Lecture 4

Last class

- Least squares
- Implicit bias

Today

- Lower bounds

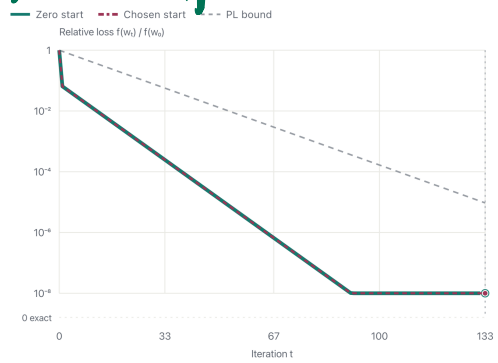
Lower bounds

In Lecture 2 we derived linear convergence for GD on PL + smooth functions, i.e.,

$$T \geq \kappa \log(1/\varepsilon) \Rightarrow \frac{f(w_T) - f^*}{f(w_0) - f^*} \leq \varepsilon$$

($\kappa = L/\mu$)

The reason why it is called linear (also geometric) is that if you plot it on a log-y scale it looks like a linear function.



This is pretty fast fast. but it is natural to wonder

Is there a faster algorithm than GD?

Today we will answer this question. To do so, we need will use a framework originally developed by Nemiroski and

Yudin 1983. The idea is the following suppose we have

A metric of progress $\mathcal{L}(\cdot)$,
a class of algorithms $\mathcal{A}_k$, and
a function class $\mathcal{F}$.

("Number of iterations")

Then our goal is to find a lower bound for

$$\inf_{A \in \mathcal{A}_k} \sup_{f \in \mathcal{F}} \mathcal{L}(w_A).$$

This is a beautiful idea, but we have to be careful with what an algorithm can do. Imagine we allowed the algorithm to query infinitely many point per iteration, that would allow us to solve any problem in one iteration.
A better "unit" is the number of oracle calls (questions or queries to the function). Today we will consider first-order oracles, i.e.,

$$w \longmapsto (f(w), \nabla f(w)).$$

We consider iterative algorithms that produce

$$x_0, x_1, \dots, x_K$$ ← One per oracle call.

For simplicity, we assume that our algorithms are provided $x_0 = 0$ and

(K) $x_{k+1} \in \text{span}\{\nabla f(x_0), \dots, \nabla f(x_k)\}$.

For the metric of progress we will use

$$\frac{f(w_K) - f^*}{\Delta}.$$

Δ ← initial function gap

We instantiate this strategy for two function classes μ-strongly convex + L-smooth functions and μ-PL + L-smooth functions.

Strongly convex functions

Theorem (N): Let $L, \mu, \Delta > 0$ with $\kappa = L/\mu \geq 1$, and $T \geq 1$. Then, there exists a μ-strongly convex, L-smooth function $f: \mathbb{R}^{2T} \to \mathbb{R}$ with $f^* = 0$, and $f(0) = \Delta$ such that every method satisfying

(K) obeys

$$\frac{f(x_T) - f^*}{\Delta} \geq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{2T}.$$

The same holds for the "best" iterate. ↑

Corollary: Under the same setting, we get that whenever $\sqrt{\kappa} \geq 4$, then

$$T \leq \min\left\{\frac{d}{2}, \frac{\sqrt{\kappa}}{8}\log\left(\frac{1}{\epsilon}\right)\right\} \Rightarrow \frac{f(w_T) - f^*}{\Delta} > \epsilon.$$

↑

This is actually important, otherwise the span can generate the entire space.

Proof: Before delving into the proof let us provide a road map.

1. Construct a minimizer with exponentially decaying entries and a "chain" function for it.
2. Verify that the chain function is in the function class.
3. Show that the chain function only reveals one entry per query.
4. Convert this fact into a lower bound.

Step 1: Build a chain

For simplicity, write

$$\rho = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) \in (0,1) \quad \text{and} \quad a = \sqrt{\frac{2\Delta}{L}\frac{(1+\rho)^2}{(1-\rho^2)}}.$$

We define our minimizer as

$$w^* = (a, a\rho, a\rho^2, \ldots, a\rho^{2T-1})$$

and the chain function as

$$f(w) = \frac{L}{2(1+\rho)^2}\left[(1-\rho^2)(a-w_1)^2 + \sum_{i=1}^{2T-1}(x_{i+1}-\rho x_i)^2\right] \quad (xx)$$

↑ The worst smooth + SC in the world.

It is easy to see that $f(w^*) = 0 \le f(w)\ \forall w$.

Further

$$f(0) = \frac{L}{2(1+\rho)^2}(1-\rho^2)a^2 = \Delta.$$

Step 2: Verify the chain is in the class

For this we will use a classical fact.

