Large-Scale Optimization for Data Science

Gradient methods for constrained problems

Yuxin Chen

Wharton Statistics & Data Science, Fall 2023

Outline

- Frank-Wolfe algorithm
- Projected gradient methods

Constrained convex problems

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{C} \end{array}$$

- $f(\cdot)$: convex function
- $\mathcal{C} \subseteq \mathbb{R}^n$: closed convex set

Feasible direction methods

Generate a feasible sequence $\{\boldsymbol{x}^t\} \subseteq \mathcal{C}$ with iterations

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \eta_t \boldsymbol{d}^t$$

where $\boldsymbol{d}^t$ is a feasible direction (s.t. $\boldsymbol{x}^t + \eta_t \boldsymbol{d}^t \in \mathcal{C}$)

- **Question:** can we guarantee feasibility while enforcing cost improvement?

Frank-Wolfe algorithm

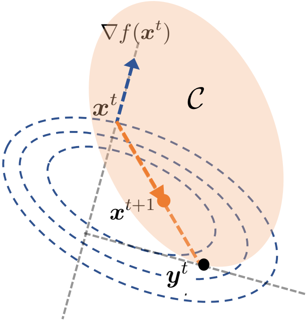
developed by Philip Wolfe and Marguerite Frank

Frank-Wolfe / conditional gradient algorithm

Algorithm 3.1 Frank-wolfe (a.k.a. conditional gradient) algorithm

```
1: for $t = 0, 1, \cdots$ do
2:     $\boldsymbol{y}^t :=$ arg $\min_{\boldsymbol{x} \in \mathcal{C}} \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} \rangle$          (direction finding)
3:     $\boldsymbol{x}^{t+1} = (1 - \eta_t)\boldsymbol{x}^t + \eta_t \boldsymbol{y}^t$          (line search and update)
```

$$\boldsymbol{y}^t = \text{arg} \min_{\boldsymbol{x} \in \mathcal{C}} \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \rangle$$



Frank-Wolfe / conditional gradient algorithm

Algorithm 3.2 Frank-wolfe (a.k.a. conditional gradient) algorithm

```
1: for $t = 0, 1, \cdots$ do
2:     $\boldsymbol{y}^t := \arg\min_{\boldsymbol{x} \in \mathcal{C}} \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} \rangle$          (direction finding)
3:     $\boldsymbol{x}^{t+1} = (1 - \eta_t)\boldsymbol{x}^t + \eta_t \boldsymbol{y}^t$          (line search and update)
```

- main step: linearization of the objective function (equivalent to $f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \rangle$)

  $\Longrightarrow$ linear optimization over a convex set

- appealing when linear optimization is cheap

- stepsize η_t determined by line search, or $\underbrace{\eta_t = \frac{2}{t+2}}_{\text{bias towards } \boldsymbol{x}^t \text{ for large } t}$

Frank-Wolfe can also be applied to nonconvex problems

Example (Luss & Teboulle '13)

$$\text{minimize}_{\boldsymbol{x}} \quad -\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} \qquad \text{subject to} \;\; \|\boldsymbol{x}\|_2 \leq 1 \tag{3.1}$$

for some $\boldsymbol{Q} \succ \boldsymbol{0}$

Frank-Wolfe can also be applied to nonconvex problems

We now apply Frank-Wolfe to solve (3.1). Clearly,

$$\boldsymbol{y}^t = \text{arg} \min_{\boldsymbol{x}:\|\boldsymbol{x}\|_2 \leq 1} \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} \rangle = -\frac{\nabla f(\boldsymbol{x}^t)}{\|\nabla f(\boldsymbol{x}^t)\|_2} = \frac{\boldsymbol{Q}\boldsymbol{x}^t}{\|\boldsymbol{Q}\boldsymbol{x}^t\|_2}$$

$$\Longrightarrow \quad \boldsymbol{x}^{t+1} = (1-\eta_t)\boldsymbol{x}^t + \eta_t \boldsymbol{Q}\boldsymbol{x}^t / \|\boldsymbol{Q}\boldsymbol{x}^t\|_2$$

