Gradient Descent: Convex and Non Convex Case

Ben Ayad Ayoub

December 12, 2022

Contents

1 Problem Setting

The goal is to minimize a differentiable function f with $\mathrm{dom}(f) = \mathbb{R}^n$, with an L-Lipschitz continuous gradient (i.e., $\exists L > 0, \forall x, y \colon \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2,$). Using the followig iterative procedure: starting from a point $x^{(0)}$, with each $t_k \leq 1/L$:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Generically written as follows: $x^+ = x - t\nabla f(x)$, where $t \leq 1/L$.

2 Non-convex Case

In this section, we will not assume that f is convex, and still manage to prove some interesting results. **Spoiler alert**: we will show that the gradient descent reaches a point x, such that $\|\nabla f(x)\|_2 \leq \epsilon$, in $O(1/\epsilon^2)$ iterations.

Show that: $f(x^+) \leq f(x) - \left(1 - \frac{Lt}{2}\right)t\|\nabla f(x)\|_2^2$

At first, we prove this "101" property of L-lipschitz functions:

$$\forall x, y \in \mathbb{R}^n \;\; f(y) \leq f(x) + \nabla f(x)^\intercal (y - x) + \frac{L}{2}\|x - y\|^2 \qquad \textbf{(1)}$$

Let x, y be in $\mathbb{R}^n$, Let's define the function $g_{x,y} : \mathbb{R} \to \mathbb{R}$ (we'll ommit the subscripts for simplicity) as $g(t) = f(tx + (1-t)y)$. The function g has some really cool properties, to me, this one property, almost feel illegal to use:

$$g'(t) = \nabla f(tx + (1-t)y)^\intercal (x - y)$$

We can also express $f(x) - f(y)$ using g, as follows: $f(y) - f(x) = g(0) - g(1) = \int_1^0 g'(t)\mathrm{d}t$. Which yileds the following:

$$
\begin{aligned}
f(y) - f(x) &= \int_1^0 g'(t)\,\mathrm{d}t \\
&= \int_1^0 \nabla f(tx + (1-t)y)^\intercal (x - y)\,\mathrm{d}t \\
&= \int_0^1 \nabla f(tx + (1-t)y)^\intercal (y - x)\,\mathrm{d}t \\
&= \int_0^1 (\nabla f(tx + (1-t)y) - \nabla f(x) + \nabla f(x))^\intercal (y - x)\,\mathrm{d}t \\
&= \int_0^1 \nabla f(x)^\intercal (y - x)\,\mathrm{d}t + \int_0^1 \nabla (f(tx + (1-t)y) - \nabla f(x))^\intercal (y - x)\,\mathrm{d}t \\
&= \nabla f(x)^\intercal (y - x) + \int_0^1 (\nabla f(tx + (1-t)y) - \nabla f(x))^\intercal (y - x)\,\mathrm{d}t \\
&\le \nabla f(x)^\intercal (y - x) + \int_0^1 \|\nabla (f(tx + (1-t)y) - \nabla f(x))\| \|(y - x)\|\,\mathrm{d}t \\
&\le \nabla f(x)^\intercal (y - x) + \int_0^1 L\|tx + (1-t)y - x\| \|(y - x)\|\,\mathrm{d}t \\
&= \nabla f(x)^\intercal (y - x) + L\|(y - x)\|^2 \int_0^1 |t - 1|\,\mathrm{d}t \\
&= \nabla f(x)^\intercal (y - x) + \frac{L}{2}\|(y - x)\|^2
\end{aligned}
$$

We used Cauchy-Schwarz and the L-smoothness of ∇f. This completes the proof of **(1)**.

Now, by plugging $x^+ = x - t\nabla f(x)$ in the placeholder y and re-arranging, we complete our proof:

$$
\begin{aligned}
f(x^+) &\le f(x) + \nabla f(x)^\intercal (-t\nabla f(x)) + \frac{L}{2}\| - t\nabla f(x)\|^2 \\
&= f(x) - (1 - \frac{Lt}{2})t\|\nabla f(x)\|^2
\end{aligned}
$$

Prove that: $\sum_{i=0}^k \|\nabla f(x^{(i)})\|_2^2 \le \frac{2}{t}(f(x^{(0)}) - f^\star)$.

From the definition of $\{x^{(i)}\}$, the past result, and the fact that $t \le \frac{1}{L}$, we get for each $i \in \{0, \ldots, k-1\}$:

$$
\|\nabla f(x^{(i)})\|_2^2 \le \frac{2}{t}(f(x^{(i)}) - f(x^{(i+1)}))
$$

By summing each term, the RHS, cancels (Telescopes?), we get:

$$
\sum_{i=0}^{k} \|\nabla f(x^{(i)})\|_2^2 \le \frac{2}{t}(f(x^{(0)}) - f(x^{(k)})) \le \frac{2}{t}(f(x^{(0)}) - f^\star)
$$

The second inequality is obtained from (by definition), $f^\star \leq f(x^{(k)})$. Which completes the proof. Which completes the proof. Which completes the proof. Which completes the proof.

