Breaking the Curse of Dimensionality (Locally) to Accelerate Conditional Gradients

Jelena Diakonikolas [1], Alejandro Carderera [2], Sebastian Pokutta [3, 4]

[1]UC Berkeley, [2]Georgia Institute of Technology, [3]Zuse Institute Berlin, [4]Technische Universität Berlin

Abstract

Conditional gradients constitute a class of *projection-free* first-order algorithms for smooth convex optimization that do not enjoy the globally optimal convergence rates achieved by projection-based accelerated methods. We present *Locally Accelerated Conditional Gradients* (LACG), which couples accelerated and conditional gradient steps to achieve *optimal accelerated local convergence* on smooth strongly convex problems and does not require projections onto the feasible set.

Motivation

$$\min_{x\in P} f(x) \tag{1}$$

Goal is L-smooth μ-strongly convex optimization over polytope P with:

1. **First-order (FO) oracle.**
2. **Linear optimization (LO) oracle.**

Focus on the *Conditional Gradients* (CG) (a.k.a. the *Frank-Wolfe*) algorithm [1, 2], and its variants, such as the *Away-step* CG algorithm.

Convergence rate of CG variants

[3] The number of steps T required to reach an ϵ-optimal solution to Problem 1:

$$T = \mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right),$$

where D and δ are the diameter and pyramidal width of P.

The rates of first-order optimal projection-based methods [4]: 1) Depend on $\sqrt{L/\mu}$ and 2) Do not depend on the dimension.

These rates cannot be achieved *globally* [5] with the LO oracle, but:

Can CG achieve these rates ***locally****?*

Locally Accelerated Conditional Gradients

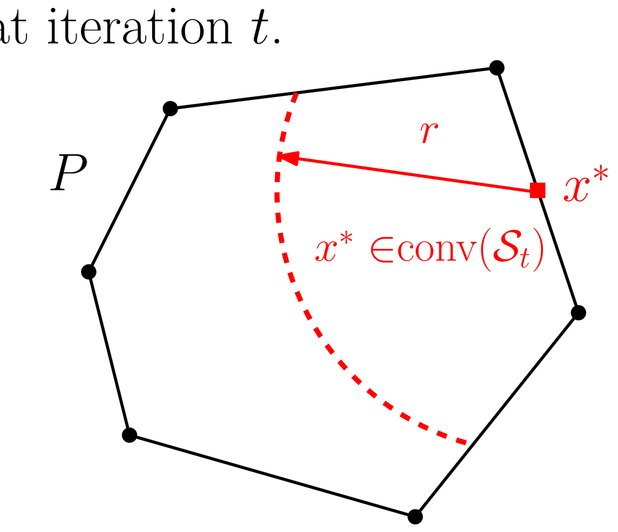
Analysis based on the *Approximate Duality Gap technique* [6] and a *Modified μAGD+* algorithm [7], that requires projections onto (typically low dimensional) simplices.

Convergence rate of the Modified μAGD+ algorithm.

Let $\{\mathcal{C}_i\}_{i=0}^t$ be a sequence of convex subsets of P such that $\mathcal{C}_i \subseteq \mathcal{C}_{i-1}$ for all i and $x^* \in \cap_{i=0}^t \mathcal{C}_i$. The number of steps T required to reach an ϵ-optimal solution to Problem 1:

$$T = \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$$

We also know: $\exists r > 0$ s.t. if $\|x^* - x_K\| \leq r \Rightarrow x^* \in \operatorname{conv}(\mathcal{S}_t)$ for all $t \geq K$, where $\mathcal{S}_t$ is the CG active set at iteration t.



If we use $\mathcal{C}_t = \mathcal{S}_t$ inside the semicircle, acceleration is possible with the Modified μAGD+ algorithm.

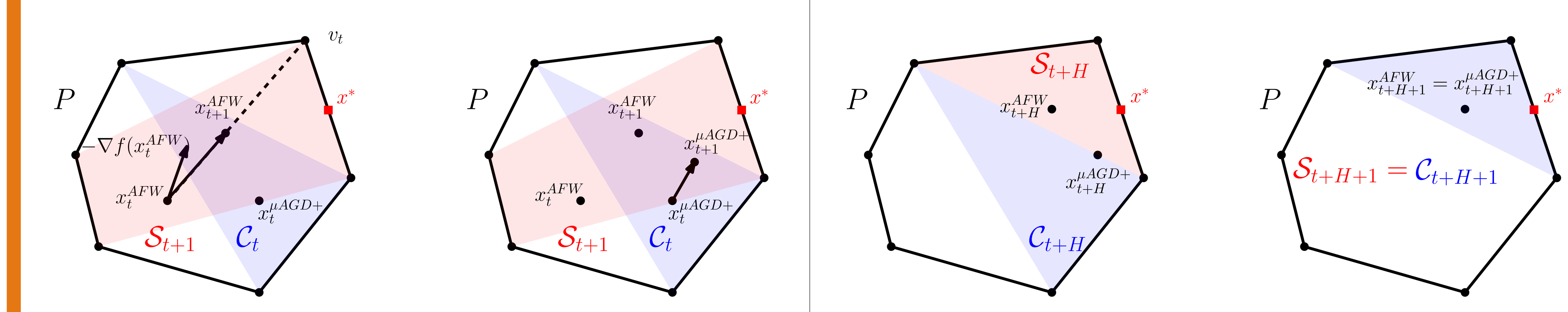
Main Idea: Combine a linearly convergent CG that maintains an active set (e.g. AFW) with a Modified μAGD+ so that when $\|x^* - x_t\| \leq r$:

Local acceleration!

