Parameter-free Locally Accelerated Conditional Gradients

Alejandro Carderera [1,5], Jelena Diakonikolas [2, 5], Cheuk Yin Lin [2, 5], Sebastian Pokutta [3, 4, 5]

[1]Georgia Institute of Technology, [2]University of Wisconsin-Madison,
[3]Zuse Institute Berlin, [4]Technische Universität Berlin,
[5]Authors ordered alphabetically

Abstract

Conditional gradient (CG) methods are the algorithms of choice for constrained optimization when projections are computationally prohibitive but linear optimization over the constraint set remains possible. Unlike in projection-based methods, globally accelerated convergence rates are in general unattainable for CG. One can achieve *local acceleration* with knowledge of the smoothness and strong convexity parameters of the function [1]. We remove this limitation by introducing the Parameter-Free Locally accelerated CG (PF-LaCG) algorithm.

Motivation

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \qquad (1)$$

Goal is L-smooth μ-strongly convex optimization over a polytope $\mathcal{X}$ with **First Order Oracles** (FOO) and **Linear Minimization Oracle** (LMO). Focus on the *Conditional Gradients* (CG) algorithm [2, 3], and its variants, such as the *Away-step Frank-Wolfe* (AFW) algorithm.

Convergence rate of CG variants

The number of steps T required to reach an ϵ-optimal solution to Problem (1) [4]:

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

where D and δ are the diameter and pyramidal width of $\mathcal{X}$, and D/δ is dimension-dependent.

The rates of first-order optimal projection-based methods [5]: 1) Depend on $\sqrt{L/\mu}$ and 2) Do not depend on the dimension.

These rates cannot be achieved *globally* [6] with the LMO, but they can be achieved locally if we know L and μ [1]:

*Can CG achieve these rates locally without knowing L and μ? **Yes!***

Parameter-Free Locally Accelerated Conditional Gradients

Our contributions are:

1. Parameter-free Locally-accelerated Conditional Gradient (PF-LaCG) algorithm.
2. Near-optimal and parameter-free accelerated algorithm (ACC) with inexact projections.

We achieve local acceleration by coupling the AFW and ACC algorithm and restarting when an **upper bound on the primal gap** is halved:

$$w(\mathbf{x},\mathcal{S}) \stackrel{\text{def}}{=} \max_{\mathbf{u}\in\mathcal{X},\mathbf{v}\in\mathcal{S}} \langle \nabla f(\mathbf{x}), \mathbf{u}-\mathbf{v}\rangle .$$

where $\mathcal{S}$ is a proper support. This allow us to maintain a computable global measure of optimality without knowing L and μ and couple the AFW and ACC algorithms while guaranteeing monotonic progress in $w(\mathbf{x},\mathcal{S})$.

Convergence rate of ACC

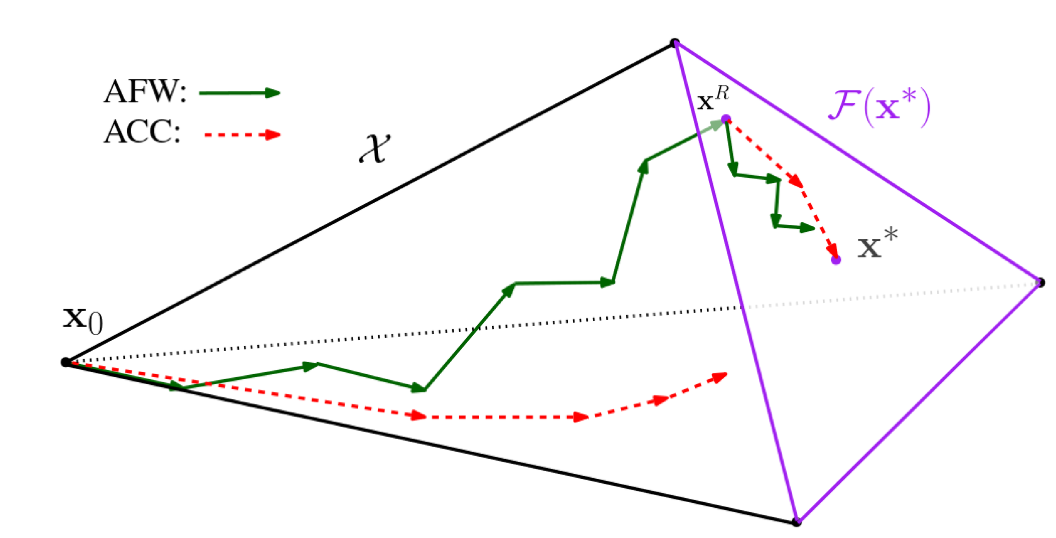
Let $C \subseteq \mathbb{R}^n$ be a closed convex set, such that $\mathbf{x}^* \subset C$. Then running the ACC with properly initialized parameters over C outputs a point with dual optimality gap smaller than ϵ with a total number of

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

queries to the FOO for f and an inexact and efficiently computable projection oracle for C, without knowledge of L or μ.

