Strict Non-Optimality in Minimax Optimization and Its Application to GAN

Donghwan Kim

KAIST

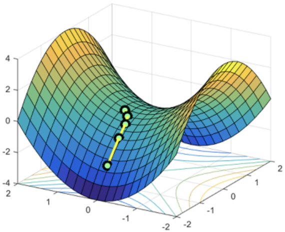
ICCOPT 2025, USC

July 24, 2025

Gradient descent (GD) and strict saddle

- GD: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)$

- Strict saddle point: its Hessian $\nabla^2 f(\boldsymbol{x})$ has a strictly negative eigenvalue



- By stable manifold theorem, for any strict saddle $\tilde{\boldsymbol{x}}$, GD satisfies[1]

$$P(\lim_k \boldsymbol{x}_k = \tilde{\boldsymbol{x}}) = 0.$$

If GD converges, then it is almost surely not a strict saddle point.

[1] Lee, Simchowitz, Jordan and Recht, Gradient descent only converges to minimizers, COLT, 2016.

Minimax and gradient descent ascent (GDA)

- Minimax optimization:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \; f(\boldsymbol{x}, \boldsymbol{y})$$

- GDA:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \end{aligned}$$

- Let $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{F} := (\nabla_{\boldsymbol{x}} f, -\nabla_{\boldsymbol{y}} f)$:

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

Standard optimality in minimax optimization

- Nash equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le f(\boldsymbol{x}, \boldsymbol{y}_*), \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

- Local Nash equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le f(\boldsymbol{x}, \boldsymbol{y}_*), \quad \forall \boldsymbol{x}, \boldsymbol{y} \text{ in a neighborhood of } (\boldsymbol{x}_*, \boldsymbol{y}_*).$$

- Second-order necessary condition:

$$\nabla f(\boldsymbol{z}_*) = \boldsymbol{0}, \quad \nabla^2_{\boldsymbol{xx}} f(\boldsymbol{z}_*) \succeq \boldsymbol{0}, \quad \text{and} \quad \nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}_*) \preceq \boldsymbol{0}$$

- **Strict non-Nash point**: (analogous to strict saddle)
  at least one of $\nabla^2_{\boldsymbol{xx}} f(\boldsymbol{z})$ or $-\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z})$ has a strictly negative eigenvalue

  (Are we done?)

Standard optimality in minimax optimization

- Nash equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le f(\boldsymbol{x}, \boldsymbol{y}_*), \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

- Local Nash equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le f(\boldsymbol{x}, \boldsymbol{y}_*), \quad \forall \boldsymbol{x}, \boldsymbol{y} \text{ in a neighborhood of } (\boldsymbol{x}_*, \boldsymbol{y}_*).$$

- Second-order necessary condition:

$$\nabla f(\boldsymbol{z}_*) = \mathbf{0}, \quad \nabla^2_{\boldsymbol{xx}} f(\boldsymbol{z}_*) \succeq \mathbf{0}, \quad \text{and} \quad \nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}_*) \preceq \mathbf{0}$$

- **Strict non-Nash point**: (analogous to strict saddle)
  at least one of $\nabla^2_{\boldsymbol{xx}} f(\boldsymbol{z})$ or $-\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z})$ has a strictly negative eigenvalue
  (Are we done?)

Standard optimality in minimax optimization

- Nash equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le f(\boldsymbol{x}, \boldsymbol{y}_*), \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

- Local Nash equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le f(\boldsymbol{x}, \boldsymbol{y}_*), \quad \forall \boldsymbol{x}, \boldsymbol{y} \text{ in a neighborhood of } (\boldsymbol{x}_*, \boldsymbol{y}_*).$$

- Second-order necessary condition:

$$\nabla f(\boldsymbol{z}_*) = \mathbf{0}, \quad \nabla^2_{\boldsymbol{xx}} f(\boldsymbol{z}_*) \succeq \mathbf{0}, \quad \text{and} \quad \nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}_*) \preceq \mathbf{0}$$

- **Strict non-Nash point**: (analogous to strict saddle)
  at least one of $\nabla^2_{\boldsymbol{xx}} f(\boldsymbol{z})$ or $-\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z})$ has a strictly negative eigenvalue

(Are we done?)

GDA can converge to non-Nash points

- Stability of a dynamic $\boldsymbol{x}_{k+1} = \boldsymbol{w}(\boldsymbol{x}_k)$ is determined by its Jacobian $D\boldsymbol{w}$.

- GD: $\boldsymbol{x}_{k+1} = \boldsymbol{w}_1(\boldsymbol{x}_k) = \boldsymbol{x}_k - \eta\nabla f(\boldsymbol{x}_k)$, $\quad D\boldsymbol{w}_1 = \boldsymbol{I} - \eta\nabla^2 f$
  GDA: $\boldsymbol{z}_{k+1} = \boldsymbol{w}_2(\boldsymbol{z}_k) = \boldsymbol{z}_k - \eta\boldsymbol{F}(\boldsymbol{z}_k)$, $\quad D\boldsymbol{w}_2 = \boldsymbol{I} - \eta D\boldsymbol{F}$

- GD almost surely escapes strict saddle, while GDA has no such guarantee[2], since

$$D\boldsymbol{F} = \begin{bmatrix} \nabla^2_{\boldsymbol{x}\boldsymbol{x}} f & \nabla^2_{\boldsymbol{x}\boldsymbol{y}} f \\ -\nabla^2_{\boldsymbol{y}\boldsymbol{x}} f & -\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f \end{bmatrix}$$

  has no direct connection to $\nabla^2_{\boldsymbol{x}\boldsymbol{x}} f$ and $\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f$ in general.

