

# CS-GY 9223 I: Lecture 6

## Smoothness, Strong convexity, and more.

---

NYU Tandon School of Engineering, Prof. Christopher Musco

# GRADIENT DESCENT ANALYSIS

Assume:

- $f$  is convex.
- Lipschitz function: for all  $\mathbf{x}$ ,  $\|\nabla f(\mathbf{x})\|_2 \leq G$ .
- Starting radius:  $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$ .

$$\|f(\mathbf{x}) - f(\mathbf{y})\|_2 \leq G \cdot \|\mathbf{x} - \mathbf{y}\|_2$$

$$\min \|A\mathbf{x} - \mathbf{b}\|_2^2$$

$$f(\mathbf{x})$$

$$\nabla f(\mathbf{x}) = 2\mathbf{x}^T A^T A \mathbf{x}$$

Gradient descent:

- Choose number of steps  $T$ .
- $\eta = \frac{R}{G\sqrt{T}}$
- For  $i = 1, \dots, T$ :
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f(\mathbf{x}^{(i)})$
- Return  $\hat{\mathbf{x}} = \arg \min_{\mathbf{x}^{(i)}} f(\mathbf{x}^{(i)})$ .

## Theorem (GD Convergence Bound)

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

Instead of a single function  $f$  to minimize, assume we have an unknown and changing set of objective functions:

$$\underline{f_1}, \dots, \underline{f_T}.$$

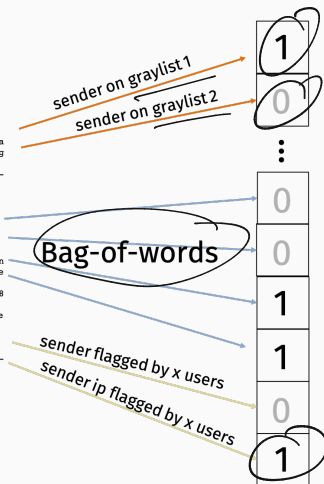
- At each time step, choose  $\mathbf{x}^{(i)}$
- $f_i$  is revealed and we pay cost  $f_i(\mathbf{x}^{(i)})$
- **Goal:** Minimize  $\sum_{i=1}^T f_i(\mathbf{x}^{(i)})$ .



# EXAMPLE

## Email spam filtering:

```
MIME-Version: 1.0 Date: Mon, 7 Oct 2019
14:51:30 -0400 Message-ID: <CANVPizUGqx=B-
39MLAnOPyJ9_jxaX60QmHWb4QCFBPgNDzA@mail.gma
il.com> Subject: 9223i Reading Group, Meeting
2, tomorrow at 10am From: Christopher Musco
<cmusco@nyu.edu> To: algmls@nyu.edu Content-
Type: multipart/alternative;
boundary="0000000000078ec240594568a53" --
0000000000078ec240594568a53 Content-Type:
text/plain; charset="UTF-8" I hope everyone
had a good weekend! Tomorrow at *10am in 370
Jay St. #1114* we will meet for the second
instantiation of the CS-GY 9223i reading
group. Nick Feng will be leading a discussion
about the paper Simple Analyses of the Sparse
Johnson-Lindenstrauss Transform
<http://drops.dagstuhl.de/opus/volltexte/2018
/8305/pdf/OASlcs-SOSA-2018-15.pdf>. Please
read the abstract and introduction before the
meeting. Best, - CM *Christopher Musco,
Assistant Professor* *New York University,
Tandon School of Engineering* *(401) 578
2541* --0000000000078ec240594568a53 Content-
Type: text/html; charset="UTF-8" Content-
Transfer-Encoding: quoted-printable
```



# SPAM FILTERING

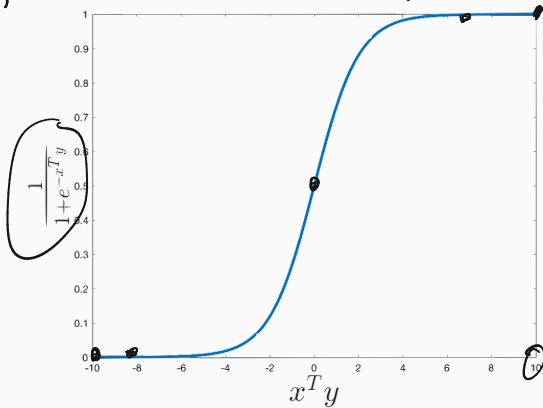
•  $M_x(y) = \frac{1}{1 + e^{-x^T y}}$

data vector

parameter

$$x^T y = \sum_{i=1}^d x_i y_i \in (-10, 10)$$

weight      20, 13



Predict  $y$  as spam if  $M_x(y) \geq \frac{1}{2}$ .

