AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

April 23, 2026

Outline

The Minimax Problem

Formulation:

$$\min_{x \in X} \max_{y \in Y} f(x, y).$$

Interpretation. Two-player zero-sum game:

- Player x minimizes; player y maximizes.
- Leader chooses x first, and the follower choose y accordingly.

This template is ubiquitous in:

- game theory and economics.
- robust optimization and control.
- Lagrangian duality (constraints as adversarial players).
- GANs, adversarial training, DRO, etc.

Rock–Paper–Scissors

Payoff matrix: row player x minimizes, column player y maximizes

$$f(x,y) = x^\top A y, \qquad A = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}.$$

Rows or columns = rock, paper, scissors.

Pure strategies: $x, y \in \{e_1, e_2, e_3\}$, nonconvex choice set.

Mixed strategies: $x, y \in \Delta^3 = \{p \in \mathbb{R}^3_+ : \sum_i p_i = 1\}$, convex simplex.
Each entry of x (resp. y) is the probability of playing that action.

Generative Adversarial Networks (GANs)

Setup:

- generator G_x maps noise $z \sim p_z$ to samples $G_x(z)$;
- discriminator D_y maps data to $[0, 1]$ (real vs. fake).

Minimax objective:

$$\min_x \max_y \; \mathbb{E}_{\xi \sim p_{\text{data}}}[\log D_y(\xi)] \; + \; \mathbb{E}_{z \sim p_z}[\log(1 - D_y(G_x(z)))].$$

Challenges:

- f is nonconvex-nonconcave in (x, y), no saddle point theory applies.
- Training is notoriously unstable; gradient methods cycle or diverge.
- Motivates many of the algorithmic questions in this lecture.

Adversarial Training

Robust classification: [Madry et al., 2017]

$$\min_x \mathbb{E}_{(\xi,c)\sim p_{\text{data}}} \left[\max_{\|\delta\|\leq\epsilon} \ell(h_x(\xi+\delta),\, c) \right].$$

Interpretation.

- Inner maximization: worst-case adversarial perturbation δ.
- Outer minimization: train classifier h_x to resist the worst case.

Structure. Inner problem is nonconcave but often approximately solvable via PGD.

Saddle Points & Duality Gap

Definition: (x^*, y^*) is a saddle point of f on $X \times Y$ if

$$f(x^*, y) \;\le\; f(x^*, y^*) \;\le\; f(x, y^*), \qquad \forall x \in X,\, y \in Y.$$

Equivalent formulation: (x^*, y^*) is a saddle point iff

$$x^* \in \arg\min_x \max_y f(x, y), \quad y^* \in \arg\max_y \min_x f(x, y),$$

and both values coincide.

Duality gap: for any $(\hat{x}, \hat{y}) \in X \times Y$,

$$\mathrm{gap}(\hat{x}, \hat{y}) := \max_y f(\hat{x}, y) - \min_x f(x, \hat{y}) \ge 0.$$

Convergence criterion: $(\hat{x}, \hat{y})$ is ε-saddle if $\mathrm{gap}(\hat{x}, \hat{y}) \le \varepsilon$.

Weak Duality

Claim. For any $f : X \times Y \to \mathbb{R}$,

$$\max_{y \in Y} \min_{x \in X} f(x, y) \leq \min_{x \in X} \max_{y \in Y} f(x, y).$$

Proof. Fix any (x, y). Then

$$\min_{x' \in X} f(x', y) \leq f(x, y) \leq \max_{y' \in Y} f(x, y').$$

Taking $\max_y$ on the left and $\min_x$ on the right:

$$\max_{y} \min_{x'} f(x', y) \leq \min_{x} \max_{y'} f(x, y').$$

Remark. The inequality can be strict. When does equality hold?

Convex-Concave Functions

Definition: $f : X \times Y \to \mathbb{R}$ is convex-concave if

- $f(\cdot, y)$ is convex on X for every fixed $y \in Y$.
- $f(x, \cdot)$ is concave on Y for every fixed $x \in X$.