[2] Daskalakis and Panageas, The limit points of (optimistic) gradient descent in min-max optimization, NeurIPS, 2018.

GDA can converge to non-Nash points

- Stability of a dynamic $\boldsymbol{x}_{k+1} = \boldsymbol{w}(\boldsymbol{x}_k)$ is determined by its Jacobian $D\boldsymbol{w}$.

- GD: $\boldsymbol{x}_{k+1} = \boldsymbol{w}_1(\boldsymbol{x}_k) = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k), \quad D\boldsymbol{w}_1 = \boldsymbol{I} - \eta \nabla^2 f$
  GDA: $\boldsymbol{z}_{k+1} = \boldsymbol{w}_2(\boldsymbol{z}_k) = \boldsymbol{z}_k - \eta \boldsymbol{F}(\boldsymbol{z}_k), \quad D\boldsymbol{w}_2 = \boldsymbol{I} - \eta D\boldsymbol{F}$

- GD almost surely escapes strict saddle, while
  GDA has no such guarantee[2], since

$$D\boldsymbol{F} = \begin{bmatrix} \nabla^2_{\boldsymbol{xx}} f & \nabla^2_{\boldsymbol{xy}} f \\ -\nabla^2_{\boldsymbol{yx}} f & -\nabla^2_{\boldsymbol{yy}} f \end{bmatrix}$$

  has no direct connection to $\nabla^2_{\boldsymbol{xx}} f$ and $\nabla^2_{\boldsymbol{yy}} f$ in general.

[2]Daskalakis and Panageas, The limit points of (optimistic) gradient descent in min-max optimization, NeurIPS, 2018.

GDA can converge to non-Nash points

- Stability of a dynamic $\boldsymbol{x}_{k+1} = \boldsymbol{w}(\boldsymbol{x}_k)$ is determined by its Jacobian $D\boldsymbol{w}$.

- GD: $\boldsymbol{x}_{k+1} = \boldsymbol{w}_1(\boldsymbol{x}_k) = \boldsymbol{x}_k - \eta\nabla f(\boldsymbol{x}_k), \quad D\boldsymbol{w}_1 = \boldsymbol{I} - \eta\nabla^2 f$
  GDA: $\boldsymbol{z}_{k+1} = \boldsymbol{w}_2(\boldsymbol{z}_k) = \boldsymbol{z}_k - \eta\boldsymbol{F}(\boldsymbol{z}_k), \quad D\boldsymbol{w}_2 = \boldsymbol{I} - \eta D\boldsymbol{F}$

- GD almost surely escapes strict saddle, while
  GDA has no such guarantee[2], since

$$D\boldsymbol{F} = \begin{bmatrix} \nabla^2_{\boldsymbol{xx}} f & \nabla^2_{\boldsymbol{xy}} f \\ -\nabla^2_{\boldsymbol{yx}} f & -\nabla^2_{\boldsymbol{yy}} f \end{bmatrix}$$

  has no direct connection to $\nabla^2_{\boldsymbol{xx}} f$ and $\nabla^2_{\boldsymbol{yy}} f$ in general.

[2] Daskalakis and Panageas, The limit points of (optimistic) gradient descent in min-max optimization, NeurIPS, 2018.

Stackelberg equilibrium

- Nash may not even exist.[3,4] Any broader alternative?

- **Stackelberg** equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}'), \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$$

Definition 1 (Jin-Netrapalli-Jordan, ICML,'20)

A point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is said to be a **local minimax point** *if there exists $\delta_0 > 0$ and a function h satisfying $h(\delta) \to 0$ as $\delta \to 0$ such that, for any $\delta \in (0, \delta_0]$ and any $(\boldsymbol{x}, \boldsymbol{y})$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}_*\| \le \delta$ and $\|\boldsymbol{y} - \boldsymbol{y}_*\| \le \delta$, we have*

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le \max_{\boldsymbol{y}' \,:\, \|\boldsymbol{y}' - \boldsymbol{y}_*\| \le h(\delta)} f(\boldsymbol{x}, \boldsymbol{y}').$$

[3] Jin, Netrapalli and Jordan, What is local optimality in nonconvex-nonconcave minimax optimization?, ICML, 2020.

[4] Farnia and Ozdaglar, Do GANs always have Nash equilibria?, ICML, 2020.

Stackelberg equilibrium

- Nash may not even exist.[3][4] Any broader alternative?
- **Stackelberg** equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}'), \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$$

Definition 1 (Jin-Netrapalli-Jordan, ICML,'20)

A point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is said to be a **local minimax point** *if there exists $\delta_0 > 0$ and a function h satisfying $h(\delta) \to 0$ as $\delta \to 0$ such that, for any $\delta \in (0, \delta_0]$ and any $(\boldsymbol{x}, \boldsymbol{y})$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}_*\| \le \delta$ and $\|\boldsymbol{y} - \boldsymbol{y}_*\| \le \delta$, we have*

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le \max_{\boldsymbol{y}' \,:\, \|\boldsymbol{y}' - \boldsymbol{y}_*\| \le h(\delta)} f(\boldsymbol{x}, \boldsymbol{y}').$$

[3] Jin, Netrapalli and Jordan, What is local optimality in nonconvex-nonconcave minimax optimization?, ICML, 2020.

[4] Farnia and Ozdaglar, Do GANs always have Nash equilibria?, ICML, 2020.