# SPAM FILTERING

Logistic loss:

Given label  $b \in \{0, 1\}$ ,

$$\begin{aligned} \underline{b} = 1 & \quad -\log(M_x(y)) \\ & = \log(1/M_x(y)) \end{aligned}$$

$$\underline{b} = 0$$

$$L(\underline{b}, \underline{M_x(y)}) = -b \log(M_x(y)) + (1-b) \log(1 - M_x(y))$$

Total cost of over time:

$$\sum_{i=1}^T L(\underline{b}^{(i)}, \underline{M_{x^{(i)}}(y^{(i)})})$$

approximation to # of mistakes

$f_i(x^{(i)})$

where  $y^{(i)}$  is the  $i^{\text{th}}$  email and  $b^{(i)}$  is the  $i^{\text{th}}$  label.

we can

$$\sum_{i=1}^T f_i(x^{(i)}) \rightarrow \text{solve w/ online gradient descent.}$$

## REGRET BOUND

How should we measure how well we did?

For some small value  $\Delta$ , can we achieve:

$$\underbrace{\sum_{i=1}^T f_i(x^{(i)})}_{\text{what?}} \leq \left[ \min_x \sum_{i=1}^T f_i(x) \right] + \underline{\underline{\Delta}}$$

I.e. can we compete with the best fixed solution in hindsight.

$\Delta = \text{"regret"}$

$$\sum_{i=1}^T f_i(x^{(i)}) \leq \Delta + \sum_{i=1}^T f_i(x^{(i)*}) = 0$$

best choice of weights  
at time  $i$

# ONLINE GRADIENT DESCENT

Assume:

- Lipschitz functions: for all  $\mathbf{x}, i, \|\nabla f_i(\mathbf{x})\|_2 \leq G$ .
- Starting radius:  $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$ .

Online Gradient descent:

- Choose number of steps  $T$ .
- $\eta = \frac{D}{G\sqrt{T}}$
- For  $i = 1, \dots, T$ :
  - $\mathbf{x}^{(i+1)} = \underline{\mathbf{x}^{(i)}} - \eta \underline{\nabla f_i(\mathbf{x}^{(i)})}$
- Play  $\mathbf{x}^{(i+1)}$ .

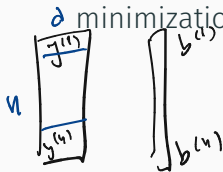
Claim (OGD Regret Bound)

$$\text{After } T \text{ steps, } \Delta = \frac{1}{T} \left[ \sum_{i=1}^T f_i(\mathbf{x}^{(i)}) \right] - \frac{1}{T} \left[ \sum_{i=1}^T f_i(\mathbf{x}^*) \right] \leq \frac{RG}{\sqrt{T}}$$



# STOCHASTIC GRADIENT DESCENT

Recall the machine learning setup. In empirical risk minimization, we can typically write:



$$\underline{f(x)} = \sum_{j=1}^n \underline{f_j(x)}$$

$$\nabla f(x) = \sum_{j=1}^n \nabla f_j(x)$$

$$\begin{aligned} & \nabla (x^T y^{(j)} - b^{(j)})^2 \\ &= 2(x^T y^{(j)} - b^{(j)}) \cdot y^{(j)} \end{aligned}$$

Annotations: An arrow points from the term  $(x^T y^{(j)} - b^{(j)})^2$  in the first line to the label "data point scalar" in the second line. Another arrow points from the term  $y^{(j)}$  in the second line to the label "vector" in the second line.

where  $f_j$  is the loss function for a particular data point.

