Introductory Lectures on Optimization

Descent Method (1)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

October 16, 2024

Outline

Part I
Gradient (Steepest) Descent

Gradient Descent Formulation

We will refer to the following scheme as a gradent method. The scalar factor of the gradient, h_k, is called the step size or learning rate. Of course, it must be positive.

Gradient Method
Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$. Iterate $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \nabla f(\boldsymbol{x}_k), k = 0, 1, \ldots$

(1)

Remark. The subscript indicates the iteration number.

Step Size

1. The sequence $\{\{h_k\}_{k=1}^{\infty}$ is chosen in advance. For example,

$$h_k = h > 0, (constant\ step) \text{ or } h_k = \frac{h}{\sqrt{k+1}}.$$

2. Full relaxation:

$$h_k = \operatorname*{argmin}_{h \geq 0} f(\boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k)).$$

3. Goldstein-Armijo Find $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k)$ such that

$$\alpha\langle\nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle \leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}), \tag{2}$$

$$\beta\langle\nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle \geq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}), \tag{3}$$

where, $0 < \alpha < \beta < 1$ are some fixed parameters.

Step Size : Geometric Interpretation of Goldstein-Armijo

Let us fix $\boldsymbol{x} \in \mathbb{R}^n$. Consider the function of one variable

$$\phi(h) = f(\boldsymbol{x} - h\nabla f(\boldsymbol{x})), h \geq 0.$$

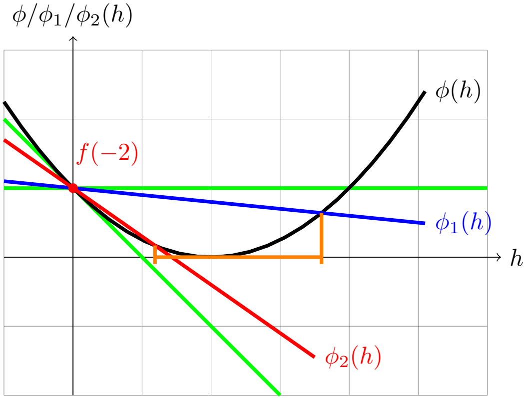
Then the step-size values acceptable for this strategy belong to the part of the graph of ϕ that is located between two linear functions:

$$\phi_1(h) = f(\boldsymbol{x}) - \alpha h \left\|\nabla f(\boldsymbol{x})\right\|^2, \; \phi_2(h) = f(\boldsymbol{x}) - \beta h \left\|\nabla f(\boldsymbol{x})\right\|^2.$$

Note that $\phi(0) = \phi_1(0) = \phi_2(0) = f(\boldsymbol{x})$, and $\phi'(0) < \phi_2'(0) < \phi_1'(0) < 0$.

Therefore, the acceptable values exist unless $\phi(h)$ is not bounded below.

Step Size : Geometric Interpretation of Goldstein-Armijo



$$f(\boldsymbol{x}) = \frac{1}{4}\boldsymbol{x}^2 \quad (\text{ We set } \boldsymbol{x}_k = -2).$$

$$\phi(h) = f(-2 - h\nabla f(-2))$$
$$= \frac{1}{4}(h-2)^2.$$

$$\phi_1(h) = f(-2) - \alpha h\|\nabla f(-2)\|^2$$
$$= 1 - \alpha h.$$

$$\phi_2(h) = f(-2) - \beta h\|\nabla f(-2)\|^2$$
$$= 1 - \beta h.$$

$$0 < \alpha = 0.1 < \beta = 0.7 < 1$$

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

Let us estimate the performance of the Gradient Method. Consider the problem

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

where $f \in C_L^{1,1}(\mathbb{R}^n)$, and assume that $f(\boldsymbol{x})$ is bounded below on $\mathbb{R}^n$.

Let us evaluate the result of one gradient step. Consider $\boldsymbol{y} = \boldsymbol{x} - h\nabla f(\boldsymbol{x})$. Then, in view of (1.2.5) of (Nesterov [2003]), we have

$$\boxed{|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle| \le \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 .}$$

Thus, we build the upper bound of $f(\boldsymbol{x} - h\nabla f(\boldsymbol{x}))$.

