Are Single-Loop Algorithms Sufficient for Unbalanced Minimax Optimization?

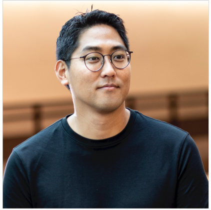
Niao He

Optimization & Decision Intelligence Group (ODI)
ETH Zürich

Simons Workshop on Adversarial Approaches in Machine Learning

Feb 24, 2022

Related Papers



Kiran Thekumparampil
(Amazon)

Sewoong Oh
(University of Washington)

- [THO22] Kiran Thekumparampil, Niao He, Sewoong Oh. **"Lifted Primal-Dual Method for Bilinearly Coupled Smooth Minimax Optimization"**. AISTATS 2022.
- [Yan+22] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, Niao He. **"Faster Single-loop Algorithms for Minimax Optimization without Strong Concavity"**. AISTATS 2022.

Minimax Optimization

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

Wide applications: game theory, reinforcement learning, robust optimization, and GANs, etc.

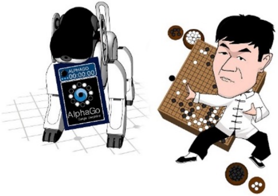


Figure: Games

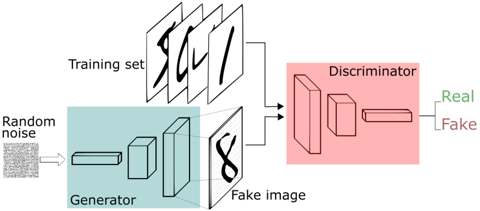
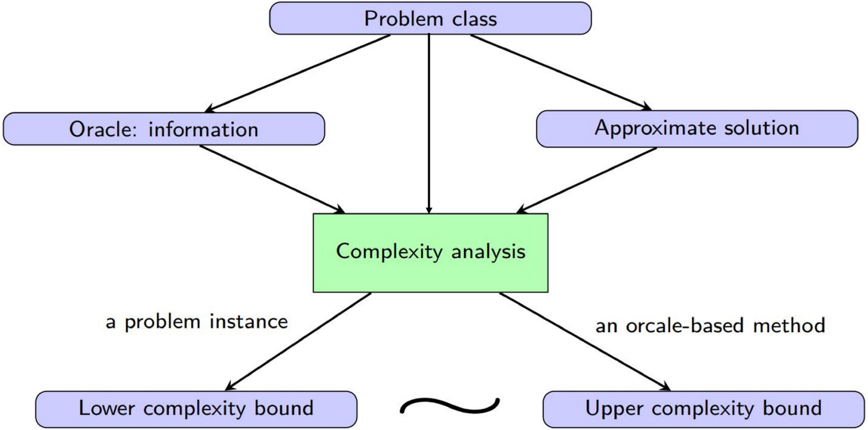


Figure: GANs

Problem Class, Oracles, Complexity



Smooth Minimax Optimization

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

- Problem Class: $\mathcal{F}(L_x, L_y, L_{xy}, \mu_x, \mu_y)$
- Smoothness constants: for all $x, x_1, x_2 \in \mathcal{X}, y, y_1, y_2 \in \mathcal{Y}$:

$$\|\nabla_x \phi(x_1, y) - \nabla_x \phi(x_2, y)\| \le L_x \|x_1 - x_2\|; \; \|\nabla_y \phi(x_1, y) - \nabla_y \phi(x_2, y)\| \le L_{xy} \|x_1 - x_2\|$$
$$\|\nabla_x \phi(x, y_1) - \nabla_x \phi(x, y_2)\| \le L_{xy} \|y_1 - y_2\|; \; \|\nabla_y \phi(x, y_1) - \nabla_y \phi(x, y_2)\| \le L_y \|y_1 - y_2\|$$

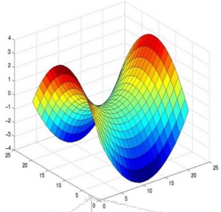
- Convexity constants: for all $x, x_1, x_2 \in \mathcal{X}, y, y_1, y_2 \in \mathcal{Y}$:

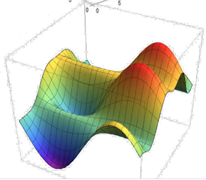
$$\mu_x \|x_1 - x_2\| \le \|\nabla_x \phi(x_1, y) - \nabla_x \phi(x_2, y)\|; \quad \mu_y \|y_1 - y_2\| \le \|\nabla_y \phi(x, y_1) - \nabla_y \phi(x, y_2)\|$$

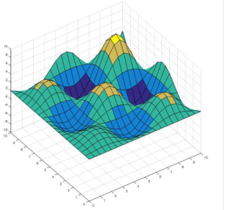
  - $\mu_x > 0$, strongly convex; $\mu_x = 0$, convex; $\mu_x < 0$, weakly convex
  - $\mu_y > 0$, strongly concave; $\mu_y = 0$, concave; $\mu_y < 0$, weakly concave

Critical Regimes

- **Convex-Concave** (C-C)
- **Strongly-Convex-Strongly-Concave** (SC-SC)
  (Extensive literature)
- **Strongly-Convex-Concave** (SC-C)
  [The+19; LJJ20b; WL20; Yan+20] ...
- **Nonconvex-Strongly-Concave** (NC-SC)
  [LJJ20a; LJJ20b; Zha+21; Li21] ...
- **Nonconvex-Concave** (NC-C)
  [The+19; LJJ20a; LJJ20b; OLR20; Yan+20] ...
- **Nonconvex-Nonconcave** (NC-NC)
  [Lin+18; DP18; FR20; JNJ20; DSZ21] ...







