

Gradient methods for constrained problems



Yuxin Chen

Princeton University, Fall 2019

Outline

- Frank-Wolfe algorithm
- Projected gradient methods

Constrained convex problems

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

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

Feasible direction methods

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

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

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

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

Frank-Wolfe algorithm

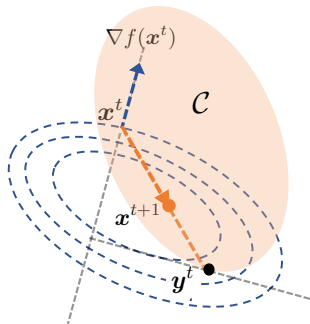
Frank-Wolfe algorithm was developed by Philip Wolfe and Marguerite Frank when they worked at / visited Princeton

Frank-Wolfe / conditional gradient algorithm

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

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

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



Frank-Wolfe / conditional gradient algorithm

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

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

- main step: linearization of the objective function (equivalent to $f(\mathbf{x}^t) + \langle \nabla f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t \rangle$)
 \implies linear optimization over a convex set
- appealing when linear optimization is cheap
- stepsize η_t determined by line search, or $\eta_t = \frac{2}{t+2}$
bias towards \mathbf{x}^t for large t

Frank-Wolfe can also be applied to nonconvex problems

Example (Luss & Teboulle '13)

$$\text{minimize}_x \quad -x^\top Qx \quad \text{subject to} \quad \|x\|_2 \leq 1 \quad (3.1)$$

for some $Q \succ 0$

Frank-Wolfe can also be applied to nonconvex problems

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

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

$$\implies \mathbf{x}^{t+1} = (1 - \eta_t)\mathbf{x}^t + \eta_t Q\mathbf{x}^t / \|Q\mathbf{x}^t\|_2$$

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

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

which is essentially **power method** for finding leading eigenvector of 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

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \frac{2Ld_C^2}{t+2}$$

where $d_C = \sup_{\mathbf{x}, \mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|_2$

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

Proof of Theorem 3.1

By smoothness,

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

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

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

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:

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

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

A negative result

Theorem 3.2 (Canon & Cullum, '68)

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

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \geq \frac{1}{t^{1+\varepsilon}} \quad \text{for infinitely many } t$$

- example: choose $\mathbf{x}^0 \in \text{interior}(\Omega)$ obeying $f(\mathbf{x}^0) < \min_i f(\mathbf{a}_i)$
- 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 $\forall \mathbf{x}, \mathbf{z} \in \mathcal{C}$:

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

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

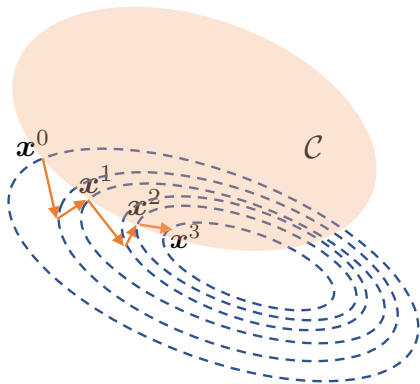
- example: ℓ_2 ball

Theorem 3.3 (Levitin & Polyak '66)

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

Projected gradient methods

Projected gradient descent



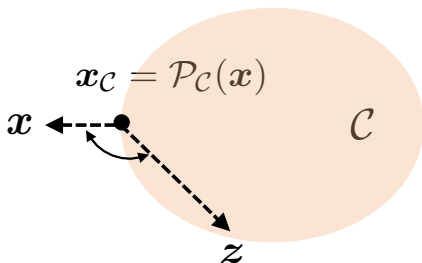
works well if projection
onto \mathcal{C} can be
computed efficiently

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

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

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

Descent direction

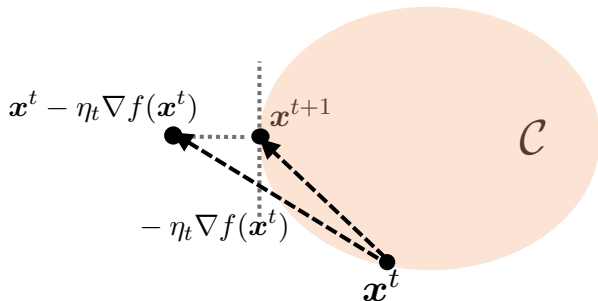


Fact 3.4 (Projection theorem)

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

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

Descent direction



From the above figure, we know

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

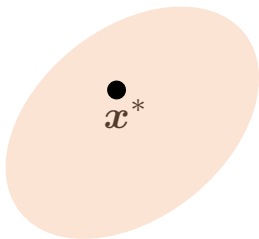
$\mathbf{x}^{t+1} - \mathbf{x}^t$ is positively correlated with the steepest descent direction

Strongly convex and smooth problems

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

- $f(\cdot)$: 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 \mathbf{x}^* lies in the interior of \mathcal{C} (so that $\nabla f(\mathbf{x}^*) = \mathbf{0}$)

Convergence for strongly convex and smooth problems

Theorem 3.5

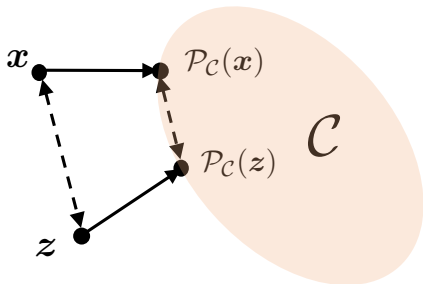
Suppose $\mathbf{x}^* \in \text{int}(\mathcal{C})$, and let f be μ -strongly convex and L -smooth. If $\eta_t = \frac{2}{\mu+L}$, then

