Gradient Descent for Unconstrained Optimization

1 Unconstrained Optimization Problem

We consider the following form of unconstrained optimization problem

$$\min_x f(x) \tag{1}$$

where $x \in \mathbb{R}^d$ is a parameter to be optimized.

The gradient descent method is stated in Algorithm 1.

Algorithm 1: Gradient Descent

Input: $f(\cdot)$, x_0, and $\eta_1, \eta_2, \ldots$

Output: x_T

for $t = 1, 2, \ldots, T$ **do**

  Let $x_t = x_{t-1} - \eta_t \nabla f(x_{t-1})$.

Return: x_T

It can be easily checked that gradient descent solves the following optimization problem at each step t:

$$x_t = \arg\min_x f_t(x), \qquad f_t(x) = \left[f(x_{t-1}) + \nabla f(x_{t-1})^\top (x - x_{t-1}) + \frac{1}{2\eta_t} \|x - x_{t-1}\|_2^2 \right]. \tag{2}$$

If we assume that $f(x)$ is L-smooth:

$$f(x) \le \left[f(x_{t-1}) + \nabla f(x_{t-1})^\top (x - x_{t-1}) + \frac{L}{2} \|x - x_{t-1}\|_2^2 \right],$$

and $\eta_t \le 1/L$, then the optimization problem of (2) is an upper bound of $f(x)$ that equals $f(x)$ at $x = x_{t-1}$. Therefore we have

$$f(x_t) \le f_t(x_t) \le f_t(x_{t-1}) = f(x_{t-1}).$$

This means that the gradient descent method produces a sequence that reduces the objective value. See Figure 1 for illustration.

2 Convergence for Smooth and Strongly Convex Functions

We will first present properties of smooth and strongly convex functions. Recall that an L-smooth function satisfies the condition

$$f(x) \le f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{L}{2} \|x - x_0\|_2^2.$$

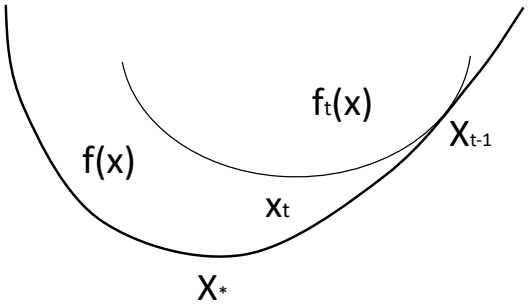

Figure 1: Illustration of Gradient Descent

An λ-strongly convex function satisfies the condition

$$f(x) \geq f(x_0) + \nabla f(x_0)^\top (x - x_0) + \frac{\lambda}{2}\|x - x_0\|_2^2.$$

We have the following results.

Proposition 1 *If $f(x)$ is L-smooth, then*

$$f(x) - \min_x f(x) \geq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2. \tag{3}$$

Proof Let $\delta = -\eta \nabla f(x)$, and use the L-smoothness definition

$$f(x + \delta) \leq f(x) + \nabla f(x)^\top \delta + \frac{L}{2}\|\delta\|_2^2,$$

we obtain

$$f(x - \eta \nabla f(x)) \leq f(x) - (\eta - 0.5\eta^2 L)\nabla f(x)^\top \nabla f(x). \tag{4}$$

Since $f(x - \eta \nabla f(x)) \geq \min_x f(x)$, we obtain the result with $\eta = 1/L$. ■

Proposition 2 *If $f(x)$ is λ-strongly convex, then*

$$f(x) - \min_x f(x) \leq \frac{1}{2\lambda}\|\nabla f(x)\|_2^2. \tag{5}$$

Proof Let x_* be the unique solution. We have from the definition of strong convexity that

$$\begin{aligned} f(x_*) \geq & f(x) + \nabla f(x)^\top (x_* - x) + \frac{\lambda}{2}\|x_* - x\|_2^2 \\ = & f(x) - \frac{1}{2\lambda}\|\nabla f(x)\|_2^2 + \frac{1}{2\lambda}\|\nabla f(x) + \lambda(x_* - x)\|_2^2 \\ \geq & f(x) - \frac{1}{2\lambda}\|\nabla f(x)\|_2^2. \end{aligned}$$

