Gradient methods for constrained problems

Yuxin Chen

Princeton University, Fall 2019
Outline

- Frank-Wolfe algorithm
- Projected gradient methods
Constrained convex problems

\[
\begin{align*}
\text{minimize}_x & \quad f(x) \\
\text{subject to} & \quad x \in C
\end{align*}
\]

- \(f(\cdot)\): convex function
- \(C \subseteq \mathbb{R}^n\): closed convex set
Feasible direction methods

Generate a feasible sequence \( \{x^t\} \subseteq C \) with iterations

\[
x^{t+1} = x^t + \eta_t d^t
\]

where \( d^t \) is a feasible direction (s.t. \( x^t + \eta_t d^t \in C \))

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

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

It follows that

\[ x_t \neq x^* \]

Lemma 2.5

Let \( f \) be convex and \( x^* \) be any minimizer with optimal value \( L \).

Proof of Lemma 2.5

Algorithm 2.3 Frank-wolfe (a.k.a. conditional gradient) algorithm

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

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

\[ \implies \text{linear optimization over a convex set} \]

- appealing when linear optimization is cheap

- stepsize \( \eta_t \) determined by line search, or \( \eta_t = \frac{2}{t + 2} \) bias towards \( x^t \) for large \( t \)

Frank-Wolfe / conditional gradient algorithm

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

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

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

\[ \implies \text{linear optimization over a convex set} \]

- appealing when linear optimization is cheap

- stepsize \( \eta_t \) determined by line search, or \( \eta_t = \frac{2}{t + 2} \) bias towards \( x^t \) for large \( t \)
Frank-Wolfe can also be applied to nonconvex problems

Example (Luss & Teboulle '13)

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

for some \( Q \succ 0 \)
Frank-Wolfe can also be applied to nonconvex problems

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

\[ y^t = \arg \min_{x: \|x\|_2 \leq 1} \langle \nabla f(x^t), x \rangle = \frac{\nabla f(x^t)}{\|\nabla f(x^t)\|_2} = \frac{Qx^t}{\|Qx^t\|_2} \]

\[ \implies x^{t+1} = (1 - \eta_t)x^t + \eta_tQx^t/\|Qx^t\|_2 \]

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

\[ x^{t+1} = Qx^t/\|Qx^t\|_2 \]

which is essentially power method for finding leading eigenvector of \( Q \)
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

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

where \( d_C = \sup_{x,y \in \mathcal{C}} \|x - y\|_2 \)

- for compact constraint sets, Frank-Wolfe attains \( \varepsilon \)-accuracy within \( O\left(\frac{1}{\varepsilon}\right) \) iterations
Proof of Theorem 3.1

By smoothness,

\[ f(x^{t+1}) - f(x^t) \leq \nabla f(x^t)^\top (x^{t+1} - x^t) + \frac{L}{2} \| x^{t+1} - x^t \|_2^2 \]

\[ = \eta_t (y^t - x^t) \]

\[ \leq \eta_t \nabla f(x^t)^\top (y^t - x^t) + \frac{L}{2} \eta_t^2 d_C^2 \]

\[ \leq \eta_t \nabla f(x^t)^\top (x^* - x^t) + \frac{L}{2} \eta_t^2 d_C^2 \quad \text{(since } y^t \text{ is minimizer)} \]

\[ \leq \eta_t (f(x^*) - f(x^t)) + \frac{L}{2} \eta_t^2 d_C^2 \quad \text{(by convexity)} \]

Letting \( \Delta_t := f(x^t) - f(x^*) \) we get

\[ \Delta_{t+1} \leq (1 - \eta_t) \Delta_t + \frac{Ld_C^2}{2} \eta_t^2 \]

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

minimize_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x + b^T x  \\
subject to \ x = [a_1, \cdots, a_k] v, \ v \geq 0, \ 1^T v = 1  \\
\ x \in \text{convex-hull}\{a_1, \cdots, a_k\}  \\

• suppose interior(\Omega) \neq \emptyset

• suppose the optimal point \ x^* \ lies on the boundary of \ Omega \ and is not an extreme point
A negative result

