Subgradient and Subgradient methods

Report for 'Teoria e Metodi dell'Ottimizzazione' course

a.y. 2023/24

Sebastiano Scardera

Contents

1 Subgradient

1.1 Definition

We say that $g \in \mathbb{R}^n$ is a *subgradient* of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ if:

$$\forall z \in \mathbb{R}^n, \quad f(z) \geq f(x) + g^T(z - x).$$

It is not necessary for f to be differentiable at x. In fact, as we can see in Figure 1, there can exist more than one subgradient at a point of non-differentiability. A function f is said to be *subdifferentiable* if it is subdifferentiable at every point. We denote the set of subgradients of f at x as $\partial f(x)$ and call it the *subdifferential*.

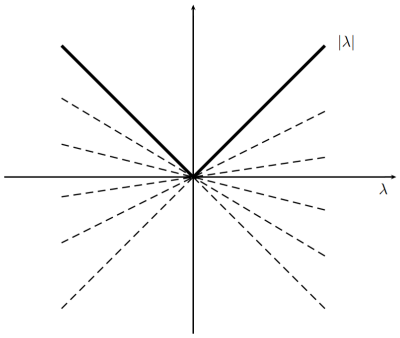

Figure 1: The function $f(\lambda) = |\lambda|$ is not differentiable at 0 but has infinitely many subgradients.

1.2 Existence of subgradients for convex functions

Theorem 1. If $f : \text{dom} f \to \mathbb{R}$ is a convex function and $x \in \text{int}(\text{dom} f)$, then $\partial f(x) \neq \emptyset$.

Before proceeding with the proof, we state a classic lemma of separation. The proof is omitted and can be found in [3] (Lemma 1.3).

Lemma (Supporting hyperplane theorem)**.** Let $C \subseteq \mathbb{R}^n$ be a convex set, $C \neq \emptyset$.

If $y \notin \text{int}(C) \Rightarrow \exists z^* \in \mathbb{R}^n, z^* \neq 0$ such that $< z^*, x > \leq < z^*, y > \quad \forall x \in C,$

where $< \cdot, \cdot >$ denotes the Euclidean inner product.

Proof (Theorem 1)**.** Without loss of generality, we can assume $\text{dom} f = \mathbb{R}^n$, the proof can be easily adapted to any domain with non-empty interior.

Let f be a convex function and $x \in \mathbb{R}^n$, we will show that $\partial f(x) \neq \emptyset$. We apply the above lemma to the convex set:

$$\mathbf{epi}(f) = \{(z,t) \in \mathbb{R}^n \times \mathbb{R} \mid t \geq f(z)\}.$$

In particular, we obtain that there exist $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, both not zero, such that:

$$\begin{bmatrix} a \\ b \end{bmatrix}^T \left(\begin{bmatrix} z \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) = a^T(z-x) + b(t - f(x)) \leq 0, \quad \forall (z,t) \in \mathbf{epi}(f).$$

If $b \neq 0$, dividing by b and taking $t = f(z)$, we obtain:

$$f(z) \geq f(x) - \frac{1}{b} a^T (z - x),$$

which shows that $-(a/b)^T \in \partial f(x)$.
Now we show that $b \neq 0$. If we assume $b = 0$ for contradiction, then:
$a^T(z-x) \leq 0 \quad \forall z \in \mathbb{R}^n$. Taking $z = a + x$ would imply $a = 0$ and $b = 0$, which is a contradiction.

1.3 Properties

1) $\partial f(x)$ is always a closed convex set, as it is the arbitrary intersection of half-spaces:

$$\partial f(x) = \bigcap_{z \in \mathrm{dom} f} \{g \mid f(z) \geq f(x) + g^T(z-x)\};$$

2) If f is continuous at x, then $\partial f(x)$ is bounded. Indeed,
let $\epsilon > 0$ be small enough. By continuity of f at x, there exists a constant $C > 0$ such that for all $y \in \mathbb{R}^n$ with $\|y - x\|_2 \leq \epsilon$, we have $\|f(y)\|_2 \leq C$.
If $\partial f(x)$ is unbounded, then there exists a sequence $g_n \in \partial f(x)$ such that $\|g_n\|_2 \to \infty$.

