Two-timescale Extragradient for Finding Local Minimax Points

Jiseok Chae[1] Kyuwon Kim[1] Donghwan Kim[1]

[1]Department of Mathematical Sciences, KAIST

2023 SAARC ColLabor Workshop

Contents

Problem Setting

We consider minimax problems:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \max_{\boldsymbol{y}\in\mathbb{R}^m} f(\boldsymbol{x}, \boldsymbol{y}).$$

Many ML applications of minimax problems:

- Generative Adversarial Networks
- Adversarial Learning
- Fair Classification
- Sharpness-aware Minimization

Optimality

Minimax problems have a hierarchical structure

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\left(\Phi(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{y}\in\mathbb{R}^m} f(\boldsymbol{x},\boldsymbol{y})\right).$$

The defintion of (local) optimal point $(\boldsymbol{x}^*,\boldsymbol{y}^*)$ reflects this hierarchy.

Definition 1 (Informal; Jin et al. (2020))

There exists $h:\mathbb{R}\to\mathbb{R}$ such that for any small δ,

$$f(\boldsymbol{x}^*,\boldsymbol{y}) \le f(\boldsymbol{x}^*,\boldsymbol{y}^*) \le \max_{\|\tilde{\boldsymbol{y}}-\boldsymbol{y}^*\|\le h(\delta)} f(\boldsymbol{x},\tilde{\boldsymbol{y}})$$

holds for any $\|\boldsymbol{x}-\boldsymbol{x}^*\|\le\delta$ and $\|\boldsymbol{y}-\boldsymbol{y}^*\|\le\delta$.

Contents

Two-timescale Methods

Gradient descent ascent (GDA) is a widely used simple modification of gradient descent

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

but the hierarchy of minimax problems is not well incorporated.

To put more emphasis on the maximization, one can introduce a timescale parameter $\tau \geq 1$ and consider the *two-timescale* GDA

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

Two-timescale GDA is Not Sufficient

Does the two-timescale GDA work well?

Yes and no...

Theorem 2 (Fiez and Ratliff (2021) and Jin et al. (2020))

If $\nabla^2_{\boldsymbol{yy}} f$ is negative definite, then for a sufficiently large τ, the limit points of the two-timescale GDA are the local optimum.

Local optimal points with $\nabla^2_{\boldsymbol{yy}} f \prec \mathbf{0}$ are called *strict* local minimax points.

For *non-strict* local optimal points, no convergence guarantees were known.

The Extragradient Method

Maybe using GDA is too naïve?

GDA already lacks convergence guarantees even for convex-concave problems. (Mescheder et al., 2018)

The extragradient(EG) method

$$\begin{aligned}
\boldsymbol{x}_{k+1/2} &= \boldsymbol{x}_k - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \\
\boldsymbol{y}_{k+1/2} &= \boldsymbol{y}_k + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \\
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2}) \\
\boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2})
\end{aligned}$$

is well known for its better convergence.

The Extragradient Method

Maybe using GDA is too naïve?

GDA already lacks convergence guarantees even for convex-concave problems. (Mescheder et al., 2018)

The *two-timescale* extragradient(EG) method

$$\begin{aligned} \boldsymbol{x}_{k+1/2} &= \boldsymbol{x}_k - {}^{\eta}/_{\tau} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \\ \boldsymbol{y}_{k+1/2} &= \boldsymbol{y}_k + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - {}^{\eta}/_{\tau} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2}) \\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k+1/2}, \boldsymbol{y}_{k+1/2}) \end{aligned}$$

should do a better job in finding local optimal points!

Dynamical Systems Approach

Using some notations...

$$\boldsymbol{\Lambda}_\tau := \begin{bmatrix} {}^1\!/\!_\tau \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}, \quad \boldsymbol{F}(\,\cdot\,) := \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\,\cdot\,) \\ -\nabla_{\boldsymbol{y}} f(\,\cdot\,) \end{bmatrix}, \quad \boldsymbol{z} := \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$$

... two-timescale EG can be viewed as a dynamical system :

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

Fact 3 (Galor (2007, Theorem 4.8))

For an equilibrium point $\boldsymbol{z}^*$ of a dynamical system $\boldsymbol{z}_{k+1} = \boldsymbol{w}(\boldsymbol{z}_k)$, the iterates locally converge to $\boldsymbol{z}^*$ if the Jacobian matrix $D\boldsymbol{w}(\boldsymbol{z}^*)$ has spectral radius smaller than 1, *i.e.*, $\rho(D\boldsymbol{w}(\boldsymbol{z}^*)) < 1$.

