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

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

Lecture 4

Caramanis & Mokhtari Monday, January 31, 2022

Goal: In this lecture, we discuss gradient descent and its convergence rate for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$, and discuss the lower bound for the functions class $\mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$.

1 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 $f \in \mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$.

Oracle: First-order Black Box.

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

We further make an important assumption on the class of algorithms that we are allowed to use.

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:

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

Most (but not all) practical first-order methods satisfy the above condition and belong to the class of algorithms that we consider.

Note that in the following lower bound result, since our complexity bound is independent of the dimension, we make the dimension large enough such that $k \leq \frac{n-1}{2}$. This requirement is essential as we are designing a function that gradient updates only capable of learning one coordinate of that at each iteration.

Theorem 1. *For any k where $1 \leq k \leq \frac{n-1}{2}$ and any $\mathbf{x}_0 \in \mathbb{R}^n$, there exists a function $f \in \mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$ such that for any algorithm in $\mathcal{A}_{lin}$ we have*

$$f(\mathbf{x}_k) - f^* \geq \frac{3L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{32(k+1)^2}.$$

Proof. Consider the following function

$$f_k(\mathbf{x}) = \frac{L}{4}\left[\frac{1}{2}\left(x(1)^2 + \sum_{i=1}^{k-1}(x(i) - x(i+1))^2 + x(k)^2\right) - x(1)\right]$$

where $x(j)$ is the j-th coordinate of $\mathbf{x}$. The value of this function basically for any vector $\mathbf{x}$ depends only on its first k elements. It can also be seen that this is quadratic function and its Hessian is given by $\nabla^2 f_k(\mathbf{x}) = \frac{L}{4}\mathbf{A}_k$ where

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

It can be also verified that the gradient of $f_k(\mathbf{x})$ is given by

$$\nabla f_k(\mathbf{x}) = \frac{L}{4}(\mathbf{A}_k\mathbf{x} - \mathbf{e}_1)$$

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

$$\hat{x}_k(i) = \begin{cases} 1 - \frac{i}{k+1} & \text{for } i = 1, \ldots, k \\ 0 & \text{for } i = k+1, \ldots, n \end{cases}$$

and therefore its optimal value is

$$\hat{f}_k = f_k(\hat{\mathbf{x}}_k) = \frac{L}{8}\left(\frac{1}{k+1} - 1\right).$$

Another important observation is that if we start from $\mathbf{x}_0 = \mathbf{0}$ (which we can do without loss of generality), then after each gradient update we make one more coordinate of the gradient non-zero. We formally prove this by induction, note that at $\mathbf{x}_0 = \mathbf{0}$ the gradient is $\nabla f_k(\mathbf{0}) = -\frac{L}{4}\mathbf{e}_1$. Therefore, only the first coordinate of $\mathbf{x}_1$ could be nonzero if our algorithm belongs to $\mathcal{A}_{lin}$. Now suppose the first l coordinates of $\mathbf{x}_l$ are possibly nonzero but the remaining are zero. Since $\mathbf{A}$ is a tri-diagonal matrix, we have that at most the first $l+1$ components of $\nabla f_k(\mathbf{x}_l)$ are non-zero. Hence, at most the first $l+1$ components of $\mathbf{x}_{l+1}$ are non-zero.
Further, for any function $f_k(\mathbf{x})$ we indeed have

$$f_k(\mathbf{x}_l) \geq \hat{f}_k$$

for $l = 1, \ldots, k$.
Now consider the following two functions

$$f_k(\mathbf{x}) = \frac{L}{4}\left[\frac{1}{2}\left(x(1)^2 + \sum_{i=1}^{k-1}(x(i) - x(i+1))^2 + x(k)^2\right) - x(1)\right]$$

and

$$f_{2k+1}(\mathbf{x}) = \frac{L}{4}\left[\frac{1}{2}\left(x(1)^2 + \sum_{i=1}^{2k}(x(i) - x(i+1))^2 + x(2k+1)^2\right) - x(1)\right]$$

It is indeed obvious that for iterate $\mathbf{x}_l$ where $l = 1, \ldots, k$ the value of these two functions are the same, as for these iterates we have $x_l(k+1) = x_l(2k+1) = 0$. Hence, we have

$$f_{2k+1}(\mathbf{x}_k) = f_k(\mathbf{x}_k) \geq \hat{f}_k = \frac{L}{8}\left(\frac{1}{k+1} - 1\right).$$

Hence, if we define $f(\mathbf{x}) := f_{2k+1}(\mathbf{x})$, then we have

$$f(\mathbf{x}_k) - f^* \geq \frac{L}{8}\left(\frac{1}{k+1} - 1\right) - \frac{L}{8}\left(\frac{1}{2k+2} - 1\right) = \frac{L}{8}\left(\frac{1}{k+1} - \frac{1}{2k+2}\right) = \frac{L}{16}\frac{1}{k+1}$$