This implies the desired bound. ■

Theorem 1 *If $f(x)$ is L-smooth and λ-strongly convex, then* (1) *has a unique solution x_*. Given any fixed learning rate $\eta_t = \eta \leq 1/L$, we have*

$$[f(x_t) - f(x_*)] \leq (1 - \eta\lambda)^t [f(x_0) - f(x_*)].$$

Proof From (4) with $x = x_{t-1}$, we obtain

$$f(x_t) \leq f(x_{t-1}) - (\eta - 0.5\eta^2 L)\|\nabla f(x_{t-1})\|_2^2.$$

Using (5), we obtain

$$f(x_t) \leq f(x_{t-1}) - (\eta - 0.5\eta^2 L)2\lambda[f(x_{t-1}) - f(x_*)].$$

Since $(\eta - 0.5\eta^2 L) = 0.5\eta$, we obtain

$$[f(x_t) - f(x_*)] \leq (1 - \eta\lambda)[f(x_{t-1}) - f(x_*)] \leq \cdots \leq (1 - \eta\lambda)^t [f(x_0) - f(x_*)].$$

This proves the desired result. ■

The rate of convergence for smooth and strongly convex functions in Theorem 1 is linear convergence.

If we set $\eta = 1/L$, then the number of iterations t to achieve an ϵ-suboptimal solution

$$f(x_t) \leq f(x_*) + \epsilon$$

is

$$t = O\left(\frac{1}{\eta\lambda}\log(1/\epsilon)\right).$$

With $\eta = 1/L$, this is

$$t = O\left(\frac{L}{\lambda}\log(1/\epsilon)\right).$$

The number $\kappa = L/\lambda$ is often called the condition number, which determines how many iterations are needed for strongly convex problems to converge.

For smooth and strongly convex problems, we can also obtain the convergence of parameters using a different proof technique.

Theorem 2 *If $f(x)$ is L-smooth and λ-strongly convex, then* (1) *has a unique solution x_*. Given any fixed learning rate $\eta_t = \eta \leq 1/L$, we have*

$$\|x_t - x_*\|_2^2 \leq (1 - \eta\lambda)^t \|x_0 - x_*\|_2^2.$$

Proof We consider the following

$$\begin{aligned}
\|x_t - x_*\|_2^2 =& \|x_{t-1} - x_* - \eta \nabla f(x_{t-1})\|_2^2 \\
=& \|x_{t-1} - x_*\|^2 - 2\eta \nabla f(x_{t-1})^\top (x_{t-1} - x_*) + \eta^2 \|\nabla f(x_{t-1})\|_2^2 \\
\leq& \|x_{t-1} - x_*\|^2 + 2\eta \left[f(x_*) - f(x_{t-1}) - \frac{\lambda}{2} \|x_{t-1} - x_*\|_2^2 \right] + \eta^2 \|\nabla f(x_{t-1})\|_2^2 \\
\leq& \|x_{t-1} - x_*\|^2 + 2\eta \left[f(x_*) - f(x_{t-1}) - \frac{\lambda}{2} \|x_{t-1} - x_*\|_2^2 \right] + \eta^2 2L [f(x_{t-1}) - f(x_*)] \\
=& (1 - \eta\lambda) \|x_{t-1} - x_*\|^2 + 2\eta(1 - \eta L) \left[f(x_*) - f(x_{t-1}) \right] \\
\leq& (1 - \eta\lambda) \|x_{t-1} - x_*\|^2.
\end{aligned}$$

In the above derivation, the first inequality is due to the definition of strong convexity. The second inequality is due to (3). The third inequality is due to $(1 - \eta L) \geq 0$ and $[f(x_*) - f(x_{t-1})] \leq 0$. Using induction on t, we obtain the desired bound. ■

3 Convergence for Smooth Convex Functions

For a function that is convex and smooth but not strongly-convex, its solution may not be finite, as shown in Figure 2.

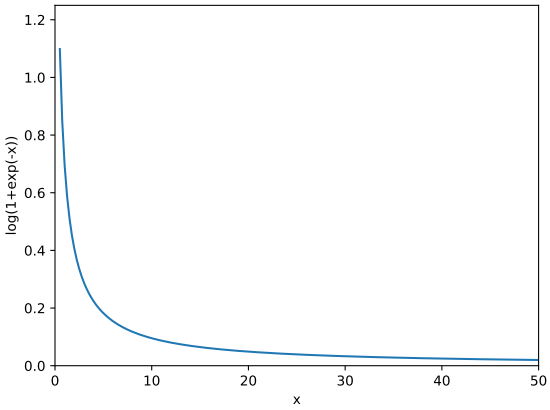

Figure 2: Solution of Smooth Convex Function

For smooth convex functions, we can not measure the convergence of x_t to x_*. Nevertheless, given any $\bar{x}$ that is finite, we have the following result.