Strongly convex-strongly concave (μ-SC-SC):

- $f(\cdot, y)$ is μ-strongly convex on X for every y.
- $f(x, \cdot)$ is μ-strongly concave on Y for every x.

Examples

- Bilinear: $f(x, y) = x^\top A y + b^\top x - c^\top y$ is C-C but not SC-SC.
- Lagrangian of a convex program with equality constraints is C-C in the primal–dual variables.

Sion's Minimax Theorem

Sion's Minimax Theorem [Sion, 1958]

Let X, Y be convex and compact, and let f be convex-concave and continuous. Then

$$\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y),$$

and a saddle point exists.

Consequence: under C-C with compactness, the minimax problem is well-posed, and

- solving $\min_x \max_y f$ is equivalent to finding a saddle point;
- the gap function is a legitimate convergence measure.

SC-SC case: under μ-SC-SC, the saddle point is unique, even without compactness.

Monotone Operator View

Define the saddle operator

$$F(z) \;=\; \begin{pmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{pmatrix}, \qquad z = (x,y).$$

Connection: saddle points are zeros of F and (x^*, y^*) is a saddle iff $F(z^*) = 0$.

Monotonicity properties:

- f is C-C $\implies$ F is monotone: $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq 0$.
- f is μ-SC-SC $\implies$ F is μ-strongly monotone.
- f is ℓ-smooth $\implies$ F is ℓ-Lipschitz.

Conclusion: minimax algorithms can be analyzed as methods for solving the variational inequality $\langle F(z^*), z - z^* \rangle \geq 0$.

Outline

Minimax Foundations

Gradient Descent Ascent (GDA)

Extragradient, OGDA, Beyond C-C

Bilevel Optimization

Gradient Descent Ascent (GDA)

GDA Update:

$$x_{t+1} = \Pi_X(x_t - \eta \nabla_x f(x_t, y_t)),$$
$$y_{t+1} = \Pi_Y(y_t + \eta \nabla_y f(x_t, y_t)).$$

Saddle-operator form:

$$z_{t+1} = \Pi_Z(z_t - \eta F(z_t)), \qquad Z = X \times Y.$$

Two variants.

- Simultaneous (Sim-GDA): update x and y in parallel.
- Alternating (Alt-GDA): update y using the fresh x_{t+1}.

In this lecture, we analyze Sim-GDA throughout.

GDA Convergence: Lipschitz C-C

Convergence of GDA for Lipschitz C-C functions

Suppose f is C-C and L-Lipschitz; X, Y have diameter D. With $\eta = \frac{D}{L\sqrt{T}}$, the averaged iterates $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$, $\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$ satisfy

$$\mathrm{gap}(\bar{x}_T, \bar{y}_T) \;\le\; \frac{2LD}{\sqrt{T}}.$$

Rate. $O(1/\sqrt{T})$ for the average-iterate gap.

Consequence. To reach $\mathrm{gap} \le \varepsilon$, we need $T = O(\varepsilon^{-2})$ iterations.

Proof Sketch

Starting from C-C:

$$f(x_t, y) - f(x, y_t) \leq \langle \nabla_y f(x_t, y_t), y - y_t \rangle + \langle \nabla_x f(x_t, y_t), x_t - x \rangle.$$

By the projection inequality $\langle z_{t+1} - (z_t - \eta F(z_t)), z - z_{t+1} \rangle \geq 0$:

$$\langle F(z_t), z_t - z \rangle \leq \frac{1}{2\eta} [\|z_t - z\|^2 - \|z_{t+1} - z\|^2] + \frac{\eta}{2} \|F(z_t)\|^2 .$$

Summing $t = 1, \ldots, T$ and using $\|F(z_t)\| \leq \sqrt{2}L$:

$$\frac{1}{T} \sum_{t=1}^{T} [f(x_t, y) - f(x, y_t)] \leq \frac{D^2}{\eta T} + \eta L^2.$$

Jensen's inequality & tuning $\eta = \frac{D}{L\sqrt{T}}$ yields the bound.

