Lecture 4. Elementary Convex Programming III

Lecturer: Jie Wang Date: April 9, 2021

Recall from the last lecture that we would like to solve the problem as follows.

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \; & f(\mathbf{x}) \\ \text{s.t.} \; & \mathbf{g}(\mathbf{x}) \leq 0, \\ & \mathbf{h}(\mathbf{x}) = 0, \\ & \mathbf{x} \in X, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$, and $X \subseteq \mathbb{R}^n$.

1 Optimality Conditions

To solve the problem in (1), we need to characterize the optimal solutions by a system of equations, which are known as the ***optimality conditions***. If we are lucky, we can directly get closed-form solutions by the optimality conditions. Even we have to solve for the solutions by iterative approaches, we need optimality conditions such that we know when to terminate the iterations.

Proposition 1. *Suppose that the problem (1) is a convex optimization problem and solvable. If f is continuously differentiable, then $\mathbf{x}$ is optimal if and only if $\mathbf{x} \in D$ and*

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0, \, \forall \, \mathbf{y} \in D. \tag{2}$$

Proof.

$\Leftarrow$ Suppose that the inequality (2) holds. Combining the convexity of f leads to

$$\begin{aligned} & f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ \Rightarrow \; & f(\mathbf{y}) \geq f(\mathbf{x}), \, \forall \, \mathbf{y} \in D. \end{aligned}$$

$\Rightarrow$ Suppose that $\mathbf{x}$ is optimal. Then,

$$\frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \geq 0, \, \forall \, t \in (0, 1].$$

Letting t goes to zero on both sides leading to

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \lim_{t \downarrow 0} \frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \geq 0.$$

This completes the proof. □

Remark 1. The inequality in (2) is the so-called ***variational inequality*** [1].

Corollary 1. *Suppose that the problem (1) is a convex optimization problem, and f is continuously differentiable. If $\mathbf{x}^*$ is an interior point of D, then*

$$\mathbf{x}^* \in \operatorname*{\mathbf{argmin}}_{\mathbf{x}\in D} f(\mathbf{x}) \Leftrightarrow \nabla f(\mathbf{x}^*) = 0.$$

2 Problem Setup

We study the unconstrained optimization problem by setting $m = p = 0$, $X = \mathbb{R}^n$ in (1):

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}). \tag{3}$$

Without any other conditions, the problem in (3) can be challenging to solve, or even impossible to analyze. Thus, we confine ourself to a smaller set of problems by requiring several assumptions.

Assumption 1. *We assume that the assumptions as follows hold for the problem in (3).*

1. *The function f attains its minimum at $\mathbf{x}^*$, i.e.,*

$$f(\mathbf{x}^*) = f^* = \min f(\mathbf{x}). \tag{4}$$

   Notice that, the point $\mathbf{x}^*$ that satisfies Eq. (4) may not be unique.

2. *The objective function f is convex and continuously differentiable, and thus*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle, \forall\, \mathbf{x}, \mathbf{y}. \tag{5}$$

3. *The gradient of function f is Lipschitz continuous, i.e.,*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \tag{6}$$

   where $L > 0$ is the so-called Lipschitz constant.

Remark 2. The first part of Assumption 1 indeed requires the existence of the optimum of the problem in (3), which is the very first thing we need to study for general optimization problems. The second the the third parts provide an upper and lower bounds for the objective function, which are useful to analyze the convergence properties of a given algorithm. Specifically, the second part of Assumption 1 provides a lower bound of the objective function—which is clearly a linear function—by its convexity; the third part provides a quadratic upper bound of the objective function (see Lemma 1).

