10-315
Introduction to ML

Optimization and Linear Regression

Instructor: Pat Virtue

Pre-reading: Optimization Formulation

Find the best $h(x) \to \hat{y}$ by searching in the space of hypothesis functions $h \in \mathcal{H}$.

Loss: $\ell(y, \hat{y})$ is the loss or cost of predicting $\hat{y}$ when the true value is y.

Risk: Risk for a hypothesis function is the expected loss over the data distribution:

$$R(h) = \mathbb{E}_{XY}\left[\ell\big(Y, h(X)\big)\right]$$

Minimizing Risk

Specifically: find a prediction function $h^*: \mathcal{X} \to \mathcal{Y}$ that minimizes the expected loss for randomly drawn test data (X, Y)

$$h^* = \operatorname*{argmin}_h \mathbb{E}_{XY}\left[\ell\big(Y, h(X)\big)\right]$$

Pre-reading: Empirical Risk Minimization

Risk

$$R(h) = \mathbb{E}_{X,Y}[\ell(Y, h(X))]$$

Empirical Risk

$$\hat{R}(h) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(y^{(i)}, h\left(x^{(i)}\right)\right)$$

$$\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$$

Empirical Risk Minimization

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \ \hat{R}(h)$$

Pre-reading: Loss Functions

$$h^* = \underset{h}{\operatorname{argmin}}\, \mathbb{E}_{XY}\big[L\big(Y, h(X)\big)\big]$$

Loss function:

$L: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Classification:

- Two-class, 0,1 loss

- Two-class, arbitrary loss

False positives and false negatives:

Poll (unused)

For a healthcare classification problem where:

$Y=0$ represents a negative test for cancer (healthy)

$Y=1$ represents a positive test for cancer (sick)

Which two-class loss value should be highest?

A. True negative
B. False negative
C. False positive
D. True positive

Poll (unused)

Does our decision tree algorithm (recursively finding the best attribute to split on a each node) find a hypothesis that minimizes empirical risk for zero-one loss?

$$\mathcal{D} = \left\{ \left(x^{(i)}, y^{(i)} \right) \right\}_{i=1}^{N}$$

$$\hat{R}(h) = \frac{1}{N} \sum_{i=1}^{N} \ell \left(y^{(i)}, h\left(x^{(i)} \right) \right)$$

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \ \hat{R}(h)$$

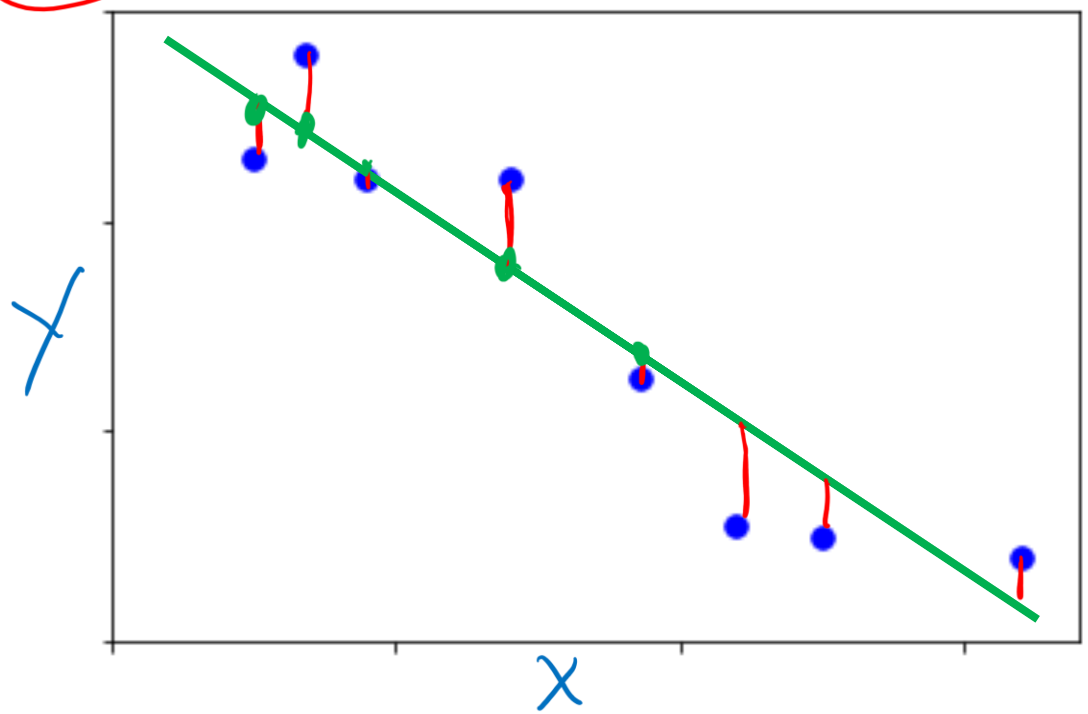
A. Yes, always
B. Yes, but only when splitting with error rate
C. Yes, but only when splitting with mutual information
D. No

Poll 1

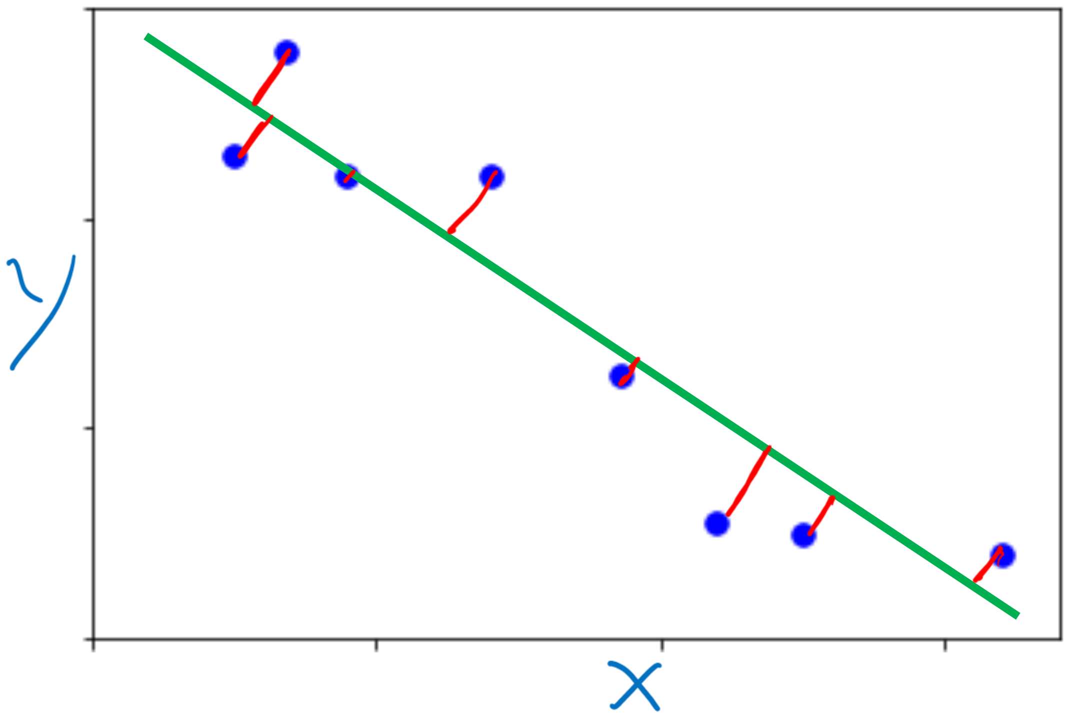
Which is a correct depiction of linear regression error lines for $x \in \mathbb{R}$, $\hat{y} = wx + b$

$y^{(i)} - \hat{y}^{(i)}$

A



B



Poll 2

Vocab: b : bias term (offset)

$\vec{w}$: weights

→ parameters

Given $\theta = \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix}$, what goes in the top entry for a linear regression

1 ← 1 to account for the bias term

vector with two features, $\mathbf{x} = \begin{bmatrix} ? \\ x_1 \\ x_2 \end{bmatrix}$?

$$\hat{y} = \theta^T \vec{x}$$

$$b \cdot 1 + w_1 x_1 + w_2 x_2$$

Poll 3: Two questions

Fill in the blanks: Linear regession optimization

$$\hat{\theta} = \underset{\text{1) ?}}{\text{argmin}} \quad \underset{\text{2) ? } \theta}{} \quad \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2$$

1) What type of optimization?

A. min
B. max
C. argmin
D. argmax

2) What are we optimizing over?

A. x
B. y
C. θ
D. x, θ
E. x, y
F. x, y, θ

Pre-reading: Risk in Regression

Squared error loss $\ell(y, h(x)) = (y - h(x))^2$ sq. err.

$$h^* = \operatorname*{argmin}_h \mathbb{E}_{XY}[\ell(Y, h(X))]$$

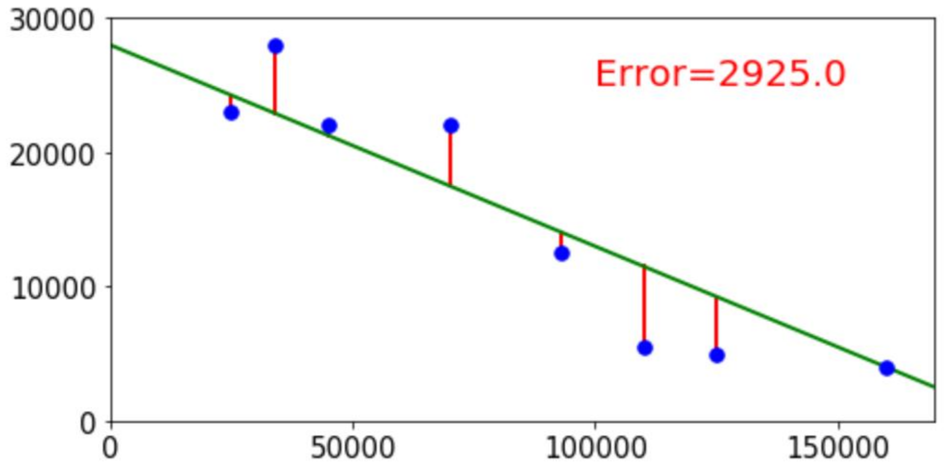
$$\begin{aligned}\hat{R}(h) &= \frac{1}{N}\sum_{i=1}^{N} \ell\left(y^{(i)}, h\left(x^{(i)}\right)\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \left(y^{(i)} - h\left(x^{(i)}\right)\right)^2 \quad \longleftarrow \text{Mean squared error!} \\ &= \frac{1}{N}\sum_{i=1}^{N} \left(y^{(i)} - \underbrace{\left(wx^{(i)} + b\right)}_{\text{linear}}\right)^2\end{aligned}$$

model for h(x)

Contructing the Design Matrix

Linear Algebra Formulation

Constructing the design matrix



Goal: Replace summations with linear algebra

$$J(w,b) = \frac{1}{N}\sum_{i}^{N}\left(y^{(i)} - \left(b + \sum_{j}^{M} w_j x_j^{(i)}\right)\right)^2$$

$$= \frac{1}{N}\sum_{i}^{N}\left(y^{(i)} - \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)}\right)\right)^2$$ ← From pre-reading

$$= \frac{1}{N}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2$$

Reminder (Lec 2): Dataset notation

Dataset notation

$$\mathcal{D} = \{(y^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^{N}$$

$$= \left\{\left(y^{(i)}, x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, x_4^{(i)}\right)\right\}_{i=1}^{N}$$

