COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2019.

Lecture 21

Last Class:

- Stochastic gradient descent (SGD).
- Online optimization and online gradient descent (OGD).
- Analysis of SGD as a special case of online gradient descent.

Last Class:

- Stochastic gradient descent (SGD).
- Online optimization and online gradient descent (OGD).
- Analysis of SGD as a special case of online gradient descent.

This Class:

- Finish discussion of SGD.
- Understanding gradient descent and SGD as applied to least squares regression.
- Connections to more advanced techniques: accelerated methods and adaptive gradient methods.

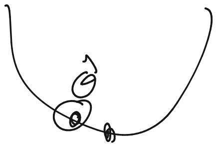
This class wraps up the optimization unit.

Three remaining classes after break. Give your feedback on Piazza about what you'd like to see.

- High ~~dimensional geometry~~ and connections to random projection.
- Randomized methods for fast approximate SVD, eigendecomposition, regression.
- Fourier methods, compressed sensing, sparse recovery.
- More advanced optimization methods (alternating minimization, k-means clustering,...)
- Fairness and differential privacy.

Gradient Descent:

- **Applies to:** Any differentiable $f : \mathbb{R}^d \rightarrow \mathbb{R}$.
- **Goal:** Find $\hat{\theta} \in \mathbb{R}^d$ with $f(\hat{\theta}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} f(\vec{\theta}) + \epsilon$.
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)})$.



QUICK REVIEW

Gradient Descent:

- **Applies to:** Any differentiable $f : \mathbb{R}^d \to \mathbb{R}$.
- **Goal:** Find $\hat{\theta} \in \mathbb{R}^d$ with $f(\hat{\theta}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} f(\vec{\theta}) + \epsilon$.
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)})$.

Online Gradient Descent:

- **Applies to:** $f_1, f_2, \ldots, f_t : \mathbb{R}^d \to \mathbb{R}$ presented online.
- **Goal:** Pick $\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(t)} \in \mathbb{R}^d$ in an ~~online fashion~~ with $\sum_{i=1}^{t} f_i(\vec{\theta}^{(i)}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} \sum_{i=1}^{t} f(\vec{\theta}) + \epsilon$ (i.e., achieve regret $\leq \epsilon$).
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f_i(\vec{\theta}^{(i)})$.

Gradient Descent:

- **Applies to:** Any differentiable $f : \mathbb{R}^d \to \mathbb{R}$.
- **Goal:** Find $\hat{\theta} \in \mathbb{R}^d$ with $f(\hat{\theta}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} f(\vec{\theta}) + \epsilon$.
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)})$.

Online Gradient Descent:

- **Applies to:** $f_1, f_2, \ldots, f_t : \mathbb{R}^d \to \mathbb{R}$ presented online.
- **Goal:** Pick $\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(t)} \in \mathbb{R}^d$ in an online fashion with $\sum_{i=1}^t f_i(\vec{\theta}^{(i)}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} \sum_{i=1}^t f(\vec{\theta}) + \epsilon$ (i.e., achieve regret $\leq \epsilon$).
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f_i(\vec{\theta}^{(i)})$.

Stochastic Gradient Descent:

- **Applies to:** $f : \mathbb{R}^d \to \mathbb{R}$ that can be written as $f(\vec{\theta}) = \sum_{i=1}^n f_i(\vec{\theta})$.
- **Goal:** Find $\hat{\theta} \in \mathbb{R}^d$ with $f(\hat{\theta}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} f(\vec{\theta}) + \epsilon$.
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f_{j_i}(\vec{\theta}^{(i)})$ where j_i is chosen uniformly at random from $1, \ldots, n$.

QUICK REVIEW

Gradient Descent:

- **Applies to:** Any differentiable $f : \mathbb{R}^d \to \mathbb{R}$.
- **Goal:** Find $\hat{\theta} \in \mathbb{R}^d$ with $f(\hat{\theta}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} f(\vec{\theta}) + \epsilon$.
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)})$.

Online Gradient Descent:

- **Applies to:** $f_1, f_2, \ldots, f_t : \mathbb{R}^d \to \mathbb{R}$ presented online.
- **Goal:** Pick $\vec{\theta}^{(1)}, \ldots, \vec{\theta}^{(t)} \in \mathbb{R}^d$ in an online fashion with $\sum_{i=1}^t f_i(\vec{\theta}^{(i)}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} \sum_{i=1}^t f(\vec{\theta}) + \epsilon$ (i.e., achieve regret $\leq \epsilon$).
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f_i(\vec{\theta}^{(i)})$.

total prediction error

Stochastic Gradient Descent:

- **Applies to:** $f : \mathbb{R}^d \to \mathbb{R}$ that can be written as $f(\vec{\theta}) = \sum_{i=1}^n f_i(\vec{\theta})$.

  prediction error on data point i.
- **Goal:** Find $\hat{\theta} \in \mathbb{R}^d$ with $f(\hat{\theta}) \leq \min_{\vec{\theta} \in \mathbb{R}^d} f(\vec{\theta}) + \epsilon$.
- **Update Step:** $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f_{j_i}(\vec{\theta}^{(i)})$ where j_i is chosen uniformly at random from $1, \ldots, n$.

Minimizing a finite sum function: $f(\vec{\theta}) = \sum_{i=1}^{n} f_i(\vec{\theta})$.