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

That is

$$\begin{aligned} f(\boldsymbol{y}) &\le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \ \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \\ &= f(\boldsymbol{x}) - h\|\nabla f(\boldsymbol{x})\|^2 + \frac{h^2}{2}L\|\nabla f(\boldsymbol{x})\|^2 \text{ since } \boldsymbol{y} = \boldsymbol{x} - h\nabla f(\boldsymbol{x}) \qquad (4) \\ &= f(\boldsymbol{x}) - h(1-\frac{h}{2}L)\|\nabla f(\boldsymbol{x})\|^2 . \end{aligned}$$

Remark. for $h \in (0, \frac{2}{L})$, $h(1-\frac{h}{2}L)\|\nabla f(\boldsymbol{x})\|^2$ is non-negative.

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

Thus, in order to get the best upper bound for the possible decrease of the objective function, we have to solve the following one-dimensional problem:

$$\Delta(h) = -h\left(1 - \frac{h}{2}L\right) \to \min_h .$$

Computing the derivative of this function, we conclude that the optimal step size must satisfy the equation $\Delta'(h) = hL - 1 = 0$. Thus, $h^* = \frac{1}{L}$, which is a minimum of $\Delta(h)$ since $\Delta''(h) = L > 0$. Thus, our considerations prove that one step of the Gradient Method decreases the value of the objective function at least as follows:

$$\boxed{f(\boldsymbol{x} - h^*\nabla f(\boldsymbol{x})) \leq f(\boldsymbol{x}) - \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|^2}. \tag{5}$$

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

Let us check what is going on with the other step-size strategies:

1. Constant Step Strategy:
Let $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \nabla f(\boldsymbol{x}_k)$. Then for the constant step strategy, $h_k = h$, we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq h \left(1 - \frac{1}{2} Lh\right) \|\nabla f(\boldsymbol{x}_k)\|^2 .$$

Therefore, if we choose $h_k = \frac{2\alpha}{L}$ with $\alpha \in (0,1)$, then

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{2}{L} \alpha(1-\alpha) \|\nabla f(\boldsymbol{x}_k)\|^2 .$$

Remark. Of course, the optimal choice is $h_k = \frac{1}{L}$. $(\alpha = \frac{1}{2})$

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

2. Full Relaxation Strategy:
   For the full relaxation strategy we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{1}{2L} \|\nabla f(\boldsymbol{x}_k)\|^2 ,$$

sicne the maximal decrease is not worse than the decrease attained by $h_k = \frac{1}{L}$.

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

3. Goldstein-Armijo:
Finally, for the Goldstein-Armijo rule, in view of (3), we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \leq \beta \langle \nabla f(\boldsymbol{x}_k), \ \boldsymbol{x}_k - \boldsymbol{x}_{k+1} \rangle = \beta h_k \left\| \nabla f(\boldsymbol{x}_k) \right\|^2 .$$

From (4), we obtain

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq h_k \left(1 - \frac{h_k}{2} L \right) \left\| \nabla f(\boldsymbol{x}_k) \right\|^2 .$$

Therefore, $h_k \geq \frac{2}{L}(1 - \beta)$.

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

3 Goldstein-Armijo (Continued.):
Further, using (2), we have

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \alpha \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}_{k+1} \rangle = \alpha h_k \left\| \nabla f(\boldsymbol{x}_k) \right\|^2 .$$

Combining this inequality with the previous one, we conclude that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{2}{L} \alpha (1 - \beta) \left\| \nabla f(\boldsymbol{x}_k) \right\|^2 .$$

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

Thus, we have proved that in all cases we have

$$\boxed{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \geq \frac{\omega}{L} \|\nabla f(\boldsymbol{x}_k)\|^2 , \qquad (6)}$$

where ω is some positive constant.

Now we are ready to estimate the performance of Gradient Method.

