MIT 6.7220/15.084 — Nonlinear Optimization Spring 2025

Problem Set 4

Due: Apr 15th 2025 11:59pm ET

Collaboration policy. We encourage working together whenever possible: in the recitations, problem sets, and general discussion of the material and assignments. Keep in mind, however, that for the problem sets the solutions you hand in should reflect your own understanding of the class material, and should be written solely by you. It is not acceptable to copy (in whole or in part) a solution that somebody else has written.

1. Strong convexity and PŁ condition [25pts]

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Suppose f is μ-strongly convex. As you have shown in the past, μ-strong convexity is equivalent to the condition

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2 \quad \forall x, y \in \mathbb{R}^n.$$

(1.a) [7pts] Show that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y \in \mathbb{R}^n$$

▸ *Hint:* Consider the function $g(z) \coloneqq f(z) - \langle \nabla f(x), z \rangle$.

Solution. Observe that due to

$$\left\| \sqrt{\frac{\mu}{2}}(x - y) - \sqrt{\frac{1}{2\mu}}(\nabla f(x) - \nabla f(y)) \right\|^2 \geq 0,$$

we have the following:

$$\frac{\mu}{2}\|x - y\|^2 + \langle x - y, \nabla f(y) - \nabla f(x) \rangle \geq -\frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|^2.$$

Let $g(z) \coloneqq f(z) - \langle \nabla f(x), z \rangle$, this function is μ strongly convex so we must have that

$$g(x) \geq g(y) + \langle \nabla g(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2,$$

substituting back the definition of g yields

$$f(x) - \langle \nabla f(x), x \rangle \geq f(y) - \langle \nabla f(x), y \rangle + \langle \nabla f(y) - \nabla f(x), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2.$$

plugging the equation at the beginning yields:

$$f(x) - \langle \nabla f(x), x \rangle \geq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|_2^2.$$

reordering results in

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|_2^2 \geq f(y).$$

◂

(1.b) [10pts] Use Problem (1.a) to show that if f is strongly convex, then it satisfies the PŁ condition.

Solution. By the last question of the midterm, we have that a minimizer x^* exists and is unique. By the first order optimility we have that $\nabla f(x^*) = 0$. Using $x := x^*$ and $y := x$ we get

$$f(x^*) + \langle 0, x - x^* \rangle + \frac{1}{2\mu} \parallel 0 - \nabla f(x) \|^2 \geq f(x)$$

which is the PŁ condition. ◂

(1.c) [8pts] Let A be an $m \times n$ matrix with $m < n$. Suppose A has full row rank. Prove that the function $f(x) := \frac{1}{2}\|Ax - b\|_2^2$ is not strongly convex, but satisfies the PŁ condition.

Solution. Since $m < n$ there must exist a nonzero v such that $Av = 0$ for some x, let $y := x + v$. We have that

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = 0 < \frac{\mu}{2} \|v\|^2$$

for all μ so it cannot be strongly convex.

Since A has full row rank $f^* := \min_x f(x) = 0$. We have that

$$\nabla f(x) = A^\top Ax - b^\top A$$

so

$$\|\nabla f(x)\|^2 = \|A^\top (Ax - b) \|^2 \geq \sigma_{\min}^2(A) \|Ax - b\|^2 \geq \sigma_{\min}^2(A)(f(x) - f^*),$$

where $\sigma_{\min}$ is the smallest non-zero singular value which must be greater than 0. ◂

2. Gradient descent with dynamic learning rates [25pts]

Let f be a differentiable convex function on $\mathbb{R}^n$. Suppose f is Lipschitz continuous, *i.e.*, there exists $L > 0$ such that

$$|f(x) - f(y)| \leq L\|x - y\|_2 \quad \forall x, y \in \mathbb{R}^n.$$

Let $x^\star$ be any optimal solution of $\min_{x\in\mathbb{R}^n} f(x)$, and let $f^* = f(x^\star)$. Consider the gradient descent steps

$$x_{k+1} := x_k - \eta_k \nabla f(x_k).$$

(2.a) [12pts] Show that for any $k \geq 1$,

$$\sum_{j=0}^{k} \eta_j (f(x_j) - f^*) \leq \frac{1}{2}\|x_0 - x^\star\|_2^2 + \frac{L^2}{2}\sum_{j=0}^{k} \eta_j^2$$

