Continuous Optimization

Chapter 2: Gradient Descent

1 Descent Direction Methods

In this chapter we consider the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^n} \quad f(x).$$

The iterative algorithms that we will consider in this chapter take the form

$$x_{k+1} = x_k + t_k d_k \quad k = 0, 1, \ldots,$$

where d_k is the so-called direction and t_k is the step size. We will limit ourselves to descent directions, whose definition is now given.

Definition 1.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ with $f \in \mathrm{C}^1(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$. A vector $0 \neq d \in \mathbb{R}^n$ is called a descent direction of f at x if the directional derivative $f'(x, d)$ is negative, i.e.,*

$$f'(x, d) = \nabla f(x)^T d < 0.$$

In particular, by taking small enough steps, descent directions lead to a decrease of the objective function.

Lemma 1.1 (descent property of descent directions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ with $f \in \mathrm{C}^1(\mathbb{R}^n)$ and let $x \in \mathbb{R}^n$. Suppose that d is a descent direction of f at x. Then, there exists $\epsilon > 0$ such that*

$$f(x + td) < f(x) \quad \forall\, t \in (0, \epsilon].$$

Proof. Since $f'(x, d) < 0$, it follows from the definition of the directional derivative that

$$\lim_{t\to 0^+} \frac{f(x + td) - f(x)}{t} = f'(x, d) < 0.$$

Therefore, there exists an $\epsilon > 0$ such that

$$\frac{f(x + td) - f(x)}{t} < 0,$$

for any $t \in (0, \epsilon)$ $\square$

Algorithm 1: Schematic Descent Directions Method

Input: $x_0 \in \mathbb{R}^n$
$k = 0$
while *Termination criterion is not satisfied* **do**
  Pick a descent direction d_k
  Find a step size t_k satisfying $f(x_k + t_k d_k) < f(x_k)$
  $x_{k+1} = x_k + t_k d_k$
  $k = k + 1$

Various choices are still unspecified: which direction to take, how to select the step size, what termination criterion to use.

2 Gradient Method

The most important choice in the algorithm above concerns the selection of the descent direction. One obvious choice is to pick the steepest (normalized) direction, i.e., $d_k = -\nabla f(x_k)/||\nabla f(x_k)||$. In fact, this direction minimizes the directional derivatives between all normalized directions.

Lemma 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ with $f \in \mathrm{C}^1(\mathbb{R}^n)$ and let $x \in \mathbb{R}^n$ be non-stationary (i.e., $\nabla f(x) \neq 0$). Then the optimal solution of the problem*

$$\begin{aligned} &\min \;\; f'(x, d), \\ &s.t. \;\;\; ||d|| = 1. \end{aligned}$$

is $d = -\frac{\nabla f(x)}{||d||}$.

Proof. As $f \in \mathrm{C}^1(\mathbb{R}^n)$ and by Cauchy-Schwarz, we have

$$f'(x, d) = \nabla f(x)^T d \geq -|| \nabla f(x)|| \cdot ||d|| = -|| \nabla f(x)||.$$

Thus, $-|| \nabla f(x)||$ is a lower bound for the optimal value of the problem. On the other hand, by plugging $d = -\nabla f(x)/|| \nabla f(x)||$ in the objective function we get

$$f' \left(x, -\frac{\nabla f(x)}{|| \nabla f(x)||} \right) = -\nabla f(x)^T \left(\frac{\nabla f(x)}{|| \nabla f(x)||} \right) = -|| \nabla f(x)||,$$

and we thus come to the conclusion that the lower bound is attained at $d = -\frac{\nabla f(x)}{||d||}$. □

Thus, the gradient method selects $d_k = -\nabla f(x_k)$ which is obviously a descent direction, i.e.,

$$\nabla f(x_k)^T d_k = -\nabla f(x_k)^T \nabla f(x_k) = -|| \nabla f(x)||^2.$$

To define an implementable method, the second important choice we have to make is the selection of the step size t. In particular, this will be clearer once we provide the Descent Lemma below, which require the gradient to be Lipschitz continuous.

