# SI251 - Convex Optimization homework 3

## Deadline: 2024-4-17 23:59:59

1. You can use Word, Latex or handwriting to complete this assignment. If you want to submit a handwritten version, scan it clearly.
2. The **report** has to be submitted as a PDF file to Gradescope, other formats are not accepted.
3. The submitted file name is **student_id+your_student_name.pdf**.
4. Late policy: You have 4 free late days for the quarter and may use up to 2 late days per assignment with no penalty. Once you have exhausted your free late days, we will deduct a late penalty of 25% per additional late day. Note: The timeout period is recorded in days, even if you delay for 1 minute, it will still be counted as a 1 late day.
5. You are required to follow ShanghaiTech's academic honesty policies. You are not allowed to copy materials from other students or from online or published resources. Violating academic honesty can result in serious sanctions.

**Any plagiarism will get Zero point.**

1. (50 pts) **L-smooth functions**. Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. Please prove that the following relations holds for all $x, y \in \mathbb{R}$ if $f$ with an $L$-Lipschitz continuous conditions,

$$[1] \Rightarrow [2] \Rightarrow [3]$$

$$\begin{aligned}
&[1] \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2, \\
&[2] \quad f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|^2, \\
&[3] \quad f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2L}\|\nabla f(y) - \nabla(x)\|^2, \forall x, y,
\end{aligned}$$

**Solution:** [1] $\Rightarrow$ [2]: Define the function $G : [0, 1] \to \mathbb{R}$

$$G(t) := f(x + t(y - x)) - f(x) - \langle \nabla f(x), t(y - x) \rangle,$$

so that $G(0) = 0$ and $G(1) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$. By the fundamental theorem of calculus, we have

$$\begin{aligned}
G(1) - G(0) = \int_0^1 G'(t)dt &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\
&= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), t(y - x) \rangle \frac{1}{t} dt \\
&\leq L\|y - x\|_2^2 \int_0^1 t dt \\
&= \frac{L}{2}\|y - x\|_2^2.
\end{aligned}$$

[2] $\Rightarrow$ [3]: We begin with a useful auxiliary lemma:

**Lemma 1.** *Consider a differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ satisfying condition [2] and with its global minimum achieved at some $v^*$. Then*

$$g(v) - g(v^*) \geq \frac{1}{2l}\|\nabla g(v)\|_2^2 \quad \textit{for all } v \in \mathbb{R}^d.$$

*Proof:* We have

$$\begin{aligned}
g(v^*) = \inf_{u \in \mathbb{R}^d} g(u) &\leq \left\{ g(v) + \langle \nabla g(v), u - v \rangle + \frac{L}{2}\|v - u\|_2^2 \right\} \\
&= g(v) - \frac{1}{L}\|\nabla g(v)\|_2^2,
\end{aligned}$$

where the last step follows by showing that the minimum of the quadratic program over $u$ is achieved at $u^* = v - \frac{1}{L}\nabla g(v)$, and then performing some algebra.

**Note:** This lemma and its proof are of independent interest, as they show how gradient descent with step size $1/L$ can be thought of as minimizing a linear approximation along

with a quadratic regularization term scaled by $L/2$. Let us now show that [2] $\Rightarrow$ [3]. For a fixed $x \in \mathbb{R}^d$, define the function

$$g_x(z) = f(z) - \langle \nabla f(x), z \rangle.$$

Note that $g_x$ is convex, differentiable and minimized when $z = x$, and it satisfies our smoothness condition. Hence, the preceding lemma with $v^* = x$ and $v = y$ implies that

$$g_x(y) - g_x(x) \geq \frac{1}{2L} \|\nabla g_x(y)\|_2^2 = \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2.$$

A little bit of calculation shows that

$$g_x(y) - g_x(x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle,$$

which completes the proof.

2. (50 pts) **Backtracking line search.** Please show the convergence of backtracking line search on a $m$-strongly convex and $M$-smooth objective function $f$ as

$$f\left(x^{(k)}\right) - p^\star \leq c^k \left(f\left(x^{(0)}\right) - p^\star\right)$$

where $c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\} < 1$.

---

**Algorithm 9.2** *Backtracking line search.*

**given** a descent direction $\Delta x$ for $f$ at $x \in \mathbf{dom}\, f$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.
$t := 1$.
**while** $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x, \quad t := \beta t.$

---

**Solution:**

Now we consider the case where a backtracking line search is used in the gradient descent method. We will show that the backtracking exit condition,

$$\tilde{f}(t) \leq f(x) - \alpha t \|\nabla f(x)\|_2^2,$$

is satisfied whenever $0 \leq t \leq 1/M$. First note that

$$0 \leq t \leq 1/M \Longrightarrow -t + \frac{Mt^2}{2} \leq -t/2$$

which follows from convexity of $-t + Mt^2/2$. Using this result and the bound

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2,$$

we have, for $0 \leq t \leq 1/M$,

$$\begin{aligned} \tilde{f}(t) &\leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2} \|\nabla(f(x))\|_2^2 \\ &\leq f(x) - (t/2)\|\nabla f(x)\|_2^2 \\ &\leq f(x) - \alpha t \|\nabla f(x)\|_2^2, \end{aligned}$$

since $\alpha < 1/2$. Therefore the backtracking line search terminates either with $t = 1$ or with a value $t \geq \beta/M$. This provides a lower bound on the decrease in the objective function. In the first case we have

$$f\left(x^+\right) \leq f(x) - \alpha\|\nabla f(x)\|_2^2,$$

and in the second case we have

$$f\left(x^+\right) \leq f(x) - (\beta\alpha/M)\|\nabla f(x)\|_2^2.$$

Putting these together, we always have

$$f\left(x^+\right) \leq f(x) - \min\{\alpha, \beta\alpha/M\}\|\nabla f(x)\|_2^2.$$

Now we can proceed exactly as in the case of exact line search. We subtract $p^\star$ from both sides to get

$$f\left(x^+\right) - p^\star \leq f(x) - p^\star - \min\{\alpha, \beta\alpha/M\}\|\nabla f(x)\|_2^2,$$

and combine this with $\|\nabla f(x)\|_2^2 \geq 2m\left(f(x) - p^\star\right)$ to obtain

$$f\left(x^+\right) - p^\star \leq \left(1 - \min\{2m\alpha, 2\beta om/M\}\right)\left(f(x) - p^\star\right).$$

From this we conclude

$$f\left(x^{(k)}\right) - p^\star \leq c^k\left(f\left(x^{(0)}\right) - p^\star\right)$$

where

$$c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\} < 1.$$

In particular, $f\left(x^{(k)}\right)$ converges to $p^\star$ at least as fast as a geometric series with an exponent that depends (at least in part) on the condition number bound $M/m$. In the terminology of iterative methods, the convergence is at least linear.