The Classical (Balanced) Setting

- Balanced setting: $\mu_x = \mu_y := \mu \geq 0, \quad L_x = L_y = L_{xy} := L$
- Variational inequalities with μ-strongly-montone and L-Lipschitz operator F:

$$\text{VI}(Z, F) \qquad \text{Find } \mathbf{z}^* \in Z : \langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0, \forall \mathbf{z} \in Z$$

$$Z = \mathcal{X} \times \mathcal{Y} \text{ and } F(\mathbf{z} = [x; y]) = [\nabla_x \phi(x, y); -\nabla_y \phi(x, y)]$$

- **Lower bound** [NY83]: $O(\frac{L}{\mu} \log \frac{1}{\epsilon})$ if $\mu > 0$ and $O(\frac{L}{\epsilon})$ if $\mu = 0$
- **Optimal first-order algorithms**:
  - **Extragradient method (EG)** [Kor76]: $z_{t+1} = z_t - \eta F(z_t - \eta F(z_t))$
  - **Optimistic GDA** [Pop80]: $z_{t+1} = z_t - \eta(2F(z_t) - F(z_{t-1}))$
  - **Reflected-Forward-Backward Splitting** [Mal15]: $z_{t+1} = z_t - \eta F(2z_t - z_{t-1})$
  - **Accelerated dual extrapolation (DE)** [NS06]

The (Unbalanced) Strongly-Convex-Strongly-Concave Seting

- Generic setting: $\mathcal{F}(L_x, L_y, L_{xy}, \mu_x, \mu_y)$ with $\mu_x > 0, \mu_y > 0$
- **Lower bound** [ZHZ19]:

$$\Omega\left(\sqrt{\frac{L_x}{\mu_x} + \frac{L_{xy}^2}{\mu_x \mu_y} + \frac{L_y}{\mu_y}} \cdot \log\frac{1}{\epsilon}\right)$$

- Consider the bilinear coupled minimax problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \ \phi(x,y) = f(x) + \langle y, Ax\rangle - h(y)$$

Here $f(x)$ is μ_x-strongly-convex and L_x-smooth, and similarly for $h(y)$, and they can only be accessed through first-order gradient oracles. Note that $L_{xy} = \|A\|$.

The (Unbalanced) Strongly-Convex-Strongly-Concave Seting

Method	Complexity for ϵ primal-dual gap	# Loops
EG/OGDA/MP/Reflective FB/DE [MOP20]	$\mathcal{O}\big(\frac{\mathcal{L}}{\min(\mu_x,\mu_y)}\big)\log\frac{1}{\epsilon}$	Single
Catalyst-EG/OGDA [Yan+20; Zha+21]	$\mathcal{O}\big(\frac{\mathcal{L}}{\sqrt{\mu_x\mu_y}}\big)\log\frac{1}{\epsilon}$	Two
Relative Lipschitz MP [CST21]	$\mathcal{O}\big(\frac{L_x}{\mu_x}+\frac{L_{xy}}{\sqrt{\mu_x\mu_y}}+\frac{L_y}{\mu_y}\big)\log\frac{1}{\epsilon}$	Single
Proximal Best Response [WL20]	$\widetilde{\mathcal{O}}\big(\sqrt{\frac{L_x}{\mu_x}+\frac{L_{xy}\mathcal{L}}{\mu_x\mu_y}+\frac{L_y}{\mu_y}}\big)\log\frac{1}{\epsilon}$	Four

$$\mathcal{L}=\max(L_x,L_{xy},L_y)$$

Question

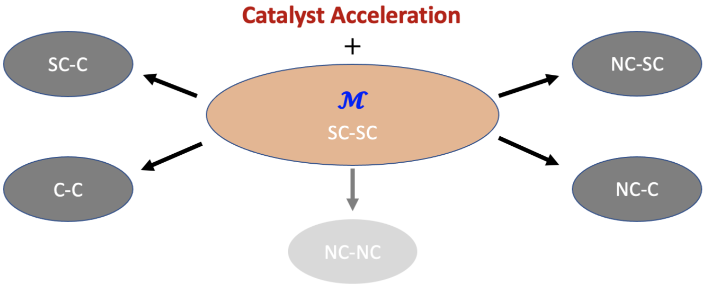
Q1: Can we close the gap?
Q2: Can we achieve it by single-loop algorithms?

$$\Omega\left(\sqrt{\frac{L_x}{\mu_x}+\frac{L_{xy}^2}{\mu_x\mu_y}+\frac{L_y}{\mu_y}}\cdot\log\frac{1}{\epsilon}\right)$$

$$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_x}{\mu_x}+\frac{L_{xy}\mathcal{L}}{\mu_x\mu_y}+\frac{L_y}{\mu_y}}\cdot\log\frac{1}{\epsilon}\right)$$

Motivation

- Why do we care about achieving optimal complexity in SC-SC setting?
- Why do we care about designing simple single-loop algorithms?