Linear regression:

$$f(x) = \|Yx - b\|_2^2$$

$$\nabla f(x) = 2Y^T(Yx - b)$$

$$\nabla f(x) = \sum_{j=1}^n 2(x^T y^{(j)} - b^{(j)}) \cdot y^{(j)}$$

$$f(x) = \sum_{j=1}^n (x^T y^{(j)} - b^{(j)})^2$$

Annotations for the linear regression equation:

- An arrow points from the label "data point" to the term  $(x^T y^{(j)} - b^{(j)})^2$ .
- An arrow points from the label "output value" to the term  $b^{(j)}$ .
- An arrow points from the label "loss function" to the term  $(x^T y^{(j)} - b^{(j)})^2$ .
- An arrow points from the label "data point scalar" to the term  $y^{(j)}$ .
- An arrow points from the label "vector" to the term  $y^{(j)}$ .

## STOCHASTIC GRADIENT DESCENT

Pick random  $j \in 1, \dots, n$ :

$$\nabla f(x) = \sum \nabla f_j(x)$$
$$n \mathbb{E} [\nabla f_j(x)] = \nabla f(x) \xrightarrow{\mathbb{E}[\nabla f_j(x)]} \text{time } O(d)$$

But  $\nabla f_j(x)$  can often be computed in a  $1/n$  fraction of the time!

time  $O(d)$

**Main idea:** Use random approximate gradient in place of actual gradient.

Trade slower convergence for cheaper iterations.

Runtime of algo = (time per step)  $\cdot$  (# of iterations) 10

# STOCHASTIC GRADIENT DESCENT

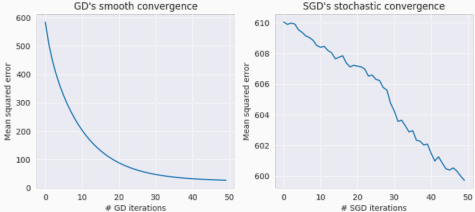
Assume:

- Lipschitz functions: for all  $\mathbf{x}, j$ ,  $\|\nabla f_j(\mathbf{x})\|_2 \leq \frac{G'}{n}$ .
- Starting radius:  $\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R$ .

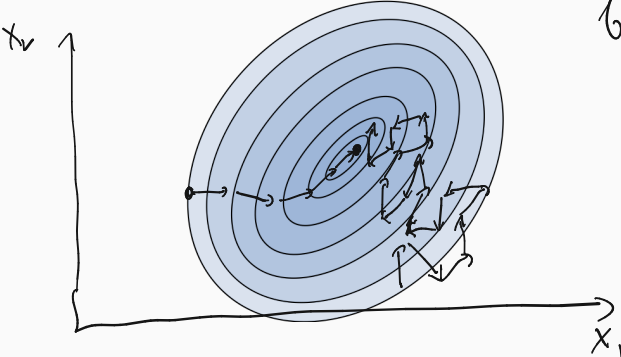
Stochastic Gradient descent:

- Choose number of steps  $T$ .
- $\eta = \frac{D}{G'\sqrt{T}}$
- For  $i = 1, \dots, T$ :
  - Pick random  $j_i \in 1, \dots, n$ .
  - $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta \nabla f_{j_i}(\mathbf{x}^{(i)})$   $\rightarrow$  stochastic gradient
- Return  $\hat{\mathbf{x}} = \frac{1}{T} \sum_{i=1}^T \mathbf{x}^{(i)}$

# VISUALIZING SGD



$\theta = [x_1, x_2]$



# STOCHASTIC GRADIENT DESCENT ANALYSIS

## Claim (SGD Convergence)

After  $T = \frac{R^2 G^2}{\epsilon^2}$  iteration:

↗ random #

$$\mathbb{E}[f(\hat{x}) - f(x^*)] \leq \epsilon.$$

with prob.  $9/10$

$$f(\hat{x}) - f(x^*) \leq 10\epsilon$$

$$\mathbb{E}[f(\hat{x}) - f(x^*)] = \mathbb{E}\left[f\left(\frac{1}{T} \sum_{i=1}^T x^{(i)}\right) - f(x^*)\right]$$

$$\leq \mathbb{E}\left[\frac{1}{T} \sum_{i=1}^T f(x^{(i)}) - f(x^*)\right]$$

$$\leq \frac{1}{T} \sum_{i=1}^T \mathbb{E} \nabla f(x^{(i)})^\top (x^{(i)} - x^*)$$

