Efficient Algorithms for Smooth Minimax Optimization

NeurIPS 2019

Kiran Koshy Thekumparampil[†], Prateek Jain[‡],
Praneeth Netrapalli[‡], Sewoong Oh[±]

[†]University of Illinois at Urbana-Champaign,
[‡]Microsoft Research, India,
[±]University of Washington, Seattle

Oct 27, 2019

Outline

- Minimax Optimization problem
- Efficient algorithm for Nonconvex–Concave minimax problem
- Optimal algorithm for Strongly-Convex—Concave minimax problem

Minimax problem

- Consider the general minimax problem

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

- Two player game: y tries to maximize and x tries to minimize.
- The order of min & max or who plays first (x above) is important

$$\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \; g(x, y) \leq \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \; g(x, y)$$

Examples of Minimax problem

1. GAN: $\min_G \max_D V(G,D)$:

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim P_X}\left[\log\left(D(x)\right)\right] + \mathop{\mathbb{E}}_{z\sim Q_Z}\left[\log\left(1 - D(G(z))\right)\right] = \mathrm{JS}(P_X||Q_X)$$

2. Constrained optimization: $\min_x f(x)$, s.t. $f_i(x) \leq 0$, $\forall\ i \in [m]$

$$\min_x \max_{y\geq 0} \left[\mathcal{L}(x,y) = f(x) + \sum_{i=1}^{m} y_i f_i(x)\right]$$

3. Robust estimation/optimization:

$$\min_x \sum_i \max_{\hat{z}_i} f(x, \hat{z}_i)$$

$$\Delta(\hat{z}_i, z_i) \leq \varepsilon\,,\ \forall\, i \in [m].$$

Nonconvex minimax

- In general $g(x, y)$ is non-convex in both x and y.
  E.g. Neural network based GAN

- Very few works on nonconvex minimax

- We focus on smooth nonconvex–concave minimax problem, i.e. $g(x, \cdot)$ is concave, and g is L-smooth:

$$\max_{a \in \{x,y\}} \left\|\nabla_a g(x, y) - \nabla_a g(x', y')\right\| \leq L \left(\left\|x - x'\right\| + \left\|y - y'\right\|\right).$$

  E.g. smooth constrained optimization.

- In general: $\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \; g(x, y) < \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \; g(x, y)$

- We focus on the non-smooth nonconvex Primal problem:
  $f(x) = \max_y g(x, y)$

$f(x) = \max_{y \in \mathcal{Y}} g(x, y)$ is non-smooth and weakly convex

- f is non-smooth due to maximization over y

ρ-weakly convex function

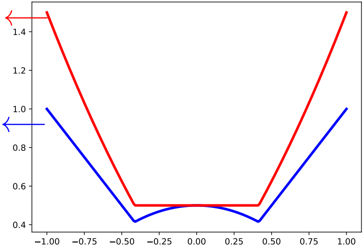
We say that f is a ρ-weakly convex f if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex, i.e.,

$$f(x) + \langle u_x, x' - x \rangle - \frac{\rho}{2}\|x' - x\|^2 \;\; \leq \;\; f(x'),$$

for all Fréchet subgradients $u_x \in \partial f(x)$, for all $x, x' \in \mathcal{X}$.

$f(x) + \frac{x^2}{2}$

$f(x) = \max\{|x|, \frac{1-x^2}{2}\}$



f is 1-weakly convex

as $f + \frac{\|\cdot\|^2}{2}$ is convex

$f(x) = \max_{y \in \mathcal{Y}} g(x, y)$ is non-smooth and weakly convex

- f is non-smooth due to maximization over y

ρ-weakly convex function

We say that f is a ρ-weakly convex f if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex, i.e.,

$$f(x) + \langle u_x, x' - x \rangle - \frac{\rho}{2}\|x' - x\|^2 \;\; \leq \;\; f(x')\,,$$

for all Fréchet subgradients $u_x \in \partial f(x)$, for all $x, x' \in \mathcal{X}$.

