CS-GY 6763: Lecture 8
Projected Gradient Descent, Second order conditions

NYU Tandon School of Engineering, Prof. Christopher Musco

Goal: Find approximate minizer for a function $f(\mathbf{x})$.

Gradient Descent Algorithm:

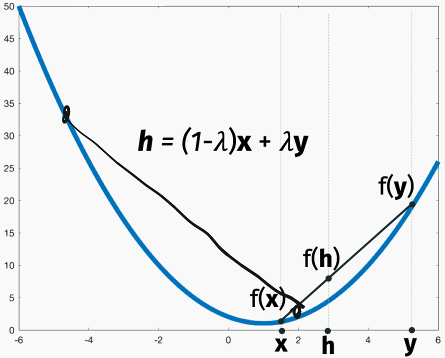
- Choose starting point $\mathbf{x}^{(0)}$.
- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\mathbf{x}^{(T)}$ (or $\arg\min_{i \leq T} f(\mathbf{x}^{(i)})$).

η is a step-size parameter.

Definition (Convex)

A function f is convex iff for any $\mathbf{x}, \mathbf{y}, \lambda \in [0, 1]$:

$$(1-\lambda)\cdot f(\mathbf{x}) + \lambda \cdot f(\mathbf{y}) \geq f((1-\lambda)\cdot \mathbf{x} + \lambda \cdot \mathbf{y})$$



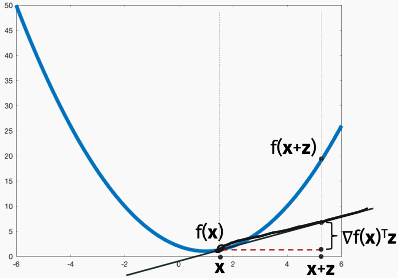
Definition (Convex function)

A function f is convex if and only if for any $\mathbf{x}, \mathbf{y}$:

$$f(\mathbf{x} + \mathbf{z}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{z}$$

Equivalently:

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{x})^T (\mathbf{x} - \mathbf{y})$$



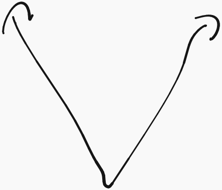
Definition (Convex function)

A twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ is convex if and only if for all x,

$$f''(x) \geq 0.$$

We will discuss the high-dimensional generalization of this fact after break.

$f : \mathbb{R}^d \to \mathbb{R}$



Assume:

- f is convex.
- Lipschitz function: for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$.
- Starting radius: $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2 \leq R$.

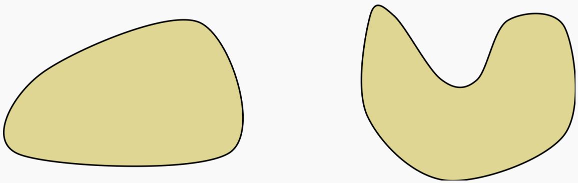
Claim (GD Convergence Bound)

If we run GD for $T \geq \frac{R^2G^2}{\epsilon^2}$ *iterations then* $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

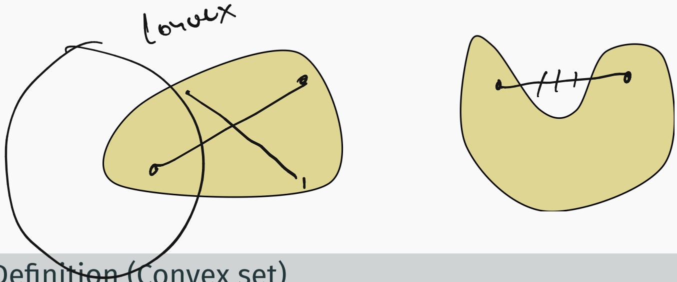
Common goal: Solve a convex minimization problem with additional convex constraints.

$$\min_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x})$$

where $\mathcal{S}$ is a convex set.



Which of these is convex?