Definition 2.1 (Lipschitz Continuous Gradient)**.** *$\nabla f(x)$ is said to be Lipschitz continuous if*

$$|| \nabla f(x) - \nabla f(y)|| \leq L||x - y|| \quad \forall x, y \in \mathbb{R}^n.$$

The class of functions with Lipschitz continuous gradient with constant L is denoted by $\mathrm{C}^{1,1}_{\mathrm{L}}(\mathbb{R}^n)$.

Theorem 2.1. *Let $f \in \mathrm{C}^2(\mathbb{R}^n)$. Then the following two claims are equivalent:*

(a) $f \in \mathrm{C}^{1,1}_{\mathrm{L}}(\mathbb{R}^n)$

(b) $|| \nabla^2 f(x)|| \leq L \quad \forall x \in \mathbb{R}^n$.

Proof. $(b) \Rightarrow (a)$. Suppose that $|| \nabla^2 f(x)|| \leq L \quad \forall x \in \mathbb{R}^n$. By the fundamental theorem of calculus we have $\forall x, y \in \mathbb{R}^n$

$$\nabla f(y) = \nabla f(x) + \int_0^1 \nabla^2 f(x + t(y-x))(y-x)dt = \nabla f(x) + \left(\int_0^1 \nabla^2 f(x + t(y-x))dt \right) \cdot (y-x)$$

Thus,

$$\begin{aligned} || \nabla f(y) - \nabla f(x)|| &= \left\| \left(\int_0^1 \nabla^2 f(x + t(y-x))dt \right) \cdot (y-x) \right\| \\ &\leq \left\| \int_0^1 \nabla^2 f(x + t(y-x))dt \right\| \cdot \|(y-x)\| \\ &\leq \left(\int_0^1 || \nabla^2 f(x + t(y-x))||dt \right) \cdot \|(y-x)\| \\ &\leq L\|(y-x)\| \end{aligned}$$

$(a) \Rightarrow (b)$. Exercise. □

Lemma 2.2 (Descent Lemma (prequel))**.** *Let $f \in \mathrm{C}^{1,1}_{\mathrm{L}}(\mathbb{R}^n)$. Then for any $x, y \in \mathbb{R}^n$*

$$f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|x-y\|^2.$$

Proof. From the fundamental theorem of calculus and differentiability of f we have

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 \nabla f((1-t)x+ty)^T(y-x)\,dt \\
&= f(x) + \int_0^1 \nabla f((1-t)x+ty)^T(y-x) - \nabla f(x)^T(y-x)\,dt + \nabla f(x)^T(y-x) \\
&\leq f(x) + \int_0^1 \|\nabla f((1-t)x+ty) - \nabla f(x)\| \cdot \|y-x\|\,dt + \nabla f(x)^T(y-x) \\
&\leq f(x) + \int_0^1 L\|t(y-x)\| \cdot \|y-x\|\,dt + \nabla f(x)^T(y-x) \\
&= f(x) + L\|y-x\|^2 \cdot \frac{t^2}{2}\Big|_0^1 + \nabla f(x)^T(y-x) \\
&= f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2,
\end{aligned}$$

where the second inequality follows from the Lipschitz continuity of ∇f. $\square$

Lemma 2.3 (Descent Lemma)**.** *Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then for any $x \in \mathbb{R}^n$ and $t > 0$*

$$f(x) - f(x - t\,\nabla f(x)) \geq t(1 - \frac{Lt}{2})\|\nabla f(x)\|^2.$$

Proof. The result simply follows by applying the descent lemma (prequel) on x and $y = x - \nabla f(x)$

$$f(x - t\,\nabla f(x)) \leq f(x) - t\|\nabla f(x)\|^2 + \frac{Lt^2}{2}\|\nabla f(x)\|^2 = f(x) - t(1 - \frac{Lt}{2})\|\nabla f(x)\|^2$$

$\square$

In particular, this holds for $x = x_k$ and $x_{k+1} = x_k - \nabla f(x_k)$,

$$f(x_k) - f(x_{k+1}) \geq t(1 - \frac{Lt}{2})\|\nabla f(x_k)\|^2,$$

which in turns implies that if we select $t \in (0, \frac{2}{L})$ we ensure a decrease of the objective function at each iteration. In particular, if we want to achieve the largest guarantee bound on the decrease, then we seek the maximum of $t(1 - \frac{Lt}{2})$ w.r.t. t, which is attained at $t = \frac{1}{L}$ with a decrease that becomes

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2. \tag{1}$$

At this point we can write down the Gradient Method in terms of an implementable algorithm.