$$= \frac{1}{T} \sum_{i=1}^T n \mathbb{E} g_i^\top (x^{(i)} - x^*)$$

$$= \frac{n}{T} \mathbb{E} \sum_{i=1}^T h(x^{(i)}) - h(x^*)$$

$$\leq R \frac{G}{n} \sqrt{T}$$

$g_i$  = the gradient at step  $i$

$$= \nabla f_i(x^{(i)})$$

$$h_i(y) = g_i^\top y$$

- convex

$$\nabla h_i(y) = g_i$$

↗ "regret" for OGD when our functions are  $h_1, \dots, h_T$

## Claim (SGD Convergence)

After  $T = \frac{R^2 G^2}{\epsilon^2}$  iteration:

$$\mathbb{E}[f(\hat{x}) - f(x^*)] \leq \epsilon.$$

$$\begin{aligned} \mathbb{E}[f(\hat{x}) - f(x^*)] &\leq \frac{\eta}{T} B \cdot \frac{G^2}{\eta} \cdot \sqrt{T} \\ &= \frac{B \cdot G^2}{\sqrt{T}} \leq \epsilon \end{aligned}$$

$$T = \frac{B^2 G^2}{\epsilon^2}$$

## COMPARISON

Number of iterations for error  $\epsilon$ :  $\|\nabla f_1(x) + \nabla f_2(x) + \dots\|_2$

• Gradient Descent:  $T = \frac{R^2 G^2}{\epsilon^2}$ .

• Stochastic Gradient Descent:  $T = \frac{R^2 G'^2}{\epsilon^2}$ .

$$\begin{aligned}\|x + \delta\|_2 &\leq \|x\|_2 + \|\delta\|_2 \\ &= 2\|x\|_2\end{aligned}$$

Always have  $G \leq G'$ :

$$\|\nabla f(x)\|_2 \leq \|\nabla f_1(x)\|_2 + \dots + \|\nabla f_n(x)\|_2 \leq n \cdot \left(\frac{G'}{n}\right) = \underline{G'}$$

Fair comparison:

• SGD cost = (# of iterations)  $\cdot O(1)$

• GD cost = (# of iterations)  $\cdot O(n)$

Cheap when  $G'$  is not much greater than  $G$ .

$$\nabla f_1(x) \approx \nabla f_2(x) \approx \nabla f_3(x)$$

Stochastic vs. Full Batch Gradient Descent:



Can the convergence bounds be tightened for certain functions? Can they guide us towards faster algorithms?

### Goals:

- Improve  $\epsilon$  dependence below  $1/\epsilon^2$ .
- Reduce or eliminate dependence on  $G$  and  $R$ .
- Etc.

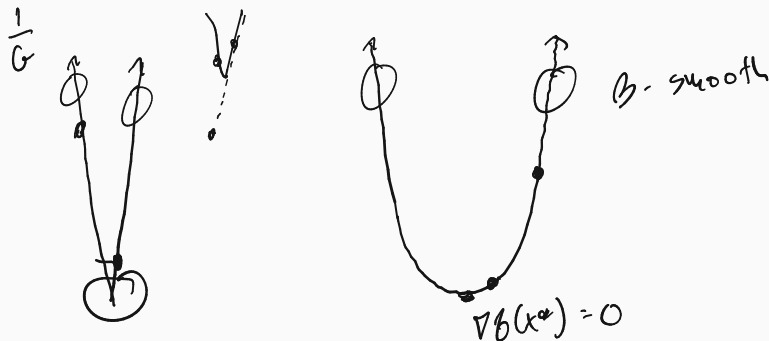
# SMOOTHNESS

## Definition ( $\beta$ -smoothness)

A function  $f$  is  $\beta$  smooth if, for all  $x, y$  Lipschitz gradients

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

$\beta$  is a parameter that will depend on our function.