Set $\eta_t = \arg\min_{0 \leq \eta \leq 1} f\Big((1-\eta)\boldsymbol{x}^t + \eta \frac{\boldsymbol{Q}\boldsymbol{x}^t}{\|\boldsymbol{Q}\boldsymbol{x}^t\|_2}\Big) = 1$ (check). This gives

$$\boldsymbol{x}^{t+1} = \boldsymbol{Q}\boldsymbol{x}^t / \|\boldsymbol{Q}\boldsymbol{x}^t\|_2$$

which is essentially power method for finding leading eigenvector of $\boldsymbol{Q}$

Convergence for convex and smooth problems

Theorem 3.1 (Frank-Wolfe for convex and smooth problems, Jaggi '13)

Let f be convex and L-smooth. With $\eta_t = \frac{2}{t+2}$, one has

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \frac{2Ld_{\mathcal{C}}^2}{t+2}$$

where $d_{\mathcal{C}} = \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{C}} \|\boldsymbol{x} - \boldsymbol{y}\|_2$

- for compact constraint sets, Frank-Wolfe attains ε-accuracy within $O(\frac{1}{\varepsilon})$ iterations

Proof of Theorem 3.1

By smoothness,

$$
\begin{aligned}
f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^t) &\leq \nabla f(\boldsymbol{x}^t)^\top (\underbrace{\boldsymbol{x}^{t+1} - \boldsymbol{x}^t}_{=\eta_t(\boldsymbol{y}^t - \boldsymbol{x}^t)}) + \frac{L}{2} \underbrace{\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2}_{=\eta_t^2 \|\boldsymbol{y}^t - \boldsymbol{x}^t\|_2^2 \leq \eta_t^2 d_{\mathcal{C}}^2} \\
&\leq \eta_t \nabla f(\boldsymbol{x}^t)^\top (\boldsymbol{y}^t - \boldsymbol{x}^t) + \frac{L}{2}\eta_t^2 d_{\mathcal{C}}^2 \\
&\leq \eta_t \nabla f(\boldsymbol{x}^t)^\top (\boldsymbol{x}^* - \boldsymbol{x}^t) + \frac{L}{2}\eta_t^2 d_{\mathcal{C}}^2 \qquad \text{(since } \boldsymbol{y}^t \text{ is minimizer)} \\
&\leq \eta_t \big(f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t)\big) + \frac{L}{2}\eta_t^2 d_{\mathcal{C}}^2 \qquad \text{(by convexity)}
\end{aligned}
$$

Letting $\Delta_t := f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)$ we get

$$
\Delta_{t+1} \leq (1 - \eta_t)\Delta_t + \frac{L d_{\mathcal{C}}^2}{2}\eta_t^2
$$

We then complete the proof by induction (which we omit here)

Strongly convex problems?

Can we hope to improve convergence guarantees of Frank-Wolfe in the presence of strong convexity?

- in general, NO
- maybe improvable under additional conditions

A negative result

Example:

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad & \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} \qquad (\boldsymbol{Q} \succ \boldsymbol{0}) & (3.2) \\ \text{subject to} \quad & \underbrace{\boldsymbol{x} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_k]\boldsymbol{v},\ \boldsymbol{v} \geq \boldsymbol{0},\ \boldsymbol{1}^\top \boldsymbol{v} = 1}_{\boldsymbol{x} \in \text{convex-hull}\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_k\}} \qquad (=: \Omega) \end{aligned}$$

- suppose interior$(\Omega) \neq \emptyset$
- suppose the optimal point $\boldsymbol{x}^*$ lies on the boundary of Ω and is not an extreme point

A negative result

Theorem 3.2 (Canon & Cullum, '68)

Let $\{\boldsymbol{x}^t\}$ be Frank-Wolfe iterates with exact line search for solving (3.2)*. Then $\exists$ an initial point $\boldsymbol{x}^0$ s.t. for every $\varepsilon > 0$,*

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \geq \frac{1}{t^{1+\varepsilon}} \qquad \textit{for infinitely many } t$$

- example: choose $\boldsymbol{x}^0 \in$ interior(Ω) obeying $f(\boldsymbol{x}^0) < \min_i f(\boldsymbol{a}_i)$
- in general, cannot improve $O(1/t)$ convergence guarantees

Positive results?