[Yan+20] Yang, Zhang, Kiyavash, He. A Catalyst Framework for Minimax Optimization. NeurIPS 2020.

Short Answer

- YES for bilinearly coupled minimax optimization (Bi-SC-SC)!
  [Thekumparampil-He-Oh, AISTATS 2022] **Primal-dual lifting.**

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \phi(x,y) = f(x) + \langle y, Ax\rangle - h(y)$$

- **Recent work:** Nearly YES for separable minimax optimization!
  [Jin-Sidford-Tian, ArXiv 2022]

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \phi(x,y) = f(x) + g(x,y) - h(y)$$

- **NB:** distinct from existing work on linear convergence of Bi-SC-SC:
  - Both f and h are proximal-friendly [CP11]
  - Only h is proximal-friendly [CP16]
  - Only h is strongly convex but A is full rank [DH19]

Example: Quadratic Minimax Problems

The most prominent example is quadratic minimax problems:

$$\min_x \max_y \phi(x, y) = x^\top Bx + y^\top Ax - y^\top Cy$$

- Numerical analysis
- Constrained matrix games
- Robust least square [EGL97]
- MSPBE for policy evaluation [DH19]
- ....

Inspiration I: Primal-Dual for Bilinear Problems

Consider composite blinear problems with simple terms:

$$\min_x \max_y \; F(x) + \langle y, Ax \rangle - H(y)$$

F, H are μ_x, μ_y-strongly convex w.r.t. Bregman divergences $V^r_{x'}(x)$, $V^s_{y'}(y)$ & proximal-friendly.

Primal-Dual [CP16]

$$\begin{cases} \widetilde{y}_{k+1} = y_k + \theta(y_k - y_{k-1}) \\ x_{k+1} = \arg\min_x \left\langle A^\top \widetilde{y}_{k+1}, x \right\rangle + \dfrac{1}{\eta_x} V^r_{x_k}(x) + F(x) \\ y_{k+1} = \arg\min_y - \langle A x_{k+1}, y \rangle + \dfrac{1}{\eta_y} V^s_{y_k}(y) + H(x) \end{cases}$$

- Can be viewed as approximation of **Proximal Point Algorithm**
- Iteration complexity is at most $\mathcal{O}\left(\frac{\|A\|}{\sqrt{\mu_x \mu_y}} \log \frac{1}{\epsilon}\right)$, which is optimal.

Inspiration II: Primal-Dual for Convex Minimization

Consider the smooth minimization with strongly convex objective:

$$\min_x f(x) \iff \min_x \max_u \frac{\mu}{2}\|x\|^2 + \langle x, u\rangle - \underline{f}^*(u)$$

where $\underline{f}(x) = f(x) - \frac{\mu}{2}\|x\|^2$, $\underline{f}^*(u) = \max_x \langle u, x\rangle - \underline{f}(x)$ is the Fenchel dual.

Primal-Dual = Accelerated Gradient Descent

$$\begin{cases} \widetilde{\nabla}_{k+1} = \nabla \underline{f}(\underline{x}_k) + \theta(\nabla \underline{f}(\underline{x}_k) - \nabla \underline{f}(\underline{x}_{k-1})) \\ x_{k+1} = (x_k - \eta_x \widetilde{\nabla}_{k+1})/(1 + \eta_x \mu) \\ \underline{x}_{k+1} = (\underline{x}_k + \eta_u x_{k+1})/(1 + \eta_u) \end{cases}$$

- Game perspective of Nesterov's acceleration [LZ18]
- Slight variation in extrapolation

Our Approach: Acceleration via Lifting

- Original problem of interest:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x) + \langle y, Ax\rangle - h(y)$$

- Reformulation on lifted space:

$$\begin{aligned}\min_{x\in\mathcal{X},v} \ \max_{y\in\mathcal{Y},u} \ \Phi(x,y;u,v) := \big[-\underline{f}^*(u) + \langle u, x\rangle + \frac{\mu_x}{2}\|x\|^2\big] \\ + \langle y, Ax\rangle \\ - \big[\frac{\mu_y}{2}\|y\|^2 + \langle v, y\rangle - \underline{h}^*(v)\big]\end{aligned}$$

- **Key feature:** only bilinear coupling + proximal-friendly terms: $\frac{\mu_x}{2}\|\cdot\|^2$, $\frac{\mu_y}{2}\|\cdot\|^2$, $\underline{f}^*$, & $\underline{h}^*$

