Parameter-free Locally Accelerated Conditional Gradients

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

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

ICML 2021

Authors in alphabetical order.

Goal is L-smooth μ-strongly convex optimization over polytope $\mathcal{X}$.

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

Main ingredients:

First Order Oracle (FOO). Given $\mathbf{x} \in \mathcal{X}$ and a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$, return:

$$\nabla f(\mathbf{x}) \in \mathbb{R}^n \text{ and } f(\mathbf{x}) \in \mathbb{R}$$

Linear Minimization Oracle (LMO). Given $\mathbf{v} \in \mathbb{R}^n$, return:

$$\operatorname*{argmin}_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{v}, \mathbf{x} \rangle$$
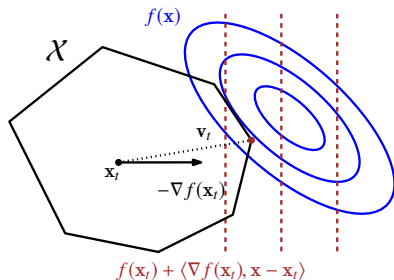
Focus on the *Conditional Gradients/Frank-Wolfe* algorithm [FW56; Pol74] and its variants. The basic variant is:

CG Algorithm.

Input: $\mathbf{x}_0 \in \mathcal{X}$, stepsizes $\gamma_t \in (0, 1]$.

1: **for** $t = 0$ to T **do**
2: $\quad \mathbf{v}_t = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} \langle \nabla f(\mathbf{x}_t), \mathbf{x} \rangle$
3: $\quad \mathbf{x}_{t+1} = \mathbf{x}_t + \gamma_t (\mathbf{v}_t - \mathbf{x}_t)$
4: **end for**

$$f(\mathbf{x})$$
$$\mathcal{X}$$
$$\mathbf{v}_t$$
$$\mathbf{x}_t$$
$$-\nabla f(\mathbf{x}_t)$$
$$f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle$$

Optimal **projection-based** methods for this class of functions achieve an ϵ-optimal in $T = O\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$ **first-order calls** [NY83; Nes83].

Q: Can CG-type methods have the same oracle complexity, both in terms of FOO to f and LMO to $\mathcal{X}$?

Optimal **projection-based** methods for this class of functions achieve an ϵ-optimal in $T = O\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$ **first-order calls** [NY83; Nes83].

Q: Can CG-type methods have the same oracle complexity, both in terms of FOO to f and LMO to $\mathcal{X}$?

Yes! Albeit locally [DCP20] if we know L and μ.

The contributions are as follows:

1. Parameter-free Locally-accelerated Conditional Gradient (PF-LaCG) algorithm.

2. Near-optimal and parameter-free accelerated algorithm (ACC) with inexact projections.

Main Ideas

The PF-LaCG algorithm [Car+21] couples the *Away-step Frank-Wolfe* (AFW) [GM86; LJ15] and ACC, a variant of the accelerated algorithm [CDO18], and periodically restarts when a measure of optimality is halved, and in this case **an upper bound on the primal gap**, denoted as:

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

where $\mathcal{S}$ is a proper support [1] of $\mathbf{x}$.

---

[1] A support $\mathcal{S}$ of $\mathbf{x}$ is a proper support of $\mathbf{x}$ if the weights associated with the convex decomposition are positive.

Main Idea

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

This allows us to:

- Maintain a computable global measure of optimality without knowing L and μ.
- Couple the AFW algorithm and the ACC algorithm while guaranteeing monotonic progress

Algorithm Sketch

At a high level:

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 optimal at an accelerated rate
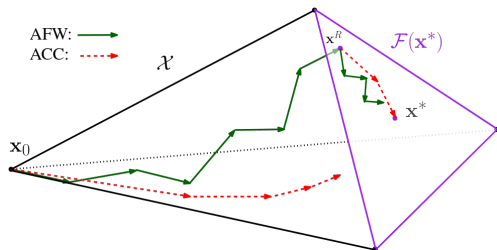
Convergence Rate of PF-LaCG

Theorem (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:

$$T = \min \left\{ \underbrace{O\left(\frac{LD^2}{\mu\delta^2} \log \frac{1}{\epsilon}\right)}_{AFW\ bound}, \right.$$

$$\left. \underbrace{K}_{Burn\text{-}in} + \underbrace{O\left(\sqrt{\frac{L}{\mu}} \log\left(\frac{L}{\mu}\right) \log\left(\frac{LD}{\mu\delta}\right) \log\frac{1}{\epsilon}\right)}_{Locally\text{-}accelerated\ convergence} \right\},$$

where K is a constant that is independent of ϵ.