Fact (Characterization) A twice differentiable function $f: \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth if, and only if,

Löwner order $A \preceq B \Leftrightarrow \lambda_{\min}(B-A) \geq 0$.

$$\mu I_d \preceq \nabla^2 f(w) \preceq L I_d \qquad \forall w \in \mathbb{R}^d$$

Our function is a quadratic that we can write as

$$f(w) = \frac{L}{2(1+\rho)^2}\left(w^T B w - (1-\rho^2)\, a\, e_1\right)$$

with a tridiagonal matrix

$$B = \begin{pmatrix} 1 & -\rho & & & & \\ -\rho & 1+\rho^2 & -\rho & & & \\ & -\rho & 1+\rho^2 & \ddots & & \\ & & \ddots & & -\rho & \\ & & & -\rho & 1+\rho^2 & \end{pmatrix}.$$

(check!)

Expanding we obtain that

$$w^T B w = w_1^2 + w_{2T}^2 + (1+\rho^2)\sum_{i=2}^{2T-1} w_i^2 - 2\rho \sum_{i=1}^{2T-1} w_i w_{i+1}$$

Using the fact that $2|uv| \leq u^2 + v^2$ we get that

$$w^T B w \geq w_1^2 + w_{2T}^2 + (1+\rho^2)\sum_{i=2}^{2T-1} w_i^2 - \rho\left(w_1^2 + w_{2T}^2 + 2\sum_{i=2}^{2T-1} w_i^2\right)$$

$$= (1-\rho)(w_1^2 + w_{2T}^2) + (1-\rho)^2 \sum_{i=1}^{2T-1} w_i^2$$

$(1-\rho) \geq (1-\rho)^2$

$$\geq (1-\rho)^2 \|w\|_2^2.$$

A similar argument, yields $w^T B w \le (1+\rho)^2 \|w\|^2$

therefore

$$\mu I = \frac{L}{K} I = L\left(\frac{1-\rho}{1+\rho}\right)^2 I \preceq \underbrace{\frac{L}{(1+\rho)^2} B}_{\nabla^2 f(w)} \preceq L I$$

So we conclude f is in our function class.

Step 3: Each query reveals one coordinate

Define the subspaces

$E_0 = \{0\}$ and $E_j = \text{span}\{e_1, \dots, e_j\}$ $\forall j$.

The gradient of f is

$$\nabla f(w) = \frac{L}{(1+\rho)^2}\left(Bx - (1-\rho^2)\, a\, e_1\right).$$

So for $w_0 \in E_0$ and $w_1 \in \text{span}\{\nabla f(0)\} = E_1$.

Now, let's show that

$$w_k \in E_k \Rightarrow w_{k+1} \in E_{k+1}.$$

Suppose $w_k \in E_k$, since B_k is tridiagonal

$$B w_k = \begin{pmatrix} \ddots & & & & \\ & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \ddots \end{pmatrix} \begin{pmatrix} \times \\ \vdots \\ \times \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \times \\ \times \\ \vdots \\ \times \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in E_{k+1}.$$

↑ k col

Hence $\nabla f(w_k) \in E_{k+1} \Rightarrow w_k \in \text{span } E_{k+1}$.

Step 4: Convert this into a lower bound

We know that for any vector $w \in E_{k+1}$ we must have that

$$|(\overbrace{w - w^*}^{z})_{T+1}| = a\rho^T.$$

Our goal is to use this as a lower bound for $f(w) - f^*$. To do so, notice that

$$(\ddot\smile) \quad f(w) - f^* = \frac{L}{2(1+\rho^2)}\left((1-\rho^2)z_1 + \sum_{j=1}^{2k-1}(z_{j+1} - \rho z_j)^2\right)$$

($f^* = 0$ and $w^*_{j+1} = \rho w^*_j$)

Let's use Cauchy-Schwarz to lower bound

$$a\rho^T = \rho^T z_1 - \sum_{j=1}^{T} \rho^{T-j}(z_{j+1} - \rho z_j)$$

Expand $z_{j+1} = \rho z_j + (z_{j+1} - \rho z_j)$

$$= \left\langle \begin{pmatrix} \frac{\rho^T}{\sqrt{1-\rho^2}} \\ \rho^{T-1} \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} \sqrt{1-\rho^2}\, z_1 \\ z_2 - \rho z_1 \\ \vdots \\ z_{T+1} - \rho z_T \end{pmatrix} \right\rangle$$

$$\leq \underbrace{\left(\frac{\rho^{2T}}{(1-\rho^2)} + \sum_{j=1}^{T}\rho^{2(T-j)}\right)^{1/2}}_{P_1} \times \overbrace{\left((1-\rho^2)z_1^2 + \sum_{j=1}^{T}(z_{j+1} - \rho z_j)^2\right)^{1/2}}^{P_2}$$

