Convergence Theorems for Gradient Descent

Robert M. Gower.

September 17, 2017

Abstract

Here you will find a growing collection of proofs of the convergence of gradient and stochastic gradient descent type method on convex, strongly convex and/or smooth functions. Important disclaimer: Theses notes do not compare to a good book or well prepared lecture notes. You should only read these notes if you have sat through my lecture on the subject and would like to see detailed notes based on my lecture as a reminder. Under any other circumstances, I highly recommend reading instead the first few chapters of the books [3] and [1].

1 Assumptions and Lemmas

1.1 Convexity

We say that f is convex if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in \mathbb{R}^d, \, t \in [0, \, 1]. \tag{1}$$

If f is differentiable and convex then every tangent line to the graph of f lower bounds the function values, that is

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathbb{R}^d. \tag{2}$$

We can deduce (2) from (1) by dividing and taking the limit on one dimensional function $g(t)$ then extending by setting $g(t) = f(tx + (1-t)y)$.

If f is twice differentiable, then taking a directional derivative in the v direction on the point x in (2) gives

$$0 \geq \langle \nabla f(x), v \rangle + \left\langle \nabla^2 f(x)v, y - x \right\rangle - \langle \nabla f(x), v \rangle = \left\langle \nabla^2 f(x)v, y - x \right\rangle, \quad \forall x, y, v \in \mathbb{R}^d. \tag{3}$$

Setting $y = x - v$ then gives

$$0 \leq \left\langle \nabla^2 f(x)v, v \right\rangle, \quad \forall x, v \in \mathbb{R}^d. \tag{4}$$

The above is equivalent to saying the $\nabla^2 f(x) \succeq 0$ is positive semi-definite for every $x \in \mathbb{R}^d$.

1.2 Smoothness

A differential function f is said to be L–smooth if its gradients are Lipschitz continuous, that is

$$\|\nabla f(x) - \nabla f(y)\| \quad \leq \quad L\|x - y\|. \tag{5}$$

If f is twice differentiable then we have, by using first order expansion

$$\nabla f(x) - \nabla f(x + \alpha d) = \int_{t=0}^{\alpha} \nabla^2 f(x + td)d,$$

followed by taking the norm gives

$$\| \int_{t=0}^{\alpha} \nabla^2 f(x + td)d\|_2 \leq L\alpha\|d\|_2.$$

Dividing by α

$$\frac{\| \int_{t=0}^{\alpha} \nabla^2 f(x + td)d\|_2}{\alpha} \leq L\|d\|_2,$$

then dividing through by $\|d\|$ with $d \neq 0$ and taking the limit as $\alpha \to 0$ we have that

$$\frac{\| \int_{t=0}^{\alpha} \nabla^2 f(x + td)d\|_2}{\alpha\|d\|} = \frac{\|\alpha\nabla^2 f(x)d\|_2}{\alpha\|d\|} + O(\alpha) \underset{\alpha\to 0}{=} \frac{\|\alpha\nabla^2 f(x)d\|_2}{d} \leq L, \quad \forall d \neq 0 \in \mathbb{R}^n$$

Taking the supremum over $d \neq 0 \in \mathbb{R}^d$ in the above gives

$$\nabla^2 f(x) \quad \preceq \quad LI. \tag{6}$$

Furthermore, using the Taylor expansion of $f(x)$ and the uniform bound over Hessian we have that

$$f(y) \quad \leq \quad f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \tag{7}$$

Some direct consequences of the smoothness are given in the following lemma.

Lemma 1.1 *If f is L–smooth then*

$$f(x - \tfrac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2, \tag{8}$$

and

$$f(x^*) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2, \tag{9}$$

hold for all $x \in \mathbb{R}^d$.

Proof: The first inequality (8) follows by inserting $y = x - \frac{1}{L}\nabla f(x)$ in the definition of smoothness (5) since

$$\begin{aligned} f(x - \tfrac{1}{L}\nabla f(x)) \quad &\leq \quad f(x) - \tfrac{1}{L}\langle \nabla f(x), \nabla f(x)\rangle + \frac{L}{2}\|\tfrac{1}{L}\nabla f(x)\|_2^2 \\ &= \quad f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2. \end{aligned}$$

Furthermore, by using (8) combined with $f(x^*) \leq f(y) \quad \forall y$, we get (9). Indeed since

$$f(x^*) - f(x) \leq f(x - \tfrac{1}{L}\nabla f(x)) - f(x) \leq -\frac{1}{2L}\|\nabla f(x)\|_2^2. \quad \blacksquare \tag{10}$$

1.3 Smooth and Convex

There are many problems in optimization where the function is both smooth and convex. Furthermore, such a combination results in some interesting consequences and Lemmas. Lemmas that we will then use to prove convergence of the Gradient method.