GDA Convergence: Smooth C-C

Convergence of GDA for smooth C-C functions

If f is C-C and ℓ-smooth, X, Y have diameter D. With $\eta = \frac{D}{G\sqrt{T}}$ where $G = 2\ell D + \|\nabla f(x_0, y_0)\|$,

$$\mathrm{gap}(\bar{x}_T, \bar{y}_T) \;\leq\; \frac{2GD}{\sqrt{T}}.$$

Proof idea: under smoothness,

$$\|\nabla f(x, y) - \nabla f(x_0, y_0)\| \;\leq\; 2\ell D,$$

so ∇f is bounded along the trajectory, reducing to the Lipschitz case.

Observation: smoothness gives no improvement. Why?

GDA Fails on Bilinear Problems

Example:

$$\min_{x\in[-1,1]} \max_{y\in[-1,1]} xy.$$

The unique saddle is $(0,0)$.

GDA dynamics:

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & -\eta \\ \eta & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix}.$$

Eigenvalues: $1 \pm i\eta$, with $|1 \pm i\eta| = \sqrt{1+\eta^2} > 1$.

Conclusion: starting from any point $\neq (0,0)$, the iterates form expanding spirals and hit the boundary. The average iterate does converge, but the last iterate diverges.

Lesson: C-C is not enough for last-iterate convergence. We need more structure, or a smarter algorithm.

GDA Convergence: SC-SC

Convergence of GDA for smooth SC-SC functions

If f is ℓ-smooth and μ-SC-SC, then with $\eta = \mu/\ell^2$, GDA satisfies

$$\|z_t - z^*\|^2 \leq \left(1 - \frac{1}{\kappa^2}\right)^t \|z_0 - z^*\|^2, \qquad \kappa = \ell/\mu.$$

Gap bound:

$$\mathrm{gap}(x_t, y_t) \leq \ell \left(1 - \tfrac{1}{\kappa^2}\right)^t \|z_0 - z^*\|^2.$$

Iteration complexity: to reach $\mathrm{gap} \leq \varepsilon$: $T = O(\kappa^2 \log(1/\varepsilon))$.

This rate is tight: consider $f(x, y) = \ell xy + \frac{\mu}{2}x^2 - \frac{\mu}{2}y^2$ with $\mu \ll \ell$.

Proof Sketch

Strong monotonicity of F: if f is μ-SC-SC, then

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu \left\| z_1 - z_2 \right\|^2 .$$

Lipschitzness of F: if f is ℓ-smooth, then

$$\left\| F(z_1) - F(z_2) \right\| \leq \ell \left\| z_1 - z_2 \right\| .$$

Consider the unconstrained setting. Using $F(z^*) = 0$:

$$\begin{aligned} \left\| z_{t+1} - z^* \right\|^2 = \left\| z_t - z^* \right\|^2 - 2\eta \langle F(z_t) - F(z^*), z_t - z^* \rangle \\ + \eta^2 \left\| F(z_t) - F(z^*) \right\|^2 . \end{aligned}$$

Since F is Lipschitz and strongly monotone:

$$\left\| z_{t+1} - z^* \right\|^2 \leq (1 - 2\eta\mu + \eta^2 \ell^2) \left\| z_t - z^* \right\|^2 .$$

Choose $\eta = \mu/\ell^2$; then

$$\left\| z_{t+1} - z^* \right\|^2 \leq \left(1 - \tfrac{\mu^2}{\ell^2}\right) \left\| z_t - z^* \right\|^2 = \left(1 - \tfrac{1}{\kappa^2}\right) \left\| z_t - z^* \right\|^2 .$$

Outline

Minimax Foundations

Gradient Descent Ascent (GDA)

Extragradient, OGDA, Beyond C-C

Bilevel Optimization

Extragradient Method (EG)

Idea: before stepping, peek at the gradient at a "look-ahead" point.

Update: using saddle-operator F,

$$z_{t+1/2} = \Pi_Z\big(z_t - \eta\, F(z_t)\big),$$
$$z_{t+1} = \Pi_Z\left(z_t - \eta\, F\left(z_{t+1/2}\right)\right).$$

Interpretation:

- take an extrapolation step to get $z_{t+1/2}$;
- from the original point z_t, step using the gradient at $z_{t+1/2}$;
- two gradient evaluations per iteration.

