The Chinese University of Hong Kong, Shenzhen
School of Data Science

DDA5001 Supplementary Note on The Convergence Analysis of Gradient Descent

In this note, We prove the $\mathcal{O}(1/k)$ convergence rate of the gradient descent method.

1 The $\mathcal{O}(1/k)$ convergence result

Suppose our task is

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) \tag{1}$$

We apply gradient descent (GD) to problem (1), which has the form

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \mu_k \nabla \mathcal{L}(\boldsymbol{\theta}_k)} \tag{2}$$

where $\mu_k > 0$ is the stepsize. When the function $\mathcal{L}$ is convex and L-smooth (i.e., its gradient is L-Lipschitz continuous with parameter L), we have the following theorem for the convergence result.

Theorem 1. *Suppose we choose constant stepsize $\mu_k = \mu = 1/L$ in (2) and the function $\mathcal{L}$ is convex and L-smooth, then we have*

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}^\star) \leq \frac{L\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_2^2}{2k}$$

where $\boldsymbol{\theta}^\star$ is any global minimizer to (1).

Theorem 1 tells us the following:

- The convergence rate of GD on convex smooth problem is $\mathcal{O}(1/k)$.
- If we want to obtain $\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}^\star) \leq \varepsilon$, we need at most $\mathcal{O}(1/\varepsilon)$ iterations.

2 Descent lemma

Before going to the proof of Theorem 1, we derive the so-called descent lemma from the L-Lipschitz gradient.

Definition 1 (L-smoothness)**.** *h is said to be L-smooth if its gradient is L-Lipschitz, i.e.,*

$$\|\nabla h(\boldsymbol{w}) - \nabla h(\boldsymbol{u})\|_2 \leq L\|\boldsymbol{w} - \boldsymbol{u}\|_2, \quad \forall \boldsymbol{w}, \boldsymbol{u} \tag{3}$$

The following is a very famous and useful lemma in the analysis of gradient-based algorithms.

Lemma 1. *Suppose the function h is L-smooth, then we have*

$$h(\boldsymbol{w}) \leq h(\boldsymbol{u}) + \langle \nabla h(\boldsymbol{u}), \boldsymbol{w} - \boldsymbol{u} \rangle + \frac{L}{2}\|\boldsymbol{w} - \boldsymbol{u}\|^2$$

Proof. To establish this descend lemma, we define $g(t) = h(\boldsymbol{u}+t(\boldsymbol{w}-\boldsymbol{u}))$, clearly $g(0) = h(\boldsymbol{u}), g(1) = h(\boldsymbol{w})$, we have

$$\begin{aligned}
h(\boldsymbol{w}) - h(\boldsymbol{u}) = g(1) - g(0) &= \int_0^1 g'(t)dt \\
&= \int_0^1 \langle \nabla h(\boldsymbol{u} + t(\boldsymbol{w}-\boldsymbol{u})) - \nabla h(\boldsymbol{u}) + \nabla h(\boldsymbol{u}), \boldsymbol{w}-\boldsymbol{u} \rangle \, dt \\
&\leq \int_0^1 \| \nabla h(\boldsymbol{u} + t(\boldsymbol{w}-\boldsymbol{u})) - \nabla h(\boldsymbol{u}) \| \cdot \| \boldsymbol{w}-\boldsymbol{u} \| \, dt + \langle \nabla h(\boldsymbol{u}), \boldsymbol{w}-\boldsymbol{u} \rangle \\
&\leq \int_0^1 tL \| \boldsymbol{w}-\boldsymbol{u} \|^2 \, dt + \langle \nabla h(\boldsymbol{u}), \boldsymbol{w}-\boldsymbol{u} \rangle \\
&= \frac{L}{2} \| \boldsymbol{w}-\boldsymbol{u} \| + \langle \nabla h(\boldsymbol{u}), \boldsymbol{w}-\boldsymbol{u} \rangle
\end{aligned}$$

where in the last inequality we have used the L-Lipschitz gradient property (3). □

3 Proof of Theorem 1

The reason that we call Lemma 1 descent lemma is by letting $h = \mathcal{L}$ and plugging $\boldsymbol{w} = \boldsymbol{\theta}_{k+1}$ and $\boldsymbol{u} = \boldsymbol{\theta}_k$. This leads to

$$\mathcal{L}(\boldsymbol{\theta}_{k+1}) \leq \mathcal{L}(\boldsymbol{\theta}_k) + \langle \nabla \mathcal{L}(\boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle + \frac{L}{2} \| \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \|^2 \tag{4}$$

According to the GD (2), we can invoke $\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\mu \nabla \mathcal{L}(\boldsymbol{\theta}_k)$ into the above inequality, this yields

$$\mathcal{L}(\boldsymbol{\theta}_{k+1}) \leq \mathcal{L}(\boldsymbol{\theta}_k) - (1 - \frac{L\mu}{2})\mu \| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \|_2^2 \tag{5}$$

If we choose $\mu < \frac{2}{L}$, we must have

$$\mathcal{L}(\boldsymbol{\theta}_{k+1}) \leq \mathcal{L}(\boldsymbol{\theta}_k) - c \| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \|_2^2 \tag{6}$$

for some constant $c > 0$, which is also called *sufficient decrease* property, this clarifies the name descent lemma.
Taking $\mu \leq \frac{1}{L}$ gives $1 - \frac{L\mu}{2} \geq \frac{1}{2}$ and

$$\mathcal{L}(\boldsymbol{\theta}_{k+1}) \leq \mathcal{L}(\boldsymbol{\theta}_k) - \frac{\mu}{2} \| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \|_2^2 \tag{7}$$

Recall that $\mathcal{L}$ is convex, we have

$$\mathcal{L}(\boldsymbol{\theta}^\star) \geq \mathcal{L}(\boldsymbol{\theta}_k) + \langle \nabla \mathcal{L}(\boldsymbol{\theta}_k), \boldsymbol{\theta}^\star - \boldsymbol{\theta}_k \rangle \tag{8}$$

which is from first-order convexity characterization.
Combing (7) and (8) provides

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_{k+1}) &\leq \mathcal{L}(\boldsymbol{\theta}_k) - \frac{\mu}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|_2^2 \\
&\leq \mathcal{L}(\boldsymbol{\theta}^\star) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\rangle - \frac{\mu}{2}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|_2^2 \\
&= \mathcal{L}(\boldsymbol{\theta}^\star) + \frac{1}{2\mu}\Big(\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\|_2^2 - \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star - \mu\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|_2^2\Big) \\
&= \mathcal{L}(\boldsymbol{\theta}^\star) + \frac{1}{2\mu}\Big(\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\|_2^2 - \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^\star\|_2^2\Big)
\end{aligned} \tag{9}$$

Now, summing over iterations, also called *telescoping*, yields

$$\begin{aligned}
\sum_{i=1}^{k}(\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}^\star)) &\leq \frac{1}{2\mu}\Big(\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_2^2 - \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star\|_2^2\Big) \\
&\leq \frac{1}{2\mu}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_2^2
\end{aligned} \tag{10}$$

From (7), we can see $\mathcal{L}(\boldsymbol{\theta}_k)$ is decreasing, hence we finally have

$$\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}^\star) \leq \frac{1}{k}\sum_{i=1}^{k}(\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}^\star)) \leq \frac{1}{2\mu k}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|_2^2 \tag{11}$$

Plugging $\mu = \frac{1}{L}$ to the above inequality provides the desired result.