CS-GY 9223 I: Lecture 6
Smoothness, Strong convexity, and more.

NYU Tandon School of Engineering, Prof. Christopher Musco

GRADIENT DESCENT ANALYSIS

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

Assume:

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

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

$f(x)$

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

Gradient descent:

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

Theorem (GD Convergence Bound)

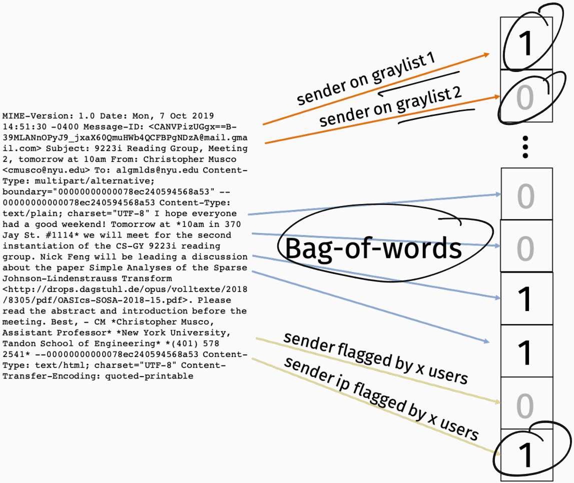
If $T \geq \frac{R^2 G^2}{\epsilon^2}$, then $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

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

$$f_1, \ldots, f_T.$$

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

Email spam filtering:



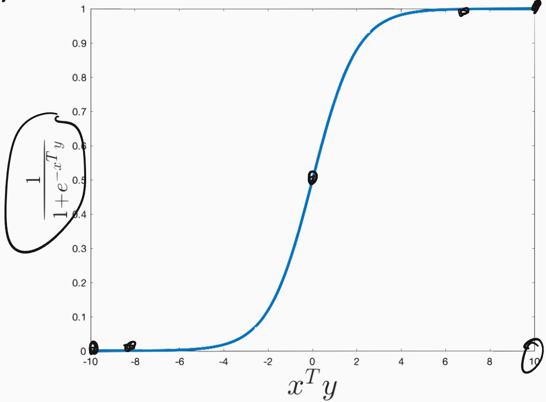
- $M_{\mathbf{x}}(\mathbf{y}) = \frac{1}{1+e^{-\mathbf{x}^T\mathbf{y}}}$

data vector

parameters

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

weight $\{0, 1\}$



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

$b=1$ $\quad -\log(M_x(y))$

$= \log(1/M_x(y))$

Logistic loss:

$b=0$

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

$$L(b, M_{\mathbf{x}}(\mathbf{y})) = -b\log\left(M_{\mathbf{x}}(\mathbf{y})\right) + (1-b)\log\left(1-M_{\mathbf{x}}(\mathbf{y})\right)$$

Total cost of over time:

approximation to # of mistakes

$$\sum_{i=1}^{T} \underbrace{L(b^{(i)}, M_{\mathbf{x}^{(i)}}(\mathbf{y}^{(i)}))}_{f_i(x^{(i)})}$$

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

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

How should we measure how well we did?

For some small value Δ, can we achieve:

$$\underbrace{\sum_{i=1}^{T} f_i(\mathbf{x}^{(i)})}_{\text{[illegible]}} \leq \left[\min_{\mathbf{x}} \sum_{i=1}^{T} f_i(\mathbf{x})\right] + \Delta.$$

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

$$\boxed{\Delta = \text{“regret”}}$$

$$\sum_{i=1}^{T} f_i(x^{(i)}) \leq \Delta + \sum_{i=1}^{T} \overbrace{f_i(\underbrace{x^{(i)*}}_{\text{best choice of weights at time } i})}^{=0}$$

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 $i = 1, \ldots, T$:
  - $\mathbf{x}^{(i+1)} = \underline{\mathbf{x}^{(i)}} - \eta \underline{\nabla f_i(\mathbf{x}^{(i)})}$
- Play $\mathbf{x}^{(i+1)}$.

Claim (OGD Regret Bound)

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

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

d, n, $y^{(1)}$, $y^{(n)}$, $b^{(1)}$, $b^{(n)}$

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

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

$$\nabla \left(x^T y^{(j)} - b^{(j)}\right)^2 = 2\left(x^T y^{(j)} - b^{(j)}\right) \cdot y^{(j)}$$

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

scalar vector

data point

Linear regression:

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

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

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

output value

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

convex funct $f_j(x)$

$\nabla f(x) = \Sigma \nabla f_i(x)$

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

$\mathbb{E}[\nabla f(x)]$

$$n\,\mathbb{E}\left[\nabla f_j(\mathbf{x})\right] = \nabla f(\mathbf{x}). \longrightarrow \text{time } O(nd)$$