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

$$\boxed{\frac{\omega}{L}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}), \qquad (6)}$$

Summing up the inequalities (6) for $k = 0, \ldots, T$, we obtain

$$\frac{\omega}{L}\sum_{k=0}^{T}\|\nabla f(\boldsymbol{x}_k)\|^2 \leq f(\boldsymbol{x}_0) - f(\boldsymbol{x}_{T+1}) \leq f(\boldsymbol{x}_0) - f^*, \qquad (7)$$

where f^* is a lower bounds for the values of objective function in our problem. As a simple consequence of the bound (7), we have （收敛级数的通项趋于 0）：

$$\boxed{\|\nabla f(\boldsymbol{x}_k)\| \to 0 \text{ as } k \to \infty.}$$

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

However, we can also say something about the rate of convergence. Indeed, define

$$g_T^* = \min_{0 \le k \le T} g_k,$$

where $g_k = \|\nabla f(\boldsymbol{x}_k)\|$. Then, in view of (7), we come to the following inequality:

$$g_T^* \le \frac{1}{\sqrt{T+1}} \left[\frac{L}{\omega} \left(f(\boldsymbol{x}_0) - f^* \right) \right]^{1/2}. \tag{8}$$

The right-hand side of this inequality describes the rate of convergence of the sequence $\{g_T^*\}$ to zero. Note that we cannot say anything about the rate of convergence of the sequences of $\{f(\boldsymbol{x}_k)\}$ and $\{\boldsymbol{x}_k\}$.

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

Example (Example 1.2.2 of Nesterov [2003])
Consider the following function of two variables:

$$f(\boldsymbol{x}) \triangleq f(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \frac{1}{2}(\boldsymbol{x}^{(1)})^2 + \frac{1}{4}(\boldsymbol{x}^{(2)})^4 - \frac{1}{2}(\boldsymbol{x}^{(2)})^2.$$

The gradient of this function is $\nabla f(\boldsymbol{x}) = (\boldsymbol{x}^{(1)}, (\boldsymbol{x}^{(2)})^3 - \boldsymbol{x}^{(2)})^\top$. Therefore, there are only three points which can pretend to be a local minimum of this function:

$$\boldsymbol{x}_1^* = (0, 0),\ \boldsymbol{x}_2^* = (0, -1),\ \boldsymbol{x}_3^* = (0, 1).$$

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

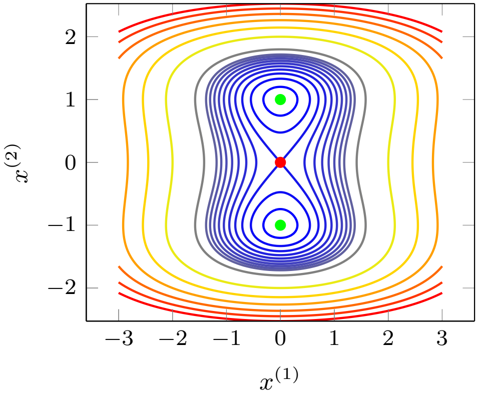
Example Continued (Example 1.2.2 of Nesterov [2003])
Computing the Hessian of this function,

$$\nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} 1 & 0 \\ 0 & 3\left(\boldsymbol{x}^{(2)}\right)^2 - 1 \end{pmatrix},$$

We conclude that $\boldsymbol{x}_2^*$ and $\boldsymbol{x}_3^*$ are isolated local minima （事实上，在我们的例子中，它们是全局解），but $\boldsymbol{x}_1^*$ is only a stationary point of our function. 确实，$f(x_1^*) = 0$，且对于足够小的 ϵ 有 $f(\boldsymbol{x}_1^* + \epsilon \boldsymbol{e}_2) = \frac{\epsilon^4}{4} - \frac{\epsilon^2}{2} < 0$。

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

$$\frac{1}{2}(x^{(1)})^2 + \frac{1}{4}(x^{(2)})^4 - \frac{1}{2}(x^{(2)})^2$$



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

Example Continued (Example 1.2.2 of Nesterov [2003])
Let us consider now the trajectory of the Gradient Method which starts at $\boldsymbol{x}_0 = (1, 0)$. Note that the second coordinate of this point is zero. Therefore, the second coordinate of $\nabla f(\boldsymbol{x}_0)$ is also zero. Consequently, the second coordinate of $\boldsymbol{x}_1$ is zero, etc.