- Any L-smooth function is L-weakly convex

$$f(x) + \langle \nabla_x f(x), x' - x \rangle - \frac{L}{2}\|x' - x\|^2 \;\; \leq \;\; f(x')$$

- $-\|x\|$ is not weakly convex (due to upward pointing cusp).

$f(x) = \max_{y \in \mathcal{Y}} g(x, y)$ is non-smooth and weakly convex

- f is non-smooth due to maximization over y

ρ-weakly convex function

We say that f is a ρ-weakly convex f if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex, i.e.,

$$f(x) + \langle u_x, x' - x \rangle - \frac{\rho}{2}\|x' - x\|^2 \;\; \leq \;\; f(x'),$$

for all Fréchet subgradients $u_x \in \partial f(x)$, for all $x, x' \in \mathcal{X}$.

- $f(x) = \max_{y \in \mathcal{Y}} g(x, y)$ is L-weakly convex, if g is L-smooth.

$$\begin{aligned} g(x, y) + \langle \nabla_x g(x, y), x' - x \rangle - \frac{L}{2}\|x' - x\|^2 \;\; &\leq \;\; g(x', y) \\ \implies \; f(x) + \langle u_x, x' - x \rangle - \frac{L}{2}\|x' - x\|^2 \;\; &\leq \;\; f(x') \end{aligned}$$

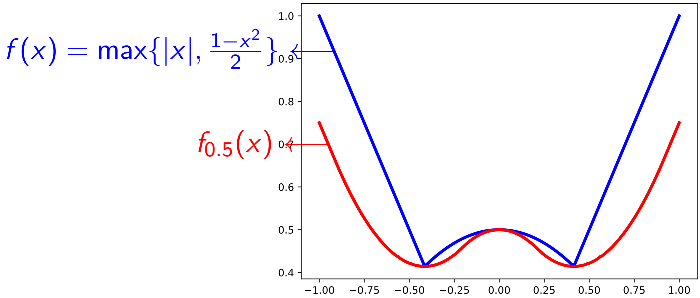
- Cannot define approx. stationary point directly using subgradients

First order stationary point of weakly-convex function

- Moreau envelope f_λ of a L-weakly convex function ($L < \frac{1}{\lambda}$):

$$f_\lambda(x) \;\; = \;\; \min_{x'} f(x') + \frac{1}{2\lambda}\|x - x'\|^2 \, .$$

- f_λ is a smooth lower bound of f: $\nabla f_\lambda(x) = 0 \implies 0 \in \partial f(x)$



First order stationary point of weakly-convex function

- Moreau envelope f_λ of a L-weakly convex function ($L < \frac{1}{\lambda}$):

$$f_\lambda(x) \;=\; \min_{x'} f(x') + \frac{1}{2\lambda}\|x - x'\|^2 \,.$$

- f_λ is a smooth lower bound of f: $\nabla f_\lambda(x) = 0 \implies 0 \in \partial f(x)$

ε-first order stationary point (ε-FOSP)

We say that x is an ε-first order stationary point of a L-weakly convex f if $\|\nabla f_{\frac{1}{2L}}(x)\| \leq \varepsilon$. Further this implies that there exists $\hat{x}$ s.t.,

$$\|\hat{x} - x\| \leq \varepsilon/2L \text{ and } \min_{u \in \partial f(\hat{x})} \|u\| \leq \varepsilon$$

- Algorithm complexity is the no. of first-order oracle calls to obtain ε-FOSP. Convergence rate is ε_k if after k oracle calls we get ε_k-FOSP.

Smooth nonconvex–concave minimax results

Setting	**Previous state-of-the-art**	**Our result**
$\max_y g(x,y)$	$O\left(\varepsilon^{-5}\right)$ [1]	$\widetilde{O}\left(\varepsilon^{-3}\right)$
$\max_i f_i(x) = \max\limits_{y\in\Delta_m} \sum_i^m y_i f_i(x)$	$O\left(\varepsilon^{-4}\right)$ [2]	$\widetilde{O}\left(\varepsilon^{-3}\right)$

Δ_m is the simplex of dimension m.