Single-loop Algorithm: Lifted Primal-Dual Method

Lifted Primal-Dual (LPD)

$$
\begin{cases}
(\widetilde{x}_{k+1}, \widetilde{y}_{k+1}) = (1+\theta)(x_k, y_k) - \theta(x_{k-1}, y_{k-1}) \\
(\widetilde{u}_{k+1}, \widetilde{v}_{k+1}) = (1+\theta)(u_k, v_k) - \theta(u_{k-1}, v_{k-1}) \\
\\
x_{k+1} = \arg\min_{x \in \mathcal{X}} \left\langle A^\top \widetilde{y}_{k+1} + \widetilde{u}_{k+1}, x \right\rangle + \|x - x_k\|^2/2\eta_x + \mu_x \|x\|^2/2 \\
y_{k+1} = \arg\min_{y \in \mathcal{Y}} - \left\langle A^\top \widetilde{x}_{k+1} + \widetilde{v}_{k+1}, y \right\rangle + \|y - y_k\|^2/2\eta_y + \mu_y \|y\|^2/2 \\
u_{k+1} = \arg\min_{u} - \langle x_{k+1}, u \rangle + \underline{f}^*(u) + V_{u_k}^{\underline{f}^*}(u)/\eta_u \\
v_{k+1} = \arg\min_{v} - \langle y_{k+1}, v \rangle + \underline{h}^*(v) + V_{v_k}^{\underline{h}^*}(v)/\eta_v
\end{cases}
$$

Single-loop Algorithm: Lifted Primal-Dual Method

Simplified Implementable Lifted Primal-Dual (LPD)

$$
\begin{cases}
\widetilde{x}_{k+1} = x_k + \theta_k(x_k - x_{k-1}) \\
\widetilde{y}_{k+1} = y_k + \theta_k(y_k - y_{k-1}) \\
\widetilde{\nabla}_{x,k+1} = \nabla \underline{f}(\underline{x}_k) + \theta_k(\nabla \underline{f}(\underline{x}_k) - \nabla \underline{f}(\underline{x}_{k-1})) \\
\widetilde{\nabla}_{y,k+1} = \nabla \underline{h}(\underline{y}_k) + \theta_k(\nabla \underline{h}(\underline{y}_k) - \nabla \underline{h}(\underline{y}_{k-1})) \\
\\
x_{k+1} = \mathcal{P}_{\mathcal{X}}((x_k - \eta_x(A^\top \widetilde{y}_{k+1} + \widetilde{\nabla}_{x,k+1})) \\
y_{k+1} = \mathcal{P}_{\mathcal{Y}}((y_k + \eta_y(A\widetilde{x}_{k+1} - \widetilde{\nabla}_{y,k+1})) \\
\underline{x}_{k+1} = (\underline{x}_k + \eta_u\, x_{k+1})/(1 + \eta_u) \\
\underline{y}_{k+1} = (\underline{y}_k + \eta_v\, y_{k+1})/(1 + \eta_v)
\end{cases}
$$

Main Result for SC-SC Setting

Theorem (Informal, Bi-SC-SC [THO22])

Let $\kappa_x = L_x/\mu_x$, $\kappa_y = L_y/\mu_y$, $\kappa_{xy} = \|A\|/\sqrt{\mu_x \mu_y}$. *Define* $\kappa = \sqrt{\kappa_x - 1} + 2\kappa_{xy} + \sqrt{\kappa_y - 1}$. *Denote*

$$\Delta(x, y) = \kappa_{xy}(\mu_x \|x - x^*\|^2 + \mu_y \|y - y^*\|^2).$$

LPD with T *iterations sastisfies*

$$\Delta(x_T, y_T) \leq \mathcal{O}(e^{-\frac{T}{\kappa}})\Delta(x_0, y_0)$$

- The gradient complexity is

$$\mathcal{O}\Big(\Big(\sqrt{\frac{L_x}{\mu_x} - 1} + \frac{\|A\|}{\sqrt{\mu_x \mu_y}} + \sqrt{\frac{L_y}{\mu_y} - 1}\Big) \log\left(\frac{1}{\epsilon}\right)\Big)$$

- Optimal as it matches exactly with lower bound in [ZHZ19].

Extension to C-SC Setting

- **LPD + Smoothing**: setting $\mu_x = \mathcal{O}(\epsilon)$ leads to gradient complexity of

$$\mathcal{O}\left(\sqrt{\frac{L_x}{\epsilon}} + \frac{\|A\|}{\sqrt{\mu_y \epsilon}} + \sqrt{\frac{L_y}{\mu_y}}\right) \log\left(\frac{1}{\epsilon}\right)$$

  Near-optimal up to logarithmic term.

- **LPD + Decaying Stepsize**: attains $\mathcal{O}\left(\frac{1}{T^2}\right)$ convergence rate and gradient complexity of

$$\mathcal{O}\left(\sqrt{\frac{L_x}{\epsilon}} + \frac{\|A\|}{\sqrt{\mu_y \epsilon}} + \sqrt{\frac{L_y - \mu_y}{\epsilon}}\right).$$

  Improve over $\mathcal{O}\left(\frac{\mathcal{L}}{\sqrt{\mu_y \epsilon}} \log \frac{1}{\epsilon}\right)$ achieved by Catalyst-EG/OGDA [Yan+20]

Summary and Open Questions

Single-loop and (near-)optimal algorithm for bilinearly coupled minimax optimization in (strongly-)convex-strongly-concave setting

Open Question: Can we extend the success to

- General non-separable minimax optimization?
- Other settings: NC-SC, NC-C?
- Stochastic and finite-sum settings?