Minimizing a finite sum function: $f(\vec{\theta}) = \sum_{i=1}^{n} f_i(\vec{\theta})$.

$\forall f_{j_i}$

- Stochastic gradient descent is identical to online gradient descent run on the sequence of t functions $f_{j_1}, f_{j_2}, \ldots, f_{j_t}$.
- These functions are picked uniformly at random, so in expectation, $\mathbb{E}\left[\sum_{i=1}^{t} f_{j_i}(\vec{\theta}^{(i)})\right] = \mathbb{E}\left[\sum_{i=1}^{t} f(\vec{\theta}^{(i)})\right]$.

Minimizing a finite sum function: $f(\vec{\theta}) = \sum_{i=1}^n f_i(\vec{\theta})$.

- Stochastic gradient descent is identical to online gradient descent run on the sequence of t functions $f_{j_1}, f_{j_2}, \ldots, f_{j_t}$.
- These functions are picked uniformly at random, so in expectation, $\mathbb{E}\left[\sum_{i=1}^t f_{j_i}(\vec{\theta}^{(i)})\right] = \mathbb{E}\left[\sum_{i=1}^t f(\vec{\theta}^{(i)})\right]$.

  $\theta^1 \quad \theta^2 \ldots \quad \theta^t$
- By convexity $\hat{\theta} = \frac{1}{t}\sum_{i=1}^t \vec{\theta}^{(i)}$ gives only a better solution. I.e.,

$\mathbb{E}\; t \cdot f(\hat{\theta})$

$\propto f(\hat{\theta})$

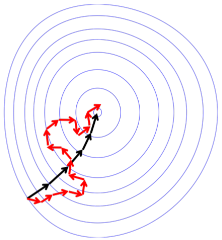
$$\mathbb{E}\left[\sum_{i=1}^t f(\hat{\theta})\right] \leq \mathbb{E}\left[\sum_{i=1}^t f(\vec{\theta}^{(i)})\right].$$

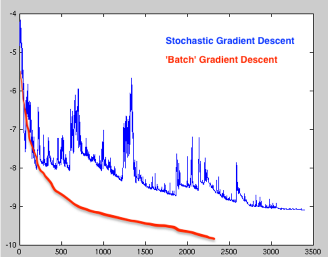
- Quality directly bounded by the regret analysis for online gradient descent!

Stochastic gradient descent generally makes more iterations than gradient descent.

Each iteration is much cheaper (by a factor of n).

$$\vec{\nabla} f(\vec{\theta}) = \vec{\nabla} \sum_{j=1}^{n} f_j(\vec{\theta}) \text{ vs. } \vec{\nabla} f_j(\vec{\theta})$$





Consider $f(\vec{\theta}) = \sum_{j=1}^{n} f_j(\vec{\theta})$ with each f_j convex.

Theorem – SGD: If $\|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq \frac{G}{n}\ \forall \vec{\theta}$, after $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying: $\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon$.

Theorem – GD: If $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq \bar{G}\ \forall \vec{\theta}$, after $t \geq \frac{R^2 \bar{G}^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying: $f(\hat{\theta}) \leq f(\theta^*) + \epsilon$.

Consider $f(\vec{\theta}) = \sum_{j=1}^{n} f_j(\vec{\theta})$ with each f_j convex.

Theorem – SGD: If $\|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq \frac{G}{n}\ \forall \vec{\theta}$, after $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying: $\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon$.

Theorem – GD: If $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq \bar{G}\ \forall \vec{\theta}$, after $t \geq \frac{R^2 \bar{G}^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying: $f(\hat{\theta}) \leq f(\theta^*) + \epsilon$.

$\frac{R^2 G^2}{\epsilon^2}$

$$\|\vec{\nabla} f(\vec{\theta})\|_2 = \|\vec{\nabla} f_1(\vec{\theta}) + \ldots + \vec{\nabla} f_n(\vec{\theta})\|_2 \leq \sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq n \cdot \frac{G}{n} \leq G.$$

SGD VS. GD

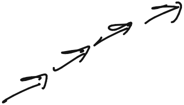
Consider $f(\vec{\theta}) = \sum_{j=1}^{n} f_j(\vec{\theta})$ with each f_j convex.

Theorem – SGD: If $\|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq \frac{G}{n}\ \forall \vec{\theta}$, after $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying: $\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon$.

Theorem – GD: If $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq \bar{G}\ \forall \vec{\theta}$, after $t \geq \frac{R^2 \bar{G}^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying: $f(\hat{\theta}) \leq f(\theta^*) + \epsilon$. $\bar{G} \leq G$

$$\|\vec{\nabla} f(\vec{\theta})\|_2 = \|\vec{\nabla} f_1(\vec{\theta}) + \ldots + \vec{\nabla} f_n(\vec{\theta})\|_2 \leq \sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq n \cdot \frac{G}{n} \leq G.$$

When would this bound be tight? I.e., SGD takes the same number of iterations as GD.



Consider $f(\vec{\theta}) = \sum_{j=1}^{n} f_j(\vec{\theta})$ with each f_j convex.

Theorem – SGD: If $\|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq \frac{G}{n} \, \forall \vec{\theta}$, after $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying: $\mathbb{E}[f(\hat{\theta})] \leq f(\theta^*) + \epsilon$.