   Consider $y_n = x + \epsilon(g_n / \|g_n\|_2)$, by the definition of subgradient:
   $f(y) \geq f(x) + g_n^T(y_b - x) = f(x) + \epsilon \|g_n\|_2 \to \infty$, which is a contradiction because $f(y_n)$ is bounded.

1.4 Characterization of minima

Lemma. x^* is a minimum of a function f if and only if f is subdifferentiable at x^* and $0 \in \partial f(x^*)$.

Proof. ($\Rightarrow$) Follows from $f(x) \geq f(x^*)\ \forall x \in \mathrm{dom} f$;
($\Leftarrow$) If $0 \in \partial f(x^*)$ then $f(x) \geq f(x^*) + 0^T(x - x^*) = f(x^*)\ \forall x \in \mathrm{dom} f$.

2 Subgradient methods

Subgradient methods are very simple algorithms for minimizing convex functions, not necessarily differentiable. The structure is very similar to that of the gradient method with two main differences:

- It is not necessary for the function to be minimized to be differentiable;
- The directions given by the subgradient are not necessarily descent directions, as can be seen in Figure 1.

Like the gradient method, the subgradient method can also be adapted to the constrained case. In the unconstrained case, it is generally slower than interior point algorithms or the Newton method. However, it can be very advantageous when applied to large-scale problems as it does not require a large amount of memory. In general, it can be used in combination with primal or dual decomposition techniques.
Next, we will discuss the constrained and unconstrained cases. We will present the structure of different algorithms and some convergence results.
For the unconstrained case, we will explore the basic subgradient method by varying different step sizes and the Polyak method.
For the constrained case, we will examine the projected subgradient method and a subgradient method in a convex problem scenario.

2.1 Subgradient Method

The subgradient method is used to minimize $f : \mathbb{R}^n \to \mathbb{R}$, where f is a convex function. The algorithm has the following structure:

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)},$$

where $x^{(k)}$ is the k-th iterate, $g^{(k)}$ is any subgradient of f at $x^{(k)}$, and $\alpha_k > 0$ is the step size at iteration k. As mentioned before, $g^{(k)}$ may not be a descent direction, so we need to keep track of the best point found so far. For this, we define:

$$f_{best}^{(k)} = \min\{f_{best}^{(k-1)}, f(x^{(k)})\} \qquad \text{and} \qquad x_{best}^{(k)} \text{ s.t. } f(x_{best}^{(k)}) = f_{best}^{(k)}.$$

Hence:

$$f_{best}^{(k)} = \min\{f(x^{(1)}), ..., f(x^{(k)})\}$$

and $f_{best}^{(k)}$ is a decreasing sequence that converges to a limit in $[-\infty, +\infty)$.

2.1.1 Step Size Selection

The behavior of the algorithm can vary greatly depending on the choice of step size. Here are some examples:

- *Constant step size.* $\alpha_k = \alpha$ is a positive constant, independent of k.
- *Constant step length.* $\alpha_k = \gamma / \|g^{(k)}\|_2$, where $\gamma > 0$, i.e.,

$$\|x^{(k+1)} - x^{(k)}\|_2 = \gamma.$$

- *Square summable but not summable.* The step sizes satisfy:

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty.$$

- *Infinitesimal but not summable.* The step sizes are of the form $\alpha_k = \gamma_k / \|g^{(k)}\|_2$, where:

$$\gamma_k \geq 0, \qquad \lim_{k \to \infty} \gamma_k = 0, \qquad \sum_{k=1}^{\infty} \gamma_k = \infty.$$

2.1.2 Convergence Results

Definition (ϵ-suboptimal)**.** Let $f : \mathrm{dom} f \to \mathbb{R}$ and f^* be the minimum of f. We say that $x^* \in \mathrm{dom} f$ is **ϵ-suboptimal** if $f(x^*) - f^* \leq \epsilon$.

