On the affine invariance of the conditional gradient algorithm

Javier Peña, CMU

Penn State University, Halloween 2023

Conditional gradient algorithm

Conditional gradient algorithm

Consider the problem

$$\min_{x \in C} f(x).$$

Conditional gradient (CG) algorithm

- pick $x_0 \in C$
- for $k = 0, 1, \ldots$ pick $\theta_k \in [0, 1]$ and let

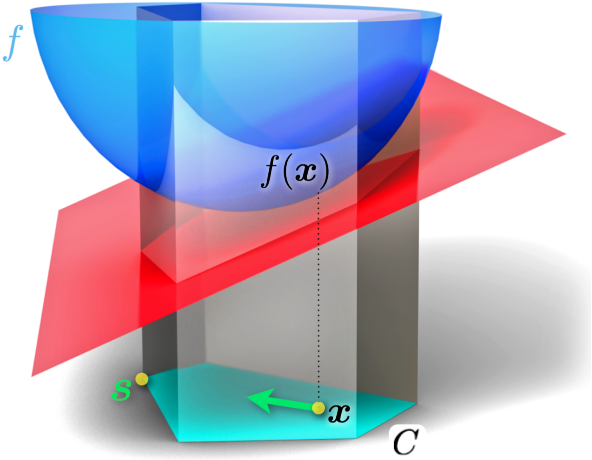
$$s_k := \operatorname*{argmin}_{y \in C} \langle \nabla f(x_k), y \rangle$$

$$x_{k+1} := x_k + \theta_k (s_k - x_k)$$

Introduced by Frank & Wolfe in 1956, and thus also known as the "Frank-Wolfe Algorithm." Very popular since around 2010.

Intuition for CG update

$x_+ = x + \theta(s - x)$ for $s = \text{argmin}_{y \in C} \langle \nabla f(x), y \rangle$ and $\theta \in [0, 1]$.



Picture from Jaggi's paper "Revisiting Frank-Wolfe" ICML 2013.

Conditional gradient algorithm

Technical assumptions

- $C \subseteq \mathbb{R}^n$ is compact and convex equipped with linear oracle:

$$g \mapsto \operatorname*{argmin}_{y \in C} \langle g, y \rangle.$$

- $f : C \to \mathbb{R}$ is convex and differentiable.

Main properties

- CG does not use projections. It uses a linear oracle instead.
- CG has nice sparsity-like properties for suitable domains and linear oracles.
- **CG is affine invariant.**

Affine invariance

Recall main problem

$$\min_{x \in C} f(x). \tag{1}$$

Consider the problem obtained after an affine change of variables:

$$\min_{\tilde{x} \in \tilde{C}} \tilde{f}(\tilde{x}) \tag{2}$$

where

$$\tilde{f} = f \circ A \text{ and } \tilde{C} = A^{-1}(C) \Leftrightarrow C = A(\tilde{C})$$

for some affine bijection $A : \mathbb{R}^n \to \mathbb{R}^n$.

Suppose CG is applied to (1) and (2) starting from $x \in C$ and $\tilde{x} \in \tilde{C}$ respectively.

Affine invariance

If $x = A\tilde{x}$ then the next iterates satisfy $x_+ = A\tilde{x}_+$ as well.
An affine change of variables does not change the algorithm.

Affine invariance details (when A is linear)

The iterates x_+ and $\tilde{x}_+$ are respectively

$$x_+ = x + \theta(s - x),\ \tilde{x}_+ = \tilde{x} + \theta(\tilde{s} - \tilde{x})$$

for

$$s = \underset{y \in C}{\operatorname{argmin}} \langle \nabla f(x), y \rangle,\ \tilde{s} = \underset{\tilde{y} \in \tilde{C}}{\operatorname{argmin}} \langle \nabla \tilde{f}(\tilde{x}), \tilde{y} \rangle.$$