Nonconvex-PL (NC-PL) Minimax Optimization

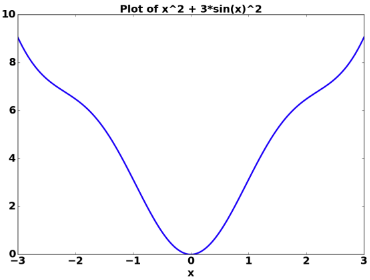
$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} f(x,y) \triangleq \mathbb{E}[F(x,y;\xi)].$$

Setting:

- f is L-Lipschitz smooth
- $-f(x,\cdot)$ satisfies μ-PL inequality, i.e., $\|\nabla_y f(x,y)\|^2 \geq 2\mu[\max_y f(x,y) - f(x,y)],\ \forall x,y.$

Note

- Does not require concavity nor strong concavity in y.
- PL inequality holds in many nonconvex applications
  - Linear–quadratic regulator [Faz+18];
  - Over-parametrized neural networks [LZB20];
  - Reinforcement learning [Mei+20].



A Single-loop Algorithm: Smoothed AGDA

Smoothed GDA

At each iteration t: draw two i.i.d. samples ξ_1^t, ξ_2^t

$$\begin{cases} x_{t+1} = x_t - \tau_1[\nabla F_x(x_t, y_t, \xi_1^t) + p(x_t - z_t)] \\ y_{t+1} = y_t + \tau_2 \nabla F_y(x_{t+1}, y_t, \xi_2^t) \\ z_{t+1} = z_t + \beta(x_{t+1} - z_t). \end{cases}$$

- Smoothed AGDA was first introduced in [Zha+20] for deterministic nonconvex-concave minimax problems
- Mimics the primal-dual method with stochastic gradients

Convergence of Smoothed AGDA

Theorem (informal, [Yan+22])

Under the NC-PL setting, Smoothed AGDA can find an ϵ-stationary point with

- *Deterministic case: $\mathcal{O}(\kappa\epsilon^{-2})$ iteration complexity*
- *Stochastic case: $\mathcal{O}(\kappa^2\epsilon^{-4})$ sample complexity*

- No need for mini-batch to achieve $\mathcal{O}(\epsilon^{-4})$ complexity unlike Stoc-GDA [LJJ20a]
- Improved dependence on κ compared to other single-loop algorithms
- Much weaker assumption

NC-PL Problems: Deterministic Case

Table: Oracle complexity to find ϵ-stationary point of Φ.

Algorithms	**Complexity**	**Loops**	**Additional assumptions**
GDA [LJJ20a]	$\mathcal{O}(\kappa^2 \Delta l \epsilon^{-2})$	1	strong concavity in y
Multi-GDA [Nou+19]	$\tilde{\mathcal{O}}(\kappa^3 \Delta l \epsilon^{-2})$[1]	2	
Catalyst-AGDA[Yan+22]	$\mathcal{O}(\kappa \Delta l \epsilon^{-2})$	2	
Smoothed-AGDA [Yan+22]	$\mathcal{O}(\kappa \Delta l \epsilon^{-2})$	1	

[1] The complexity is derived by translating from another stationary measure.

NC-PL Problems: Stochastic Case

Table: Sample complexity to find ϵ-stationary point of Φ.

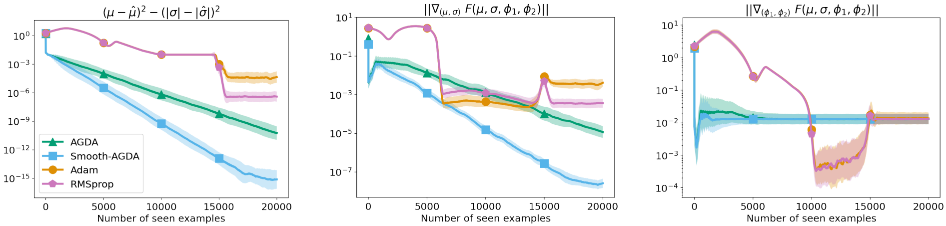
Algorithms	Complexity	Batch size	Additional assumptions
Stoc-GDA [LJJ20a]	$\mathcal{O}(\kappa^3 \Delta l \epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	strong concavity in y
Stoc-GDA [LJJ20a]	$\mathcal{O}(\kappa^3 \Delta l \epsilon^{-5})$	$\mathcal{O}(1)$	strong concavity in y
PDSM [Guo+21]	$\mathcal{O}(\kappa^3 \Delta l \epsilon^{-4})$	$\mathcal{O}(1)$	strong concavity in y
ALSET [CSY21]	$\mathcal{O}(\kappa^3 \Delta l \epsilon^{-4})$	$\mathcal{O}(1)$	strong concavity in y, Lipschitz[1]
Stoc-AGDA[Yan+22]	$\mathcal{O}(\kappa^4 \Delta l \epsilon^{-4})$	$\mathcal{O}(1)$	
Stoc-Smoothed-AGDA[Yan+22]	$\mathcal{O}(\kappa^2 \Delta l \epsilon^{-4})$	$\mathcal{O}(1)$	

[1] It assumes f is Lipschitz continuous about x and its Hessian is Lipschitz continuous.

