1 Lecture 3: Convergence of gradient descent algorithm

Gradient Descent (GD): It is an iterative method. The optimization procedure proceeds in iterations, each improving the objective value. The basic method amounts to iteratively moving the current point in the direction of the gradient, which is a linear operation if the gradient is given explicitly!
Unconstrained Gradient Descent Algorithm with constant step size
Input: Time horizon: T, initial point: x_0, step size: a
for $t = 0, \dots, T-1$ **do**

$$x_{t+1} = x_t - a\nabla f(x_t) \tag{1.1}$$

end for The idea of GD is to takes step in the direction of the steepest descent, which is $\nabla f(x)$. The following statement proves the convergence to a global solution of the gradient descent algorithm when the objective f is convex and Lipschitz continuous.

Theorem 1.1 *Suppose the function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, differentiable, and its gradient is Lipschitz continuous, i.e., there exists $L > 0$ such that for any x, y $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$. Then if we run gradient descent algorithm for k iterations with a fixed step size $a = 1/L$ it will yield a solution $f^{(k)}$ which satisfies:*

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|_2^2}{2ak} \leq \frac{L}{2k}\|x^{(0)} - x^*\|_2^2,$$

where $f(x^*)$ is the optimal value. Intuitively, this means that GD is guaranteed to converge and it converges with sublinear rate $\mathcal{O}(1/k)$.

Proof. Since f is Lipschitz and convex we can perform a convexity second order approximation of f around $f(x)$:

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}\nabla^2 f(x)\|y - x\|_2^2 \\ &\leq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}L\|y - x\|_2^2. \end{aligned} \tag{1.2}$$

In the second line we used the gradient Lipschitzness, i.e., $\nabla^2 f(x) \leq LI$. Now we plug in the GD update **??** by letting

$$y = x_{t+1} = x_t - a\nabla f(x_t)$$

Then

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^\top (x_{t+1} - x_t) + \frac{1}{2}L\|x_{t+1} - a\nabla f(x_t) - x_t\|_2^2 \\ &= f(x_t) + \nabla f(x_t)^\top (x_t - a\nabla f(x_t) - x_t) + \frac{1}{2}L\|a\nabla f(x_t)\|_2^2 \\ &= f(x_t) - a\|\nabla f(x_t)\|_2^2 + \frac{1}{2}La^2\|\nabla f(x_t)\|_2^2 \\ &= f(x_t) - (1 - \frac{1}{2}La)a\|\nabla f(x_t)\|_2^2 \\ &= f(x_t) - \frac{1}{2}a\|\nabla f(x_t)\|_2^2. \end{aligned} \tag{1.3}$$

To sum up we get the following inequality

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2}a\|\nabla f(x_t)\|_2^2. \tag{1.4}$$

Remark 1.2. $\|\nabla f(x_t)\|_2^2$ will always be positive or zero if we reach the optimal. This implies that f wil strictly decrease at each iteration of GD until it reaches the optimal value $f(x) = f(x^*)$. This convergence result holds only if we choose step size small enough, i.e., smaller or equal to $1/L$. This explains why we observe in practice that GD deiverges if the step size is too large!

The second step of the proof requires to bound $f(x_{t+1})$ in terms of $f(x^*)$. Since f is convex Then

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^\top (x^* - x) \\ f(x) &\leq f(x^*) + \nabla f(x)^\top (x - x^*). \end{aligned} \tag{1.5}$$

We plug the last inequality into (**??**) and we get

$$\begin{aligned} f(x_{t+1}) &\leq f(x^*) + \nabla(x)^\top (x - x^*) = \frac{a}{2}\|\nabla f(x)\|_2^2 \\ f(x_{t+1}) - f(x^*) &\leq \frac{1}{2a}\left(2a\nabla f(x)^\top (x - x^*) - a^2\|\nabla f(x)\|_2^2\right). \end{aligned} \tag{1.6}$$

So we get

$$\begin{aligned} f(x_{t+1}) - f(x^*) &\leq \frac{1}{2a}\left(2a\nabla f(x)^\top (x - x^*) - a^2\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \\ &\leq \frac{1}{2a}\left(\|x - x^*\|_2^2 - \|x - a\nabla f(x) - x^*\|_2^2\right) \end{aligned} \tag{1.7}$$

The last inequality holds because of

$$\|x - a\nabla f(x) - x^*\|_2^2 = \|x - x^*\|_2^2 - 2a\nabla f(x)^\top (x - x^*) + a^2\|\nabla f(x)\|_2^2.$$

By the definition of GD update **??** and plugging the update into (**??**) we have

$$f(x_{t+1}) - f(x^*) \leq \frac{1}{2a}\left(\|x_t - x^*\|_2^2 + \|x_{t+1} - x^*\|_2\right.$$

2.(1.8)This inequality holds for every iteration of GD. The final step of the proof is to sum over all iterations. Henceforth, we obtain

$$\begin{aligned} \sum_{i=0}^{k} f(x^{(i)}) - f(x^*) &\leq \sum_{i=1}^{k} \frac{1}{2a}\left(\|x^{(i-1)} - x^*\|_2^2 + \|x^{(i)} - x^*\|_2^2\right) \\ &= \frac{1}{2a}\left(\|x^{(0)} - x^*\|_2^2 + \|x^{(k)} - x^*\|_2^2\right) \\ &\leq \frac{1}{2a}\|x^{(0)} - x^*\|_2^2. \end{aligned} \tag{1.9}$$

The sum in the right-hand side of the first line is a telescoping sum. So all the terms will cancel apart from the first term and the last one. Since we know that f is decreasing at each iteration, we conclude that

$$\begin{aligned} f(x^{(k)}) - f(x^*) &\leq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) \\ &\leq \frac{1}{2ak}\|x^{(0)} - x^*\|_2^2. \end{aligned} \tag{1.10}$$

This means that the GD updates will converge to the solution at a sublinear rate $\mathcal{O}(1/k)$. This concludes the proof. □

This theorem assumes that we know the Lipschitz constant L of the gradient beforehand. In practice the gradient constant is not known in most of the times. However, there exists the Armijo rule, which is capable of adjusting the step size adequetely without knowing the Lipschitz constant.