Asymptotic Behavior of Eigenvalues

Fact 4 (Chae, Kim, and Kim (2024))

The condition for $\rho(D\boldsymbol{w}_\tau(\boldsymbol{z}^*)) < 1$ can be characterized by the spectrum of $\boldsymbol{\Lambda}_\tau D\boldsymbol{F}(\boldsymbol{z}^*)$.

Theorem 5 (Chae, Kim, and Kim (2024))

For $\epsilon := {}^1\!/_\tau$, the complex eigenvalues $\lambda_j(\epsilon)$ of $\boldsymbol{\Lambda}_\tau D\boldsymbol{F}(\boldsymbol{x}^*, \boldsymbol{y}^*)$ have one of the following three asymptotics as $\tau \to +\infty$:

(i) $|\lambda_j(\epsilon) \pm i\sigma_j\sqrt{\epsilon}| = o(\sqrt{\epsilon})$,

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

(iii) $|\lambda_j(\epsilon) - \nu_j| = o(1)$.

Here, σ_j, ν_j, and μ_j are some values determined by $D\boldsymbol{F}(\boldsymbol{x}^*, \boldsymbol{y}^*)$.

Limit Points of the Two-timescale EG

Theorem 6

For an equilibrium point $\boldsymbol{z}^*$ of the two-timescale EG

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

TFAE:

- $\boldsymbol{S}_{\text{res}}(D\boldsymbol{F}(\boldsymbol{z}^*)) \succeq \mathbf{0}$, $\nabla^2_{\boldsymbol{yy}} f \preceq \mathbf{0}$, and $s_0 < \frac{1}{2L}$.
- There exists a number $0 < \eta^\star < \frac{1}{L}$ such that the two-timescale EG locally converges to $\boldsymbol{z}^*$ for all sufficiently large τ and $\eta^\star < \eta < \frac{1}{L}$.

Here, s_0 is a scalar depending on $D\boldsymbol{F}(\boldsymbol{z}^*)$, and $\boldsymbol{S}_{\text{res}}$ is a matrix-valued function. (Details omitted.)

Relating to Local Optimality

What points satisfy $\boldsymbol{S}_{\text{res}}(D\boldsymbol{F}(\boldsymbol{z}^*)) \succeq \mathbf{0}$ and $\nabla^2_{\boldsymbol{yy}} f \preceq \mathbf{0}$?

Proposition 7 (Chae, Kim, and Kim (2024))

If $f \in \mathcal{C}^2$, then any local optimal point satisfies $\nabla^2_{\boldsymbol{yy}} f \preceq \mathbf{0}$.

If we further assume that $\limsup_{\delta \to 0} {}^{h(\delta)}/_{\delta} < \infty$, then any local optimal point satisfies $\boldsymbol{S}_{\text{res}}(D\boldsymbol{F}(\boldsymbol{z}^*)) \succeq \mathbf{0}$.

Therefore, under those mild conditions...

Two-timescale EG can find non-strict *local optimal points.*

This is the first ever known convergence result of a first-order method to non-strict local optimal points.

Thank you for your attention.

References I

Chae, Jiseok, Kyuwon Kim, and Donghwan Kim (2024). "Two-timescale Extragradient for Finding Local Minimax Points". In: *The Twelfth International Conference on Learning Representations*.

Fiez, Tanner and Lillian J. Ratliff (2021). "Local convergence analysis of gradient descent ascent with finite timescale separation". In: *International Conference on Learning Representations*.

Galor, Oded (2007). *Discrete dynamical systems*. Springer-Verlag.

Jin, Chi, Praneeth Netrapalli, and Michael Jordan (2020). "What is local optimality in nonconvex-nonconcave minimax optimization?" In: *International Conference on Machine Learning*. PMLR, pp. 4880–4889.

Mescheder, Lars, Andreas Geiger, and Sebastian Nowozin (2018). "Which training methods for GANs do actually converge?" In: *International Conference on Machine Learning*. PMLR, pp. 3481–3490.