We will prove the convergence results under the following **assumptions**:

a) There exists x^*, a minimum point of f, such that $f(x^*) = f^*$.

b) $\partial f(x)$ is a uniformly (in x) bounded set, i.e., there exists a constant $G > 0$ such that for all $x \in \mathrm{dom} f$ and all $g \in \partial f(x)$: $\|g\|_2 \leq G$.

c) We know $R > 0$ such that $\|x^{(1)} - x^*\|_2 \leq R$.

As we have seen in Section 1.3, if f is continuous, then $\partial f(x)$ is pointwise bounded. Assumption b) can be relaxed by requiring uniform boundedness only on the subgradients of the iterates. In general, this assumption is not necessary. For example, the subgradient method with infinitesimal but not summable step sizes works even without this assumption.

The following convergence theorem holds:

Theorem 2 (Convergence)**.** Under the above assumptions, different convergence results are obtained depending on the choice of step size for the subgradient method.

For *constant step size and constant step length*, the algorithm converges in an ϵ-suboptimal manner, i.e.:

$$\lim_{k\to\infty} f_{best}^{(k)} - f^* < \epsilon,$$

for some $\epsilon > 0$, which decreases with the step size parameter.

By choosing the *square summable but not summable or infinitesimal but not summable* step sizes, we can guarantee convergence to the optimal value:

$$\lim_{k\to\infty} f(x_{best}^{(k)}) = f^*.$$

Before proceeding with the proof, we present a supporting lemma.

Lemma. Under the previous assumptions, for any choice of step sizes α and any iterate k of the algorithm, the following inequality holds:

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2 \sum_{i=1}^{k} \alpha_i}. \tag{1}$$

Proof (Lemma)**.** Let x^* be a minimum point of f, we have:

$$\begin{aligned}
\|x^{(k+1)} - x^*\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^*\|_2^2 \\
&= \|x^{(k)} - x^*\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^*) + \alpha_k^2 \|g^{(k)}\|_2^2 \\
&\leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) + \alpha_k^2 \|g^{(k)}\|_2^2,
\end{aligned}$$

where the first equality is by definition of $x^{(k+1)}$, and the last inequality is by definition of subgradient. Applying the inequality recursively, we obtain:

$$\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(1)} - x^*\|_2^2 - 2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^*) + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2.$$

Using that $0 \leq \|x^{(k+1)} - x^*\|_2^2$ and $\|x^{(1)} - x^*\|_2 \leq R$, we have:

$$2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^*) \leq R^2 + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2. \tag{2}$$

Note that:

$$\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^*) \geq \left(\sum_{i=1}^{k} \alpha_i\right) \min_{i=1,\ldots,k} (f(x^{(i)}) - f^*) = \left(\sum_{i=1}^{k} \alpha_i\right)(f_{best}^{(k)} - f^*),$$

which implies:

$$f_{best}^{(k)} - f^* = \min_{i=1,\dots,k} (f(x^{(i)}) - f^*) \leq \frac{R^2 + \sum_{i=1}^k \alpha_i^2 \|g^{(i)}\|_2^2}{2\sum_{i=1}^k \alpha_i} \tag{3}$$

and using $\|g^{(i)}\|_2^2 \leq G^2$, we have:

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2\sum_{i=1}^k \alpha_i}.$$

□

Proof (Theorem 2)**.** Using the lemma, we prove the convergence results by distinguishing the type of step size.

Constant step size $\alpha_k = \alpha$:

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2\alpha^2 k}{2\alpha k}.$$

The right-hand side of the equation converges to $G^2\alpha/2$ as $k \to \infty$. In particular, $f_{best}^{(k)}$ converges in an $G^2\alpha/2$-suboptimal manner.

Constant step length: if $\alpha_k = \gamma/\|g^{(k)}\|_2$, using the inequality (3):

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + \gamma^2 k}{2\sum_{i=1}^k \alpha_i} \leq \frac{R^2 + \gamma^2 k}{2\gamma k/G},$$

where we used $\alpha_i \geq \gamma/G$. As before, the right-hand side converges to $G\gamma/2$ as $k \to \infty$, and the algorithm converges in an $G\gamma/2$-suboptimal manner.