EG Convergence: C-C

Convergence of EG for C-C functions

Suppose f is C-C and ℓ-smooth, $Z = X \times Y$ has diameter D. With $\eta = 1/\ell$,

$$\operatorname{gap}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t+1/2},\ \frac{1}{T}\sum_{t=1}^{T} y_{t+1/2} \right) \leq \frac{\ell D^2}{2T}.$$

Rate: $O(1/T)$ on the half-iterate average, a quadratic improvement over GDA's $O(1/\sqrt{T})$.

Iteration complexity: $T = O(\varepsilon^{-1})$ to reach $\operatorname{gap} \leq \varepsilon$.

Reason: the look-ahead step corrects the rotational component that GDA amplifies.

EG Proof Sketch

Starting point. By C-C and convexity:

$$f(x_{t+1/2}, y) - f(x, y_{t+1/2}) \leq \langle F(z_{t+1/2}), z_{t+1/2} - z \rangle.$$

Decomposition: using the update rule:

$$\langle F(z_{t+1/2}), z_{t+1/2} - z \rangle = (a) + (b) + (c),$$

where

$$(a) = \tfrac{1}{2\eta}[\|z_t - z\|^2 - \|z_t - z_{t+1}\|^2 - \|z_{t+1} - z\|^2],$$

$$(b) = \tfrac{1}{2\eta}\left[\|z_t - z_{t+1}\|^2 - \left\|z_t - z_{t+1/2}\right\|^2 - \left\|z_{t+1/2} - z_{t+1}\right\|^2\right],$$

$$(c) = \langle F(z_t) - F(z_{t+1/2}), z_{t+1/2} - z_{t+1} \rangle.$$

By smoothness: $(c) \leq \tfrac{\ell}{2}\left(\left\|z_t - z_{t+1/2}\right\|^2 + \left\|z_{t+1/2} - z_{t+1}\right\|^2)\right)$.

Summing: $(a) + (b) + (c) \leq \tfrac{1}{2\eta}[\|z_t - z\|^2 - \|z_{t+1} - z\|^2]$. Telescope.

EG: SC-SC

Convergence of EG for SC-SC functions

If f is ℓ-smooth and μ-SC-SC, with $\eta = \Theta(1/\ell)$, EG achieves

$$\|z_t - z^*\|^2 \leq \left(1 - \frac{1}{\kappa}\right)^t \|z_0 - z^*\|^2 .$$

Iteration complexity: $T = O(\kappa \log(1/\varepsilon))$.

Comparison:

- GDA: $O(\kappa^2 \log(1/\varepsilon))$.
- EG: $O(\kappa \log(1/\varepsilon))$, tight.
- EG matches the rate of standard GD on μ-SC functions.

Conclusion: the look-ahead step buys us a factor of κ in the condition number, and buys last-iterate convergence on bilinear problems.

Optimistic Gradient Descent Ascent (OGDA)

Motivation: EG requires two gradient evaluations per step. Can we achieve the same rates with one?

OGDA update:

$$z_{t+1} = \Pi_Z \left(z_t - 2\eta\, F(z_t) + \eta\, F(z_{t-1}) \right).$$

Interpretation:

- A "negative momentum" term: $-\eta F(z_{t-1})$.
- Predicts the next gradient using the previous one.
- Only one new gradient evaluation per step, reusing $F(z_{t-1})$.

Origin: online learning [Rakhlin and Sridharan, 2013]; later applied to GANs [Daskalakis et al., 2017].

OGDA & Proximal Point Method

Proximal point method:

$$z_{t+1} = z_t - \eta F(z_{t+1}).$$

EG approximates $F(z_{t+1})$ by $F(z_{t+1/2})$ where $z_{t+1/2} = z_t - \eta F(z_t)$.

OGDA approximates $F(z_{t+1})$ by $2F(z_t) - F(z_{t-1})$.

Both EG and OGDA can be analyzed under a unified proximal-point template, giving the same rates [Mokhtari et al., 2020].