Lemma 1.2 *If $f(x)$ is convex and L–smooth then*

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2. \tag{11}$$

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\| \quad (\textit{Co-coercivity}). \tag{12}$$

Proof: To prove (11), let $z = x - \frac{1}{L}(\nabla f(x) - \nabla f(y))$. It follows that

$$\begin{aligned}
f(y) - f(x) &= f(y) - f(z) + f(z) - f(x) \\
&\overset{(2)+(5)}{\leq} \langle \nabla f(y), y - z \rangle + \langle \nabla f(x), z - x \rangle + \frac{L}{2}\|z - x\|_2^2 \\
&\overset{\text{subs. } z}{=} \langle \nabla f(y), y - x + x - z \rangle + \langle \nabla f(x), z - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\
&= \langle \nabla f(y), y - x \rangle + \langle \nabla f(y) - \nabla f(x), x - z \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \\
&\overset{\text{subs. } z}{=} \langle \nabla f(y), y - x \rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2.
\end{aligned}$$

Finally (12) follows from applying (11) once

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2,$$

then interchanging the roles of x and y to get

$$f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2.$$

Finally adding together the two above inequalities gives

$$0 \leq \langle \nabla f(y) - \nabla f(x), y - x \rangle - \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|_2^2. \quad \blacksquare$$

1.4 Strong convexity

We can "strengthen" the notion of convexity by defining μ–strong convexity, that is

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d. \tag{13}$$

Minimizing both sides of (13) in y proves the following lemma

Lemma 1.3 *If f is μ–strongly convex then it also satisfies the* Polyak- Lojasiewicz *condition, that is*

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f(x^*). \tag{14}$$

Rearranging (13) we have that

$$\langle \nabla f(x), x - y \rangle \quad \geq \quad f(x) - f(y) + \frac{\mu}{2}\|y - x\|_2^2, \tag{15}$$

Rearranging (13) and substituting $y = x^*$ we have that

$$\langle \nabla f(x), x - x^* \rangle \quad \geq \quad f(x) - f(x^*) + \frac{\mu}{2}\|x^* - x\|_2^2 \;\geq\; \frac{\mu}{2}\|x^* - x\|_2^2, \tag{16}$$

where we used that $f(x) - f(x^*) \geq 0$. The inequality (16) is of such importance in optimization that is merits its own name.

2 Gradient Descent

Consider the problem

$$x^* = \arg\min_{x \in \mathbb{R}^d} f(x), \tag{17}$$

and the following gradient method

$$x^{t+1} = x^t - \frac{1}{L}\nabla f(x^t), \tag{18}$$

where f is L–smooth. We will now prove that the iterates (23) converge. In Theorem 2.1 we will prove sublinear convergence under the assumption that f is convex. In Theorem 2.2 we will prove linear convergence (a stronger form of convergence) under the assumption that f is μ–strongly convex.

2.1 Convergence for convex and smooth functions

Theorem 2.1 *Let f be convex and L–smooth and let x^t for $t = 1, \ldots, n$ be the sequence of iterates generated by the gradient method* (23)*. It follows that*

$$f(x^n) - f(x^*) \leq \frac{2L\|x^1 - x^*\|^2}{n - 1}. \tag{19}$$

Proof: Let f be convex and L–smooth. It follows that

$$\begin{aligned} \|x^{t+1} - x^*\|_2^2 &= \|x^t - x^* - \tfrac{1}{L}\nabla f(x^t)\|_2^2 \\ &= \|x^t - x^*\|_2^2 - 2\tfrac{1}{L}\left\langle x^t - x^*, \nabla f(x^t)\right\rangle + \tfrac{1}{L^2}\|\nabla f(x^t)\|_2^2 \\ &\overset{(12)}{\leq} \|x^t - x^*\|_2^2 - \tfrac{1}{L^2}\left\langle x^t - x^*, \nabla f(x^t)\right\rangle. \end{aligned}$$

