Rate of Convergence of Gradient Descent

Orestis Plevrakis

Finishing what we left

Last time we showed that the function decreases in each step:

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

If we knew that $\|\nabla f(x_t)\|$ was always large, we would be in good shape, because we would decrease by a lot in each step. What if $\|\nabla f(x_t)\|$ becomes small though? Let's take the extreme case: $\|\nabla f(x_t)\| = 0$, i.e., $\nabla f(x_t) = 0$. Well, we have seen in class that this implies that we reached the optimal value, so this a perfect scenario! We showed this using that at x_*, the function f is above its linearization (at x_t), i.e.,

$$f(x_*) \geq f(x_t) + \nabla f(x_t) \cdot (x_* - x_t)$$

We will use this again here. By rearranging, we get

$$f(x_t) - f(x_*) \leq -\nabla f(x_t) \cdot (x_* - x_t) \leq \|\nabla f(x_t)\| \cdot \|x_t - x_*\| \leq \|\nabla f(x_t)\| \cdot \|x_1 - x_*\|$$

where in the second-to-last step, we used Cauchy-Scharz, and in the last step, we used Problem 4. The resulting inequality: $f(x_t) - f(x_*) \leq \|\nabla f(x_t)\| \cdot \|x_1 - x_*\|$ tells us something very helpful: if $\|\nabla f(x_t)\|$ is small, then the current value is close to optimal! By replacing in (1), we get

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\frac{(f(x_t) - f(x_*))^2}{\|x_1 - x_*\|^2} \tag{2}$$

Subtracting $f(x_*)$ from both sides, letting $C := 2\beta\|x_1 - x_*\|^2$ and $\Delta_t := f(x_t) - f(x_*)$ we get

$$\Delta_{t+1} \leq \Delta_t - \Delta_t^2/C \tag{3}$$

Before analyzing how quickly Δ_t goes to zero, we first derive a bound on Δ_1 in terms of β and $\|x_1 - x_*\|$: by convexity again, we have $\Delta_1 \leq \nabla f(x_1) \cdot (x_1 - x_*) \leq \|\nabla f(x_1)\|\|x_1 - x_*\|$. Now, from Theorem 5 in Lecture 5, we get $\|\nabla f(x_1)\| \leq \beta\|x_1 - x_*\|$, and so $\Delta_1 \leq \beta\|x_1 - x_*\|^2 = C/2$.

The recursive inequality

How can we analyze the recursive inequality (3)? Let's rewrite it as $\Delta_{t+1} - \Delta_t \leq \Delta_t^2/C$. If Δ_t is large we decrease by a lot. If it is very small, we are good. Here is a way to translate this intuition into a proof of speed of convergence: bound the number of steps to go from Δ_1 to $\Delta_1/2$, then from $\Delta_1/2$ to $\Delta_1/4$, then from $\Delta_1/4$ to $\Delta_1/8$ etc. Let $k \geq 1$ be an integer. Suppose that the decreasing sequence $\Delta_1, \Delta_2, \Delta_3, \ldots$ spends τ_k steps inside the interval $(\Delta_1/2^k, \Delta_1/2^{k-1}]$. Then,

$$(\tau_k - 1) \cdot \frac{1}{C} \cdot \left(\frac{\Delta_1}{2^k}\right)^2 \leq \frac{\Delta_1}{2^k}$$

(why?). So, $\tau_k \le 1 + C \cdot 2^k/\Delta_1$. Let $\epsilon > 0$. We want to bound the number of steps it takes for Δ_t to become smaller or equal than ϵ. Let $k_* := \lceil \log_2 \frac{\Delta_1}{\epsilon} \rceil$. Then, we will start having $\Delta_t \le \epsilon$ after at most $\sum_{k=1}^{k_*} \tau_k$ steps, and

$$\sum_{k=1}^{k_*} \tau_k \le k_* + \frac{C}{\Delta_1} \sum_{k=1}^{k_*} 2^k = k_* + \frac{C}{\Delta_1}(2^{k_*+1} - 1) = O\left(\frac{C}{\epsilon}\right).$$

Since $C = \beta \|x_1 - x_*\|$, we are done.