Algorithm 2: Gradient Descent (GD) Method

Input: Pick $x_0 \in \mathbb{R}^n$ arbitrarly, chose $\epsilon > 0$ (e.g., 10^{-4}).
$k = 0$
while $\|\nabla f(x_k)\| > \epsilon$ **do**
    $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$
    $k = k + 1$

Let us now prove convergence for GD, in particular that $\nabla f(x_k)$ goes to zero.

Theorem 2.2 (Convergence of GD)**.** *Let $f \in C_L^{1,1}(\mathbb{R}^n)$ and let $\{x_k\}_k$ be a sequence generated by Algorithm 1 for solving $\min_{x \in \mathbb{R}^n} f(x)$. Assume that f is bounded below over $\mathbb{R}^n$, i.e., there exists $m \in \mathbb{R}$ such that $f(x) > m \;\; \forall\, x \in \mathbb{R}^n$. Then we have the following*

(a) The sequence $\{f(x_k)\}_k$ is nonincreasing. In addition, for any $k \geq 0$, $f(x_{k+1}) < f(x_k)$ unless $\nabla f(x_k) = 0$.

(b) $\nabla f(x_k) \to 0$ as $k \to \infty$.

Proof. (a) directly follows from (1), as $f(x_{k+1}) < f(x_k)$ and the equality $f(x_{k+1}) = f(x_k)$ only holds when $\nabla f(x_k) = 0$. (b) Since the sequence $\{f(x_k)\}_k$ is nonincreasing and bounded from below, it converges. Thus, $f(x_k) - f(x_{k+1}) \to 0$ as $k \to \infty$, which combined with (1) implies that $\|\nabla f(x_k)\| \to 0$ as $k \to \infty$. $\square$

Moreover, we can provide the rate of convergence of GD.

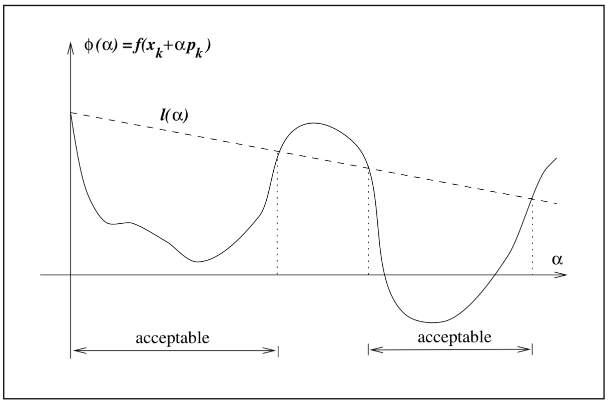

Figure 1: The figure represents the Armijo line search condition (the notation in this figure is different from the text, replace α in the figure with t from the text.)

Theorem 2.3 (Rate of Convergence of GD)**.** *Under the setting of Theorem 2.2, let f^* be the limit of the convergent sequence $\{f(x_k)\}_k$. Then for any $T = 0, 1, \dots$*

$$\min_{k=0,1,\dots,T} \| \nabla f(x_k)\| \leq \sqrt{\frac{L(f(x_0) - f^*)}{T+1}}$$