Thus, the entire sequence of points generated by the Gradient Method will have the second coordinate equal to zero. This means that this sequence converges to $\boldsymbol{x}_1^*$. $\square$

最后，注意这种情况对于所有的一阶无约束的最小化问题是典型的。没有附加的更严格的假设，不可能保证它们能全局收敛到一个局部最小，只能保证收敛到一个静态点。

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

Consider the following problem class:

Model:	1. Unconstrained minmization. 2. $f \in C_L^{1,1}(\mathbb{R}^n)$. 3. $f(\boldsymbol{x})$ is bounded below.
Oracle	First-order Black Box
ϵ-solution:	$f(\bar{\boldsymbol{x}}) \leq f(\boldsymbol{x}_0), \ \|\nabla f(\bar{\boldsymbol{x}})\| \leq \epsilon$。

(9)

Note that inequality (8) can be used in order to obtain an upper bound for the number of steps (= calls of the oracle), which is necessary to find a point where the norm of the gradient is small.

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

For that, let us write down the following inequality:

$$g_T^* \leq \frac{1}{\sqrt{T+1}} \left[\frac{1}{\omega} L \left(f(\boldsymbol{x}_0) - f^* \right) \right]^{1/2} \leq \epsilon.$$

Therefore, if $T + 1 \geq \frac{1}{\omega\epsilon^2} \left(f(\boldsymbol{x}_0) - f^* \right)$, then we necessarily have $g_T^* \leq \epsilon$.

Thus, we can use the value $\frac{1}{\omega\epsilon^2} \left(f(\boldsymbol{x}_0) - f^* \right)$ as an upper complexity bound for our problem class (T take at most $\frac{1}{\omega\epsilon^2} \left(f(\boldsymbol{x}_0) - f^* \right)$).

Comparing this estimate with the result of Theorem (1.1.2) of Nesterov [2003] (UGM), we can see that it is much better. At least it does not depend on dimension n.

Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Theorem 1 (Theorem 2.1.14 of Nesterov [2003])

Let $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$ and $0 < h < \frac{2}{L}$. Then the Gradient Method generates a sequence of points $\{\boldsymbol{x}_k\}$, with function values satisfying the inequality:

$$f(\boldsymbol{x}_k) - f^* \leq \frac{2(f(\boldsymbol{x}_0) - f^*) \left\| \boldsymbol{x}_0 - x^* \right\|^2}{2 \left\| \boldsymbol{x}_0 - \boldsymbol{x}^* \right\|^2 + kh(2 - Lh)(f(\boldsymbol{x}_0) - f^*)}.$$

Proof. Define $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|$, $\Delta_k = f(\boldsymbol{x}_k) - f^*$ and $\omega = h(1 - \frac{L}{2}h)$. Now our problem is

$$\Delta_k \leq \frac{\Delta_0 r_0^2}{r_0^2 + \omega k \Delta_0}.$$

Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

$$\boxed{\Delta_k \le \frac{\Delta_0 r_0^2}{r_0^2 + \omega k \Delta_0}}$$

Proof. (Continued.) We have that $r_{k+1} \le r_k$ (thus $r_k \le r_0$), since

$$\begin{aligned}
r_{k+1}^2 &= \|\boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k) - \boldsymbol{x}^*\|^2 = \|(\boldsymbol{x}_k - \boldsymbol{x}^*) - h\nabla f(\boldsymbol{x}_k)\|^2 \\
&= r_k^2 - 2h\langle \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^*),\ \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + h^2 \|\nabla f(\boldsymbol{x}_k)\|^2 \quad (\nabla f(\boldsymbol{x}^*) = 0) \\
&\le r_k^2 - h(\frac{2}{L} - h) \|\nabla f(\boldsymbol{x}_k)\|^2 \left(\frac{1}{L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \le \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}),\ \boldsymbol{x} - \boldsymbol{y}\rangle\right) \\
&\le r_k^2.
\end{aligned}$$

Remark. The second last inequality comes from 2.1.8 of Nesterov [2003].

Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

$$\Delta_k \leq \frac{\Delta_0 r_0^2}{r_0^2 + \omega k \Delta_0}$$

$$0 \leq f(y) - f(x) - \langle \nabla f(x),\ y - x \rangle \leq \frac{L}{2} \|x - y\|^2, \tag{2.1.6}$$

Proof. (Continued.) In view of (2.1.6) of Nesterov [2003], we have

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{L}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \\ &= f(\boldsymbol{x}_k) - \omega \|\nabla f(\boldsymbol{x}_k)\|^2 \ \text{(Using } \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k)\text{)}, \end{aligned}$$

$$\Rightarrow \underbrace{f(\boldsymbol{x}_{k+1}) - f^*}_{\Delta_{k+1}} \leq \underbrace{f(\boldsymbol{x}_k) - f^*}_{\Delta_k} - \omega \boxed{\|\nabla f(\boldsymbol{x}_k)\|}^2. \tag{10}$$

Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

$$\boxed{\Delta_k \leq \frac{\Delta_0 r_0^2}{r_0^2 + \omega k \Delta_0}}$$

Proof. (Continued.) For $\boxed{\|\nabla f(\boldsymbol{x}_k)\|}$, we have

$$\underbrace{\Delta_k \leq \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_k - \boldsymbol{x}^* \rangle}_{\text{convex}} \leq r_k \|\nabla f(\boldsymbol{x}_k)\| \leq r_0 \|\nabla f(\boldsymbol{x}_k)\| .$$

Thus we have $\|\nabla f(\boldsymbol{x}_k)\| \geq \frac{\Delta_k}{r_0}$, which can be combined with (10) to get

$$\Delta_{k+1} \leq \Delta_k - \frac{\omega}{r_0^2} \Delta_k^2. \tag{11}$$

Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

$$\boxed{\begin{aligned} \Delta_k &\le \frac{\Delta_0 r_0^2}{r_0^2 + \omega k \Delta_0} \\ \Delta_{k+1} &\le \Delta_k - \frac{\omega}{r_0^2}\Delta_k^2 \qquad (11) \end{aligned}}$$

Proof. (Continued.) Thus, both sides of (11) are divided by $\Delta_{k+1}\Delta_k$, we obtain

$$\begin{aligned} \frac{1}{\Delta_{k+1}} &\ge \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}\frac{\Delta_k}{\Delta_{k+1}} \\ &\ge \frac{1}{\Delta_k} + \frac{\omega}{r_0^2}. \left(\text{Using } \frac{\Delta_k}{\Delta_{k+1}} \ge 1\right) \end{aligned}$$

Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

$$\boxed{\Delta_k \leq \frac{\Delta_0 r_0^2}{r_0^2 + \omega k \Delta_0}}$$

Proof. (Continued.) Summing up these inequalities, we get

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_0} + \frac{\omega}{r_0^2}(k+1).$$

$$\Rightarrow \Delta_k \leq \frac{\Delta_0 r_0^2}{r_0^2 + \omega k \Delta_0}.$$

□

Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

In order to choose the optimal step size, we need to maximize the function $\phi(h) = h(2 - Lh)$ with respect to h. The first-order optimality condition $\phi'(h) = 2 - 2Lh = 0$ provides us with the value $h^* = \frac{1}{L}$.

In this case, we get the following rate of convergence for the Gradient Method:

$$f(\boldsymbol{x}_k) - f^* \leq \frac{2L(f(\boldsymbol{x}_0) - f^*)\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 + k(f(\boldsymbol{x}_0) - f^*)}. \tag{12}$$

Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Since the right-hand side of inequality (12) is increasing in $f(\boldsymbol{x}_0) - f^*$, we obtain the following result.

Corollary 2 (Corollary of Nesterov [2003])

If $h = \frac{1}{L}$ and $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, then

$$f(\boldsymbol{x}_k) - f^* \leq \frac{2L \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{k+4}. \tag{13}$$

Remark. in view of Lemma 1.2.3 of Nesterov [2003]:

$$|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle| \leq \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2$$

and $\nabla f(\boldsymbol{x}^*) = 0$, we have $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 / (f(\boldsymbol{x}_0) - f^*) \geq 2/L$.