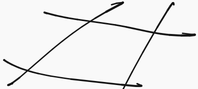


Definition (Convex set)

A set $\mathcal{S}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}, \lambda \in [0, 1]$:

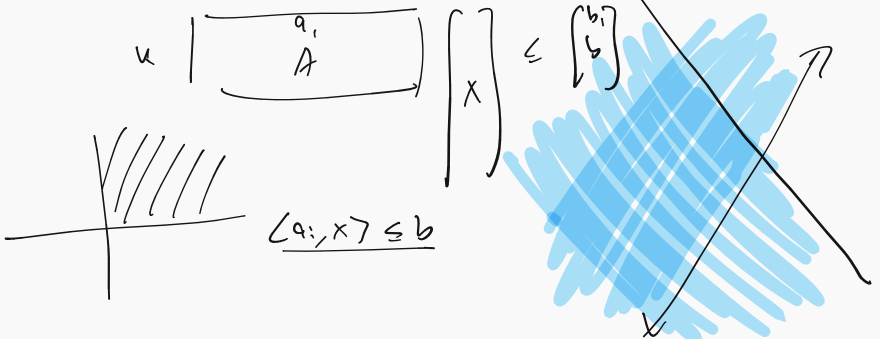
$$(1 - \lambda)\mathbf{x} + \lambda\mathbf{y} \in \mathcal{S}.$$

$S \subseteq \mathbb{R}^d$



Examples:

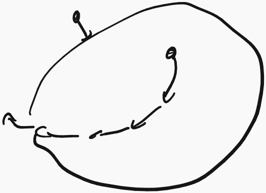
- **Norm constraint:** minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ subject to $\|\mathbf{x}\|_2 \leq \lambda$. Used e.g. for regularization, finding a sparse solution, etc.
- **Positivity constraint:** minimize $f(\mathbf{x})$ subject to $\mathbf{x} \geq 0$.
- **Linear constraint:** minimize $\mathbf{c}^T\mathbf{x}$ subject to $\mathbf{A}\mathbf{x} \leq \mathbf{b}$.



$\min_{S} f(x)$

Gradient descent:

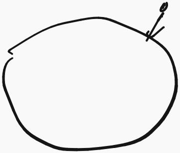
- For $i = 0, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_i f(\mathbf{x}^{(i)})$.



Even if we start with $\mathbf{x}^{(0)} \in \mathcal{S}$, there is no guarantee that $\mathbf{x}^{(0)} - \eta \nabla f(\mathbf{x}^{(0)})$ will remain in our set.

Extremely simple modification: Force $\mathbf{x}^{(i)}$ to be in $\mathcal{S}$ by **projecting** onto the set.

Given a function f to minimize and a convex constraint set $\mathcal{S}$, assume we have:



- **Function oracle**: Evaluate $f(\mathbf{x})$ for any $\mathbf{x}$.
- **Gradient oracle**: Evaluate $\nabla f(\mathbf{x})$ for any $\mathbf{x}$.
- **Projection oracle**: Evaluate $P_{\mathcal{S}}(\mathbf{x})$ for any $\mathbf{x}$.

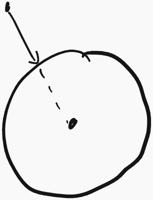
$$P_{\mathcal{S}}(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathcal{S}} \|\mathbf{x} - \mathbf{y}\|_2$$

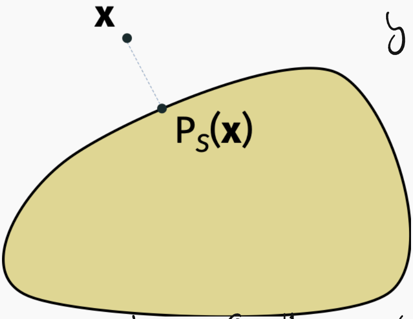
$P_S(x) = x / \|x\|_2$

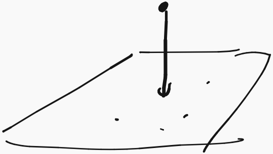
- How would you implement $P_{\mathcal{S}}$ for $\mathcal{S} = \{\mathbf{y} : \|\mathbf{y}\|_2 \leq 1\}$.
- How would you implement $P_{\mathcal{S}}$ for $\mathcal{S} = \{\mathbf{y} : \mathbf{y} = \mathbf{Q}\mathbf{z}\}$.

