Locally Accelerated Conditional Gradients

Alejandro Carderera

Joint work with J. Diakonikolas and S. Pokutta
Georgia Institute of Technology

alejandro.carderera@gatech.edu

November 20th, 2019

Georgia Tech H. Milton Stewart School of Industrial and Engineering Systems

Goal is smooth strongly-convex optimization.

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

Goal is smooth strongly-convex optimization.

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

Main ingredients:

First-order (FO) oracle. Given $x \in \mathcal{X}$ return:

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

Linear optimization (LO) oracle. Given $v \in \mathbb{R}^n$, return:

$$\operatorname*{argmin}_{x \in \mathcal{X}} \langle v, x \rangle$$

Focus of our work is on the *Conditional Gradients* algorithm (CG) [1], also known as the *Frank-Wolfe* algorithm (FW) [2] and its variants.

Focus of our work is on the *Conditional Gradients* algorithm (CG) [1], also known as the *Frank-Wolfe* algorithm (FW) [2] and its variants.

Theorem (Convergence rate of CG variants.)

[3] For the problem at hand the number of steps T required to reach an ϵ-optimal solution to the minimization problem verifies,

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

where D and δ are the diameter and pyramidal width of polytope $\mathcal{X}$.

CG Global Acceleration.

However, we know that optimal projected methods for this class of functions achieve an ϵ solution in $T = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$ first-order calls [4, 5].

Can CG achieve these convergence rates **globally**?

CG Global Acceleration.

However, we know that optimal projected methods for this class of functions achieve an ϵ solution in $T = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$ first-order calls [4, 5].

Can CG achieve these convergence rates **globally**?

No: global acceleration in Nesterov's sense is not possible.

Objectives:

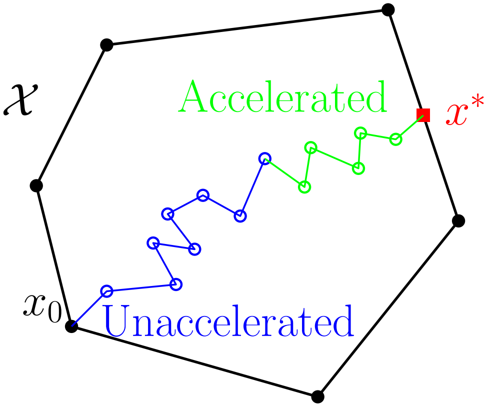
- Dimension independent global acceleration.

Objectives:

- ~~Dimension independent global acceleration.~~
- Dimension independent local acceleration.

Locally Accelerated Conditional Gradients (LaCG).

What do we mean by **local acceleration**?



After a constant number of iterations, accelerate the convergence.

Locally Accelerated Conditional Gradients (LaCG).

The key ingredients is a *Modified μAGD* algorithm [6].

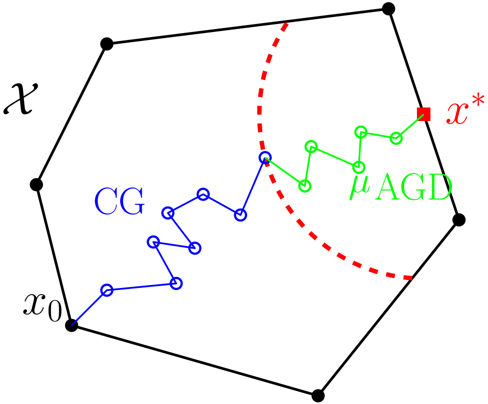
Theorem (Convergence rate of μAGD.)

Let $\{\mathcal{C}_i\}_{i=0}^t$ be a sequence of convex subsets of $\mathcal{X}$ such that $\mathcal{C}_i \subseteq \mathcal{C}_{i-1}$ for all i and $x^* \in \cap_{i=0}^t \mathcal{C}_i$, then the μAGD achieves an ϵ-optimal solution in:

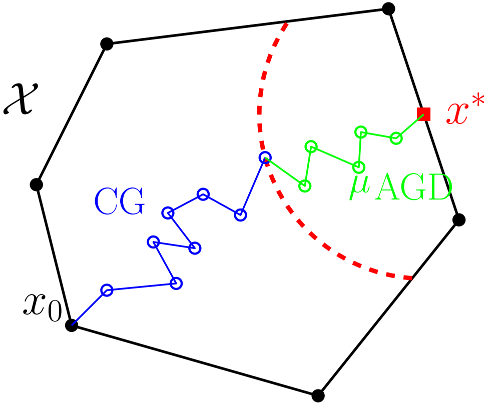
$$T = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$$

How do we build $\{\mathcal{C}_i\}_{i=0}^t$ in an efficient way?

Naively, what we would like:



Naively, what we would like:



But since the value of r is not known, we don't know when to switch from CG to μAGD.