Since $\tilde{f} = f \circ A$ we have $\nabla \tilde{f}(\tilde{x}) = A^* \nabla f(A\tilde{x})$ and so

$$\langle \nabla \tilde{f}(\tilde{x}), \tilde{y} \rangle = \langle A^* \nabla f(A\tilde{x}), \tilde{y} \rangle = \langle \nabla f(A\tilde{x}), A\tilde{y} \rangle.$$

Furthermore, $C = A(\tilde{C})$ and $x = A\tilde{x}$ imply that

$$A\tilde{s} = \underset{A\tilde{y} \in A(\tilde{C})}{\operatorname{argmin}} \langle \nabla f(A\tilde{x}), A\tilde{y} \rangle = \underset{y \in C}{\operatorname{argmin}} \langle \nabla f(x), y \rangle = s.$$

Therefore

$$A\tilde{x}_+ = A(\tilde{x} + \theta(\tilde{s} - \tilde{x})) = x + \theta(s - x) = x_+.$$

Theme of this talk

Recall main problem

$$f^\star := \min_{x \in C} f(x)$$

and conditional gradient algorithm

$$s_k := \underset{y \in C}{\mathsf{argmin}} \langle \nabla f(x_k), y \rangle$$

$$x_{k+1} := x_k + \theta_k (s_k - x_k) \text{ for } \theta_k \in [0,1]$$

Theme of this talk

Affine invariant convergence rates for $f(x_k) \to f^\star$ via a *growth property* of the pair (f, C).

Convergence rates range from sublinear to linear depending on the *degree* of the growth property.

Starting point

Key property

Affine-invariant finite curvature property (to be defined soon)

Theorem (Jaggi 2013)

If f has finite curvature on C then

$$f(x_k) - f^\star = \mathcal{O}\left(\frac{1}{k}\right).$$

Main development

Generalize finite curvature to an affine-invariant r-growth property for $r \in [0, 1]$. Then show that

$$f(x_k) - f^\star = \mathcal{O}\left(\frac{1}{k^{\frac{1}{1-r}}}\right).$$

Growth property and affine invariant convergence

Bregman distance, curvature

Recall main problem

$$\min_{x \in C} f(x).$$

Bregman distance D_f

$$D_f(y, x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

Finite curvature (adapted from Clarkson 2010 and Jaggi 2013)

Say that f has finite curvature on C if there exists $M < \infty$ such that for $x \in C, s = \mathrm{argmin}_{y \in C} \langle \nabla f(x), y \rangle$, and $\theta \in [0, 1]$

$$D_f(x + \theta(s - x), x) \leq \frac{M\theta^2}{2}.$$

This property holds when ∇f is Lipschitz continuous on C.

Suboptimality gap and Wolfe gap

Recall main problem

$$f^\star := \min_{x \in C} f(x).$$

Define $\texttt{subopt} : C \to \mathbb{R}$ and $\texttt{gap} : C \to \mathbb{R}$ as follows

$$\begin{aligned} \texttt{subopt}(x) &:= f(x) - f^\star \\ \texttt{gap}(x) &:= \langle \nabla f(x), x - s \rangle \end{aligned}$$

for $s = \text{argmin}_{y \in C} \langle \nabla f(x), y \rangle$.

Key facts

For all $x \in C$ we have $\texttt{gap}(x) \geq \texttt{subopt}(x)$.

For $x \in C,\ s = \text{argmin}_{y \in C} \langle \nabla f(x), y \rangle$, and $\theta \in [0, 1]$ we have

$$\texttt{subopt}(x + \theta(s - x)) = \texttt{subopt}(x) - \theta \cdot \texttt{gap}(x) + D_f(x + \theta(s - x), x).$$

Growth property (simplified version)

Suppose $r \in [0, 1]$.