Toy WGAN with linear generator

- Linear generator $G_{\mu,\sigma}(z) = \mu + \sigma z$ and quadratic discriminator $D_\phi(x) = \phi_1 x + \phi_2 x^2$

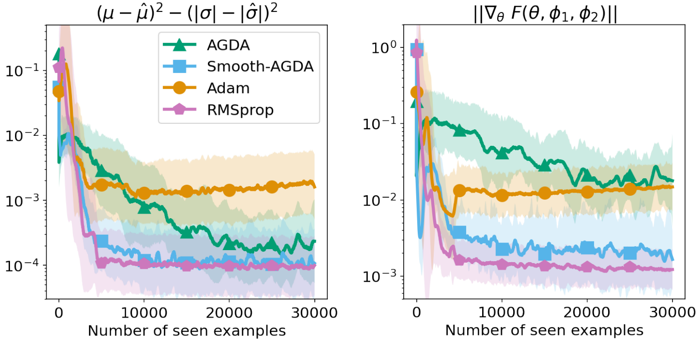
$$\min_{\mu,\sigma} \max_{\phi_1,\phi_2} \mathbb{E}_{(x_{real},z)\sim\mathcal{D}}\ \phi_1 x_{real} + \phi_2 x_{real}^2 - \phi_1 \cdot (\mu + \sigma z) - \phi_2 \cdot (\mu + \sigma z)^2 - \lambda \|\phi\|^2.$$



Toy WGAN with Neural Generator

- One hidden layer neural network generator G_θ and quadratic discriminator

$$\min_{\theta} \max_{\phi_1,\phi_2} \mathbb{E}_{(x_{real},z)\sim\mathcal{D}} \; \phi_1 x_{real} + \phi_2 x_{real}^2 - \phi_1 \cdot G_\theta(z) - \phi_2 \cdot (G_\theta(z))^2 - \lambda\|\phi\|^2.$$



References I

[CP11] Antonin Chambolle and Thomas Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging". In: Journal of mathematical imaging and vision 40.1 (2011), pp. 120–145.

[CP16] Antonin Chambolle and Thomas Pock. "On the ergodic convergence rates of a first-order primal–dual algorithm". In: Mathematical Programming 159.1 (2016), pp. 253–287.

[CST21] Michael B. Cohen, Aaron Sidford, and Kevin Tian. Relative Lipschitzness in Extragradient Methods and a Direct Recipe for Acceleration. 2021. arXiv: 2011.06572 [math.OC].

[CSY21] Tianyi Chen, Yuejiao Sun, and Wotao Yin. "Tighter Analysis of Alternating Stochastic Gradient Method for Stochastic Nested Problems". In: arXiv preprint arXiv:2106.13781 (2021).

[DH19] Simon S Du and Wei Hu. "Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity". In: The 22nd International Conference on Artificial Intelligence and Statistics. PMLR. 2019, pp. 196–205.

[DP18] Constantinos Daskalakis and Ioannis Panageas. "The limit points of (optimistic) gradient descent in min-max optimization". In: Advances in Neural Information Processing Systems 31 (2018).

[DSZ21] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. "The complexity of constrained min-max optimization". In: Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing. 2021, pp. 1466–1478.

References II

[EGL97] Laurent El Ghaoui and Hervé Lebret. "Robust solutions to least-squares problems with uncertain data". In: SIAM Journal on matrix analysis and applications 18.4 (1997), pp. 1035–1064.

[Faz+18] Maryam Fazel et al. "Global convergence of policy gradient methods for the linear quadratic regulator". In: arXiv preprint arXiv:1801.05039 (2018).

[FP07] Francisco Facchinei and Jong-Shi Pang. Finite-dimensional variational inequalities and complementarity problems. Springer Science & Business Media, 2007.

[FR20] Tanner Fiez and Lillian Ratliff. "Gradient descent-ascent provably converges to strict local minmax equilibria with a finite timescale separation". In: arXiv preprint arXiv:2009.14820 (2020).

[Guo+21] Zhishuai Guo et al. "On Stochastic Moving-Average Estimators for Non-Convex Optimization". In: arXiv preprint arXiv:2104.14840 (2021).

[HXZ21] Yuze Han, Guangzeng Xie, and Zhihua Zhang. "Lower complexity bounds of finite-sum optimization problems: The results and construction". In: arXiv preprint arXiv:2103.08280 (2021).

[JNJ20] Chi Jin, Praneeth Netrapalli, and Michael Jordan. "What is local optimality in nonconvex-nonconcave minimax optimization?" In: International Conference on Machine Learning. PMLR. 2020, pp. 4880–4889.

References III

[Kor76] Galina M Korpelevich. "The extragradient method for finding saddle points and other problems". In: Matecon 12 (1976), pp. 747–756.

[Li21] Haochuan Li. "On the Complexity of Nonconvex-Strongly-Concave Smooth Minimax Optimization Using First-Order Methods". PhD thesis. Massachusetts Institute of Technology, 2021.