Convergence Rate of OGDA

Convergence rate OGDA for smooth C-C functions

If f is C-C and ℓ-smooth, Z has diameter D, then OGDA with $\eta = \Theta(1/\ell)$ satisfies

$$\mathrm{gap}(\bar{x}_T, \bar{y}_T) \;\leq\; O\Big(\frac{\ell D^2}{T}\Big)\,.$$

Convergence rate OGDA for smooth SC-SC functions

If f is ℓ-smooth and μ-SC-SC, OGDA satisfies

$$\|z_t - z^*\|^2 \;\leq\; \Big(1 - \frac{1}{\kappa}\Big)^t \,\|z_0 - z^*\|^2\,.$$

OGDA rate matches EG at half the gradient cost per iteration.

GDA vs. EG vs. OGDA

	GDA	EG	OGDA
C-C, smooth (avg)	$O(1/\sqrt{T})$	$O(1/T)$	$O(1/T)$
SC-SC linear	$(1-1/\kappa^2)^t$	$(1-1/\kappa)^t$	$(1-1/\kappa)^t$
Bilinear (last)	diverges	converges	converges
Gradients / step	1	2	1

Observation:

- EG and OGDA are the workhorses for C-C and SC-SC minimax;
- OGDA is often preferred in practice: same rates, cheaper per step;
- both have strong last-iterate convergence guarantees.

Beyond Convex-Concave

Motivation. Modern ML is not convex-concave:

- GANs: nonconvex-nonconcave.
- Adversarial training: nonconvex in x, constrained nonconcave in y.

Hierarchy of tractability.

- SC-SC: linear convergence.
- C-C: sublinear convergence.
- NC-SC: stationarity in $O(\varepsilon^{-2})$.
- NC-C: stationarity in $O(\varepsilon^{-6})$.
- NC-NC: intractable in general (Daskalakis et al. 2021).

Nonconvex-Concave Minimax

Problem: f nonconvex in x, concave in y, and Y convex bounded.

$$\min_x \max_{y \in Y} f(x, y)$$

Convergence criterion: stationarity of

$$\Phi(x) := \max_{y \in Y} f(x, y).$$

Goal: find x with $\|\nabla \Phi(x)\|^2 \leq \varepsilon$, or a Moreau-envelope stationary point.

Danskin Theorem: if $f(x, \cdot)$ is μ-strongly concave, Φ is smooth and $\nabla \Phi(x) = \nabla_x f(x, y^*(x))$.

Two-Timescale GDA

Idea. Use different step sizes for x and y [Lin et al., 2020]:

$$x_{t+1} = x_t - \eta_x \nabla_x f(x_t, y_t)$$
$$y_{t+1} = \Pi_Y(y_t + \eta_y \nabla_y f(x_t, y_t))$$

with $\eta_x \ll \eta_y$.

Rationale. Inner y updates are much faster $\Rightarrow y_t \approx y^*(x_t) \Rightarrow$ outer loop behaves like GD on Φ.

Theorem (informal).

- NC-SC: complexity $\tilde{O}(\kappa^2 \varepsilon^{-2})$.
- NC-C: complexity $\tilde{O}(\varepsilon^{-6})$.

Outline

Minimax Foundations

Gradient Descent Ascent (GDA)

Extragradient, OGDA, Beyond C-C

Bilevel Optimization

Wrap-up and Next Week

Week 13: Key takeaways

-

Thank you for your attendance!

Questions?

Next week (Week 14): zeroth-order method & final review

References I


Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2017).
Training gans with optimism.
arXiv preprint arXiv:1711.00141.


Lin, T., Jin, C., and Jordan, M. (2020).
On gradient descent ascent for nonconvex-concave minimax problems.
In *International conference on machine learning*, pages 6083–6093. PMLR.


Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017).
Towards deep learning models resistant to adversarial attacks.
arXiv preprint arXiv:1706.06083.

References II


Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2020).
A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach.
In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR.


Rakhlin, S. and Sridharan, K. (2013).
Optimization, learning, and games with predictable sequences.
Advances in Neural Information Processing Systems, 26.


Sion, M. (1958).
On general minimax theorems.
Pacific Journal of Mathematics, 8(1):171–176.