The University of Texas at Austin
Department of Electrical and Computer Engineering

ECE 381V: Large-Scale Optimization II — Spring 2022

Lecture 5

Caramanis & Mokhtari Wednesday, February 2, 2022

Goal: In this lecture, we discuss strongly convex functions and their properties. Then, we establish a lower bound for the functions class $\mathcal{F}^{\infty,1}_{\mu,L}(\mathbb{R}^n)$. Finally, we show the convergence rate of gradient descent for $\mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$.

1 The Class of Strongly Convex Functions

Definition 1. *If there exists a constant $\mu > 0$ such that*

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

*for all $\mathbf{x}, \mathbf{y} \in S$, then the function f is μ-*strongly convex *on S.*

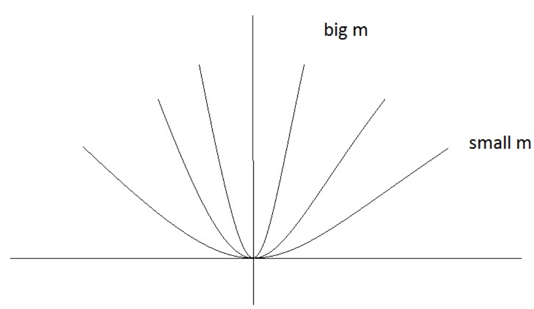

Figure 1: A strongly convex function with different parameter μ. The larger μ is, the steeper the function looks like.

When $\mu = 0$, we recover the basic inequality characterizing convexity; for $\mu > 0$, we obtain a better lower bound on $f(\mathbf{y})$ than that from convexity alone.

Typically as shown in Figure (1) , a small μ corresponds to a 'flat' convex function while a large μ corresponds to a 'steep' convex function.

Remark 1. *We denote the class of differentiable functions that are strongly convex by $\mathcal{F}^1_\mu(\mathbb{R}^n, \|.\|)$.*

1.1 Side results of strong convexity

Strong convexity has several interesting consequences. We will see that we can bound both $f^* - f(\mathbf{x})$ and $||\mathbf{x} - \mathbf{x}^*||_2$ in this section.

Lemma 1. *If $f \in \mathcal{F}^1_\mu(\mathbb{R}^n)$, then*

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|_2^2$$

Proof. The righthand side of the strong convexity inequality is a convex quadratic function of $\mathbf{y}$ (for fixed $\mathbf{x}$). Setting the gradient with respect to $\mathbf{y}$ equal to zero, we can find the $\tilde{y}$ that minimizes the right hand side.

$$\begin{aligned}
\frac{\partial}{\partial x}(f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2) &= 0 \\
\nabla f(\mathbf{x}) - \mu(\mathbf{y} - \mathbf{x}) &= 0 \\
\mathbf{y} = \mathbf{x} - \frac{1}{\mu}\nabla f(\mathbf{x})
\end{aligned}$$

So $\tilde{\mathbf{y}} = \mathbf{x} - (1/\mu)\nabla f(\mathbf{x})$ minimizes the righthand side. If we minimize the left hand side with respect to $\mathbf{y}$ we obtain $\mathbf{y} = \mathbf{x}^*$ and the left hand side becomes f^*. Hence, we have

$$\begin{aligned}
f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\
f^* &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \tilde{\mathbf{y}} - \mathbf{x}\rangle + \frac{\mu}{2}\|\tilde{\mathbf{y}} - \mathbf{x}\|^2 \\
&= f(\mathbf{x}) - \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2
\end{aligned}$$

and the claim follows. □

This result allows us to realize how fast you get to a minimum as a function of gradient. If the gradient is small at a point, then the point is nearly optimal. This upper bound also implies that if we find a point $\hat{\mathbf{x}}$ such that, $\|\nabla f(\hat{\mathbf{x}})\|_2 \leq \sqrt{2\mu\epsilon}$, then we can conclude that $\hat{\mathbf{x}}$ is ϵ-suboptimal, i.e., $f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon$.

Similarly, we can also derive an upper bound on $\|\mathbf{x} - \mathbf{x}^*\|_2$, the distance between x and any optimal point x^*, in terms of $\|\nabla f(\mathbf{x})\|_2$:

Lemma 2. *If $f \in \mathcal{F}^1_\mu(\mathbb{R}^n)$, then*

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \frac{2}{\mu}\|\nabla f(\mathbf{x})\|_2 \tag{2}$$

where $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$ is the unique minimizer of f.

