SIGMA XPLORE

AI Letters

Journal homepage: www.ail.sigmaxplore.com

AI Letters 2, Article ID: 2951, 1–2, July 2025 ISSN 3065-3525

A Knife-Edge Stability in Dynamics of Generative Adversarial Learning

Tetsuya Saito*

Abstract

This paper analyzes strategic dynamics between generators and discriminators in adversarial learning through a simplified strategic framework. A knife-edge stability, a fragile balance where small deviations destabilize learning, is identified in the adversarial training process. A deviation from this stability results in failures: mode collapse occurs at mixed-strategy Nash equilibria, where discriminators become overly dominant, while oscillatory behavior arises, when generators receive insufficient learning signals. Practical implementations require careful tuning of learning parameters to maintain this delicate equilibrium, offering a theoretical account of the instability frequently encountered in adversarial training.

Keywords: Generative adversarial network (GAN), mode collapse, oscillatory behavior, stability analysis, Nash equilibrium.

1. Preliminaries

Generative adversarial networks (GANs) [1] have become an important technique in machine learning [2]. However, several instability outcomes resulting in mode collapse and oscillatory behavior are reported. Some solutions are provided by applications of alternative loss metrics [3] and introductions of alternative structures [4], [5], [6], [7].

In standard GANs [1], a generator network G attempts to produce synthetic data $x' \sim g$ from an arbitrary input z using the generator function $G: z \longmapsto x'$ that resembles the real data distribution $x \sim f$, while a discriminator network D attempts to distinguish between real and synthetic samples. The global convergence of distribution is achieved by $g \mapsto f$ in the learning process. This adversarial process is formalized as a two-player minimax game where the generator minimizes, and the discriminator maximizes a value function involving the log-likelihood of correctly identifying real and fake samples.

The basic algorithm is illustrated by Algorithm 1, where the discriminator aims to maximize its ability to distinguish between real and generated data by V_D, while the generator aims to minimize the discriminator's success by V_G. ∇ indicates the corresponding gradient vector. The value functions in the standard GAN [1] are provided by

$$V_D = \mathbb{E}_{x\sim f} \log D(x) + \mathbb{E}_{x'\sim g} \log(1 - D(G(z))) \tag{1}$$

$$V_G = \mathbb{E}_{x'\sim g} \log D(G(z)), \tag{2}$$

where D is the discriminator function assigning the probability that the given data is real or synthetic.

The training process often suffers from instability [8], [9], with gradient-based updates leading to oscillations or mode collapse, where the generator produces limited varieties of samples [10]. To analyze these strategic dynamics, we provide a simplified two-player two-action non-cooperative game that captures the essential equilibrium properties underlying GANs [11] as follows.

A strategic interaction between two players is considered, as illustrated in Table 1, where D^* is the probability of identification when the two players optimize their actions and $D' = D^* - \varepsilon$ for $\varepsilon \geq 0$ when both do not optimize. For simplicity, the player is guaranteed to win this game if one optimizes its action, but the opponent does not.

Algorithm 1 Standard GAN Training

```
Require: $\beta_D$: discriminator learning rate, $\beta_G$: generator learning rate
while not converged do
    $\theta_D \longleftarrow \theta_D + \beta_D \nabla V_D$
    $\theta_G \longleftarrow \theta_G + \beta_G \nabla V_G$
end while
```

Table 1. Game structure and payoffs, where the discriminator (D) is the *row player*, and the generator (G) is the *column player*.

Payoff = (D, G)	*Optimize*	*Not Optimize*
Optimize	$(D^*, 1 - D^*)$	$(1, 0)$
Not Optimize	$(0, 1)$	$(D', 1 - D')$

2. Nash Equilibrium

Let π_D and π_G be the mixed strategies of D and G, respectively, take optimization action, with values between 0 and 1. The expected payoffs for both players can be calculated based on the probability of each player optimizing their strategy and the payoff matrix in Table 1. When we set $D^* = (1 - \varepsilon)/2$, which creates a balanced relationship where $D^* + D' = 1$, analyzing the first-order conditions leads to a dominant strategy where both players choose to optimize ($\pi_D^* = \pi_G^* = 1$).

For the case where $D^* + D' > 1$, analyzing the players' best responses reveals two possible equilibria below:

- A *pure strategy Nash equilibrium* (PSNE) where neither player optimizes ($\pi_D^* = \pi_G^* = 0$).
- A *mixed strategy Nash equilibrium* (MSNE) where the discriminator may optimize with some probability ($\pi_D^* = p$) but the generator never optimizes ($\pi_G^* = 0$), which occurs specifically when $D^* = 1$ indicating $\varepsilon = 0$ (mode collapse condition).

Similarly, when $D^* + D' < 1$, we can find the following two equilibria:

- The same PSNE where neither player optimizes.
- A different MSNE where the discriminator never optimizes ($\pi_D^* = 0$) but the generator may optimize with some probability ($\pi_G^* = q$), occurring when $D^* = \varepsilon$.

A summary of these five (or four) equilibrium classes is provided in Figure 1 to guide the subsequent investigation.