Square summable but not summable: let

$$\|\alpha\|_2^2 = \sum_{i=1}^{\infty} \alpha_i^2 < \infty, \qquad\qquad \sum_{i=1}^{\infty} \alpha_i = \infty.$$

Then we have:

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2\|\alpha\|_2^2}{2\sum_{i=1}^k \alpha_i},$$

which converges to zero as the numerator is bounded and the denominator goes to infinity. This gives us convergence to the optimal value.

Infinitesimal but not summable step size: we see that the right-hand side of inequality (1) converges to zero.

Let $\epsilon > 0$, there exists an integer N_1 such that $\alpha_i \leq \epsilon/G^2$ for every $i > N_1$. Furthermore, there exists an integer N_2 such that:

$$\sum_{i=1}^{N_2} \alpha_i \geq \frac{1}{\epsilon}\left(R^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2\right),$$

since $\sum_{i=1}^{\infty} \alpha_i = \infty$. Let $N = \max\{N_1, N_2\}$, then for $k > N$, we have:

$$\begin{aligned}
\frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} &\leq \frac{R^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} + \frac{G^2 \sum_{i=N_1+1}^{k} \alpha_i^2}{2\sum_{i=1}^{N_1} \alpha_i + 2\sum_{i=N_1+1}^{k} \alpha_i} \\
&\leq \frac{R^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2}{(2/\epsilon)(R^2 + G^2 \sum_{i=1}^{N_1} \alpha_i^2)} + \frac{G^2 \sum_{i=N_1+1}^{k} (\epsilon\alpha_i/G^2)}{2\sum_{i=N_1+1}^{k} \alpha_i} \\
&= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

□

2.2 Polyak's Subgradient Method

Polyak's subgradient method is similar to the basic subgradient method, but it differs in the choice of step size. In particular, we will analyze the "exact" case, where we assume the optimal value of f is known, and compare it to the approximate case. We will work with the same assumptions and notations as in Section 2.1.2.

2.2.1 Exact Case with Known f^*

Assuming that the optimal value f^* is known, the step size is given by:

$$\alpha_k = \frac{f(x^{(k)}) - f^*}{\|g^{(k)}\|_2^2}.$$

One reason to choose this α_k is as follows. Suppose the "Taylor expansion" holds:

$$f(x^{(k)} - \alpha g^{(k)}) \approx f(x^{(k)}) + g^{(k)T}(x^{(k)} - \alpha g^{(k)} - x^{(k)}) = f(x^{(k)}) - \alpha g^{(k)T} g^{(k)},$$

where $g^{(k)}$ plays the same role as the gradient in the case of differentiable f. Substituting f^* on the left-hand side gives us α_k. Another reason, more analytical, comes from the following inequality:

$$\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k(f(x^{(k)}) - f^*) + \alpha_k^2 \|g^{(k)}\|_2^2,$$

and it can be observed that with the choice of Polyak's step, we minimize the right-hand term.

Theorem 3 (Convergence Theorem)**.** If $x^{(k)}$ are the iterates of the exact Polyak's method, then:

$$f(x^{(k)}) \to f^*, k \to \infty.$$

Proof. To obtain convergence results, we rely on the previous Lemma. In particular, by substituting the expression of α_k into the inequality (2), we have:

$$2\sum_{i=1}^{k} \frac{(f(x^{(i)}) - f^*)^2}{\|g^{(i)}\|_2^2} \leq R^2 + \sum_{i=1}^{k} \frac{(f(x^{(i)}) - f^*)^2}{\|g^{(i)}\|_2^2},$$

and therefore:

$$\sum_{i=1}^{k} \frac{(f(x^{(i)}) - f^*)^2}{\|g^{(i)}\|_2^2} \leq R^2 \iff \sum_{i=1}^{k} (f(x^{(i)}) - f^*)^2 \leq R^2 G^2,$$

using again that $\|g^{(i)}\|_2 \leq G$. Since the series converges, we obtain that $f(x^{(k)}) \to f^*$, and furthermore, to achieve ϵ-suboptimality, at most $(RG/\epsilon)^2$ steps are required. □