The first term in this product simplifies

$$P_1 = \frac{\rho^{2T}}{1-\rho^2} + \frac{1-\rho^{2T}}{1-\rho^2} = \frac{1}{1-\rho^T}.$$

On the other hand, using $(\because)$ we get

$$P_2 \leq \frac{2(1+\rho)^2}{L}\left(f(w) - f^*\right)$$

Squaring $(\circ\circ)$ and using these two bounds yields

$$a^2\rho^{2T} \leq \frac{2(1+\rho)^2}{L(1-\rho^2)}\left(f(w) - f^*\right)$$

Rearranging and simplifying gives

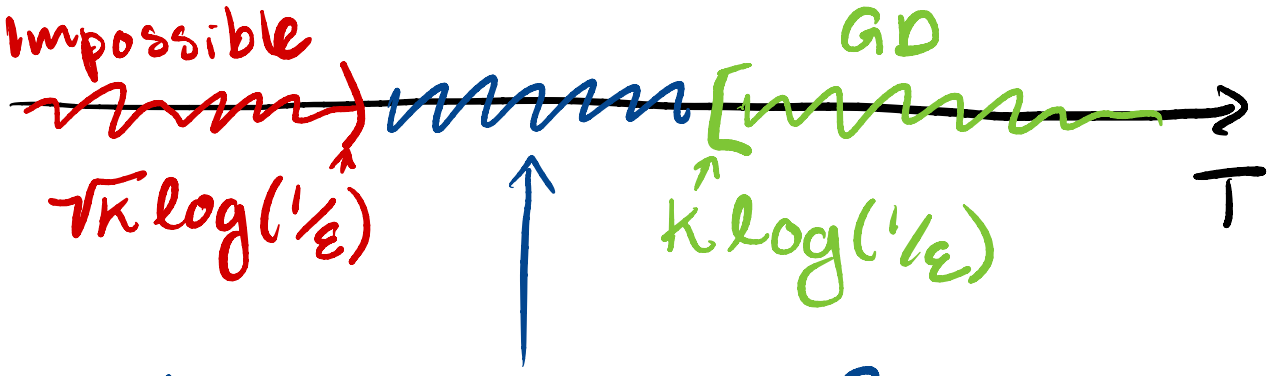
$$f(w) - f^* \geq \frac{L(1-\rho^2)}{2(1+\rho)^2}\, a\, \rho^{2T} = \Delta\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2.$$

□

Functions satisfying PL

The previous result also applies for PL because SC $\Rightarrow$ PL.

Thus, we know the following



What happens here?

The answer to this question depends on the family of function we consider.

For, functions satisfying PL, the lower bound we derived today is loose.

Theorem (Yue, Fang, Lin 2023): Fix $L, \mu, \Delta > 0$ with $\kappa = L/\mu \geq 10^5$ and $T \geq 1$. Then, there exists an L-smooth function $f: \mathbb{R}^{2T} \to \mathbb{R}$ with $f^* = 0$ and $f(0) = \Delta$ such that every method satisfying (K) obeys

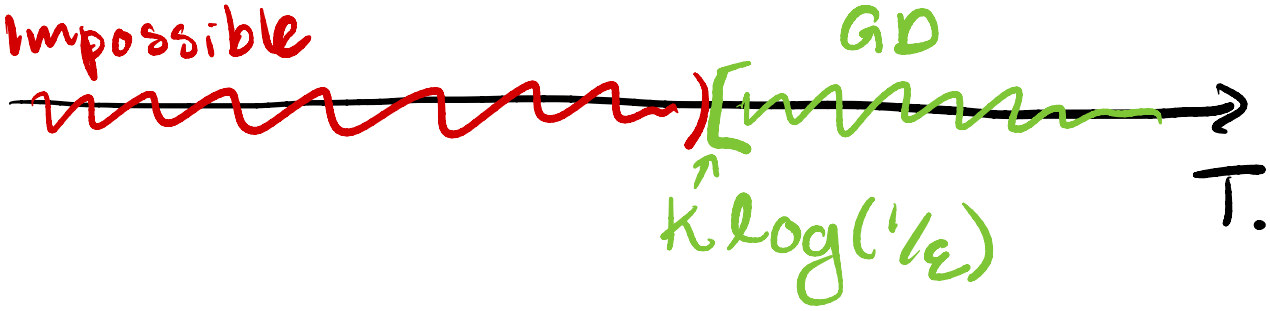
$$\frac{f(w_K) - f^*}{\Delta} \geq \exp(-10^5 T/\kappa)$$

The same bound holds for the best iterate.

Thus we get

$$T \le \min\left\{\frac{d}{2}, \frac{\kappa}{105}\log(1/\varepsilon)\right\} \Rightarrow \frac{f(w_T) - f^*}{\Delta} \ge \varepsilon.$$

Hence, up to constants we have



Next class we will see that for strongly convex + smooth function,