But $\nabla f_j(\mathbf{x})$ 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) · (# of iterations)

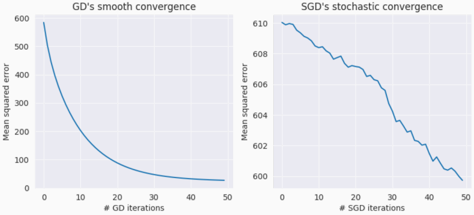
Assume:

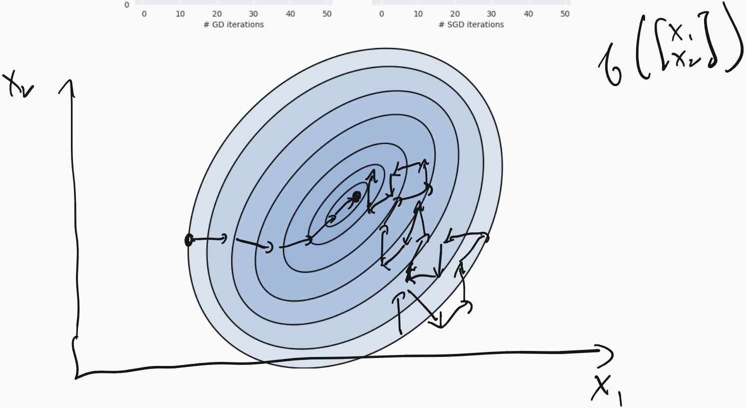
- Lipschitz functions: for all $\mathbf{x}$, j, $\|\nabla f_j(\mathbf{x})\|_2 \leq \boxed{\frac{G'}{n}}$.
- Starting radius: $\boxed{\|\mathbf{x}^* - \mathbf{x}^{(1)}\|_2 \leq R.}$

Stochastic Gradient descent:

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

VISUALIZING SGD





Claim (SGD Convergence)

After $\boxed{T = \frac{R^2 G'^2}{\epsilon^2}}$ *iteration:*

random #

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

with prob. 9/10

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

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

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

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

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

$$= \frac{n}{T}\, \mathbb{E} \underbrace{\sum_{i=1}^{T} h(x^{(i)}) - h(x^*)}_{\leq R\frac{G'}{n}\sqrt{T}}$$

g_i = the gradient at step i $= \nabla f_{j_i}(x^{(i)})$

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

$h_i(y) = g_i^T y$

- convex

$\nabla h_i(y) = g_i$

Claim (SGD Convergence)

After $T = \frac{R^2 G'^2}{\epsilon^2}$ iteration:

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

$$\mathbb{E}[f(\hat{x}) - f(x^*)] \leq \frac{\eta}{T} R \cdot \frac{G'}{\eta} \cdot \sqrt{T}$$

$$= \frac{R \cdot G'}{\sqrt{T}} \leq \varepsilon$$

$$T = \frac{R^2 G'^2}{\varepsilon^2}$$

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

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

$\|x+y\|_2 \leq \|x\|_2 + \|y\|_2$
$= 2\|x\|_2$

Always have $G \leq G'$:

$$\|\nabla f(x)\|_2 \leq \|\nabla f_1(x)\|_2 + \ldots + \|\nabla f_n(x)\|_2 \leq n \cdot \frac{G'}{n} = G'.$$

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:

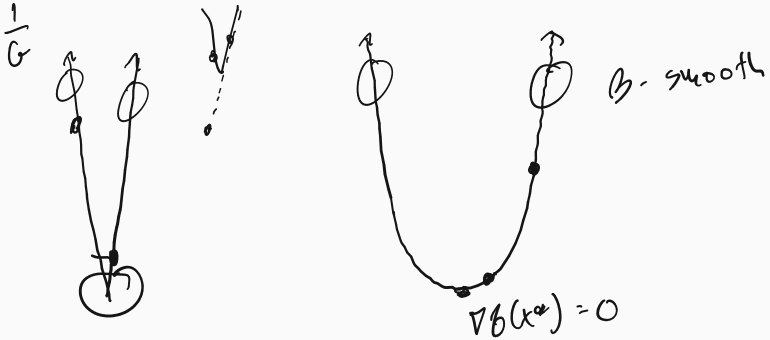
- Improve ϵ dependence below $1/\epsilon^2$.
- Reduce or eliminate dependence on G and R.
- Etc.

Definition (β-smoothness)

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

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

β is a parameter that will depend on our function.