Say that (f, C) satisfies the r-*growth property* if there is $M < \infty$ such that for $x \in C, s = \text{argmin}_{y \in C} \langle \nabla f(x), y \rangle$, and $\theta \in [0, 1]$

$$D_f(x + \theta(s - x), x) \cdot \texttt{subopt}(x)^{2-r} \leq \frac{M\theta^2}{2} \cdot \text{gap}(x)^2.$$

Observe

- Growth property is affine invariant.
- Finite *curvature* (Clarkson 2010, Jaggi 2013) $\Rightarrow$ 0-growth. Indeed, 0-growth follows from $\texttt{subopt}(x)^2 \leq \text{gap}(x)^2$ and

$$D_f(x + \theta(s - x), x) \leq \frac{M\theta^2}{2}.$$

Main theorem: affine invariant convergence

Consider the iterates $x_k,\ k = 0, 1, \ldots$ generated by CG. Let

$$\texttt{subopt}_k := \texttt{subopt}(x_k) = f(x_k) - f^\star.$$

To ease notation, suppose CG chooses θ_k via

$$\theta_k := \operatorname*{argmin}_{\theta\in[0,1]} f(x_k + \theta(s_k - x_k)).$$

(This assumption can be relaxed. More on this matter later.)

Main theorem: affine invariant convergence

Theorem (P. 2022)

Suppose (f, C) satisfy the r-growth property. Then
For $r = 1$ we get linear convergence

$$\texttt{subopt}_k \leq \texttt{subopt}_0 \left(1 - \frac{1}{2} \cdot \min\left\{1, \frac{1}{M}\right\}\right)^k.$$

For $r = 0$ we get sublinear convergence (as in Jaggi 2013)

$$\texttt{subopt}_k \leq \frac{2M}{k+3}.$$

For $r \in [0, 1)$ we get

$$\texttt{subopt}_k = \mathcal{O}\left(\frac{1}{k^{\frac{1}{1-r}}}\right).$$

Sufficient conditions for r-growth

Lipschitz continuity

Recall our main problem and introduce some notation. Let

$$f^\star := \min_{x \in C} f(x) \text{ and } X^\star := \{x \in C : f(x) = f^\star\}.$$

Suppose $\mathbb{R}^n$ is endowed with a norm $\|\cdot\|$.

Say that ∇f is Lipschitz continuous on C if there exists $L < \infty$ such that for all $x, y \in C$

$$\|\nabla f(y) - \nabla f(x)\|^* \leq L\|y - x\|.$$

In this case, it readily follows that for all $x, s \in C$ and $\theta \in [0, 1]$

$$D_f(x + \theta(s - x), x) \leq \frac{L \cdot \mathrm{diam}(C)^2}{2} \cdot \theta^2$$

and so (f, C) satisfies the 0-growth property.

Error bound and uniform convexity

Suppose $\gamma \in [0, 1/2]$. Say that (f, C) satisfies the γ-error bound condition if there exists $K < \infty$ such that for all $x \in C$

$$\|x - X^\star\| \leq K \cdot (f(x) - f^\star)^\gamma.$$

Suppose $p \geq 2$. Say that $C \subseteq \mathbb{R}^n$ is p-uniformly convex if there exists $\mu > 0$ such that for all $x, y \in C,\ \theta \in [0, 1]$, and $\|z\| \leq 1$

$$x + \theta(y - x) + \frac{\mu}{p}\theta(1 - \theta)\|y - x\|^p z \in C.$$

Remark

Lipschitz continuity, γ-error bound, and p-uniform convexity properties are affine invariant.

The corresponding constants L, K, μ are not.

Canonical examples

- Suppose $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ and

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2.$$

  Then ∇f is Lipschitz continuous and (f, C) satisfies the $1/2$-error bound condition for any closed convex $C \subseteq \mathbb{R}^n$.

- Suppose $p > 1$ and $C = \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$.
  - If $1 < p \leq 2$ then C is 2-uniformly convex.
  - If $p > 2$ then C is p-uniformly convex.

