Adaptive Extra-Gradient Methods for Min-Max Optimization and Games

K. Antonakopoulos[1]
joint with
E. Veronica Belmega[2] Panayotis Mertikopoulos[13]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG
[2]ETIS, CY Cergy Paris Université, ENSEA, CNRS, UMR 8051
[3]Criteo AI Lab

International Conference on Learning Representations, 3-7 May 2021

Problem setup

- $\mathcal{X} \subset \mathbb{R}^d$ *convex*
- $A : \mathcal{X} \to \mathcal{X}^*$ monotone:

$$\langle A(x) - A(x') \,|\, x - x' \rangle \geq 0 \text{ for all } x, x' \in \mathcal{X}$$

Main goal:

$$\text{Find } x^* \in \mathcal{X} \text{ s.t. } \langle A(x^*) \,|\, x - x^* \rangle \geq 0 \text{ for all } x \in \mathcal{X} \tag{VI}$$

Applications: **Convex Minimization, Saddle points, Nash equilibria...**

Performance Evaluation

Restricted gap function [Auslender,. . .]

$$Gap_{\mathcal{C}}(x) = \sup_{x' \in \mathcal{C}} \langle A(x') \,|\, x - x' \rangle \tag{RGF}$$

for every $\mathcal{C}$ compact neighbourhood x^*

Its zeros characterize the solution of (VI)

Standard regularity conditions

- **Boundedness:**

$$\|A(x)\|_* \leq G \ \text{ for all } \ x \in \mathcal{X}$$

- **Smoothness:**

$$\|A(x) - A(x')\|_* \leq L\|x - x'\|$$

- Boundedness:

$$Gap(X_T) = \mathcal{O}(1/\sqrt{T}) \qquad \text{[tight]}$$

- Smoothness: Nemirovski [2004], Nesterov [2007]

$$Gap(X_T) = \mathcal{O}(1/T) \qquad \text{[tight]}$$

Standard regularity conditions

- **Boundedness:**

$$\|A(x)\|_* \leq G \text{ for all } x \in \mathcal{X}$$

- **Smoothness:**

$$\|A(x) - A(x')\|_* \leq L\|x - x'\|$$

- Boundedness:

$$Gap(X_T) = \mathcal{O}(1/\sqrt{T}) \qquad \text{[tight]}$$

- Smoothness: Nemirovski [2004], Nesterov [2007]

$$Gap(X_T) = \mathcal{O}(1/T) \qquad \text{[tight]}$$

- Adaptive: Bach and Levy [2019]

Best of both worlds!

Limitations of SOTA

2 types of Boundedness:

Limitations of SOTA

2 types of Boundedness:

- Boundedness of the operator or its variation (smoothness)

> Problems with singularities are excluded
> examples: Poisson Inverse Problems, D-Optimal Design, Support Vector Machines etc

Limitations of SOTA

2 types of Boundedness:

- Boundedness of the operator or its variation (smoothness)

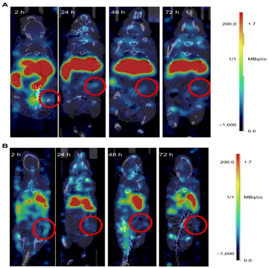
  Problems with singularities are excluded
  examples: Poisson Inverse Problems, D-Optimal Design, Support Vector Machines etc

- Boundedness of the domain for adaptive methods

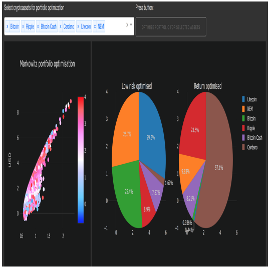
  Inappropriate for problems with unbounded and/or non-compact domains

Why bother?

Various real-life problems are not ”bounded”!



Positron emission tomography



Portfolio selection

Image denoising

Main objective

Design a adaptive algorithm that transcends the limitations of SOTA

Main objective

Design a adaptive algorithm that transcends the limitations of SOTA

Our approach:

- A broader class to account problems with singularities!

Main objective

Design a adaptive algorithm that transcends the limitations of SOTA

Our approach:

- A broader class to account problems with singularities!
- A tailor-made Bregman methods to achieve order optimal rate interpolation/ domain indifference!