Algorithm Main Ideas

1. Perform AFW step and Modified μAGD+ step (the latter over current $\mathcal{C}$):

2. Every $H = 2\sqrt{\frac{2L}{\mu}}\log(L/\mu - 1)$ iterations, restart the Modified μAGD+ algorithm and update $\mathcal{C}$ if a vertex was added to $\mathcal{S}$ since the last update.



(a) AFW step (b) Modified μAGD+ step (c) Before restart. (d) After restart.

3. For every iteration: $x_{t+1} = \operatorname{argmin}_{x\in\{x_t^{AFW}, x_t^{\mu AGD+}\}} f(x)$. This ensures *monotonicity*.

Convergence rate of LaCG

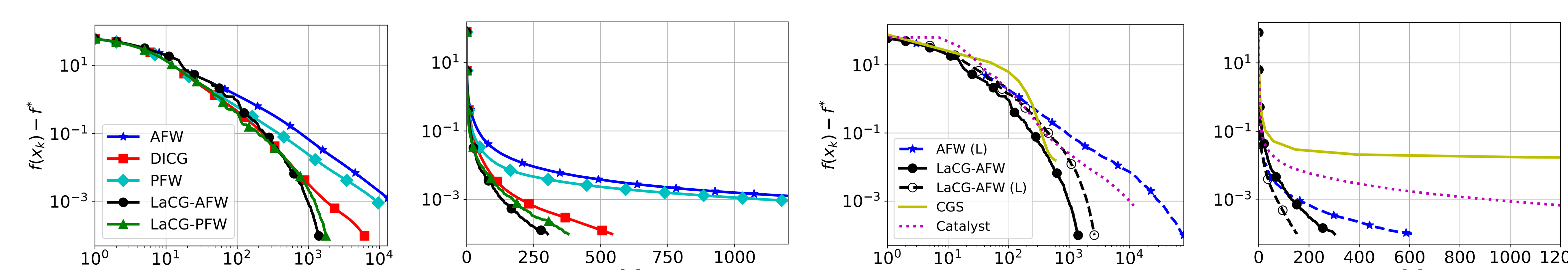
Considering Problem 1, let r be the critical radius associated with P. If:

$$t = \min\left\{\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2\log\frac{1}{\epsilon}\right), K + \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)\right\}$$

where $K = \frac{8L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\left(\frac{2(f(x_0)-f^*)}{\mu r^2}\right)$, then $f(x_t) - f(x^*) \leq \epsilon$

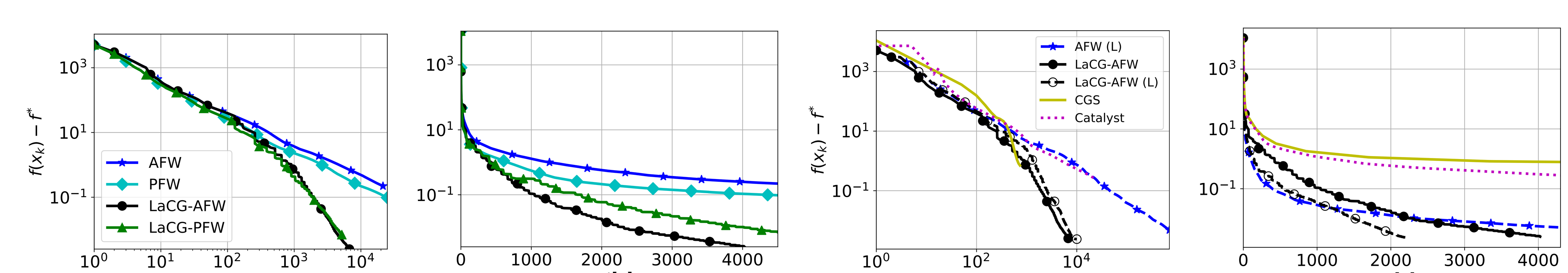
LaCG achieves local acceleration

Computational Results



(e) Iteration (f) Wall-clock time (g) Iteration (h) Wall-clock time

Figure 1: Birkhoff polytope: Algorithm comparison in terms of (e),(g) iteration count and (f),(h) wall-clock time.



(a) Iteration (b) Wall-clock time (c) Iteration (d) Wall-clock time

Figure 2: MIPLIB polytope: Algorithm comparison in terms of (a),(c) iteration count and (b),(d) wall-clock time for the `ran14x18-disj-8` polytope from the MIPLIB library.

Paper: https://opt-ml.org/papers/2019/paper_26.pdf

Code: https://github.com/pokutta/lacg

References

[1] B. T. Polyak, "Minimization methods in the presence of constraints," *Itogi Nauki i Tekhniki. Seriya" Matematicheskii Analiz"*, 1974.

[2] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[3] S. Lacoste-Julien and M. Jaggi, "On the global linear convergence of Frank-Wolfe optimization variants," in *Advances in Neural Information Processing Systems*, 2015.

[4] Y. Nesterov, *Lectures on convex optimization*, vol. 137, Springer, 2018.

[5] G. Lan, "The complexity of large-scale convex programming under a linear optimization oracle," 2013.

[6] J. Diakonikolas and L. Orecchia, "The approximate duality gap technique: A unified theory of first-order methods," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 660–689, 2019.

[7] M. B. Cohen, J. Diakonikolas, and L. Orecchia, "On acceleration with noise-corrupted gradients," *35nd International Conference on Machine Learning, ICML 2018*, 2018.