**Theorem 3.2 (Canon & Cullum, ’68)**

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

\[
f(x^t) - f(x^*) \geq \frac{1}{t^{1+\varepsilon}}
\]

for infinitely many \( t \)

- example: choose \( x^0 \in \text{interior}(\Omega) \) obeying \( f(x^0) < \min_i f(a_i) \)
- in general, cannot improve \( O(1/t) \) convergence guarantees
To achieve faster convergence, one needs additional assumptions

- example: strongly convex feasible set $C$
- active research topics
An example of positive results

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

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

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

- example: $\ell_2$ ball

**Theorem 3.3 (Levitin & Polyak ’66)**

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

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

$$x^{t+1} = P_C(x^t - \eta_t \nabla f(x^t))$$

where $P_C(x) := \arg \min_{z \in C} \|x - z\|_2^2$ is Euclidean projection onto $C$

works well if projection onto $C$ can be computed efficiently
Fact 3.4 (Projection theorem)

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

$$(x - x_\mathcal{C})^T (z - x_\mathcal{C}) \leq 0, \quad \forall z \in \mathcal{C}$$
Descent direction

Let $y_t = x_t - \frac{1}{t} \nabla f(x_t)$ be gradient update before projection. Then Fact 3.2 implies $\nabla f(x_t) \in (x_t \neq x_{t+1}) = \frac{1}{t} (x_t \neq y_t) \in (x_t \neq x_{t+1}) \Rightarrow 0$.

From the above figure, we know

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

$x_{t+1} - x_t$ is positively correlated with the steepest descent direction.
Strongly convex and smooth problems

\[
\begin{align*}
\text{minimize}_x & \quad f(x) \\
\text{subject to} & \quad x \in C
\end{align*}
\]

- \(f(\cdot)\): \(\mu\)-strongly convex and \(L\)-smooth
- \(C \subseteq \mathbb{R}^n\): closed and convex
Convergence for strongly convex and smooth problems

Let’s start with the simple case when $x^*$ lies in the interior of $C$ (so that $\nabla f(x^*) = 0$)
Convergence for strongly convex and smooth problems

Theorem 3.5

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

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

where $\kappa = \frac{L}{\mu}$ 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 $\|\mathcal{P}_C(x) - \mathcal{P}_C(z)\|_2 \leq \|x - z\|_2$
Proof of Theorem 3.5

We have shown for the unconstrained case that

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

From the nonexpansiveness of $P_C$, we know

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

Apply it recursively to conclude the proof.
Convergence for strongly convex and smooth problems

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

- main issue: $\nabla f(\mathbf{x}^*)$ may not be 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 $\mu$-strongly convex and $L$-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then

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

- slightly weaker convergence guarantees than Theorem 3.5
Proof of Theorem 3.7

Let $x^+ := P_C(x - \frac{1}{L} \nabla f(x))$ and $g_C(x) := \frac{1}{\eta} (x - x^+) = L(x - x^+)$

- $g_C(x)$ generalizes $\nabla f(x)$ and obeys $g_C(x^*) = 0$

Main pillar:

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

- this generalizes the regularity condition for GD

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

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

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

\[
0 \leq f(x^+) - f(x^*) = f(x^+) - f(x) + f(x) - f(x^*)
\]

\[
\leq \nabla f(x)^\top (x^+ - x) + \frac{L}{2} \|x^+ - x\|_2^2 + \nabla f(x)^\top (x - x^*) - \frac{\mu}{2} \|x - x^*\|_2^2
\]

\underline{smoothness} \hspace{2cm} \underline{strong convexity}

\[
= \nabla f(x)^\top (x^+ - x^*) + \frac{1}{2L} \|g_C(x)\|_2^2 - \frac{\mu}{2} \|x - x^*\|_2^2,
\]

which would establish (3.3) if

\[
\nabla f(x)^\top (x^+ - x^*) \leq \underbrace{g_C(x)^\top (x^+ - x^*)}_{\text{projection only makes it better}}
\]

\[
= g_C(x)^\top (x - x^*) - \frac{1}{L} \|g_C(x)\|_2^2
\]