$\mathbf{z} \in \mathbb{R}^k$, $k < d$

$y = Q\,\beta_2$







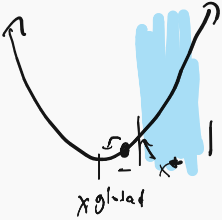
$\min_{y \in S} \|x - y\|_2 = \min_z \|x - Qz\|_2 \qquad z = (Q^T Q)^{-1} Q^T x$

$y = Qz$

Given function $f(\mathbf{x})$ and set $\mathcal{S}$, such that $\|\nabla f(\mathbf{x})\|_2 \leq G$ for all $\mathbf{x} \in \mathcal{S}$ and starting point $\mathbf{x}^{(0)}$ with $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R$.

Projected gradient descent:

- Select starting point $\mathbf{x}^{(0)}$, $\eta = \frac{R}{G\sqrt{T}}$.
- For $i = 0, \ldots, T$:
  - $\mathbf{z} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
  - $\mathbf{x}^{(i+1)} = P_{\mathcal{S}}(\mathbf{z})$
- Return $\hat{\mathbf{x}} = \arg\min_i f(\mathbf{x}^{(i)})$.



Claim (PGD Convergence Bound)

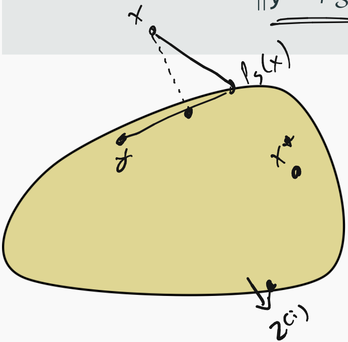
If $f, \mathcal{S}$ are convex and $T \geq \frac{R^2 G^2}{\epsilon^2}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

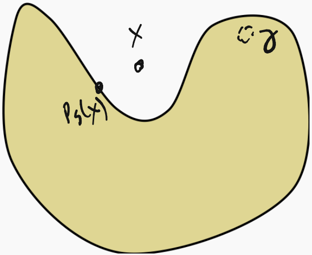
Analysis is almost identical to standard gradient descent! We just need one additional claim:

Claim (Contraction Property of Convex Projection)

If $\mathcal{S}$ is convex, then for <u>any</u> $\mathbf{y} \in \mathcal{S}$,

$$\|\mathbf{y} - P_{\mathcal{S}}(\mathbf{x})\|_2 \leq \|\mathbf{y} - \mathbf{x}\|_2.$$





Claim (PGD Convergence Bound)

If $f, \mathcal{S}$ are convex and $T \geq \frac{R^2G^2}{\epsilon^2}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Claim 1: For all $i = 0, \ldots, T$, let $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} - \eta\nabla f(\mathbf{x}^{(i)})$. Then:

$$f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{z}^{(i)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

$$\leq \frac{\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(i+1)} - \mathbf{x}^*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$$

$x^* = \min_S f(x)$

Same telescoping sum argument:

$$\left[\frac{1}{T}\sum_{i=0}^{T-1} f(\mathbf{x}^{(i)})\right] - f(\mathbf{x}^*) \leq \frac{R^2}{2T\eta} + \frac{\eta G^2}{2}.$$

Conditions:

- **Convexity:** f is a convex function, $\mathcal{S}$ is a convex set.
- **Bounded initial distant:**

$$\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R$$

- **Bounded gradients (Lipschitz function):**

$$\|\nabla f(\mathbf{x})\|_2 \leq G \text{ for all } \mathbf{x} \in \mathcal{S}.$$

Theorem (GD Convergence Bound)

(Projected) Gradient Descent returns $\hat{\mathbf{x}}$ with $f(\hat{\mathbf{x}}) \leq \min_{\mathbf{x}\in\mathcal{S}} f(\mathbf{x}) + \epsilon$ after

$$T = \frac{R^2 G^2}{\epsilon^2} \text{ iterations.}$$

The previous bounds are optimal for convex first order optimization in general.