Linear algebra can represent all data

$\boxed{\mathbf{y}} \in \{0,1,2\}^N$

$\boxed{X} \in \mathbb{R}^{N\times 4}$ (design matrix)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

Linear Algebra Formulation

$$\hat{y}^{(i)} = b + \sum_{j}^{M} w_j x_j^{(i)}$$

Constructing the design matrix

y

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
2	5.9	3.0	5.1	1.8

$\vec{y}$

Linear Algebra Formulation

$$\hat{y}^{(i)} = b + \sum_{j}^{M} w_j x_j^{(i)}$$

Constructing the design matrix

Species
0
0
0
1
1
1
2

$\vec{y}$

Sepal Length	Sepal Width	Petal Length	Petal Width
4.3	3.0	1.1	0.1
4.9	3.6	1.4	0.1
5.3	3.7	1.5	0.2
4.9	2.4	3.3	1.0
5.7	2.8	4.1	1.3
6.3	3.3	4.7	1.6
5.9	3.0	5.1	1.8

Linear Algebra Formulation

Constructing the design matrix

$$\hat{y}^{(i)} = b + \sum_{j}^{M} w_j x_j^{(i)} \quad \text{orig}$$

$$\hat{y}^{(i)} = \underline{\boldsymbol{\theta}}^{\top} \underline{\mathbf{x}^{(i)}}$$

$$\theta = \begin{bmatrix} b \\ \vec{w} \end{bmatrix} \quad \vec{x}^{(i)} = \begin{bmatrix} 1 \\ \vec{x}^{(i)}_{orig} \end{bmatrix}$$

Species
0
0
0
1
1
1
2

$\vec{y}$

(account for bias)	Sepal Length	Sepal Width	Petal Length	Petal Width
1	4.3	3.0	1.1	0.1
1	4.9	3.6	1.4	0.1
1	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
1	5.9	3.0	5.1	1.8

$X_{orig} \in \mathbb{R}^{N \times M_{orig}}$

$X \in \mathbb{R}^{N \times M}$, where $M = M_{orig} + 1$

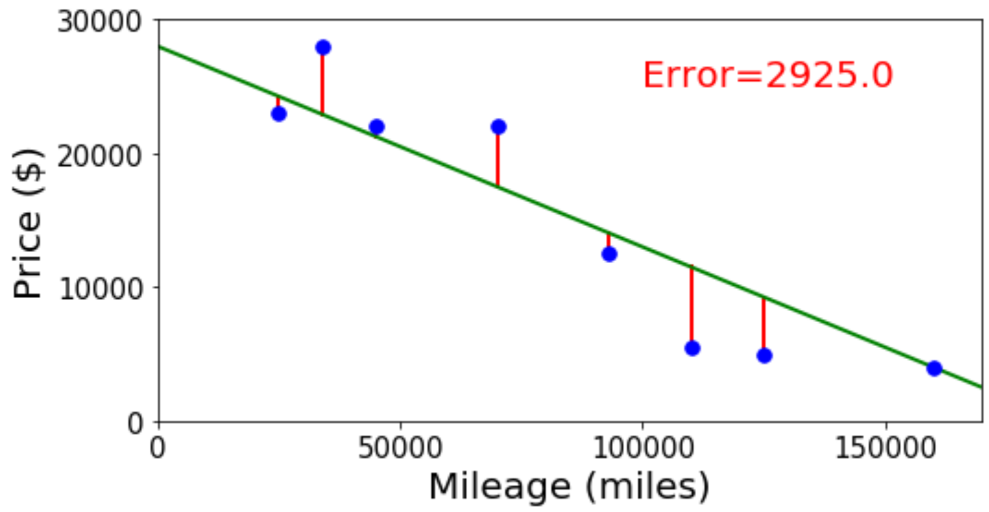
Linear Regression

System of Equations

Poll 4

$[N \times 1] = [N \times M]\,[M \times 1]$

Now that we have $\mathbf{y} = X\boldsymbol{\theta}$, we can solve this system of equations to get the best $\boldsymbol{\theta}$?



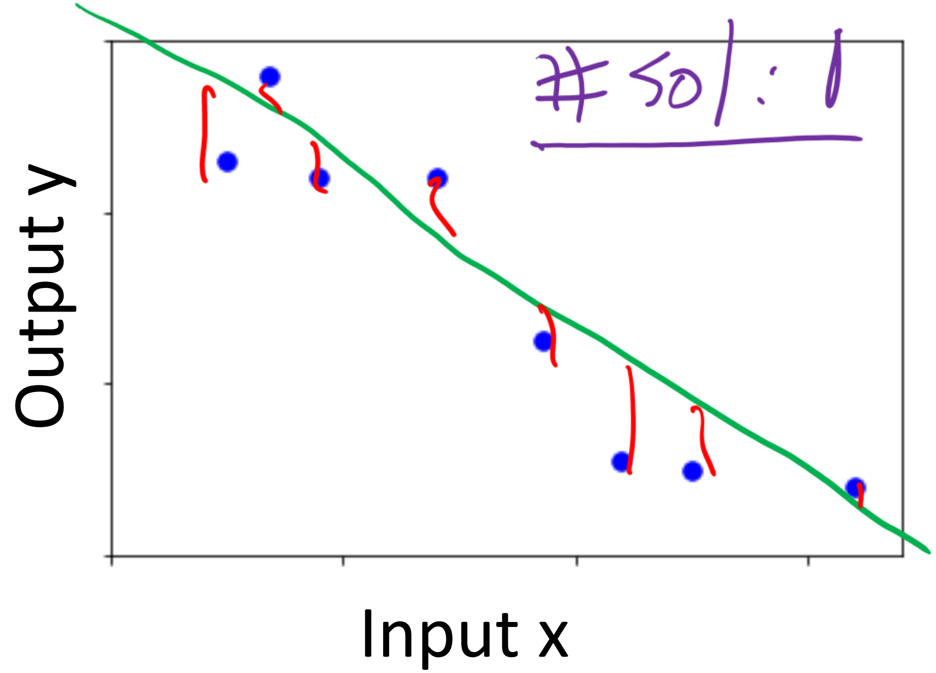
A. Yes
B. No
C. I have no idea

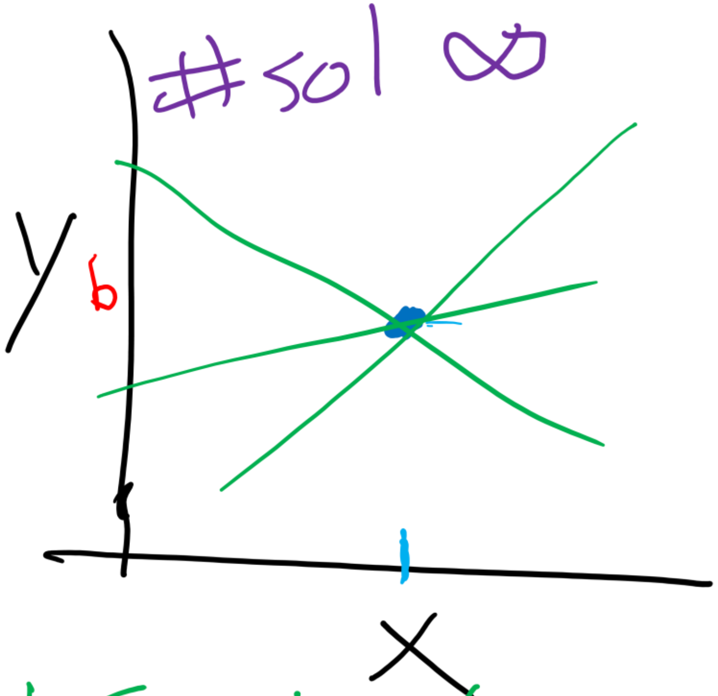
Linear Regression

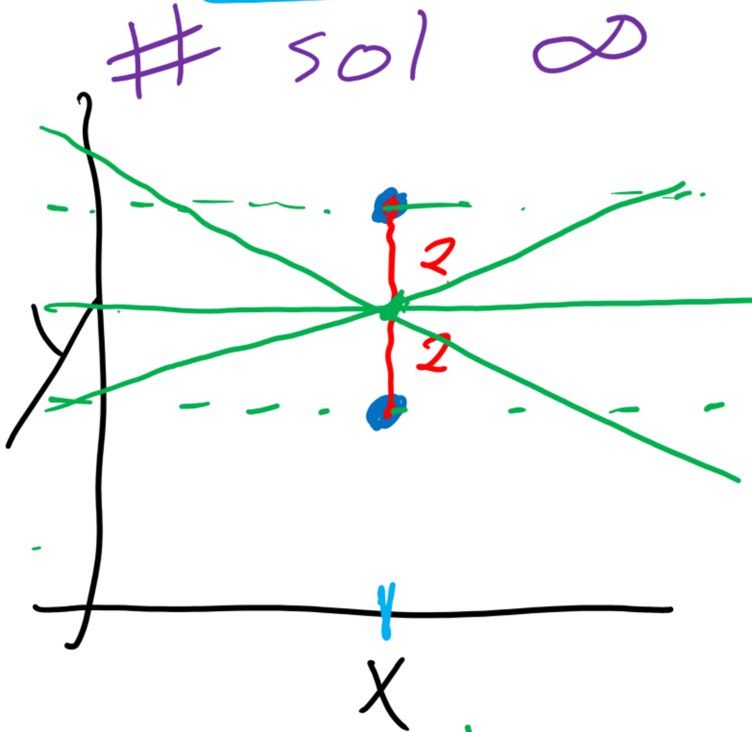
$\vec{y} = X\vec{\theta}$

$\hat{y} = h(x)$
5

Number of solutions? Objective function value, $J(\theta)$?