Lemma 1. *Suppose that a function f is continuously differentiable. If the gradient of f is Lipschitz continuous with Lipschitz constant L, i.e., the inequality (6) holds, then we have*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{7}$$

Proof. Suppose that the inequality (6) holds. Then,

$$\begin{aligned}
f(\mathbf{y}) =& f(\mathbf{x}) + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x}\rangle dt \\
=& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle dt \\
\leq& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|\|\mathbf{y} - \mathbf{x}\| dt \\
\leq& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \int_0^1 Lt\|\mathbf{y} - \mathbf{x}\|^2 dt \\
\leq& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2,
\end{aligned}$$

which completes the proof. □

The Lipschitz continuity of the gradient implies some useful information about the Hessian if $f(\cdot)$ is twice continuously differentiable. Indeed, we have the result as follows.

Theorem 1. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Then, the gradient of f is Lipschitz continuous with constant $L > 0$ if and only if*

$$\|\nabla^2 f(\mathbf{x})\|_2 \leq L, \forall \mathbf{x} \in \mathbb{R}^n, \tag{8}$$

that is, the largest eigenvalues in absolute value of the Hessian matrix is no larger than L.

Proof. $\Rightarrow$ For $t > 0$ and an arbitrary vector $\mathbf{v} \in \mathbb{R}^n$ such that $\|\mathbf{v}\| = 1$, we have

$$Lt = L\|t\mathbf{v}\| \geq \|\nabla f(x + t\mathbf{v}) - \nabla f(\mathbf{x})\| = \left\| \int_0^t \nabla^2 f(\mathbf{x} + \tau\mathbf{v}) \cdot \mathbf{v} d\tau \right\|. \tag{9}$$

The inequality (8) follows by dividing both sides of the above inequity by t and let $t \downarrow 0$.

$\Leftarrow$

$$\begin{aligned}
\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| &= \left\| \int_0^1 \langle \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt \right\| \\
&\leq \left(\int_0^1 \left\| \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \right\| dt \right) \|\mathbf{y} - \mathbf{x}\| \\
&\leq L\|\mathbf{y} - \mathbf{x}\|.
\end{aligned}$$

□

3 The Gradient Descent Algorithm

We propose to solve the problem (3) by gradient descent in Algorithm 1 as follows.

Algorithm 1 Gradient Descent

Input: An initial point $\mathbf{x}_0$, a constant $\alpha \in (0, 2/L)$, and $k = 0$.

1: **while** the *termination condition* does not hold **do**

2: $$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k). \tag{10}$$

3: $k \leftarrow k + 1$,

4: **end while**

Question 1. How to measure the performance of the above iterative algorithms?

Definition 1. Suppose that the sequence (a_k) converges to a number L.

- The sequence is said to converge linearly to L, if there exists a number $\mu \in (0, 1)$ such that

$$\lim_{k \to \infty} \frac{|a_{k+1} - L|}{|a_k - L|} = \mu,$$

where the number μ is called the ***rate of convergence***.

- The sequence is said to converge superlinearly to L, if

$$\lim_{k\to\infty} \frac{|a_{k+1} - L|}{|a_k - L|} = 0.$$

- The sequence is said to converge sublinearly to L, if

$$\lim_{k\to\infty} \frac{|a_{k+1} - L|}{|a_k - L|} = 1.$$

- The sequence is said to converge to L with order q, if there exists a number M such that

$$\lim_{k\to\infty} \frac{|a_{k+1} - L|}{|a_k - L|^q} < M.$$

  If $q = 2$ ($q = 3$), the sequence is quadratic (cubic) convergence.

4 Convergence property

In this section, we analyze the convergence property of Algorithm 1 applied to the problem in (3) introduced in Section 2. We first show that the function values generated by Algorithm 1 monotonically decrease. This is where "descent" in gradient descent comes from. Then, we show that the function values approach f^* with a rate of $O(1/k)$. We will see that the convexity and Lipschitz continuity play a central role in deriving the convergence properties.

4.1 Convergence in terms of the Function Values

In this section, we show that $f(\mathbf{x}_k) \to f(\mathbf{x}^*)$ with a convergence rate $O(1/k)$. We first show a ***descent lemma*** as follows.