Computational Results
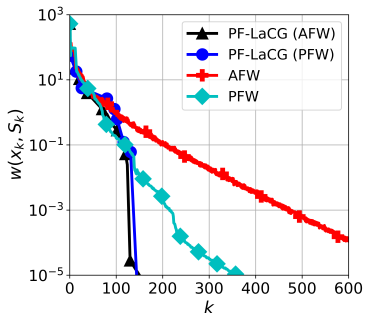
Structured LASSO Regression[2]

Figure: $w(\mathbf{x}_t, \mathcal{S}_t)$ vs. iteration
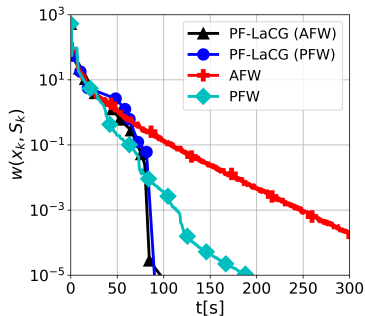
Figure: $w(\mathbf{x}_t, \mathcal{S}_t)$ vs. time

---

[2]Note: Other measures of optimality can be bounded above by $w(\mathbf{x}, \mathcal{S})$.

Summary

- We introduce PF-LaCG, a novel parameter-free projection-free algorithm for minimizing smooth and strongly convex functions over convex polytopes.
- After a finite burn-in phase independent of ϵ, PF-LaCG achieves a near-optimal accelerated convergence rate without knowledge of any problem parameters.
- We demonstrate PF-LaCG's practical improvements over non-accelerated algorithms, both in iteration count and wall-clock time.

References I

[FW56]   Marguerite Frank and Philip Wolfe. "An algorithm for quadratic programming". In: *Naval research logistics quarterly* 3.1-2 (1956), pp. 95–110.

[Pol74]   Boris Teodorovich Polyak. "Minimization methods in the presence of constraints". In: *Itogi Nauki i Tekhniki. Seriya" Matematicheskii Analiz"* 12 (1974), pp. 147–197.

[NY83]   Arkadii Semenovich Nemirovsky and David Borisovich Yudin. "Problem complexity and method efficiency in optimization". In: *Wiley-Interscience Series in Discrete Mathematics* 15 (1983).

[Nes83]   Y Nesterov. "A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$". In: *Soviet Math. Dokl.* Vol. 27. 1983.

References II

[DCP20]   Jelena Diakonikolas, Alejandro Carderera, and
          Sebastian Pokutta. "Locally accelerated conditional
          gradients". In: *International Conference on Artificial
          Intelligence and Statistics*. PMLR. 2020, pp. 1737–1747.

[Car+21] Alejandro Carderera, Jelena Diakonikolas, Cheuk Yin Lin,
          and Sebastian Pokutta. "Parameter-free Locally Accelerated
          Conditional Gradients". In: *To appear in International
          Conference on Machine Learning*. 2021.

[GM86]   Jacques Guélat and Patrice Marcotte. "Some comments on
          Wolfe's 'away step'". In: *Mathematical Programming* 35.1
          (1986), pp. 110–119.

[LJ15]   Simon Lacoste-Julien and Martin Jaggi. "On the Global
          Linear Convergence of Frank-Wolfe Optimization Variants".
          In: *Advances in Neural Information Processing Systems*.
          2015, pp. 496–504.

[CDO18]   Michael B. Cohen, Jelena Diakonikolas, and
           Lorenzo Orecchia. "On Acceleration with Noise-Corrupted
           Gradients". In: *Proc. ICML'18.* 2018.