Theorem – GD: If $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq \bar{G} \, \forall \vec{\theta}$, after $t \geq \frac{R^2 \bar{G}^2}{\epsilon^2}$ iterations outputs $\hat{\theta}$ satisfying: $f(\hat{\theta}) \leq f(\theta^*) + \epsilon$.

$$\|\vec{\nabla} f(\vec{\theta})\|_2 = \|\vec{\nabla} f_1(\vec{\theta}) + \ldots + \vec{\nabla} f_n(\vec{\theta})\|_2 \leq \sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta})\|_2 \leq n \cdot \frac{G}{n} \leq G.$$

When would this bound be tight? I.e., SGD takes the same number of iterations as GD.

When is it loose? I.e., SGD performs very poorly compared to GD.

$f(\theta) = c\theta - c\theta = 0$

$f_1(\theta) + f_2(\theta)$

$f_1(\theta) = c\theta \quad f_1'(\theta) = c \quad f'(\theta) = 0$

$f_2(\theta) = -c\theta \quad f_2'(\theta) = -c$

Roughly: SGD performs well compared to GD when $\sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta})\|_2$ is small compared to $\|\vec{\nabla} f(\vec{\theta})\|_2$.

Roughly: SGD performs well compared to GD when $\sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta})\|_2^{2}$ is small compared to $\|\vec{\nabla} f(\vec{\theta})\|_2^{2}$.

$$\sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta})\|_2^2 - \|\vec{\nabla} f(\vec{\theta})\|_2^2 = \sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta}) - \vec{\nabla} f(\vec{\theta})\|_2^2 \text{ (good exercise)}$$

SGD VS. GD

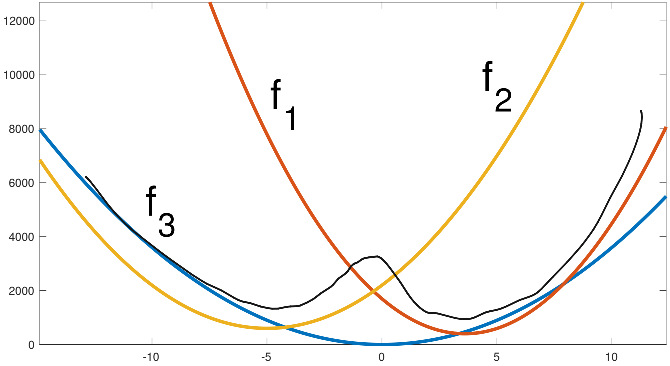
Roughly: SGD performs well compared to GD when $\sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta})\|_2$ is small compared to $\|\vec{\nabla} f(\vec{\theta})\|_2$.

$$\sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta})\|_2^2 - \|\vec{\nabla} f(\vec{\theta})\|_2^2 = \sum_{j=1}^{n} \|\vec{\nabla} f_j(\vec{\theta}) - \frac{1}{n}\vec{\nabla} f(\vec{\theta})\|_2^2 \text{ (good exercise)}$$

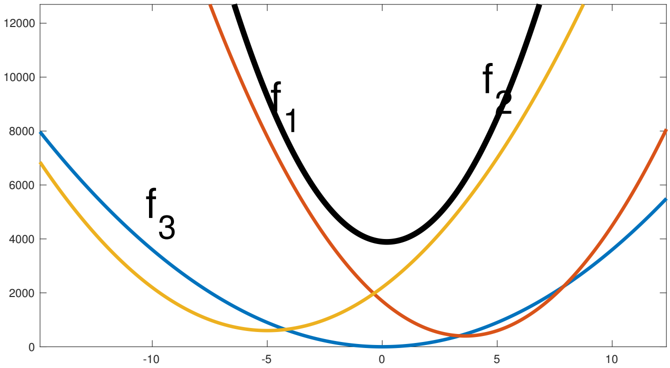
Reducing this variance is a key technique used to increase performance of SGD.

- mini-batching
- stochastic variance reduced gradient descent (SVRG)
- stochastic average gradient (SAG)

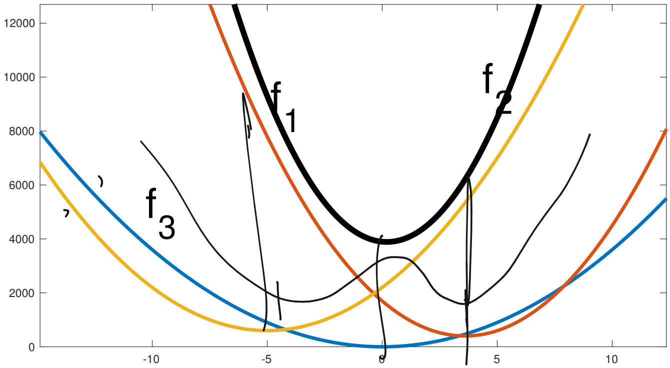
What does $f_1(\theta) + f_2(\theta) + f_3(\theta)$ look like?



What does $f_1(\theta) + f_2(\theta) + f_3(\theta)$ look like?



What does $f_1(\theta) + f_2(\theta) + f_3(\theta)$ look like?



A sum of convex functions is always convex (good exercise).

