We show that the Hessian with respect to the discriminator parameters

$$\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f := \begin{bmatrix} \nabla^2_{\boldsymbol{w}\boldsymbol{w}} f & \nabla^2_{\boldsymbol{w}\boldsymbol{a}_1} f & \cdots & \nabla^2_{\boldsymbol{w}\boldsymbol{a}_r} f \\ \nabla^2_{\boldsymbol{a}_1\boldsymbol{w}} f & \nabla^2_{\boldsymbol{a}_1\boldsymbol{a}_1} f & \cdots & \nabla^2_{\boldsymbol{a}_1\boldsymbol{a}_r} f \\ \vdots & \vdots & & \vdots \\ \nabla^2_{\boldsymbol{a}_r\boldsymbol{w}} f & \nabla^2_{\boldsymbol{a}_r\boldsymbol{a}_1} f & \cdots & \nabla^2_{\boldsymbol{a}_r\boldsymbol{a}_r} f \end{bmatrix} \in \mathbb{R}^{(1+d_1)r\times(1+d_1)r}.$$

has a positive eigenvalue, which implies that these are strict non-minimax points. Assume that $w_l = 0$ and $\sum_{i=1}^{n}[\sigma'(\boldsymbol{a}_l^{\top}\boldsymbol{x}_i)\boldsymbol{x}_i - \sigma'(\boldsymbol{a}_l^{\top}\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g})\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g}] \neq \boldsymbol{0}$ for some $l$. Then, we have $\nabla^2_{\boldsymbol{a}_l\boldsymbol{a}_l} f = \boldsymbol{0}$ and $\nabla^2_{\boldsymbol{w}\boldsymbol{a}_l} f \neq \boldsymbol{0}$. Consider a submatrix of $\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f$:

$$\boldsymbol{M} := \begin{bmatrix} \nabla^2_{\boldsymbol{w}\boldsymbol{w}} f & \nabla^2_{\boldsymbol{w}\boldsymbol{a}_l} f \\ \nabla^2_{\boldsymbol{a}_l\boldsymbol{w}} f & \nabla^2_{\boldsymbol{a}_l\boldsymbol{a}_l} f \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \nabla^2_{\boldsymbol{w}\boldsymbol{a}_l} f \\ \nabla^2_{\boldsymbol{a}_l\boldsymbol{w}} f & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{r+d_1\times r+d_1},$$

which is real symmetric with all diagonal entries equal to zero and at least one non-zero off-diagonal entry. Let $\{\lambda_i\}_{i=1}^{r+d_1}$ denote the eigenvalues of the matrix $\boldsymbol{M}$, which is real-valued since $\boldsymbol{M}$ is real symmetric. Moreover, since $\mathrm{tr}(\boldsymbol{M}) = \sum_{i=1}^{n}\lambda_i = 0$ and $\boldsymbol{M}$ has at least one non-zero entry, it must have a positive eigenvalue.

We now show that $\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f$ has a positive eigenvalue. Let $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2) \in \mathbb{R}^{r+d_1}$ be a vector such that $\boldsymbol{u}^{\top}\boldsymbol{M}\boldsymbol{u} > 0$, where $\boldsymbol{u}_1 \in \mathbb{R}^{r}$ corresponds to the $\boldsymbol{w}$-coordinates and $\boldsymbol{u}_2 \in \mathbb{R}^{d_1}$ corresponds to the $\boldsymbol{a}_l$-cooridnates. Define the vector $\boldsymbol{v} \in \mathbb{R}^{(1+d_1)r}$ by placing $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ into the slots corresponding to $\boldsymbol{w}$ and $\boldsymbol{a}_l$, respectively, and setting all other entries to zero. Then, $\boldsymbol{v}^{\top}\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f\boldsymbol{v} = \boldsymbol{u}^{\top}\boldsymbol{M}\boldsymbol{u} > 0$, the Hessian $\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f$ has a positive eigenvalue.