[Lin+18] Qihang Lin et al. "Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality". In: arXiv preprint arXiv:1810.10207 5 (2018).

[LJJ20a] Tianyi Lin, Chi Jin, and Michael Jordan. "On gradient descent ascent for nonconvex-concave minimax problems". In: International Conference on Machine Learning. PMLR. 2020, pp. 6083–6093.

[LJJ20b] Tianyi Lin, Chi Jin, and Michael I Jordan. "Near-optimal algorithms for minimax optimization". In: Conference on Learning Theory. PMLR. 2020, pp. 2738–2779.

[LZ18] Guanghui Lan and Yi Zhou. "An optimal randomized incremental gradient method". In: Mathematical programming 171.1 (2018), pp. 167–215.

[LZB20] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks". In: arXiv preprint arXiv:2003.00307 (2020).

[Mal15] Yu Malitsky. "Projected reflected gradient methods for monotone variational inequalities". In: SIAM Journal on Optimization 25.1 (2015), pp. 502–520.

References IV

[Mei+20] Jincheng Mei et al. "On the global convergence rates of softmax policy gradient methods". In: International Conference on Machine Learning. PMLR. 2020, pp. 6820–6829.

[MOP20] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. "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. PMLR. 2020, pp. 1497–1507.

[Nou+19] Maher Nouiehed et al. "Solving a class of non-convex min-max games using iterative first order methods". In: arXiv preprint arXiv:1902.08297 (2019).

[NS06] Yurii Nesterov and Laura Scrimali. "Solving strongly monotone variational and quasi-variational inequalities". In: (2006).

[NY83] Arkadi. S. Nemirovski and David. B. Yudin. Problem Complexity and Method Efficiency in Optimization. Wiley, New York, 1983.

[OLR20] Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. "Efficient Search of First-Order Nash Equilibria in Nonconvex-Concave Smooth Min-Max Problems". In: arXiv preprint arXiv:2002.07919 (2020).

[Pop80] Leonid Denisovich Popov. "A modification of the Arrow-Hurwicz method for search of saddle points". In: Mathematical notes of the Academy of Sciences of the USSR 28.5 (1980), pp. 845–848.

References V

[The+19] Kiran K Thekumparampil et al. "Efficient algorithms for smooth minimax optimization". In: Advances in Neural Information Processing Systems. 2019, pp. 12659–12670.

[THO22] Kiran Thekumparampil, Niao He, and Sewoong Oh. "Lifted Primal-Dual Method for Bilinearly Coupled Smooth Minimax Optimization". In: To appear in AISTATS. 2022.

[WL20] Yuanhao Wang and Jian Li. "Improved algorithms for convex-concave minimax optimization". In: Advances in Neural Information Processing Systems 33 (2020), pp. 4800–4810.

[Yan+20] Junchi Yang et al. "A catalyst framework for minimax optimization". In: Advances in Neural Information Processing Systems 33 (2020), pp. 5667–5678.

[Yan+22] Junchi Yang et al. "Faster Single-loop Algorithms for Minimax Optimization without Strong Concavity". In: To appear in AISTATS. 2022.

[Zha+20] Jiawei Zhang et al. "A Single-Loop Smoothed Gradient Descent-Ascent Algorithm for Nonconvex-Concave Min-Max Problems". In: arXiv preprint arXiv:2010.15768 (2020).

[Zha+21] Siqi Zhang et al. "The complexity of nonconvex-strongly-concave minimax optimization". In: Uncertainty in Artificial Intelligence. PMLR. 2021, pp. 482–492.

[ZHZ19] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. "On lower iteration complexity bounds for the saddle point problems". In: arXiv preprint arXiv:1912.07481 (2019).

Supplementary: Catalyst Acceleration

	SC-SC	SC-C	NC-SC	NC-C
GDA	$\mathcal{O}\left(\frac{L^2}{\min\{\mu_x^2,\mu_y^2\}}\log\frac{1}{\epsilon}\right)$ [FP07]	?	$\mathcal{O}\left(\frac{L^3}{\mu^2}\epsilon^{-2}\right)$ [LJJ20a]	$\mathcal{O}\left(L^3\ell^2\epsilon^{-6}\right)$ [LJJ20a]
SOTA (before ours)	$\mathcal{O}\left(\frac{L}{\sqrt{\mu_x\mu_y}}\log^3\frac{1}{\epsilon}\right)$ [LJJ20b]	$\mathcal{O}\left(\frac{L}{\sqrt{\mu\epsilon}}\log^3\frac{1}{\epsilon}\right)$ [LJJ20b]	$\mathcal{O}\left(\frac{L^{3/2}}{\sqrt{\mu}}\epsilon^{-2}\log^2\frac{1}{\epsilon}\right)$ [LJJ20b]	$\mathcal{O}\left(L^2\epsilon^{-3}\log^2\frac{1}{\epsilon}\right)$ [LJJ20b] [The+19]
Lower bound	$\Omega\left(\frac{L}{\sqrt{\mu_x\mu_y}}\log\frac{1}{\epsilon}\right)$ [ZHZ19]	$\Omega\left(\frac{L}{\sqrt{\mu\epsilon}}\right)$ [HXZ21]	$\Omega\left(\frac{L^{3/2}}{\sqrt{\mu}}\epsilon^{-2}\right)$	?
Catalyst-EG/OGDA	$\mathcal{O}\left(\frac{L}{\sqrt{\mu_x\mu_y}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\sqrt{\mu\epsilon}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L^{3/2}}{\sqrt{\mu}}\epsilon^{-2}\right)$	$\mathcal{O}\left(L^2\epsilon^{-3}\log\frac{1}{\epsilon}\right)$

