Introductory Lectures on Optimization

General Convex Problem (3)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

November 27, 2024

Outline

Part I
General Lower Bound

Problem Class

In the previous section we have introduced a class of general convex functions. These functions can be non-smooth and therefore the corresponding minimization problem can be quite difficult. As for smooth problems, let us try to derive a lower complexity bounds, which will help us to evaluate the performance of numerical methods.

In this section we derive such bounds for the following unconstrained minimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) \tag{1}$$

where f is a convex function.

Problem Class

Our problern dass is as follows:

Model:	1 Unconstrained minimization. 2 f is convex on $\mathbb{R}^n$ and Lipschitz continuous on a bounded set.
Oracle:	First-order black box: at each point $\hat{\boldsymbol{x}}$ we can compute $f(\hat{\boldsymbol{x}}), g(\hat{\boldsymbol{x}}) \in \partial f(\hat{\boldsymbol{x}})$, $g(\hat{\boldsymbol{x}})$ is an arbitrary subgradient.
Approximate solution:	Find $\bar{\boldsymbol{x}} \in \mathbb{R}^n : f(\bar{\boldsymbol{x}}) - f^* \leq \epsilon$。
Methods	Generate a sequence $\{\boldsymbol{x}_k\} : \boldsymbol{x}_k \in \boldsymbol{x}_0 + \mathrm{Lin}\{g(\boldsymbol{x}_0), \ldots, g(\boldsymbol{x}_{k-1})\}$.

(2)

Special Function family

Let us fix some constants $\mu > 0$ and $\gamma > 0$. Consider the family of functions

$$f_k(\boldsymbol{x}) = \gamma \max_{1 \le i \le k} \boldsymbol{x}^{(i)} + \frac{\mu}{2} \|\boldsymbol{x}\|^2, \quad k = 1 \dots n.$$

Remark. (1) k is smaller than the dimension. (2) $f_k(0) = 0$. (3) The first term of f_k is negative if the first k components are negative.

Using the rules of subdifferential calculus, described in last Section (3.1.6), we can write down an expression for the subdifferential of f_k at x. That is

$$\partial f_k(\boldsymbol{x}) = \mu \boldsymbol{x} + \gamma \mathrm{Conv}\{e_i | i \in I(\boldsymbol{x})\},$$
$$I(\boldsymbol{x}) = \{1 \le j \le k, \boldsymbol{x}^{(j)} = \max_{1 \le i \le k} \boldsymbol{x}^{(i)}\}.$$

Special Function family

1. Local Lipschitz Continuity of f_k

Therefore for any $\boldsymbol{x}, \boldsymbol{y} \in B_2(0, \rho), \rho > 0$, and $g_k(\boldsymbol{y}) \in \partial f_k(\boldsymbol{y})$ we have

$$\begin{aligned} f_k(\boldsymbol{y}) - f_k(\boldsymbol{x}) &\leq \langle g_k(\boldsymbol{y}),\ \boldsymbol{y} - \boldsymbol{x} \rangle \\ &\leq \|g_k(\boldsymbol{y})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \leq (\mu\rho + \gamma) \|\boldsymbol{y} - \boldsymbol{x}\| . \end{aligned}$$

Thus, f_k is Lipschitz continuous on $B_2(0, \rho)$ with Lipschitz constant

$$M = \mu\rho + \gamma.$$

Special Function family

2. minimum of f_k:

Further, consider the point $\boldsymbol{x}_k^*$ with the coordinates

$$(\boldsymbol{x}_k^*)^{(i)} = \begin{cases} -\frac{\gamma}{\mu k}, & 1 \le i \le k, \\ 0, & k+1 \le i \le n. \end{cases}$$

It is easy to check that $0 \in \partial f_k(\boldsymbol{x}_k^*)$, and therefore $\boldsymbol{x}_k^*$ is the minimum of $f_k(\boldsymbol{x})$ (See theorem 3.1.15). Note that

$$R_k \equiv \|\boldsymbol{x}_k^*\| = \frac{\gamma}{\mu\sqrt{k}}, \quad f_k^* = -\frac{\gamma^2}{\mu k} + \frac{\mu}{2} R_k^2 = -\frac{\gamma^2}{2\mu k}.$$

Resisting Oracle

3. The subgradient of f_k:

Let us describe now a resisting oracle for function $f_k(\boldsymbol{x})$. Since the analytical form of this function is fixed, the resistance of this oracle consists in providing us with the worst possible subgradient at each test point. The algorithmic scheme of this oracle is as follows.

Input:	$\boldsymbol{x} \in \mathbb{R}^n$.
Main Loop:	$f := -\infty;\ \ i^* := 0;$ for $j := 1$ to k do if $\boldsymbol{x}^{(j)} > f$ then $\{f := \boldsymbol{x}^{(j)}; i^* := j\};$ $f := \gamma f + \frac{\mu}{2} \|\boldsymbol{x}\|^2;\ \ g := e_{i^*} + \mu \boldsymbol{x};$
Output:	$f_k(\boldsymbol{x}) := f,\ \ g_k(\boldsymbol{x}) := g \in \mathbb{R}^n.$

Resisting Oracle

At the first glance, there is nothing special in this scheme.

1. Its main loop is just a standard process for finding a maximal coordinate of a vector from $\mathbb{R}^n$.
2. However, the main feature of this loop is that we always form the subgradient as a coordinate vector.
3. Moreover, this coordinate corresponds to i^*, which is the first maximal component of vector x.

Let us check what happens with a minimizing sequence, which uses such an oracle.

Resisting Oracle

4. Characteristic of $\boldsymbol{x}_{i+1}$ generated by resisting oracle:

Let us choose starting point $\boldsymbol{x}_0 = 0$. Denote

$$\mathbb{R}^{p,n} = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x}^{(i)} = 0, p+1 \leq i \leq n\}.$$

Since $\boldsymbol{x}_0 = 0$, the answer of the oracle is $f_k(\boldsymbol{x}_0) = 0$ and $g_k(\boldsymbol{x}_0) = e_1$. Therefore the next point of the sequence, $\boldsymbol{x}_1$, necessarily belongs to $\mathbb{R}^{1,n}$.

