Gradient methods for constrained problems

Yuxin Chen
Princeton University, Spring 2018
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 feasible sequence \( \{x^t\} \subseteq C \) with iterations

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

where \( d^t \): 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: \quad $y^t := \arg\min_{x \in \mathcal{C}} \langle \nabla f(x^t), x \rangle$ \hfill (direction finding)
3: \quad $x^{t+1} = (1 - \eta_t)x^t + \eta_t y^t$ \hfill (line search and update)

\[ y^t = \arg\min_{x \in \mathcal{C}} \langle \nabla f(x^t), x - x^t \rangle \]
Frank-Wolfe / conditional gradient algorithm

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

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

- main step: linearization of objective function (equivalent to $f(x^t) + \langle \nabla f(x^t), x - x^t \rangle$)
  \[ \Longrightarrow \text{linear optimization over convex set} \]
- 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)

\[
\begin{align*}
\text{minimize}_x & \quad -x^\top Qx \\
\text{subject to} & \quad \|x\|_2 \leq 1
\end{align*}
\] (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,

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

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

Set \( \eta_t = \arg \min_{0 \leq \eta \leq 1} f\left((1 - \eta)x^t + \eta \frac{Qx^t}{\|Qx^t\|_2}\right) = 1. \) 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 $d_C = \sup_{x,y \in C} \|x - y\|_2$

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

By smoothness,

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

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

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

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

\[ \leq \eta_t (f(x^*) - f(x^t)) + \frac{L}{2} \eta_t^2 d_C^2 \] (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 proof by induction (which we omit here)
Proof of Theorem 3.1

Before proceeding, we pause to get a sense of recursive inequality

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

It might be convenient to look at continuous limit (with “≤” replaced by “=” for simplicity)

\[ \delta'(t) = -\eta_t \delta(t) + O(\eta_t^2) \]

Suppose \( \eta_t \approx t^{-\alpha} \) and \( \delta(t) \approx t^{-\beta} \). Then above three terms are orderwise equivalent to

\[ -\beta t^{-\beta - 1}, -t^{-\alpha - \beta}, t^{-2\alpha} \]

Matching all exponents, we have \( -\beta - 1 = -\alpha - \beta = -2\alpha \) and hence \( \alpha = \beta = 1 \). This is helpful when guessing convergence rate and stepsize
Can we hope to improve convergence guarantees of Frank-Wolfe in presence of strong convexity?

- in general, NO
- maybe improvable under additional conditions
A negative result

Example:

\[
\begin{align*}
\text{minimize}_{x \in \mathbb{R}^n} & \quad \frac{1}{2} x^\top Q x + b^\top x \\
\text{subject to} & \quad x = [a_1, \ldots, a_k] v, \quad v \geq 0, \quad 1^\top v = 1 \\
& \quad x \in \text{convex-hull}\{a_1, \ldots, a_k\}
\end{align*}
\] (3.2)

- suppose interior(Ω) ≠ ∅
- suppose optimal point \(x^*\) lies on boundary of Ω and is not 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 \( \exists \) 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 rate
Positive results?

To achieve faster convergence, one needs additional assumptions

- example: strongly convex feasible set $C$
- still active research topics (e.g. see E. Hazan’s group)
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

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

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

for $t = 0, 1, \cdots$

works well if projection onto $C$ can be computed efficiently
Descent direction

**Fact 3.4 (Projection theorem)**

Let $\mathcal{C}$ be closed convex set. Then $x_C$ is projection of $x$ onto $\mathcal{C}$ iff

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

\[ \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) \]

\[ \mathbf{x}^t \]

\[ \mathbf{x}^{t+1} \]

\[ \mathbf{C} \]

From above figure, we know

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

\[ \mathbf{x}^{t+1} - \mathbf{x}^t \text{ is positively correlated with steepest descent direction} \]
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 convex set
Let’s start with simple case when $x^*$ lies in 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 $\kappa = L/\mu$ is condition number

• same convergence rate as for 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 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 nonexpansiveness of \( \mathcal{P}_C \), we know

\[ \| \mathbf{x}^{t+1} - \mathbf{x}^* \|_2 = \| \mathcal{P}_C(\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t)) - \mathcal{P}_C(\mathbf{x}^*) \|_2 \]
\[ \leq \| \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) - \mathbf{x}^* \|_2 \]
\[ \leq \frac{\kappa - 1}{\kappa + 1} \| \mathbf{x}^t - \mathbf{x}^* \|_2 \]

Apply it recursively to concludes proof
Convergence for strongly convex and smooth problems

Proof of Lemma 2.5

It follows that
\[ x_{t+1} \neq x_{\bar{u}} = \frac{\nabla f(x_t) - \nabla f(x_{\bar{u}})}{\| \nabla f(x_t) - \nabla f(x_{\bar{u}}) \|} \]

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

- main issue: \( \nabla f(x^*) \) 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 rate than Theorem 3.5
Proof of Theorem 3.7

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

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

Main pillar:

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

(3.3)

- this generalizes regularity condition for GD

With (3.3) in place, repeating GD analysis under 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
It remains to justify (3.3). To this end, it is seen that
\[
0 \leq f(x^+) - f(x^*) = f(x^+) - f(x) + f(x) - f(x^*) \\
\leq \nabla f(x)^\top (x^+ - x) + \frac{L}{2} \|x^+ - x\|^2 + \nabla f(x)^\top (x - x^*) - \frac{\mu}{2} \|x - x^*\|^2
\]

which would establish (3.3) if
\[
\nabla f(x)^\top (x^+ - x^*) \leq g_C(x)^\top (x^+ - x^*) - \frac{1}{L} \|g_C(x)\|^2
\]

This inequality is equivalent to
\[
(x^+ - (x - L^{-1}\nabla f(x)))^\top (x^+ - x^*) \leq 0,
\]
which is direct consequence of Fact 3.4
Remark

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

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

(3.5)

This proves useful for subsequent analysis
Convex and smooth problems

\[ \text{minimize}_{\mathbf{x}} \quad f(\mathbf{x}) \]
\[ \text{subject to} \quad \mathbf{x} \in C \]

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

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

- similar convergence rate as for unconstrained case
Proof of Theorem 3.8

We first recall our main steps when establishing unconstrained case

**Step 1:** show cost improvement

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

**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 proof by induction
Proof of Theorem 3.8 (cont.)

We will modify these steps for 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 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 \( x^+ = P_C(x - \frac{1}{L}\nabla f(x)) \) 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 \( y = x^\ast \) in Lemma 3.9 to get

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

\[
\geq g_C(x^t)^\top (x^\ast - x^t)
\]

which together with Cauchy-Schwarz yields

\[
\|g_C(x^t)\|_2 \geq \frac{f(x^{t+1}) - f(x^\ast)}{\|x^t - x^\ast\|_2}
\]
Proof of Theorem 3.8 (cont.)

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

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

which reveals

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

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

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

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

\[ f(y) - f(x^+) = f(y) - f(x) - (f(x^+) - f(x)) \]

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

By convexity

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

By smoothness

\[ \geq g_c(x)^T (y - x^+) - \frac{L}{2} \| x^+ - x \|^2 \]

(by (3.5))

\[ = g_c(x)^T (y - x) + g_c(x)^T (x - x^+) - \frac{L}{2} \| x^+ - x \|^2 \]

\[ = \frac{1}{L} g_c(x) + \frac{1}{2L} \| g_c(x) \|^2 \]

Gradient methods (constrained case)
### 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