LS # sol : 0

$J(\theta) > 0$

LS # sol : ∞

$J(\theta) = 0$

LS # sol : 0

$J(\theta) > 0$

Linear Regression

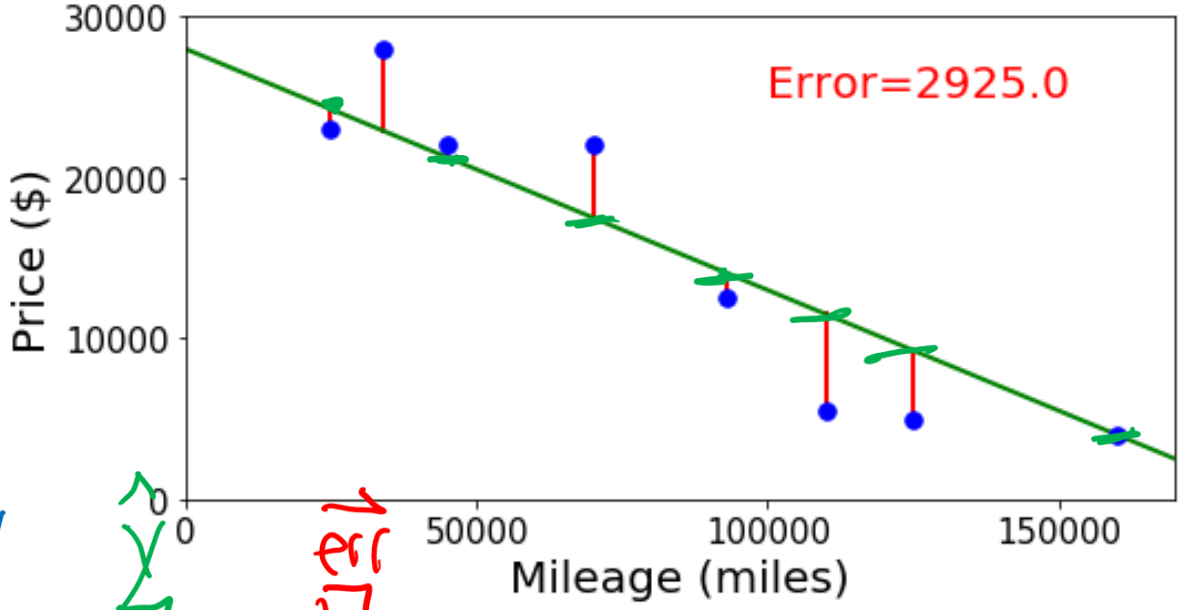
Linear algebra formulation

$$\vec{\hat{y}} = X\vec{\theta}$$

$$\underset{\theta}{\operatorname{argmin}}\ J(\theta)$$

$$= " \frac{1}{N} \|\vec{y} - \vec{\hat{y}}\|_2^2$$

$$= " \frac{1}{N} \|\vec{y} - X\vec{\theta}\|_2^2$$



$$\vec{y} \; [\;] - \vec{\hat{y}} \; [\;] = \vec{err} \; [\;]$$

Closed-form Solution

Linear Regression

$\rightarrow \|\vec{z}\|_2^2 = \sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} z_i z_i = \vec{z}^T \vec{z}$

Expanding objective before computing gradient

$$J(\boldsymbol{\theta}) = \frac{1}{N}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \quad \checkmark$$

$$= \frac{1}{N}(\vec{y} - X\theta)^T(\vec{y} - X\theta)$$

$$= \frac{1}{N}(\vec{y}^T - \theta^T X^T)(\vec{y} - X\theta)$$

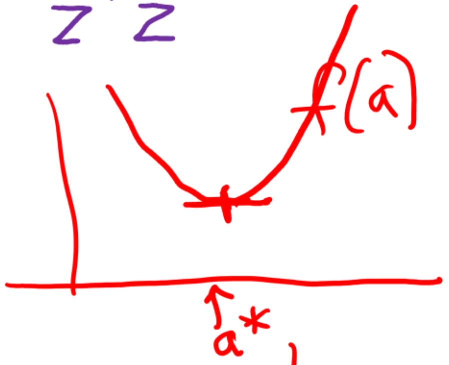
$$= \frac{1}{N}(\vec{y}^T y - \theta^T X^T y - \vec{y}^T X\theta + \theta^T X^T X\theta) \qquad \frac{\partial J}{\partial \vec{\theta}} = 0$$

$[1 \times M][M \times N][N \times 1]$

$[1 \times 1]$

$\rightarrow (\vec{y}^T X\theta)^T = (\theta^T X^T y) \rightarrow$ solve for $\vec{\theta}$

$f(a)$



$\frac{df}{da} = 0 \rightarrow$ solve for a

$$J(\theta) = \frac{1}{N}(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T X^T \mathbf{y} + \boldsymbol{\theta}^T X^T X \boldsymbol{\theta})$$

Linear Regression

$\|\vec{z}\|_2^2 = \sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} z_i z_i = \vec{z}^T\vec{z}$

Expanding objective before computing gradient

$$J(\boldsymbol{\theta}) = \frac{1}{N}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$$

$$= \frac{1}{N}(\mathbf{y} - X\boldsymbol{\theta})^T(\mathbf{y} - X\boldsymbol{\theta})$$

$$= \frac{1}{N}\left(\mathbf{y}^T - \boldsymbol{\theta}^T X^T\right)(\mathbf{y} - X\boldsymbol{\theta})$$

$$= \frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - \boldsymbol{\theta}^T X^T\mathbf{y} - \mathbf{y}^T X\boldsymbol{\theta} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\right)$$

$$= \frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T X^T\mathbf{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\right)$$

Calculus Time-out

Partial derivatives

Calculus with Linear Algebra

Vector in, scalar out

Gradient

$y = f(\mathbf{x}) \qquad y \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^M$

$$\nabla_x f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_M} \end{bmatrix}$$

denom

$$\frac{\partial f}{\partial \vec{x}} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

numer

$$\frac{\partial f}{\partial \vec{x}} = \begin{bmatrix} \partial f / \partial x_1 & \cdots & \partial f / \partial x_M \end{bmatrix}$$

Calculus with Linear Algebra

Functions with linear algebra

$J(\vec{\theta})$

$X\theta$

$$y = \mathrm{f}(x)$$

$$y = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

$$y = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

$f(X)$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}(x)$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

$$\begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,K} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}(x)$$

$$\begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,K} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

$$\begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,K} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

Calculus with Linear Algebra

One way to think of it: bag of partial derivatives

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

$$\frac{\partial y_1}{\partial x_1} \qquad \frac{\partial y_1}{\partial x_2}$$

$$\frac{\partial y_1}{\partial x_M}$$

$$\frac{\partial y_N}{\partial x_M} \qquad \frac{\partial y_2}{\partial x_1} \qquad \frac{\partial y_2}{\partial x_M}$$

Calculus with Linear Algebra

$C = f(A) = AB$

$\frac{\partial C}{\partial A}$

One way to think of it: bag of partial derivatives

$$\begin{bmatrix} Y_{1,1} & \cdots & Y_{1,K} \\ \vdots & \ddots & \vdots \\ Y_{N,1} & \cdots & Y_{N,K} \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

$\frac{\partial Y_{11}}{\partial X_{11}}$ $\frac{\partial Y_{1K}}{\partial X_{M,1}}$

Numerator Layout

$y = f(x)$	$x \in \mathbb{R}$	$\mathbf{x} \in \mathbb{R}^{\mathrm{M}}$	$X \in \mathbb{R}^{\mathrm{M} \times L}$
$y \in \mathbb{R}$	$\frac{\partial y}{\partial x} \in \mathbb{R}$	$\frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^{1 \times M}$ Output by input	$\frac{\partial y}{\partial X} \in \mathbb{R}^{L \times M}$ Transpose of input
$\mathbf{y} \in \mathbb{R}^{N}$	$\frac{\partial \mathbf{y}}{\partial x} \in \mathbb{R}^{N \times 1}$ Output by input	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{N \times M}$ Output by input	$\frac{\partial \mathbf{y}}{\partial X} \in \mathbb{R}^{?}$?
$Y \in \mathbb{R}^{N \times K}$	$\frac{\partial Y}{\partial x} \in \mathbb{R}^{N \times K}$ Size of output	$\frac{\partial Y}{\partial \mathbf{x}} \in \mathbb{R}^{?}$?	$\frac{\partial Y}{\partial X} \in \mathbb{R}^{?}$?

Denominator Layout

$y = f(x)$	$x \in \mathbb{R}$	$\mathbf{x} \in \mathbb{R}^{\mathrm{M}}$	$X \in \mathbb{R}^{\mathrm{M} \times L}$
$y \in \mathbb{R}$	$\frac{\partial y}{\partial x} \in \mathbb{R}$	$\frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^{M \times 1}$ Input by output	$\frac{\partial y}{\partial X} \in \mathbb{R}^{M \times L}$ Size of input
$\mathbf{y} \in \mathbb{R}^N$	$\frac{\partial \mathbf{y}}{\partial x} \in \mathbb{R}^{1 \times N}$ Input by output	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{M \times N}$ Input by output	$\frac{\partial \mathbf{y}}{\partial X} \in \mathbb{R}^{?}$?
$Y \in \mathbb{R}^{N \times K}$	$\frac{\partial Y}{\partial x} \in \mathbb{R}^{K \times N}$ Transpose of output	$\frac{\partial Y}{\partial \mathbf{x}} \in \mathbb{R}^{?}$?	$\frac{\partial Y}{\partial X} \in \mathbb{R}^{?}$?

$\frac{\partial f}{\partial x}$

$\frac{\partial f}{\partial X}$

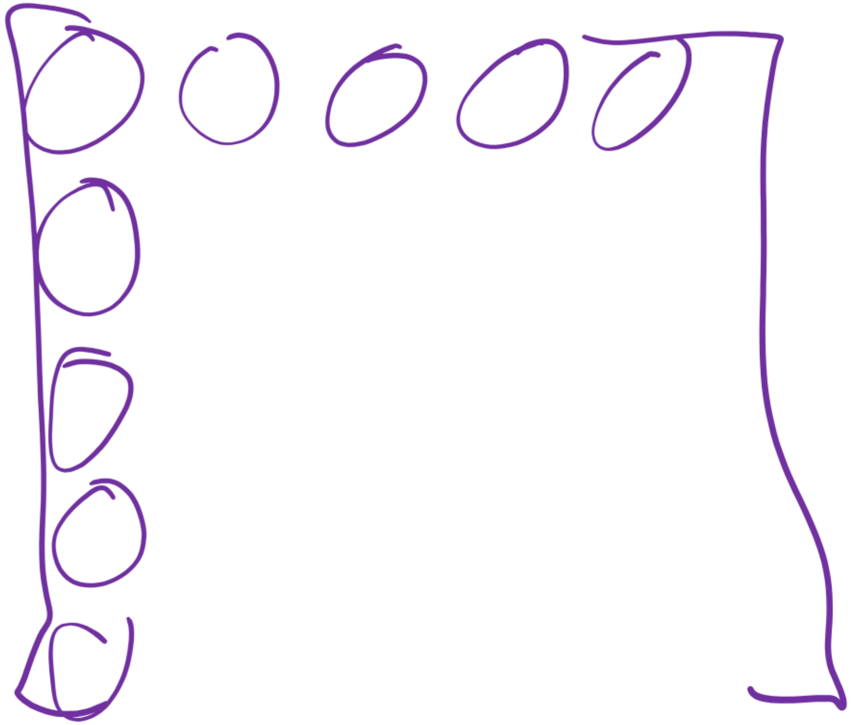
Calculus with Linear Algebra

Jacobian: Vector in, vector out

Numerator-layout

$\boldsymbol{y} = f(\boldsymbol{x}) \quad \boldsymbol{y} \in \mathbb{R}^N, \quad \boldsymbol{x} \in \mathbb{R}^M, \quad \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \in \mathbb{R}^{N \times M}$

$\frac{\partial \vec{y}}{\partial \vec{x}} =$



Calculus with Linear Algebra

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

Numerator-layout vs denominator-layout

Vector in, vector out

$\mathbf{y} = f(\mathbf{x})$ $\quad$ $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{x} \in \mathbb{R}^M$ (assume $N > M$ for illustrative purposes)

Numerator layout	Denominator layout
Number of outputs × number of inputs	Number of inputs × number of outputs
$\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{N \times M}$	$\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{M \times N}$

Calculus with Linear Algebra

$$y = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

Vector in, scalar out

Numerator-layout

$$y = f(\mathbf{x}) \qquad y \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^M, \quad \frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^{1\times M}$$

Calculus with Linear Algebra

Numerator-layout vs denominator-layout

$$y = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

Vector in, scalar out

$y = f(\mathbf{x}) \qquad y \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^M$

Numerator layout	Denominator layout
Number of outputs × number of inputs	Number of inputs × number of outputs
$\frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^{1 \times M}$	$\frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^{M \times 1}$

Calculus with Linear Algebra

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \mathrm{f}(x)$$

Scalar in, vector out

Numerator-layout

$$\mathbf{y} = f(x) \qquad \mathbf{y} \in \mathbb{R}^N, \quad x \in \mathbb{R}, \quad \frac{\partial \mathbf{y}}{\partial x} \in \mathbb{R}^{N\times 1}$$

Calculus with Linear Algebra

$$y = \mathrm{f}\left(\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}\right)$$

Gradient: Vector in, scalar out

Transpose of numerator-layout

$$y = f(\mathbf{x}) \qquad y \in \mathbb{R}, \;\; \mathbf{x} \in \mathbb{R}^M, \;\; \frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^{1\times M}, \;\; \nabla_{\mathbf{x}} f \in \mathbb{R}^{M\times 1}$$

Calculus with Linear Algebra

$$y = \mathrm{f}\left(\begin{bmatrix} X_{1,1} & \cdots & X_{1,L} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \cdots & X_{M,L} \end{bmatrix}\right)$$

Matrix in, scalar out

Numerator layout: Transpose of matrix dimensions

Denominator layout: Keep same dimensions as matrix

$y = f(\boldsymbol{X}) \quad y \in \mathbb{R}, \; \boldsymbol{X} \in \mathbb{R}^{N \times M}, \quad \frac{\partial y}{\partial \boldsymbol{X}} \in \mathbb{R}^{N \times M}$ (denominator layout)

Exercise

Suppose we have a function that takes in a vector and squares each element individually, returning another vector, $\mathbf{y} = f(\mathbf{x})$.