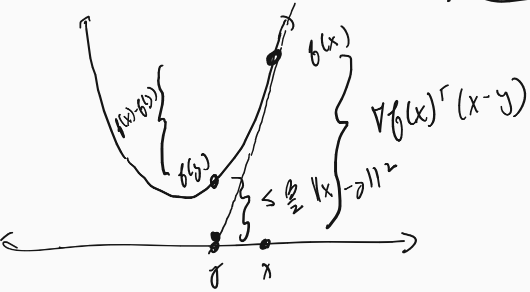


Recall from definition of convexity that:

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

How much smaller can left hand side be?

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



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

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

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

Progress per step of gradient descent:

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

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

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

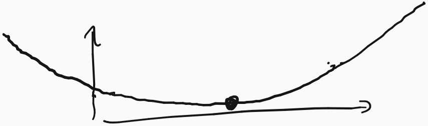
Theorem (GD convergence for β-smooth functions.)

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

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

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

$$T = O\left(\frac{G^2 R^2}{\varepsilon^2}\right)$$



Definition (~~β-smoothness~~ α - strong convexity)

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

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

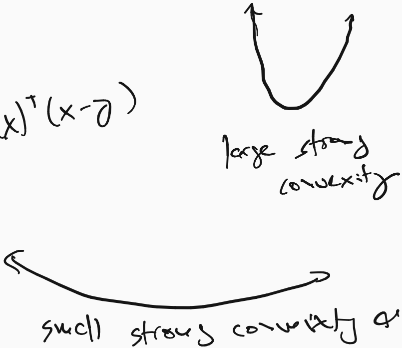
α is a parameter that will depend on our function.

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

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

Completing the picture: If f is α strongly convex and β smooth,

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



Gradient descent for strongly convex functions:

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

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:

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

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

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

What if f is both β-smooth and α-strongly convex?

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

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

What if $\alpha = \beta$:

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

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

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

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

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

What if f is both β-smooth and α-strongly convex?

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

What if $\alpha = \beta$:

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

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

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

$\nabla h(y) = 0$ for maximizing y

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

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:

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

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

Is it better if κ is large or small?

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

Converting to more familiar form:

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

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

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

0

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

$$\leq \frac{\beta}{2}\cdot\left(e^{-} \| \quad \|\right)$$

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:

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

$$\|x^{(1)} - x^*\|_2^2 = R$$

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

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

(apologies for typos on this page

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

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

Let $f(\mathbf{x}) = \|\mathbf{D}\mathbf{x} - \mathbf{b}\|_2^2$ where $\mathbf{D}$ is a diagaonl matrix. For now imagine we're in two dimensions: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$.

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

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

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

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

In other words: What is smallest β so that for all $\mathbf{x}, \mathbf{y}$,

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

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

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

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

smallest we can choose β?

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

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

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

In other words: What is largest α so that for all $\mathbf{x}, \mathbf{y}$,

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

$$2D(Dx-b)^T(x-y) - [x^TD^2x - 2x^TDb - b^Tb$$

$$- y^TD^2y + 2y^TDb + b^Tb]$$

$$\frac{\alpha}{2}\|x-y\|_2^2 \leq 2x^TD^2x - 2x^TDb - 2x^TD^2y + 2y^TDb - [\ \ldots\ ]$$

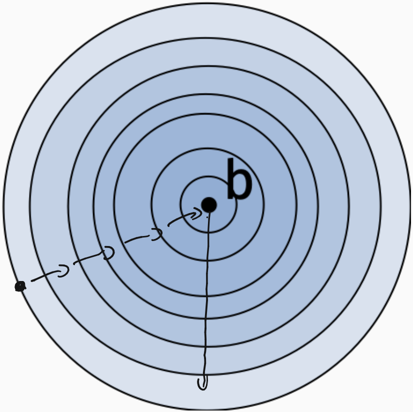
$$= x^TD^2x - 2x^TD^2y + y^TD^2y$$

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

what's the largest α so that

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

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

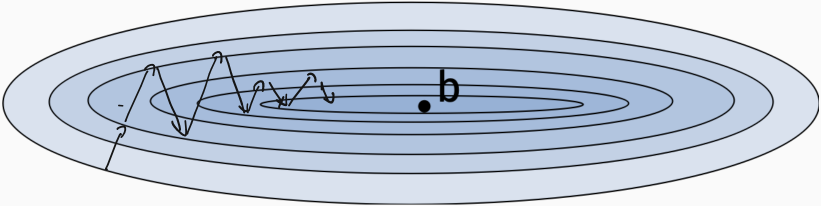


Level sets of $\|\mathbf{Dx} - \mathbf{b}\|_2^2$ when $d_1 = 1, d_2 = 1$.

$\|x - b\|_2^2$

$\kappa = 1$

$\max(d_1, d_2)$
$= \min(d_1, d_2)$



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

$\beta = 2$

$\alpha = 1/3$

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

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

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

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

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:

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

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

IN-CLASS EXERCISE

IN-CLASS EXERCISE