[1] Jin, C., Netrapalli, P., & Jordan, M. I. (2019). Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. arXiv preprint arXiv:1902.00618.

[2] Davis, D., & Drusvyatskiy, D. (2018). Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. arXiv preprint arXiv:1802.02988.

Baseline: Subgradient method $O(\varepsilon^{-5})$ [1, 2]

- Apply (inexact) subgradient method

$$u_{x_k} = \nabla_x g(x_k, y_k)\,, \text{ where, } y_k \approx y^*(x) = \arg\max_{y \in \mathcal{Y}} g(x_k, y)$$
$$x_{k+1} = \mathcal{P}_{\mathcal{X}}(x_k - \eta u_{x_k})$$

- Sufficient condition: $\max_y g(x_k, y) - g(x_k, y_k) \le O(\varepsilon^2)$ [1]

Setting	Per-step (AGD)	# iterations (Subgrad. method)	Total complexity
$\max_y g(x, y)$	$O\left(\varepsilon^{-1}\right)$	$O\left(\varepsilon^{-4}\right)$	$O\left(\varepsilon^{-5}\right)$
$\max_i f_i(x)$	$O\left(1\right)$	$O\left(\varepsilon^{-4}\right)$	$O\left(\varepsilon^{-4}\right)$

- Does not utilize the smooth minimax structure of $f(x) = max_y g(x, y)$

Proximal Point method (PPM)

- (Inexact) Proximal point method

$$x_{k+1} \approx \arg\min_{x \in \mathcal{X}} f(x) + L\|x - x_k\|^2$$
$$\iff \quad x_{k+1} \approx x_k - 2L\, u_{x_{k+1}},\ u_{x_{k+1}} \in \partial f(x_{k+1})$$

- Iterations complexity to get ε-FOSP is $O(\frac{1}{\varepsilon^2})$

Proof sketch.

- L-weak convexity implies,
  $$f(x_{k+1}) + \left\langle u_{x_{k+1}}, x_k - x_{k+1} \right\rangle - L/2\|x_k - x_{k+1}\|^2 \leq f(x_k)$$
- Using update $x_{k+1} = x_k - 2L\, u_{x_{k+1}}$ we get a Descent Lemma:
  $$f(x_{k+1}) - f(x_k) \leq -3L/2\|u_{x_{k+1}}\|^2$$
- After $O(\frac{f(x_0) - \min_x f(x)}{\varepsilon^2})$ steps, $\min_k \|u_{x_{k+1}}\| = O(\varepsilon)$.
- Generalized to $\|\nabla_{\frac{1}{2L}} f(x_k)\|$ due to inexact update and non-smooth f.

$\square$

Per-step complexity of PPM

- L-weakly convex + $2L$-strongly convex = L-strongly convex
  $f(x) \quad + \quad L\|x - x_k\|^2$
- Each iteration solves L-strongly-convex–concave problem:

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \left[\tilde{g}_k(x,y) = g(x,y) + 2L/2\|x - x_k\|^2\right]$$

- Primal dual gap of $O(\varepsilon^2)$ is sufficient:

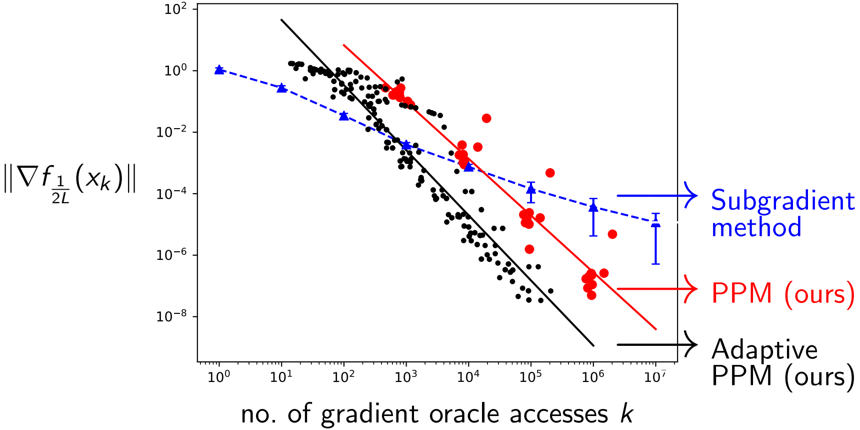
$$\max_{y\in\mathcal{Y}} \tilde{g}_k(x_{k+1}, y) - \min_{x\in\mathcal{X}} \tilde{g}_k(x, y_{k+1}) = O(\varepsilon^2)$$