Main ideas of LaCG:

Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

Main ideas of LaCG:

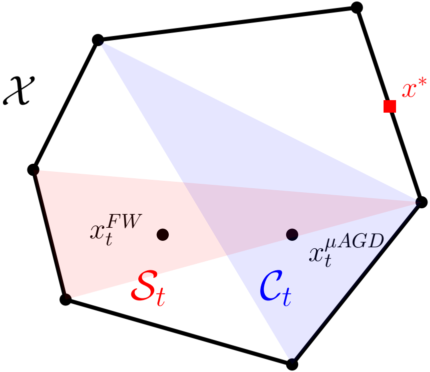
- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

CG Step

Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

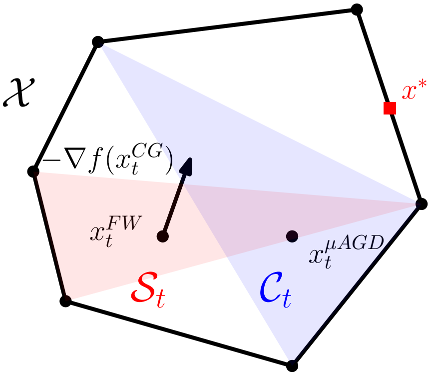
CG Step



Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

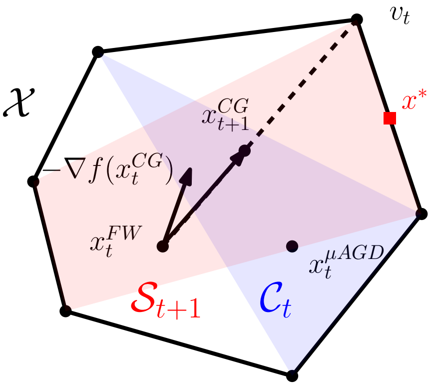
CG Step



Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

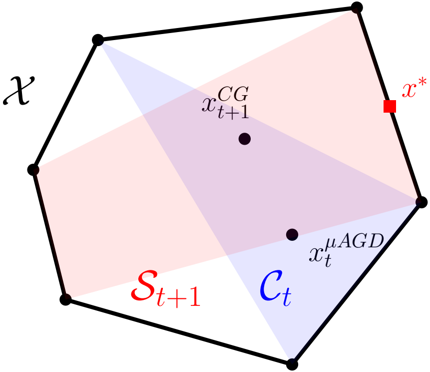
CG Step



Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

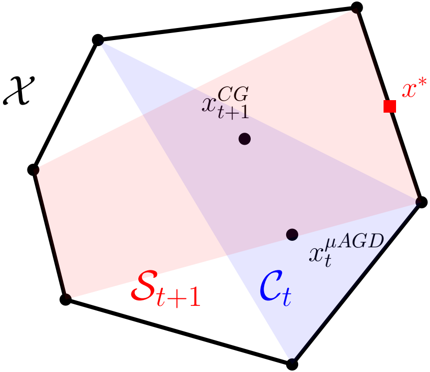
CG Step



Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

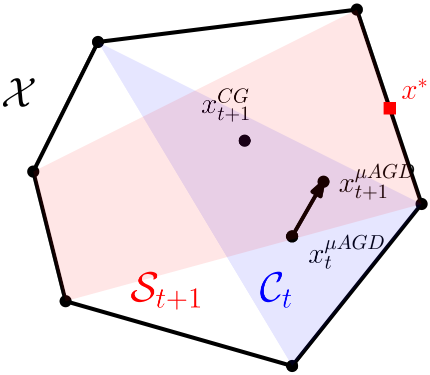
μAGD Step



Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

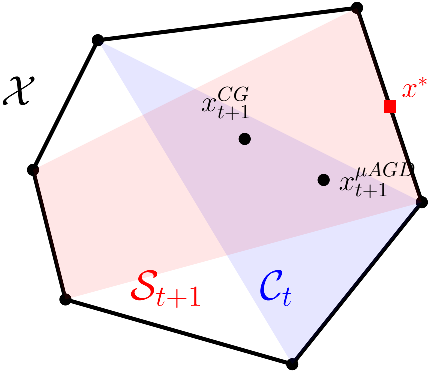
μAGD Step



Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

μAGD Step



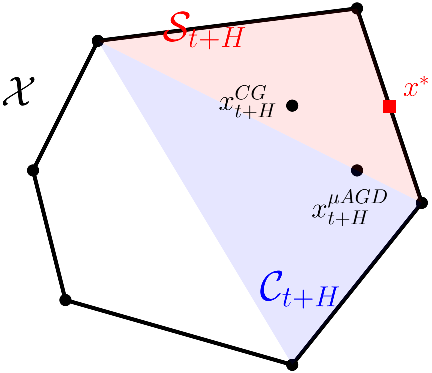
Main ideas of LaCG:

- Every H iterations restart: use $\mathcal{S}_t$ to update C_t if a vertex was added to $\mathcal{S}_t$ since the last update.