2.2.2 Approximate Case with an Estimate of f^*

In the case where f^* is not known, we want to reuse the previous step by approximating $f^* \approx f_{best}^{(k)} - \gamma_k$, where $\gamma_k > 0$, $\gamma_k \to 0$, and we also assume $\sum_{k=1}^{\infty} \gamma_k = \infty$, resulting in:

$$\alpha_k = \frac{f(x^{(k)}) - f_{best}^{(k)} + \gamma_k}{\|g^{(k)}\|_2^2}.$$

Theorem 4 (Convergence Theorem)**.** Let $x^{(k)}$ be the iterates of the approximate Polyak's method and $f_{best}^{(k)}$ as above, then:

$$f_{best}^{(k)} \to f^*, \quad k \to \infty.$$

Proof. As before, we substitute α_i into the inequality (2):

$$\begin{aligned}
R^2 &\geq \sum_{i=1}^{k} (2\alpha_i (f(x^{(i)}) - f^*) - \alpha_i^2 \|g^{(i)}\|_2^2) \\
&= \sum_{i=1}^{k} \frac{2(f(x^{(i)}) - f_{best}^{(i)} + \gamma_i)(f(x^{(i)}) - f^*) - (f(x^{(i)}) - f_{best}^{(i)} + \gamma_i)^2}{\|g^{(i)}\|_2^2} \\
&= \sum_{i=1}^{k} \frac{(f(x^{(i)}) - f_{best}^{(i)} + \gamma_i)((f(x^{(i)}) - f^*) + (f_{best}^{(i)} - f^*) - \gamma_i)}{\|g^{(i)}\|_2^2}.
\end{aligned}$$

By contradiction, assume $f(x^{(k)}) - f^* \geq \epsilon > 0$. Then for $i = 1, ..., k$, we have $f(x^{(i)}) - f^* \geq \epsilon$. Let N be such that $\gamma_i \leq \epsilon$ for $i \geq N$. Then the second term in the numerator is positive:

$$(f(x^{(i)}) - f^*) + (f_{best}^{(i)} - f^*) - \gamma_i \geq \epsilon.$$

Therefore, all terms in the summation for $i > N$ are positive. Let S denote the summation up to $i = N - 1$. We have:

$$\sum_{i=N}^{k} \frac{(f(x^{(i)}) - f_{best}^{(i)} + \gamma_i)((f(x^{(i)}) - f^*) + (f_{best}^{(i)} - f^*) - \gamma_i)}{\|g^{(i)}\|_2^2} \leq R^2 - S.$$

Using the fact that $f(x^{(i)}) - f_{best}^{(i)} + \gamma_i \geq \gamma_i$ and $\|g^{(i)}\|_2 \leq G$:

$$(\epsilon/G^2) \sum_{i=N}^{k} \gamma_i \leq R^2 - S.$$

This leads to a contradiction because the left-hand term tends to infinity as $k \to \infty$, while the right-hand term is constant in k. □

2.3 Projected Subgradient Method

The projected subgradient method is the first method in this discussion that deals with constrained problems. Specifically, it addresses convex problems of the form:

$$\begin{aligned} &\min f(x) \\ &x \in C, \end{aligned}$$

where C is a closed convex set. The method is given by:

$$x^{(k+1)} = \Pi\Big(x^{(k)} - \alpha_k g^{(k)}\Big),$$

where Π is the (Euclidean) projector onto C, and $g^{(k)}$ is any subgradient of f at $x^{(k)}$. Note that $x^{(k)}$ is always feasible.
Depending on the choice of α_k, we have different convergence results.

Theorem 5 (Convergence Theorem)**.** For the projected subgradient method, by choosing:

constant step size or constant step length, the algorithm converges in an ϵ-suboptimal manner, i.e.:

$$\lim_{k\to\infty} f_{best}^{(k)} - f^* < \epsilon,$$

for some $\epsilon > 0$, which decreases with the step size parameter;

square summable step size or non-summable step size or infinitesimal non-summable step size, we can guarantee convergence to the optimal value:

$$\lim_{k\to\infty} f(x_{best}^{(k)}) = f^*.$$

Remark. The theorem just stated is practically identical to Theorem **??** on the convergence of the base subgradient method. This is because in the proof, we will refer to the proof of Theorem **??**.