Proximal Best Response [WL20]

Algorithm 4 Proximal Best Response

Require: Initial point $\mathbf{z}_0 = [\mathbf{x}_0; \mathbf{y}_0]$

$\beta_1 \leftarrow \max\{m_{\mathbf{x}}, L_{\mathbf{xy}}\}$, $M_1 \leftarrow \frac{80L^3}{m_{\mathbf{x}}^{1.5} m_{\mathbf{y}}^{1.5}}$

$\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_0$, $\kappa \leftarrow \beta_1 / m_{\mathbf{x}}$, $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$, $\tau \leftarrow \frac{1}{2\sqrt{\kappa}+4\kappa}$

for $t = 1, \cdots, T$ **do**

  $(\mathbf{x}_t, \mathbf{y}_t) \leftarrow \text{APPA-ABR}(f(\mathbf{x}, \mathbf{y}) + \beta_1 \|\mathbf{x} - \hat{\mathbf{x}}_{t-1}\|^2, [\mathbf{x}_{t-1}, \mathbf{y}_{t-1}], M_1)$

  $\hat{\mathbf{x}}_t \leftarrow \mathbf{x}_t + \theta(\mathbf{x}_t - \mathbf{x}_{t-1}) + \tau(\mathbf{x}_t - \hat{\mathbf{x}}_{t-1})$

end for

Algorithm 3 APPA-ABR

Require: $g(\cdot,\cdot)$, Initial point $\mathbf{z}_0 = [\mathbf{x}_0; \mathbf{y}_0]$, precision parameter M_1

$\beta_2 \leftarrow \max\{m_{\mathbf{y}}, L_{\mathbf{xy}}\}$, $M_2 \leftarrow \frac{96L^{2.5}}{m_{\mathbf{x}} m_{\mathbf{y}}^{1.5}}$

$\hat{\mathbf{y}}_0 \leftarrow \mathbf{y}_0$, $\kappa \leftarrow \beta_2 / m_{\mathbf{y}}$, $\theta \leftarrow \frac{2\sqrt{\kappa}-1}{2\sqrt{\kappa}+1}$, $\tau \leftarrow \frac{1}{2\sqrt{\kappa}+4\kappa}$, $t \leftarrow 0$

repeat

  $t \leftarrow t + 1$

  $(\mathbf{x}_t, \mathbf{y}_t) \leftarrow \text{ABR}(g(\mathbf{x}, \mathbf{y}) - \beta_2 \|\mathbf{y} - \hat{\mathbf{y}}_{t-1}\|^2, [\mathbf{x}_{t-1}; \mathbf{y}_{t-1}], 1/M_2, 2\beta_1, 2\beta_2, 3L, 3L)$

  $\hat{\mathbf{y}}_t \leftarrow \mathbf{y}_t + \theta(\mathbf{y}_t - \mathbf{y}_{t-1}) + \tau(\mathbf{y}_t - \hat{\mathbf{y}}_{t-1})$

until $\|\nabla g(\mathbf{x}_t, \mathbf{y}_t)\| \le \frac{\min\{m_{\mathbf{x}}, m_{\mathbf{y}}\}}{9LM_1} \|\nabla g(\mathbf{x}_0, \mathbf{y}_0)\|$

Algorithm 1 Alternating Best Response (ABR)

Require: $g(\cdot,\cdot)$, Initial point $\mathbf{z}_0 = [\mathbf{x}_0; \mathbf{y}_0]$, precision ϵ, parameters $m_{\mathbf{x}}$, $m_{\mathbf{y}}$, $L_{\mathbf{x}}$, $L_{\mathbf{y}}$

$\kappa_{\mathbf{x}} := L_{\mathbf{x}} / m_{\mathbf{x}}$, $\kappa_{\mathbf{y}} := L_{\mathbf{y}} / m_{\mathbf{y}}$, $T \leftarrow \left\lceil \log_2 \left(\frac{4\sqrt{\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}}}}{\epsilon} \right) \right\rceil$

for $t = 0, \cdots, T$ **do**

  Run AGD on $g(\cdot, \mathbf{y}_t)$ from $\mathbf{x}_t$ for $\Theta(\sqrt{\kappa_{\mathbf{x}}} \ln(\kappa_{\mathbf{x}}))$ steps to get $\mathbf{x}_{t+1}$

  Run AGD on $-g(\mathbf{x}_{t+1}, \cdot)$ from $\mathbf{y}_t$ for $\Theta(\sqrt{\kappa_{\mathbf{y}}} \ln(\kappa_{\mathbf{y}}))$ steps to get $\mathbf{y}_{t+1}$

end for