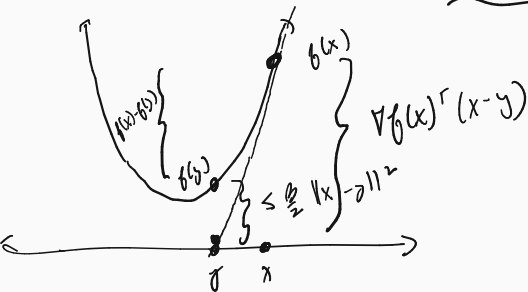
# SMOOTHNESS

Recall from definition of convexity that:

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

How much smaller can left hand side be?

$$\boxed{\nabla f(x)^T (x - y)} - \boxed{f(x) - f(y)} \leq \underline{\underline{\frac{\beta}{2} \|x - y\|_2^2}}$$



## GUARANTEED PROGRESS

Previously learning rate/step size  $\eta$  depended on  $G$ . Now choose it based on  $\beta$ :

$$\eta = \frac{1}{\beta}$$

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

Progress per step of gradient descent:

$$\underbrace{\nabla f(\mathbf{x}^{(t)})^\top}_{\frac{1}{\beta} \nabla f(\mathbf{x}^{(t)})} (\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}) - [f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t+1)})] \leq \underbrace{\frac{\beta}{2} \|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\|_2^2}_{\frac{1}{\beta} \nabla f(\mathbf{x}^{(t)})}$$

$$\frac{1}{\beta} \|\nabla f(\mathbf{x}^{(t)})\|_2^2 - [f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t+1)})] \leq \underbrace{\frac{\beta}{2} \cdot \|\frac{1}{\beta} \nabla f(\mathbf{x}^{(t)})\|_2^2}_{\frac{1}{2\beta}}$$

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

# CONVERGENCE GUARANTEE

Theorem (GD convergence for  $\beta$ -smooth functions.)

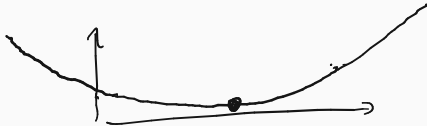
Let  $f$  be a  $\beta$  smooth convex function and assume we have

$\|x^* - x^{(1)}\|_2 \leq R$ . If we run GD for  $T$  steps with  $\eta = \frac{1}{\beta}$  we have:

$$f(x^{(T)}) - f(x^*) \leq \frac{2\beta R^2}{T-1}$$

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

$$T = O\left(\frac{G^2 B^2}{\epsilon^2}\right)$$



## STRONG CONVEXITY

$\alpha$ -strong convexity  
Definition (~~smoothness~~)

A ~~convex~~ function  $f$  is  $\alpha$  ~~smooth~~  
 $\alpha$ -strongly convex if, for all  $x, y$

$$\underline{f(y)} \geq \underline{f(x)} + \nabla f(x)^T (y - x) + \frac{\alpha}{2} \|x - y\|_2^2$$

$\alpha$  is a parameter that will depend on our function.

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

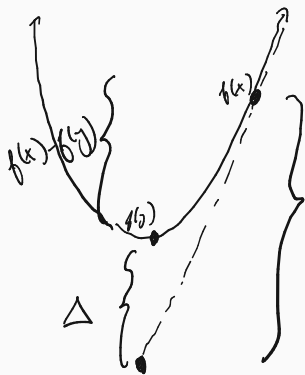
$$f(x) - f(y) \leq \nabla f(x)^T (x - y) - \frac{\alpha}{2} \|x - y\|_2^2$$

---

# STRONG CONVEXITY

Completing the picture: If  $f$  is  $\alpha$  strongly convex and  $\beta$  smooth,

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



$\nabla f(x)^T(x - y)$



small strong convexity  $\alpha$

## Gradient descent for strongly convex functions:

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



## CONVERGENCE GUARANTEE

Theorem (GD convergence for  $\alpha$ -strongly convex functions.)

Let  $f$  be an  $\alpha$ -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:

$$\underline{f(\hat{\mathbf{x}})} - \underline{f(\mathbf{x}^*)} \leq \frac{2G^2}{\alpha(T-1)}$$

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

$$O\left(\frac{G^2 B^2}{\epsilon^2}\right)$$

## SMOOTH AND STRONGLY CONVEX

What if  $f$  is both  $\beta$ -smooth and  $\alpha$ -strongly convex?

$$T \sim \beta \quad T \sim \frac{1}{\alpha} \quad T \sim \frac{\beta}{\alpha}$$

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

What if  $\alpha = \beta$ :

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

$$\underline{f(x) - f(y)} = \nabla f(x)^T(x - y) - \frac{\beta}{2} \|x - y\|_2^2$$

$$h(y) = \nabla f(x)^T(x - y) - \frac{\beta}{2} \|x - y\|_2^2$$

$$x^* = \arg \max_y h(y)$$

## SMOOTH AND STRONGLY CONVEX

$$x^* = x - \frac{1}{\beta} \nabla f(x) \quad (\text{optimize in 1 step.})$$

What if  $f$  is both  $\beta$ -smooth and  $\alpha$ -strongly convex?