Stackelberg equilibrium

- Nash may not even exist.[3][4] Any broader alternative?
- **Stackelberg** equilibrium:

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}'), \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$$

Definition 1 (Jin-Netrapalli-Jordan, ICML,'20)

A point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is said to be a **local minimax point** *if there exists $\delta_0 > 0$ and a function h satisfying $h(\delta) \to 0$ as $\delta \to 0$ such that, for any $\delta \in (0, \delta_0]$ and any $(\boldsymbol{x}, \boldsymbol{y})$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}_*\| \le \delta$ and $\|\boldsymbol{y} - \boldsymbol{y}_*\| \le \delta$, we have*

$$f(\boldsymbol{x}_*, \boldsymbol{y}) \le f(\boldsymbol{x}_*, \boldsymbol{y}_*) \le \max_{\boldsymbol{y}' \,:\, \|\boldsymbol{y}' - \boldsymbol{y}_*\| \le h(\delta)} f(\boldsymbol{x}, \boldsymbol{y}').$$

---

[3] Jin, Netrapalli and Jordan, What is local optimality in nonconvex-nonconcave minimax optimization?, ICML, 2020.

[4] Farnia and Ozdaglar, Do GANs always have Nash equilibria?, ICML, 2020.

Loose second-order necessary condition

- Second-order necessary condition: (Jin-Netrapalli-Jordan, ICML, '20)

$$\nabla f(\boldsymbol{z}_*) = \mathbf{0}, \quad \nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}_*) \preceq \mathbf{0},$$

and if $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}_*) \prec \mathbf{0}$, then in addition,

$$\boldsymbol{S}(\boldsymbol{z}_*) := \underbrace{[\nabla^2_{\boldsymbol{xx}} f - \nabla^2_{\boldsymbol{xy}} f (\nabla^2_{\boldsymbol{yy}} f)^{-1} \nabla^2_{\boldsymbol{yx}} f]}_{\text{Schur complement of } D\boldsymbol{F}}(\boldsymbol{z}_*) \succeq \mathbf{0}$$

(Loose when $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}_*)$ is not invertible...)

Restricted Schur Complement

- By similarity transform, we may assume that $\nabla^2_{\boldsymbol{yy}} f$ is diagonal such that

$$D\boldsymbol{F} = \left[\begin{array}{c|cccccc} \nabla^2_{\boldsymbol{xx}} f & & & \nabla^2_{\boldsymbol{xy}} f & & & \\ \hline & \beta_1 & & & \vdots & & \\ & & \ddots & & \vdots & & \\ -\nabla^2_{\boldsymbol{yx}} f & & & \beta_r & \vdots & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & \vdots & 0 & \\ & & & & \vdots & \ddots & \\ & & & & \vdots & & 0 \end{array}\right]$$

- Let $\boldsymbol{\Gamma}$ be the submatrix in the shaded part above.
- Let $\boldsymbol{U}$ be a matrix whose columns form an orthonormal basis of $\mathcal{R}(\boldsymbol{\Gamma})^\top$.

Definition 2 (Restricted Schur Complement, Chae-Kim-**K.**, ICLR, '24)

$$\boldsymbol{S}_{\text{res}} = \boldsymbol{U}^\top \underbrace{(\nabla^2_{\boldsymbol{xx}} f - \nabla^2_{\boldsymbol{xy}} f (\nabla^2_{\boldsymbol{yy}} f)^\dagger \nabla^2_{\boldsymbol{yx}} f)}_{(\textit{Generalized Schur complement of } D\boldsymbol{F}) =: \boldsymbol{S}} \boldsymbol{U}$$

Chae-Kim-**K.**, Two-timescale extragradient for finding local minimax points, ICLR, 2024.

Improved necessary condition and strict non-minimax point

Proposition 1 (Chae-Kim-**K.**, ICLR, '24)

$\boldsymbol{S}_{\mathrm{res}} \succeq \mathbf{0}$ *if and only if* $\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v} \geq 0$ *for any* $\boldsymbol{v} \in \mathcal{R}(\boldsymbol{U})$,

or equivalently, for any $(\nabla^2_{\boldsymbol{yx}} f)\boldsymbol{v} \in \mathcal{R}(\nabla^2_{\boldsymbol{yy}} f)$.

Proposition 2 (Second-order necessary, Chae-Kim-**K.**, ICLR, '24)

Any local minimax point satisfies $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}) \preceq \mathbf{0}$ *and if* $h(\delta)$ *satisfies* $\limsup_{\delta\to 0+} \frac{h(\delta)}{\delta} < \infty$, [a] *then* $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{z}) \succeq \mathbf{0}$.

[a] Ma, Yao, Ye and Zhang, Calm local optimality for nonconvex-nonconcave minimax problems, Set-Valued and Variational Analysis, 2025

Definition 3 (Chae-Kim-K., ICLR, '24)

A stationary point $\boldsymbol{z}$ *is said to be a* **strict non-minimax** *point if at least one of* $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{z})$ *or* $-\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z})$ *has a strictly negative eigenvalue.*

Improved necessary condition and strict non-minimax point

Proposition 1 (Chae-Kim-**K.**, ICLR, '24)

$\boldsymbol{S}_{\mathrm{res}} \succeq \mathbf{0}$ *if and only if* $\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v} \geq 0$ *for any* $\boldsymbol{v} \in \mathcal{R}(\boldsymbol{U})$,

or equivalently, for any $(\nabla^2_{\boldsymbol{y}\boldsymbol{x}} f)\boldsymbol{v} \in \mathcal{R}(\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f)$.