Linear Algebra + Convex Optimization

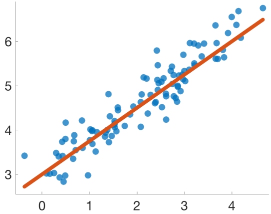
Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$(\langle x_i, \theta \rangle - y_i)^2$

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$



Least Squares Regression: Given data matrix $X \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$X = U \Sigma V^T$

$$f(\vec{\theta}) = \|X\vec{\theta} - \vec{y}\|_2^2.$$

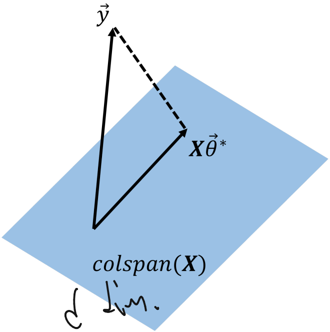
Optimum given by $\vec{\theta}^* = V\mathbf{\Sigma}^{-1}U^T y$. Have $X\vec{\theta}^* = \underbrace{U \Sigma V^T}_{X} V \Sigma^{-1} U^T y$

$= UU^T y$

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$

Optimum given by $\vec{\theta}^* = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^T\mathbf{y}$. Have $\mathbf{X}\vec{\theta}^* =$

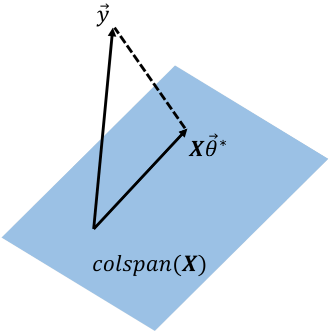


Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$

Optimum given by $\vec{\theta}^* = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{y}$. Have $\mathbf{X}\vec{\theta}^* =$

SVD $O(nd^2)$



Why solve with an iterative method (e.g., gradient descent)?

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$

Claim 1: $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \mathbf{X}\vec{\theta}^*\|_2^2 + c = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2 + c.$

$$\|X\vec{\theta}\|_2^2 + \|X\theta^*\|_2^2 - 2\theta^T X^T X \theta^*$$

$$\|X\theta\|_2^2 + \|X\theta^*\|_2^2 - \underbrace{2\theta^T X^T U \Sigma \cancel{V^T V} \Sigma^{-1} U^T y}$$

$$2\theta^T X^T U U^T y$$

$$2\theta^T V \Sigma U^T \cancel{U} U^T y$$

$$2\theta^T X^T y$$

$$\|X\theta\|_2^2 + \|X\theta^*\|_2^2 - 2\theta^T X^T y = \underbrace{\|X\theta\|_2^2 + \|y\|_2^2 - 2\theta^T X^T y}_{\|X\theta - y\|_2^2} + \overbrace{(\|X\theta^*\|_2^2 - \|y\|}^{c}$$

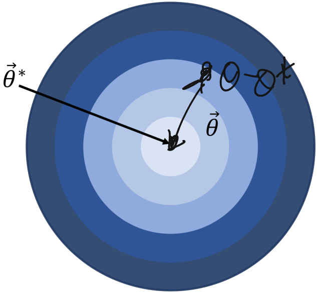
Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$

Claim 1: $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \mathbf{X}\vec{\theta}^*\|_2^2 + c = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2 + c.$

$X = I$

$\underline{\|\vec{\theta} - \vec{\theta}^*\|_2^2}$



Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

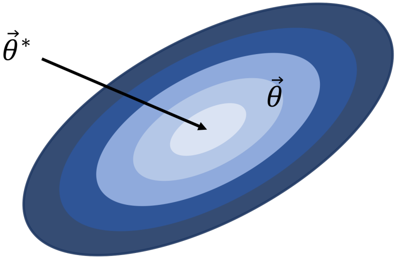
$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$

Claim 1: $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \mathbf{X}\vec{\theta}^*\|_2^2 + c = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2 + c.$

$\|\theta - \theta^*\|_2^2$

$\|X\theta - X\theta^*\|_2^2$

$\|X(\theta - \theta^*)\|_2^2$



Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$

Claim 1: $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \mathbf{X}\vec{\theta}^*\|_2^2 + c = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2 + c.$

Claim 2: $\vec{\nabla} f(\theta) = 2\mathbf{X}^T\mathbf{X}\vec{\theta} - 2\mathbf{X}^T\vec{y} = 2\mathbf{X}^T \underbrace{(\mathbf{X}\vec{\theta} - \vec{y})}_{\text{residual}}$

$\|X\theta - y\|_2^2 = \underbrace{\theta^T X^T X \theta}_{\|X\theta\|_2^2} - 2\theta^T X^T y + \cancel{y^T y}^{\,0}$

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2.$$

Claim 1: $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \mathbf{X}\vec{\theta}^*\|_2^2 + c = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2 + c.$

$\theta^{i+1} = \theta^i - \eta \vec{\nabla} f(\theta^i)$

Claim 2: $\vec{\nabla} f(\theta) = 2\mathbf{X}^T\mathbf{X}\vec{\theta} - 2\mathbf{X}^T\vec{y} = 2\mathbf{X}^T \underbrace{(\mathbf{X}\vec{\theta} - \vec{y})}_{\text{residual}}$

Gradient Descent Update:

jth entry

$$\boxed{\vec{\theta}^{(i+1)}} = \vec{\theta}^{(i)} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(i)} - \vec{y}) \sim \mathbb{R}^d$$

$\langle x_j, \theta^{(i)} + 2\eta x_j \rangle \leq y_j$

$$= \vec{\theta}^{(i)} - 2\eta \sum_{j=1}^{n} \vec{x}_j \cdot r_{i,j}. \quad -1 \quad +2\eta \cdot x_j$$

where $r_{i,j} = (\vec{x}_j^T \vec{\theta}^{(i)} - y_j)$ is the residual for data point j at step i.

