Locally Accelerated Conditional Gradients

Alejandro Carderera

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

alejandro.carderera@gatech.edu

July 29th, 2019
Goal is smooth convex optimization.

\[
\min_{x \in \mathcal{X}} f(x)
\]
Goal is smooth convex optimization.

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

Main ingredients:
**First-order (FO) oracle.** Given \( x \in \mathcal{X} \) and a differentiable convex function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), 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].

Algorithm 1 Conditional Gradients algorithm.

**Input:** $x_0 \in \mathcal{X}$, stepsizes $\gamma_1 \cdots \gamma_T \in [0, 1]$.  

1: **for** $t = 0$ to $T$ **do**  
2: $v_t = \arg\min_{x \in \mathcal{X}} \langle \nabla f(x_t), x \rangle$  
3: $x_{t+1} = x_t + \gamma_t(v_t - x_t)$  
4: **end for**
Advantages of CG.

First-order. Dimensionality of modern problems makes computing second-order information infeasible.

Projection-free. Projection into certain feasible regions is computationally expensive: Birkhoff polytope and flow polytope are a few examples.

Sparse solutions. Solution is a convex combination of (a typically sparse set of) extreme points.
Advantages of CG.

**First-order.** Dimensionality of modern problems makes computing second-order information infeasible.
Advantages of CG.

**First-order.** Dimensionality of modern problems makes computing second-order information infeasible.

**Projection-free.** Projection into certain feasible regions is computationally expensive: Birkhoff polytope and flow polytope are a few examples.
Advantages of CG.

**First-order.** Dimensionality of modern problems makes computing second-order information infeasible.

**Projection-free.** Projection into certain feasible regions is computationally expensive: Birkhoff polytope and flow polytope are a few examples.

**Sparse solutions.** Solution is a convex combination of (a typically sparse set of) extreme points.
Disadvantages of CG.
Disadvantages of CG.

**Sublinear convergence.** For $L$-smooth and $\mu$-strongly convex $f$ when $x^*$ is in a face of $\mathcal{X}$.
Disadvantages of CG.

**Sublinear convergence.** For \( L \)-smooth and \( \mu \)-strongly convex \( f \) when \( x^* \) is in a face of \( \mathcal{X} \).

**Example (CG Convergence.)**

\( L \)-smooth and \( \mu \)-strongly convex \( f \) with \( x \in \mathbb{R}^2 \), and \( x^* \) in boundary of \( \mathcal{X} \).
Linear convergence is achieved by allowing steps that decrease the weight of *bad* vertices [3]. This has led to various CG variants:
Linear convergence is achieved by allowing steps that decrease the weight of *bad* vertices [3]. This has led to various CG variants:

**Away-step Conditional Gradients (AFW)**

Allow steps in the direction of:

\[ x - \arg\max_{y \in S} \langle \nabla f(x), y \rangle, \quad y \in S \]

where \( S \) is the active set of \( x \).

Figure: Away-step CG (AFW)
Pairwise CG

Fully-Corrective CG

Figure: PFW

Figure: FCFW
Convergence rate for $L$-smooth $\mu$-strongly convex $f$.

**Theorem (Convergence rate of AFW, PFW and FCFW.)**

[4] Suppose that $f$ is $L$-smooth $\mu$-strongly convex over a polytope $\mathcal{X}$, 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 $\mathcal{X}$.
Example (CG Variant Convergence.)

$L$-smooth and $\mu$-strongly convex $f$ ($L/\mu \approx 10^8$) over the probability simplex in $\mathbb{R}^{100}$, and $x^*$ a convex combination of 13 vertices.
However, we know that optimal 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 [5, 6]. Can CG achieve these convergence rates **globally**?
However, we know that optimal 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 [5, 6].

Can CG achieve these convergence rates \textit{globally}?

Example ([7, 8] $f(x) = \|x\|^2$ over unit simplex in $\mathbb{R}^n$.)

We know the optimal solution is given by $x^* = 1/n$. CG can incorporate at most one vertex in each iteration, if we start from a vertex $x_0$, in iteration $t < n$ we have that:

$$f(x_t) - f(x^*) \geq \frac{1}{t} - \frac{1}{n}.$$
Considering iterations such that \( t \leq \lfloor n/2 \rfloor \) and rearranging into a linear convergence contraction we have:

\[
T = \Omega \left( \frac{1}{r} \log \frac{1}{\epsilon} \right),
\]

where \( r \leq 2 \frac{\log 2t}{2t} \).
Considering iterations such that $t \leq \lfloor n/2 \rfloor$ and rearranging into a linear convergence contraction we have:

$$T = \Omega \left( \frac{1}{r} \log \frac{1}{\epsilon} \right),$$

where $r \leq 2 \frac{\log 2t}{2t}$.