Proposition 2 (Second-order necessary, Chae-Kim-**K.**, ICLR, '24)

Any local minimax point satisfies $\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f(\boldsymbol{z}) \preceq \mathbf{0}$ *and if* $h(\delta)$ *satisfies* $\limsup_{\delta \to 0+} \frac{h(\delta)}{\delta} < \infty$, [a] *then* $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{z}) \succeq \mathbf{0}$.

[a] Ma, Yao, Ye and Zhang, Calm local optimality for nonconvex-nonconcave minimax problems, Set-Valued and Variational Analysis, 2025

Definition 3 (Chae-Kim-K., ICLR, '24)

A stationary point $\boldsymbol{z}$ *is said to be a* **strict non-minimax** *point if at least one of* $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{z})$ *or* $-\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f(\boldsymbol{z})$ *has a strictly negative eigenvalue.*

Improved necessary condition and strict non-minimax point

Proposition 1 (Chae-Kim-**K.**, ICLR, '24)

$\boldsymbol{S}_{\mathrm{res}} \succeq \boldsymbol{0}$ *if and only if* $\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v} \geq 0$ *for any* $\boldsymbol{v} \in \mathcal{R}(\boldsymbol{U})$,

or equivalently, for any $(\nabla^2_{\boldsymbol{yx}} f)\boldsymbol{v} \in \mathcal{R}(\nabla^2_{\boldsymbol{yy}} f)$.

Proposition 2 (Second-order necessary, Chae-Kim-**K.**, ICLR, '24)

Any local minimax point satisfies $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}) \preceq \boldsymbol{0}$ *and if* $h(\delta)$ *satisfies* $\limsup_{\delta\to 0+} \frac{h(\delta)}{\delta} < \infty$, [a] *then* $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{z}) \succeq \boldsymbol{0}$.

[a] Ma, Yao, Ye and Zhang, Calm local optimality for nonconvex-nonconcave minimax problems, Set-Valued and Variational Analysis, 2025

Definition 3 (Chae-Kim-K., ICLR, '24)

A stationary point $\boldsymbol{z}$ *is said to be a* **strict non-minimax** *point if at least one of* $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{z})$ *or* $-\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z})$ *has a strictly negative eigenvalue.*

Two-timescale GDA

- Two-timescale τ-GDA[5] for $\tau > 1$:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \frac{\eta}{\tau} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \end{aligned}$$

- In a compact form:

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta \Lambda_\tau \boldsymbol{F}(\boldsymbol{z}_k), \quad \text{where} \quad \Lambda_\tau = \begin{bmatrix} \frac{1}{\tau}\boldsymbol{I} & \\ & \boldsymbol{I} \end{bmatrix}$$

- Two-timescaled Jacobian:

$$\Lambda_\tau D\boldsymbol{F} = \begin{bmatrix} \frac{1}{\tau}\nabla^2_{\boldsymbol{xx}} f & \frac{1}{\tau}\nabla^2_{\boldsymbol{xy}} f \\ -\nabla^2_{\boldsymbol{yx}} f & -\nabla^2_{\boldsymbol{yy}} f \end{bmatrix}$$

(But why two-timescale?)

[5] Heusel, Ramsauer, Unterthiner, Nessler and Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, NeurIPS, 2017.

Two-timescale GDA

- Two-timescale τ-GDA[5] for $\tau > 1$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\eta}{\tau}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k),$$
$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \eta\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)$$

- In a compact form:

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta\Lambda_\tau \boldsymbol{F}(\boldsymbol{z}_k), \quad \text{where} \quad \Lambda_\tau = \begin{bmatrix} \frac{1}{\tau}\boldsymbol{I} & \\ & \boldsymbol{I} \end{bmatrix}$$

- Two-timescaled Jacobian:

$$\Lambda_\tau D\boldsymbol{F} = \begin{bmatrix} \frac{1}{\tau}\nabla^2_{\boldsymbol{xx}} f & \frac{1}{\tau}\nabla^2_{\boldsymbol{xy}} f \\ -\nabla^2_{\boldsymbol{yx}} f & -\nabla^2_{\boldsymbol{yy}} f \end{bmatrix}$$

(But why two-timescale?)

[5] Heusel, Ramsauer, Unterthiner, Nessler and Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, NeurIPS, 2017.

Spectrum of two-timescaled Jacobian

Lemma 1 (Jin-Netrapalli-Jordan, ICML, '20, Lemma 40)

If $\nabla^2_{\boldsymbol{yy}} f$ is invertible, *the $d_1 + d_2$ complex eigenvalues $\{\lambda_j\}$ of $\Lambda_\tau D\boldsymbol{F}$ have one of the following asymptotics as $\epsilon = \frac{1}{\tau} \to 0+$:*

$$|\lambda_j - \epsilon\mu_j| = o(\epsilon),$$
$$|\lambda_j - \nu_j| = o(1),$$

where $\{\mu_j\}$ and $\{\nu_j\}$ are the eigenvalues of $\boldsymbol{S}$ and $-\nabla^2_{\boldsymbol{yy}} f$, respectively.

(Invertibility of $\nabla^2_{\boldsymbol{yy}} f$ is too restrictive.)