Another way to prove with $h^* = \frac{1}{L}$

The first line:

$$
\begin{aligned}
f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 \qquad (5)\\
&\leq \underbrace{f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}_k - x^* \rangle}_{\text{since convexity: } f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k),\ \boldsymbol{x}^* - \boldsymbol{x}_k \rangle} - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 \\
&= f(\boldsymbol{x}^*) + \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_k - \boldsymbol{x}^* - (1/L)\nabla f(\boldsymbol{x}_k)\|^2\right) \\
&= f(\boldsymbol{x}^*) + \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right)
\end{aligned}
$$

Another way to prove with $h^* = \frac{1}{L}$

$$\boxed{f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}^*) + \frac{L}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right)}$$

The second line: by suming over $k = 0, 1, 2, \ldots, T-1$, we have

$$\begin{aligned}
\sum_{k=0}^{T-1} f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) &\leq \frac{L}{2}\sum_{k=0}^{T-1}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2\right) \\
&= \frac{L}{2}\left(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2\right) \\
&\leq \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2
\end{aligned}$$

Another way to prove with $h^* = \frac{1}{L}$

$$\boxed{\sum_{k=0}^{T-1} f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq \frac{L}{2} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}$$

The third line: since $\{f(\boldsymbol{x}_k)\}$ is a non-increasing, we have

$$\begin{aligned} f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) &\leq \frac{1}{T} \sum_{k=0}^{T-1} (f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*)) \\ &\leq \frac{L}{2T} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 . \end{aligned}$$

$\square$

Performance for $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$

Theorem 3

If $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$ and $0 < h \le \frac{2}{\mu+L}$, then the Gradient Method generates a sequence $\{x_k\}$ such that

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \le \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

If $h = \frac{2}{\mu+L}$, then

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \le \left(\frac{Q_f - 1}{Q_f + 1}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\| .$$

$$f(\boldsymbol{x}_k) - f^* \le \frac{L}{2}\left(\frac{Q_f - 1}{Q_f + 1}\right)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 ,$$

where $Q_f = L/\mu$。

Performance for $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$

Proof. Let $r_k = \|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. Then

$$\begin{aligned} r_{k+1}^2 &= \|\boldsymbol{x}_k - \boldsymbol{x}^* - h\nabla f(\boldsymbol{x}_k)\|^2 \\ &= r_k^2 - 2h\langle \nabla f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}^*), \boldsymbol{x}_k - \boldsymbol{x}^* \rangle + h^2 \|\nabla f(\boldsymbol{x}_k)\|^2 \text{ Using } \nabla f(\boldsymbol{x}^*) = 0 \\ &\le \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2 + \underbrace{h\left(h - \frac{2}{\mu + L}\right)}_{\le 0} \|\nabla f(\boldsymbol{x}_k)\|^2 \le \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2. \end{aligned}$$

$$\boxed{\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \ge \frac{\mu L}{\mu + L} \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \frac{1}{\mu + L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 . \quad (2.1.24)}$$

Performance for $\mathcal{S}^{1,1}_{\mu,L}(\mathbb{R}^n)$

$$\boxed{r_{k+1}^2 \le \left(1 - \frac{2h\mu L}{\mu + L}\right) r_k^2.}$$

Proof. (Continued.) It is easy to have

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \le \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

If If $h = \frac{2}{\mu+L}$, we also have

$$\left(1 - \frac{\frac{4}{\mu+L}\mu L}{\mu + L}\right) = \left(\frac{(\mu+L)^2 - 4\mu L}{(\mu+L)^2}\right) = \frac{(\mu - L)^2}{(\mu+L)^2} = \left(\frac{Q_f - 1}{Q_f + 1}\right)^2 .$$

Performance for $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$

$$\boxed{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \le \left(\frac{Q_f - 1}{Q_f + 1}\right)^{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .}$$

Proof. (Continued) Thus, we have

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \le \left(\frac{Q_f - 1}{Q_f + 1}\right)^{k} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\| .$$

Using $f^* = 0$, and

$$0 \le f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x} \rangle \le \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 , \tag{2.1.6}$$

last inequality holds. □

Part II
General Descent Directions

Descent Directions

Formally, the descent dierection is defined as follows.

Definition 4 (Descent Direction)

$\boldsymbol{d}$ is a descent direction for f at x if $f(\boldsymbol{x}+t\boldsymbol{d}) < f(\boldsymbol{x})$ for all $t > 0$ sufficiently small.