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2$$

SGD FOR REGRESSION

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \sum_{j=1}^{n} \overbrace{\left(\vec{x}_j^T \vec{\theta} - y_j\right)^2}^{f_j(\vec{\theta})}$$

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \sum_{j=1}^{n} \left(\vec{x}_j^T\vec{\theta} - y_j\right)^2 = \sum_{j=1}^{n} f_j(\vec{\theta}).$$

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \sum_{j=1}^{n} \left(\vec{x}_j^T\vec{\theta} - y_j\right)^2 = \sum_{j=1}^{n} f_j(\vec{\theta}).$$

Claim 3: $\vec{\nabla} f_j(\theta) = \underbrace{2(\vec{x}_j^T\vec{\theta} - \vec{y}_j)}_{\text{residual}} \cdot \vec{x}_j$

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \sum_{j=1}^{n} \left(\vec{x}_j^T \vec{\theta} - y_j\right)^2 = \sum_{j=1}^{n} f_j(\vec{\theta}).$$

Claim 3: $\vec{\nabla} f_j(\theta) = \underbrace{2(\vec{x}_j^T \vec{\theta} - \vec{y}_j)}_{\text{residual}} \cdot \vec{x}_j$

SGD Update: Pick random $j \in \{1, \ldots, n\}$ and set:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f_j(\theta^{(i)}) = \vec{\theta}^{(i)} - 2\eta \vec{x}_j \cdot r_{i,j}$$

where $r_{i,j} = (\vec{x}_j^T \vec{\theta}^{(i)} - y_j)$ is the residual for data point j at step i.

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$f(x,z) = x^2 + z^2$

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \sum_{j=1}^{n} \left(\vec{x}_j^T\vec{\theta} - y_j\right)^2 = \sum_{j=1}^{n} f_j(\vec{\theta}).$$

$f_z(x) = f(x,c) = x^2 + c^2$

Claim 3: $\vec{\nabla} f_j(\theta) = \underbrace{2(\vec{x}_j^T\vec{\theta} - \vec{y}_j)}_{\text{residual}} \cdot \vec{x}_j$

SGD Update: Pick random $j \in \{1, \ldots, n\}$ and set:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta\vec{\nabla} f_j(\theta^{(i)}) = \vec{\theta}^{(i)} - 2\eta\vec{x}_j \cdot r_{i,j} \text{ verses } -2\eta\sum_{j=1}^{n}\vec{x}_j r_{i,j}$$

where $r_{i,j} = (\vec{x}_j^T\vec{\theta}^{(i)} - y_j)$ is the residual for data point j at step i.

Least Squares Regression: Given data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\vec{y} \in \mathbb{R}^n$:

$$f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \sum_{j=1}^{n} \left(\vec{x}_j^T \vec{\theta} - y_j\right)^2 = \sum_{j=1}^{n} f_j(\vec{\theta}).$$

Claim 3: $\vec{\nabla} f_j(\theta) = \underbrace{2(\vec{x}_j^T \vec{\theta} - \vec{y}_j)}_{\text{residual}} \cdot \vec{x}_j$

SGD Update: Pick random $j \in \{1, \ldots, n\}$ and set:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f_j(\theta^{(i)}) = \vec{\theta}^{(i)} - 2\eta \vec{x}_j \cdot r_{i,j} \text{ verses } -2\eta \sum_{j=1}^{n} \vec{x}_j r_{i,j}$$

where $r_{i,j} = (\vec{x}_j^T \vec{\theta}^{(i)} - y_j)$ is the residual for data point j at step i.

Make a small correction for a single data point in each step. In the direction of the data point.

Gradient Descent for Regression:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta\vec{\nabla}f(\vec{\theta}^{(i)}) = \vec{\theta}^{(i)} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(i)} - \vec{y}).$$

Initialize $\vec{\theta}^{(1)} = \vec{0}$.

Gradient Descent for Regression:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)}) = \vec{\theta}^{(i)} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(i)} - \vec{y}).$$

Initialize $\vec{\theta}^{(1)} = \vec{0}$.

$$\vec{\theta}^{(2)} = \vec{0} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{0} - \vec{y}) = 2\eta \mathbf{X}^T \vec{y}.$$

Gradient Descent for Regression:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)}) = \vec{\theta}^{(i)} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(i)} - \vec{y}).$$

Initialize $\vec{\theta}^{(1)} = \vec{0}$.

$$\vec{\theta}^{(2)} = \vec{0} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{0} - \vec{y}) = 2\eta \mathbf{X}^T\vec{y}.$$

$$\vec{\theta}^{(3)} = 2\eta \mathbf{X}^T\vec{y} - 2\eta \mathbf{X}^T(2\eta \mathbf{X}\mathbf{X}^T\vec{y} - \vec{y}) = 4\eta \mathbf{X}^T\vec{y} - 4\eta^2(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} = 4\eta(\mathbf{I} - \eta \mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y}$$