Spectrum of two-timescaled Jacobian (cont'd)

Theorem 1 (Chae-Kim-**K.**, ICLR '24)

If at least one of S_{res} and $\nabla^2_{\boldsymbol{yy}} f$ is invertible, *the $d_1 + d_2$ complex eigenvalues $\{\lambda_j\}$ of $\Lambda_\tau D\boldsymbol{F}$ have one of the following asymptotics as $\epsilon = \frac{1}{\tau} \to 0+$:*

$$
\begin{aligned}
\text{(i)} \quad & |\lambda_j \pm i\sqrt{\epsilon}\sigma_j| = o(\sqrt{\epsilon}), \\
\text{(ii)} \quad & |\lambda_j - \epsilon\mu_j| = o(\epsilon), \\
\text{(iii)} \quad & |\lambda_j - \nu_j| = o(1),
\end{aligned}
$$

where $\{\sigma_j\}$ are the singular values of $\boldsymbol{\Gamma}$, $\{\mu_j\}$ are the eigenvalues of $\boldsymbol{S}_{\text{res}}$, and $\{\nu_j\}$ are the nonzero eigenvalues of $-\nabla^2_{\boldsymbol{yy}} f$.

(type (i) makes GDA unstable)

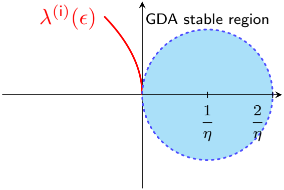
Spectrum of two-timescaled Jacobian (cont'd)

Theorem 1 (Chae-Kim-**K.**, ICLR '24)

If at least one of S_{res} and $\nabla^2_{\boldsymbol{yy}} f$ is invertible, *the $d_1 + d_2$ complex eigenvalues $\{\lambda_j\}$ of $\Lambda_\tau D\boldsymbol{F}$ have one of the following asymptotics as $\epsilon = \frac{1}{\tau} \to 0+$:*

$$
\begin{aligned}
&\text{(i)} && |\lambda_j \pm i\sqrt{\epsilon}\sigma_j| = o(\sqrt{\epsilon}), \\
&\text{(ii)} && |\lambda_j - \epsilon\mu_j| = o(\epsilon), \\
&\text{(iii)} && |\lambda_j - \nu_j| = o(1),
\end{aligned}
$$

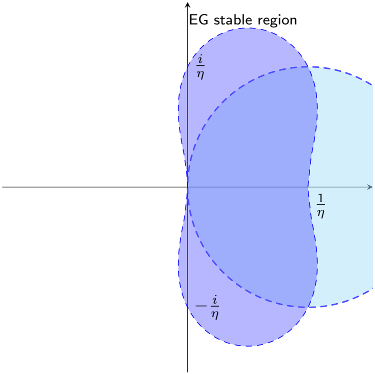
where $\{\sigma_j\}$ are the singular values of $\boldsymbol{\Gamma}$, $\{\mu_j\}$ are the eigenvalues of $\boldsymbol{S}_{\text{res}}$, and $\{\nu_j\}$ are the nonzero eigenvalues of $-\nabla^2_{\boldsymbol{yy}} f$.



(type (i) makes GDA unstable)

Extragradient

- Extragradient (EG): $\boldsymbol{z}_{k+1} = \boldsymbol{w}(\boldsymbol{z}_k) = \boldsymbol{z}_k - \eta \boldsymbol{F}(\boldsymbol{z}_k - \eta \boldsymbol{F}(\boldsymbol{z}_k))$

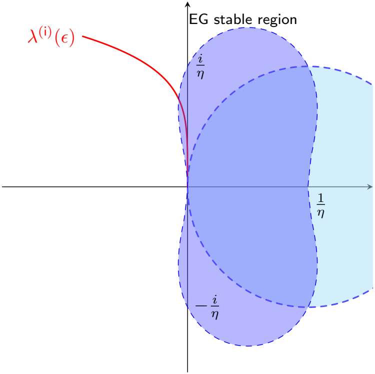


- Kyuwon Kim[6], Wed 10:30am - 11:45am
  (Alternating update + Chambolle-Pock extrapolation helps!)

[6] Kim-**K.**, Double-step alternating extragradient with increasing timescale separation for finding local minimax points: Provable improvements, ICML, 2024

Extragradient

- τ-EG: $z_{k+1} = w(z_k) = z_k - \eta\Lambda_\tau F(z_k - \eta\Lambda_\tau F(z_k))$

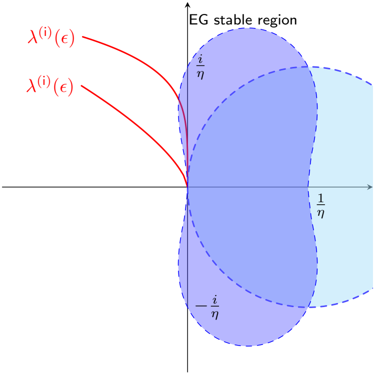


- Kyuwon Kim[6], Wed 10:30am - 11:45am
  (Alternating update + Chambolle-Pock extrapolation helps!)

[6]Kim-**K.**, Double-step alternating extragradient with increasing timescale separation for finding local minimax points: Provable improvements, ICML, 2024

Extragradient

- τ-EG: $\boldsymbol{z}_{k+1} = \boldsymbol{w}(\boldsymbol{z}_k) = \boldsymbol{z}_k - \eta\Lambda_\tau \boldsymbol{F}(\boldsymbol{z}_k - \eta\Lambda_\tau \boldsymbol{F}(\boldsymbol{z}_k))$



- Kyuwon Kim[6], Wed 10:30am - 11:45am

  (Alternating update + Chambolle-Pock extrapolation helps!)