Metric Boundedness and Smoothness

Key observations

- Lipschitz continuity is a metric space property
- Add geometry awareness via a suitable family of local norms

Metric Boundedness and Smoothness

Key observations

- Lipschitz continuity is a metric space property
- Add geometry awareness via a suitable family of local norms

- **Metric Boundedness:** (A, Belmega, Mertikopoulos, ICLR 2020)

$$\|A(x)\|_{x,*} \leq G \tag{MB}$$

- **Metric Smoothness:** (A, Belmega, Mertikopoulos, NeurIPS 2019)

$$\|A(x) - A(x')\|_{x,*} \leq L\|x - x'\|_{x'} \tag{MS}$$

Our Method: AdaProx

AdaProx=Mirror-Prox + Local Norm Twist + Adaptive step-size

Our Method: AdaProx

AdaProx=Mirror-Prox + Local Norm Twist + Adaptive step-size

Key elements: Mirror-Prox template:

$$\begin{aligned} x_{t+1/2} &= \arg\min_{x\in\mathcal{X}}\left\{\langle A(X_t)\,|\,x - X_t\rangle + \tfrac{1}{\gamma_t}D_h(x, X_t)\right\} \\ x_{t+1} &= \arg\min_{x\in\mathcal{X}}\left\{\langle A(X_{t+1/2})\,|\,x - x_t\rangle + \tfrac{1}{\gamma_t}D_h(x, x_t)\right\} \end{aligned} \tag{MP}$$

Bregman divergence: $D(x', x) = h(x') - h(x) - \langle \nabla h(x), x' - x\rangle$

Our Method: AdaProx

AdaProx=Mirror-Prox + Local Norm Twist + Adaptive step-size

Key elements: Mirror-Prox template:

$$
\begin{aligned}
x_{t+1/2} &= \arg\min_{x\in\mathcal{X}}\left\{\langle A(X_t)\,|\,x - X_t\rangle + \tfrac{1}{\gamma_t} D_h(x, X_t)\right\} \\
x_{t+1} &= \arg\min_{x\in\mathcal{X}}\left\{\langle A(X_{t+1/2})\,|\,x - x_t\rangle + \tfrac{1}{\gamma_t} D_h(x, x_t)\right\}
\end{aligned}
\tag{MP}
$$

Bregman divergence: $D(x', x) = h(x') - h(x) - \langle \nabla h(x), x' - x\rangle$

Local Norm twist: Take h **strongly convex w.r.t. the local norms**, i.e.

$$h(x') \geq h(x) + \langle \nabla h(x), x' - x\rangle + \frac{K}{2}\|x - x'\|_x^2$$

Defining the adaptive step-size

Adaptive Step-size:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|A(X_j) - A(X_{j+1/2})\|^2_{X_{j+1/2},*}}} \tag{1}$$

Defining the adaptive step-size

Adaptive Step-size:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|A(X_j) - A(X_{j+1/2})\|^2_{X_{j+1/2},*}}} \tag{1}$$

- "Worst Case" (MB)

$$\|A(X_j) - A(X_{j+1/2})\|^2_{X_{j+1/2},*} \approx \text{"constant"}$$

and hence $\gamma_t \propto 1/\sqrt{t}$

Defining the adaptive step-size

Adaptive Step-size:

$$\gamma_t = \frac{1}{\sqrt{1 + \sum_{j=1}^{t-1} \|A(X_j) - A(X_{j+1/2})\|^2_{X_{j+1/2},*}}} \tag{1}$$

- "Worst Case" (MB)

$$\|A(X_j) - A(X_{j+1/2})\|^2_{X_{j+1/2},*} \approx \text{"constant"}$$

  and hence $\gamma_t \propto 1/\sqrt{t}$
- "Best Case" (MS)

$$\|A(X_j) - A(X_{j+1/2})\|^2_{X_{j+1/2},*} \propto \|X_{j+1/2} - X_j\|_{X_j}$$

  which converges to 0 if the algorithm converges → Bigger step-size

Results

Theorem

If $\overline{X}_T = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t X_{t+1/2}$, *then:*

- *Under* (MB), *then:*

$$\mathit{Gap}_{\mathcal{C}}(\overline{X}_T) = \mathcal{O}(1/\sqrt{T})$$

- *Under* (MS) *then:*

$$\mathit{Gap}_{\mathcal{C}}(\overline{X}_T) = \mathcal{O}(1/T)$$

- *Under* (MB) *or* (MS), *the iterates of* **(AdaProx)** *converge to the solution set* $\mathcal{X}^*$*:*

$$\mathit{dist}(X_t, \mathcal{X}^*) \to 0 \quad \mathit{dist}(X_{t+1/2}, \mathcal{X}^*)$$

F. Bach and K. Y. Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *COLT '19: Proceedings of the 32nd Annual Conference on Learning Theory*, 2019.

A. S. 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. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.