Thus if $\alpha \leq \frac{2}{L}$ then $\|x^t - x^*\|_2^2$ is a decreasing sequence in t. Calling upon (8) and subtracting $f(x^*)$ from both sides gives

$$f(x^{t+1}) - f(x^*) \quad \leq \quad f(x^t) - f(x^*) - \frac{1}{2L}\|\nabla f(x^t)\|_2^2. \tag{20}$$

Applying convexity we have that

$$\begin{aligned} f(x^t) - f(x^*) &\leq \left\langle \nabla f(x^t), x^t - x^* \right\rangle \\ &\leq \|\nabla f(x^t)\|_2 \|x^t - x^*\| \overset{(20)}{\leq} \|\nabla f(x^t)\|_2 \|x^1 - x^*\|. \end{aligned} \tag{21}$$

Isolating $\|\nabla f(x^t)\|_2$ in the above and inserting in (20) gives

$$f(x^{t+1}) - f(x^*) \overset{(20)+(21)}{\leq} f(x^t) - f(x^*) - \underbrace{\frac{1}{2L} \frac{1}{\|x^1 - x^*\|^2}}_{\beta} (f(x^t) - f(x^*))^2 \tag{22}$$

Let $\delta_t = f(x^t) - f(x^*)$. Manipulating (22) we have that

$$\delta_{t+1} \leq \delta_t - \beta \delta_t^2 \overset{\times \frac{1}{\delta_t \delta_{t+1}}}{\Leftrightarrow} \beta \frac{\delta_t}{\delta_{t+1}} \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \Leftrightarrow \beta \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}.$$

Summing up both sides over $t = 1, \ldots, n-1$ and using telescopic cancellation we have that

$$(n-1)\beta \leq \frac{1}{\delta_n} - \frac{1}{\delta_1} \leq \frac{1}{\delta_n}. \quad \blacksquare$$

2.2 Convergence for strongly convex (PL) and smooth convex functions

Now we prove some bounds that hold for strongly convex and smooth functions. In fact, if you observe, we will only use PL inequality (14) to establish the convergence result. Assuming a function satisfies the PL condition is a strictly weaker assumption then assuming strong convexity [2]. This proof is taken from [2].

Theorem 2.2 *Let f be L–smooth and μ–strongly convex. From a given $x_0 \in \mathbb{R}^d$ and $\frac{1}{L} \geq \alpha > 0$, the iterates*

$$x^{t+1} = x^t - \alpha \nabla f(x^t), \tag{23}$$

converge according to

$$\|x^{t+1} - x^*\|_2^2 \leq (1 - \alpha\mu)^{t+1} \|x^0 - x^*\|_2^2. \tag{24}$$

In particular, or $\alpha = \frac{1}{L}$ the iterates (23) enjoy a linear convergence with a rate of μ/L.

Proof: From (23) we have that

$$\begin{aligned} \|x^{t+1} - x^*\|_2^2 &= \|x^t - x^* - \alpha \nabla f(x^t)\|_2^2 \\ &= \|x^t - x^*\|_2^2 - 2\alpha \left\langle \nabla f(x^t), x^t - x^* \right\rangle + \alpha^2 \|\nabla f(x^t)\|_2^2 \\ &\overset{(15)}{\leq} (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha(f(x^t) - f(x^*)) + \alpha^2 \|\nabla f(x^t)\|_2^2 \\ &\overset{(10)}{\leq} (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha(f(x^t) - f(x^*)) + 2\alpha^2 L (f(x^t) - f(x^*) \\ &= (1 - \alpha\mu)\|x^t - x^*\|_2^2 - 2\alpha(1 - \alpha L)(f(x^t) - f(x^*)). \end{aligned} \tag{25}$$

Since $\frac{1}{L} \geq \alpha$ we have that $-2\alpha(1-\alpha L)$ is negative, and thus can be safely dropped to give

$$\|x^{t+1} - x^*\|_2^2 \leq (1-\alpha\mu)\|x^t - x^*\|_2^2.$$

It now remains to unroll the recurrence. ∎