Gradient Descent for Regression:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)}) = \vec{\theta}^{(i)} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(i)} - \vec{y}).$$

Initialize $\vec{\theta}^{(1)} = \vec{0}$.

$$\vec{\theta}^{(2)} = \vec{0} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{0} - \vec{y}) = \boxed{2\eta \mathbf{X}^T\vec{y}.}$$

$$\vec{\theta}^{(3)} = 2\eta \mathbf{X}^T\vec{y} - 2\eta \mathbf{X}^T(\underbrace{2\eta \mathbf{X}\mathbf{X}^T\vec{y}}_{X\theta^{(2)}} - \vec{y}) = 4\eta \mathbf{X}^T\vec{y} - 4\eta^2(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} = 4\eta(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y}$$

$$\vec{\theta}^{(4)} = \theta^{(3)} - \eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(3)} - \vec{y}) = 6\eta \mathbf{X}^T\vec{y} - 16\eta(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} + 8\eta^2(\mathbf{X}^T\mathbf{X})^2\mathbf{X}^T\vec{y}.$$

$$\left[6\eta I - 16\eta(X^TX) + 8\eta^2(X^TX)^2\right]X^Ty$$

Gradient Descent for Regression:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)}) = \vec{\theta}^{(i)} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(i)} - \vec{y}).$$

Initialize $\vec{\theta}^{(1)} = \vec{0}$.

$$\vec{\theta}^{(2)} = \vec{0} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{0} - \vec{y}) = 2\eta \mathbf{X}^T \vec{y}.$$

$$\vec{\theta}^{(3)} = 2\eta \mathbf{X}^T\vec{y} - 2\eta \mathbf{X}^T(2\eta \mathbf{X}\mathbf{X}^T\vec{y} - \vec{y}) = 4\eta \mathbf{X}^T\vec{y} - 4\eta^2(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} = 4\eta(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y}$$

$$\vec{\theta}^{(4)} = \theta^{(3)} - \eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(3)} - \vec{y}) = 6\eta \mathbf{X}^T\vec{y} - 16\eta(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} + 8\eta^2(\mathbf{X}^T\mathbf{X})^2\mathbf{X}^T\vec{y}.$$

d x d   n d   lin

$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \qquad .$$

where p_t is a degree $t-2$ polynomial.

GRADIENT DESCENT AS POLYNOMIAL APPROXIMATION

Gradient Descent for Regression:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta\vec{\nabla}f(\vec{\theta}^{(i)}) = \vec{\theta}^{(i)} - 2\eta\mathbf{X}^T(\mathbf{X}\vec{\theta}^{(i)} - \vec{y}).$$

Initialize $\vec{\theta}^{(1)} = \vec{0}$.

$$\vec{\theta}^{(2)} = \vec{0} - 2\eta\mathbf{X}^T(\mathbf{X}\vec{0} - \vec{y}) = 2\eta\mathbf{X}^T\vec{y}.$$

$$\vec{\theta}^{(3)} = 2\eta\mathbf{X}^T\vec{y} - 2\eta\mathbf{X}^T(2\eta\mathbf{X}\mathbf{X}^T\vec{y} - \vec{y}) = 4\eta\mathbf{X}^T\vec{y} - 4\eta^2(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} = 4\eta(\mathbf{I} - \eta\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y}$$

$$\vec{\theta}^{(4)} = \theta^{(3)} - \eta\mathbf{X}^T(\mathbf{X}\vec{\theta}^{(3)} - \vec{y}) = 6\eta\mathbf{X}^T\vec{y} - 16\eta(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} + 8\eta^2(\mathbf{X}^T\mathbf{X})^2\mathbf{X}^T\vec{y}.$$

$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \approx \hat{\theta}^* = V^T\Sigma^{-1}U^Ty = (X^TX)^{-1}X^Ty$$

$$= (V\Sigma^2V^T)^{-1}V\Sigma U^Ty$$

where p_t is a degree $t-2$ polynomial.

$$(V\Sigma^2V^T)^{-1} = V\Sigma^{-2}V^T$$

$$= V\Sigma^{-2}V^TV\Sigma U^Ty = V\Sigma^{-1}U^Ty$$

Gradient Descent for Regression:

$$\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} - \eta \vec{\nabla} f(\vec{\theta}^{(i)}) = \vec{\theta}^{(i)} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(i)} - \vec{y}).$$

Initialize $\vec{\theta}^{(1)} = \vec{0}$.

$$\vec{\theta}^{(2)} = \vec{0} - 2\eta \mathbf{X}^T(\mathbf{X}\vec{0} - \vec{y}) = 2\eta \mathbf{X}^T\vec{y}.$$

$$\vec{\theta}^{(3)} = 2\eta \mathbf{X}^T\vec{y} - 2\eta \mathbf{X}^T(2\eta \mathbf{X}\mathbf{X}^T\vec{y} - \vec{y}) = 4\eta \mathbf{X}^T\vec{y} - 4\eta^2(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} = 4\eta(\mathbf{I} - \eta \mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y}$$

$$\vec{\theta}^{(4)} = \theta^{(3)} - \eta \mathbf{X}^T(\mathbf{X}\vec{\theta}^{(3)} - \vec{y}) = 6\eta \mathbf{X}^T\vec{y} - 16\eta(\mathbf{X}^T\mathbf{X})\mathbf{X}^T\vec{y} + 8\eta^2(\mathbf{X}^T\mathbf{X})^2\mathbf{X}^T\vec{y}.$$

$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \approx \theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y}.$$

where p_t is a degree $t-2$ polynomial.