On the other hand we have

$$\begin{aligned}
\|\mathbf{x}_0 - \mathbf{x}^*\|^2 = \|\mathbf{x}^*\|^2 &= \sum_{i=1}^{2k+1}\left(1 - \frac{i}{2k+2}\right)^2 \\
&= \sum_{i=1}^{2k+1} 1 - \frac{i}{k+1} + \frac{i^2}{(2k+2)^2} \\
&= 2k+1 - \frac{(2k+1)(2k+2)}{2(k+1)} + \frac{(2k+1)(2k+2)(4k+3)}{6(2k+2)^2} \\
&= \frac{(2k+1)(2k+2)(4k+3)}{24(k+1)^2} \\
&= \frac{(2k+1)(4k+3)}{12(k+1)} \\
&\leq \frac{(2k+2)(4k+4)}{12(k+1)} \\
&\leq \frac{2(k+1)}{3}
\end{aligned}$$

Hence, we have

$$\frac{f(\mathbf{x}_k) - f^*}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2} \geq \frac{\frac{L}{16}\frac{1}{k+1}}{\frac{2(k+1)}{3}} = \frac{3L}{32(k+1)^2}$$

and the proof is complete. □

2 Gradient Descent for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

As we learned in the second lecture, if the gradient at a point is nonzero by following the negative gradient direction and choosing a small enough stepsize we can always decrease the function value, in unconstrained problem. 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. Another interpretation of the gradient descent method is the minimizer of the following function:

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n}\left\{f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top(\mathbf{x} - \mathbf{x}_k) + \frac{1}{\eta}\|\mathbf{x} - \mathbf{x}_k\|^2\right\}$$

Note that if $1/\eta > L/2$, i.e., $\eta < 2/L$, then the above function in the right hand side is an upper bound for the objective function. Hence, it means that in gradient descent, we use the linear

approximation and the smoothness to come up with an upper bound for the function, and then use its minimizer as the new iterate.

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

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

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

Proof. Note that using the smoothness property we can write

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

Hence, we have

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

Now to show at what rate $f(\mathbf{x}_k) - f^*$ is decreasing we need to study the behavior of the gradient norm.
Note that using convexity we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}^* - \mathbf{x}_k)$$

By regrouping the terms we obtain

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_k - \mathbf{x}^*) \leq \|\nabla f(\mathbf{x}_k)\|\|\mathbf{x}_k - \mathbf{x}^*\|$$

The above inequality provides a nice lower bound on norm of gradient in terms of objective function suboptimality which is exactly what we need. Now we need to ensure that $\|\mathbf{x}_k - \mathbf{x}^*\|$ is uniformly bounded above. This can be shown using the fact that

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

where the first inequality follows from $\frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2 \leq (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y})$. Hence, $\|\mathbf{x}_k - \mathbf{x}^*\|^2$ for any k is uniformly bounded by $\|\mathbf{x}_0 - \mathbf{x}^*\|^2$.
By combining our results we have

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

Divide both sides by $(f(\mathbf{x}_{k+1}) - f^*)(f(\mathbf{x}_k) - f^*)$ to obtain

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

By regrouping the terms we have

$$\begin{aligned}\frac{1}{f(\mathbf{x}_{k+1}) - f^*} &\geq \frac{1}{f(\mathbf{x}_k) - f^*} + \eta \left(1 - \frac{L\eta}{2}\right) \frac{f(\mathbf{x}_k) - f^*}{(f(\mathbf{x}_{k+1}) - f^*)\|\mathbf{x}_0 - \mathbf{x}^*\|^2} \\ &\geq \frac{1}{f(\mathbf{x}_k) - f^*} + \eta \left(1 - \frac{L\eta}{2}\right) \frac{1}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}\end{aligned}$$

where the last inequality holds since $f(\mathbf{x}_{k+1}) - f^* \leq f(\mathbf{x}_k) - f^*$. Now if we sum up both sides we obtain

$$\frac{1}{f(\mathbf{x}_{k+1}) - f^*} \geq \frac{1}{f(\mathbf{x}_0) - f^*} + \eta \left(1 - \frac{L\eta}{2}\right) \frac{k+1}{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}$$

Hence,

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

□

Corollary 1. *If we set $\eta = 1/L$ we obtain the best rate which is*

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

Corollary 2. *If we set $\eta = 1/L$ we have*

$$f(\mathbf{x}_k) - f^* \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k+4}$$

Proof. It simply follows from the fact that

$$f(\mathbf{x}_0) \leq f(\mathbf{x}^*) + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^*\|^2.$$

□

Remark 1. *Our scales with $1/k^2$ while our upper bound scales with $1/k$. 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.*