[6]Kim-**K.**, Double-step alternating extragradient with increasing timescale separation for finding local minimax points: Provable improvements, ICML, 2024

Two-timescale method avoids strict non-minimax points

- By the stable manifold theorem, if at least of one of $\boldsymbol{S}_{\mathrm{res}}$ and $\nabla^2_{\boldsymbol{y}\boldsymbol{y}} f$ is invertible, for any **strict non-minimax** point $\tilde{z}$, τ-EG satisfies (Chae-Kim-**K.**, ICLR '24)

$$P(\lim_k \boldsymbol{z}_k = \tilde{\boldsymbol{z}}) = 0,$$

for sufficiently large τ

If τ-EG (and τ-GDA) converges to a point, then it is almost surely not a **strict non-minimax** point.

(But τ-GDA also escapes some optimal points)

Generative models and GAN

- $\mathbb{P}_{\text{true}}$ and $\mathbb{P}_\theta$: True and generated data distributions
- Generative models aim to train θ so that $\mathbb{P}_{\text{true}} \approx \mathbb{P}_\theta$.
- GAN: minimizes the distance between them via minimax optimization

Wasserstein GAN

- Wasserstein GAN[7]: minimizes the Wasserstein-1 distance

$$\begin{aligned}\min_{\theta} \mathrm{W}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_{\theta}) &= \inf_{\gamma \in \Pi(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_{\theta})} \mathbb{E}_{(x,z)\sim\gamma}[\|x - z\|] \\ &= \min_{\theta} \left(\max_{f : \|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_{\mathrm{true}}}[f(x)] - \mathbb{E}_{z\sim\mathbb{P}_{\theta}}[f(z)] \right).\end{aligned}$$

- Since this is informative even when $\mathbb{P}_{\mathrm{true}}$ and $\mathbb{P}_{\theta}$ have disjoint supports, it is considered well-suited for gradient-based training.

- Yet enforcing the constraint required heuristics, and optimization remained difficult, leading the community to favor diffusion models for their stable training.

- Avoidance in constrained problem?: projected GD may converge to a strict saddle even when there is only a single linear constraint.[8]

[7] Arjovsky, Chintala and Bottou, Wasserstein generative adversarial networks, ICML, 2017

[8] Nouiehed, Lee and Razaviyayn, Convergence to second-order stationarity for constrained non-convex optimization, arXiv, 2018.

Wasserstein GAN

- Wasserstein GAN[7]: minimizes the Wasserstein-1 distance

$$\min_\theta \mathrm{W}(\mathbb{P}_{\text{true}}, \mathbb{P}_\theta) = \inf_{\gamma \in \Pi(\mathbb{P}_{\text{true}}, \mathbb{P}_\theta)} \mathbb{E}_{(x,z)\sim\gamma}[\|x - z\|]$$
$$= \min_\theta \left(\max_{f:\|f\|_L \le 1} \mathbb{E}_{x\sim\mathbb{P}_{\text{true}}}[f(x)] - \mathbb{E}_{z\sim\mathbb{P}_\theta}[f(z)] \right).$$

- Since this is informative even when $\mathbb{P}_{\text{true}}$ and $\mathbb{P}_\theta$ have disjoint supports, it is considered well-suited for gradient-based training.

- Yet enforcing the constraint required heuristics, and optimization remained difficult, leading the community to favor diffusion models for their stable training.

- Avoidance in constrained problem?: projected GD may converge to a strict saddle even when there is only a single linear constraint.[8]

[7] Arjovsky, Chintala and Bottou, Wasserstein generative adversarial networks, ICML, 2017

[8] Nouiehed, Lee and Razaviyayn, Convergence to second-order stationarity for constrained non-convex optimization, arXiv, 2018.

Wasserstein GAN

- Wasserstein GAN[7]: minimizes the Wasserstein-1 distance

$$\min_\theta \mathrm{W}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) = \inf_{\gamma \in \Pi(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta)} \mathbb{E}_{(x,z)\sim\gamma}[\|x - z\|]$$
$$= \min_\theta \left(\max_{f : \|f\|_L \le 1} \mathbb{E}_{x\sim\mathbb{P}_{\mathrm{true}}}[f(x)] - \mathbb{E}_{z\sim\mathbb{P}_\theta}[f(z)] \right).$$

- Since this is informative even when $\mathbb{P}_{\mathrm{true}}$ and $\mathbb{P}_\theta$ have disjoint supports, it is considered well-suited for gradient-based training.

- Yet enforcing the constraint required heuristics, and optimization remained difficult, leading the community to favor diffusion models for their stable training.

- Avoidance in constrained problem?: projected GD may converge to a strict saddle even when there is only a single linear constraint.[8]

[7] Arjovsky, Chintala and Bottou, Wasserstein generative adversarial networks, ICML, 2017

[8] Nouiehed, Lee and Razaviyayn, Convergence to second-order stationarity for constrained non-convex optimization, arXiv, 2018.

Wasserstein GAN

- Wasserstein GAN[7]: minimizes the Wasserstein-1 distance

$$\min_{\theta} \mathrm{W}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_{\theta}) = \inf_{\gamma \in \Pi(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_{\theta})} \mathbb{E}_{(x,z)\sim\gamma}[\|x - z\|]$$
$$= \min_{\theta} \left(\max_{f:\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_{\mathrm{true}}}[f(x)] - \mathbb{E}_{z\sim\mathbb{P}_{\theta}}[f(z)] \right).$$

- Since this is informative even when $\mathbb{P}_{\mathrm{true}}$ and $\mathbb{P}_{\theta}$ have disjoint supports, it is considered well-suited for gradient-based training.

- Yet enforcing the constraint required heuristics, and optimization remained difficult, leading the community to favor diffusion models for their stable training.