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) \rightarrow \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{bmatrix} \qquad \text{Example: } f\left(\begin{bmatrix} 7 \\ 3 \\ 5 \end{bmatrix}\right) \rightarrow \begin{bmatrix} 49 \\ 9 \\ 25 \end{bmatrix}$$

What is $\partial\mathbf{y}/\partial\mathbf{x}$? (use numerator layout)

No!

$$\cancel{\frac{\partial f}{\partial \vec{x}} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{bmatrix}} \rightarrow \begin{bmatrix} 2x_1 & 0 & 0 \\ 0 & 2x_2 & 0 \\ 0 & 0 & 2x_3 \end{bmatrix}$$

Derivatives of Functions with Respect to Vectors

	Numerator layout	Denominator layout	Notes
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}$	I_N	I_N	$\mathbf{v} \in \mathbb{R}^N$
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T$	I_N	I_N	$\mathbf{v} \in \mathbb{R}^N$
$\frac{\partial}{\partial \mathbf{v}}\ t\mathbf{v}$	tI_N	tI_N	$\mathbf{v} \in \mathbb{R}^N$
$\frac{\partial}{\partial \mathbf{u}}\ \mathbf{u}^T\mathbf{v}$	$\mathbf{v}^T$	$\mathbf{v}$	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{u}^T\mathbf{v}$	$\mathbf{u}^T$	$\mathbf{u}$	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T\mathbf{v}$	$2\mathbf{v}^T$	$2\mathbf{v}$	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T A\mathbf{v}$	$\mathbf{v}^T(A + A^T)$	$(A + A^T)\mathbf{v}$	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T A\mathbf{v}$	$2\mathbf{v}^T A$	$2A\mathbf{v}$	If $A = A^T$
$\frac{\partial}{\partial \mathbf{v}}\ A\mathbf{v}$	A	A^T	
$\frac{\partial}{\partial \mathbf{v}}\ \mathbf{v}^T A$	A^T	A	

Numerator Layout

$y = f(x)$	$x \in \mathbb{R}$	$\mathbf{x} \in \mathbb{R}^{\mathrm{M}}$	$X \in \mathbb{R}^{\mathrm{M} \times L}$
$y \in \mathbb{R}$	$\frac{\partial y}{\partial x} \in \mathbb{R}$	$\frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^{1 \times M}$ Output by input	$\frac{\partial y}{\partial X} \in \mathbb{R}^{L \times M}$ Transpose of input
$\mathbf{y} \in \mathbb{R}^{N}$	$\frac{\partial \mathbf{y}}{\partial x} \in \mathbb{R}^{N \times 1}$ Output by input	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{N \times M}$ Output by input	$\frac{\partial \mathbf{y}}{\partial X} \in \mathbb{R}^{?}$?
$Y \in \mathbb{R}^{N \times K}$	$\frac{\partial Y}{\partial x} \in \mathbb{R}^{N \times K}$ Size of output	$\frac{\partial Y}{\partial \mathbf{x}} \in \mathbb{R}^{?}$?	$\frac{\partial Y}{\partial X} \in \mathbb{R}^{?}$?

Denominator Layout

$y = f(x)$	$x \in \mathbb{R}$	$\mathbf{x} \in \mathbb{R}^{\mathrm{M}}$	$X \in \mathbb{R}^{\mathrm{M} \times L}$
$y \in \mathbb{R}$	$\frac{\partial y}{\partial x} \in \mathbb{R}$	$\frac{\partial y}{\partial \mathbf{x}} \in \mathbb{R}^{M \times 1}$ Input by output	$\frac{\partial y}{\partial X} \in \mathbb{R}^{M \times L}$ Size of input
$\mathbf{y} \in \mathbb{R}^{N}$	$\frac{\partial \mathbf{y}}{\partial x} \in \mathbb{R}^{1 \times N}$ Input by output	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{M \times N}$ Input by output	$\frac{\partial \mathbf{y}}{\partial X} \in \mathbb{R}^{?}$?
$Y \in \mathbb{R}^{N \times K}$	$\frac{\partial Y}{\partial x} \in \mathbb{R}^{K \times N}$ Transpose of output	$\frac{\partial Y}{\partial \mathbf{x}} \in \mathbb{R}^{?}$?	$\frac{\partial Y}{\partial X} \in \mathbb{R}^{?}$?

Linear Regression

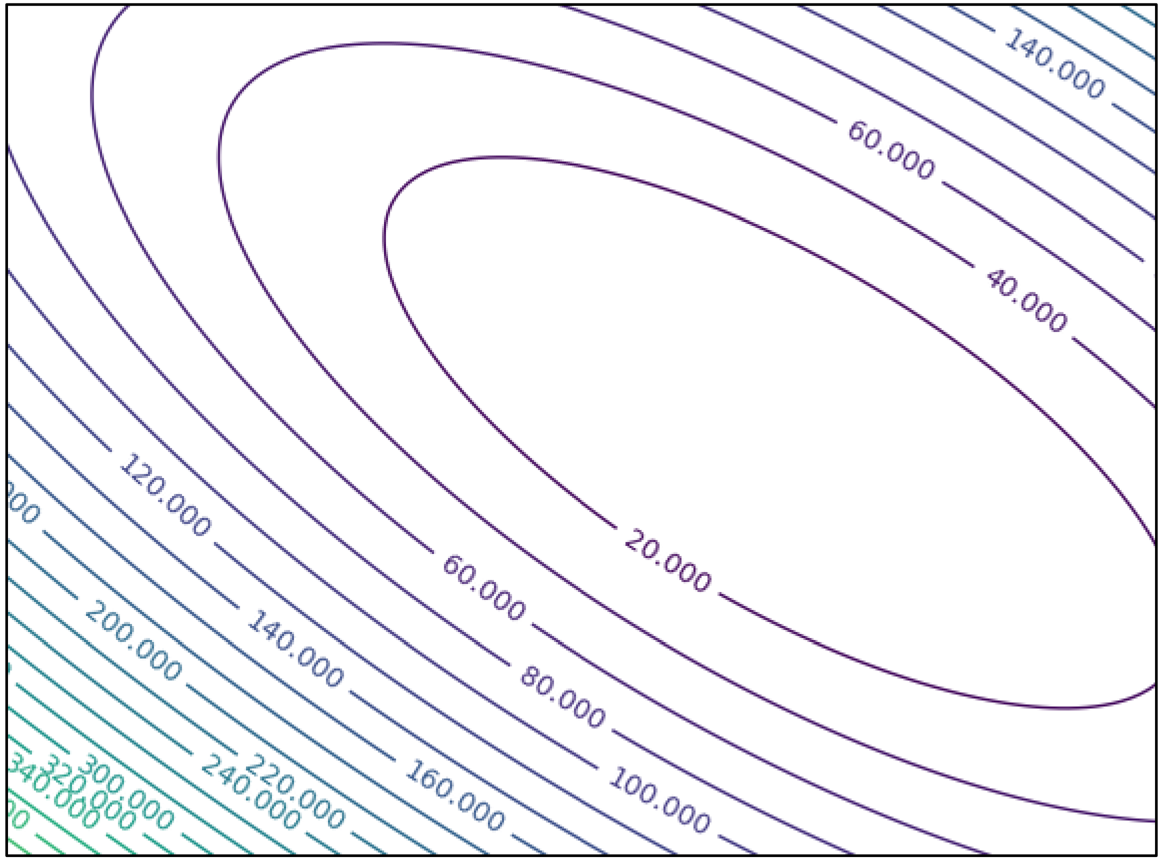
Closed-form Solution

Linear Regression

Methods for optimizing the objective

- Closed-form solution

$J(w, b)$



b

w

Linear Regression

Expanding objective before computing gradient

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= \frac{1}{N}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \\
&= \frac{1}{N}(\mathbf{y} - X\boldsymbol{\theta})^T(\mathbf{y} - X\boldsymbol{\theta}) \\
&= \frac{1}{N}\left(\mathbf{y}^T - \boldsymbol{\theta}^T X^T\right)(\mathbf{y} - X\boldsymbol{\theta}) \\
&= \frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - \boldsymbol{\theta}^T X^T\mathbf{y} - \mathbf{y}^T X\boldsymbol{\theta} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\right) \\
&= \frac{1}{N}\left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T X^T\mathbf{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\right)
\end{aligned}
$$

$$\nabla_{\theta} J(\theta) = \frac{\partial J}{\partial \theta}$$

Linear Regression

$y \in \mathbb{R}^N \quad X \in \mathbb{R}^{N \times M}$

$\theta \in \mathbb{R}^M$

Gradient of objective with respect to parameters

$$J(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2$$

$$= \frac{1}{N}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\theta}^T X^T\boldsymbol{y} + \boldsymbol{\theta}^T X^T X\boldsymbol{\theta}\right)$$

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{N}\left(0 - 2X^T y + 2X^T X\theta\right)$$

$(M \times N) \times (N \times 1) + (M \times M) \times (M \times 1)$

$M \times 1$

$$\frac{2}{N}\left(-X^T y + X^T X\theta\right)$$

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}^T$$

$2(A)z$

$$\frac{\partial \boldsymbol{z}^T A\boldsymbol{z}}{\partial \boldsymbol{z}} = (A + A^T)\boldsymbol{z}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T A\boldsymbol{z}}{\partial \boldsymbol{z}} = \boldsymbol{z}^T(A + A^T)$$

Linear Regression

Gradient of objective with respect to parameters

$$J(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2$$

$$= \frac{1}{N}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\theta}^T X^T \boldsymbol{y} + \boldsymbol{\theta}^T X^T X \boldsymbol{\theta}\right)$$

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{N}\left(0 - 2X^T\boldsymbol{y} + 2\boldsymbol{\theta}^T X^T X\right)$$

$$\frac{\partial \boldsymbol{z}^T \boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T \boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}^T$$

$$\frac{\partial \boldsymbol{z}^T A \boldsymbol{z}}{\partial \boldsymbol{z}} = (A + A^T)\boldsymbol{z}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T A \boldsymbol{z}}{\partial \boldsymbol{z}} = \boldsymbol{z}^T(A + A^T)$$