But in practice, the dependence on $1/\epsilon^2$ is pessimistic: gradient descent typically requires far fewer steps to reach ϵ error.

Previous bounds only make a very weak first order assumption:

$$\|\nabla f(x)\|_2 \leq G.$$

In practice, many function satisfy stronger assumptions.

$f: \mathbb{R} \to \mathbb{R}$ $f''(0) \geq 0$.

Often possible to place assumptions on the second derivative of f.

In particular, we say that a scalar function f is α-strongly convex and β-smooth if for all x:

$$0 < \alpha \leq |f''(x)| \leq \beta.$$

We will give an appropriate generalization of these conditions to multi-dimensional functions shortly.

Take away: Having either an upper and lower bound on the second derivative helps convergence. Having both helps a lot.

Take away: Having either an upper and lower bound on the second derivative helps convergence. Having both helps a lot.

Number of iterations for ϵ error:

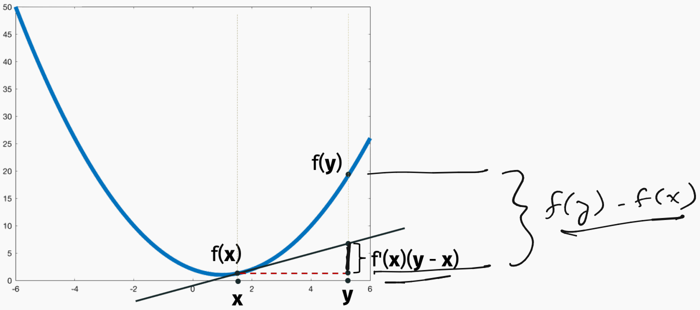
	G-Lipschitz	β-smooth
R bounded start	$O\left(\frac{G^2R^2}{\epsilon^2}\right)$	$O\left(\frac{\beta R^2}{\epsilon}\right)$
α-strong convex	$O\left(\frac{G^2}{\alpha\epsilon}\right)$	$O\left(\frac{\beta}{\alpha}\log(1/\epsilon)\right)$

As we defined them so far, smoothness and strong convexity require f to be twice differentiable. On the other hand, gradient descent only requires first order differentiability.

Equivalent conditions:

$$f''(x) \leq \beta \Rightarrow [f(y) - f(x)] - f'(x)(y - x) \leq \frac{\beta}{2}(y - x)^2$$

$$f''(x) \geq \alpha \Rightarrow [f(y) - f(x)] - f'(x)(y - x) \geq \frac{\alpha}{2}(y - x)^2$$



Recall: For all convex functions $[f(y) - f(x)] - f'(x)(y - x) \geq 0$.

Proof that $f''(x) \leq \beta \Rightarrow [f(y) - f(x)] - f'(x)(y-x) \leq \frac{\beta}{2}(y-x)^2$:

$$f(y) - f(x) = \int_x^y f'(t)\, dt$$

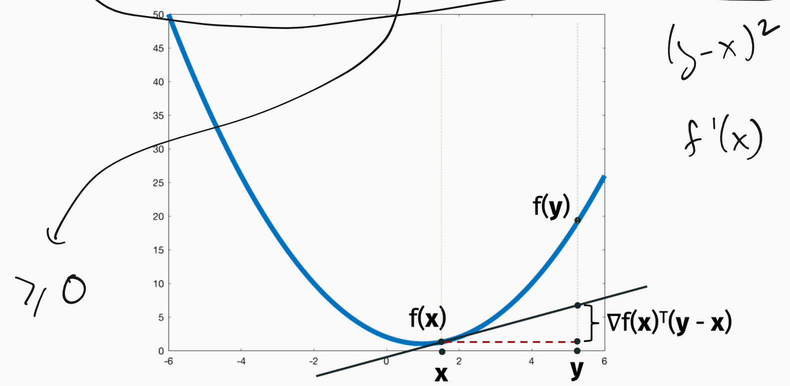
$$\leq \int_x^y f'(x) + \beta(x-t)$$

$$\vdots$$

Proof for α-strongly convex is similar, as are the other directions when f is twice differentiable.