- Avoidance in constrained problem?: projected GD may converge to a strict saddle even when there is only a single linear constraint.[8]

[7] Arjovsky, Chintala and Bottou, Wasserstein generative adversarial networks, ICML, 2017
[8] Nouiehed, Lee and Razaviyayn, Convergence to second-order stationarity for constrained non-convex optimization, arXiv, 2018.

Jensen-Shannon GAN

- Original GAN[9]: minimizes the Jensen-Shannon (JS) divergence:

$$\mathrm{JS}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) := \frac{1}{2}\mathrm{KL}\left(\mathbb{P}_{\mathrm{true}} \middle\| \frac{\mathbb{P}_{\mathrm{true}} + \mathbb{P}_\theta}{2}\right) + \frac{1}{2}\mathrm{KL}\left(\mathbb{P}_\theta \middle\| \frac{\mathbb{P}_{\mathrm{true}} + \mathbb{P}_\theta}{2}\right),$$

which is

$$\min_\theta \left(2\mathrm{JS}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) - 2\log 2\right) = \min_\theta \left(\max_f -\mathbb{E}_{x\sim\mathbb{P}_{\mathrm{true}}}[l(f(x))] - \mathbb{E}_{z\sim\mathbb{P}_\theta}[l(-f(z))]\right)$$

where $l(t) = \log(1 + \exp(-t))$ is the logistic loss.

- This is unconstrained, but suffers from a vanishing gradient issue, since for each θ, near-optimal f stays in the tail of the logistic loss.

- Can we design a loss (or distance) that mitigates vanishing gradients without imposing constraints?

[9] Goodfellow et al., Generative adversarial nets, NeurIPS, 2014.

Jensen-Shannon GAN

- Original GAN[9]: minimizes the Jensen-Shannon (JS) divergence:

$$\mathrm{JS}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) := \frac{1}{2}\mathrm{KL}\left(\mathbb{P}_{\mathrm{true}} \middle\| \frac{\mathbb{P}_{\mathrm{true}} + \mathbb{P}_\theta}{2}\right) + \frac{1}{2}\mathrm{KL}\left(\mathbb{P}_\theta \middle\| \frac{\mathbb{P}_{\mathrm{true}} + \mathbb{P}_\theta}{2}\right),$$

which is

$$\min_\theta (2\mathrm{JS}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) - 2\log 2) = \min_\theta \left(\max_f -\mathbb{E}_{x\sim\mathbb{P}_{\mathrm{true}}}[l(f(x))] - \mathbb{E}_{z\sim\mathbb{P}_\theta}[l(-f(z))]\right)$$

where $l(t) = \log(1+\exp(-t))$ is the logistic loss.

- This is unconstrained, but suffers from a vanishing gradient issue, since for each θ, near-optimal f stays in the tail of the logistic loss.

- Can we design a loss (or distance) that mitigates vanishing gradients without imposing constraints?

[9] Goodfellow et al., Generative adversarial nets, NeurIPS, 2014.

Jensen-Shannon GAN

- Original GAN[9]: minimizes the Jensen-Shannon (JS) divergence:

$$\mathrm{JS}(\mathbb{P}_{\text{true}}, \mathbb{P}_\theta) := \frac{1}{2}\mathrm{KL}\left(\mathbb{P}_{\text{true}} \middle\| \frac{\mathbb{P}_{\text{true}} + \mathbb{P}_\theta}{2}\right) + \frac{1}{2}\mathrm{KL}\left(\mathbb{P}_\theta \middle\| \frac{\mathbb{P}_{\text{true}} + \mathbb{P}_\theta}{2}\right),$$

  which is

$$\min_\theta \left(2\mathrm{JS}(\mathbb{P}_{\text{true}}, \mathbb{P}_\theta) - 2\log 2\right) = \min_\theta \left(\max_f -\mathbb{E}_{x\sim\mathbb{P}_{\text{true}}}[l(f(x))] - \mathbb{E}_{z\sim\mathbb{P}_\theta}[l(-f(z))]\right)$$

  where $l(t) = \log(1 + \exp(-t))$ is the logistic loss.

- This is unconstrained, but suffers from a vanishing gradient issue, since for each θ, near-optimal f stays in the tail of the logistic loss.

- Can we design a loss (or distance) that mitigates vanishing gradients without imposing constraints?

[9] Goodfellow et al., Generative adversarial nets, NeurIPS, 2014.

Zero-Infinity GAN

- Zero-Infinity distance:

$$\mathrm{ZI}(\mathbb{P}_{\text{true}}, \mathbb{P}_\theta) = \begin{cases} 0, & \mathbb{P}_{\text{true}} = \mathbb{P}_\theta, \\ \infty, & \text{otherwise.} \end{cases}$$

- Zero-Infinity (ZI) GAN:[10] (= WGAN w/o constraint)

$$\min_\theta \mathrm{ZI}(\mathbb{P}_{\text{true}}, \mathbb{P}_\theta) = \min_\theta \left(\max_f \mathbb{E}_{x \sim \mathbb{P}_{\text{true}}}[f(x)] - \mathbb{E}_{z \sim \mathbb{P}_\theta}[f(z)] \right)$$

- We argue that this least informative distance yields the simplest minimax loss, one that is potentially solvable by gradient methods. (But... really?)