Also , the following proposition is easy to obtain:

Propsition 5 (Descent Direction)

if f is continuously differentiable in a neighborhood of $\boldsymbol{x}$, then any $\boldsymbol{d}$ such that $\boldsymbol{d}^\top \nabla f(\boldsymbol{x}) < 0$ is a descent direction.

Descent Directions: Steepest Descent Direction

The rate of change of f at $\boldsymbol{x}$ along a vector $\boldsymbol{v} \in \mathbb{R}^n$ can be measured by the directional derivative:

$$\nabla_{\boldsymbol{v}} f(x) \triangleq \lim_{\epsilon \to 0} \frac{f(\boldsymbol{x} + \epsilon \boldsymbol{v}) - f(\boldsymbol{x})}{\epsilon} = \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}) \rangle$$

where the second equality can be verified. Essentially, we obtain the steepest descent direction of f at $\boldsymbol{x}$ by

$$\Delta_{\|\cdot\|} \boldsymbol{x} \triangleq \operatorname*{argmin}_{\boldsymbol{v}: \|\boldsymbol{v}\| \leq 1} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}) \rangle$$

where we constrain the length of $\boldsymbol{v}$ with some norm $\|\cdot\|$. We obtain various directions depending on the choice of the norm.

This section follows https://karlstratos.com/notes/descent.pdf

Descent Directions: Steepest Descent Direction

Propsition 6

Assuming $\nabla f(\boldsymbol{x}) \neq 0$, we have

$$\Delta_{\|\cdot\|_2}\boldsymbol{x} = -\left\|\nabla f(\boldsymbol{x})\right\|_2^{-1}\nabla f(\boldsymbol{x}),$$
$$\Delta_{\|\cdot\|_1}\boldsymbol{x} = -\mathrm{sign}\left(\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_l}\right)\boldsymbol{e}_l, \qquad l = \operatorname*{argmax}_{i\in 1,\ldots,n}\left|\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_i}\right|,$$
$$\Delta_{\|\cdot\|_A}\boldsymbol{x} = -\left\|\nabla f(\boldsymbol{x})\right\|_{A^{-1}}^{-1} A^{-1}\nabla f(\boldsymbol{x}),$$

where A is a symmetric and $A \succ 0$. And the A-norm $\|\cdot\|_A$ is defined by $\|\boldsymbol{v}\|_A \triangleq \sqrt{\boldsymbol{v}^\top A\boldsymbol{v}}$.

Descent Directions: Steepest Descent Direction

- Gradient Descent, the 2-norm direction of f at $\boldsymbol{x}$ is given by

$$d_{gd} \triangleq -\nabla f(\boldsymbol{x}).$$

- The 1-norm direction of f at $\boldsymbol{x}$ is given by

$$d_{cd} \triangleq -\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_l}\boldsymbol{e}_l, \quad l = \operatorname*{argmax}_{i \in 1,\ldots,n} \left| \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \right|.$$

- The A-norm direction of f at $\boldsymbol{x}$ is given by

$$d_A \triangleq -A^{-1}\nabla f(\boldsymbol{x}).$$

Descent Directions: Randomized Schemes

- For coordinate descent method, the direction of f at $\boldsymbol{x}$ can be randomly chosen, and is given by
$$d_{cd-rand} \triangleq -[\nabla f(\boldsymbol{x})]_{i_k} \boldsymbol{e}_{i_k},$$
where i_k chosen uniformly at random from $\{1, 2, \ldots, n\}$ at each k.
- For stochastic gradient method, the direction of f at $\boldsymbol{x}$ is given by
$$d_{sgc} \triangleq -g(\boldsymbol{x}_k, \xi_k),$$
where ξ_k is a random variable, such that $\mathbb{E}_{\xi_k} g(\boldsymbol{x}_k, \xi_k) = \nabla f(\boldsymbol{x}_k)$. That is, $g(\boldsymbol{x}_k, \xi_k)$ is an unbiased (but often very noisy) estimate of the true gradient $\nabla f(\boldsymbol{x}_k)$.

References I

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

Stephen J Wright and Benjamin Recht. *Optimization for data analysis*. Cambridge University Press, 2022.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Thank You!

Email:qianhui@zju.edu.cn