Gradient descent for unconstrained convex optimization problems

Huanle Xu *

November 29, 2018

1 Basic Elements of Iterative Algorithms

To fix ideas, let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Consider the unconstrained optimization problem:

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

In general, it may be too ambitious to find a global minimum of f. Hence, we will just look for a **stationary point** of f; i.e., a point $\bar{x} \in \mathbb{R}^n$ that satisfies $\nabla f(\bar{x}) = \mathbf{0}$. To begin, let $x^0 \in \mathbb{R}^n$ be an initial iterate with $\nabla f(x^0) \neq \mathbf{0}$. In order to achieve progress, we need to proceed in some **search direction** $d^k \in \mathbb{R}^n$. For instance, we can update the iterates according to the following rule:

$$x^{k+1} = x^k + \alpha_k d^k \qquad \text{for } k = 0, 1, \ldots. \tag{2}$$

Here, $\alpha_k > 0$ is called the **step size** and controls how far we proceed in the direction d^k. Note that (2) actually defines a *family* of update rules that are parametrized by the search directions $\{d^k\}_{k \geq 0}$ and step sizes $\{\alpha_k\}_{k \geq 0}$. There are many possibilities in choosing the search directions and step sizes. Below are some common choices.

*Huanle Xu is with the College of Computer Science and Technology, Dongguan University of Technology. E-mail: {xuhl}@dgut.edu.cn.

1.1 Choosing the search directions

Roughly speaking, the method of steepest descent is based on minimizing a linear approximation of f at the current iterate $x^k \in \mathbb{R}^n$. Specifically, suppose that the current iterate x^k satisfies $\nabla f(x^k) \neq \mathbf{0}$. Then, we may construct a *linear approximation* of f at x^k, which is given by

$$f^k(x) \equiv f(x^k) + \nabla f(x^k)^T(x - x^k).$$

Now, recall that if there exists a $d \in \mathbb{R}^n$ such that $\nabla f(x^k)^T d < 0$, then there exists an $\alpha_0 > 0$ such that $f(x^k + \alpha d) < f(x^k)$ for all $\alpha \in (0, \alpha_0)$ (Proposition 1 of Handout 7). Thus, in order to guarantee descent, we need to choose $x \in \mathbb{R}^n$ such that $\nabla f(x^k)^T(x - x^k) < 0$. Of course, if $\nabla f(x^k) \neq \mathbf{0}$, then we can make $\nabla f(x^k)^T(x - x^k)$ as negative as possible. Thus, we need to restrict the length of the direction $d = x - x^k$. In particular, we may consider the following:

$$\begin{aligned} \text{minimize} \quad & \nabla f(x^k)^T d \\ \text{subject to} \quad & \|d\|_2^2 \leq \|\nabla f(x^k)\|_2^2. \end{aligned}$$

By the Cauchy–Schwarz inequality, we see that the optimal solution to the above problem is $d^k = -\nabla f(x^k)$. The direction d^k is called the **direction of steepest descent**, and the resulting iterative algorithm

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) \qquad \text{for } k = 0, 1, \ldots \tag{3}$$

is called the **method of steepest descent** or simply the **gradient method**.

1.2 Choosing the stepsize

The simplest choice for the step sizes is to use the same value for all iterations; i.e., set $\alpha_k = \alpha$ for some $\alpha > 0$ and for $k = 0, 1, \ldots$. Unfortunately, an iterative method with constant step size may perform very poorly. On the one hand, if the step size is too large, then we may not be able to guarantee that the descent condition is satisfied, and the method may not converge. On the other hand, if the step size is too small, then the method may converge very slowly. An alternative approach is to perform a **line search** at each iterate. Specifically, given the current iterate $x^k \in \mathbb{R}^n$ and a search direction $d^k \in \mathbb{R}^n$, we compute the step size α_k by

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^k + \alpha d^k). \tag{5}$$

In fact, it is not necessary that we find the exact minimum, as the goal of the line search (5) is simply to achieve a substantial decrease in the objective value by proceeding in the direction d^k. For various implementations of (approximate) line search, we refer the reader to [1, Appendix C].

