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

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

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 $f : \mathbb{R}^n \to \mathbb{R}$ with $f \in \mathrm{C}^1(\mathbb{R}^n)$. 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 $f : \mathbb{R}^n \to \mathbb{R}$ 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 $t \in (0, \epsilon)$ □

Algorithm 1: Schematic Descent Directions Method

```
Input: x_0 ∈ 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 are still unspecified.

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*

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

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

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

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

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. $\square$

Theorem 2.2 (Descent Lemma)**.**

References