Proof. Summing the inequality (1) over $k = 0, 1, \dots, T$, we obtain

$$f(x_0) - f(x_{T+1}) = \frac{1}{L} \sum_{k=0}^{T} \| \nabla f(x_k)\|^2 \geq \frac{T+1}{L} \min_{k=0,1,\dots,T} \| \nabla f(x_k)\|^2$$

which concludes the proof. □

2.1 Line search methods

The gradient method as defined above can only be employed when we know or we can compute the Lipschitz constant L, on the other hand, we would like to have a general method that can be applied on any unconstrained optimization problem. An alternative for selecting the step size is provided by line search methods. Consider a direction d_k, one option would be to exactly minimize along the direction d_k, i.e., **exact line search**

$$t_k \in \operatorname*{argmin}_{t>0} f(x_k + td_k).$$

However, this approach is not always viable and even when it is, it might be costly. Another option is instead that of accepting a step that will make the function value decrease "sufficiently", namely to apply an **inexact line search**. In particular, the first line search proposed in the literature is called Armijo line search [1] and it requires the following

$$f(x_k + t_k d_k) \leq f(x_k) + \alpha t_k \nabla f(x_k)^T d_k. \tag{2}$$

Notice that if we define $\phi(t) = f(x_k + td_k)$ we can rewrite the inequality above as

$$\phi(t_k) \leq \phi(0) + \alpha t_k \phi'(0) \qquad \text{with } \alpha \in (0, 1).$$

As depicted in Figure 1, the condition requires that the new function value $\phi(t_k)$ stays below the line passing for $(0, \phi(0))$ and with $\alpha\phi'(0)$ as inclination. Notice that as $\phi'(0) < 0$ and $\alpha < 1$, the line $y = \phi(0) + \alpha t_k \phi'(0)$ is not as inclined as the tangent in 0. The way for selecting a step t_k that satisfies (2) is suggested by the figure. In particular, the method is called backtracking and it is described below.

Algorithm 3: Backtracking on Armijo line search

Input: Pick $s > 0$, $\alpha, \beta \in (0, 1)$.

$i = 0$

do

  $t_k = s\beta^i$

  $i = i + 1$

while $f(x_k + t_k d_k) > f(x_k) + \alpha t_k \nabla f(x_k)^T d_k$;

Let us first show that this method terminates in a finite amount of steps

Lemma 2.4. *Let $f \in \mathrm{C}^{1,1}_{\mathrm{L}}(\mathbb{R}^n)$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ be a descent direction. Then Algorithm 3 terminates in a finite amount of steps with a $t_k > 0$ that satisfies (2). Moreover, one of the following holds*

(a) $t_k = s$

(b) $t_k \le \beta s$ *such that* $f(x_k + \frac{t_k}{\beta} d_k) > f(x_k) + \alpha \frac{t_k}{\beta} \nabla f(x_k)^T d_k$

Consequently, with $d_k = -\nabla f(x_k)$ *we get* $t_k \ge \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\}$.

Proof. Let us first prove that the algorithm terminates in a finite amount of steps. By contradiction there are no finite value of i for which (2) is satisfied, that is

$$\frac{f(x_k + s\beta^i d_k) - f(x_k)}{s\beta^i} > \alpha \nabla f(x_k)^T d_k.$$

Given $\beta < 1$ we have that $\lim_{i\to\infty} \beta^i = 0$ and thus, with $i \to \infty$ the LHS of the inequality above is the directional derivative of f along d_k. In particular, we get

$$\nabla f(x_k)^T d_k \ge \alpha \nabla f(x_k)^T d_k,$$

which is a contradiction, as $\nabla f(x_k)^T d_k < 0$ and $\alpha < 1$. Following the steps of the algorithm, either the first guess s is accepted or $t_k \le s\beta$. In the second case, given t_k the outcome of the algorithm, the step size before the last backtracking ($\frac{t_k}{\beta}$) was surely not accepted, from which (b) follows.

Now, we can replace $\frac{t_k}{\beta}$ in Lemma 2.3 with $x = x_k$ to get

$$f(x_k) - f(x_k - \frac{t_k}{\beta} \nabla f(x_k)) \ge \frac{t_k}{\beta}\left(1 - \frac{Lt_k}{2\beta}\right) \| \nabla f(x_k)\|^2$$

which combined with (b) with $d_k = -\nabla f(x_k)$ implies

$$\frac{t_k}{\beta}\left(1 - \frac{Lt_k}{2\beta}\right) < \alpha \frac{t_k}{\beta}$$

and consequently $t_k > \frac{2(1-\alpha)\beta}{L}$, which together with (a) concludes the proof. □

We can now provide a version of the GD method that is independent from L.