Algorithm for $\min_{\mathbf{x}} \max_{\mathbf{y}} \tilde{\mathbf{g}}_{\mathbf{k}}(\mathbf{x}, \mathbf{y})$	**Per-step complexity**	**Total complexity**
$O\left(k^{-1}\right)$ Cvx–Cve [Mirror-Prox, 3]	$O\left(\varepsilon^{-2}\right)$	$O\left(\varepsilon^{-4}\right)$
$O\left(k^{-2}\right)$ Strongly-Cvx–Cve [ours]	$O\left(\varepsilon^{-1}\right)$	$O\left(\varepsilon^{-3}\right)$

[3] A. Nemirovski. "Prox-method with rate of convergence O (1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex–concave saddle point problems". In: SIAM Journal on Optimization 15.1 (2004), pp. 229–251.

Nonconvex–concave experiment

$\min_{x\in\mathbb{R}^2}\left[f(x) = \max_{1\le i\le m=9} f_i(x)\right]$, where $f_i(x) = a_i\|x - b_i\|_2^2 + c_i$.



Smooth Convex–Concave minimax problem

- $g(\cdot, y)$ is convex and $g(x, \cdot)$ is concave, and g is L-smooth:

$$\max_{a\in\{x,y\}} \left\|\nabla_a g(x,y) - \nabla_a g(x',y')\right\| \le L\left(\left\|x-x'\right\| + \left\|y-y'\right\|\right).$$

- Primal: $f(x) = \max_{y\in\mathcal{Y}} g(x,y)$. Dual: $h(y) = \min_{x\in\mathcal{X}} g(x,y)$
- If $\mathcal{X}$, $\mathcal{Y}$ are compact, then there is a saddle point (x^*, y^*) (Sion's minimax theorem):

$$\min_{x\in\mathcal{X}} f(x) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x,y) = g(x^*,y^*) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x,y) = \max_{y\in\mathcal{Y}} h(y)$$

ε-primal dual pair (ε-PD pair)

$(\hat{x}, \hat{y})$ is an ε-primal dual pair if the primal-dual gap is less than ε

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

Optimal algorithms for smooth Convex–Concave minimax

Setting	Previous state-of-the-art	Our results	Lower bound
Strongly convex	$O\left(k^{-1}\right)$ [3]	$\widetilde{O}\left(k^{-2}\right)$	$\Omega(k^{-2})$ [4]

[3] A. Nemirovski. "Prox-method with rate of convergence O (1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems". In: SIAM Journal on Optimization 15.1 (2004), pp. 229–251.
[4] Y. Ouyang, & Y. Xu (2018). Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. arXiv preprint arXiv:1808.02901.

Mirror-Descent (MD) algorithm [5]

- For Euclidean norm MD has the following iteration

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}\left(x_k - \eta \nabla_x g\left(x_k, y_k\right)\right).$$
$$y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(y_k + \eta \nabla_y g\left(x_k, y_k\right)\right).$$

- Iterates and function value do not converge. Let $z = (x, y)$.