Lemma 2. *Suppose a function f is continuously differentiable and its gradient is Lipschitz continuous with constant $L > 0$. Then, for the sequence generated by Algorithm 1, we have*

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \alpha\left(1 - \frac{L}{2}\alpha\right)\|\nabla f(\mathbf{x}_k)\|^2, \forall\, k = 0, 1, 2\ldots, \tag{11}$$

i.e., the sequence of function values $\{f(\mathbf{x}_k)\}_k$ is monotonically decreasing.

Proof. In view of Lemma 1 and the update rule in (10), we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\le f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= f(\mathbf{x}_k) - \alpha\|\nabla f(\mathbf{x}_k)\|^2 + \frac{L}{2}\alpha^2\|\nabla f(\mathbf{x}_k)\|^2 \\ &= f(\mathbf{x}_k) - \alpha\left(1 - \frac{L}{2}\alpha\right)\|\nabla f(\mathbf{x}_k)\|^2, \end{aligned} \tag{12}$$

which completes the proof. □

The next result tells us that the sequence of gradients tends to be zero.

Lemma 3. *Suppose that the conditions in Lemma 2 hold. Then,*

$$\nabla f(\mathbf{x}_k) \to 0. \tag{13}$$

Proof. The inequality (11) leads to

$$\begin{aligned}
&\|\nabla f(\mathbf{x}_k)\|^2 \le \frac{f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})}{\alpha\left(1 - \frac{L}{2}\alpha\right)} \\
\Rightarrow &\sum_{i=0}^{k} \|\nabla f(\mathbf{x}_i)\|^2 \le \frac{f(\mathbf{x}_0) - f(\mathbf{x}_{k+1})}{\alpha\left(1 - \frac{L}{2}\alpha\right)} \\
\Rightarrow &\sum_{i=0}^{\infty} \|\nabla f(\mathbf{x}_i)\|^2 \le \frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\alpha\left(1 - \frac{L}{2}\alpha\right)}.
\end{aligned}$$

The claim follows immediately. □

Remark 3. In both Lemmas 2 and 3, we do not assume convexity of f. Moreover, Lemma 3 implies that, if the sequence $\{\mathbf{x}_k\}$ converges to a point $\bar{\mathbf{x}}$, we would have $\nabla f(\bar{\mathbf{x}}) = 0$.

We are now ready to state the following theorem.

Theorem 2. *Consider the problem in Section 2 and the sequence generated by Algorithm 1. Then, the sequence of function values $f(\mathbf{x}_k)$ tends to the optimal function value $f(\mathbf{x}^*)$ in a rate of $O(1/k)$. Specifically,*

1. *if $\alpha \in (0, 1/L]$, we have*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \frac{1}{k}\left(\frac{1}{2\alpha}\|\mathbf{x}_0 - \mathbf{x}^*\|^2\right); \tag{14}$$

2. *if $\alpha \in (1/L, 2/L)$, we have*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \frac{1}{k}\left(\frac{1}{2\alpha}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{L\alpha - 1}{2 - L\alpha}(f(\mathbf{x}_0) - f(\mathbf{x}^*))\right). \tag{15}$$

Proof. Combining the inequality (7) and (10) leads to

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \tag{16}$$

$$\Leftrightarrow f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \le f(\mathbf{x}_k) - f(\mathbf{x}^*) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \tag{17}$$

Due to the convexity of f, we have

$$f(\mathbf{x}^*) \ge f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k\rangle. \tag{18}$$

Combining (17) and (18) leads to

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) &\le \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= -\frac{1}{\alpha}\langle \mathbf{x}_{k+1} - \mathbf{x}_k, \mathbf{x}_k - \mathbf{x}^*\rangle - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= -\frac{1}{2\alpha}\left(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2\right) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= \frac{1}{2\alpha}\left(\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2\right) - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.
\end{aligned}$$