Assurne now that the current test point of the sequence, $\boldsymbol{x}_i$, belongs to $\mathbb{R}^{p,n}, 1 \leq p \leq k$. Then the oracle will return a subgradient

$$g = \mu \boldsymbol{x}_i + \gamma e_{i^*},$$

where $i^* \leq p+1$. Therefore, if $\boldsymbol{x}_i$ belongs to $\mathbb{R}^{p,n}, 1 \leq p \leq k$, the next test point $\boldsymbol{x}_{i+1}$ belongs to $\mathbb{R}^{p+1,n}$.

Resisting Oracle

4. Characteristic of x_{i+1} generated by resisting oracle:

This simple reasoning proves that for all $i, 1 \leq i \leq k$, we have $x_i \in \mathbb{R}^{i,n}$.

Consequently, for $i : 1 \leq i \leq k-1$, we cannot improve the starting value of the objective function:

$$f_k(x_i) \geq \gamma \max_{1 \leq j \leq k} x_i^{(j)} \geq 0 = f_k(x_0).$$

Remark. For $f = f_{k+1}$, we know $f(x_k) \geq 0$.

Resisting Oracle

Let us convert this observation in a lower complexity bound. Let us fix some parameters of our problern class $\mathcal{P}(\boldsymbol{x}_0, R, M)$, that is $R > 0$ and $M > 0$. In addition to (2) we assume that

- the solution of problem (1), $\boldsymbol{x}^*$, exists and $\boldsymbol{x}^* \in B_2(\boldsymbol{x}_0, R)$,
- f is Lipschitz continuous on $B_2(\boldsymbol{x}_0, R)$, with constant $M > 0$.

Lower Bound

Theorem 1 (Theorem 3.2.1)

For any class $\mathcal{P}(\boldsymbol{x}_0, R, M)$ and any $k, 0 \leq k \leq n-1$, there exists a function $f \in \mathcal{P}(\boldsymbol{x}_0, R, M)$ such that

$$\boldsymbol{x}_k \in \boldsymbol{x}_0 + \text{Lin}\{g(\boldsymbol{x}_0), \ldots, g(\boldsymbol{x}_{k-1})\}$$

for any optimization scheme, which generates a sequence $\{\boldsymbol{x}_k\}$ satisfying the condition

$$f(\boldsymbol{x}_k) - f^* \geq \frac{MR}{2(1+\sqrt{k+1})}.$$

Lower Bound

Proof. Without lass of generality we can assume that $\boldsymbol{x}_0 = 0$. Let us choose $f(\boldsymbol{x}) = f_{k+1}(\boldsymbol{x})$ with

$$\gamma = \frac{\sqrt{k+1}\, M}{1+\sqrt{k+1}}, \quad \mu = \frac{M}{(1+\sqrt{k+1})R}.$$

Then

$$f^* = f^*_{k+1} = -\frac{\gamma^2}{2\mu(k+1)} = -\frac{MR}{2(1+\sqrt{k+1})},$$
$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| = R_{k+1} = \frac{\gamma}{\mu\sqrt{k+1}} = R,$$

and $f(\boldsymbol{x})$ is Lipschitz continuous on $B_2(\boldsymbol{x}_0, R)$, with constant $\mu R + \gamma = M$. Note that $\boldsymbol{x}_k \in \mathbb{R}^{k,n}$. Hence $f(\boldsymbol{x}_k) - f^* \geq -f^*$. □

Lower Bound

The lower complexity bound presented in Theorem 3.2.1 is uniform in the dimension of the space of variables. As for the lower bound of Theorem 2.1.7, it can be applied to problems with very large dimension, or to efficiency analysis of starting iterations of a minimization scheme ($k \leq n - 1$).

We will see that our lower estimate is exact: There exist minimization methods, which have the rate of convergence proportional to this lower bound.

Comparing this bound with the lower bound for smooth minimization problems, we can see that now the possible convergence rate is much slower.

Part II
Subgradient Descent

Problem

At this moment we are interested in the following problem:

$$\min \{f(\boldsymbol{x}) | x \in Q\}, \tag{3}$$

where Q is a closed convex set, and f is a function, which is convex on $\mathbb{R}^n$.

We are going to study some methods for solving (3), which employ subgradients $g(\boldsymbol{x})$ of the objective function. As compared with the smooth problem, our goal now is much more complicated. Indeed, even in the simplest situation, when $Q \equiv \mathbb{R}^n$ be a poor replacement for the gradient of smooth function.

For example, we cannot be sure that the value of the objective function is decreasing in the direction $-g(\boldsymbol{x})$. We cannot expect that $g(\boldsymbol{x}) \to 0$ as $\boldsymbol{x}$ approaches a solution of our problem, etc.

Property of Subgradients

We have proved this property in Corollary 3.1.4: At any $\boldsymbol{x} \in Q$ the following inequality holds:

$$\langle g(\boldsymbol{x}),\ \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0. \tag{4}$$

This simple inequality leads to two consequences, which form a basis for any nonsmooth minimization method. Namely:

- The direction from $\boldsymbol{x}^*$ to $\boldsymbol{x}$ is opposite to the direction $-g(\boldsymbol{x})$.
- Inequality (4) cuts $\mathbb{R}^n$ into two half-spaces. Only one of them contains $\boldsymbol{x}^*$.

Main Lemma

Nonsmooth minimization methods cannot employ the idea of relaxation or approximation. There is another concept, underlying all these schemes. That is the concept of localization.

However, to go forward with this concept, we have to develop some special technique, which allows us to estimate the quality of an approximate solution to problem (3).

Main Lemma

Let us fix some $\bar{\boldsymbol{x}} \in \mathbb{R}^n$. For $\boldsymbol{x} \in \mathbb{R}^n$ with $g(\boldsymbol{x}) \neq 0$ define

$$v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{\|g(\boldsymbol{x})\|} \langle g(\boldsymbol{x}),\ \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle.$$

If $g(\boldsymbol{x}) = 0$, then define $v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}) = 0$. Clearly, $v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}) \leq \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|$.