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

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

- the same convergence rate as for the unconstrained case

Aside: nonexpansiveness of projection operator



Fact 3.6 (Nonexpansiveness of projection)

For any x and z , one has $\|P_C(x) - P_C(z)\|_2 \leq \|x - z\|_2$

Proof of Theorem 3.5

We have shown for the unconstrained case that

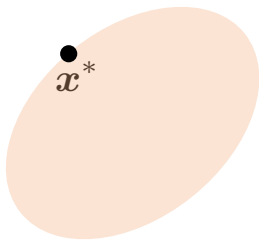
$$\|\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) - \mathbf{x}^*\|_2 \leq \frac{\kappa - 1}{\kappa + 1} \|\mathbf{x}^t - \mathbf{x}^*\|_2$$

From the nonexpansiveness of \mathcal{P}_C , we know

$$\begin{aligned} \|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2 &= \|\mathcal{P}_C(\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t)) - \mathcal{P}_C(\mathbf{x}^*)\|_2 \\ &\leq \|\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) - \mathbf{x}^*\|_2 \\ &\leq \frac{\kappa - 1}{\kappa + 1} \|\mathbf{x}^t - \mathbf{x}^*\|_2 \end{aligned}$$

Apply it recursively to conclude the proof

Convergence for strongly convex and smooth problems



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

- main issue: $\nabla f(\mathbf{x}^*)$ 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

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

- slightly weaker convergence guarantees than Theorem 3.5

Proof of Theorem 3.7

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

- $\mathbf{g}_C(\mathbf{x})$ generalizes $\nabla f(\mathbf{x})$ and obeys $\mathbf{g}_C(\mathbf{x}^*) = \mathbf{0}$

Main pillar:

$$\langle \mathbf{g}_C(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2 + \frac{1}{2L} \|\mathbf{g}_C(\mathbf{x})\|_2^2 \quad (3.3)$$

- this generalizes the regularity condition for GD

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

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

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

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

which would establish (3.3) if

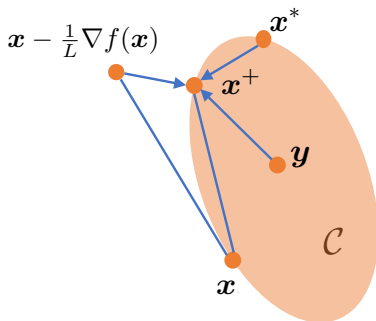
$$\begin{aligned} \nabla f(\mathbf{x})^\top (\mathbf{x}^+ - \mathbf{x}^*) &\leq \underbrace{\mathbf{g}_C(\mathbf{x})^\top (\mathbf{x}^+ - \mathbf{x}^*)}_{= \mathbf{g}_C(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^*) - \frac{1}{L} \|\mathbf{g}_C(\mathbf{x})\|_2^2} \quad (\text{projection only makes it better}) \end{aligned} \tag{3.4}$$

This inequality is equivalent to

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

This fact (3.5) follows directly from Fact 3.4

Remark



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

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

This proves useful for subsequent analysis

Convex and smooth problems

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

- $f(\cdot)$: 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

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

- similar convergence rate as for the unconstrained case

Proof of Theorem 3.8

We first recall our main steps when handling the unconstrained case

Step 1: show cost improvement

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

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

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

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

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

and complete the proof by induction

Proof of Theorem 3.8 (cont.)

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

Step 1: show cost improvement

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

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

$$\|g_C(\mathbf{x}^t)\|_2 \geq \frac{f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*)}{\|\mathbf{x}^t - \mathbf{x}^*\|_2} \geq \frac{f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*)}{\|\mathbf{x}^0 - \mathbf{x}^*\|_2}$$

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

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

and complete the proof by induction

Proof of Theorem 3.8 (cont.)

Main pillar: generalize smoothness condition as follows

Lemma 3.9

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

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

Proof of Theorem 3.8 (cont.)

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

$$f(\mathbf{x}^t) \geq f(\mathbf{x}^{t+1}) + \frac{1}{2L} \|g_C(\mathbf{x}^t)\|_2^2$$

as desired

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

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

which together with Cauchy-Schwarz yields

$$\|g_C(\mathbf{x}^t)\|_2 \geq \frac{f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*)}{\|\mathbf{x}^t - \mathbf{x}^*\|_2} \quad (3.7)$$

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)

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

which combined with (3.7) reveals

$$\|\mathbf{g}_C(\mathbf{x}^t)\|_2 \geq \frac{f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*)}{\|\mathbf{x}^0 - \mathbf{x}^*\|_2}$$

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

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

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

Proof of Lemma 3.9

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

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

- [1] "*Nonlinear programming (3rd edition)*," D. Bertsekas, 2016.
- [2] "*Convex optimization: algorithms and complexity*," S. Bubeck, Foundations and trends in machine learning, 2015.
- [3] "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.
- [4] "*Convex optimization and algorithms*," D. Bertsekas, 2015.
- [5] "*Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint*," R. Luss, M. Teboulle, SIAM Review, 2013.
- [6] "*Revisiting Frank-Wolfe: projection-free sparse convex optimization*," M. Jaggi, ICML, 2013.
- [7] "*A tight upper bound on the rate of convergence of Frank-Wolfe algorithm*," M. Canon and C. Cullum, SIAM Journal on Control, 1968.

Reference

- [8] "*Constrained minimization methods*," E. Levitin and B. Polyak, USSR Computational mathematics and mathematical physics, 1966.