Solution.

$$\begin{aligned}
\|x_{j+1} - x^*\|^2 &= \|x_j - \eta_j \nabla f(x_j) - x^*\|^2 \\
&= \|x_j - x^*\|^2 + \eta_j^2 \|\nabla f(x_j)\|^2 - 2\eta_j \langle \nabla f(x_j), x_j - x^* \rangle \\
&\leq \|x_j - x^*\|^2 + \eta_j^2 L^2 + 2\eta_j \langle \nabla f(x_j), x^* - x_j \rangle \\
&\leq \|x_j - x^*\|^2 + \eta_j^2 L^2 + 2\eta_j (f^* - f(x_j)),
\end{aligned}$$

where we've used Lipschitz continuity to bound $\|\nabla f\| \leq L$ and convexity to bound $\langle \nabla f(x_j), x^* - x_j \rangle \leq f^* - f(x_j)$.

Rearranging yields

$$\eta_j (f(x_j) - f^*) \leq \frac{1}{2}\left[\|x_j - x^*\|^2 - \|x_{j+1} - x^*\|^2 + \eta_j^2 L^2\right].$$

Summing and cancelling the telescoping terms yields

$$\begin{aligned}
\sum_{j=0}^{k} \eta_j (f(x_j) - f^*) &\leq \frac{1}{2}\left[\|x_0 - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \sum_{j=0}^{k} \eta_j^2 L^2\right] \\
&\leq \frac{1}{2}\left[\|x_0 - x^*\|^2 + \sum_{j=0}^{k} \eta_j^2 L^2\right],
\end{aligned}$$

where we used the nonnegativity of the norm. ◂

(2.b) [13pts] Suppose we take $\eta_k := \frac{1}{L\sqrt{k+1}}$. Show that

$$\min_{0\le j\le k}\left\{f(x_j)-f^*\right\}\le \frac{L}{2}\cdot\frac{\|x_0-x^\star\|_2^2+1+\log(k+1)}{\sqrt{k+1}}$$

for all $k\ge 1$.

▸ *Hint:* $\sum_{i=0}^{k}\frac{1}{i+1}\le 1+\log(k+1)$

Solution. Observe

$$\sum_{j}^{k}\frac{1}{\sqrt{j+1}}\ge\sum_{j}^{k}\frac{1}{\sqrt{k+1}}=\sqrt{k+1},$$

and

$$\sum_{j=0}^{k}\frac{1}{j+1}\le 1+\log(k+1).$$

$$\begin{aligned}\left\{\min_j f(x_j)-f^*\right\}\frac{\sqrt{k+1}}{L}&\le\left\{\min_j f(x_j)-f^*\right\}\sum_j\frac{1}{L\sqrt{j+1}}\\&\le\sum_j\frac{1}{L\sqrt{j+1}}(f(x_j)-f^*)\\&\le\frac{1}{2}\|x_0-x^*\|^2+\frac{L^2}{2}\sum_j\frac{1}{L^2(j+1)}\\&\le\frac{1}{2}\|x_0-x^*\|^2+\frac{1}{2}(1+\log(k+1)).\end{aligned}$$

Multiplying both sides by L yields the desired result. ◂

3. Bounding the effect of errors in gradient descent [50pts]

In class, we tacitly assumed that the gradient descent update

$$x_{t+1} := x_t - \eta \nabla f(x_t)$$

can be performed exactly. Of course, infinite precision is just an abstraction, and real machines instead perform arithmetic with finite precision, and accumulate numerical errors. To model this, we could consider an update of the form $x_{t+1} := x_t - \eta(\nabla f(x_t) + \delta_t)$ where δ_t is some error term of bounded norm $\|\delta_t\|_2 \leq \delta$, for some $\delta > 0$, that models the arithmetic imprecision of the machine. One might then wonder if the results seen in class still apply even under this update model, or whether instead the numerical error explodes.