The value $v_f(\bar{\boldsymbol{x}}, \boldsymbol{x})$ have a natural geometric interpretation. Consider a point $\boldsymbol{x}$ such that $g(\boldsymbol{x}) \neq 0$ and $\langle g(\boldsymbol{x}),\ \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle \geq 0$. Let us Iook at the point $\boldsymbol{y} = \bar{\boldsymbol{x}} + v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}) g(\boldsymbol{x}) / \|g(\boldsymbol{x})\|$. Then

$$\langle g(\boldsymbol{x}),\ \boldsymbol{x} - \boldsymbol{y} \rangle = \langle g(\boldsymbol{x}),\ \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle - v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}) \|g(\boldsymbol{x})\| = 0$$

and $\|\boldsymbol{y} - \bar{\boldsymbol{x}}\| = v_f(\bar{\boldsymbol{x}}, \boldsymbol{x})$. Thus, $v_f(\bar{\boldsymbol{x}}, \boldsymbol{x})$ is a distance from the point $\bar{\boldsymbol{x}}$ to hyperplane $\{\boldsymbol{z} : \langle g(\boldsymbol{x}),\ \boldsymbol{x} - \boldsymbol{z} \rangle = 0\}$.

Main Lemma

Let us introduce a function that measures the variation of function f with respect to the point $\bar{\boldsymbol{x}}$. For $t \geq 0$ define

$$\omega_f(\bar{\boldsymbol{x}}; t) = \max\{f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) | \, \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq t\}.$$

If $t < 0$, we set $\omega_f(\bar{\boldsymbol{x}}; t) = 0$. Clearly, the function ω_f possesses the following properties:

- For all $t \leq 0$, $\omega_f(\bar{\boldsymbol{x}}; 0) = 0$.
- $\omega_f(\bar{\boldsymbol{x}}; t)$ is a nondecreasing function of t, $t \in \mathbb{R}^1$.
- $f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq \omega_f(\bar{\boldsymbol{x}}; \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|)$。

Main Lemma

It is important that in the convex situation the last inequality can be strengthened.

Lemma 2 (Lemma 3.2.1)

For any $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq \omega_f(\bar{\boldsymbol{x}}; v_f(\bar{\boldsymbol{x}}; \boldsymbol{x})). \tag{5}$$

If $f(\boldsymbol{x})$ is Lipschitz continuous on $B_2(\bar{\boldsymbol{x}}, R)$ with some M, then

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq M(v_f(\bar{\boldsymbol{x}}; \boldsymbol{x}))_+, \tag{6}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$ with $\omega_f(\bar{\boldsymbol{x}}; \boldsymbol{x}) \leq \mathbb{R}$.

Main Lemma

Proof.

If $\langle g(\boldsymbol{x}),\ \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle \leq 0$, then $f(\bar{\boldsymbol{x}}) \geq f(\boldsymbol{x}) + \langle g(\boldsymbol{x}),\ \bar{\boldsymbol{x}} - \boldsymbol{x} \rangle \geq f(\boldsymbol{x})$. This implies that $v_f(\bar{\boldsymbol{x}}; \boldsymbol{x}) \leq 0$. Hence, $\omega_f(\bar{\boldsymbol{x}}; v_f(\bar{\boldsymbol{x}}; \boldsymbol{x})) = 0$ and (5) holds.

Let $\langle g(\boldsymbol{x}),\ \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle > 0$. For

$$\boldsymbol{y} = \bar{\boldsymbol{x}} + \frac{1}{\|g(\boldsymbol{x})\|} v_f(\bar{\boldsymbol{x}}; \boldsymbol{x}) g(\boldsymbol{x})$$

we have $\langle g(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle = 0$ and $\|\boldsymbol{y} - \bar{\boldsymbol{x}}\| = v_f(\bar{\boldsymbol{x}}; \boldsymbol{x})$.

Main Lemma

Proof. (Continued)

Therefore,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle g(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle = f(\boldsymbol{x}),$$

and

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{y}) - f(\bar{\boldsymbol{x}}) \leq \omega_f(\bar{\boldsymbol{x}}; \|\boldsymbol{y} - \bar{\boldsymbol{x}}\|) = \omega_f(\bar{\boldsymbol{x}}; v_f(\bar{\boldsymbol{x}}; \boldsymbol{x})).$$

That is the (5).

If f is Lipschitz continuous on $B_2(\bar{\boldsymbol{x}}, R)$ and $0 \leq v_f(\bar{\boldsymbol{x}}; \boldsymbol{x}) \leq R$, then $\boldsymbol{y} \in B_2(\bar{\boldsymbol{x}}, R)$. Hence

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{y}) - f(\bar{\boldsymbol{x}}) \leq M \|\boldsymbol{y} - \bar{\boldsymbol{x}}\| = M v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}).$$

□

Main Lemma

Let us fix some x^*, a solution to problem (3). The values $v_f(\boldsymbol{x}^*; \boldsymbol{x})$ allow us to estimate the quality of localization sets.

Definition 3
Let $\{\boldsymbol{x}_i\}_{i=0}^{\infty}$ be a sequence in Q. Define

$$S_k = \{\boldsymbol{x} \in Q | \langle g(\boldsymbol{x}_i),\ \boldsymbol{x}_i - \boldsymbol{x} \rangle \geq 0, i = 0 \dots k\}.$$

We call this set the localization set of problem (3) generated by sequence $\{\boldsymbol{x}_i\}_{i=0}^{\infty}$.

Main Lemma

Note that in view of inequality (4), for all $k \geq 0$ we have $\boldsymbol{x}^* \in S_k$. Denote

$$v_i = v_f(\boldsymbol{x}^*; \boldsymbol{x}_i) (\geq 0), \quad v_k^* = \min_{0 \leq i \leq k} v_i.$$

Thus,

$$v_k^* = \max\{r | \langle g(\boldsymbol{x}_i),\ \boldsymbol{x}_i - \boldsymbol{x} \rangle \geq 0, i = 0 \dots k, \forall \boldsymbol{x} \in B_2(\boldsymbol{x}^*, r)\}.$$

Main Lemma

Lemma 4
Let $f_k^* = \min_{0 \le i \le k} f(\boldsymbol{x}_i)$. Then $f_k^* - f^* \le \omega_f(\boldsymbol{x}^*; v_k^*)$.

证明.