Conclude that this lower bound holds:

$$\min_{i=0,\ldots,k} \|\nabla f(x^{(i)})\|_2 \leq \sqrt{\frac{2}{t(k+1)}(f(x^{(0)}) - f^\star)},$$

We have $\forall i \in \{0, \ldots, k-1\}$: $\min_{i=0,\ldots,k} \|\nabla f(x^{(i)})\|^2 \leq \|\nabla f(x^{(i)})\|^2$. Which implies the follwing:

$$\begin{aligned}
(k+1) \min_{i=0,\ldots,k} \|\nabla f(x^{(i)})\|^2 &\leq \sum_{i=0}^{k} \|\nabla f(x^{(i)})\|_2^2 \\
&\leq \frac{2}{t}(f(x^{(0)}) - f^\star) \\
&\implies \\
\min_{i=0,\ldots,k} \|\nabla f(x^{(i)})\|^2 &\leq \frac{2}{t(k+1)}(f(x^{(0)}) - f^\star) \\
&\implies \quad \text{(Since } x \mapsto x^2 \text{ is strictly increasing on } \mathbb{R}^+\text{)} \\
\Big(\min_{i=0,\ldots,k} \|\nabla f(x^{(i)})\|\Big)^2 &\leq \frac{2}{t(k+1)}(f(x^{(0)}) - f^\star) \\
&\implies \\
\min_{i=0,\ldots,k} \|\nabla f(x^{(i)})\| &\leq \sqrt{\frac{2}{t(k+1)}(f(x^{(0)}) - f^\star)}
\end{aligned}$$

Which proves that we could achieve ϵ-substationarity in $O(1/\epsilon^2)$ iterations.

3 Convex case

Assuming now that f is convex. We prove that we can achieve ϵ-optimality in $O(1/\epsilon)$ steps of SGD, i.e., $f(x) - f^\star \leq \epsilon$

Question : Show that: $f(x^+) \leq f^\star + \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|^2$.

We have proved that: $f(x^+) \leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2$ (add link here to question b). Also, using the first-order condition of convexity, we get:

$$f(x) + \nabla f(x)^T(x^\star - x) \leq f^\star \implies f(x) \leq f^\star + \nabla f(x)^T(x - x^\star) \qquad \mathbf{[2]}$$

Which yileds:

$$f(x^+) \leq f^\star + \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|_2^2$$

Show the following: $\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \leq \frac{1}{2t}\|x^{(0)} - x^\star\|^2$.

We first prove the following:

$$f(x^+) \leq f^\star + \frac{1}{2t}\big(\|x - x^\star\|^2 - \|x^+ - x^\star\|^2\big).$$

Starting from the result we have proved in the past question, we use the generic update $t\nabla f(x) = x - x^+$, and a few arrangements to get the following:

$$\begin{aligned}
f(x^+) &\leq f^\star + \nabla f(x)^T(x - x^\star) - \frac{t}{2}\|\nabla f(x)\|^2 \\
&= f^\star + \frac{1}{2t}(2t\nabla f(x)^T(x - x^\star) - t^2\|\nabla f(x)\|^2) \\
&= f^\star + \frac{1}{2t}(2(x - x^+)^\intercal(x - x^\star) - \|x - x^+\|^2) \\
&= f^\star + \frac{1}{2t}(\|x\|^2 - 2x^\intercal x^\star + 2(x^+)^\intercal x^\star - \|x^+\|^2) \\
&= f^\star + \frac{1}{2t}((\|x\|^2 - 2x^\intercal x^\star + \|x^*\|^2) - (\|x^*\|^2 - 2(x^+)^\intercal x^\star + \|x^+\|^2)) \\
&= f^\star + \frac{1}{2t}(\|x - x^*\|^2 - \|x^+ - x^*\|^2)
\end{aligned}$$

By summing the inequalities for each $i \in \{0, \ldots, k-1\}$, the RHS "telescopes", we are left with: $\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \leq \frac{1}{2t}(\|x^{(0)} - x^\star\|^2 - \|x^{(k)} - x^\star\|^2)$
Since $0 \leq \|x^{(k)} - x^\star\|^2$, we upper bound the RHS to complete the proof.

Conclude that: $f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|^2}{2tk}$,

Every update from $x^{(i)}$ to $x^{(i+1)}$ makes the gap $f(x^{(i)}) - f^\star$ smaller (see first question), and hence, $f(x^{(k)}) - f^\star = \min\{f(x^{(i)}) - f^\star\}_{i=1}^{i=k}$, which yields, using what we established in the past question:

$$\begin{aligned}
k(f(x^{(k)}) - f^\star) &\leq \sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \\
&\leq \frac{1}{2t}\|x^{(0)} - x^\star\|^2.
\end{aligned}$$

Dividing by k completes the proof. This proves the $O(1/\epsilon)$ rate for achieving ϵ-suboptimality.