TD: Gradient Descent for convex and smooth functions

Faïcel Chamroukhi

week 4-5 - Nov. 28 (lecture). Dec 05. 2024

Convergence Analysis

We study the convergence for a fixed step size α. Prove the following result.

Theorem Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth. If x^* is a critical point of f, i.e., $\nabla f(x^*) = 0$, then the the sequence $\{x^{(k)}\}$ generated by gradient descent

$$x^{(k+1)} = x^{(k)} + \alpha \nabla f(x^{(k)}),$$

with fixed step size $0 \leq \alpha \leq \frac{1}{L}$ satisfies:

$$f(x^{(k)}) - f(x^\star) \leq \frac{\|x^{(0)} - x^\star\|^2}{2\alpha k}.$$

Theorem Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth. If x^* is a critical point of f, i.e., $\nabla f(x^*) = 0$, then the the sequence $\{x^{(k)}\}_{k=0}^{\infty}$ generated by a gradient descent

$$x^{(k+1)} = x^{(k)} + \alpha \nabla f(x^{(k)}),$$

with fixed step size $\alpha \leq \frac{1}{L}$ satisfies:

$$f(x^{(k)}) - f(x^\star) \leq \frac{\|x^{(0)} - x^\star\|^2}{2\alpha k}.$$

i.e., This implies that gradient descent has a convergence rate of $O\left(\frac{1}{k}\right)$.
i.e., To achieve $f(x^{(k)}) - f(x^\star) \leq \epsilon$, we need $O\left(\frac{1}{\epsilon}\right)$ iterations.

Proof: Using the smoothness property, we can write:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2 \quad \text{for any } x, y$$

Proving this property:

- Since f is L-smooth, then ∇f is L-Lipschitz continuous, this means there exists a constant $L > 0$ such that

$$\nabla^2 f \preceq LI, \quad \text{or equivalently,} \quad \nabla^2 f(z) - LI \preceq 0$$

- i.e., $\nabla^2 f(z) - LI$ is semi-definite negative, which means $\forall x, y, z$ we have:

$$(x - y)^\top (\nabla^2 f(z) - LI)(x - y) \leq 0$$

  which means:

$$(x - y)^\top \nabla^2 f(z)(x - y) = (x - y)^\top \nabla^2 f(z)(x - y) - L\|x - y\|^2 \leq 0$$

  Rearranging this inequality, we get the bound:

$$(x - y)^\top \nabla^2 f(z)(x - y) \leq L\|x - y\|^2$$

- Based on Taylor's Remainder Theorem, we have $\forall x, y, \exists z \in [x, y]$:

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(x - y)^\top \nabla^2 f(z)(x - y)$$

  where $\nabla f(x)$ is the gradient of f at x, $\nabla^2 f(z)$ is the Hessian matrix of f evaluated at some intermediate point $z \in [x, y]$, and the notation $z \in [x, y]$ (i.e., z lies on the line segment between x and y, i.e., $z = x + t(y - x)$ for some $t \in (0, 1)$).

- Substituting the bound from the previous step into Taylor's expansion, we get:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2 \quad \text{for any } x, y$$

- Plugging in $y = x^{(k+1)}$ and $x = x^{(k)}$ with $x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)})$. To simplify notation, let's use $x^+ = x - \alpha \nabla f(x)$:

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)^\top (x^+ - x) + \frac{L}{2} \|x^+ - x\|^2 \\ &= f(x) + \nabla f(x)^\top (x - \alpha \nabla f(x) - x) + \frac{L}{2} \|x - \alpha \nabla f(x) - x\|^2 \\ &= f(x) - \alpha \nabla f(x)^\top \nabla f(x) + \frac{L}{2} \alpha^2 \|\nabla f(x)\|^2 \\ &= f(x) - \left(1 - \frac{L\alpha}{2}\right) \alpha \|\nabla f(x)\|^2 \end{aligned}$$

- Taking $0 < \alpha \leq \frac{1}{L}$, we have $1 - \frac{L\alpha}{2} \geq \frac{1}{2}$. Therefore:

$$f(x^+) \leq f(x) - \frac{\alpha}{2} \|\nabla f(x)\|^2.$$

- Since f is convex, $f(x) \leq f(x^*) + \nabla f(x)^\top (x - x^*)$, we have:

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{\alpha}{2} \|\nabla f(x)\|^2 \\ &\leq f(x^\star) + \nabla f(x)^\top (x - x^\star) - \frac{\alpha}{2} \|\nabla f(x)\|^2 \\ &= f(x^\star) + \frac{1}{2\alpha} \left(2\alpha \nabla f(x)^\top (x - x^\star) - \alpha^2 \|\nabla f(x)\|^2\right) \end{aligned}$$