Proof. We apply (1) with $\mathbf{y} = \mathbf{x}^*$ to obtain:

$$\begin{aligned}
f^* = f(\mathbf{x}^*) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}\|_2^2 \\
&\geq f(\mathbf{x}) - \|\nabla f(\mathbf{x})\|_2\|\mathbf{x}^* - \mathbf{x}\|_2 + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}\|_2^2,
\end{aligned}$$

Since $f^* \leq f(\mathbf{x})$, the terms following $f(\mathbf{x})$ on the righthand side must be negative. We have

$$\begin{aligned}
-\|\nabla f(\mathbf{x})\|_2\|\mathbf{x}^* - \mathbf{x}\|_2 + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}\|_2^2 &\leq 0 \\
\|\mathbf{x} - \mathbf{x}^*\|_2 &\leq \tfrac{2}{\mu}\|\nabla f(\mathbf{x})\|_2
\end{aligned}$$

from which (2) follows. □

One consequence of (2) is the solution locates within a ball of radius of $\frac{2}{\mu}\|\nabla f(\mathbf{x})\|_2$ around the optimal solution.

Lemma 3. *If $f \in \mathcal{F}^1_\mu(\mathbb{R}^n)$, then for any $\mathbf{x}$ and $\mathbf{y}$ and $\alpha \in [0,1]$ we have*

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) - \frac{\alpha(1-\alpha)\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

$$\mu\|\mathbf{x}-\mathbf{y}\|^2 \leq (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y})$$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{1}{2\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))\top(\mathbf{x}-\mathbf{y}) \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

Proof. Exercise. □

Remark 2. *We denote the class of twice differentiable functions that are strongly convex by $\mathcal{F}^2_\mu(\mathbb{R}^n)$.*

Theorem 1. *Consider a twice differentiable function f. Then, f is μ-strongly convex if and only if $\nabla^2 f(\mathbf{x}) \succeq \mu\mathbf{I}$.*

2 The class of $\mathcal{F}^{1,1}_{\mu,L}$ functions

This is an important class functions that are differentiable, strongly convex, and has Lipschitz continuous gradients. Basically, for $\mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$ we have

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y}) \geq \mu\|\mathbf{x}-\mathbf{y}\|^2$$

and

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

We further define the value $\kappa = L/\mu$ as the condition number of the problem.

Theorem 2. *If $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$, then we have*

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y}) \geq \frac{\mu L}{\mu+L}\|\mathbf{x}-\mathbf{y}\|^2 + \frac{1}{\mu+L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

3 The class of $\mathcal{F}^{2,1}_{\mu,L}$ functions

For this class of functions that are twice differentiable, strongly convex, and has Lipschitz continuous gradients, the most important property that we have is

$$\mu\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I} \tag{3}$$

In this case, we observe that the condition number $\kappa = L/\mu$ is a uniform (and hence upper) bound on the condition number of the matrix $\nabla^2 f(\mathbf{x})$ at any given $\mathbf{x}$.

Definition 2. *When the ratio is close to 1, we call it* **well-conditioned**. *When the ratio is very large, we call it* **ill-conditioned**.

4 Lower bound for $\mathcal{F}_{L}^{\infty,1}(\mathbb{R}^n)$

We first formally define the class of functions, query/oracle, and the required approximation error.

Model: $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$, where $f \in \mathcal{F}_{\mu,L}^{\infty,1}(\mathbb{R}^n)$.

Oracle: First-order Black Box.

Approximation Solution: Find $\hat{\mathbf{x}} \in \mathbb{R}^n$ such that $f(\hat{\mathbf{x}}) - f^* \leq \epsilon$

Assumption on the class of algorithms. We define the class of algorithms $\mathcal{A}_{lin}$ as methods that generate a sequence of test points $\{\mathbf{x}_k\}$ according to the following condition:

$$\mathbf{x}_k \in \mathbf{x}_0 + \mathbf{Span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_{k-1})\} \qquad k \geq 1.$$

Our lower bound will depend on the problem condition number $\kappa = L/\mu$. Note that infinite dimensional problems also belongs to the class of problems that we consider, as we don't have any restrictions on problem dimension. A similar argument can be shown for the finite dimension, but for our lower bound we focus on infinite dimensional problems to make our reasoning simpler.

Theorem 3. *For any $\mathbf{x}_0 \in \mathbb{R}^n$, there exists a function $f \in \mathcal{F}_{\mu,L}^{\infty,1}(\mathbb{R}^n)$ such that for any algorithm in $\mathcal{A}_{lin}$ we have*

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

and

$$f(\mathbf{x}_k) - f^* \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Proof. Consider the following function

$$f(\mathbf{x}) = \frac{L-\mu}{8}\left[\left(x(1)^2 + \sum_{i=1}^{\infty}(x(i) - x(i+1))^2 + x(k)^2\right) - 2x(1)\right] + \frac{\mu}{2}\|\mathbf{x}\|^2$$

where $x(j)$ is the j-th coordinate of $\mathbf{x}$. It can also be seen that this is quadratic function and its Hessian is given by $\nabla^2 f(\mathbf{x}) = \frac{L-\mu}{4}\mathbf{A} + \mu\mathbf{I}$ where