A function is α-strongly convex and β-smooth if for all **x**, **y**:

$$\frac{\alpha}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \leq [f(\mathbf{y})-f(\mathbf{x})] - \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \leq \frac{\beta}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$



Definition (β-smoothness)

A function f is β smooth if and only if, for all $\mathbf{x}, \mathbf{y}$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \beta \|\mathbf{x} - \mathbf{y}\|_2$$

I.e., the gradient function is a β-Lipschitz function.

We won't use this definition directly, but it's good to know. Easy to prove equivalency to previous definition (see Lem. 3.4 in Bubeck's book).

Theorem (GD convergence for β-smooth functions.)

Let f be a β smooth convex function and assume we have $\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_2 \leq R$. If we run GD for T steps, we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \frac{2\beta R^2}{T}$$

Corollary: If $T = O\left(\frac{\beta R^2}{\epsilon}\right)$ we have $f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \epsilon$.

Compare this to $T = O\left(\frac{G^2 R^2}{\epsilon^2}\right)$ without a smoothness assumption.

Why do you think gradient descent might be faster when a function is β-smooth?

Previously learning rate/step size η depended on G. Now choose it based on β:

$$\mathbf{x}^{(t+1)} \leftarrow \mathbf{x}^{(t)} - \frac{1}{\beta}\nabla f(\mathbf{x}^{(t)})$$

Progress per step of gradient descent:

1. $\left[f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)})\right] - \nabla f(\mathbf{x}^{(t)})^T(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}) \leq \frac{\beta}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\|_2^2$.

2. $\left[f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)})\right] + \frac{1}{\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2 \leq \frac{\beta}{2}\|\frac{1}{\beta}\nabla f(\mathbf{x}^{(t)})\|_2^2$.

3. $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t+1)}) \geq \frac{1}{2\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2$.

Once we have the bound from the previous page, proving a convergence result isn't hard, but not obvious. A concise proof can be found in Page 15 in Garrigos and Gower's notes.

Theorem (GD convergence for β-smooth functions.)

Let f be a β smooth convex function and assume we have $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$. If we run GD for T steps with $\eta = \frac{1}{\beta}$ we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \frac{2\beta R^2}{T}$$

Corollary: If $T = O\left(\frac{\beta R^2}{\epsilon}\right)$ we have $f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \epsilon$.

Note: This is not optimal! Can be improved to depend on $O(1/T^2)$ using a technique called acceleration.

Where did we use convexity in this proof?

Progress per step of gradient descent:

1. $\left[f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)})\right] - \nabla f(\mathbf{x}^{(t)})^T(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}) \leq \frac{\beta}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\|_2^2$.
2. $\left[f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^{(t)})\right] + \frac{1}{\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2 \leq \frac{\beta}{2}\|\frac{1}{\beta}\nabla f(\mathbf{x}^{(t)})\|_2^2$.
3. $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t+1)}) \geq \frac{1}{2\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2$.

Definition (Stationary point)

For a differentiable function f, a stationary point is any $\mathbf{x}$ with:

$$\nabla f(\mathbf{x}) = \mathbf{0}$$

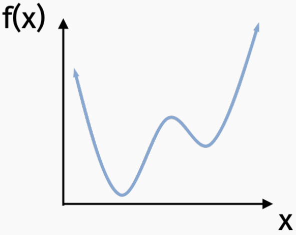
local/global minima - local/global maxima - saddle points

Theorem (Convergence to Stationary Point)

For any β-smooth differentiable function f (convex or not), if we run GD for T steps, we can find a point $\hat{\mathbf{x}}$ such that:

$$\|\nabla f(\hat{\mathbf{x}})\|_2^2 \leq \frac{2\beta}{T}\left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)\right)$$

Corollary: If $T \geq \frac{2\beta}{\epsilon}$, then $\|\nabla f(\hat{\mathbf{x}})\|_2^2 \leq \epsilon\left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)\right)$.