[10] Lee-**K.**, Zero-Infinity GAN and implicit bias of extragradient (anchored at zero), 2025

Zero-Infinity GAN

- Zero-Infinity distance:

$$\mathrm{ZI}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) = \begin{cases} 0, & \mathbb{P}_{\mathrm{true}} = \mathbb{P}_\theta, \\ \infty, & \text{otherwise.} \end{cases}$$

- Zero-Infinity (ZI) GAN:[10] (= WGAN w/o constraint)

$$\min_\theta \mathrm{ZI}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) = \min_\theta \left(\max_f \mathbb{E}_{x\sim\mathbb{P}_{\mathrm{true}}}[f(x)] - \mathbb{E}_{z\sim\mathbb{P}_\theta}[f(z)] \right)$$

- We argue that this least informative distance yields the simplest minimax loss, one that is potentially solvable by gradient methods. (But... really?)

[10] Lee-**K.**, Zero-Infinity GAN and implicit bias of extragradient (anchored at zero), 2025

Zero-Infinity GAN

- Zero-Infinity distance:

$$\mathrm{ZI}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) = \begin{cases} 0, & \mathbb{P}_{\mathrm{true}} = \mathbb{P}_\theta, \\ \infty, & \text{otherwise.} \end{cases}$$

- Zero-Infinity (ZI) GAN:[10] (= WGAN w/o constraint)

$$\min_\theta \mathrm{ZI}(\mathbb{P}_{\mathrm{true}}, \mathbb{P}_\theta) = \min_\theta \left(\max_f \mathbb{E}_{x\sim\mathbb{P}_{\mathrm{true}}}[f(x)] - \mathbb{E}_{z\sim\mathbb{P}_\theta}[f(z)] \right)$$

- We argue that this least informative distance yields the simplest minimax loss, one that is potentially solvable by gradient methods. (But... really?)

[10] Lee-**K.**, Zero-Infinity GAN and implicit bias of extragradient (anchored at zero), 2025

Toy Example: Dirac GAN

- Consider the following simple setting:[11]

  True and generated data: 0 and θ in $\mathbb{R}$ (Linear generator)

  True and generated data distributions: $\mathbb{P}_{\text{true}} = \delta_0$ and $\mathbb{P}_\theta = \delta_\theta$

  Linear discriminator: $f(x) = wx$

- Dirac GAN:

$$\min_{\theta} \min_{w \in \mathcal{W}} -l(0) - l(-w\theta)$$

  W: $l(t) = -t$ and $\mathcal{W} = \{|w| \le 1\}$

  JS: $l(t) = \log(1 + \exp(-t))$ and $\mathcal{W} = \mathbb{R}$

  ZI: $l(t) = -t$ and $\mathcal{W} = \mathbb{R}$ (Unconstrained bilinear)

[11] Mescheder, Geiger and Nowozin. Which training methods for GANs do actually converge?, ICML, 2018.

Toy Example: Dirac GAN

- Consider the following simple setting:[11]

  True and generated data: 0 and θ in $\mathbb{R}$ (Linear generator)

  True and generated data distributions: $\mathbb{P}_{\text{true}} = \delta_0$ and $\mathbb{P}_\theta = \delta_\theta$

  Linear discriminator: $f(x) = wx$

- Dirac GAN:

$$\min_{\theta} \min_{w \in \mathcal{W}} -l(0) - l(-w\theta)$$

  W: $l(t) = -t$ and $\mathcal{W} = \{|w| \le 1\}$

  JS: $l(t) = \log(1 + \exp(-t))$ and $\mathcal{W} = \mathbb{R}$

  ZI: $l(t) = -t$ and $\mathcal{W} = \mathbb{R}$ (Unconstrained bilinear)

[11] Mescheder, Geiger and Nowozin. Which training methods for GANs do actually converge?, ICML, 2018.

Toy Example: Dirac GAN (cont'd)

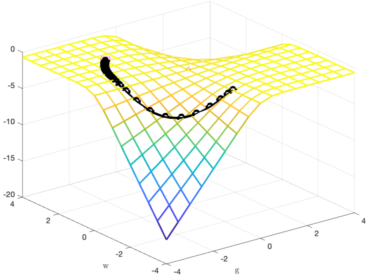


Figure: EG trajectories for Dirac GAN: (L) JS divergence and (R) ZI distance

Strict Non-Optimality in Zero-Infinity GAN

- We investigated the loss landscape of the ZIGAN for a linear generator and a two-layer **neural network** discriminator f. (Lee-**K.**, 2025)
- There are strict non-minimax points in ZIGAN, which two-timescale methods (τ-GDA and τ-EG) can almost surely escape.
- τ-GDA vs. τ-EG?: the latter locally converges to global solutions, while the former does not.

Strict Non-Optimality in Zero-Infinity GAN

- We investigated the loss landscape of the ZIGAN for a linear generator and a two-layer **neural network** discriminator f. (Lee-**K.**, 2025)
- There are strict non-minimax points in ZIGAN, which two-timescale methods (τ-GDA and τ-EG) can almost surely escape.
- τ-GDA vs. τ-EG?: the latter locally converges to global solutions, while the former does not.

Strict Non-Optimality in Zero-Infinity GAN

- We investigated the loss landscape of the ZIGAN for a linear generator and a two-layer **neural network** discriminator f. (Lee-**K.**, 2025)
- There are strict non-minimax points in ZIGAN, which two-timescale methods (τ-GDA and τ-EG) can almost surely escape.
- τ-GDA vs. τ-EG?: the latter locally converges to global solutions, while the former does not.

Conclusion

1. We defined a new strict non-optimality in minimax optimization, named strict non-minimax points, which the two-timescale gradient methods can almost surely escape.

2. We introduced the Zero-Infinity (ZI) GAN that neither requires Lipschitz contraint nor suffers from gradient vanishing.

3. We showed that the ZIGAN has strict non-minimax points.