Using Lemma 2, we have

$$\omega_f(\boldsymbol{x}^*; v_k^*) = \min_{0 \le i \le k} \omega_f(\boldsymbol{x}^*; v_i) \ge \min_{0 \le i \le k} [f(\boldsymbol{x}_i) - f^*] = f_k^* - f^*.$$

□

Remark 1: The first inequality comes from that $\omega_f(\bar{\boldsymbol{x}}; t)$ is a non-decreasing function on t, $t \in \mathbb{R}^1$.

Problem

Consider the problem

$$\min\{f(\boldsymbol{x})|\boldsymbol{x} \in Q\}, \tag{7}$$

where f is a convex on $\mathbb{R}^n$, Q is a simple closed convex set. The term "simple" means that we can solve explicitly some simple mini- mization problems over Q.

In accordance to the goals of this section, we have to be able to find in a reasonably cheap way a Euclidean projection of any point onto Q.

Problem

We assume that problem (7) is equipped with a first-order oracle, which at any test point $\bar{\boldsymbol{x}}$ provides us with the value of objective function $f(\bar{\boldsymbol{x}})$ and with one of its subgradients $g(\bar{\boldsymbol{x}})$.

As usual, we try first a version of a gradient method. Note that for nonsmooth problems the norm of the subgradient, $\|g(\boldsymbol{x})\|$, is not very informative. Therefore in the subgradient scheme we use a normalized direction $g(\bar{\boldsymbol{x}}) / \|g(\bar{\boldsymbol{x}})\|$.

Scheme

Subgradient method. Unconstrained minimization

0. Choose $\boldsymbol{x}_0 \in Q$ and a sequence $\{h_k\}_{k=0}^{\infty}$:

$$h_k > 0, h_k \to 0, \sum_{k=0}^{\infty} h_k = \infty \tag{8}$$

1. The k-th iteration ($k \geq 0$). Compute $f(\boldsymbol{x})$, $g(\boldsymbol{x}_k)$ and set

$$\boldsymbol{x}_{k+1} = \pi_Q \left(\boldsymbol{x}_k - h_k \frac{g(\boldsymbol{x}_k)}{\|g(\boldsymbol{x}_k)\|} \right).$$

Main Theorem

Let us estimate the rate of convergence of this scheme.

Theorem 5
Let f be Lipschitz continuous on $B_2(\boldsymbol{x}^*, R)$ with constant M and $\boldsymbol{x}_0 \in B(\boldsymbol{x}^*, R)$. Then

$$f_k^* - f^* \leq M \frac{R^2 + \sum_{i=0}^k h_i^2}{2 \sum_{i=0}^k h_i}. \tag{9}$$

Proof. Denote $r_i = \|\boldsymbol{x}_i - \boldsymbol{x}^*\|$. Then, in view of Lemma 3.1.5, we have

$$\begin{aligned} r_{i+1}^2 &= \left\| \pi_Q \left(\boldsymbol{x}_i - h_i \frac{g(\boldsymbol{x}_i)}{\|g(\boldsymbol{x}_i)\|} \right) - \boldsymbol{x}^* \right\|^2 \\ &\leq \left\| \boldsymbol{x}_i - h_i \frac{g(\boldsymbol{x}_i)}{\|g(\boldsymbol{x}_i)\|} - \boldsymbol{x}^* \right\|^2 = r_i^2 - 2h_i v_i + h_i^2. \end{aligned}$$

Main Theorem

Proof. (Continued)

Summing up these inequalities for $i = 0 \dots k$, we arrive at

$$r_0^2 + \sum_{i=0}^{k} h_i^2 = 2 \sum_{i=0}^{k} h_i v_i + r_{k+1}^2 \geq 2 v_k^* \sum_{i=0}^{k} h_i.$$

Thus,

$$v_k^* \leq \frac{R^2 + \sum_{i=0}^{k} h_i^2}{2 \sum_{i=0}^{k} h_i}.$$

It remains to use Lemma 3.2.2. □

Remark. $f_k^* - f^* \leq \omega_f(x^*, v_k^*) = f(x_m) - f(x^*)$, where the distance between $f(x_m)$ and f^* is biggest, and we also have $\|x_m - x^*\| \leq v_k^*$. In view of the local Lipschitz continuity, we arrive at the result.

Main Theorem

Thus Theorem 5 demonstrate the convergence rate of subgradient method(8) depends on the values

$$\Delta_k = \frac{R^2 + \sum_{i=0}^{k} h_i^2}{2 \sum_{i=0}^{k} h_i}.$$

We can easily see that $\Delta_k \to 0$ if the series $\sum_{i=0}^{\infty} h_i$ diverges. However, Iet us try to choose h_k in an optimal way.

Main Theorem

Let us assume that we have to perform a fixed number of steps of the subgradient method, say, N.

Then, minimizing Δ_k as a function of $\{h_k\}_{k=0}^N$, we find that the optimal strategy is as follows.

$$h_i = \frac{R}{\sqrt{N+1}}, \quad i = 0 \dots N. \tag{10}$$

(From Example 3.1.2, we can see that Δ_k is a convex function on $\{h_i\}$.)

In this case $\Delta_N = \frac{R}{\sqrt{N+1}}$ and we obtain the following rate of convergence:

$$f_k^* - f^* \leq \frac{MR}{\sqrt{N+1}}.$$

Conclusion

Comparing this result with the Iower bound of Theorem 3.2.1, we conclude:

> The subgradient method (8), (10) is optimal for problem (7) uniformly in the dimension n.

I f we do not want to fix the number of iterations apriori, we can choose

$$h_i = \frac{r}{\sqrt{i+1}}, i = 0, \ldots .$$

Then it is easy to see that Δ_k is proportional to

$$\frac{R^2 + r \ln(k+1)}{2r\sqrt{k+1}},$$

and we can classify the rate of convergence of this scheme as sub-optimal.

Conclusion

Thus, the simplest method for solving the problem (3.2.3) appears to be optimal. This indicates that the problems from our class are too complicated to be solved efficiently. However, we should remember, that our conclusion is valid uniformly in the dimension of the problem.

We will see that a moderate dimension of the problem, taken into account in a proper way, helps to develop much more efficient schemes.