Linear Regression

Gradient of objective with respect to parameters

$$J(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2$$

$$= \frac{1}{N}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\theta}^T X^T \boldsymbol{y} + \boldsymbol{\theta}^T X^T X \boldsymbol{\theta}\right)$$

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{N}\left(0 - 2X^T\boldsymbol{y} + \cancel{2\boldsymbol{\theta}^T X^T X}\right)$$

$$= \frac{1}{N}\left(0 - 2X^T\boldsymbol{y} + 2X^T X\boldsymbol{\theta}\right)$$

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}^T$$

$$\frac{\partial \boldsymbol{z}^T A \boldsymbol{z}}{\partial \boldsymbol{z}} = (A + A^T)\boldsymbol{z}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T A \boldsymbol{z}}{\partial \boldsymbol{z}} = \boldsymbol{z}^T(A + A^T)$$

Linear Regression

Gradient of objective with respect to parameters

$$J(\boldsymbol{\theta}) = \frac{1}{N}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2$$

$$= \frac{1}{N}\left(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\theta}^T X^T \boldsymbol{y} + \boldsymbol{\theta}^T X^T X \boldsymbol{\theta}\right)$$

$$\nabla J(\boldsymbol{\theta}) = \frac{1}{N}\left(0 - 2X^T\boldsymbol{y} + \cancel{2\boldsymbol{\theta}^T X^T X}\right)$$

$$= \frac{1}{N}\left(0 - 2X^T\boldsymbol{y} + 2X^T X\boldsymbol{\theta}\right)$$

$$= \frac{2}{N}\left(-X^T\boldsymbol{y} + X^T X\boldsymbol{\theta}\right)$$

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T\boldsymbol{u}}{\partial \boldsymbol{z}} = \boldsymbol{u}^T$$

$$\frac{\partial \boldsymbol{z}^T A\boldsymbol{z}}{\partial \boldsymbol{z}} = (A + A^T)\boldsymbol{z}$$

-- or --

$$\frac{\partial \boldsymbol{z}^T A\boldsymbol{z}}{\partial \boldsymbol{z}} = \boldsymbol{z}^T(A + A^T)$$

Linear Regression

$J(\theta)$

Closed-form solution

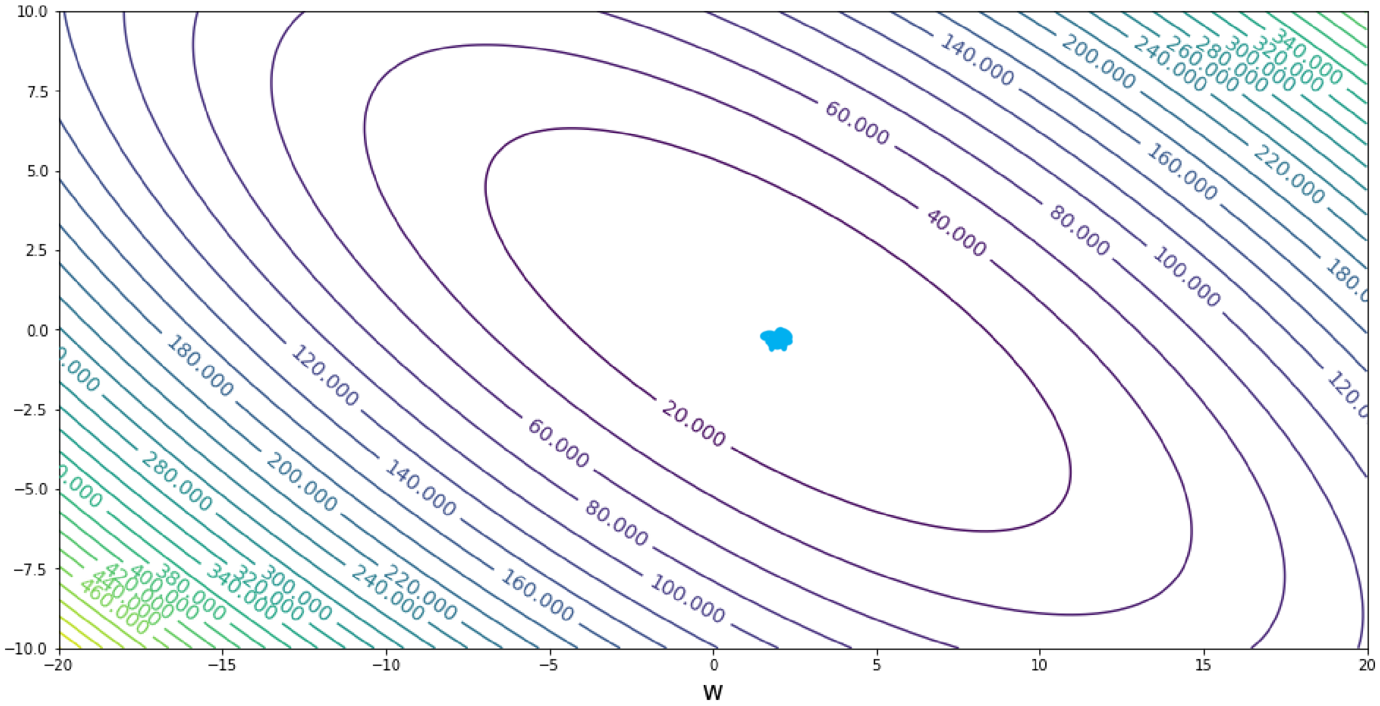
$$\nabla J(\boldsymbol{\theta}) = \frac{2}{N}\left(-X^T\boldsymbol{y} + X^T X\boldsymbol{\theta}\right)$$

$$= 0$$

$$0 = -X^T y + X^T X \theta$$

$$(X^T X)^{-1} X^T X \theta = (X^T X)^{-1} X^T y$$

$$\vec{\theta} = (X^T X)^{-1} X^T \vec{y}$$



b

w

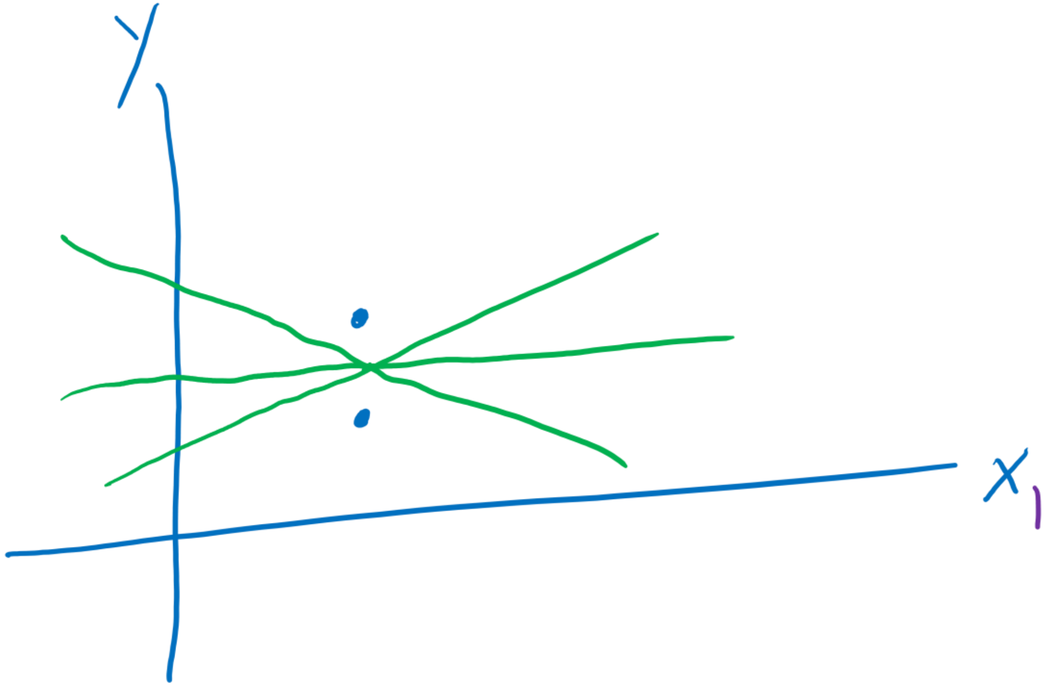
$$\cancel{X^T} X \theta = \cancel{X^T} y$$

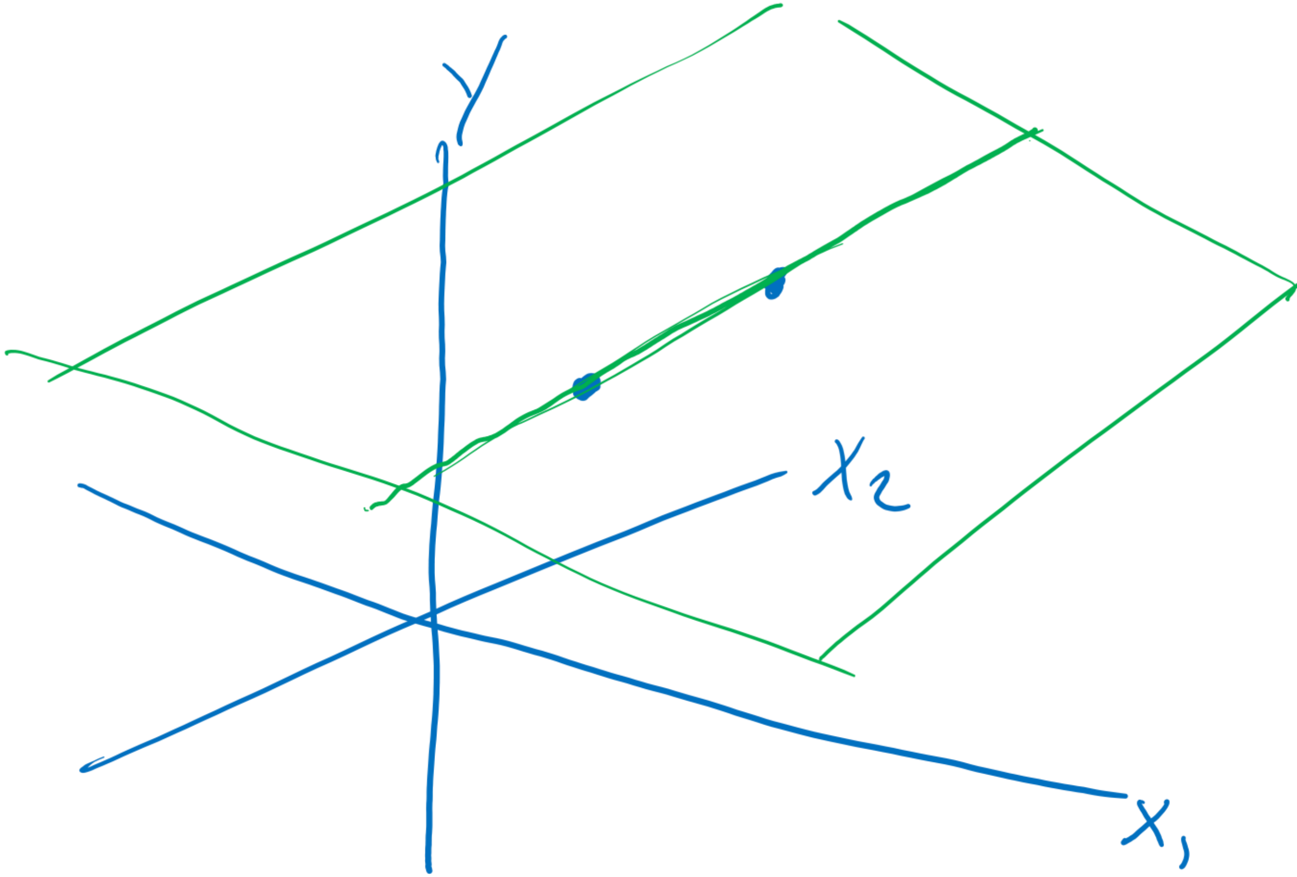
Normal equations

Linear Regression

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

Number of solutions





A note on matrix rank:

$$\theta = \underline{(X^T X)}^{-1} X^T y$$

Not solving this directly

$$y = X\theta$$

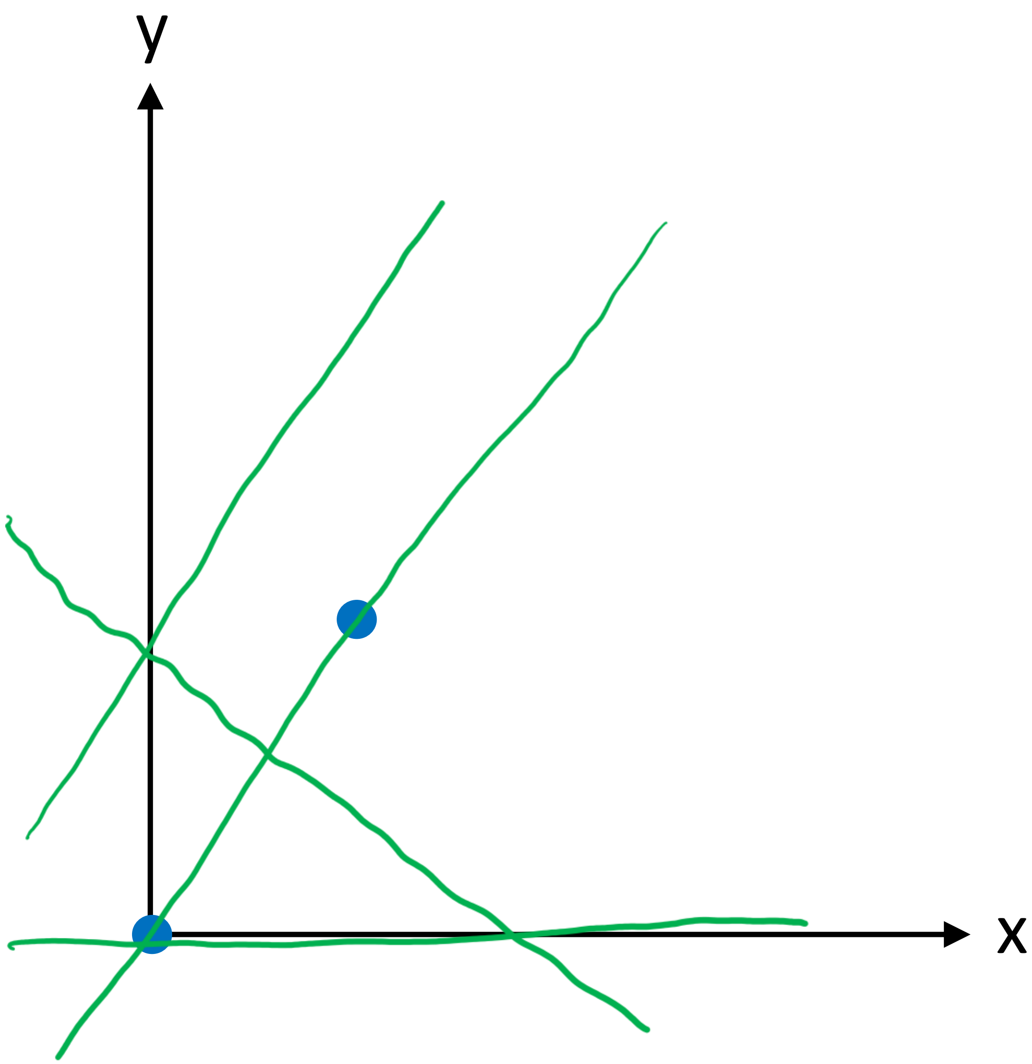
Optimization

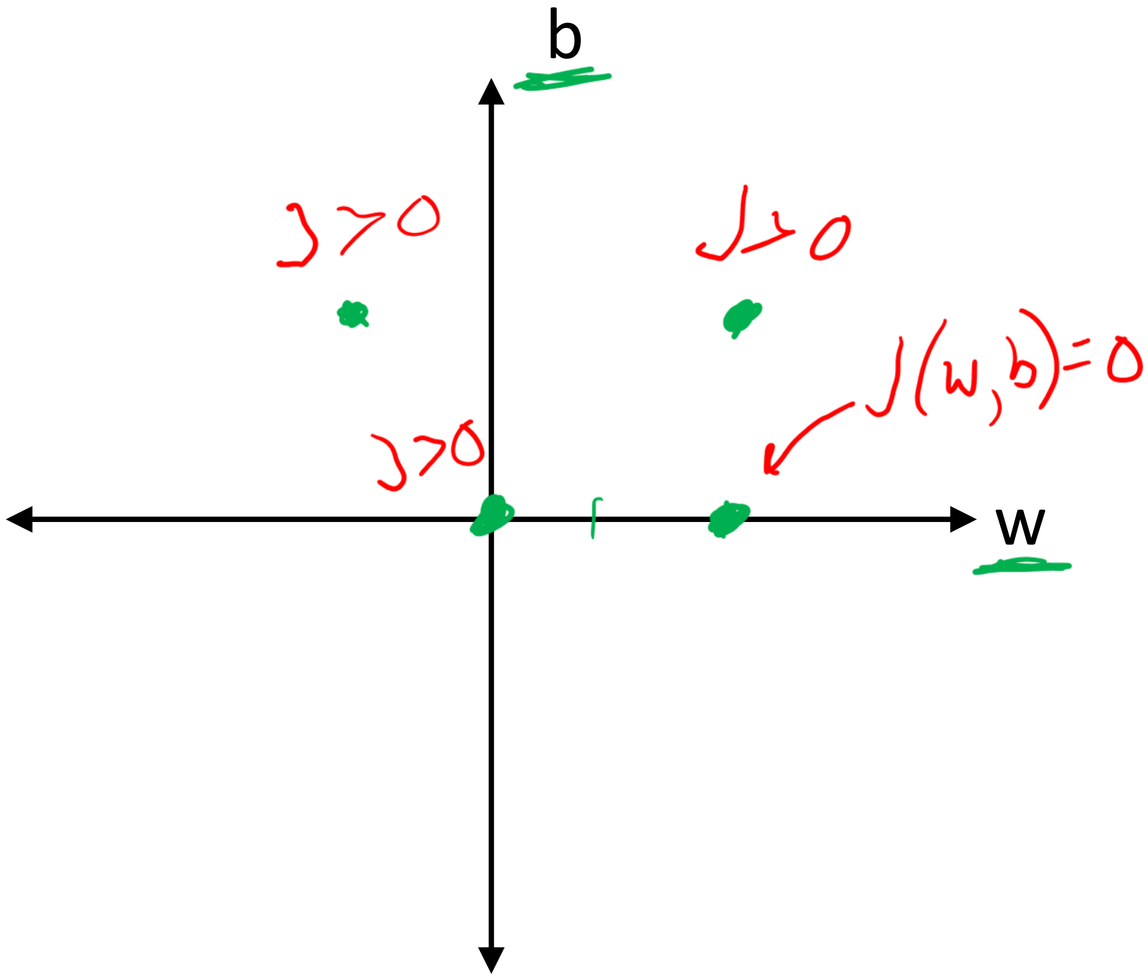
Grid Search, Gradient Descent, Stochastic Gradient Descent

Linear Regression

Optimizing the objective

$$J(w,b) = \frac{1}{2}\left[\left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2 + \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2\right]$$





Poll 5

$$J(w,b) = \frac{1}{2}\left[\left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2 + \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2\right]$$

$$\left(c_1 - (c_2 + b)\right)^2 + \left(c_3 - (c_4 + b)\right)^2$$

For fixed data and fixed slope, w, what shape do we get by plotting MSE objective vs intercept, b?