Sufficient conditions for r-growth

Proposition (P. 2022)

Suppose ∇f is Lipschitz continuous on C. Then

(a) *(f, C) satisfies the r-growth property for $r = 2/p$ if C is p-uniformly convex and ∇f is bounded away from zero in C.*

(b) *(f, C) satisfies the r-growth property for $r = 2\gamma/p$ if C is p-uniformly convex and the γ-error bound holds.*

(c) *(f, C) satisfies the r-growth property for $r = 2\gamma$ if $X^\star \subseteq$ ri(C) and the γ-error bound holds.*

Recall main theorem: if r-growth holds then CG iterates satisfy

$$\texttt{subopt}_k = \mathcal{O}\left(\frac{1}{k^{\frac{1}{1-r}}}\right).$$

Consequences of main theorem and proposition

Corollary (Kerdreux et al. 2021)

Suppose C is p-uniformly convex, ∇f is Lipschitz continuous on C, and ∇f is bounded away from zero in C. Then the CG iterates satisfy

(a) $\texttt{subopt}_k = \mathcal{O}\left(\frac{1}{k^{\frac{p}{p-2}}}\right)$ *if* $p > 2$.

(b) $\texttt{subopt}_k \to 0$ *linearly if* $p = 2$.

Corollary (Garber-Hazan 2015, Xu-Yang 2018, Kerdreux et al. 2021)

Suppose C is p-uniformly convex, ∇f is Lipschitz continuous on C, and the γ-error bound holds. Then the CG iterates satisfy

$$\texttt{subopt}_k = \mathcal{O}\left(\frac{1}{k^{\frac{p}{p-2\gamma}}}\right).$$

Consequences of main theorem and proposition

Corollary (Guélat-Marcotte 1986 extended)

Suppose ∇f is Lipschitz continuous on C, the γ-error bound holds, and $X^\star \subseteq ri(C)$. Then the CG iterates satisfy

(a) $\texttt{subopt}_k = \mathcal{O}\left(\frac{1}{k^{\frac{1}{1-2\gamma}}}\right)$ *if* $\gamma \in [0, 1/2)$.

(b) $\texttt{subopt}_k \to 0$ *linearly if* $\gamma = 1/2$.

Remark

In all cases the constants in the $\mathcal{O}(\cdot)$ bounds are at least as sharp as previous ones.

Canonical examples again

Consider the problem

$$\min_{x \in C} f(x)$$

where $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ for some $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $C = \{x \in \mathbb{R}^n : \|x\|_p \le 1\}$ for some $p > 1$.

Then the CG iterates satisfy

- $\mathtt{subopt}_k \to 0$ linearly when $\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \notin C$ and $1 < p \le 2$
- $\mathtt{subopt}_k = \mathcal{O}(1/k^{\frac{p}{p-2}})$ when $\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \notin C$ and $p > 2$
- $\mathtt{subopt}_k = \mathcal{O}(1/k^{\frac{p}{p-1}})$ when $\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \in \mathrm{rbd}(C)$
- $\mathtt{subopt}_k \to 0$ linearly when $\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) \in \mathrm{ri}(C)$.

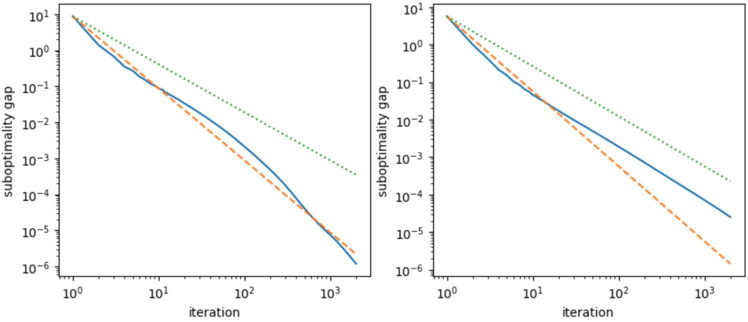
A simple numerical experiment

Consider the problem

$$\min_{x \in C} f(x)$$

where $f(x) = \frac{1}{2}\|x - b\|_2^2$ and $C = \{x \in \mathbb{R}^n : \|x\|_4 \leq 1\}$.