Part III
Condition Descent or Franke Wolfe

The Frank-Wolfe Algorithm

The Frank-Wolfe (FW) or conditional gradient algorithm is one of the oldest methods for nonlinear constrained optimization and has seen an impressive revival in recent years due to its low memory requirement and projection-free iterations. In its classical form, the Frank-Wolfe algorithm can solve problems of the form

$$\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x}) \tag{11}$$

where f is differentiable with L-Lipschitz gradient and the domain $\mathcal{D}$ is a convex and compact set.

Originally published as *Frank, Marguerite, and Philip Wolfe. "An algorithm for quadratic programming." Naval Research Logistics (1956).*

The Frank-Wolfe Algorithm

Frank-Wolfe is a remarkably simple algorithm that given an initial guess $\boldsymbol{x}_0$ construct a sequence of estimates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ that converges towards a solution of optimization problem. The algorithm is define as follows:

Input: intial guess $\boldsymbol{x}_0$, tolerance $\delta > 0$
For $k = 0, \ldots$ do
    $\boldsymbol{s}_k \in \text{argmax}_{\boldsymbol{s} \in \mathcal{D}} \langle -\nabla f(\boldsymbol{x}_k),\ \boldsymbol{s} \rangle$
    $\boldsymbol{d}_k = \boldsymbol{s}_k - \boldsymbol{x}_k$
    $\boldsymbol{g}_k = -\langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{d}_k \rangle$
    if $\boldsymbol{g}_k < \delta$:
        return $\boldsymbol{x}_k$
    update $\gamma_k(\boldsymbol{g}_k)$
    $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \gamma_k \boldsymbol{d}_k$
End Loop.

The Frank-Wolfe Algorithm

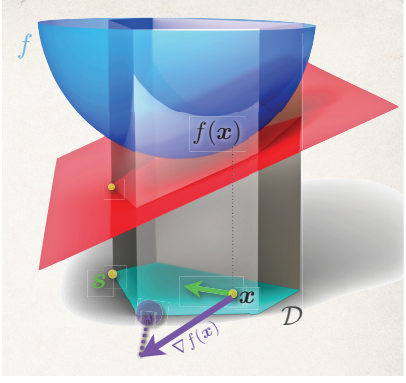
Remark for Frank-Wolfe:

1. The Linear Minimization Oracle:

$$\mathrm{LMO}_{\mathcal{D}}(\boldsymbol{d}) = \operatorname*{argmin}_{\boldsymbol{s}\in\mathcal{D}} \langle \boldsymbol{d},\ \boldsymbol{s}\rangle$$

2. Linearization:

$$\begin{aligned}\boldsymbol{s}_k &= \operatorname*{argmin}_{\boldsymbol{s}\in\mathcal{D}} f(\boldsymbol{x}_k) \\ &\quad + \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{s}-\boldsymbol{x}_k\rangle \\ &= \operatorname*{argmin}_{\boldsymbol{s}\in\mathcal{D}} \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{s}\rangle \\ &= \mathrm{LMO}_{\mathcal{D}}(\nabla f(\boldsymbol{x}_k)).\end{aligned}$$



The Frank-Wolfe Algorithm

There are many methods to update the γ_k. We list two popular variants of them as follows.

1. Variant 1: set step size as

$$\gamma_k = \min\left\{\frac{\boldsymbol{g}_k}{L\left\|\boldsymbol{d}_k\right\|^2}, 1\right\}. \tag{12}$$

2. Variant 2: set step size by line search

$$\gamma_k = \underset{\gamma\in[0,1]}{\operatorname{argmin}} f(\boldsymbol{x}_k + \gamma \boldsymbol{d}_k). \tag{13}$$

Contrary to projected gradient descent methods, the Frank-Wolfe does not require access to a projection (a projection-free algorithm). It instead relies on a routine that solves a linear problem over the domain (commonly referred to as a linear minimization oracle).

The Frank-Wolfe Algorithm

We condense the algorithm as two basics steps:

$$\textbf{Step 1:}\ \ \boldsymbol{s}_k \in \text{Argmin}_{\boldsymbol{s}\in\mathcal{D}}\langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{s}\rangle; \tag{14}$$

$$\textbf{Step 2:}\ \ \boldsymbol{x}_{k+1} = (1-\gamma_k)\boldsymbol{x}_t + \gamma_k \boldsymbol{s}_k. \tag{15}$$

Remark: What happens when $\mathcal{D} = \{\|\boldsymbol{x}\| \le \tau\}$ for norm $\|\cdot\|$? Then

$$\begin{aligned}\boldsymbol{s}_k \in \text{Argmin}_{\|\boldsymbol{s}\|\le\tau}\langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{s}\rangle &= -\tau\left(\text{Argmax}_{\|\boldsymbol{s}\|\le 1}\langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{s}\rangle\right)\\ &= -\tau\partial\left\|\nabla f(\boldsymbol{x}_k)\right\|_*\end{aligned}$$

where $\|\cdot\|_*$ is the corresponding dual norm. In other words, if we know how to compute subgradients of the dual norm, then we can easily perform Frank-Wolfe steps. More detailed examples are listed as follows.

The Frank-Wolfe Algorithm

Dual norm: Let $\|\cdot\|$ be a norm in $\mathbb{R}$, The associated dual norm is defined as

$$\|\boldsymbol{z}\|_* = \sup_{\boldsymbol{x}}\{\boldsymbol{z}^\top \boldsymbol{x}| \, \|\boldsymbol{x}\| \leq 1\}.$$

That is, given a linear function(al) $f_{\boldsymbol{z}}(\boldsymbol{x}) = \boldsymbol{z}^\top \boldsymbol{x}$, how big is the number $f_{\boldsymbol{z}}(\boldsymbol{x})$ relative to the size (norm) of $\boldsymbol{x}$. This is exactly the number $\boldsymbol{z}^\top \boldsymbol{x}/\|\boldsymbol{x}\|$. Then the largest quantity can possibly be:

$$\|\boldsymbol{z}\|_* = \sup_{\boldsymbol{x}\neq 0} \frac{\boldsymbol{z}^\top \boldsymbol{x}}{\|\boldsymbol{x}\|}$$

Subdifferential of a norm: $\partial \|\boldsymbol{z}\|_* \equiv \{x : \langle \boldsymbol{x}, \, \boldsymbol{z}\rangle = \|\boldsymbol{z}\|_*, \|\boldsymbol{x}\| \leq 1\}$, thus $\boldsymbol{x}^* \in \{\boldsymbol{x} : \langle \boldsymbol{x}, \, \boldsymbol{z}\rangle = \|\boldsymbol{z}\|_*, \|\boldsymbol{x}\| \leq 1\}$, that is $\boldsymbol{x}^* \in \partial \|\boldsymbol{z}\|_*$.

