Soo Min Kwon

Notes on the Convergence of Gradient Descent

1 Definitions

Definition 1 (Taylor Series)**.** The Taylor series of a real or complex-valued function $f(y)$ that is infinitely differentiable at a value x is the power series

$$f(x) + \frac{f'(x)}{1!}(y-x) + \frac{f''(x)}{2!}(y-x)^2 + \frac{f'''(x)}{3!}(y-x)^3 + \dots \tag{1}$$

Definition 2 (Convex Function [1])**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if the domain of f is a convex set and if for all $x, y \in \mathbb{R}^n$ with $0 \leq \lambda \leq 1$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y). \tag{2}$$

Geometrically, this definition is saying that the line segment between $(x, f(x))$ and $(y, f(y))$ lies above the graph of f.

Definition 3 (First-Order Condition [1])**.** Suppose a function f is differentiable (i.e. its gradient ∇f exists at each point in **dom** f). Then f is convex if and only if **dom** f is convex and

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) \tag{3}$$

holds for all $x, y \in \mathbf{dom}\, f$.

Note that the function $f(x) + \nabla f(x)^\top (y-x)$ is the first-order Taylor approximation of f near x. The inequality in the definition above is saying that the function f is convex if the first-order Taylor approximation is a *global underestimator* of the function. If we change the inequality to $>$, then we have strong convexity.

Definition 4 (Lipschitz Continuity)**.** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous at $x \in S$, where $S \subset \mathbb{R}^n$, if there is a constant C such that

$$\|f(y) - f(x)\| \leq C\|y-x\|, \tag{4}$$

for all $y \in S$ sufficiently near x.

Definition 5 (Lipschitz Smooth)**.** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz smooth with constant C if its derivatives are Lipschitz continuous with constant C,

$$\|\nabla f(y) - \nabla f(x)\| \leq C\|y-x\| \tag{5}$$

for any $x, y \in \mathbf{dom}\, f$.

One can think of the last two definitions as a "stretching" bound. The constant C bounds the function f from growing (or stretching) too fast. This goes the same for the gradients of f for the definition of smoothness.

Definition 6 (Cauchy-Schwarz Inequality)**.** The Cauchy–Schwarz inequality states that for all vectors u and v of an inner product space,

$$|\langle u, v\rangle| \leq \|u\| \cdot \|v\|. \tag{6}$$

2 Properties

Proposition 1. A differentiable function f is convex if and only if $\mathbf{dom}\, f$ is convex and

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq 0, \tag{7}$$

for all $x, y \in \mathbf{dom}\, f$ (i.e. the gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is a **monotone** mapping.

Proof. If the function f is differentiable and convex, then we have the two inequalities:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \tag{8}$$
$$f(x) \geq f(y) + \nabla f(y)^\top (x - y), \tag{9}$$

where $x, y \in \mathbf{dom}\, f$. Combining (summing) these two inequalities gives us what we need. Note that if ∇f is monotone, then $g'(t) \geq g'(0)$ for $t \geq 0$ and $t \in \mathbf{dom}\, g$, where

$$g(t) = f(x + t(y - x)) \tag{10}$$
$$g'(t) = \nabla f(x + t(y - x))^\top (y - x). \tag{11}$$

Hence,

$$\begin{aligned} f(y) &= g(1) && (12) \\ &= g(0) + \int_0^1 g'(t) dt && (13) \\ &\geq g(0) + g'(0) && (14) \\ &= f(x) + \nabla f(x)^\top (y - x), && (15) \end{aligned}$$

which is the first-order condition for convexity.

□

Proposition 2. If a function f is Lipschitz smooth with parameter C, then from the Cauchy-Schwarz inequality, we have that

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \leq C \|x - y\|^2, \tag{16}$$

for all $x, y \in \mathbf{dom}\ f$.

Proof. The proof is quite straightforward. The Cauchy-Schwarz inequality implies that

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \leq \|\nabla f(x) - \nabla f(y)\| \cdot \|x - y\|$$

Combining this inequality with Lipschitz smoothness, we get

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \leq C \|x - y\|^2.$$

□

Proposition 3. If a **convex** function f is Lipschitz smooth with parameter C, we have that

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{C}{2} \|y - x\|^2, \tag{17}$$

for all $x, y \in \mathbf{dom}\ f$.

