Optimization for Data Science
ETH Zürich, FS 2022 261-5110-00L

Lecture 14: Min-Max Optimization, Part II

(will not be covered in Exam)

Bernd Gärtner
Niao He

https://www.ti.inf.ethz.ch/ew/courses/ODS22/index.html

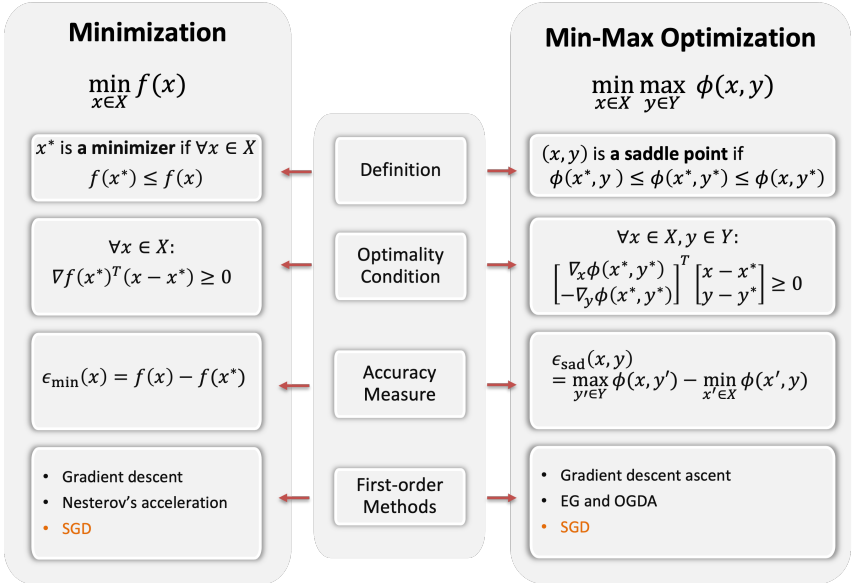
May 30, 2022

Recap

“ Minimization is to current AI,
_____ is to future AI”

Min-max Optimization
Equilibrium Learning

Recap: Minimization vs. Minimax Optimization



Recap: Unified Variational Inequality Framework

Variational Inequality Problem VI $(\mathcal{Z}, F)$

Find $\mathbf{z}^* \in \mathcal{Z}$ such that $\langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0$ for all $\mathbf{z} \in \mathcal{Z}$.

- Convex minimization:
$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \Longleftrightarrow \mathsf{VI}(\mathcal{Z}, F) \text{ with } \mathbf{z} = \mathbf{x}, \mathcal{Z} = \mathcal{X}, F(\cdot) = \nabla f(\cdot)$$

- Convex-concave min-max optimization:
$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}) \Longleftrightarrow \mathsf{VI}(\mathcal{Z}, F) \text{ with } \mathbf{z} = [\mathbf{x}; \mathbf{y}], \mathcal{Z} = \mathcal{X} \times \mathcal{Y}, F(\cdot) = [\nabla_{\mathbf{x}} \phi(\cdot); -\nabla_{\mathbf{y}} \phi(\cdot)]$$

- Concave games:
$$\min_{\mathbf{x}_i \in \mathcal{X}_i} u_i(\mathbf{x}_i, \mathbf{x}_{-i}), \forall i \Longleftrightarrow \mathsf{VI}(\mathcal{Z}, F) \text{ with } \mathbf{z} = [\mathbf{x}_i]_{i=1}^n, \mathcal{Z} = \prod_i \mathcal{X}_i, F(\mathbf{z}) = [\nabla_{\mathbf{x}_i} u(\mathbf{x}_i, \mathbf{x}_{-i})]_{i=1}^n$$

Recap: Algorithms and Convergence

Setting (using constant stepsize)	**GDA**	**OGDA and EG**
Convex-Concave	non-convergence	$O\left(\frac{L}{t}\right)$
Strongly-Convex-Strongly-Concave	$O\left((1-\frac{4\mu^2}{L^2})^t\right)$	$O\left((1-\frac{\mu}{4L})^t\right)$

- GDA: gradient descent ascent
- EG: extragradient method
- OGDA: optimistic GDA

Min-Max Optimization

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

Q: Going beyond the golden convex-concave regime, what do we know about it?

Global Solutions

$(\mathbf{x}^*, \mathbf{y}^*)$ is a global saddle point if

$$\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \phi(\mathbf{x}, \mathbf{y}^*),$$

for any $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$.