(3.a) [15pts] For arbitrary valid L, δ and η (*i.e.*, keep these as variable quantities), find:

- A convex function f with unrestricted domain, which is L-smooth and attains a minimum
- A starting position x_0
- A sequence δ_t where $\forall t : \|\delta_t\|_2 \leq \delta$

such that noisy gradient descent will result in $\lim_{t\to\infty} f(x_t) = \infty$.

Solution. Consider the following piecewise function:

$$f(x) := \begin{cases} 0 & \text{if } \ x < 0 \\ \frac{L}{2}x^2 & \text{if } \ 0 \leq x \leq \frac{\delta}{2L} \\ -\frac{\delta^2}{8L} + \delta\frac{x}{2} & \text{o.w.} \end{cases}$$

Note f is differentiable and hence its convexity can be checked by seen by taking the first derivative. Namely, we have

$$f'(x) := \begin{cases} 0 & \text{if } \ x < 0 \\ Lx & \text{if } \ 0 \leq x \leq \frac{\delta}{2L} \\ \frac{\delta}{2} & \text{o.w.} \end{cases}$$

which is clearly non-decreasing. To prove L-smoothness, we need to check whether f' is L-Lipschitz, *i.e.* $f'(y) - f'(x) \leq L(y - x)\ \forall x, y \in \mathbb{R}$ with $x \leq y$. To do this we consider 6 cases:

- $x \leq y \leq 0$: Clearly $f'(x) = f'(y) = 0$. So the condition is trivially satisfied.
- $x \leq 0 \leq y \leq \frac{\delta}{2L}$: $f'(x) = 0$ and $f'(y) = Ly$. So $f'(y) - f'(x) = Ly \leq L(y - x)$.
- $x \leq 0, \frac{\delta}{2L} \leq y$: $f'(x) = 0$ and $f'(y) = \frac{\delta}{2}$. $f'(y) - f'(x) = \frac{\delta}{2} = \delta\frac{L}{2}L \leq Ly \leq L(y - x)$.
- $0 \leq x, y \leq \frac{\delta}{2L}$: $f'(x) = Lx$ and $f'(y) = Ly$. $f'(y) - f'(x) = L(y - x)$.

- $0 \le x \le \frac{\delta}{2L} \le y$: $f'(x) = Lx$ and $f'(y) = \frac{\delta}{2}$. $f'(y) - f'(x) = \frac{\delta}{2} - Lx = \delta\frac{L}{2L} - Lx \le L(y - x)$.
- $\frac{\delta}{2L} \le x, y$: Clearly $f'(x) = f'(y) = \frac{\delta}{2}$. So the condition is trivially satisfied.

As for the starting position, x_0, pick $x_0 = \frac{\delta}{L}$ with noise $\delta_t = \delta$. With these choices, we have

$$x_{t+1} = x_t - \eta\left(\frac{\delta}{2} - \delta\right) = x_t + \eta\frac{\delta}{2}.$$

Hence $\lim_{t\to\infty} f(x_t) = \infty$. ◂

(3.b) [35pts] Now show that when f is convex, L-smooth, and attains a minimum $x^\star$, when $\eta \le \frac{1}{2L}$ one will still have:

$$\min_{t=1\dots T} f(x_t) \le f(x^\star) + \frac{\|x^\star - x_0\|_2^2}{2\eta T} + h\Big(L, \delta, \max_{t=1\dots T} \|x_t - x^\star\|_2\Big)$$

For some function h such that

$$\lim_{\delta\downarrow 0} \frac{h\Big(L, \delta, \max_{t=1\dots T} \|x_t - x^\star\|_2\Big)}{\delta} < \infty.$$

Basically, find some $h \in O(\delta)$, so that the error disappears linearly as δ goes to zero.