To achieve faster convergence, one needs additional assumptions

- example: strongly convex feasible set $\mathcal{C}$
- active research topics

An example of positive results

A set $\mathcal{C}$ is said to be μ-strongly convex if $\forall \lambda \in [0,1]$ and $\forall \boldsymbol{x}, \boldsymbol{z} \in \mathcal{C}$:

$$\mathcal{B}\Big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{z},\ \frac{\mu}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{z}\|_2^2\Big) \in \mathcal{C},$$

where $\mathcal{B}(\boldsymbol{a}, r) := \{\boldsymbol{y} \mid \|\boldsymbol{y}-\boldsymbol{a}\|_2 \le r\}$

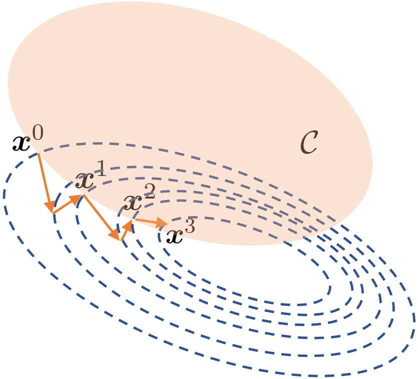
- example: ℓ_2 ball

Theorem 3.3 (Levitin & Polyak '66)

Suppose f is convex and L-smooth, and $\mathcal{C}$ is μ-strongly convex. Suppose $\|\nabla f(\boldsymbol{x})\|_2 \ge c > 0$ for all $\boldsymbol{x} \in \mathcal{C}$. Then under mild conditions, Frank-Wolfe with exact line search converges linearly

Projected gradient methods

Projected gradient descent



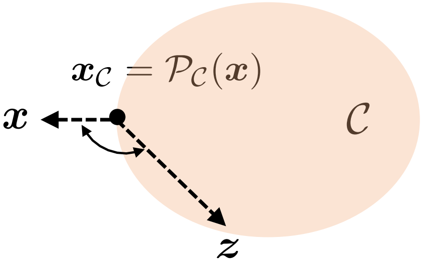
works well if projection onto $\mathcal{C}$ can be computed efficiently

for $t = 0, 1, \cdots$:

$$\boldsymbol{x}^{t+1} = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t))$$

where $\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}\in\mathcal{C}} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2$ is $\underbrace{\text{Euclidean projection}}_{\text{quadratic minimization}}$ onto $\mathcal{C}$

Descent direction

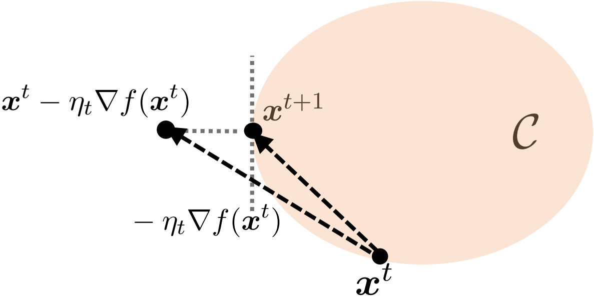


Fact 3.4 (Projection theorem)

Let $\mathcal{C}$ be closed & convex. Then $x_{\mathcal{C}}$ is the projection of x onto $\mathcal{C}$ iff

$$(x - x_{\mathcal{C}})^\top (z - x_{\mathcal{C}}) \leq 0, \qquad \forall z \in \mathcal{C}$$

Descent direction


$\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)$
$\boldsymbol{x}^{t+1}$
$\mathcal{C}$
$-\eta_t \nabla f(\boldsymbol{x}^t)$
$\boldsymbol{x}^t$

From the above figure, we know

$$-\nabla f(\boldsymbol{x}^t)^\top (\boldsymbol{x}^{t+1} - \boldsymbol{x}^t) \geq 0$$

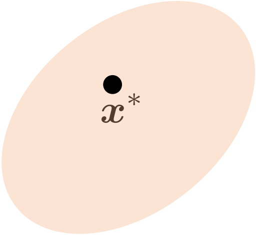
$\boldsymbol{x}^{t+1} - \boldsymbol{x}^t$ is positively correlated with the steepest descent direction

Strongly convex and smooth problems

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{C} \end{array}$$

- $f(\cdot)$: μ-strongly convex and L-smooth
- $\mathcal{C} \subseteq \mathbb{R}^n$: closed and convex

Convergence for strongly convex and smooth problems



Let's start with the simple case when x^* lies in the interior of $\mathcal{C}$ (so that $\nabla f(x^*) = \mathbf{0}$)