Algorithm 4: Gradient Descent (GD) Method with Armijo Line Search

---

Input: Pick $x_0 \in \mathbb{R}^n$ arbitrarily, chose $\epsilon > 0$ (e.g., 10^{-4}).

$k = 0$
while $\| \nabla f(x_k)\| > \epsilon$ **do**
  $t_k \leftarrow$ Armijo Line Search (Algorithm 3)
  $x_{k+1} = x_k - t_k \nabla f(x_k)$
  $k = k + 1$

Notice that to prove convergence of Algorithm 4 it suffices to show that also in this case we can derive a decrease as in (1), where the step size is replaced by a constant term. In particular, from the Lemma above, we get

$$f(x_k) - f(x_{k+1}) \ge \alpha \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\} \| \nabla f(x_k)\|^2.$$

Moreover, the asymptotic convergence of Algorithm 4 can also be proven if we assume $f \in \mathrm{C}^1(\mathbb{R}^n)$ instead of $f \in \mathrm{C}^{1,1}_{\mathrm{L}}(\mathbb{R}^n)$ (Exercise).

2.1.1 Nonmonotone line search

Until now we focused on methods that always ensures/requires a decrease of the objective function at every iteration. In fact, looking at the proof of Theorem 2.2, we strongly rely on (1). In some case, however, this requirement is too tight and a few methods (e.g., Newton [3], Barziai-Borwein [4]) get great (numerical) advantages from some additional freedom. In fact, it is still possible to prove convergence also when the objective function does not decrease at every step, but every $M > 0$ ($\in \mathbb{N}$) steps. In particular, we can consider the following nonmonotone condition, originally proposed in [3],

$$f(x_k + t_k d_k) \le \max_{0 \le j \le \min\{k,M\}} f(x_{k-j}) + \alpha t_k \nabla f(x_k)^T d_k. \tag{3}$$

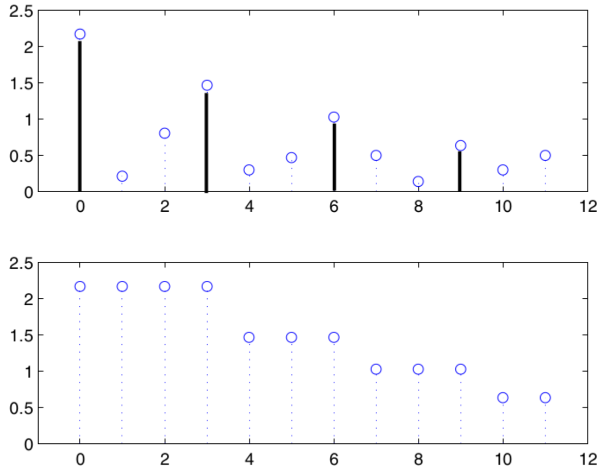

Figure 2: The sequences $\{f(x_k)\}_k$ and $\{R_k\}_k$.

and design a new backtracking line search that employs (3) instead of (2). If we now call $R_k := \max_{0\le j\le \min\{k,M\}} f(x_{k-j})$ the reference value at iteration k and notice that $R_k \ge f(x_k)$ at each iteration, it is easy to prove a lemma similar to Lemma 2.4, i.e., also this backtracking line search terminates in a finite amount of internal iterations and the step size t_k has the same lower bound. Moreover, we can achieve the following result for the sequence $\{R_k\}_k$.