- Game interpretation: Nash equilibrium
- No player has the incentive to make unilateral change at NE.
- Simultaneous game

$(\mathbf{x}^*, \mathbf{y}^*)$ is a global minimax point if

$$\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}' \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y}'),$$

for any $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$.

- Game interpretation: Stackelberg equilibrium
- Best response to the best response.
- Sequential game

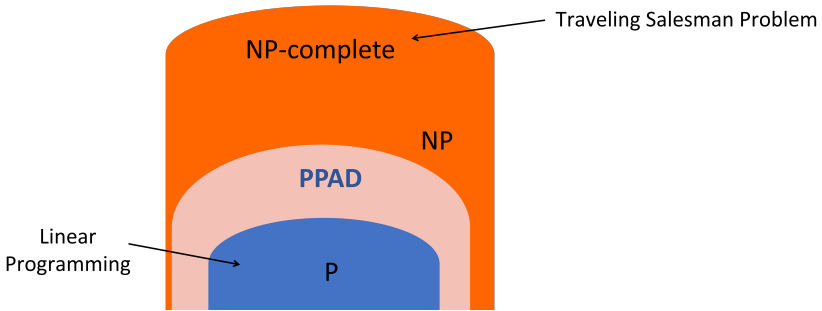
Global Solutions

- Global saddle point (Nash equilibrium) may not always exist.
  - Recall the example of $\phi(x, y) = (x - y)^2, x, y \in [-1, 1]$.
- Even if a global saddle point exists, finding it is NP-hard in general.
  - This is proven to be PPAD-complete in [Daskalakis et al. 2008].
- Global minimax point (Stackelberg equilibrium) exists under mild conditions, but finding it is also NP-hard in general.
  - This is because even the nonconvex minimization is already NP-hard.

PPAD

PPAD (Polynomial Parity Arguments on Directed graphs) is a complexity class introduced by Christos Papadimitriou in 1994. List of PPAD-complete problems:

- finding Nash equilibria for 2-player game
- computing approximate Brouwer fixed points of Lipschitz functions
- computing mixed Nash equilibria in normal-form games



Local Surrogate Solutions

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})$$

Q: What is a good notion of local optimality for nonconvex min-max optimization?

For simplicity, focus on unconstrained setting.

Local Solutions

$(\mathbf{x}^*, \mathbf{y}^*)$ is a local saddle point if

$$\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \phi(\mathbf{x}, \mathbf{y}^*),$$

in the neighborhood of $(\mathbf{x}^*, \mathbf{y}^*)$.

$(\mathbf{x}^*, \mathbf{y}^*)$ is a local minimax point if

$$\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}' : \|\mathbf{y}' - \mathbf{y}^*\| \leq \delta} \phi(\mathbf{x}, \mathbf{y}'),$$

in the neighborhood of $(\mathbf{x}^*, \mathbf{y}^*)$.

Note: Local (resp. global) saddle point is a local (resp. global) minimax point.

Local Solutions

- Local saddle point implies that (for unconstrained problems)
  - $\nabla_{\mathbf{x}}\phi(\mathbf{x}^*, \mathbf{y}^*) = 0, \nabla_{\mathbf{y}}\phi(\mathbf{x}^*, \mathbf{y}^*) = 0.$
  - $\nabla^2_{\mathbf{xx}}\phi(\mathbf{x}^*, \mathbf{y}^*) \succeq 0, \nabla^2_{\mathbf{yy}}\phi(\mathbf{x}^*, \mathbf{y}^*) \preceq 0.$

- Local minimax point implies that (for unconstrained problems)
  - $\nabla_{\mathbf{x}}\phi(\mathbf{x}^*, \mathbf{y}^*) = 0, \nabla_{\mathbf{y}}\phi(\mathbf{x}^*, \mathbf{y}^*) = 0.$
  - $\nabla^2_{\mathbf{yy}}\phi(\mathbf{x}^*, \mathbf{y}^*) \preceq 0.$
  - If $\nabla^2_{\mathbf{yy}}\phi(\mathbf{x}^*, \mathbf{y}^*) \prec 0$, then $\nabla^2_{\mathbf{xx}}\phi(\mathbf{x}^*, \mathbf{y}^*) - \left[\nabla^2_{\mathbf{xy}}\phi(\nabla^2_{\mathbf{yy}}\phi)^{-1}\nabla^2_{\mathbf{yx}}\phi\right](\mathbf{x}^*, \mathbf{y}^*) \succeq 0.$

Caveat: Local saddle point or local minimax point may not exist! [Jin et al, 2020]

- Counterexample: consider $\phi(x, y) = \sin(x + y)$.
- Counterexample: consider $\phi(x, y) = y^2 - 2xy$ on $[-1, 1] \times [-1, 1]$.