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

What if  $\alpha = \beta$ :

$$h(y) = \nabla f(x)^T (x - y) - \frac{\beta}{2} \|x - y\|_2^2$$

$$x^* = \arg \max_y h(y) \quad -\frac{\beta}{2} x^T x + \beta x^T y - \frac{\beta}{2} y^T y$$

$$\nabla h(y) = -\nabla f(x) + \beta x - \beta y$$

$$\nabla h(y) = 0 \quad \text{for maximization of } y$$

$$0 = \nabla f(x) + \beta x - \beta y$$

$$\boxed{y^* = x - \frac{1}{\beta} \nabla f(x)} \quad 27$$

## CONVERGENCE GUARANTEE

Theorem (GD for  $\beta$ -smooth,  $\alpha$ -strongly convex.)

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

$$\underbrace{\|x^{(t)} - x^*\|_2^2}_{\text{B}} \leq \underbrace{e^{-\underbrace{(T-1)\frac{\alpha}{\beta}}_{\text{B}}}}_{\text{B}} \underbrace{\|x^{(1)} - x^*\|_2^2}_{\text{B}}$$

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

Is it better if  $\kappa$  is large or small?

$$T = \underbrace{\left( \frac{\beta}{\alpha} \log(1/\epsilon) \right)}_{\text{“condition number”}} \rightarrow \|x^{(t)} - x^*\|_2^2 \leq \epsilon \|x^{(1)} - x^*\|_2^2$$

# SMOOTH AND STRONGLY CONVEX

Converting to more familiar form:

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

$$x = x^* \quad y = x^{(t)}$$

$$\frac{\alpha}{2} \|x^* - x^{(t)}\|_2^2 \leq \underbrace{\nabla f(x^*)^T (x - x^{(t)})}_0 + f(x^{(t)}) - f(x^*) \leq \frac{\beta}{2} \|x^* - x^{(t)}\|_2^2$$

$$\frac{\alpha}{2} \|x^* - x^{(t)}\|_2^2 \leq f(x^{(t)}) - f(x^*) \leq \frac{\beta}{2} \|x^* - x^{(t)}\|_2^2$$

$$\leq \frac{\beta}{2} \cdot (e^{-1} \cdot \|\cdot\|)$$

## CONVERGENCE GUARANTEE

Corollary (GD for  $\beta$ -smooth,  $\alpha$ -strongly convex.)

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

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{\beta}{2} e^{-(T-1)\frac{\alpha}{\beta}} \left[ f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*) \right]$$

$\|\mathbf{x}^{(1)} - \mathbf{x}^*\|_2 = R$

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

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

logologies for  
types on this  
page

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

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^*) \leq \epsilon \left[ \underline{f(\mathbf{x}^{(1)}) - f(\mathbf{x}^*)} \right]$$

## UNDERSTANDING CONDITIONING

Let  $f(x) = \|\underline{D}x - \mathbf{b}\|_2^2$  where  $\mathbf{D}$  is a diagonal matrix. For now

imagine we're in two dimensions:  $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ ,  $\mathbf{D} = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$ .

$$\min \|Dx - b\|_2^2$$

$$f(x) = x^T D^T x - 2x^T D b + \cancel{b^T b}$$

$$\nabla f(x) = 2D^2 x - 2Db = 2D(Dx - b),$$

What is  $\beta$  for  $f(x) = \|Dx - b\|_2^2$ ?

In other words: What is smallest  $\beta$  so that for all  $x, y$ ,

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

$$\|2D(Dx - b) - 2D(Dy - b)\|_2 \leq \beta \|x - y\|_2$$

$$\|2D^2x - 2D^2y\|_2 \leq \beta \|x - y\|_2$$

$$\|2D^2(x - y)\|_2 \leq \beta \|x - y\|_2.$$

Smallest we can choose  $\beta$ ?