Theorem (Convergence to Stationary Point)

For any β-smooth differentiable function f (convex or not), if we run GD for T steps, we can find a point $\hat{\mathbf{x}}$ such that:

$$\|\nabla f(\hat{\mathbf{x}})\|_2^2 \leq \frac{2\beta}{T}\left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)\right)$$

We have that $\frac{1}{2\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2 \leq f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t+1)})$. So:

$$\sum_{t=0}^{T-1} \frac{1}{2\beta}\|\nabla f(\mathbf{x}^{(t)})\|_2^2 \leq f(\mathbf{x}^{(0)}) - f(\mathbf{x}^{(t)})$$

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}^{(t)})\|_2^2 \leq \frac{2\beta}{T}\left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)\right)$$

$$\min_t \|\nabla f(\mathbf{x}^{(t)})\|_2^2 \leq \frac{2\beta}{T}\left(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^*)\right)$$

I said it was a bit tricky to prove that $f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \frac{2\beta R^2}{T}$ for convex functions. But we just easily proved that $\|\nabla f(\hat{\mathbf{x}})\|_2^2$ is small. Why doesn't this show we are close to the minimum?

Definition (α-strongly convex)

A convex function f is α-strongly convex if, for all $\mathbf{x}, \mathbf{y}$

$$[f(\mathbf{y}) - f(\mathbf{x})] - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \geq \frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

Compare to smoothness condition.

$$[f(\mathbf{y}) - f(\mathbf{x})] - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \leq \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|_2^2.$$

For a twice-differentiable scalar function f, equivalent to $f''(x) \geq \alpha$.

When f is convex, we always have that $f''(x) \geq 0$, so larger values of α correspond to a "stronger" condition.

Gradient descent for strongly convex functions:

- Choose number of steps T.
- For $i = 0, \ldots, T$:
  - $\eta = \frac{2}{\alpha \cdot (i+1)}$
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}^{(i)}} f(\mathbf{x}^{(i)})$.

Theorem (GD convergence for α-strongly convex functions.)

Let f be an α-strongly convex function and assume we have that, for all $\mathbf{x}$, $\|\nabla f(\mathbf{x})\|_2 \leq G$. If we run GD for T steps (with adaptive step sizes) we have:

$$f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \frac{2G^2}{\alpha T}$$

Corollary: If $T = O\left(\frac{G^2}{\alpha\epsilon}\right)$ we have $f(\hat{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon$

We could also have that f is both β-smooth and α-strongly convex.

Theorem (GD for β-smooth, α-strongly convex.)

Let f be a β-smooth and α-strongly convex function. If we run GD for T steps (with step size $\eta = \frac{1}{\beta}$) we have:

$$\|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2 \leq e^{-T\frac{\alpha}{\beta}}\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$$

$\kappa = \frac{\beta}{\alpha}$ is called the "condition number" of f.

Is it better if κ is large or small?

Converting to more familiar form: Using that fact the $\nabla f(\mathbf{x}^*) = \mathbf{0}$ along with

$$\frac{\alpha}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq [f(\mathbf{y}) - f(\mathbf{x})] - \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \leq \frac{\beta}{2}\|\mathbf{x}-\mathbf{y}\|_2^2,$$

we have:

$$\frac{2}{\beta}\left[f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*)\right] \leq \|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2^2$$

We also assume

$$\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2 \leq R^2.$$

Corollary (GD for β-smooth, α-strongly convex.)

Let f be a β-smooth and α-strongly convex function. If we run GD for T steps (with step size $\eta = \frac{1}{\beta}$) we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \frac{\beta}{2} e^{-T\frac{\alpha}{\beta}} \cdot R^2$$

Corollary: If $T = O\left(\frac{\beta}{\alpha}\log(R\beta/\epsilon)\right)$ we have:

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \epsilon$$

Only depend on $\log(1/\epsilon)$ instead of on $1/\epsilon$ or $1/\epsilon^2$!

After break or on homework we will prove the guarantee for the special case of:

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

Goal: Get some of the key ideas across, introduces important concepts like the Hessian, and show the connection between conditioning and linear algebra.

But first we will talk about online gradient descent and stochastic gradient descent next week.