Efficient Algorithms for Smooth Minimax Optimization

Kiran Thekumparampil[†] Prateek Jain[‡] Praneeth Netrapalli[‡] Sewoong Oh[±]
[†]University of Illinois at Urbana-Champaign [‡]Microsoft Research, India [±]University of Washington, Seattle

https://github.com/POLane16/DIAG

Smooth Minimax problem

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

where g is L-smooth

$$\|\nabla_x g(x,y) - \nabla_x g(x',y')\| \le L_{xx}\|x-x'\| + L_{xy}\|y-y'\|$$
$$\|\nabla_y g(x,y) - \nabla_y g(x',y')\| \le L_{yx}\|x-x'\| + L_{yy}\|y-y'\|$$

Convex–Concave Minimax problem

- $g(\cdot,y)$ is convex in x and $g(x,\cdot)$ is concave in y
- Minimax Theorem: if $\mathcal{X}/\mathcal{Y}$ is compact or if $g(\cdot,y)/g(x,\cdot)$ is strongly-convex/concave:

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

- ε-primal dual pair $(\tilde{x},\tilde{y})$

$$\max_{y\in\mathcal{Y}} g(\tilde{x},y) - \min_{x\in\mathcal{X}} g(x,\tilde{y}) \le \varepsilon$$

- For a L-smooth convex function $f(x)$
  - Gradient Descent (GD): $x_{k+1} = \mathcal{P}_{\mathcal{X}}(x_k - \eta\nabla f(x_k))$
  - Proximal Point Method (PPM): $x_{k+1} = \mathcal{P}_{\mathcal{X}}(x_k - \eta\nabla f(x_{k+1}))$

Algo.	Update	Step	Rate
Mirror Descent	$x_k - \eta\nabla_x g(x_k,y_k)$	GD on $g(\cdot,y_k)$	$O\left(k^{-1/2}\right)$
Mirror-Prox [3]	$x_k - \eta\nabla_x g(x_{k+1},y_{k+1})$	PPM on $g(\cdot,y_{k+1})$	$\tilde{O}\left(k^{-1}\right)$
C-MD	$x_k - \eta\nabla_x g(x_k,y_{k+1})$	GD on $g(\cdot,y_{k+1})$	$\tilde{O}\left(k^{-1}\right)$

- Looking ahead in the other variable accelerates the minimax optimization

Strongly Convex–Concave Minimax problem

- $g(\cdot,y)$ is σ_x-strongly convex in x

$$g(x,y) + \langle\nabla_x g(x,y), x'-x\rangle + \frac{\sigma_x}{2}\|x'-x\|^2 \le g(x',y)$$

- Dual $h(y) = \min_{x\in\mathcal{X}} g(x,y)$ a $L_{xx} + L_{xy}^2/\sigma_x$-smooth concave function
- Apply Accelerated Gradient Ascent (AGA) on the dual function $h(y)$

DIAG (Dual Implicit Accelerated Gradient)

DAG (Dual Accel. Gradient)

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

DIAG (Dual Implicit Accel. Gradient)

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

Algo.	Gradient used	Step	Dual Optimality $h(y_k) - h(y^*)$	Primal Dual Gap $f(x_k) - h(y_k)$
DAG	$\nabla_y g(x_k,w_k)$	AGA on $g(x_k,\cdot)$	$O\left(k^{-2}\right)$	$O\left(k^{-1}\right)$
DIAG	$\nabla_x g(x_{k+1},w_k)$	AGA on $g(x_{k+1},\cdot)$	$\tilde{O}\left(k^{-2}\right)$	$\tilde{O}\left(k^{-2}\right)$

Implementatble DIAG

- Mirror-Prox: $(x_{k+1},y_{k+1}) = (x_k,y_k) - \eta(\nabla_x g(x_k,y_k), -\nabla_y g(x_{k+1},y_{k+1}))$
  - $\mathcal{O}(x,y) = (x_k,y_k) - \eta(\nabla_x g(x,y), -\nabla_y g(x,y))$ is contraction if $\eta L < 1$
  - Fixed point of $\mathcal{O}$, (x_{k+1},y_{k+1}) can be found in $O\left(\log 1/\varepsilon\right)$ steps
- DIAG: $x_{k+1} = \arg\min_{x\in\mathcal{X}} g(x,y_{k+1}), y_{k+1} = w_k + \eta\nabla_y g(x_{k+1},w_k))$
  - $\mathcal{O}(y) = \mathcal{P}_{\mathcal{Y}}(w_k + \eta\nabla_y g(x^*(y),w_k))$ is contraction if $2\,\eta L_{xy}^2/\sigma_x < 1$, where $x^*(y) = \min_{x\in\mathcal{X}} g(x,y)$
  - $x^*(y)$ can be found in $O\left(\sqrt{L_{xx}/\sigma_x}\log 1/\varepsilon\right)$ steps using AGD
  - Fixed point of $\mathcal{O}$, y_{k+1} can be found in $O\left(\sqrt{L_{xx}/\sigma_x}\log^2 1/\varepsilon\right)$ steps