Theorem 2.4. *Let $f \in \mathrm{C}^{1,1}_{\mathrm{L}}(\mathbb{R}^n)$ be limited from below. Let $\{x_k\}_k$ be a sequence generated by Algorithm 4 with* (3) *used in Algorithm 3 (instead of* (2)*) for solving $\min_{x\in\mathbb{R}^n} f(x)$. Then we have*

- *(a)* $x_k \in \mathcal{L}_0 \;\; \forall k$ *(where $\mathcal{L}_0 := \{x \in \mathbb{R}^n : f(x) \le f(x_0)\}$)*
- *(b)* $\lim_{k\to\infty} R_k = \lim_{k\to\infty} f(x_k) = f^*$
- *(c)* $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0.$

Proof. First, we show that the sequence $\{R_k\}_k$ is monotonically nonincreasing. For each k, let $r(k) \in [k - \min\{k, M\}, k]$ be the smallest integer such that

$$R_k = f(x_{r(k)}) = \max_{0\le j\le \min\{k,M\}} f(x_{k-j}).$$

Thus, we rewrite (2) as

$$f(x_{k+1}) \le f(x_{r(k)}) - \alpha t_k \|\nabla f(x_k)\|^2 = f(x_{r(k)}) - \frac{\alpha}{t_k}\|x_{k+1} - x_k\|^2, \tag{4}$$

Since $\min(k+1, W) \le \min(k, W) + 1$, we have

$$\begin{aligned} f(x_{r(k+1)}) &= \max_{0\le j\le \min(k+1,W)} f(x_{k-j+1}) \\ &\le \max_{0\le j\le \min(k,W)+1} f(x_{k-j+1}) \\ &= \max\{f(x_{r(k)}), f(x_{k+1})\} = f(x_{r(k)}), \end{aligned}$$

where last equality follows from (4). Since $\{f(x_{r(k)})\}$ is nonincreasing and $x_{r(0)} = x_0$, we have that $f(x_k) \le f(x^0)\ \forall k$, which proves (a).

Since f is limited from below, the monotone nonincreasing sequence $\{f(x_{r(k)})\}$ admits a limit f^* for $k \to \infty$. By induction on j, with $1 \le j \le W + 1$, let us prove that the two limits below are satisfied:

$$\lim_{k\to\infty} \|x_{r(k)-j+1} - x_{r(k)-j}\| = 0 \tag{5}$$

$$\lim_{k\to\infty} f(x_{r(k)-j}) = \lim_{k\to\infty} f(x_{r(k)}) \tag{6}$$

where k is assumed to be large enough to have $r(k) \ge k - W > 1$.

If $j = 1$, using (4) with $k = r(k) - 1$, we have

$$f(x_{r(k)}) \le f(x_{r(r(k)-1)}) - \frac{\alpha}{t_{r(k)-1}}\|x_{r(k)} - x_{r(k)-1}\|^2.$$

Thus, together with convergence of $\{f(x_{r(k)})\}$ and the fact that $t_k \le s$, we obtain

$$\lim_{k\to\infty} \|x_{r(k)} - x_{r(k)-1}\| = 0$$

From Lipschitz continuity of f and the above limit we obtain that

$$\lim_{k\to\infty} f(x_{r(k)-1}) = \lim_{k\to\infty} f(x_{r(k)}),$$

which means that induction has been proved for the case $j = 1$.

Now assume that (5) and (6) are valid for a given j. From (4) used with $k = r(k) - j - 1$, we have that

$$f(x_{r(k)-j}) \le f(x_{r(r(k)-j-1)}) - \frac{\alpha}{t_{r(k)-j-1}} \|x_{r(k)-j} - x_{r(k)-j-1}\|^2.$$

Thus, together with (6), where the left limit is used for the LHS and the right limit is used for the RHS and remembering that $k = r(k) - j - 1$, and that $t_k \le s$ we obtain that

$$\lim_{k\to\infty} \|x_{r(k)-j} - x_{r(k)-j-1}\| = 0.$$

From the limit above, Lipschitz continuity of f and again (6) we obtain that

$$\lim_{k\to\infty} f(x_{r(k)-j-1}) = \lim_{k\to\infty} f(x_{r(k)-j}) = \lim_{k\to\infty} f(x_{r(k)}),$$

which means that induction has been proved from a generic j to $j + 1$.