■

### B.5. Proof of Theorem 3

**Proof** Consider a global solution that satisfies $\boldsymbol{\Phi}_i(\boldsymbol{\theta})\boldsymbol{g} = \boldsymbol{x}_{\pi(i)}$ and $\boldsymbol{w}$ is zero. Then, we have all Hessian is zero except $\nabla^2_{\boldsymbol{g}\boldsymbol{w}} f$, $\nabla^2_{\boldsymbol{\theta}\boldsymbol{w}} f$.

$$D\boldsymbol{F} = \begin{bmatrix} \nabla^2_{\boldsymbol{x}\boldsymbol{x}} f & \nabla^2_{\boldsymbol{x}\boldsymbol{y}} f \\ -\nabla^2_{\boldsymbol{y}\boldsymbol{x}} f & -\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \nabla^2_{\boldsymbol{g}\boldsymbol{w}} f & \nabla^2_{\boldsymbol{\theta}\boldsymbol{w}} f & \cdots & \nabla^2_{\boldsymbol{g}\boldsymbol{a}_r} f \\ -\nabla^2_{\boldsymbol{w}\boldsymbol{g}} f & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ -\nabla^2_{\boldsymbol{w}\boldsymbol{\theta}} f & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ -\nabla^2_{\boldsymbol{a}_r\boldsymbol{g}} f & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}$$

As studied in [6, 16], the stability near a stationary point is characterized by the spectrum of the following timescaled matrix:

$$\boldsymbol{\Lambda}_\tau D\boldsymbol{F} = \begin{bmatrix} \mathbf{0} & \frac{1}{\tau}\boldsymbol{B} \\ -\boldsymbol{B}^\top & \mathbf{0} \end{bmatrix},$$

where $\boldsymbol{B} = \begin{bmatrix} \nabla^2_{\boldsymbol{g}\boldsymbol{w}} f & \nabla^2_{\boldsymbol{g}\boldsymbol{a}_1} f & \cdots & \nabla^2_{\boldsymbol{g}\boldsymbol{a}_r} f \end{bmatrix}$. By Schur's formula:

$$\det\left(\begin{bmatrix} -\lambda I & \frac{1}{\tau}\boldsymbol{B} \\ -\boldsymbol{B}^\top & -\lambda I \end{bmatrix}\right) = \det\left(\lambda^2 I + \frac{1}{\tau}\boldsymbol{B}^\top\boldsymbol{B}\right),$$

we can show that the eigenvalues $\{\lambda_n\}$ of the matrix $\boldsymbol{\Lambda}_\tau D\boldsymbol{F}$ is either zero or pure imaginary. Specifically, let $\{\mu_n\}$ be the (nonnegative) eigenvalues of positive semidefinite matrix $\boldsymbol{B}^\top\boldsymbol{B}$, then we have

$$\lambda_n = i\sqrt{\frac{\mu_n}{\tau}}.$$

Provided that $\{\mu_n\}$ are not all zero, [6, Theorem 4.4] implies that the two-timescale GDA almost surely escapes a global solution where either $\boldsymbol{w}$ or $\boldsymbol{A}$ is zero for any $\tau$, while the two-timescale EG does not. ■

## Appendix C. Experiment detail and additional experiments

We presample the latent variable $\boldsymbol{z}_i \in \mathbb{R}^{100}$ from standard Gaussian distribution $\mathcal{N}(\mathbf{0}, I_{100})$, where the number of samples is equivalent to that of training samples. We employed two-layer sigmoid neural networks for both the generator $G : \mathbb{R}^{100} \to \mathbb{R}^{784}$ and discriminator $D : \mathbb{R}^{784} \to \mathbb{R}$, with the hidden layer widths set to 500 and 3000, respectively. Each experimental result shows the images generated by the generator from the presampled noise after 100,000 iterations of full-batch training.

We ran five additional experiments with different initializations for the eight settings in Figure 2. The results, shown in Figures 3 and 4, closely match those in Figure 2.