The ℓ_1-regularized problem

Example 1 (ℓ_1-regularization): For the ℓ_1-regularized problem

$$\min_x f(\boldsymbol{x}) \quad \text{subject to } \|\boldsymbol{x}\|_1 \le t,$$

we have $\boldsymbol{s}_k \in -t\partial \|\nabla f(\boldsymbol{x}_k)\|_\infty$. Frank-Wolfe update is thus

$$i_k \in \operatorname*{argmax}_{i=1,\dots,p} \left|\nabla^{(i)} f(\boldsymbol{x}_k)\right|$$
$$x_{k+1} = (1-\gamma_k)\boldsymbol{x}_k - \gamma_k\, \tau\, \text{sign}\left(\nabla^{(i_k)} f(\boldsymbol{x}_k)\right) e_{i_k}.$$

Just like the greedy coordinate descent.

Note: this is a lot simpler than projection onto the ℓ_1 ball, though both require $O(n)$ operations. (actually, the original projection need $n \log n$ due to the sorting procedure)

The ℓ_1-regularized problem

Remark: Since

$$\partial \|\boldsymbol{z}\|_\infty \equiv \{\boldsymbol{x} : \langle \boldsymbol{x},\ \boldsymbol{z} \rangle = \|\boldsymbol{z}\|_\infty, \|\boldsymbol{x}\| \leq 1\},$$

It is clear that $\boldsymbol{x} = e_j$ or $-e_j$ where $|\boldsymbol{z}_j|$ is the largest.

The ℓ_p-regularized problem

Example 1 (ℓ_p-regularization): For the ℓ_p-regularized problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{subject to } \|\boldsymbol{x}\|_p \leq t, \quad 1 \leq p \leq \text{inf},$$

we have $\boldsymbol{s}_k \in -t\partial \|\nabla f(\boldsymbol{x}_k)\|_q$, where p, q are dual, i.e. $1/p + 1/q = 1$. Frank-Wolfe update can thus be

$$\begin{aligned} \boldsymbol{s}_k^{(i)} &= -\alpha \text{ sign}\left(\nabla^{(i)} f(\boldsymbol{x}_k)\right) \left|\nabla^{(i)} f(\boldsymbol{x}_k)\right|^{p/q}, \\ i &= 1, \ldots, n, \end{aligned}$$

where α is a constant such that $\|\boldsymbol{s}_k\|_p = t$, and then Frank-Wolfe update are as usual.

Note: this is a lot simpler projection onto the ℓ_p ball, for general p! Aside from special cases ($p = 1, 2, \infty$), these projections cannot be directly computed (must be treated as an optimization).

The ℓ_p-regularized problem

Remark: 直接根据 $\|\nabla f(\boldsymbol{x}_k)\|_q$ 定义，求微分。

The trace-regularized problem

Example 1 (Trace-regularization): For the trace-regularized problem

$$\min_X f(X) \quad \text{subject to } \|X\|_{tr} \leq t,$$

we have $S_k \in -t\partial \|\nabla f(X_k)\|_{op}$. Frank-Wolfe update can thus be

$$S_k = -t\, \mu\nu^\top$$

where μ, ν are leading left, right singular vectors of $\nabla f(X_k)$, and then Frank-Wolfe updates are as usual.

Note: this is a lot simpler and more efficient than projection onto the trace norm ball, which requires a singular value decomposition.

The trace-regularized problem

Remark:
(1) 矩阵内积：

$$\langle X,\ Y\rangle = \text{trace}(X^\top Y) = \text{trace}(Y^\top X),\ \text{where } X, Y \in \mathbb{R}^{n\times m}.$$

(2)$\|X\|_{tr}$ 的对偶是 $\|X\|_{op}$。前者是 X 奇异值的和，类似于 ℓ_1；后者类似于 ℓ_∞，是最大奇异值 σ_1
(3) 测试 $\langle Z,\ u_1v_1^\top\rangle = \text{trace}(Z^\top u_1v_1^\top)$。我们有

$$\begin{aligned}
\text{trace}(Z^\top u_1v_1^\top) &= \text{trace}((\sum_i \sigma_i u_i v_i^\top)^\top u_1v_1^\top) \\
&= \text{trace}((\sum_i \sigma_i v_i u_i^\top) u_1v_1^\top) \\
&= \text{trace}(v_1^\top(\sum_i \sigma_i v_i u_i^\top)u_1)\ (\text{since trace}(ABC) = \text{trace}(CAB)) \\
&= \sigma_1 = \|Z\|_{op}
\end{aligned}$$

Example: compared to proximal form

Recall that solution of the constrained problem

$$\min_x f(\boldsymbol{x}) \quad \text{subject to } \|\boldsymbol{x}\| \leq t,$$

are often equivalent to those of the Lagrange problem

$$\min_x f(\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|$$

as we let the tuning parameters t and λ vary over $[0, \infty]$. Typically in statistics and ML problems, we would just solve whichever form is easiest, over wide range of parameter values.

Example: compared to proximal form

We should also compare the Frank-Wolfe update under $\|\cdot\|$ to the proximal operator of $\|\cdot\|$.

1. ℓ_1 norm: Frank-Wolfe update scans for maximum of gradient direction; proximal operator soft-thresholds the gradient step; both use $O(n)$ flops.
2. ℓ_p norm: Frank-Wolfe update computes raise each entry of gradient to power and sums, in $O(n)$ flops; proximal operator is not generally directly computable.
3. Trace norm: Frank-Wolfe update computes top left and right singular vector of gradient; proximal operator soft-thresholds the gradient step, requiring a singular value decomposition.