Theorem 3 *If $f(x)$ is L-smooth and convex, then given an arbitrary finite $\bar{x}$, and a fixed learning rate $\eta_t = \eta \leq 1/L$, we have*

$$\frac{1}{T} \sum_{t=1}^{T} f(x_t) \leq f(\bar{x}) + \frac{\|x_0 - \bar{x}\|_2^2}{2\eta T}.$$

Proof The technique in this proof starts with a similar argument as that of Theorem 2 with $\lambda = 0$. We first obtain from (4) that

$$\|\nabla f(x_{t-1})\|_2^2 \leq \frac{1}{\eta - 0.5\eta^2 L}[f(x_{t-1}) - f(x_t)] \leq \frac{2}{\eta}[f(x_{t-1}) - f(x_t)].$$

Now we obtain

$$\begin{aligned}
\|x_t - \bar{x}\|_2^2 =& \|x_{t-1} - \bar{x} - \eta \nabla f(x_{t-1})\|_2^2 \\
=& \|x_{t-1} - \bar{x}\|^2 - 2\eta \nabla f(x_{t-1})^\top (x_{t-1} - \bar{x}) + \eta^2 \|\nabla f(x_{t-1})\|_2^2 \\
\leq& \|x_{t-1} - \bar{x}\|^2 + 2\eta \left[f(\bar{x}) - f(x_{t-1})\right] + \eta^2 \|\nabla f(x_{t-1})\|_2^2 \\
\leq& \|x_{t-1} - \bar{x}\|^2 + 2\eta \left[f(\bar{x}) - f(x_{t-1})\right] + 2\eta [f(x_{t-1}) - f(x_t)] \\
=& \|x_{t-1} - \bar{x}\|^2 + 2\eta \left[f(\bar{x}) - f(x_t)\right].
\end{aligned}$$

Now by summing $t = 1$ to T, we obtain

$$\|x_T - \bar{x}\|_2^2 \leq \|x_0 - \bar{x}\|_2^2 + 2\eta \sum_{t=1}^{T} \left[f(\bar{x}) - f(x_t)\right].$$

This implies the desired bound. ■

This shows that for smooth but not strongly convex functions, the gradient descent method has a slower convergence rate of $O(1/T)$ on average from $t = 1$ to $t = T$. Therefore to obtain an average sub-optimality of ϵ, we need

$$T = \frac{\|x_0 - \bar{x}\|_2^2}{2\eta} \cdot \frac{1}{\epsilon}$$

steps. This rate of convergence is sub-linear, and slower than the rate of smooth and strongly convex functions.

4 Example

We will use the following example to illustrate the convergence result.

Consider regression problem, where we have training data (x_i, y_i), where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We assume a linear model

$$y_i = x_i^\top w + \text{noise},$$

and we solve the problem for w using ridge regression:

$$\hat{w} = \arg\min_w f(w), \qquad f(w) = \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^\top w - y_i)^2 + \lambda_0 \|w\|_2^2\right].$$

Let X be the matrix with x_i being its rows ($i = 1, \ldots, n$). Let $\lambda_{\max}$ be the largest eigenvalue of $n^{-1}X^\top X$, and $\lambda_{\min}$ be the smallest eigenvalue of $n^{-1}X^\top X$, then the smoothness parameter of $f(x)$ is $L = \lambda_{\max} + \lambda_0$, and the strong-convexity parameter is $\lambda = \lambda_{\min} + \lambda_0$. If $d > n$, $\lambda_{\min} = 0$, and thus the strong-convexity parameter is λ.

With $\eta = 1/L$, the convergence of gradient descent to ϵ-optimality is

$$O\left(\frac{\lambda_{\max} + \lambda_0}{\lambda_{\min} + \lambda_0} \log \frac{1}{\epsilon}\right)$$

if $\lambda_0 > 0$. When $\lambda_0 = 0$, then the number of steps is

$$O\left((\lambda_{\max} + \lambda_0)\frac{1}{\epsilon}\right).$$