$$g(x_k, y) - g(x, y_k) \leq \frac{1}{2\eta}\Big(\underbrace{\|z - z_k\|^2 - \|z - z_{k+1}\|^2}_{\text{telescopes}} + \underbrace{\|z_k - z_{k+1}\|^2}_{\text{residual}}\Big)$$

$$g\Big(\frac{1}{k}\sum_{i=0}^{k-1} x_i, y\Big) - g\Big(x, \frac{1}{k}\sum_{i=0}^{k-1} y_i\Big) \leq \frac{1}{2\,k\,\eta}\Big(\|z - z_0\|^2 + \sum_{i=0}^{k-1}\|z_i - z_{i+1}\|^2\Big)$$

$$\eta = O\Big(\frac{1}{\sqrt{k}}\Big) \implies g\Big(\frac{1}{k}\sum_{i=0}^{k-1} x_i, y\Big) - g\Big(x, \frac{1}{k}\sum_{i=0}^{k-1} y_i\Big) = O\Big(\frac{1}{\sqrt{k}}\Big)$$

[5] A. Nemirovski, D. Yudin, Problem complexity and Method Efficiency in Optimization, Wiley, New York, 1983

(Conceptual) Mirror-Prox (MP) algorithm [3]

- For Euclidean norm MP has the following iteration

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}\left(x_k - \eta \nabla_x g\left(x_{k+1}, y_{k+1}\right)\right)$$
$$y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(y_k + \eta \nabla_y g\left(x_{k+1}, y_{k+1}\right)\right)$$

- Iterates and function value converge $(z = (x, y))$

$$g(x_{k+1}, y) - g(x, y_{k+1}) \leq \frac{1}{2\eta}\Big(\underbrace{\|z - z_k\|^2 - \|z - z_{k+1}\|^2}_{\text{telescopes}} - \underbrace{\|z_k - z_{k+1}\|^2}_{\text{neg. residual}}\Big)$$

$$g\Big(\frac{1}{k}\sum_{i=0}^{k-1} x_i, y\Big) - g\Big(x, \frac{1}{k}\sum_{i=0}^{k-1} y_i\Big) \leq \frac{1}{2\,k\,\eta}\left(\|z - z_0\|^2\right)$$

- Implementable since $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}((x_k, y_k) - \eta\nabla g(x, y))$ is contraction when $\eta L < 1$

Smooth Strongly-Convex–Concave minimax problem

- $g(\cdot,\cdot)$ is L-smooth and $g(x,\cdot)$ is concave

- Additionally, assume that $g(\cdot,y)$ is σ-strongly-convex ($\sigma < L$)

$$g(x,y) + \left\langle \nabla_x g(x,y), x' - x \right\rangle + \frac{\sigma}{2}\|x' - x\|^2 \leq g(x',y)$$

- Then by duality of strong convexity and smoothness the dual problem $h(y) = \min_{x\in\mathcal{X}} g(x,y)$ is $\frac{2L^2}{\sigma}$-smooth and hence differentiable

- Further by Danskin's theorem [6, Section 6.11] $\nabla h(y) = \nabla_y g(x^*(y), y)$ where $x^*(y) = \arg\min_{x\in\mathcal{X}} g(x,y)$

- Dual problem $\min_{y\in\mathcal{Y}} h(y)$ is smooth concave minimization problem.

[6] D. P. Bertsekas. Convex optimization theory. Athena Scientific Belmont, 2009.

Dual Accelerated Gradient Ascent (AGA) method [ours]

- $O(k^{-2})$ AGA [7] on $h(y)$ with $\eta < \sigma/2\,L^2$:

$$\begin{aligned}
\tau_k &= \frac{2}{(k+2)},\ \eta_k = \frac{(k+1)\eta}{2} \\
w_k &= (1-\tau_k)y_k + \tau_k v_k \\
x_k &= \min_{x\in\mathcal{X}} g(x, w_k), \text{ and } y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(w_k + \eta\nabla_y g\left(x_k, w_k\right)\right) \\
v_{k+1} &= \mathcal{P}_{\mathcal{Y}}\left(v_k + \eta_k\nabla_y g\left(x_k, w_k\right)\right)
\end{aligned}$$

- AGA on $g(x_k, \cdot)$ at y_k where $x_k = \arg\min_{x\in\mathcal{X}} g(x, w_k)$
- Accelerated rate on the dual, $h(y_k) - h(y^*) = O(k^{-2})$.
- Still slow rate for primal-dual gap, $f(x_k) - h(y_k) = O(k^{-1})$.

