

# Continuous Optimization

## Chapter 2: Gradient Descent

### 1 Descent Direction Methods

In this chapter we consider the unconstrained minimization problem

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

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

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

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  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  with  $f \in C^1(\mathbb{R}^n)$ . A vector  $0 \neq d \in \mathbb{R}^n$  is called a descent direction of  $f$  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  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  with  $f \in C^1(\mathbb{R}^n)$  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

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

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

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

for any  $t \in (0, \epsilon)$  □

---

**Algorithm 1:** Schematic Descent Directions Method

---

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

---

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

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

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

*Proof.* As  $f \in C^1(\mathbb{R}^n)$  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  $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$ .  $\square$

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

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

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  $C_L^{1,1}(\mathbb{R}^n)$ .

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

- (a)  $f \in C_L^{1,1}(\mathbb{R}^n)$
- (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.  $\square$

**Lemma 2.2** (Descent Lemma (prequel)). *Let  $f \in C_L^{1,1}(\mathbb{R}^n)$ . 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  $\nabla f$ . □

**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  $y = x - \nabla f(x)$

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

□

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

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

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

---

**Algorithm 2:** Gradient Descent (GD) Method

---

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

---

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 GD 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)  *$\nabla f(x_k) \rightarrow 0$  as  $k \rightarrow \infty$ .*

*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,  $f(x_k) - f(x_{k+1}) \rightarrow 0$  as  $k \rightarrow \infty$ , which combined with (1) implies that  $\|\nabla f(x_k)\| \rightarrow 0$  as  $k \rightarrow \infty$ .  $\square$

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

**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.  $\square$

## 2.1 Line search methods

The gradient method as defined above can only be implemented when we can compute the Lipschitz function  $L$ , 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 it might be sometimes costly. Another option is instead that of accepting a step that will make our function value decrease "sufficiently", namely to apply an inexact line search.

and now let us write the linear approximation of  $f$  along  $d_k$  as a function of the step size  $t_k$ , i.e.,

$$f(x_k + t_k d_k) = f$$