▸ *Hint:* Since δ is small, $O(\delta^2)$ terms are considered smaller than $O(\delta)$ terms.

Solution. From L-smoothness and convexity,

$$\begin{aligned} f(x_{t+1}) &\le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &\le f(x^*) + \langle \nabla f(x_t), x_t - x^*\rangle + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \end{aligned}$$

Plugging in

$$x_{t+1} = x_t - \eta(\nabla f(x_t) + \delta_t)$$

gives

$$\begin{aligned} f(x_{t+1}) &\le f(x^*) - \left\langle \frac{x_{t+1} - x_t}{\eta} - \delta_t, x_{t+1} - x^*\right\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &\le f(x^*) - \frac{1}{\eta}\langle x_{t+1} - x_t, x_{t+1} - x^*\rangle - \langle \delta_t, x_{t+1} - x^*\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \end{aligned}$$

Observing that

$$\begin{aligned}\|x_t - x^*\|^2 &= \|(x_{t+1} - x_t) - (x_{t+1} - x^*)\|^2 \\ &= \|x_{t+1} - x_t\|^2 - 2\langle x_{t+1} - x_t, x_{t+1} - x^*\rangle + \|x_{t+1} - x^*\|^2,\end{aligned}$$

together with Cauchy-Schwarz inequality

$$-\langle \delta_t, x_{t+1} - x^*\rangle \le \|\delta_t\|\,\|x_{t+1} - x^*\| \le \delta\|x_{t+1} - x^*\|,$$

we have that

$$\begin{aligned}f(x_{t+1}) &\le f(x^*) + \frac{1}{\eta}\left(\frac{1}{2}\|x_t - x^*\|^2 - \frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{1}{2}\|x_{t+1} - x^*\|^2\right) \\ &\quad + \delta\|x_{t+1} - x^*\| + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= f(x^*) + \frac{1}{2\eta}\|x_t - x^*\|^2 - \frac{1}{2\eta}\|x_{t+1} - x^*\|^2 \\ &\quad + \delta\|x_{t+1} - x^*\| + \left(\frac{L}{2} - \frac{1}{2\eta}\right)\|x_{t+1} - x_t\|^2.\end{aligned}$$

When $\eta \le \frac{1}{L}$, $\frac{L}{2} - \frac{1}{2\eta} \le 0$, thus

$$f(x_{t+1}) \le f(x^*) + \frac{1}{2\eta}\|x_t - x^*\|^2 - \frac{1}{2\eta}\|x_{t+1} - x^*\|^2 + \delta\|x_{t+1} - x^*\|.$$

Using telescoping, averaging the inequality over $t = 0, ..., T-1$,

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} f(x_{t+1}) &\le \frac{1}{T}\sum_{t=0}^{T-1}\left(f(x^*) + \frac{1}{2\eta}\|x_t - x^*\|^2 - \frac{1}{2\eta}\|x_{t+1} - x^*\|^2 + \delta\|x_{t+1} - x^*\|\right) \\ &= f(x^*) + \frac{1}{2\eta T}\|x_0 - x^*\|^2 - \frac{1}{2\eta T}\|x_T - x^*\|^2 + \frac{\delta}{T}\sum_{t=0}^{T-1}\delta\|x_{t+1} - x^*\|\end{aligned}$$

Since minimal is no-greater than the average, we get

$$\min_{t=1}^{T} f(x_t) \le f(x^*) + \frac{1}{2\eta T}\|x_0 - x^*\|^2 - \frac{1}{2\eta T}\|x_T - x^*\|^2 + \delta \max_{t=1}^{T}\|x_t - x^*\|$$

as desired. ◂

6.7220 Nonlinear Optimization

Spring 2025