First-order Stationary Points

- ϵ-approximate first-order local Nash equilibrium: $(\mathbf{x}^*, \mathbf{y}^*)$ such that

$$\|\nabla_{\mathbf{x}} \phi(\mathbf{x}^*, \mathbf{y}^*)\| \leq \epsilon, \qquad \|\nabla_{\mathbf{y}} \phi(\mathbf{x}^*, \mathbf{y}^*)\| \leq \epsilon.$$

Fundamental hardness:

> [Daskalakis, Skoulakis & Zampetakis STOC'21]: For constrained min-max optimization, any gradient-based method needs exponentially many queries in the dimension and $1/\epsilon$ to compute even an ϵ-approximate first-order local Nash equilibrium.

Note: this is fundamentally different from nonconvex minimization.

Nonconvex-Nonconcave Min-Max Optimization

In general, even finding a first-order local Nash equilibrium is NP-hard.

Q: What should be a reasonable goal here?

Well, either change the assumption or relax the goal:

- Weaker solutions such as stable points of gradient dynamics?
- Problems with one-sided convexity?
- Problems with benign nonconvex-nonconcavity?
  (Lojasiewicz conditions, pseudo-monotonicity, variational coherence, interaction dominant regime, etc.)

GDA Dynamics

Gradient Descent Ascent (GDA)

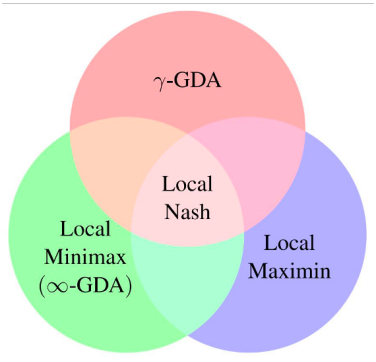
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_x \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)$$
$$\mathbf{y}_{t+1} = \mathbf{y}_t + \eta_y \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t)$$

γ-GDA flow (let $\gamma = \eta_y / \eta_x$ and $\eta_y \to 0$):

$$\frac{d\mathbf{x}}{dt} = -\frac{1}{\gamma} \nabla_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y})$$
$$\frac{d\mathbf{y}}{dt} = \nabla_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y})$$

Where does GDA dynamics converge to?

[Daskalakis and Panageas, 2018; Jin et al., 2020; Fiez and Ratli, 2020]



- Here γ-GDA stands for the linearly stable points at which the Jacobian matrix $J_\gamma = \begin{bmatrix} -\nabla^2_{\mathbf{xx}}\phi(\mathbf{x},\mathbf{y})/\gamma & -\nabla^2_{\mathbf{xy}}\phi(\mathbf{x},\mathbf{y})/\gamma \\ \nabla^2_{\mathbf{yx}}\phi(\mathbf{x},\mathbf{y}) & \nabla^2_{\mathbf{yy}}\phi(\mathbf{x},\mathbf{y}) \end{bmatrix}$ has negative real eigenvalues.

Problems with One-sided Convexity

[Lots of literature]

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \phi(\mathbf{x},\mathbf{y})$$

- $\phi(\mathbf{x},\mathbf{y})$ is L-smooth;
- $\phi(\mathbf{x},\mathbf{y})$ is concave or strongly concave in $\mathbf{y}$ (resp. NC-C or NC-SC setting).

[Lin, Jin, Jordan, ICML'20] With proper choice of stepsize, GDA converges to an ϵ-stationary point such that $\|\nabla\bar{\phi}(\mathbf{x})\| \leq \epsilon$, with at most

$$T(\epsilon) = \begin{cases} O(\kappa^2/\epsilon^2) & \text{NC-SC} \\ O(\epsilon^{-6}) & \text{NC-C} \end{cases}$$

iterations.

GDA and GDmax

GDA

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta_x \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t))$$
$$\mathbf{y}_{t+1} = \Pi_{\mathcal{Y}}(\mathbf{y}_t + \eta_y \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))$$

- Single loop: $\mathbf{x}$ and $\mathbf{y}$ are updated simultaneously
- Different stepsizes for $\mathbf{x}$ and $\mathbf{y}$ – two-time-scale!