Convergence for strongly convex and smooth problems

Theorem 3.5

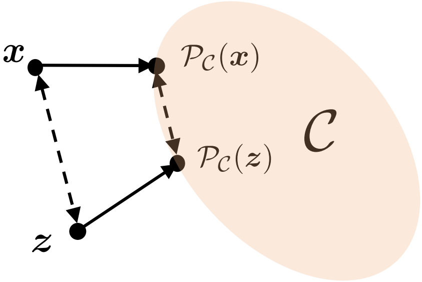
Suppose $\boldsymbol{x}^* \in \mathrm{int}(\mathcal{C})$, *and let* f *be* μ*-strongly convex and* L*-smooth. If* $\eta_t = \frac{2}{\mu+L}$, *then*

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2$$

where $\kappa = L/\mu$ *is condition number*

- the same convergence rate as for the unconstrained case

Aside: nonexpansiveness of projection operator



$\boldsymbol{x}$
$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$
$\mathcal{C}$
$\mathcal{P}_{\mathcal{C}}(\boldsymbol{z})$
$\boldsymbol{z}$

Fact 3.6 (Nonexpansivness of projection)

For any $\boldsymbol{x}$ *and* $\boldsymbol{z}$, *one has* $\|\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{z})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{z}\|_2$

Proof of Theorem 3.5

We have shown for the unconstrained case that

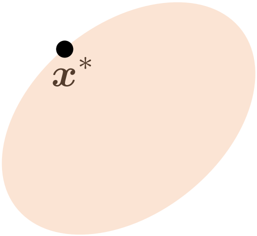
$$\|\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t) - \boldsymbol{x}^*\|_2 \le \frac{\kappa - 1}{\kappa + 1}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$$

From the nonexpansiveness of $\mathcal{P}_{\mathcal{C}}$, we know

$$\begin{aligned}
\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 &= \|\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*)\|_2 \\
&\le \|\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t) - \boldsymbol{x}^*\|_2 \\
&\le \frac{\kappa - 1}{\kappa + 1}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2
\end{aligned}$$

Apply it recursively to conclude the proof

Convergence for strongly convex and smooth problems



What happens if we don't know whether $x^* \in \text{int}(\mathcal{C})$?

- main issue: $\nabla f(x^*)$ may not be $\mathbf{0}$ (so prior analysis might fail)

Convergence for strongly convex and smooth problems

Theorem 3.7 (projected GD for strongly convex and smooth problems)

Let f be μ-strongly convex and L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2$$

- slightly weaker convergence guarantees than Theorem 3.5

Proof of Theorem 3.7

Let $\boldsymbol{x}^+ := \mathcal{P}_\mathcal{C}(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x}))$ and $\underbrace{\boldsymbol{g}_\mathcal{C}(\boldsymbol{x}) := \frac{1}{\eta}(\boldsymbol{x} - \boldsymbol{x}^+) = L(\boldsymbol{x} - \boldsymbol{x}^+)}_{\text{negative descent direction}}$

- $\boldsymbol{g}_\mathcal{C}(\boldsymbol{x})$ generalizes $\nabla f(\boldsymbol{x})$ and obeys $\boldsymbol{g}_\mathcal{C}(\boldsymbol{x}^*) = \boldsymbol{0}$

Main pillar:

$$\langle \boldsymbol{g}_\mathcal{C}(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \frac{1}{2L}\|\boldsymbol{g}_\mathcal{C}(\boldsymbol{x})\|_2^2 \tag{3.3}$$

- this generalizes the regularity condition for GD

With (3.3) in place, repeating GD analysis under the regularity condition gives

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2$$

which immediately establishes Theorem 3.7

Proof of Theorem 3.7 (cont.)

It remains to justify (3.3). To this end, it is seen that

$$\begin{aligned} 0 &\leq f(\boldsymbol{x}^+) - f(\boldsymbol{x}^*) = f(\boldsymbol{x}^+) - f(\boldsymbol{x}) + f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \\ &\leq \underbrace{\nabla f(\boldsymbol{x})^\top (\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|_2^2}_{\text{smoothness}} + \underbrace{\nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{x}^*) - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2}_{\text{strong convexity}} \\ &= \nabla f(\boldsymbol{x})^\top (\boldsymbol{x}^+ - \boldsymbol{x}^*) + \frac{1}{2L}\|\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x})\|_2^2 - \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2, \end{aligned}$$