By summing up the above inequality for $i = 0, 1, \ldots, k-1$, we have

$$k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) \leq \sum_{i=0}^{k-1} f(\mathbf{x}_{i+1}) - f(\mathbf{x}^*) \tag{19}$$
$$\leq \frac{1}{2\alpha}(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_k - \mathbf{x}^*\|^2) - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\sum_{i=0}^{k-1}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2.$$

By noting (10) and a similar argument in Lemma 3, we have

$$\sum_{i=0}^{\infty}\|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 \leq \frac{2\alpha}{2 - L\alpha}(f(\mathbf{x}_0) - f(\mathbf{x}^*)). \tag{20}$$

Thus, for $\alpha \in (0, 1/L]$, the inequality in (19) implies that

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{2\alpha k}\|\mathbf{x}_0 - \mathbf{x}^*\|^2. \tag{21}$$

If $\alpha \in (1/L, 2/L)$, by (20), we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{k}\left(\frac{1}{2\alpha}\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{L\alpha - 1}{2 - L\alpha}(f(\mathbf{x}_0) - f(\mathbf{x}^*))\right). \tag{22}$$

This completes the proof that $f(\mathbf{x}_k) \to f(\mathbf{x}^*)$ with a convergence rate $O(1/k)$. □

Remark 4. Why do not show that $\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \to 0$? Indeed, this can be wrong, as there might exist multiple optimal solutions, and we do not know in advance which optimal solution the sequence $\mathbf{x}_k$ will converge to.

5 GD for Strongly Convex Optimization Problems

In this section, we analyze the convergence property of Algorithm 1 for problem (1) when the objective function is strongly convex. We first introduce the concept of strongly convex.

Definition 2. A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called strongly convex if there exists a constant $\mu > 0$ such that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{23}$$

The constant μ is called the convexity parameter of function f.

To avoid any confusion, we explicitly state the problem we shall analyze as follows.

$$\min_{\mathbf{x}} f(\mathbf{x}), \tag{24}$$

where f is strongly convex with convexity parameter $\mu > 0$ and its gradient is Lipschitz continuous with constant $L > 0$.

Remark 5. Problem (24) always admits a unique solution. Why?

We first show a useful result as follows.

Lemma 4. *Suppose f is strongly convex with convexity parameter $\mu > 0$ and continuously differentiable. Then,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2, \forall \mathbf{x}, \mathbf{y}. \tag{25}$$

Proof. Suppose that $\mathbf{x}$ is fixed. Then, the right hand side of (23) is a function of $\mathbf{y}$, which is denoted by

$$q(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{26}$$

Then,

$$f(\mathbf{y}) \geq q(\mathbf{y}) \geq \min_{\mathbf{z}} q(\mathbf{z}) = f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2, \tag{27}$$

which completes the proof. □

We are now ready to analyze the convergence property of Algorithm 1 on problem (24).

Theorem 3. *Consider the problem (24) and the sequence generated by Algorithm 1. Then,*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq (1 - \mu\alpha(2 - L\alpha))^k (f(\mathbf{x}_0) - f(\mathbf{x}^*)). \tag{28}$$

Proof. In view of the inequality (17), we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq & f(\mathbf{x}_k) - f(\mathbf{x}^*) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\alpha^2 \|\nabla f(\mathbf{x}_k)\|^2 \\ \leq & f(\mathbf{x}_k) - f(\mathbf{x}^*) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\alpha^2 2\mu (f(\mathbf{x}_k) - f(\mathbf{x}^*)) \quad \text{by Lemma 4} \\ \leq & (1 - \mu\alpha(2 - L\alpha))^k (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \end{aligned}$$

which completes the proof. □

Remark 6. When $\alpha \in (0, 2/L)$, the coefficient $1 - \mu\alpha(2 - L\alpha)$ in (28) is in $[1 - \mu/L, 1)$, which means that the function values converge linearly to its optimal value (what if $\mu = L$?).

References

[1] O. Güler. *Foundations of optimization*. Springer, 2010.