Proof. Recall the fundamental theorem of calculus:

$$\int_a^b h'(x)\, dx = h(b) - h(a).$$

Consider arbitrary $x, y \in \mathbf{dom}\, f$ and define $g(\tau) = f(\tau y + (1-\tau)x) = f(x + \tau(y-x))$. Then, the function $g(\tau)$ is defined for $\tau \in [0,1]$ since the $\mathbf{dom}\, f$ is convex. Then,

$$g'(\tau) = \nabla f(x + \tau(y-x))^\top (y-x) \tag{18}$$
$$g'(0) = \nabla f(x)^\top (y-x). \tag{19}$$

Subtracting these two, we get

$$\begin{aligned} g'(\tau) - g'(0) &= (\nabla f(x + \tau(y-x)) - \nabla f(x))^\top (y-x) & (20)\\ &\le \|(\nabla f(x + \tau(y-x)) - \nabla f(x))\| \cdot \|y-x\| & (21)\\ &\le C\|\tau(y-x)\| \cdot \|y-x\| & (22)\\ &= \tau C\|y-x\| \cdot \|y-x\| & (23)\\ &= \tau C\|y-x\|^2. & (24) \end{aligned}$$

We obtain the inequalities from using the Cauchy-Schwarz inequality and from the Lipschitz smoothness property. Now, integrating $g'(\tau)$ from 0 to 1,

$$\int_0^1 g'(\tau) = g(1) - g(0). \tag{25}$$

Note that $g(1) = f(y) = g(0) + \int_0^1 g'(\tau)$. Solving for $f(y)$, we have

$$\begin{aligned} f(y) &= g(0) + \int_0^1 g'(\tau) & (26)\\ &\le g(0) + g'(\tau) & (27)\\ &\le g(0) + g'(0) + \tau C\|y-x\|^2 & (28)\\ &= f(x) + \nabla f(x)^\top (y-x) + \tau C\|y-x\|^2. & (29) \end{aligned}$$

Setting $\tau = \frac{1}{2}$, we get the inequality we want. Note that we get the first inequality from using the Mean Value Theorem.

□

Proposition 4 (Gradient Properties)**.** Let f be a function that is convex, differentiable, and L-Lipschitz continuous with $L > 0$. Suppose that f^* is the optimal value of the objective function, i.e.

$$f^* = \inf_x f(\theta), \tag{30}$$

and f^* is attained at θ^*. Then, if a gradient step is given by

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t), \tag{31}$$

we have the inequalities

$$f(\theta_{t+1}) < f(\theta_t) \tag{32}$$
$$\|\theta_{t+1} - \theta^*\| < \|\theta_t - \theta^*\|. \tag{33}$$

This is a long proposition, but the idea behind it is simple. If our objective function f satisfies the properties we have been talking about thus far, then the objective function value and error between the current θ and the optimal θ decreases over consecutive iterations.

Proof. Let the gradient step be defined as

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t). \tag{34}$$

Since f is L-Lipschitz and convex, by the previous proposition, we have

$$
\begin{aligned}
f(\theta_{t+1}) &\leq f(\theta_t) + \nabla f(\theta_t)^\top (\theta_{t+1} - \theta_t) + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 && (35)\\
&= f(\theta_t) + \nabla f(\theta_t)^\top (\theta_t - \eta \nabla f(\theta_t) - \theta_t) + \frac{L}{2}\|(\theta_t - \eta \nabla f(\theta_t) - \theta_t)\|^2 && (36)\\
&= f(\theta_t) + \nabla f(\theta_t)^\top (-\eta \nabla f(\theta_t)) + \frac{L}{2}\|-\eta \nabla f(\theta_t)\|^2 && (37)\\
&= f(\theta_t) - \eta(\nabla f(\theta_t)^\top \nabla f(\theta_t)) + \frac{\eta^2 L}{2}\|\nabla f(\theta_t)\|^2 && (38)\\
&= f(\theta_t) - \eta\|\nabla f(\theta_t)\|^2 + \frac{\eta^2 L}{2}\|\nabla f(\theta_t)\|^2 && (39)\\
&= f(\theta_t) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(\theta_t)\|^2. && (40)
\end{aligned}
$$