GDmax

find $\mathbf{y}_t \in \mathcal{Y}$ such that
$\phi(\mathbf{x}_t, \mathbf{y}_t) \geq \max_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}_{t-1}, \mathbf{y}) - \zeta$

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t - \eta_x \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t))$$

- Double loop: approximately solve $\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}_t, \mathbf{y})$ in each iteration.
- Projected gradient ascent can be used to solve the inner problem.

Problems with Polyak-Łojasiewicz Condition

[Yang et al., 2020]

One-sided PL condition:

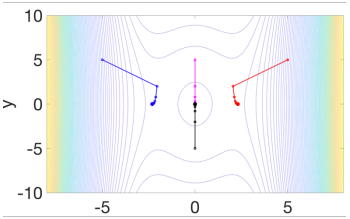
$$\|\nabla_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y})\|^2 \geq 2\mu_2[\max_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y}) - \phi(\mathbf{x},\mathbf{y})].$$

- Alternating GDA converges to an ϵ-stationary point such that $\|\nabla\bar{\phi}(\mathbf{x})\| \leq \epsilon$, with at most $O(\kappa^2/\epsilon^2)$ iterations.

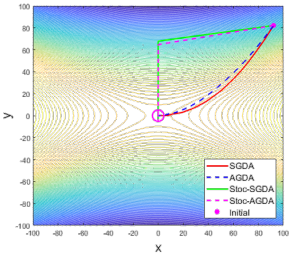
Two-sided PL condition:

$$\|\nabla_{\mathbf{x}}\phi(\mathbf{x},\mathbf{y})\|^2 \geq 2\mu_1[\phi(\mathbf{x},\mathbf{y}) - \min_{\mathbf{x}}\phi(\mathbf{x},\mathbf{y})],$$
$$\|\nabla_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y})\|^2 \geq 2\mu_2[\max_{\mathbf{y}}\phi(\mathbf{x},\mathbf{y}) - \phi(\mathbf{x},\mathbf{y})].$$

- Alternating GDA converges linearly to an saddle point.



(NC-PL)

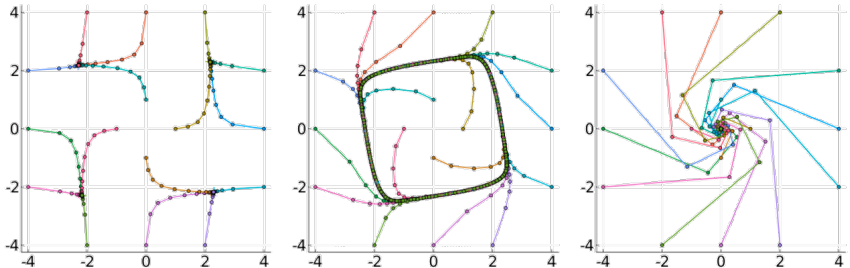


(PL-PL)

Problems with Interaction Dominance

[Grimmer et al., 2020]

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}) + \mathbf{x}^T A\mathbf{y} - g(\mathbf{y})$$



(a) $A = 1$ (b) $A = 10$ (c) $A = 100$

Proximal point algorithm: local linear convergence, divergence, global linear convergence

Recent Advances and Open Questions

- Non-monotone VIs and weakly-convex-weakly-concave problems
- Second-order method for local Nash equilibrium or local minimax points
- Stochastic and adaptive methods
- Landscape of the stable limit points for other algorithms
- Avoid spurious critical points
- Lower complexity bounds for minimax problems
- ...

References

Further reference: see Daskalakis's tutorial on min-max optimization at Simons Institute: `https://simons.berkeley.edu/workshops/games2022-boot-camp`

T. Lin, C. Jin, & M. Jordan.
On gradient descent ascent for nonconvex-concave minimax problems.
International Conference on Machine Learning (pp. 6083-6093), 2020

C. Jin, P. Netrapalli,& M. Jordan.
What is local optimality in nonconvex-nonconcave minimax optimization?
International Conference on Machine Learning (pp. 4880-4889), 2020

Yang, J., Kiyavash, N., & He, N.
Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems.
Advances in Neural Information Processing Systems, 33, 1153-1165, 2020.

Grimmer, B., Lu, H., Worah, P., & Mirrokni, V.
Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems.
In International Conference on Algorithmic Learning Theory (pp. 465-487), 2022

Last Lecture

Q&A Session: May 31, 10am-12pm