Proof (Theorem 5)**.** Let $z^{(k)} = x^{(k)} - \alpha_k g^{(k)}$, which may not be in C. As in the proof of the base subgradient method:

$$\begin{aligned}\|z^{(k+1)} - x^*\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^*\|_2^2 \\ &= \|x^{(k)} - x^*\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^*) + \alpha_k^2 \|g^{(k)}\|_2^2 \\ &\leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) + \alpha_k^2 \|g^{(k)}\|_2^2.\end{aligned}$$

The key observation is that when we project a point onto C, we get closer to every point in C, especially x^*:

$$\|x^{(k+1)} - x^*\|_2^2 = \|\Pi(z^{(k+1)}) - x^*\|_2^2 \leq \|z^{(k+1)} - x^*\|_2^2.$$

Combining the two expressions above, we obtain:

$$\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) + \alpha_k^2 \|g^{(k)}\|_2^2.$$

This is the starting point of the proof of inequality (**??**), which allows us to exactly reduce to the proof in the base case.

□

2.4 Subgradient Method for Constrained Problems

In this section, we present a method for constrained problems of the form:

$$\begin{aligned} &\min f_0(x) \\ &x \in \{f_i(x) \leq 0, \quad i = 1, ..., m\}, \qquad\qquad f_i \text{ convex.} \end{aligned}$$

The algorithm we present has the following form:

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)},$$

where $\alpha_k > 0$ is the step size and $g^{(k)}$, unlike the previous algorithms, is chosen as follows:

$$g^{(k)} \in \begin{cases} \partial f_0(x^{(k)}) & \text{if } f_i(x^{(k)}) \leq 0 \ \ \forall i = 1, ..., m; \\ \partial f_j(x^{(k)}) & \text{with } f_j(x^{(k)}) > 0. \end{cases}$$

In particular, if $x^{(k)}$ is feasible, we proceed as if the problem were unconstrained; otherwise, we choose the direction opposite to the subgradient of one of the violated constraints.
Often, the iterates will be infeasible, so it will be useful to redefine:

$$f_{best}^{(k)} = \min\{f_0(x^{(i)} \mid x^{(i)} \text{ feasible}, \ i = 1, ..., k\}.$$

(If $x^{(0)}, ..., x^{(k)}$ are infeasible, then $f_{best}^{(k)} = \infty$.)

In addition to the assumptions made in Section 2.1.2, we assume the existence of a Slater point x^{sf}, i.e., a point for which $f_i(x^{sf}) < 0, \ i = 1, ..., m$.
We now prove a convergence result in which the step size is chosen to be infinitesimal and non-summable. Similar results can be obtained for other choices of α_k.

Theorem 6 (Convergence Theorem)**.** Let $x^{(k)}$ be the iterates of the method described above, where we choose α_k to be infinitesimal and non-summable. Then:

$$f_{best}^{(k)} \to f^*, \ k \to \infty.$$

Proof. Suppose, for contradiction, that $f_{best}^{(k)} \nrightarrow f^*$. Then, there exists $\epsilon > 0$ such that $f_{best}^{(k)} \geq f^* + \epsilon$ for every k, which implies $f(x^{(k)}) \geq f^* + \epsilon$ for every k for which $x^{(k)}$ is feasible.

Step 1. Find $\tilde{x}$ (not necessarily among the iterates) and $\mu > 0$ such that:

$$f_0(\tilde{x}) \leq f^* + \epsilon/2, \qquad\qquad f_1(\tilde{x}) \leq -\mu, ..., f_m(\tilde{x}) \leq -\mu.$$

$\tilde{x}$ is $\epsilon/2$-suboptimal and satisfies the constraints with a margin of μ.
We search for $\tilde{x}$ in the segment between x^* and x^{sf}: $\tilde{x} = (1-\theta)x^* + \theta x^{sf}$, where $\theta \in (0, 1)$. Due to the convexity of f_0, we have:

$$f_0(\tilde{x}) \leq (1 - \theta)f^* + \theta f_0(x^{sf}),$$

so by choosing $\theta = \min\{1, (\epsilon/2)/(f_0(x^{sf}) - f^*)\}$, we have:

$$f_0(\tilde{x}) \leq f^* + \epsilon/2.$$

Furthermore, by the convexity of f_i and the fact that x^* is feasible, we have:

$$f_i(\tilde{x}) \leq (1-\theta) f_i(x^*) + \theta f_i(x^{sf}) \leq \theta f_i(x^{sf}),$$

and we take $\mu = -\theta \min_i f_i(x^{sf})$.