Main ideas of LaCG:

- Every H iterations restart: use $\mathcal{S}_t$ to update C_t if a vertex was added to $\mathcal{S}_t$ since the last update.

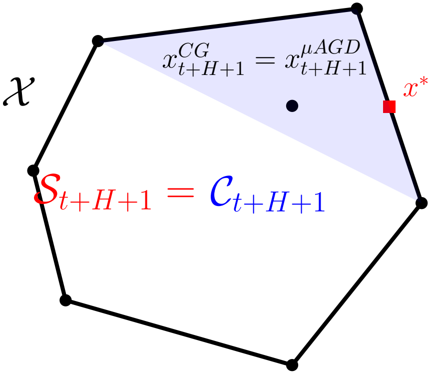
Restart



Main ideas of LaCG:

- Every H iterations restart: use $\mathcal{S}_t$ to update C_t if a vertex was added to $\mathcal{S}_t$ since the last update.

Restart



Convergence rate of LaCG.

Theorem (Convergence rate of LaCG.)

Let f be L-smooth and μ-strongly convex and let r be the critical radius, for:

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

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

Computational Results.

Despite the faster convergence rate after the burn-in phase, how does LaCG perform with respect to other projection-free algorithms?

Simplex in $\mathbb{R}^{1500}$ with $L/\mu = 1000$.

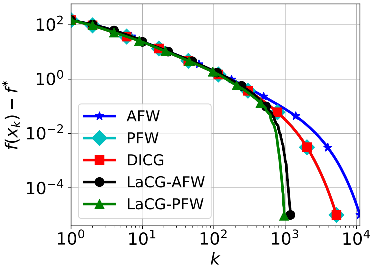


Figure: Primal gap vs. iteration

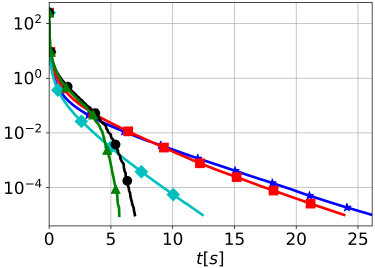


Figure: Primal gap vs. time

When close enough to $x*$ (after burn-in phase), there is a significant speedup in the convergence rate.

Birkhoff polytope in $\mathbb{R}^{400x400}$ with $L/\mu = 100$.

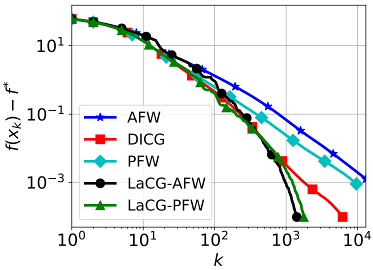


Figure: Primal gap vs. iteration

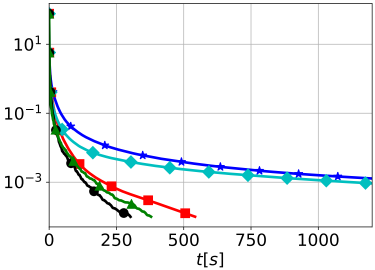


Figure: Primal gap vs. time

Structured Regression over MIPLIB Polytope (`ran14x18-disj-8`).

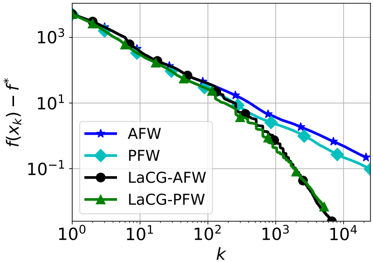


Figure: Primal gap vs. iteration

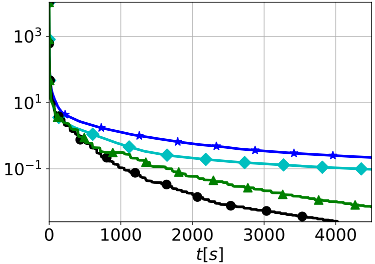


Figure: Primal gap vs. time

Thank you
for your attention.

References I

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

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

[3] Simon Lacoste-Julien and Martin Jaggi. "On the Global Linear Convergence of Frank-Wolfe Optimization Variants". In: *Advances in Neural Information Processing Systems 28*. 2015, pp. 496–504.

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

[5] 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

[6] Jelena Diakonikolas, Alejandro Carderera, and Sebastian Pokutta. "Locally Accelerated Conditional Gradients". In: *arXiv preprint arXiv:1906.07867* (2019).