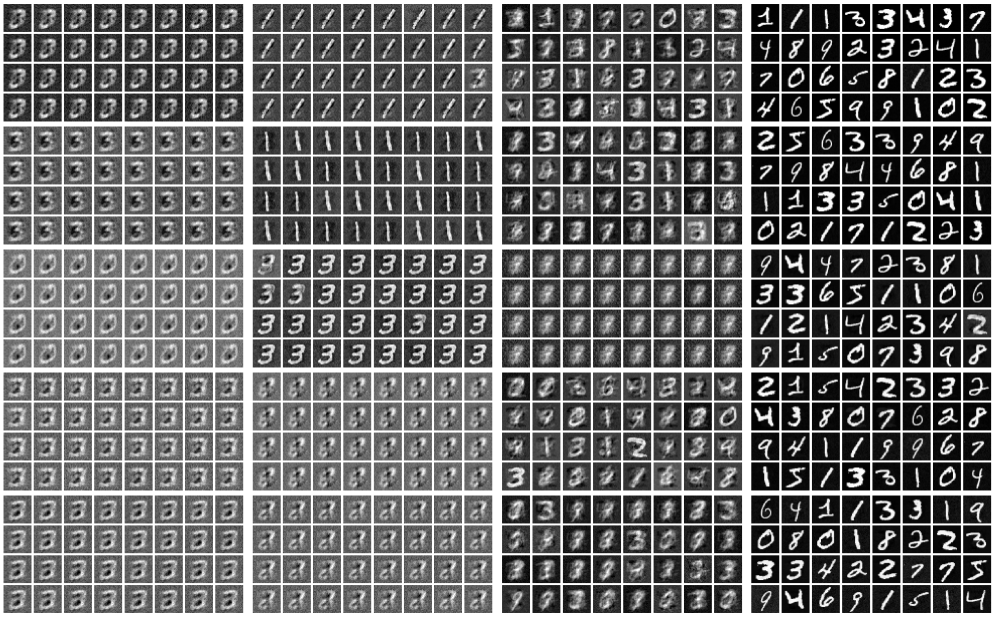


Figure 3: (i) JS-GDA($\tau = 1$), (ii) JS-EG($\tau = 1$), (iii) ZI-GDA($\tau = 1$), (iv) ZI-EG($\tau = 1$)

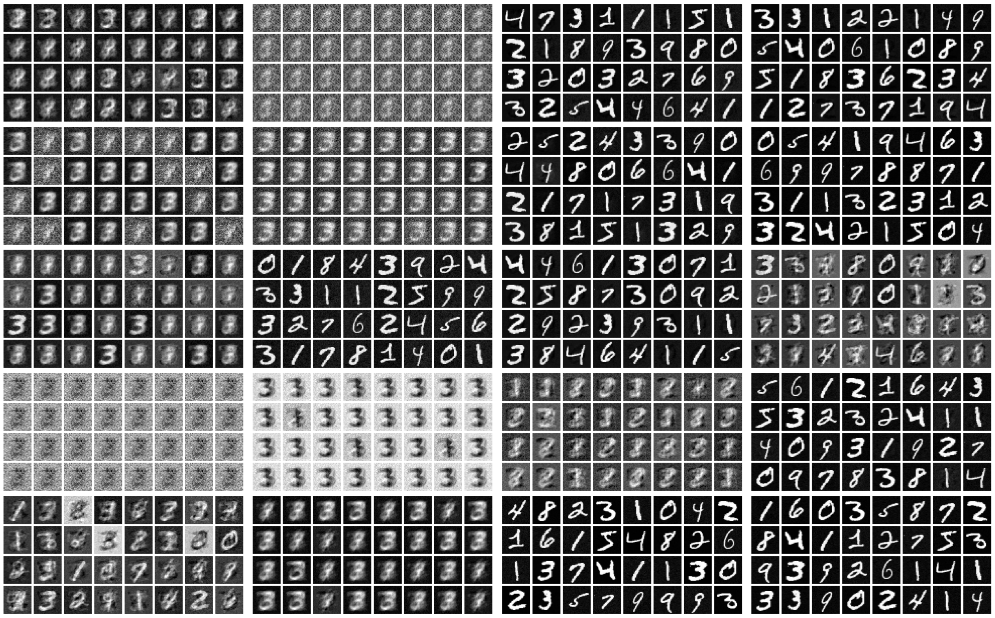


Figure 4: (i) ZI-GDA($\tau = 1$), (ii) ZI-GDA($\tau = 5$), (iii) ZI-EG($\tau = 1$), (iv) ZI-EG($\tau = 5$)