[7] Y.E. Nesterov. "A method for solving the convex programming problem with convergence rate $O(1/k^2)$". In: Dokl. akad. nauk Sssr. Vol. 269. 1983, pp. 543–547.

Dual Accelerated Gradient Ascent (AGA) method is slow

- Consider $\min_{x\in[-1,1]} \max_{y\in[-1,1]} g(x,y) = x^2/2 + xy$.
- Then $h(y) = -y^2/2$, $f(x) = x^2/2 + |x|$, and $(x^*, y^*) = (0,0)$
- Let $h(y_k) - h(y^*) = \Theta(k^{-2}) \implies |y_k| = \Theta(k^{-1})$.
- Let $x_k = \arg\min_{x\in\mathcal{X}} g(x, y_k) = -y_k, \implies |x_k| = |y_k| = \Theta(k^{-1})$.
- Thus $f(x_k) - f(x^*) = x_k^2/2 + |x_k| = \Theta(k^{-1})$

Dual Implicit Accelerated Gradient (DIAG) method [ours]

- For each k, apply AGA step of $g(x_{k+1}, \cdot)$

$$
\begin{aligned}
\tau_k &= \frac{2}{(k+2)},\ \eta_k = \frac{(k+1)\eta}{2} \\
w_k &= (1-\tau_k) y_k + \tau_k v_k \\
x_{k+1} &= \arg\min_{x \in \mathcal{X}} g(x, y_{k+1}), \text{ and } y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(w_k + \eta \nabla_y g\left(x_{k+1}, w_k\right)\right) \\
v_{k+1} &= \mathcal{P}_{\mathcal{Y}}\left(v_k + \eta_k \nabla_y g\left(x_{k+1}, w_k\right)\right)
\end{aligned}
$$

- AGA on $g(x_k, \cdot)$ at y_k where $x_k = \arg\min_{x \in \mathcal{X}} g(x, y_{k+1})$
- Primal-dual gap inherits the accelerated $O(k^{-2})$ convergence of dual $h(y_k) = \min_{x \in \mathcal{X}} g(x, y_k)$

$$
g\left(\frac{1}{k}\sum_{i=1}^{k}(2i) \cdot x_i, y\right) - g(x, y_k) \leq \frac{2\left\|y - y_0\right\|^2}{k\,(k+1)\,\eta}
$$

Implementable DIAG

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} g(x, y_{k+1}), \text{ and } y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(w_k + \eta \nabla_y g\left(x_{k+1}, w_k\right)\right)$$

- Since $\eta < 2L^2/\sigma$, the following operator $(\cdot)^+ : \mathcal{Y} \to \mathcal{Y}$ is a $1/2$-contraction

$$\begin{aligned} x^*(y) &= \arg\min_{x\in\mathcal{X}} g(x, y) \\ (y)^+ &= \mathcal{P}_{\mathcal{Y}}\left(w_k + \eta \nabla_y g\left(x^*(y), w_k\right)\right). \end{aligned}$$

- Thus $(x_k^{(i)}, y_k^{(i)})$ converges approximately to (x_{k+1}, y_{k+1}) in $O(\log(\frac{1}{\varepsilon}))$ steps

$$\begin{aligned} x_k^{(i)} &= \arg\min_{x\in\mathcal{X}} g(x, y_k^{(i)}) \\ y_k^{(i+1)} &= \mathcal{P}_{\mathcal{Y}}\left(w_k + \eta \nabla_y g\left(x_k^{(i)}, w_k\right)\right). \end{aligned}$$

Summary and my contributions

- We studied smooth minimax opitmization problem
- Improved $O(\varepsilon^{-3})$ algorithm for smooth Nonconvex–Concave problem
- Optimal $O(k^{-2})$ algorithm for smooth Strongly-convex–Concave problem