Typical convergence rate of $\mathtt{subopt}_k \to 0$



Dotted line: $\mathtt{subopt}_0 / k^{\frac{p}{p-1}}$, dashed line: $\mathtt{subopt}_0 / k^{\frac{p}{p-2}}$

Left plot: $b \notin C$, right plot: $b \in \mathrm{rbd}(C)$.

Conditional gradient with other stepsizes

(joint work with Wirth and Pokutta, ZIB)

Recall main problem

$$\min_{x \in C} f(x)$$

and conditional gradient algorithm

$$\begin{aligned} s_k &:= \operatorname*{argmin}_{y \in C} \langle \nabla f(x_k), y \rangle \\ x_{k+1} &:= x_k + \theta_k (s_k - x_k) \text{ for } \theta_k \in [0,1] \end{aligned}$$

Stepsize via line-search

Exact line-search

Main theorem holds provided the stepsize θ_k is chosen via

$$\begin{aligned}\theta_k &:= \operatorname*{argmin}_{\theta\in[0,1]} f(x_k + \theta(s_k - x_k)) \\ &= \operatorname*{argmin}_{\theta\in[0,1]}\{(1-\theta)\mathtt{gap}(x_k) + D_f(x_k + \theta(s_k - x_k), x_k)\}.\end{aligned}$$

Approximate line-search (Armijo-like)

Main theorem also holds (with larger constants) if θ_k is chosen so that $\rho \cdot \hat{\theta} \le \theta_k \le \hat{\theta}$ where $\hat{\theta}$ is the largest $\theta \in [0, 1]$ such that

$$(1-\theta)\mathtt{gap}(x_k) + D_f(x_k + \theta(s_k - x_k), x_k) \le (1 - c\cdot\theta)\mathtt{gap}(x_k)$$

for $c, \rho \in (0, 1)$ with $c + \rho > 1$.

For $c = 1/2, \rho = 1$ get the main theorem.

Open-loop stepsizes

Wirth-Kerdreux-Pokutta 2022:

As an alternative to line-search, use pre-determined stepsizes, like $\theta_k = \frac{2}{k+2}$ or more generally $\theta_k = \frac{\ell}{k+\ell}$ for $\ell \in \mathbb{N}$.

Theorem (Wirth, Pokutta, P. 2023)

Suppose (f, C) *satisfy the* **strong** r*-growth property and* $\theta_k = \frac{\ell}{k+\ell}$ *for* $\ell \in \mathbb{N}$. *Then for all* $\epsilon \in (0,1)$ *the CG iterates satisfy*

$$\mathtt{subopt}_k = \mathcal{O}\left(\frac{1}{k^{\frac{1-\epsilon}{1-r}}} + \frac{1}{k^\ell}\right).$$

Remark

Stepsize $\theta_k = \frac{2}{k+2}$ yields $\mathcal{O}\left(\frac{1}{k^2}\right)$ convergence if $r \in (1/2, 1]$.

Strong growth property

Suppose $r \in [0, 1]$.

Recall r-growth property

There is $M < \infty$ such that for $x \in C, s = \text{argmin}_{y \in C} \langle \nabla f(x), y \rangle$, and $\theta \in [0, 1]$

$$D_f(x + \theta(s - x)) \cdot \texttt{subopt}(x)^{2-r} \leq \frac{M\theta^2}{2} \cdot \texttt{gap}(x)^2.$$

Strong r-growth property

There is $M < \infty$ such that for $x \in C, s = \text{argmin}_{y \in C} \langle \nabla f(x), y \rangle$, and $\theta \in [0, 1]$

$$D_f(x + \theta(s - x)) \leq \frac{M\theta^2}{2} \cdot \texttt{gap}(x)^r.$$

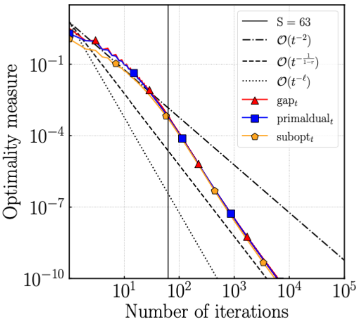
Numerical experiments with $\theta_k = 4/(k+4)$