Theorem 1 (Convergence rate of DIAG). *After K iterations, DIAG finds $(\frac{1}{K}\sum_{k=1}^K x_k, y_K)$ s.t.:*

$$\max_{\tilde{y}\in\mathcal{Y}} g\left(\frac{1}{K}\sum_{k=1}^K x_k,\tilde{y}\right) - \min_{\tilde{x}\in\mathcal{X}} g(\tilde{x},y_K) \,\tilde{\le}\, \frac{4\max\{L_{yy}, 2\frac{L_{xy}^2}{\sigma}\}D_{\mathcal{Y}}^2}{K(K+1)},$$

and these K iterations require $O(\sqrt{\frac{L_{xx}}{\sigma_x}}K\log^2(K))$ first order gradient oracle calls.

- Total complexity $\tilde{O}\left(\sqrt{\frac{L_{xx}}{\sigma_x}}\sqrt{L_{yy} + \frac{L_{xy}^2}{\sigma_x}}\frac{1}{\sqrt{\varepsilon}}\right)$, matches lower bound $O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ [4]
- Our rate can also be obtained by a simpler smoothing technique [5]

Nonconvex–Concave Minimax problem

- $g(\cdot,y)$ is nonconvex, but $g(\cdot,y)$ is concave.
- We focus on the primal problem $f(x) = \max_{y\in\mathcal{Y}} g(x,y)$, and not the dual $\min_{x\in\mathcal{X}} g(x,y)$
- As f is nonsmooth, optimality defined using L-smoothness of g, which implies L_{xx}-weak convexity of f

$$L_{xx}\text{-smoothness of } g(\cdot) \implies f(x) + \langle\partial f(x), x'-x\rangle - \frac{L_{xx}}{2}\|x'-x\|^2 \le f(x')$$

- ε-FOSP (First Order Stationary Point)

$$\|\nabla f_{\frac{1}{2L_{xx}}}(x)\| \le \varepsilon, \text{ where, } \quad f_{\frac{1}{2L_{xx}}}(x) = \min_{x'} f(x') + L_{xx}\|x'-x\|^2$$

Prox-DIAG (Proximal DIAG)

	Subgrad method [1, 2]	Proximal point method
Exact	$x_{k+1} = x_k - \eta\partial f(x_k)$	$x_{k+1} = x_k - \eta\partial f(x_{k+1})$
Approx.	$\max_y g(x_k,y) - O(\varepsilon^2) \le g(x_k,y_k)$ $x_{k+1} = x_k - \eta\nabla g(x_k,y_k)$	$f_k(x) = \max_y g(x,y) + L_{xx}\|x-x_k\|^2$ $f_k(x_{k+1}) \le \min_x f_k(x) + O\left(\varepsilon^2\right)$
#iter.	$O\left(1/\varepsilon^4\right)$	$O\left(1/\varepsilon^2\right)$
per-step	$O\left(1/\varepsilon\right)$ [AGD]	$O\left(1/\varepsilon\right)$ [DIAG]
total	$O\left(1/\varepsilon^5\right)$	$O\left(1/\varepsilon^3\right)$

Implementing Prox-DIAG

- Prox-DIAG step finds x_{k+1} such that,

$$\max_{y\in\mathcal{Y}} g(x_{k+1},y) + L_{xx}\|x_{k+1}-x_k\|^2 \le \min_x \max_{y\in\mathcal{Y}} g(x_{k+1},y) + L_{xx}\|x-x_k\|^2 + O\left(\varepsilon^2\right)$$