$J(b)$

A. Line

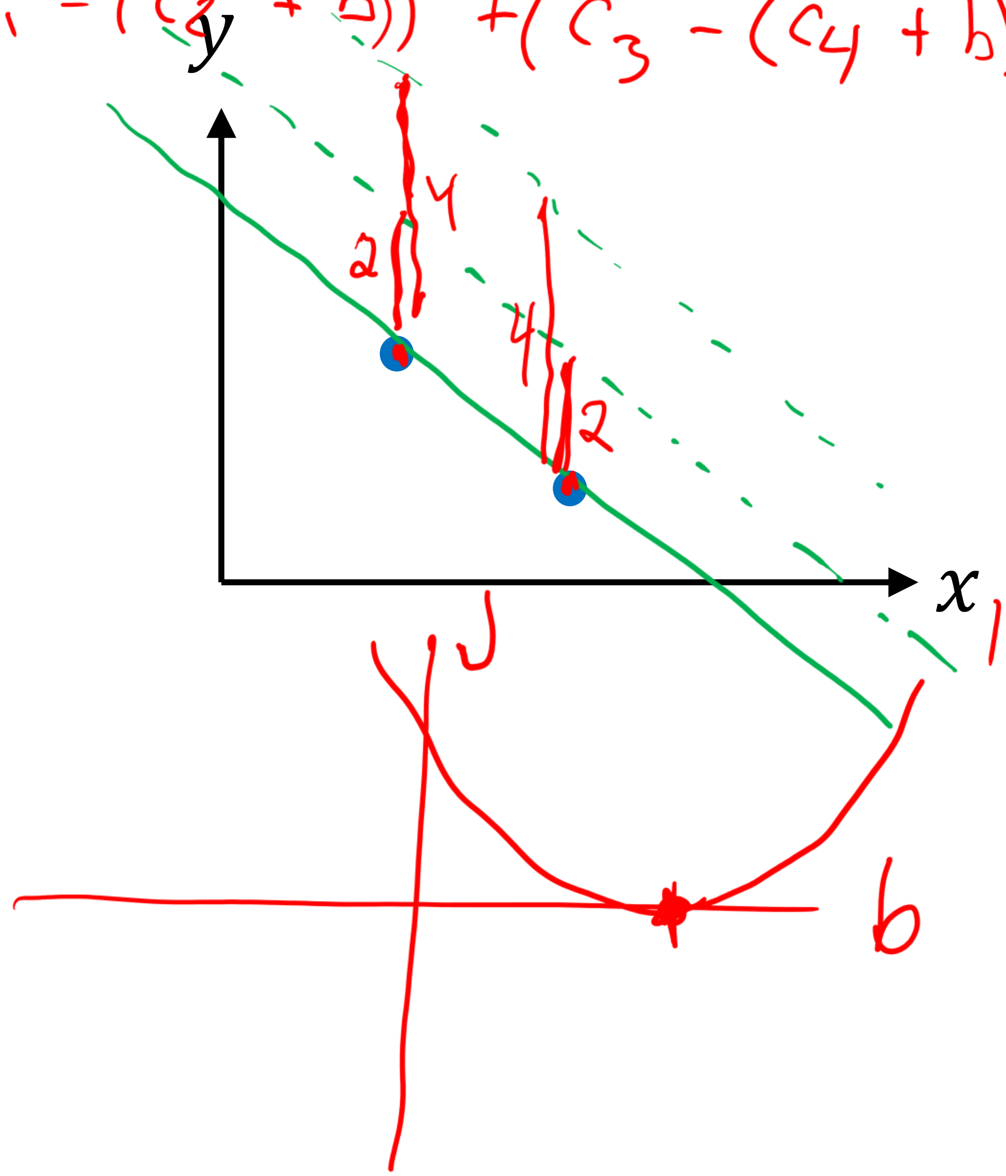
B. Plane

C. Half-plane

D. Convex parabola (U-shape)

E. Concave parabola (up-side-down U)

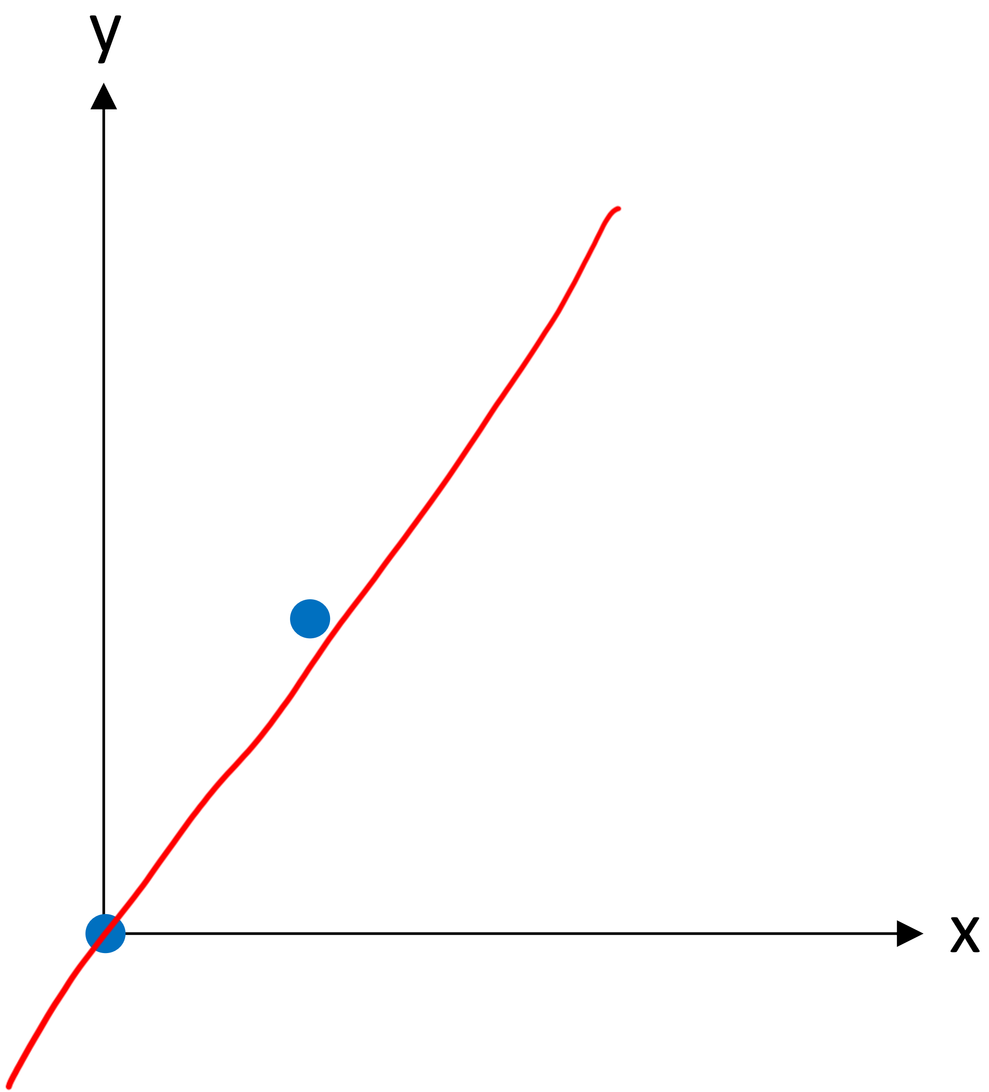
F. None of the above

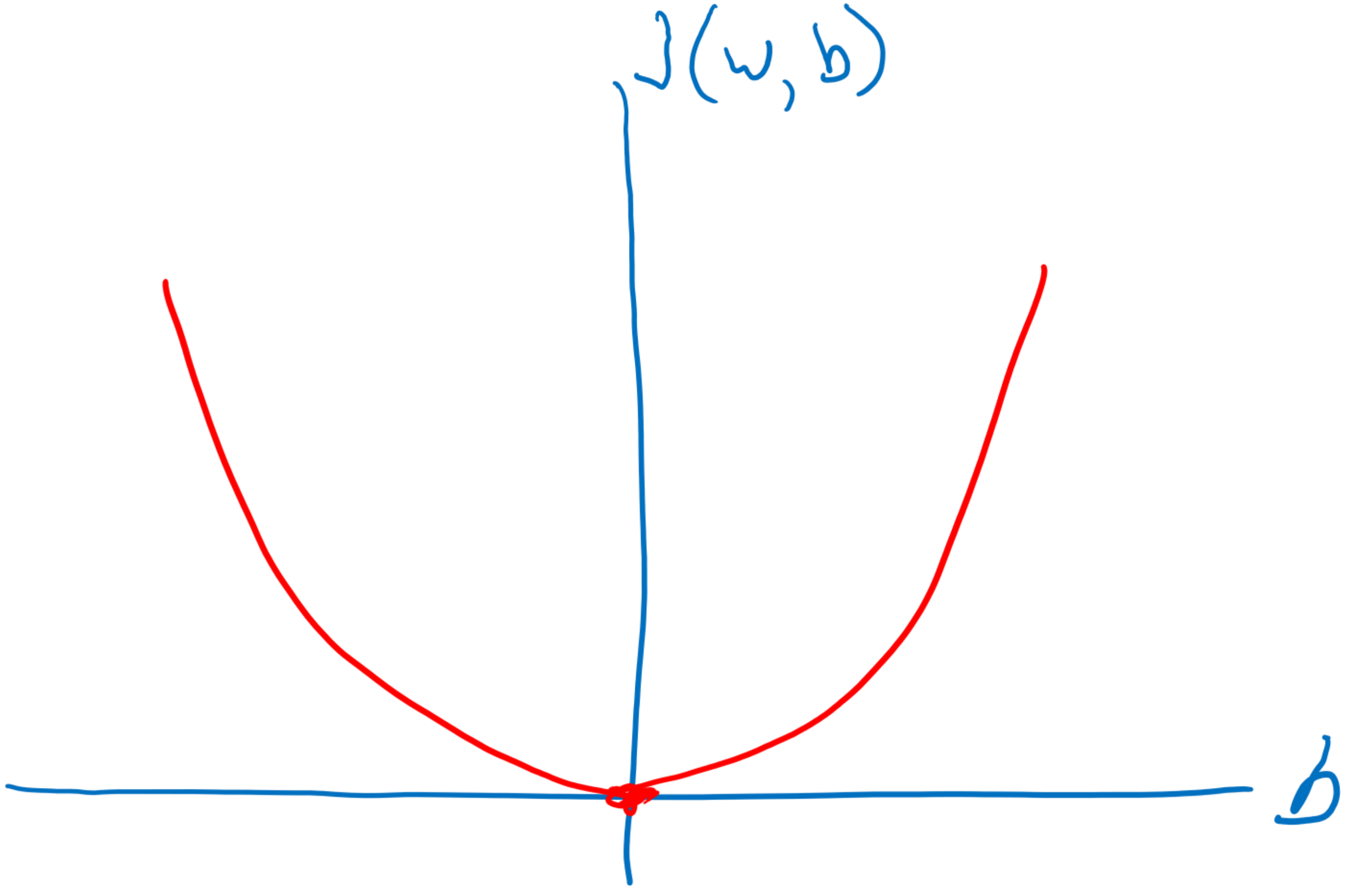


Linear Regression

Optimizing the objective

$$J(w,b) = \frac{1}{2}\left[\left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2 + \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2\right]$$

