Advanced Algorithms: Solution of Problem 4

Comment. By no means your solutions are expected to be as long as the ones I am providing. Mine are long because I describe the discovery process.

I will give two solutions.

Solution 1

We will employ the 2^{nd} and the 4^{th} strategy (special cases and formulating questions).

Special case: Quadratics

Let $f(x) = (x-x_*)^\top A(x-x_*)$ for some $A \succcurlyeq 0$, $x_* \in \mathbb{R}^n$. The point x_* is the minimum (why?). Note that $\nabla f(x) = 2A(x-x_*), \nabla^2 f(x) = 2A$. The condition of β-smoothness means that $\lambda_{\max}(2A) \le \beta$, i.e., $\lambda_{\max}(A) \le \beta/2$. The GD update is $x_{t+1} = x_t - \frac{1}{\beta}2A(x_t - x_*)$, and thus

$$\|x_{t+1} - x_*\| = \left\|(x_t - x_*) - \frac{2}{\beta}A(x_t - x_*)\right\| = \left\|\left(I - \frac{2}{\beta}A\right)(x_t - x_*)\right\| \le \left\|I - \frac{2}{\beta}A\right\|_2 \|x_t - x_*\| \quad (1)$$

where the last step follows from HW2, Problem 1.5. The matrix $M := I - \frac{2}{\beta}A$ is symmetric, so from HW2, Problem 1.6, we have $\|M\|_2 = \max(|\lambda_{\max}(M)|, |\lambda_{\min}(M)|)$. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. We know that all $\lambda_i \in [0, \beta/2]$. Now, from HW2, Problem 1.1, the eigenvalues of M are $1 - \frac{2}{\beta}\lambda_1, \ldots, 1 - \frac{2}{\beta}\lambda_n$. Since $\lambda_i \in [0, \beta/2]$ implies that $\left|1 - \frac{2}{\beta}\lambda_i\right| \le 1$, we are done.

The General Case

Here $x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$ and

$$\|x_{t+1} - x_*\| = \left\|(x_t - x_*) - \frac{1}{\beta}\nabla f(x_t)\right\|$$

Now we need to ask

- **Q**: What was the key fact that allowed us to solve the special case?
- **A**: First of all, we had

$$\nabla f(x_t) = B(x_t - x_*) \quad (2)$$

  where B was a matrix, and thus we could factorize. Second, we had bounds on the eigenvalues of B.
- **Q**: Do we know an analog of (2) for general functions?

- **A**: Yes! The fundamental theorem of calculus (FTC):

$$\forall x, y, \quad \nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x_\tau) d\tau (y - x) \tag{3}$$

  where $x_\tau := (1-\tau)x + \tau y$. For $y = x_t$ and $x = x_*$, we get $\nabla f(x_t) = \int_0^1 \nabla^2 f(x_\tau) d\tau \ (x_t - x_*)$.

So now, we get

$$\begin{aligned} \|x_{t+1} - x_*\| = \left\| (x_t - x_*) - \frac{1}{\beta} \int_0^1 \nabla^2 f(x_\tau) d\tau (x_t - x_*) \right\| &= \left\| \left(I - \frac{1}{\beta} \int_0^1 \nabla^2 f(x_\tau) d\tau \right) (x_t - x_*) \right\| \\ &\leq \left\| I - \frac{1}{\beta} \int_0^1 \nabla^2 f(x_\tau) d\tau \right\|_2 \|x_t - x_*\| \end{aligned}$$

and observe that the matrix $H := \int_0^1 \nabla^2 f(x_\tau) d\tau$ is symmetric (why?). Now, if we show that all eigenvalues of H lie inside the interval $[0, \beta]$, we can just repeat the steps of the special case and finish the proof.

Claim 1. *All the eigenvalues of H lie inside the interval $[0, \beta]$.*

Proof. From HW2, Problem 1.4, it suffices to show that for all $v \in \mathbb{R}^n$ with $\|v\| = 1$, we have $v^\top H v \in [0, \beta]$. But, $v^\top \int_0^1 \nabla^2 f(x_\tau) d\tau \ v = \int_0^1 v^\top \nabla^2 f(x_\tau) \ v \ d\tau$. This is because

$$\begin{aligned} v^\top \int_0^1 \nabla^2 f(x_\tau) d\tau \ v &= \sum_{i,j} \left(\int_0^1 \nabla^2 f(x_\tau) d\tau \right)_{ij} v_i v_j = \sum_{i,j} \int_0^1 \frac{\partial^2 f}{\partial x_i \partial x_j} (x_\tau) d\tau \ v_i v_j \\ &= \int_0^1 \sum_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j} (x_\tau) \ v_i v_j d\tau = \int_0^1 v^\top \nabla^2 f(x_\tau) \ v \ d\tau \end{aligned}$$

From convexity and β-smoothness, $v^\top \nabla^2 f(x_\tau) \ v \in [0, \beta]$, for all unit vectors v. This completes the proof. □

Thus, we have proven that $\|x_{t+1} - x_*\| \leq \|x_t - x_*\|$.

Note. One more special case that could point you towards using FTC is the case $n = 1$. There, $x_{t+1} = x_t - \frac{1}{\beta} f'(x_t)$. The convexity and β-smoothness mean that $0 \leq f''(x) \leq \beta$ for all x. We need to connect f' and f''. How to do it? FTC!

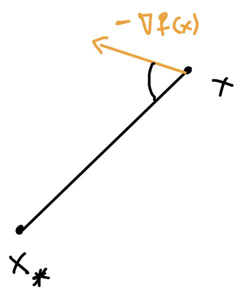
The techniques we just saw are very important and worth knowing. However, there is a shorter solution:

Solution 2

The convexity criterion: $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$ has an important consequence: for $y = x_*$, we get

$$(-\nabla f(x)) \cdot (x_* - x) \geq f(x) - f(x_*) \tag{4}$$

If $f(x) > f(x_*)$, we have $(-\nabla f(x)) \cdot (x_* - x) > 0$ which means that this angle is acute:



and this implies that if we start from x and we take a small enough step in the direction of $-\nabla f(x)$, we will get closer to x_*. The question now is, is the step $\eta = 1/\beta$ small enough? Back to GD,

$$\|x_{t+1} - x_*\|^2 = \|x_t - \frac{1}{\beta}\nabla f(x_t) - x_*\|^2 = \|x_t - x_*\|^2 - \frac{2}{\beta}\nabla f(x_t) \cdot (x_t - x_*) + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2$$

So, it suffices to prove that $-\frac{2}{\beta}\nabla f(x_t) \cdot (x_t - x_*) + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2 \leq 0$, i.e.,

$$(-\nabla f(x_t)) \cdot (x_* - x_t) \geq \frac{1}{2\beta}\|\nabla f(x_t)\|^2$$

From (4), it suffices to show that $f(x_t) - f(x_*) \geq \frac{1}{2\beta}\|\nabla f(x_t)\|^2$. But, we know that $f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|^2$, and since $f(x_{t+1}) \geq f(x_*)$ we are done. Note that the last argument says that since in the next step, the value decreases by $G_t := \frac{1}{2\beta}\|\nabla f(x_t)\|^2$, the maximum possible decrease: $f(x_t) - f(x_*)$ will be at least G_t.