If we let $0 \leq \eta \leq 1/L$, then

$$
\begin{aligned}
f(\theta_{t+1}) &\leq f(\theta_t) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(\theta_t)\|^2 && (41)\\
&\leq f(\theta_t) - \frac{\eta}{2}\|\nabla f(\theta_t)\|^2. && (42)
\end{aligned}
$$

From this, we can conclude that

$$
f(\theta_{t+1}) \leq f(\theta_t). \tag{43}
$$

Next, from the convexity of f, we have

$$
f(\theta_t) \leq f(\theta^*) + \nabla f(\theta_t)^\top (\theta_t - \theta^*). \tag{44}
$$

Combining the equation above with equation (42), we have

$$
\begin{aligned}
f(\theta_{t+1}) &\leq f(\theta^*) + f(\theta_t)^\top (\theta_t - \theta^*) - \frac{\eta}{2}\|\nabla f(\theta_t)\|^2 && (45)\\
&= f(\theta^*) + \frac{1}{\eta}(\theta_t - \theta_{t+1})^\top (\theta_t - \theta^*) - \frac{1}{2\eta}\|\theta_t - \theta_{t+1}\|^2 && (46)\\
&= f(\theta^*) + \frac{1}{2\eta}\|\theta_t - \theta^*\|^2 - \frac{1}{2\eta}(\|\theta_t - \theta^*\|^2 - 2(\eta \nabla f(\theta_t))^\top (\theta_t - \theta^*) + \|\eta \nabla f(\theta_t)\|^2) && (47)\\
&= f(\theta^*) + \frac{1}{2\eta}\|\theta_t - \theta^*\|^2 - \frac{1}{2\eta}\|\theta_t - \theta^* - \eta \nabla f(\theta_t)\|^2 && (48)\\
&= f(\theta^*) + \frac{1}{2\eta}\left(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right). && (49)
\end{aligned}
$$

This last equation implies that

$$
\|\theta_{t+1} - \theta^*\| < \|\theta_t - \theta^*\|. \tag{50}
$$

□

Now we have everything at our fingertips for the convergence analysis.

3 Convergence Analysis of Gradient Descent

Theorem 1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a L-Lipschitz convex, differentiable function with

$$
\theta^* = \operatorname*{argmin}_\theta f(\theta). \tag{51}
$$

Then, gradient descent with step-size $\eta \leq 1/L$ satisfies

$$f(\theta_J) \leq f(\theta^*) + \frac{\|\theta_0 - \theta^*\|^2}{2\eta J}, \tag{52}$$

where J is the total number of iterations.

Proof. Recall from Proposition 4 that

$$f(\theta_{t+1}) - f(\theta^*) \leq \frac{1}{2\eta}\left(\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right). \tag{53}$$

If we sum the equation above for $t = 0, \ldots, J-1$, we have

$$\sum_{t=0}^{J-1} f(\theta_{t+1}) - f(\theta^*) \leq \frac{1}{2\eta}\left(\|\theta_0 - \theta^*\|^2 - \|\theta_J - \theta^*\|^2\right) \tag{54}$$

$$\leq \frac{\|\theta_0 - \theta^*\|^2}{2\eta}. \tag{55}$$

This easily translates to

$$J \cdot (f(\theta_J) - f(\theta^*)) \leq \frac{\|\theta_0 - \theta^*\|^2}{2\eta} \tag{56}$$

$$(f(\theta_J) - f(\theta^*)) \leq \frac{\|\theta_0 - \theta^*\|^2}{2\eta J}. \tag{57}$$

□

The conclusion here is that the number of iterations to reach $f(\theta_J) - f(\theta^*) \leq \epsilon$.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.