1.3 Handing Constraints

So far our discussion has focused on the unconstrained optimization problem (1). Let us now turn our attention to constrained optimization problems of the form

$$v^* = \min_{x \in X} f(x), \tag{6}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is as before and $X \subset \mathbb{R}^n$ is non–empty and closed. To develop an iterative algorithm for solving (6), a natural idea is to modify the update rule (2) to

$$x^{k+1} = \Pi_X \left(x^k + \alpha_k d^k \right) \qquad \text{for } k = 0, 1, \ldots, \tag{7}$$

where $\Pi_X : \mathbb{R}^n \to X$ is the projection operator onto X. Note that if the projection in (7) is well–defined for each $k \geq 0$ and if $x^0 \in X$, then the sequence of iterates $\{x^k\}_{k \geq 0}$ will all be feasible for (6). In particular, by taking $d^k = -\nabla f(x^k)$, we obtain the **projected gradient method**:

$$x^{k+1} = \Pi_X \left(x^k - \alpha_k \nabla f(x^k) \right) \qquad \text{for } k = 0, 1, \ldots. \tag{8}$$

2 Convergence Analysis of the Gradient Method

As mentioned in the Introduction, an important issue concerning iterative methods is their convergence behavior. In this section we will develop the machinery for analyzing the projected gradient method (8) when it is applied to solve a class of constrained convex optimization problems. Specifically, consider the following assumptions concerning (6):

Assumption 1

(a) The function f is continuously differentiable and strongly convex with parameter $\sigma > 0$, and the gradient ∇f is Lipschitz continuous with parameter $L > 0$; i.e., for all $x, y \in \mathbb{R}^n$,

$$\sigma\|x - y\|_2^2 \leq (\nabla f(x) - \nabla f(y))^T(x - y), \tag{9}$$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{10}$$

(b) The set X is non–empty, convex, and closed.

Under Assumption 1, it can be shown that the set $\{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$ is bounded for any given $x^0 \in \mathbb{R}^n$, and hence the optimal solution set X^* of problem (6) is non–empty. This motivates us to ask whether the sequence of iterates $\{x^k\}_{k\geq 0}$ generated by the projected gradient method (8) converges to an element of X^*. Towards that end, let us first study the convergence behavior of the objective values $\{f(x^k)\}_{k\geq 0}$ associated with the iterates $\{x^k\}_{k\geq 0}$. We begin with the following proposition:

Proposition 1 *For any $x, y \in \mathbb{R}^n$,*

$$\left|f(y) - f(x) - \nabla f(x)^T(y - x)\right| \leq \frac{L}{2}\|x - y\|_2^2.$$

Proof Consider the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(t) = f(x + t(y - x))$. By the Fundamental Theorem of Calculus and the Chain Rule, we have

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)\,dt = \int_0^1 \nabla f(x + t(y - x))^T(y - x)\,dt.$$

Upon writing

$$\int_0^1 \nabla f(x + t(y - x))^T(y - x)\,dt = \int_0^1 \left[\nabla f(x + t(y - x)) - \nabla f(x)\right]^T(y - x)\,dt + \nabla f(x)^T(y - x)$$

and using the Cauchy–Schwarz inequality together with the fact that ∇f is Lipschitz continuous with constant $L > 0$, we have

$$\begin{aligned}
\left|f(y) - f(x) - \nabla f(x)^T(y - x)\right| &\leq \int_0^1 \left|\left[\nabla f(x + t(y - x)) - \nabla f(x)\right]^T(y - x)\right| dt \\
&\leq L\|y - x\|_2^2 \int_0^1 t\,dt \\
&= \frac{L}{2}\|y - x\|_2^2,
\end{aligned}$$

2.1 convergence behavior

An important consequence of Proposition 1 is that the projected gradient method (8) is a descent method when the step sizes $\{\alpha_k\}_{k\geq 0}$ are sufficiently small. Specifically, we have the following corollary:

Corollary 1 *Suppose that Assumption 1 holds for problem (6). Then, the sequence of iterates* $\{x^k\}_{k\geq 0}$ *generated by the projected gradient method (8) satisfies*

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2}\left(\frac{1}{\alpha_k} - L\right)\|x^k - x^{k+1}\|_2^2. \tag{11}$$

In particular, if $\alpha_k < 1/L$ *for all* $k \geq 0$*, then the sequence* $\{f(x^k)\}_{k\geq 0}$ *is monotonically decreasing and hence convergent.*