which would establish (3.3) if

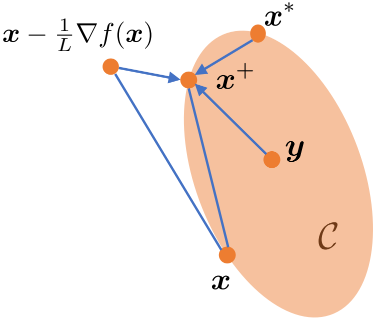
$$\nabla f(\boldsymbol{x})^\top (\boldsymbol{x}^+ - \boldsymbol{x}^*) \leq \underbrace{g_{\mathcal{C}}(\boldsymbol{x})^\top (\boldsymbol{x}^+ - \boldsymbol{x}^*)}_{= g_{\mathcal{C}}(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{x}^*) - \frac{1}{L}\|g_{\mathcal{C}}(\boldsymbol{x})\|_2^2} \quad \text{(projection only makes it better)} \tag{3.4}$$

This inequality is equivalent to

$$\left(\boldsymbol{x}^+ - \left(\boldsymbol{x} - L^{-1}\nabla f(\boldsymbol{x})\right)\right)^\top (\boldsymbol{x}^+ - \boldsymbol{x}^*) \leq 0 \tag{3.5}$$

This fact (3.5) follows directly from Fact 3.4

Remark



One can easily generalize (3.4) to (via the same proof)

$$\nabla f(\boldsymbol{x})^\top(\boldsymbol{x}^+ - \boldsymbol{y}) \leq g_{\mathcal{C}}(\boldsymbol{x})^\top(\boldsymbol{x}^+ - \boldsymbol{y}), \qquad \forall \boldsymbol{y} \in \mathcal{C} \tag{3.6}$$

This proves useful for subsequent analysis

Convex and smooth problems

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{C} \end{aligned}$$

- $f(\cdot)$: convex and L-smooth
- $\mathcal{C} \subseteq \mathbb{R}^n$: closed and convex

Convergence for convex and smooth problems

Theorem 3.8 (projected GD for convex and smooth problems)

Let f be convex and L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \frac{3L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{t+1}$$

- similar convergence rate as for the unconstrained case
- cannot replace $f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)$ with $\frac{1}{2}L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2$ since in general $\nabla f(\boldsymbol{x}^*) \neq 0$

Proof of Theorem 3.8

We first recall our main steps when handling the unconstrained case

Step 1: show cost improvement

$$f(\boldsymbol{x}^{t+1}) \leq f(\boldsymbol{x}^t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}^t)\|_2^2$$

Step 2: connect $\|\nabla f(\boldsymbol{x}^t)\|_2$ with $f(\boldsymbol{x}^t)$

$$\|\nabla f(\boldsymbol{x}^t)\|_2 \geq \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2} \geq \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}$$

Step 3: let $\Delta_t := f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)$ to get

$$\Delta_{t+1} - \Delta_t \leq -\frac{\Delta_t^2}{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}$$

and complete the proof by induction

Proof of Theorem 3.8 (cont.)

We then modify these steps for the constrained case. As before, set $\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x}^t) = L(\boldsymbol{x}^t - \boldsymbol{x}^{t+1})$, which generalizes $\nabla f(\boldsymbol{x}^t)$ in constrained case

Step 1: show cost improvement

$$f(\boldsymbol{x}^{t+1}) \leq f(\boldsymbol{x}^t) - \frac{1}{2L}\|\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x}^t)\|_2^2$$

Step 2: connect $\|\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x}^t)\|_2$ with $f(\boldsymbol{x}^t)$

$$\|\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x}^t)\|_2 \geq \frac{f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2} \geq \frac{f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}$$

Step 3: let $\Delta_t := f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)$ to get

$$\Delta_{t+1} - \Delta_t \leq -\frac{\Delta_{t+1}^2}{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}$$

and complete the proof by induction

Proof of Theorem 3.8 (cont.)

Main pillar: generalize smoothness condition (under convexity) as follows

Lemma 3.9

Suppose f *is convex and* L*-smooth. For any* $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$*, let* $\boldsymbol{x}^+ = \mathcal{P}_{\mathcal{C}}(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x}))$ *and* $g_{\mathcal{C}}(\boldsymbol{x}) = L(\boldsymbol{x} - \boldsymbol{x}^+)$*. Then*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^+) + g_{\mathcal{C}}(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2L}\|g_{\mathcal{C}}(\boldsymbol{x})\|_2^2$$

Proof of Theorem 3.8 (cont.)