Step 2. Consider $i \in \{1, ..., k\}$ for which $x^{(i)}$ is feasible.
Then we have $g^{(i)} \in \partial f_0(x^{(i)})$ and, by contradiction, $f_0(x^{(i)}) \geq f^* + \epsilon$. Since $\tilde{x}$ is ϵ-suboptimal, we have $f_0(x^{(i)}) - f_0(\tilde{x}) \geq \epsilon/2$ and we obtain:

$$\begin{aligned}
\|x^{(i+1)} - \tilde{x}\|_2^2 &= \|x^{(i)} - \tilde{x}\|_2^2 - 2\alpha_i g^{(i)T}(x^{(i)} - \tilde{x}) + \alpha_i^2 \|g^{(i)}\|_2^2 \\
&\leq \|x^{(i)} - \tilde{x}\|_2^2 - 2\alpha_i (f_0(x^{(i)}) - f_0(\tilde{x})) + \alpha_i^2 \|g^{(i)}\|_2^2 \\
&\leq \|x^{(i)} - \tilde{x}\|_2^2 - \alpha_i \epsilon + \alpha_i^2 \|g^{(i)}\|_2^2;
\end{aligned}$$

in the second line, we use the definition of subgradient.

Step 2bis. Suppose $i \in \{1, ..., k\}$ for which $x^{(i)}$ is infeasible and $g^{(i)} \in \partial f_p(x^{(i)})$, with $f_p(x^{(i)}) > 0$. Since $f_p(\tilde{x}) \leq -\mu$, we have $f_p(x^{(i)}) - f_p(\tilde{x}) \geq \mu$ and:

$$\begin{aligned}
\|x^{(i+1)} - \tilde{x}\|_2^2 &= \|x^{(i)} - \tilde{x}\|_2^2 - 2\alpha_i g^{(i)T}(x^{(i)} - \tilde{x}) + \alpha_i^2 \|g^{(i)}\|_2^2 \\
&\leq \|x^{(i)} - \tilde{x}\|_2^2 - 2\alpha_i (f_p(x^{(i)}) - f_p(\tilde{x})) + \alpha_i^2 \|g^{(i)}\|_2^2 \\
&\leq \|x^{(i)} - \tilde{x}\|_2^2 - 2\alpha_i \mu + \alpha_i^2 \|g^{(i)}\|_2^2.
\end{aligned}$$

Step 3. By choosing $\delta = \min\{\epsilon, 2\mu\} > 0$:

$$\|x^{(i+1)} - \tilde{x}\|_2^2 \leq \|x^{(i)} - \tilde{x}\|_2^2 - \alpha_i \delta + \alpha_i^2 \|g^{(i)}\|_2^2,$$

and recursively:

$$\|x^{(k+1)} - \tilde{x}\|_2^2 \leq \|x^{(1)} - \tilde{x}\|_2^2 - \delta \sum_{i=1}^{k} \alpha_i + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2.$$

It follows that:

$$\delta \sum_{i=1}^{k} \alpha_i \leq R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2 \iff \delta \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{\sum_{i=1}^{k} \alpha_i}$$

but the right-hand side tends to zero as $k \to \infty$. Contradiction. □

Bibliography

[1] S. Boyd et al. "Subgradients". In: *lecture notes of EE364b, Stanford University* (Spring 2021-2022).

[2] Stephen Boyd. "Subgradient methods". In: *lecture notes of EE364b, Stanford University* (Spring 2013-2014).

[3] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, 2010. ISBN: 9780387709130.