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
Goal is smooth strongly-convex optimization.

\[
\min_{x \in X} f(x)
\]
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:

\[
\arg\min_{x \in \mathcal{X}} \langle v, x \rangle
\]
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 $\epsilon$-optimal solution to the minimization problem verifies,

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

where $D$ and $\delta$ are the diameter and pyramidal width of polytope $\mathcal{X}$. 
However, we know that optimal projected methods for this class of functions achieve an $\epsilon$ solution in $T = O \left( \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon} \right)$ 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 $\epsilon$ 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.
The key ingredients is a Modified $\mu$AGD algorithm [6].

**Theorem (Convergence rate of $\mu$AGD.)**

Let $\{C_i\}_{i=0}^t$ be a sequence of convex subsets of $\mathcal{X}$ such that $C_i \subseteq C_{i-1}$ for all $i$ and $x^* \in \bigcap_{i=0}^t C_i$, then the $\mu$AGD achieves an $\epsilon$-optimal solution in:

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

How do we build $\{C_i\}_{i=0}^t$ 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 $\mu$AGD.
Main ideas of LaCG:

At each iteration perform a CG variant step and a $\mu$AGD step over $C_{t+1}$ and select $x_{t+1} = \arg\min\{x_{CG_{t+1}}, x_{\mu AGD_{t+1}}\}$. 

CG Step

$X_{FW_{t}}$ $X_{\mu AGD_{t}}$ $x^{*}$ $C_{t}$ $S_{t}$
Main ideas of LaCG:

- At each iteration perform a CG variant step and a $\mu$AGD step over $C_{t+1}$ and select $x_{t+1} = \arg\min\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$. 
Main ideas of LaCG:

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

CG Step
Main ideas of LaCG:

- At each iteration perform a CG variant step and a $\mu$AGD step over $C_{t+1}$ 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 $\mu$AGD step over $C_{t+1}$ and select $x_{t+1} = \arg\min\{x_{CGt+1}, x_{\mu AGD t+1}\}$.

**CG Step**

![Diagram showing the CG Step process]
Main ideas of LaCG:

- At each iteration perform a CG variant step and a $\mu$AGD step over $C_{t+1}$ 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 $\mu$AGD step over $C_{t+1}$ and select $x_{t+1} = \arg\min\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

CG Step
Main ideas of LaCG:

- At each iteration perform a CG variant step and a $\mu$AGD step over $C_{t+1}$ and select $x_{t+1} = \arg\min\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

$\mu$AGD Step
Main ideas of LaCG:

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

$\mu$AGD Step

![Diagram](image.png)
Main ideas of LaCG:

- At each iteration perform a CG variant step and a $\mu$AGD step over $C_{t+1}$ and select $x_{t+1} = \arg\min\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

$\mu$AGD Step
Main ideas of LaCG:

- Every $H$ iterations restart: use $S_t$ to update $C_t$ if a vertex was added to $S_t$ since the last update.
Main ideas of LaCG:

- Every $H$ iterations restart: use $S_t$ to update $C_t$ if a vertex was added to $S_t$ since the last update.

Restart
Main ideas of LaCG:

- Every $H$ iterations restart: use $S_t$ to update $C_t$ if a vertex was added to $S_t$ since the last update.
Convergence rate of LaCG.

**Theorem (Convergence rate of LaCG.)**

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

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

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

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

Figure: Primal gap vs. time
Birkhoff polytope in $\mathbb{R}^{400 \times 400}$ 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