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Hence, we can easily show that this function is μ-strongly convex and L-smooth. It can be also verified that the gradient of $f(\mathbf{x})$ is given by

$$\nabla f(\mathbf{x}) = \frac{L-\mu}{4}(\mathbf{A}\mathbf{x} - \mathbf{e}_1) + \mu\mathbf{x}$$

By setting the gradient to zero we obtain that the optimal solution of $f(\mathbf{x})$ is

$$\frac{L-\mu}{4}(\mathbf{A}\hat{\mathbf{x}}-\mathbf{e}_1)+\mu\mathbf{x}=\mathbf{0} \iff \left(\frac{L-\mu}{4}\mathbf{A}+\mu\mathbf{I}\right)\hat{\mathbf{x}}=\frac{L-\mu}{4}\mathbf{e}_1 \iff \left(\mathbf{A}+\frac{4}{\kappa-1}\mathbf{I}\right)\hat{\mathbf{x}}=\mathbf{e}_1$$

Hence, the coordinates of the optimal solution should satisfy the following conditions:

$$2\frac{\kappa+1}{\kappa-1}\hat{x}(1)-\hat{x}(2)=1$$

and for $k\geq 2$ we have

$$\hat{x}(k+1)-2\frac{\kappa+1}{\kappa-1}\hat{x}(k)+\hat{x}(k-1)=0$$

If we assume that the solution for the above system of equations has the form $\hat{x}(k)=cq^k$ then we obtain that

$$q^2-2\frac{\kappa+1}{\kappa-1}q+1=0$$

and, hence, the smaller root that makes the norm finite has the form

$$q^*=\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$

To figure out c we can use the first equation which implies that

$$2c\times\frac{\kappa+1}{\kappa-1}\times\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}-c\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2=1$$

It can be easily verified that $c=1$ and hence

$$\hat{x}(k)=\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$$

Now if (WLOG) we set $\mathbf{x}_0=\mathbf{0}$, then we have

$$\|\mathbf{x}_0-\mathbf{x}^*\|^2=\|\mathbf{x}^*\|^2=\sum_{k=1}^{\infty}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}=\frac{\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2}{1-\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2}$$

Moreover, since all coordinates of $\mathbf{x}_k$ are nonzero for indices larger than k we can show that

$$\|\mathbf{x}_k-\mathbf{x}^*\|^2\geq\sum_{i=k+1}^{\infty}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2i}=\frac{\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(k+1)}}{1-\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2}=\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\|\mathbf{x}_0-\mathbf{x}^*\|^2$$

and the proof for the first statement is complete. To prove the second claim, we simply use the fact that

$$f(\mathbf{x})-f(\mathbf{x}^*)\geq\frac{\mu}{2}\|\mathbf{x}-\mathbf{x}^*\|^2.$$

Hence, the proof is complete. $\square$

5 Gradient Descent for $\mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$

In the gradient descent method we follow the update

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)$$

where η is the stepsize or learning rate.

In the following theorem, we characterize the convergence rate of gradient descent for $\mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$.

Theorem 4. *Let $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$ and $0 < \eta \leq \frac{2}{\mu+L}$. Then, the iterates of GD satisfy*

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^2 \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Proof. We can simply show that

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_k - \mathbf{x}^*) + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \\
&= \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))^\top (\mathbf{x}_k - \mathbf{x}^*) + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \\
&\leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \frac{\mu L}{\mu + L} \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\eta \frac{1}{\mu + L} \|\nabla f(\mathbf{x})\|^2 + \eta^2 \|\nabla f(\mathbf{x}_k)\|^2 \\
&= \left(1 - 2\eta \frac{\mu L}{\mu + L}\right) \|\mathbf{x}_k - \mathbf{x}^*\|^2 - \eta \left(\frac{2}{\mu + L} - \eta\right) \|\nabla f(\mathbf{x}_k)\|^2 \\
&\leq \left(1 - 2\eta \frac{\mu L}{\mu + L}\right) \|\mathbf{x}_k - \mathbf{x}^*\|^2
\end{aligned}$$

where the first inequality holds due to

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq \frac{\mu L}{\mu + L} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{1}{\mu + L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

and the second one holds since $0 < \eta \leq \frac{2}{\mu+L}$. □

Corollary 1. *If we set $\eta = 2/(\mu + L)$ we obtain the best rate which is*

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

and hence

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

Remark 3. *Our lower bound shows a linear convergence rate of $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}$ while our upper bound scales with $\left(\frac{\kappa-1}{\kappa+1}\right)^{2k}$. Indeed, there is a gap between these two bounds. Now the question is, can we improve the lower bound or the upper bound? In other words, which one is possible? Deriving a harder instance to improve our lower bound? Or presenting an algorithm that converges faster and improves our upper bound? We will answer this question later.*