Step 1: set $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{x}^t$ in Lemma 3.9 to reach

$$f(\boldsymbol{x}^t) \geq f(\boldsymbol{x}^{t+1}) + \frac{1}{2L} \| g_{\mathcal{C}}(\boldsymbol{x}^t) \|_2^2$$

as desired

Step 2: set $\boldsymbol{x} = \boldsymbol{x}^t$ and $\boldsymbol{y} = \boldsymbol{x}^*$ in Lemma 3.9 to get

$$\begin{aligned} 0 \geq f(\boldsymbol{x}^*) - f(\boldsymbol{x}^{t+1}) &\geq g_{\mathcal{C}}(\boldsymbol{x}^t)^\top (\boldsymbol{x}^* - \boldsymbol{x}^t) + \frac{1}{2L} \| g_{\mathcal{C}}(\boldsymbol{x}^t) \|_2^2 \\ &\geq g_{\mathcal{C}}(\boldsymbol{x}^t)^\top (\boldsymbol{x}^* - \boldsymbol{x}^t) \end{aligned}$$

which together with Cauchy-Schwarz yields

$$\| \boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x}^t) \|_2 \geq \frac{f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*)}{\| \boldsymbol{x}^t - \boldsymbol{x}^* \|_2} \tag{3.7}$$

Proof of Theorem 3.8 (cont.)

It also follows from our analysis for the strongly convex case that (by taking $\mu = 0$ in Theorem 3.7)

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2$$

which combined with (3.7) reveals

$$\|\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x}^t)\|_2 \geq \frac{f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}$$

Step 3: letting $\Delta_t = f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)$, the previous bounds together give

$$\Delta_{t+1} - \Delta_t \leq -\frac{\Delta_{t+1}^2}{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}$$

Use induction to finish the proof (which we omit here)

Proof of Lemma 3.9

$$
\begin{aligned}
f(\boldsymbol{y}) - f(\boldsymbol{x}^+) &= f(\boldsymbol{y}) - f(\boldsymbol{x}) - \big(f(\boldsymbol{x}^+) - f(\boldsymbol{x})\big) \\
&\geq \underbrace{\nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})}_{\text{convexity}} - \underbrace{\left(\nabla f(\boldsymbol{x})^\top (\boldsymbol{x}^+ - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|_2^2\right)}_{\text{smoothness}} \\
&= \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}^+) - \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|_2^2 \\
&\geq \boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}^+) - \frac{L}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|_2^2 \qquad \text{(by (3.6))} \\
&= \boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x})^\top \underbrace{(\boldsymbol{x} - \boldsymbol{x}^+)}_{=\frac{1}{L}\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x})} - \frac{L}{2}\| \underbrace{\boldsymbol{x}^+ - \boldsymbol{x}}_{=-\frac{1}{L}\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x})} \|_2^2 \\
&= \boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2L}\|\boldsymbol{g}_{\mathcal{C}}(\boldsymbol{x})\|_2^2
\end{aligned}
$$

Summary

- Frank-Wolfe: projection-free

	stepsize rule	convergence rate	iteration complexity
convex & smooth problems	$\eta_t \asymp \frac{1}{t}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

- projected gradient descent

	stepsize rule	convergence rate	iteration complexity
convex & smooth problems	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$
strongly convex & smooth problems	$\eta_t = \frac{1}{L}$	$O\left(\left(1-\frac{1}{\kappa}\right)^t\right)$	$O\left(\kappa \log \frac{1}{\varepsilon}\right)$

Reference

- "*Nonlinear programming (3rd edition)*," D. Bertsekas, 2016.
- "*Convex optimization: algorithms and complexity*," S. Bubeck, Foundations and trends in machine learning, 2015.
- "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.
- "*Convex optimization and algorithms*," D. Bertsekas, 2015.
- "*Conditional gradient algorithmsfor rank-one matrix approximations with a sparsity constraint*," R. Luss, M. Teboulle, SIAM Review, 2013.
- "*Revisiting Frank-Wolfe: projection-free sparse convex optimization*," M. Jaggi, ICML, 2013.
- "*A tight upper bound on the rate of convergence of Frank-Wolfe algorithm*," M. Canon and C. Cullum, SIAM Journal on Control, 1968.
- "*Constrained minimization methods*," E. Levitin and B. Polyak, USSR Computational mathematics and mathematical physics, 1966.