- using the fact that $2\alpha \nabla f(x)^\top (x - x^\star) - \alpha^2 \|\nabla f(x)\|^2$ is a part of a remarkable identity $\|a - b\|^2 = \|a\|^2 - 2a^\top b + \|b\|^2$ where

$$a = x - x^\star, \quad b = \alpha \nabla f(x),$$

since $\|x - x^\star - \alpha \nabla f(x)\|^2 = \|x - x^\star\|^2 - 2\alpha (x - x^\star)^\top \nabla f(x) + \alpha^2 \|\nabla f(x)\|^2$. Then we have

$$2\alpha (x - x^\star)^\top \nabla f(x) - \alpha^2 \|\nabla f(x)\|^2 = \|x - x^\star\|^2 - \|x - x^\star - \alpha \nabla f(x)\|^2.$$

- The previous inequality becomes

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{\alpha}{2} \|\nabla f(x)\|^2 \\ &\leq f(x^\star) + \frac{1}{2\alpha} \left(\|x - x^\star\|^2 - \|x - x^\star - \alpha \nabla f(x)\|^2\right) \\ &= f(x^\star) + \frac{1}{2\alpha} \left(\|x - x^\star\|^2 - \|x^+ - x^\star\|^2\right) \end{aligned}$$

and we finally get

$$f(x^+) - f(x^\star) \leq \frac{1}{2\alpha} \left(\|x - x^\star\|^2 - \|x^+ - x^\star\|^2\right)$$

- This inequality holds for x^+ on every iteration of gradient descent. Summing over iterations, we have:

$$\begin{aligned}\sum_{i=1}^{k}\left(f(x^{(i)})-f(x^*)\right) &\leq \sum_{i=1}^{k}\frac{1}{2\alpha}\left(\|x^{(i-1)}-x^*\|_2^2-\|x^{(i)}-x^*\|_2^2\right)\\ &\overset{\text{telescoping series}}{=} \frac{1}{2\alpha}\left(\|x^{(0)}-x^*\|_2^2-\|x^{(k)}-x^*\|_2^2\right)\\ &\leq \frac{1}{2\alpha}\left(\|x^{(0)}-x^*\|_2^2\right)\end{aligned}$$

So we obtain:

$$\sum_{i=1}^{k}\left(f(x^{(i)})-f(x^*)\right)\leq\frac{1}{2\alpha}\|x^{(0)}-x^*\|_2^2$$

- Since $f(x^{(k)})$ is nonincreasing,

$$kf(x^{(k)})\leq\sum_{i=1}^{k}f(x^{(i)})$$

which implies

$$k(f(x^{(k)})-f(x^\star))\leq\sum_{i=1}^{k}(f(x^{(i)})-f(x^\star)),$$

equivalently,

$$f(x^{(k)})-f(x^\star)\leq\frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)})-f(x^\star)).$$

Thus:

$$f(x^{(k)})-f(x^\star)\leq\frac{1}{k}\sum_{i=1}^{k}\left(f(x^{(i)})-f(x^\star)\right)\leq\frac{\|x^{(0)}-x^\star\|^2}{2\alpha k}$$

We then finally have:

$$f(x^{(k)})-f(x^\star)\leq\frac{\|x^{(0)}-x^\star\|^2}{2\alpha k}.$$

appendix
Telescoping Series:
To understand why

$$\sum_{i=1}^{k} \frac{1}{2\alpha} \left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2 \right) = \frac{1}{2\alpha} \left(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2 \right)$$

let's expand the summation to observe the telescoping effect:

$$\sum_{i=1}^{k} \frac{1}{2\alpha} \left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2 \right)$$

Expanding this explicitly, we have:

$$\frac{1}{2\alpha} \left(\|x^{(0)} - x^*\|_2^2 - \|x^{(1)} - x^*\|_2^2 \right) + \frac{1}{2\alpha} \left(\|x^{(1)} - x^*\|_2^2 - \|x^{(2)} - x^*\|_2^2 \right) + \cdots$$

$$+ \frac{1}{2\alpha} \left(\|x^{(k-1)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2 \right)$$

Notice that most intermediate terms **cancel**:

- The term $\|x^{(1)} - x^*\|_2^2$ appears as a positive value in the first part and cancels with the negative value in the next part.
- Similarly, the term $\|x^{(2)} - x^*\|_2^2$ cancels out, and this pattern continues.

Thus, the only terms that do not cancel are the **first** term $\|x^{(0)} - x^*\|_2^2$ and the **ast** negative term $-\|x^{(k)} - x^*\|_2^2$, which results in:

$$\frac{1}{2\alpha} \left(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2 \right)$$