In particular (5) and (6) are also valid if we replace $r(k)$ with $R(k) := r(k + W + 1)$. Now, for k sufficiently large, we have that

$$\begin{aligned} x_{R(k)} &= x_k + (x_{k+1} - x_k) + \cdots + (x_{R(k)} - x_{R(k)-1}) \\ &= x_k + \sum_{j=1}^{R(k)-k} \left(x_{R(k)-j+1} - x_{R(k)-j}\right) \end{aligned} \tag{7}$$

Since $r(k + W + 1) \le k + W + 1$, we have $R(k) - k \le W + 1$ and, thus, from (7) and (5) used replacing $r(k)$ with $R(k)$, we obtain

$$\lim_{k\to\infty} \|x_k - x_{R(k)}\| = 0.$$

From convergence of $\{f(x_{r(k)})\}$ and Lipschitz continuity, it follows that

$$\lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} f(x_{R(k)}) = \lim_{k\to\infty} f(x_{r(k+W+1)}) = f^*,$$

which complete proof of (b). Thesis (c) follows from (b), (4) and the fact that $t_k \ge \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\}$. □

2.2 Barzilai-Borwein Method

Let us first consider a quadratic objective function,

$$f(x) = x^T Q x + c^T x + b, \tag{8}$$

where $x, c \in \mathbb{R}^n$, $b \in \mathbb{R}$, $Q \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. In this case, we get the following result for GD.

Proposition 2.1 (Finite convergence of GD in the quadratic case)**.** *Let $\lambda_1 \ge \lambda_2, \dots, \lambda_n$ be the eigenvalue of the Hessian of the quadratic function* (8)*. Then, GD defined by the iteration*

$$x_{k+1} = x_k - \frac{1}{\lambda_k} \nabla f(x_k) \quad \text{with}\, k = 1, \dots, n, \tag{9}$$

determines in at maximum n iterations the minimizer of x_k.

Proof. As gradient of (8) is $\nabla f(x) = Qx + c$, we get

$$\nabla f(x_{k+1}) = Qx_{k+1} + c = Qx_k + c - \frac{1}{\lambda_k} Q\, \nabla f(x_k) = \left(I - \frac{1}{\lambda_k} Q\right) \nabla f(x_k)$$

Repeating the application of the above formula, we get

$$\nabla f(x_k) = \left(\prod_{j=1}^{k-1} (I - \frac{1}{\lambda_j} Q) \right) \nabla f(x_1)$$

Let $u_h \in \mathbb{R}^n : h = 1, \ldots, n$ be the set of eigenvectors of Q, associated to the eigenvalues λ_h, in particular they form a basis of $\mathbb{R}^n$, meaning that we can represent $\nabla f(x_1)$ as a linear combination of these vectors, i.e.,

$$\nabla f(x_1) = \sum_{h=1}^{n} \beta_h u_h, \quad \text{with}\, \beta_h \in \mathbb{R}.$$

Thus, for $k \geq 2$, we get

$$\nabla f(x_k) = \left(\prod_{j=1}^{k-1} (I - \frac{1}{\lambda_j} Q) \right) \left(\sum_{h=1}^{n} \beta_h u_h \right),$$

and consequently, setting $k = n + 1$, from $Iu_h = u_h$ and $Qu_h = \lambda_h u_h$, we have

$$\nabla f(x_{n+1}) = \sum_{h=1}^{n} \beta_h \left(\prod_{j=1}^{n} (1 - \frac{1}{\lambda_j} \lambda_h) \right) u_h = 0,$$

from which we get that GD converges in at most n steps. □

Unfortunately, the eigenvalues of Q are usually not available and obtaining them is roughly as expensive as applying *GD* on (8). For this reason we consider instead of (9) to approximate them on the fly via the Rayleigh quotient

$$R_Q(x) := \frac{x^T Q x}{||x||^2} \quad \forall x \neq 0,$$

given that

$$\lambda_n(Q) \leq R_Q(x) \leq \lambda_1(Q) \quad \text{with}\, \forall x \neq 0.$$

In particular, if the image of Q is small (i.e., the interval $[\lambda_n(Q), \lambda_1(Q)]$ is small), $R_Q(x)$ approximates the eigenvalues of Q.

References

[1] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

[2] Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[3] Luigi Grippo, Francesco Lampariello, and Stefano Lucidi. A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

[4] M. Raydan. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM Journal on Optimization*, 7(1):26–33, 1997.