Many other regularizers yield efficient Frank-Wolfe updtes, e.g., special polyhedra or cone constraints, sum-of-norms (group-based) regularization, atomic norms. See [Jaggi, 2011, 2013].

Convergence Theory

Before diving into the convergence analysis, we list two key definition and technical Lemma.

Definition 6 (Stationary Point)

We will say that $\boldsymbol{x}^* \in \mathcal{D}$ is a stationary point if

$$\langle \nabla f(\boldsymbol{x}^*),\ \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0 \quad \text{for all } \boldsymbol{x} \in \mathcal{D}.$$

Definition 7 (Frank-Wolfe Gap)

We denote by g_k the Frank-Wolfe gap, defined as

$$g_k \equiv \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_k - \boldsymbol{s}_k \rangle \tag{16}$$

Convergence Theory

Note that by the definition of $\boldsymbol{s}_k$ in (14) we always have $\langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{s}_k \rangle \leq \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_k \rangle$ and so the Frank-Wolfe gap is always non-negative, and zero only at a stationary point.

When f is convex we also have that the FW gap verifies

$$g_k = \max_{\boldsymbol{s}} \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_k - \boldsymbol{s} \rangle \tag{17}$$

$$\geq \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_k - \boldsymbol{x}^* \rangle \tag{18}$$

$$\geq f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \tag{19}$$

where the last inequality follows from the definition of convexity and so can be used as a function suboptimality certificate.

Convergence Theory

The next lemma relates the objective function value at two consecutive iterates and will be key to prove convergence results, both for convex and non-convex objectives.

Lemma 8
Let $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots\}$ be the iterates produced by the Frank Wolfe algorithm (in either variants). Then we have the following inequality, valid for any $\xi \in [0, 1]$:

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{1}{2}\xi^2 L \operatorname{diam}(\mathcal{D})^2. \tag{20}$$

Convergence Theory

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{1}{2}\xi^2 L \operatorname{diam}(\mathcal{D})^2.$$

Proof. A consequence of the Lipschitz gradient assumption on f is that we can upper bound the function f at every point $y \in \mathcal{D}$ by the following quadratic:

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{21}$$

We can apply this inequality, valid for any $\boldsymbol{x}$, $\boldsymbol{y}$ in the domain, to the special case $\boldsymbol{x} = \boldsymbol{x}_k$, $\boldsymbol{y} = (1-\gamma)\boldsymbol{x}_k + \gamma \boldsymbol{s}_k$ with $\gamma \in [0,1]$ so that $\boldsymbol{y}$ remains in the domain, and so we have

$$f((1-\gamma)\boldsymbol{x}_k + \gamma \boldsymbol{s}_k) \leq f(\boldsymbol{x}_k) + \gamma \overbrace{\langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{s}_k - \boldsymbol{x}_k \rangle}^{-g_k} + \frac{L\gamma^2}{2} \|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2. \tag{22}$$

Convergence Theory

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{1}{2}\xi^2 L \operatorname{diam}(\mathcal{D})^2.$$

Proof. (continued) We will now minimize the right hand side with respect to $\gamma \in [0, 1]$. This is a quadratic function of γ and its minimum, which we denote γ_k^* is given by

$$\gamma_k^* = \min\left\{\frac{g_k}{L\left\|\boldsymbol{x}_k - \boldsymbol{s}_k\right\|^2}, 1\right\}. \tag{23}$$

We now use the value $\gamma = \gamma_k^*$ in the inequality (22) to get the following sequence of inequalities:

Convergence Theory

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{1}{2}\xi^2 L \operatorname{diam}(\mathcal{D})^2.$$

Proof. (continued)

$$f((1-\gamma_k^*)\boldsymbol{x}_k + \gamma_k^* \boldsymbol{s}_k) \leq f(\boldsymbol{x}_k) - \gamma_k^* g_k + \frac{L(\gamma_k^*)^2}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 \tag{24}$$

$$= f(\boldsymbol{x}_k) + \min_{\xi \in [0,1]} \left\{ -\xi g_k + \frac{L\xi^2}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 \right\} \tag{25}$$

(by optimality of γ_k^*)

$$\leq f(\boldsymbol{x}_k) - \xi g_k + \frac{L\xi^2}{2}\|\boldsymbol{s}_k - \boldsymbol{x}_k\|^2 \tag{26}$$

(for any $\xi \in [0,1]$) (27)

Convergence Theory

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{1}{2}\xi^2 L \operatorname{diam}(\mathcal{D})^2.$$

Proof. (continued) Finally, we obtain

$$f((1-\gamma_k^*)\boldsymbol{x}_k + \gamma_k^* \boldsymbol{s}_k) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{L\xi^2}{2}\operatorname{diam}(\mathcal{D})^2. \tag{28}$$

The right hand side of the above inequality already contains the terms claimed in the Lemma. We will now bound the right hand side.

Convergence Theory

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \xi g_k + \frac{1}{2}\xi^2 L \operatorname{diam}(\mathcal{D})^2.$$

Proof. (continued) For Variant 1 of the algorithm we have $f(\boldsymbol{x}_{k+1}) = f((1-\gamma_k^*)\boldsymbol{x}_k + \gamma_k^* \boldsymbol{s}_k)$, since γ_k and γ_k^* coincide in this case. For Variant 2 we have $f(\boldsymbol{x}_{k+1}) \leq f((1-\gamma_k^*)\boldsymbol{x}_k + \gamma_k^* \boldsymbol{s}_k)$ since by definition of line search $f(\boldsymbol{x}_{k+1})$ is the point that minimizes the objective value in the segment $(1-\gamma)\boldsymbol{x}_k + \gamma \boldsymbol{s}_k$. Hence, in either case we have

$$f(\boldsymbol{x}_{k+1}) \leq f((1-\gamma_k^*)\boldsymbol{x}_k + \gamma_k^* \boldsymbol{s}_k).$$

Changing this last inequality with Eq. (28) yields the claimed inequality. □

Convergence Theory

The following is our first convergence rate result and is valid for objectives with L-Lipschitz gradient but not necessarily convex. This was first proven by Simon Lacoste-Julien [Lacoste-Julien, 2016].