Algorithm Main Ideas

1. PF-LaCG runs AFW and ACC in parallel, and restarts every time AFW halves $w(\mathbf{x},\mathcal{S})$. After every restart choose point with lower value of $w(\mathbf{x},\mathcal{S})$ and potentially update active set of ACC
2. After a finite number of iterations independent of ϵ, the active set of AFW contains $\mathbf{x}^*$ **and** ACC converges to the optimum at an accelerated rate



Convergence rate of PF-LaCG

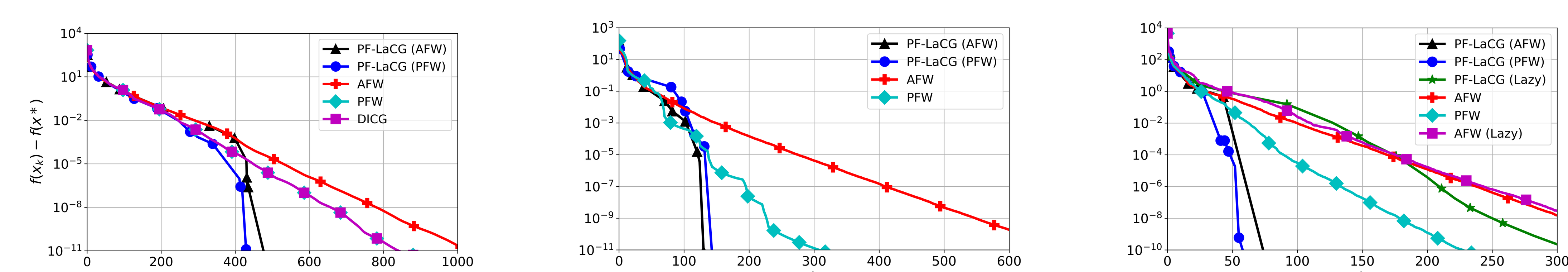
Let f be L-smooth and μ-strongly convex. The number of calls to FOO and LMO required to reach an ϵ-optimal solution, measured in terms of $w(\mathbf{x},\mathcal{S})$, to the minimization problem satisfies:

$$T = \min\left\{\underbrace{O\left(\frac{LD^2}{\mu\delta^2}\log\frac{1}{\epsilon}\right)}_{\text{AFW bound}},\ \underbrace{K}_{\text{Burn-in}} + \underbrace{O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{L}{\mu}\right)\log\left(\frac{LD}{\mu\delta}\right)\log\frac{1}{\epsilon}\right)}_{\text{Locally-accelerated convergence}}\right\},$$

where K is a constant that is independent of ϵ.

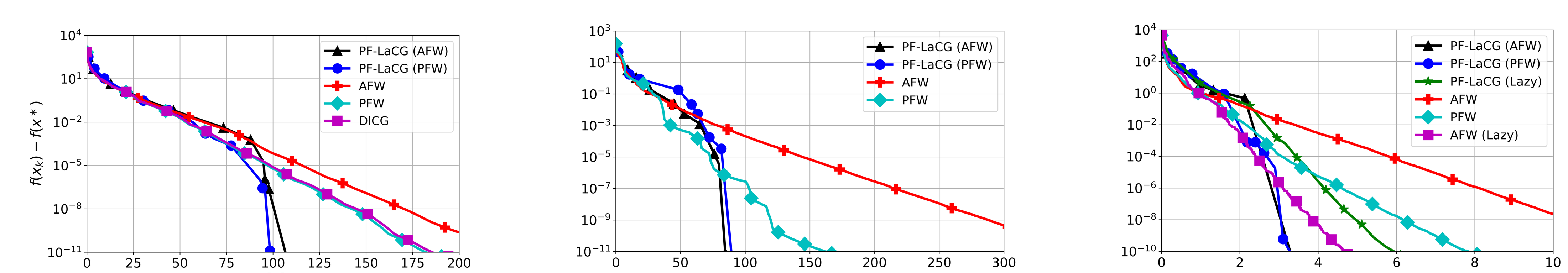
PF-LaCG achieves parameter-free local acceleration (optimal up to poly-log factors)

Computational Results



(b) Probability simplex problem (c) Structured lasso problem (d) Structured Birkhoff problem

Figure 1: Performance w.r.t. iteration count: Algorithm comparison for a strongly-convex and smooth quadratic problem over the probability simplex (b), a structured lasso feasible region (c), and a structured Birkhoff polytope domain (d) with respect to iteration count.



(a) Probability simplex problem (b) Structured lasso problem (c) Structured Birkhoff problem

Figure 2: Performance w.r.t. wall-clock time: Algorithm comparison for a strongly-convex and smooth quadratic problem over the probability simplex (a), a structured lasso feasible region (b), and a structured Birkhoff polytope domain (c) with respect to wall-clock time.

Paper: https://arxiv.org/pdf/2102.06806.pdf **Code**: https://github.com/ericlincc/Parameter-free-LaCG

References

[1] J. Diakonikolas, A. Carderera, and S. Pokutta, "Locally accelerated conditional gradients," in *Proc. AISTATS'20*, 2020.

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

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

[4] S. Lacoste-Julien and M. Jaggi, "On the global linear convergence of Frank-Wolfe optimization variants," in *Proc. NIPS'15*, 2015.

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

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