**Convergence rate of the CG variants for this problem instance:** $r = \frac{1}{4t}$.

At best a global logarithmic improvement in the convergence rate, therefore global acceleration in Nesterov’s sense is not possible.
**Idea:** Run Nesterov’s Accelerated Gradient Descent, use CG to solve the projection subproblems approximately [9].
Conditional Gradient Sliding

**Idea:** Run Nesterov’s Accelerated Gradient Descent, use CG to solve the projection subproblems approximately [9].

**Results:**
- Separate LO and FO oracle calls.
- Globally optimal \( O \left( \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon} \right) \) calls to FO and \( O \left( \frac{LD^2}{\epsilon} + \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon} \right) \) calls to LO oracles.
- Convergence rates independent of the dimension \( n \).
Catalyst Augmented AFW.

Idea: Run Accelerated Proximal Method and solve proximal problems with a linearly convergent CG [10].
Catalyst Augmented AFW.

**Idea:** Run Accelerated Proximal Method and solve proximal problems with a linearly convergent CG [10].

**Results:**

- $O \left( \sqrt{\frac{L-\mu}{\mu}} \left( \frac{D}{\delta} \right)^2 \log \frac{1}{\epsilon} \right)$ Calls to FO and LO oracles.
- Convergence rates dependent of the dimension $n$. 
Summary

Complexity for $L$-smooth $\mu$-strongly convex $f$.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LO Calls</th>
<th>FO Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG Variants</td>
<td>$\mathcal{O}\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
<td>$\mathcal{O}\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>CGS</td>
<td>$\mathcal{O}\left(\frac{LD^2}{\epsilon} + \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$</td>
<td>$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>Catalyst</td>
<td>$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
<td>$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
</tr>
</tbody>
</table>
**Summary**

**Complexity for** $L$-smooth $\mu$-strongly convex $f$.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LO Calls</th>
<th>FO Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG Variants</td>
<td>$O\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
<td>$O\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>CGS</td>
<td>$O\left(\frac{L D^2}{\epsilon} + \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$</td>
<td>$O\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>Catalyst</td>
<td>$O\left(\sqrt{\frac{L-\mu}{\mu}} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
<td>$O\left(\sqrt{\frac{L-\mu}{\mu}} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>What we want:</td>
<td>$O\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$</td>
<td>$O\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$</td>
</tr>
</tbody>
</table>
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 [11].

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

Let $f$ be $L$-smooth and $\mu$-strongly convex and 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?
[12] CG: \( \exists r > 0 \) (that depends only on \( f \) and \( \mathcal{X} \)) s.t. if \( \| x^* - x_K \| \leq r \Rightarrow x^* \in \text{conv} \left( S_t \right) \) for all \( t \geq K \), where \( S_t \) is the active set at iteration \( t \).

So when we are inside the red semicircle and we use \( C_t = S_t \), acceleration is possible.
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:
Main ideas of LaCG:

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

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

- 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:

- At each iteration perform a CG variant step and a $\mu$AGD step over $C_{t+1}$ and select $x_{t+1} = \arg\min\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.
- Every $H$ iterations restart: use $S_t$ to update $C_t$ if a vertex was added to $S_t$ since the last update.
- After a constant burn-in phase, acceleration will be achieved.
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 $f(x_t) - f(x^*) \leq \epsilon$
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:

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

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

In fact, we often observe faster convergence even for $\|x_t - x^*\| \geq r$
If $\|x_T - x^*\| \geq r$

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LO Calls</th>
<th>FO Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG Variants</td>
<td>$\mathcal{O}\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
<td>$\mathcal{O}\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>CGS</td>
<td>$\mathcal{O}\left(\frac{LD^2}{\epsilon} + \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$</td>
<td>$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>Catalyst</td>
<td>$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
<td>$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>LaCG</td>
<td>$\mathcal{O}\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
<td>$\mathcal{O}\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$</td>
</tr>
</tbody>
</table>

Table: Complexity for $L$-smooth $\mu$-strongly convex $f$. 

Recap
Recap

If \( \|x_T - x^*\| \leq r \)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LO Calls</th>
<th>FO Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG Variants</td>
<td>( \mathcal{O}\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right) )</td>
<td>( \mathcal{O}\left(\frac{L}{\mu} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right) )</td>
</tr>
<tr>
<td>CGS</td>
<td>( \mathcal{O}\left(\frac{LD^2}{\epsilon} + \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right) )</td>
<td>( \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right) )</td>
</tr>
<tr>
<td>Catalyst</td>
<td>( \mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right) )</td>
<td>( \mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}} \left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right) )</td>
</tr>
<tr>
<td>LaCG</td>
<td>( K + \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right) )</td>
<td>( K + \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right) )</td>
</tr>
</tbody>
</table>

Table: Complexity for \( L \)-smooth \( \mu \)-strongly convex \( f \).

\( K \) is independent of \( \epsilon \), so asymptotically optimal.
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 $\mathbb{R}^{2000}$ 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**
$\ell_1$ unit ball in $\mathbb{R}^{2000}$ with $L/\mu = 100$. 

**Figure**: Primal gap vs. iteration

**Figure**: Primal gap vs. time
Birkhoff polytope in $\mathbb{R}^{40 \times 40}$ with $L/\mu = 100$. 

**Figure:** Primal gap vs. iteration

**Figure:** Primal gap vs. time
Video co-localization problem over flow polytope [13].

Figure: Primal gap vs. iteration

Figure: Primal gap vs. time
Thank you for your attention.
References I