Consider the problem

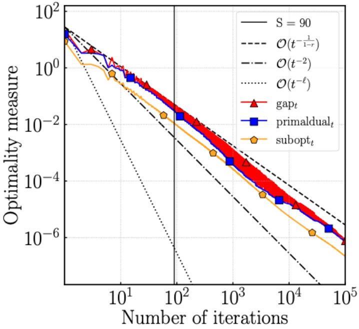
$$\min_{x \in C} f(x)$$

where $f(x) = \frac{1}{2}\|x - b\|_2^2$ and $C = \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$.

Convergence rate when $b \notin C$



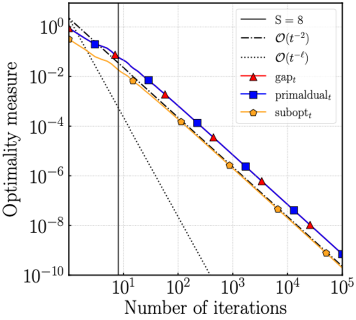
(b) ℓ_3-ball, $r = 2/3$.



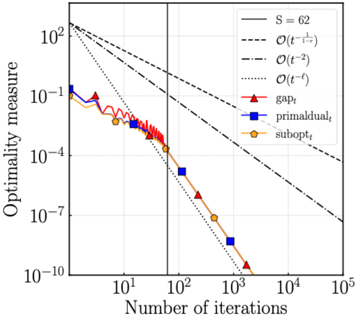
(c) ℓ_7-ball, $r = 2/7$.

Numerical experiments with $\theta_k = 4/(k+4)$

Convergence rate when $b \in \mathrm{rbd}(C)$



(b) ℓ_2-ball, $r = 1/2$.

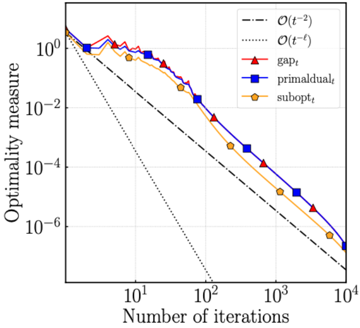


(c) ℓ_7-ball, $r = 1/7$.

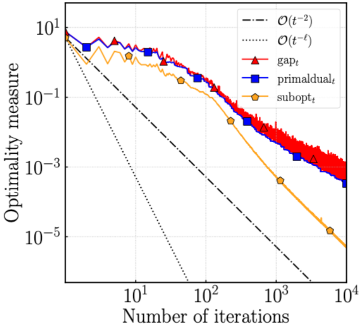
A more interesting experiment with $\theta_k = 4/(k+4)$

Collaborative filtering (Mehta et al. 2007)

$$\min_{X \in \mathbb{R}^{m \times n},\ \|X\|_{\mathsf{nuc}} \le \beta} \sum_{(i,j) \in \mathcal{I}} H_\rho(A_{ij} - X_{ij})$$



(b) $\beta = 2000$.



(c) $\beta = 3000$.

Conclusions

Conclusions

Conditional gradient method for $\min_{x \in C} f(x)$.

- Affine invariant convergence rates via a *growth property.*
- Sublinear to linear range of convergence rates depending on the degree of the growth property.
- Similar results for open-loop step-sizes $\theta_k = \ell/(k+\ell)$.
- Similar developments for conditional gradient variants, e.g., away steps, blended pairwise steps, in-face steps, etc.

Main references

- P. *"Affine invariant convergence rates of the conditional gradient method,"* `https://arxiv.org/abs/2112.06727`
- Wirth, P., Pokutta *"Accelerated Affine-Invariant Convergence Rates of the Frank-Wolfe Algorithm with Open-Loop Step-Sizes,"* `https://arxiv.org/abs/2310.04096`