Worst case is when  $x - y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$  or  $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ .

$$\rightarrow \text{need } \beta = 2 \max(d_1, d_2) = 2 \underline{\underline{\max(D)}}.$$



# UNDERSTANDING CONDITIONING

What is  $\alpha$  for  $f(x) = \|\underline{D}x - \underline{b}\|_2^2$ ?

In other words: What is largest  $\alpha$  so that for all  $x, y$ ,

$$\frac{\alpha}{2} \|x - y\|_2^2 \leq \underbrace{\nabla f(x)^T}_{2D(Dx - b)^T} \underbrace{(x - y)}_{(x - y)} - [f(x) - f(y)]$$

$$2D(Dx - b)^T (x - y) - [x^T D^2 x - 2x^T D b - b^T b - y^T D^2 y + 2y^T D b + b^T b]$$

$$\frac{\alpha}{2} \|x - y\|_2^2 \leq 2x^T D^2 x - 2x^T D b - 2x^T D^2 y + 2y^T D b - \{ \dots \}$$

$$= x^T D^2 x - 2x^T D^2 y + y^T D^2 y$$

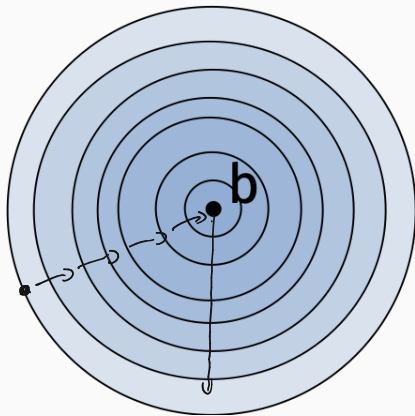
$$= \|\underline{D}(x - y)\|_2^2$$

What's the largest  $\alpha$  so that

$$\frac{\alpha}{2} \|x - y\|_2^2 \leq \|\underline{D}(x - y)\|_2^2 ?$$

$$\boxed{\alpha = 2 \min(\lambda(D))}$$

# UNDERSTANDING CONDITIONING



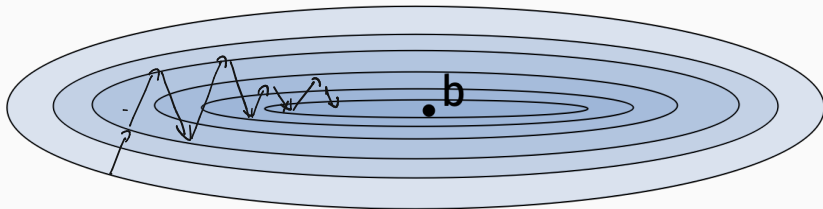
Level sets of  $\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$  when  $d_1 = 1, d_2 = 1$ .

$$\|\mathbf{x} - \mathbf{b}\|_2^2$$

$$u = 1$$

$$\text{norm}(d_1, d_2) \\ = \max(d_1, d_2)$$

# UNDERSTANDING CONDITIONING



Level sets of  $\|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$  when  $d_1 = \frac{1}{3}, d_2 = 2$ .

$$\beta = 2$$
$$\alpha = \frac{1}{3} \quad \kappa = \frac{2}{1/3} = 6$$

Steps to convergence  $\approx O(\kappa \log(1/\epsilon)) = O\left(\frac{\max(\mathbf{D}^2)}{\min(\mathbf{D}^2)} \log(1/\epsilon)\right)$ .

For general regression problems  $\|\mathbf{Ax} - \mathbf{b}\|_2^2$ ,

$$\beta = \lambda_{\max}(\mathbf{A}^T \mathbf{A})$$

$$\alpha = \lambda_{\min}(\mathbf{A}^T \mathbf{A})$$

## IN-CLASS EXERCISE

Theorem (GD for  $\beta$ -smooth,  $\alpha$ -strongly convex.)

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

$$\|x^{(t)} - x^*\|_2^2 \leq e^{-(t-1)\frac{\alpha}{\beta}} \|x^{(1)} - x^*\|_2^2$$

Prove for  $f(x) = \underbrace{\|Dx - b\|_2^2}$ .

## IN-CLASS EXERCISE

## IN-CLASS EXERCISE