Mathematical Derivation of the Steepest Descent Method

1 Steepest Descent Method

The formula of the steepest descent method, also known as the gradient descent method, can be written as:

$$x_{t+1} = x_t - \eta \nabla f(x_t), \tag{1}$$

where $f(x)$ is the loss function.

2 Convergence Analysis

Choose $\eta \leq \frac{1}{L}$. Then:

$$-(1 - \frac{1}{2}L\eta) = \frac{1}{2}L\eta - 1 \tag{2}$$

which implies:

$$\leq \frac{1}{2}L\left(\frac{1}{L}\right) - 1 = -\frac{1}{2}. \tag{3}$$

Then, we obtain:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2}\eta\|\nabla f(x_t)\|^2. \tag{4}$$

Since $\|\nabla f(x_t)\|^2$ is always non-negative, this ensures that the sequence $\{f(x_t)\}$ is monotonically decreasing.

3 Convex Functions and Global Minimum

Assume that $f(x)$ is convex and attains its minimum at x^*, then for any x:

$$f(x) \geq f(x^*) + \nabla f(x^*)^T(x - x^*). \tag{5}$$

Equivalently, we write:

$$f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*). \tag{6}$$

Applying this to the gradient descent update:

$$f(x_{t+1}) \leq f(x^*) + \nabla f(x_t)^T (x_{t+1} - x^*). \tag{7}$$

Using the step-size condition:

$$-\frac{1}{2}\eta \|\nabla f(x_t)\|^2 \leq 0. \tag{8}$$

Thus, we can derive the sequence:

$$f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta} \left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 \right). \tag{9}$$

which implies that the sequence $\{f(x_t)\}$ is bounded.

4 Lipschitz Continuity

Assume that $f(x)$ is Lipschitz continuous with constant $L > 0$, i.e.,

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

Using quadratic expansion around $f(x_t)$, we obtain the inequality:

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{1}{2} L \|x_{t+1} - x_t\|^2. \tag{11}$$

Substituting $x_{t+1} = x_t - \eta \nabla f(x_t)$, we get:

$$f(x_{t+1}) \leq f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{1}{2} L\eta^2 \|\nabla f(x_t)\|^2. \tag{12}$$

Rearranging:

$$f(x_{t+1}) \leq f(x_t) - \eta \left(1 - \frac{1}{2} L\eta \right) \|\nabla f(x_t)\|^2. \tag{13}$$

Since $\|x_{t+1} - x_t\|$ must be sufficiently small, this implies that η must also be small.

5 Gradient Computation with LSE Loss

Consider the function:

$$g(w) = \|b - Aw\|^2, \tag{14}$$

which represents the Least Squares Estimation (LSE) loss.

Taking the gradient:

$$\frac{\partial g}{\partial w} = 2A^T Aw - 2A^T b. \tag{15}$$

Thus, the gradient is:

$$2A^T (Aw - b). \tag{16}$$

6 Regularization Using L_1-Norm

On the other hand, for the regularized term in L_1-norm, the gradient can be written using the sign function:

$$\text{sign}(w_i) = \begin{cases} 1, & \text{if } w_i > 0 \\ -1, & \text{if } w_i < 0 \\ 0, & \text{if } w_i = 0 \end{cases} \tag{17}$$

Thus, the total gradient becomes:

$$2A^T(Aw - b) + \lambda \text{sign}(w). \tag{18}$$