Upshot: Gradient descent computes

$$\vec{\theta}^{(t)} = \underline{p_t(\mathbf{X}^T\mathbf{X})} \cdot \mathbf{X}^T\vec{y} \approx \underline{(\mathbf{X}^T\mathbf{X})^{-1}}\mathbf{X}^T\vec{y} = \theta^*.$$

Upshot: Gradient descent computes

$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \approx (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} = \theta^*.$$

One of the most basic Krylov subspace methods. Chebyshev iteration, Jacobi iteration, conjugate gradient, accelerated gradient descent, heavy ball methods....

Upshot: Gradient descent computes

$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \approx (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} = \theta^*.$$

One of the most basic Krylov subspace methods. Chebyshev iteration, Jacobi iteration, conjugate gradient, accelerated gradient descent, heavy ball methods....

Upshot: Gradient descent computes

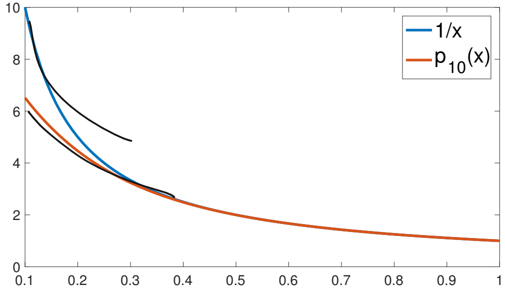
$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \approx (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} = \theta^*.$$

View in Eigendecomposition:

One of the most basic Krylov subspace methods. Chebyshev iteration, Jacobi iteration, conjugate gradient, accelerated gradient descent, heavy ball methods....

Upshot: Gradient descent computes

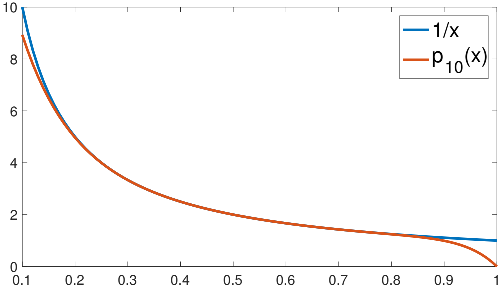
$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \approx (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} = \theta^*.$$



One of the most basic Krylov subspace methods. Chebyshev iteration, Jacobi iteration, conjugate gradient, accelerated gradient descent, heavy ball methods....

Upshot: Gradient descent computes

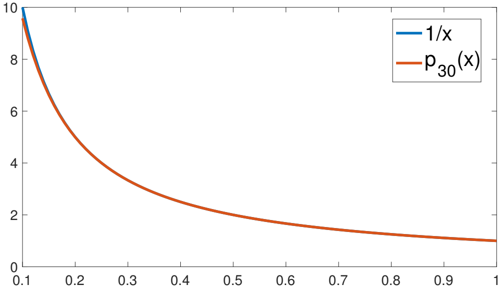
$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \approx (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} = \theta^*.$$



One of the most basic Krylov subspace methods. Chebyshev iteration, Jacobi iteration, conjugate gradient, accelerated gradient descent, heavy ball methods....

Upshot: Gradient descent computes

$$\vec{\theta}^{(t)} = p_t(\mathbf{X}^T\mathbf{X}) \cdot \mathbf{X}^T\vec{y} \approx (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} = \theta^*.$$



One of the most basic Krylov subspace methods. Chebyshev iteration, Jacobi iteration, conjugate gradient, accelerated gradient descent, heavy ball methods....

Gradient descent for least squares regression requires a lot of iterations when the eigenvalues of $X^T X$ are spread out. Formally:

Gradient descent for least squares regression requires a lot of iterations when the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are spread out. Formally:

- Is $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$ Lipschitz?

Gradient descent for least squares regression requires a lot of iterations when the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are spread out. Formally:

- Is $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$ Lipschitz?
- A convex function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth and α-strongly convex if $\forall \vec{\theta}_1, \vec{\theta}_2$:

$$\frac{\alpha}{2}\|\vec{\theta}_1 - \vec{\theta}_2\|_2^2 \leq \vec{\nabla} f(\vec{\theta}_1)^T(\vec{\theta}_1 - \vec{\theta}_2) - [f(\vec{\theta}_1) - f(\vec{\theta}_2)] \leq \frac{\beta}{2}\|\vec{\theta}_1 - \vec{\theta}_2\|_2^2.$$

Gradient descent for least squares regression requires a lot of iterations when the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are spread out. Formally:

- Is $f(\vec{\theta}) = \|\mathbf{X}\vec{\theta} - \vec{y}\|_2^2 = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$ Lipschitz?
- A convex function $f: \mathbb{R}^d \to \mathbb{R}$ is β-smooth and α-strongly convex if $\forall \vec{\theta}_1, \vec{\theta}_2$:

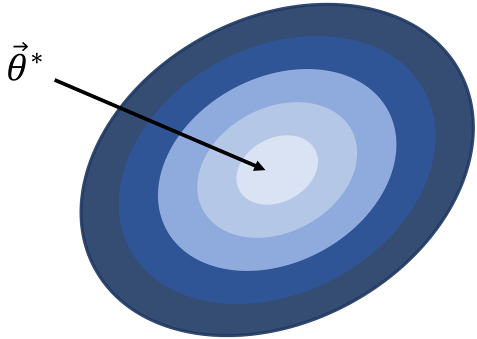
$$\frac{\alpha}{2}\|\vec{\theta}_1 - \vec{\theta}_2\|_2^2 \leq \vec{\nabla} f(\vec{\theta}_1)^T(\vec{\theta}_1 - \vec{\theta}_2) - [f(\vec{\theta}_1) - f(\vec{\theta}_2)] \leq \frac{\beta}{2}\|\vec{\theta}_1 - \vec{\theta}_2\|_2^2.$$

- $f(\theta)$ is $\beta = \lambda_{max}(\mathbf{X}^T\mathbf{X})$ smooth and $\alpha = \lambda_{min}(\mathbf{X}^T\mathbf{X})$ strongly convex.

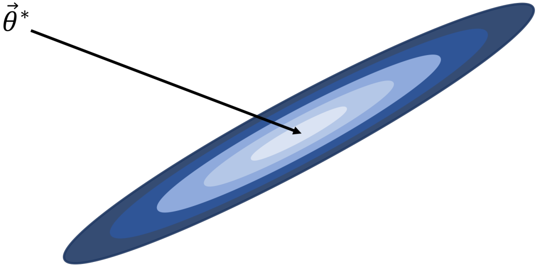
Theorem: For any α-strongly convex and β-smooth function $f(\vec{\theta})$, *GD* initialized with $\vec{\theta}^{(1)}$ within a radius R of $\vec{\theta}^*$ and run for $t = O\left(\frac{\beta}{\alpha} \cdot \log(1/\epsilon)\right)$ iterations returns $\hat{\theta}$ with $\|\hat{\theta} - \theta^*\|_2 \leq \epsilon R$.

For least squares regression, $\alpha = \lambda_{min}(\mathbf{X}^T\mathbf{X})$, $\beta = \lambda_{max}(\mathbf{X}^T\mathbf{X})$, and $\frac{\beta}{\alpha}$ is called the condition number κ.

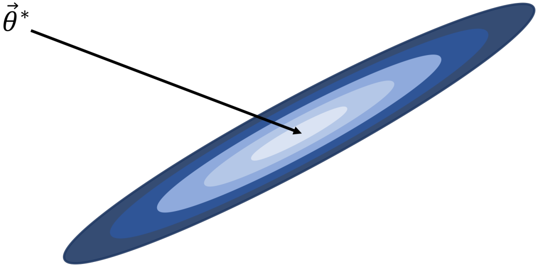
Recall: $f(\vec{\theta}) = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$.



Recall: $f(\vec{\theta}) = \|\mathsf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$.

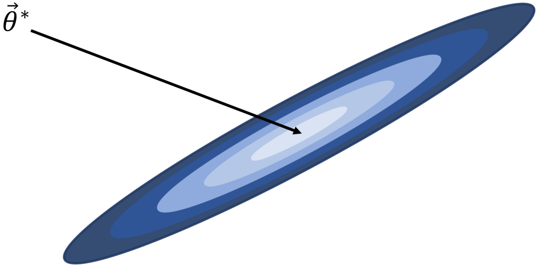


Recall: $f(\vec{\theta}) = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$.



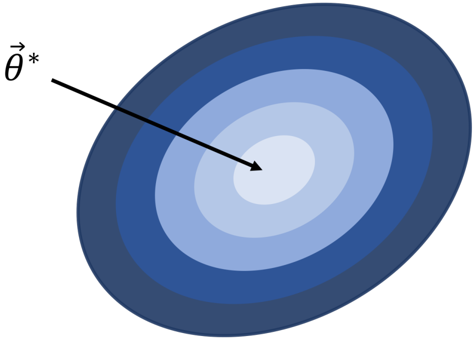
How can we mitigate this issue?

Recall: $f(\vec{\theta}) = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$.



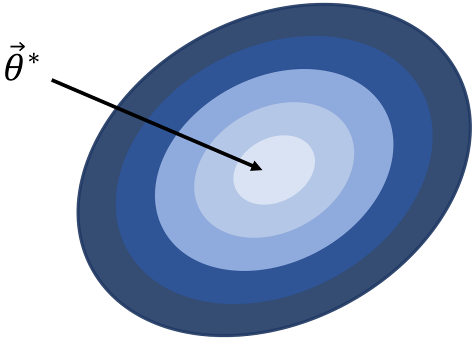
How can we mitigate this issue? Scale the directions to make the surface more 'round'.

Recall: $f(\vec{\theta}) = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$.



How can we mitigate this issue? Scale the directions to make the surface more 'round'.

Recall: $f(\vec{\theta}) = \|\mathbf{X}(\vec{\theta} - \vec{\theta}^*)\|_2^2$.



How can we mitigate this issue? Scale the directions to make the surface more 'round'.

Idea of adaptive gradient methods: AdaGrad, RMSprop, Adam. And quasi-Newton methods: BFGS, L-BFGS,...

MATHEMATICAL VIEW OF PRECONDITIONING – IF TIME