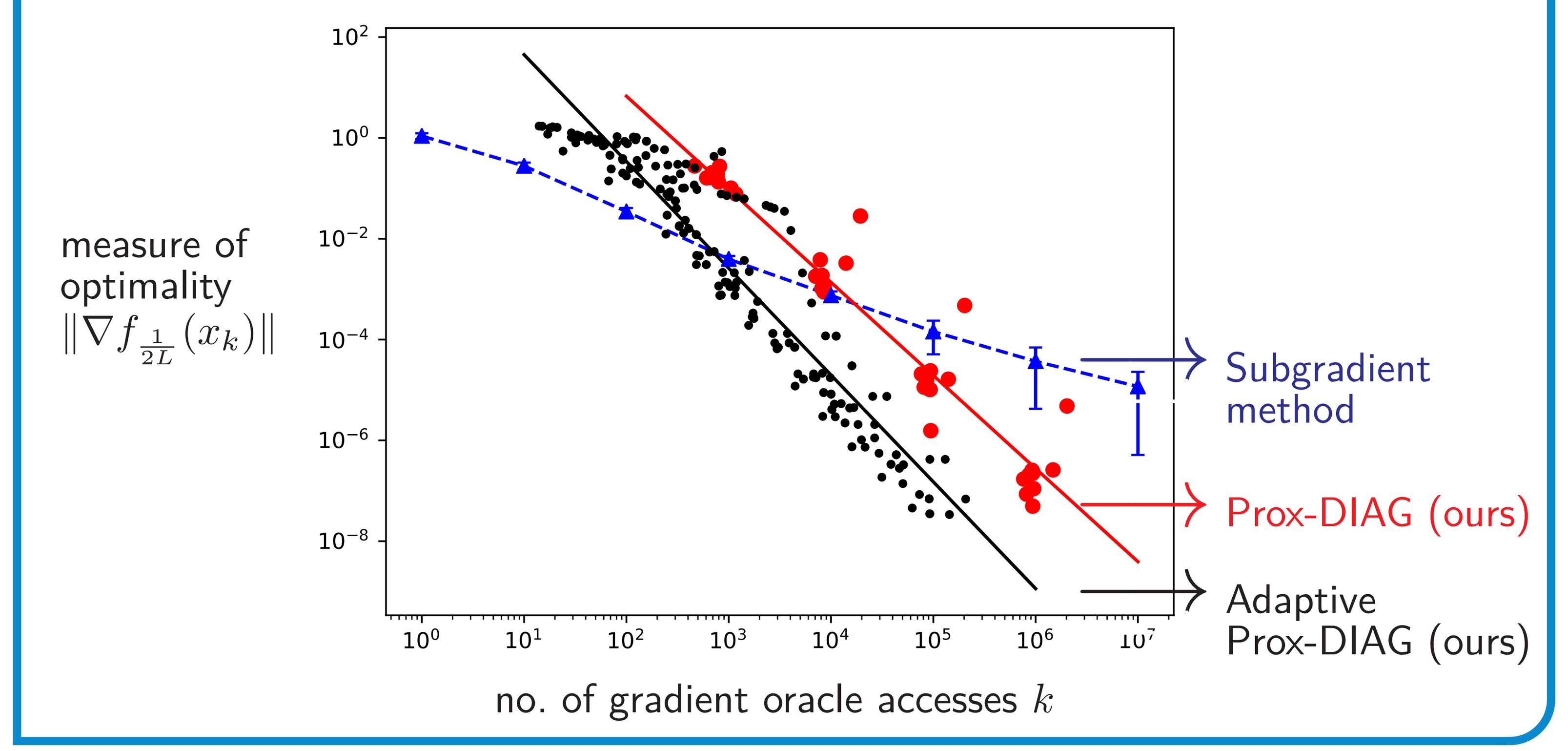
- L_{xx}-weak convexity of $g(\cdot,y) \implies L_{xx}$-strong convexity of $g(\cdot,y) + L_{xx}\|\cdot - x_k\|^2$
- DIAG solves L-smooth, L_{xx}-strongly-convex–concave problem in $O\left(1/\varepsilon\right)$ steps
- By weak-convexity outer loop find a ε-FOSP in $O\left(1/\varepsilon^2\right)$ steps.

$$x_{k+1} = x_k - L_{xx}\partial f(x_{k+1}) \overset{L_{xx}\text{-weakly convex}}{\Longrightarrow} f(x_{k+1}) \le f(x_k) - 3L_{xx}/2\|\partial f(x_{k+1})\|^2$$

- Total first order (gradient) oracle complexity is $O\left(1/\varepsilon^3\right)$
- Similar rate obtained using smoothing technique [6]

Experiments: Nonconvex–Concave

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



References

[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.
[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.
[5] M. Alkousa, D. Dvinskikh, F. Stonyakin, and A. Gasnikov, 2019. Accelerated methods for composite non-bilinear saddle point problem. arXiv preprint arXiv:1906.03620.
[6] W. Kong & R. Monteiro. "An accelerated inexact proximal point method for solving nonconvex-concave min-max problems." arXiv preprint arXiv:1905.13433 (2019).