Globiz Professional University, Kawasaki, Japan
* Corresponding Author: Tetsuya Saito
E-mail address: t_saito@gpu.ac.jp, ORCID: https://orcid.org/0000-0002-9117-6751

PATENT LAB

Received: 15 March 2025; Revised: 2 April 2025; Accepted: 5 April 2025

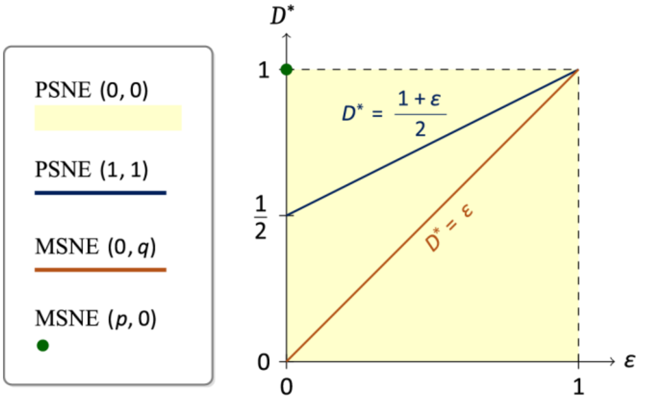

Figure 1. Nash equilibria in adversarial learning

3. Knife-Edge Learning Stability

Knife-edge stability refers to the narrow set of conditions under which equilibrium can be sustained. Outside this range, adversarial training typically diverges into failure modes such as oscillation or collapse.

PSNE $(0, 0)$ indicates that both players have no incentive to update their functions due to insufficient rewards. MSNE $(p, 0)$ indicates an excessively strong D to eliminate the incentive of G to make an effort. These states can represent *mode collapse* in GAN implementations. MSNE $(0, q)$ indicates $D' = 0$ and there is no incentive for D to update the discriminator function anymore, yet G still struggles with updating its generation function. However, G cannot obtain any informative response from D, resulting in an *oscillation*. The distinction of the two modes suggests that mode collapse requires techniques to prevent G from becoming *overly dominant*, while oscillation to ensure the provision of consistent learning signals.

PSNE $(1, 1)$ indicates successful enforcement of the learning process, as both players optimize their objective functions, resulting in the provision of meaningful signals to opponents. The convergence indicates that there are no further updates: $D^* = D'$ $(\varepsilon = 0)$. Ideally, D must be completely myopic about the identification of the fake: $D^* = 1/2$. In this case, $D^* + D' = 1$ is applied to find PSNE $(1, 1)$. If ε represents the convergence path, the stable evolution is required to conform to $D^* = (1 + \epsilon)/2$, or it causes mode collapse, which is a **knife-edge stable process**. In practice, ε is almost impossible to observe; thus, this stability critically depends on the choice of learning parameters, especially the coefficients (β_D, β_G) on each gradient, $\nabla_\theta D$ and $\nabla_\theta G$ in Algorithm 1 (learning rates).

4. Conclusion

This paper introduced a game-theoretic model identifying a knife-edge equilibrium in adversarial learning. The analysis accounts for both mode collapse and oscillation as distinct mixed-strategy Nash equilibria that deviate from the ideal training path. This narrow stability region helps explain the sensitivity of adversarial training to parameter choices. The structured approach highlights core strategic tensions underlying practical implementation.

Learning rates emerge as potentially critical parameters associated with knife-edge stability, suggesting directions for studying how training hyperparameters influence equilibrium outcomes. This study contributes to structural approaches in adversarial learning, offering a formal lens to interpret strategic dynamics and complement empirical methods aimed at improving stability.

Acknowledgement

The author thanks the anonymous reviewer for the constructive feedback, which helped improve the clarity, precision, and overall presentation of this work.

Conflict of Interest Statement

The author declare no conflict of interest.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adver- sarial networks," in *Advances in Neural Information Processing Systems*, 27 (2672–2680) 2014.

[2] M. Moghaddam, B. Boroomand, M. Jalali *et al.*, "Games of GANs: Game-theoretical models for generative adversarial networks," *Artificial Intelligence Review*, 56 (9771–9807) 2023.

[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, 70 (214–223) 2017.

[4] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, andS. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 30, 2017.

[5] H. Zhang, S. Xu, J. Jiao, P. Xie, R. Salakhutdinov, and E. Xing, "Stackelberg GAN: Towards provable minimax equilibrium via multi-generator architectures," *arXiv*, vol. arXiv:1811.08010, 2018.

[6] T. Fiez, B. Chasnov, and L. Ratliff, "Conver- gence of learning dynamics in stackelberg games," *arXiv*, vol. arXiv:1906.01217, 2019.

[7] T. Saito, "TIOLI-GAN: A bargaining for efficiency," *TechRxiv*, March 2025.

[8] L. Mescheder, S. Nowozin, and A. Geiger, "The numerics of GANs," *arXiv preprint arXiv:1705.10461*, 2017.

[9] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge?" in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, (3481–3490) 2018.

[10] F. Farnia and A. Ozdaglar, "GANs may have no nash equilibria," *arXiv*, vol. arXiv:2002.09124, 2020.

[11] T. Hazra and K. Anjaria, "Applications of game theory in deep learning: a survey," *Neural Computing and Applications*, vol. 81, pp. 8963–8994, February 2022.