Theorem 9
If f is differentiable with L-Lipschitz gradient, then we have the following $O(1/\sqrt{k})$ bound on the best Frank Wolfe gap:

$$\min_{0 \leq i \leq k} g_i \leq \frac{\max\{2h_0, L \operatorname{diam}(\mathcal{D})^2\}}{\sqrt{k+1}}, \tag{29}$$

where $h_0 = f(\boldsymbol{x}_0) - \min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$ is the initial global suboptimality.

Convergence Theory

Proof. By Lemma 8 we have the following sequence of inequalities, valid for any $\xi \in [0, 1]$:

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - \xi g_k + \frac{\xi^2 L}{2} \text{diam}(\mathcal{D})^2 \\ &= f(\boldsymbol{x}_k) - \xi g_k + \frac{\xi^2 C}{2}, \end{aligned}$$

with $C = L \text{ diam}(\mathcal{D})^2$. We consider the value of ξ that minimizes the right hand size and we obtain $\xi^* = \min\{g_k/C, 1\}$.

Convergence Theory

Proof. (continued.) We will now make a distinction of cases based on the value of ξ^*:

1. If $g_k \leq C$, then $\xi^* = g_k/C$ and using this value in the previous inequality we obtain the bound

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{g_k^2}{2C} \tag{30}$$

2. If $g_k > C$, then $\xi^* = 1$ and we have the following sequence of inequalities

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - g_k + \frac{C}{2} \tag{31}$$

$$< f(\boldsymbol{x}_k) - \frac{g_k}{2} \text{ (since } C < g_k) \tag{32}$$

Convergence Theory

Proof. (continued.) combining both cases we have

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{g_k}{2} \min\left\{\frac{g_k}{C}, 1\right\}. \tag{33}$$

Adding the previous inequality from iterate 0 to t and rearranging we have

$$-h_0 \leq f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_0) \leq -\sum_{i=0}^{k} \frac{g_i}{2} \min\left\{\frac{g_i}{C}, 1\right\} \tag{34}$$

$$\leq -(k+1)\frac{g_k^*}{2}\left\{\frac{g_k^*}{C}, 1\right\}, \tag{35}$$

where $g_k^* = \min_{0 \leq i \leq t} g_i$.

Convergence Theory

Proof. (continued.) Again, we make a distinction of cases, this time on g_k^*:

1. If $g_k^* \leq C$, then $\min\{g_k^*/C, 1\} = g_k^*/C$ and solving for g_k^* in the previous inequality we have

$$g_k^* \leq \sqrt{\frac{2Ch_0}{k+1}} \leq \frac{2h_0 + C}{2\sqrt{k+1}} \tag{36}$$

$$\leq \frac{\max\{2h_0, C\}}{\sqrt{k+1}} \tag{37}$$

where in the second inequality we have used Young's inequality $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$ with $a = \sqrt{2h_0}$, $b = \sqrt{C}$.

Convergence Theory

Proof. (continued.)

2. if $g_k^* > C$, then $\min\{g_k^*/C, 1\} = 1$ and rearranging(35) we have the following inequality with the stronger $O(1/\sqrt{k})$ bound:

$$g_k^* \leq \frac{2h_0}{k+1} \leq \frac{2h_0}{\sqrt{k+1}} \leq \frac{\max\{2h_0, C\}}{\sqrt{k+1}} \tag{38}$$

Hence, in both case we have

$$g_k^* \leq \frac{\max\{2h_0, C\}}{\sqrt{k+1}}, \tag{39}$$

and the claimed bound follows from the definition of g_k^*.

Convergence Theory

Theorem 10
If f is convex and differentiable with L-Lipschitz gradient, then we have the following convergence rate for the function suboptimality:

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{2L \operatorname{diam}(\mathcal{D})^2}{k+1}. \tag{40}$$

Proof. Because of convexity we can obtain a tighter bound using the following simple inequality, mentioned earlier (see (17)):

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq \langle \nabla f(\boldsymbol{x}),\ \boldsymbol{x} - \boldsymbol{x}^* \rangle. \tag{41}$$

Let $e_k = A_k(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*))$ for a positive A_k that will fix later and $C = L \operatorname{diam}(\mathcal{D})^2$.

Convergence Theory

Proof. (continued.) Then we have the following sequence of inequalities.

$$
\begin{aligned}
e_{k+1} - e_k &= A_{k+1}(f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*)) - A_k(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)) && (42)\\
&\le A_{k+1}(f(\boldsymbol{x}_k) - \xi g_k + \frac{\xi^2 C}{2} - f(x^*)) \\
&\quad - A_k(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)) && (43)\\
&\quad \text{(by Lemma 8, for any } \xi \in [0,1]) \\
&\le A_{k+1}\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) - \xi\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)\right) + \frac{\xi^2 C}{2}\right) \\
&\quad - A_k(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)) && (44)\\
&\quad \text{(by convexity (41))} \\
&= \left((1-\xi)A_{k+1} - A_k\right)\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)\right) + A_{k+1}\frac{\xi^2 C}{2}. && (45)
\end{aligned}
$$

Convergence Theory

Proof. (continued.) Now, choosing $A_k = \frac{k(k+1)}{2}$, $\xi = 2/(k+2)$ we have

$$(1-\xi)A_{k+1} - A_k = \frac{k(k+1)}{2} - \frac{k(k+1)}{2} = 0 \tag{46}$$

$$A_{k+1}\frac{\xi^2}{2} = \frac{k+1}{k+2} \leq 1, \tag{47}$$

and so replacing with these values of A_k and ξ in Eq. (45) gives

$$e_{k+1} - e_k \leq C. \tag{48}$$

Convergence Theory

Proof. (continued.) Adding this inequality from 0 to $k-1$ and using $e_0 = 0$ we have for $k > 0$:

$$e_k \leq kC \quad \Rightarrow f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{2C}{k+1}, \tag{49}$$

where in the last implication we have divided by $k(k+1)$ and so we need $k > 0$. The claimed bound now follows from definition of C. □

References I

Martin Jaggi. Convex optimization without projection steps. *arXiv preprint arXiv:1108.1170*, 2011.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Thank You!

Email:qianhui@zju.edu.cn