(3.4)

This inequality is equivalent to

\[
(x^+ - (x - L^{-1}\nabla f(x)))^\top (x^+ - x^*) \leq 0
\]

(3.5)

This fact (3.5) follows directly from Fact 3.4
Remark

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

\[ \nabla f(x)^\top (x^+ - y) \leq g_C(x)^\top (x^+ - y), \quad \forall y \in C \]  

(3.6)

This proves useful for subsequent analysis.
Convex and smooth problems

\[ \begin{align*} 
\text{minimize}_x & \quad f(x) \\
\text{subject to} & \quad x \in C 
\end{align*} \]

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

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

- 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

\[ f(x^{t+1}) \leq f(x^t) - \frac{1}{2L} \|\nabla f(x^t)\|^2 \]

**Step 2:** connect \( \|\nabla f(x^t)\|_2 \) with \( f(x^t) \)

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

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

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

and complete the proof by induction
Proof of Theorem 3.8 (cont.)

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

**Step 1:** show cost improvement

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

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

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

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

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

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 $x, y \in C$, let $x^+ = \mathcal{P}_C(x - \frac{1}{L} \nabla f(x))$ and $g_C(x) = L(x - x^+)$. Then

$$f(y) \geq f(x^+) + g_C(x)^\top(y - x) + \frac{1}{2L} \|g_C(x)\|_2^2$$
Proof of Theorem 3.8 (cont.)

**Step 1:** set $x = y = x^t$ in Lemma 3.9 to reach

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

as desired

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

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

$$\geq g_C(x^t)^\top (x^* - x^t)$$

which together with Cauchy-Schwarz yields

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

$$\|x^t - x^*\|_2 \leq \|x^0 - x^*\|_2$$

which combined with (3.7) reveals

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

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

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

Use induction to finish the proof (which we omit here)
Proof of Lemma 3.9

\[ f(y) - f(x^+) = f(y) - f(x) - (f(x^+) - f(x)) \]
\[ \geq \nabla f(x)^\top (y - x) - \left( \nabla f(x)^\top (x^+ - x) + \frac{L}{2} \|x^+ - x\|^2_2 \right) \]

convexity

\[ = \nabla f(x)^\top (y - x^+) - \frac{L}{2} \|x^+ - x\|^2_2 \]

smoothness

\[ \geq g_C(x)^\top (y - x^+) - \frac{L}{2} \|x^+ - x\|^2_2 \quad \text{(by (3.6))} \]

\[ = g_C(x)^\top (y - x) + g_C(x)^\top (x - x^+) - \frac{L}{2} \|x^+ - x\|^2_2 \]

\[ = g_C(x)^\top (y - x) + \frac{1}{L} g_C(x) + \frac{1}{2L} \|g_C(x)\|^2 \]

Gradient methods (constrained case) 3-37
### Summary

- **Frank-Wolfe: projection-free**

<table>
<thead>
<tr>
<th></th>
<th>stepsize rule</th>
<th>convergence rate</th>
<th>iteration complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>convex &amp; smooth problems</td>
<td>( \eta_t \approx \frac{1}{t} )</td>
<td>( O\left(\frac{1}{t}\right) )</td>
<td>( O\left(\frac{1}{\varepsilon}\right) )</td>
</tr>
</tbody>
</table>

- **Projected gradient descent**

<table>
<thead>
<tr>
<th></th>
<th>stepsize rule</th>
<th>convergence rate</th>
<th>iteration complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>convex &amp; smooth problems</td>
<td>( \eta_t = \frac{1}{L} )</td>
<td>( O\left(\frac{1}{t}\right) )</td>
<td>( O\left(\frac{1}{\varepsilon}\right) )</td>
</tr>
<tr>
<td>strongly convex &amp; smooth problems</td>
<td>( \eta_t = \frac{1}{L} )</td>
<td>( O\left((1 - \frac{1}{\kappa})^t\right) )</td>
<td>( O\left(\kappa \log \frac{1}{\varepsilon}\right) )</td>
</tr>
</tbody>
</table>


Reference