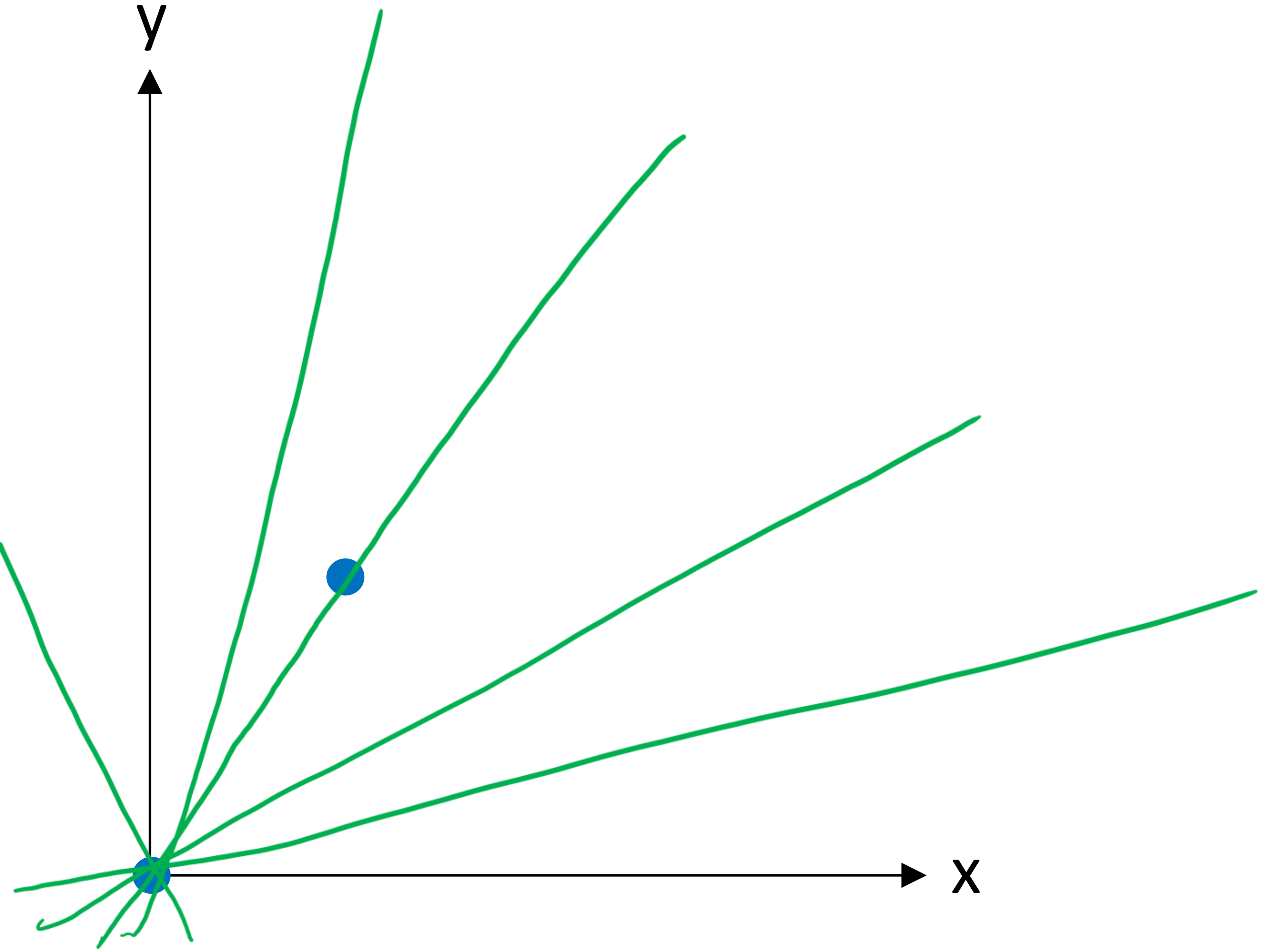


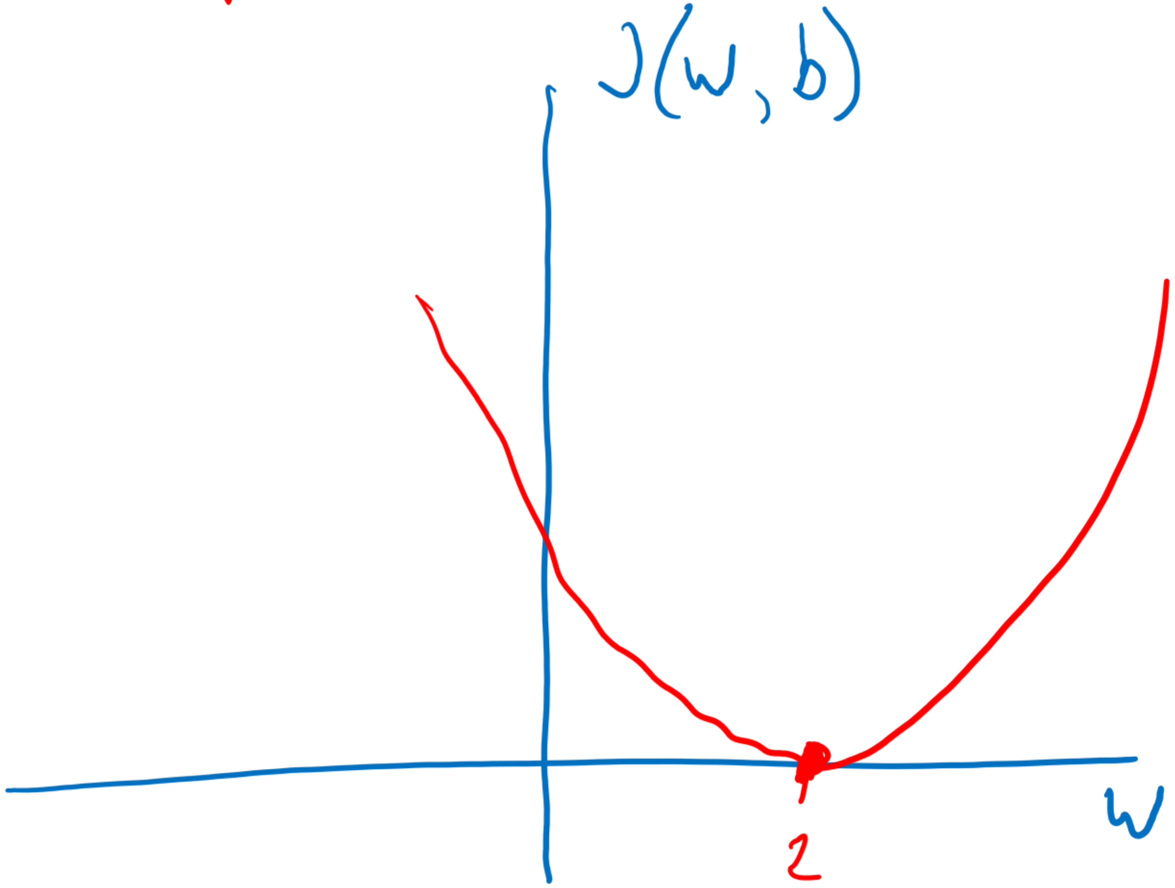


Linear Regression

Optimizing the objective

$$J(w,b) = \frac{1}{2}\left[\left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2 + \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2\right]$$

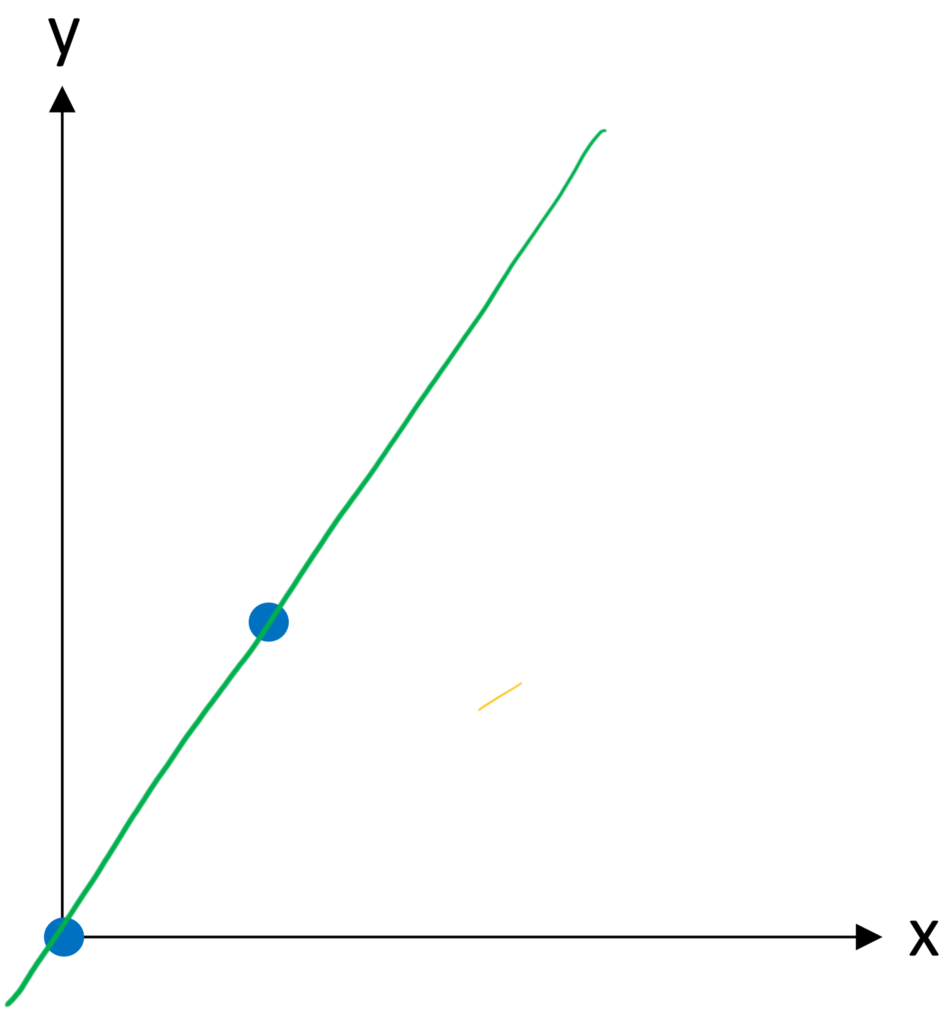


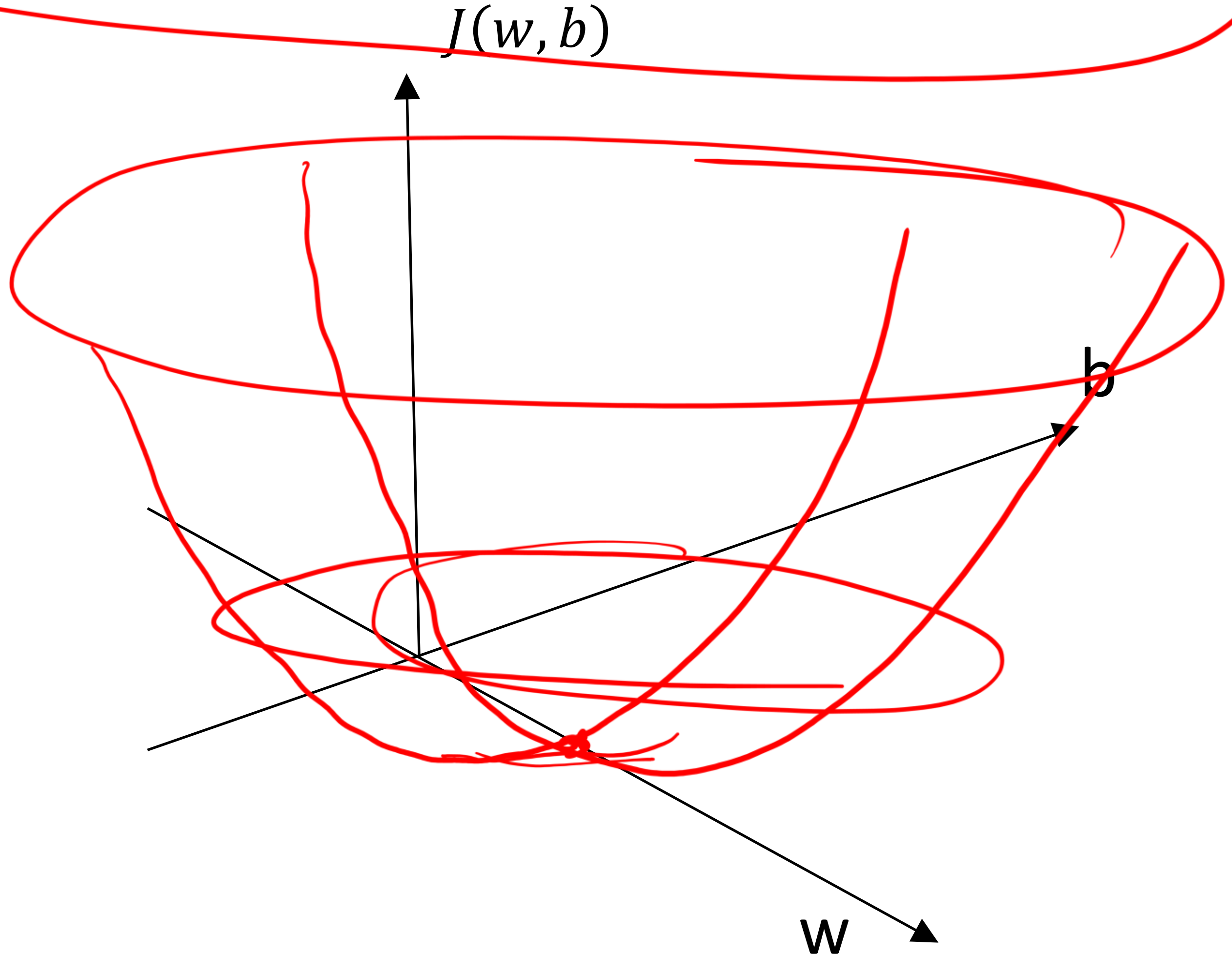


Linear Regression

$$J(w,b) = \frac{1}{2}\left[\left(y^{(1)} - (wx^{(1)} + b)\right)^2 + \left(y^{(2)} - (wx^{(2)} + b)\right)^2\right]$$

What shape is the obejective function when we consider both parameters w and b?

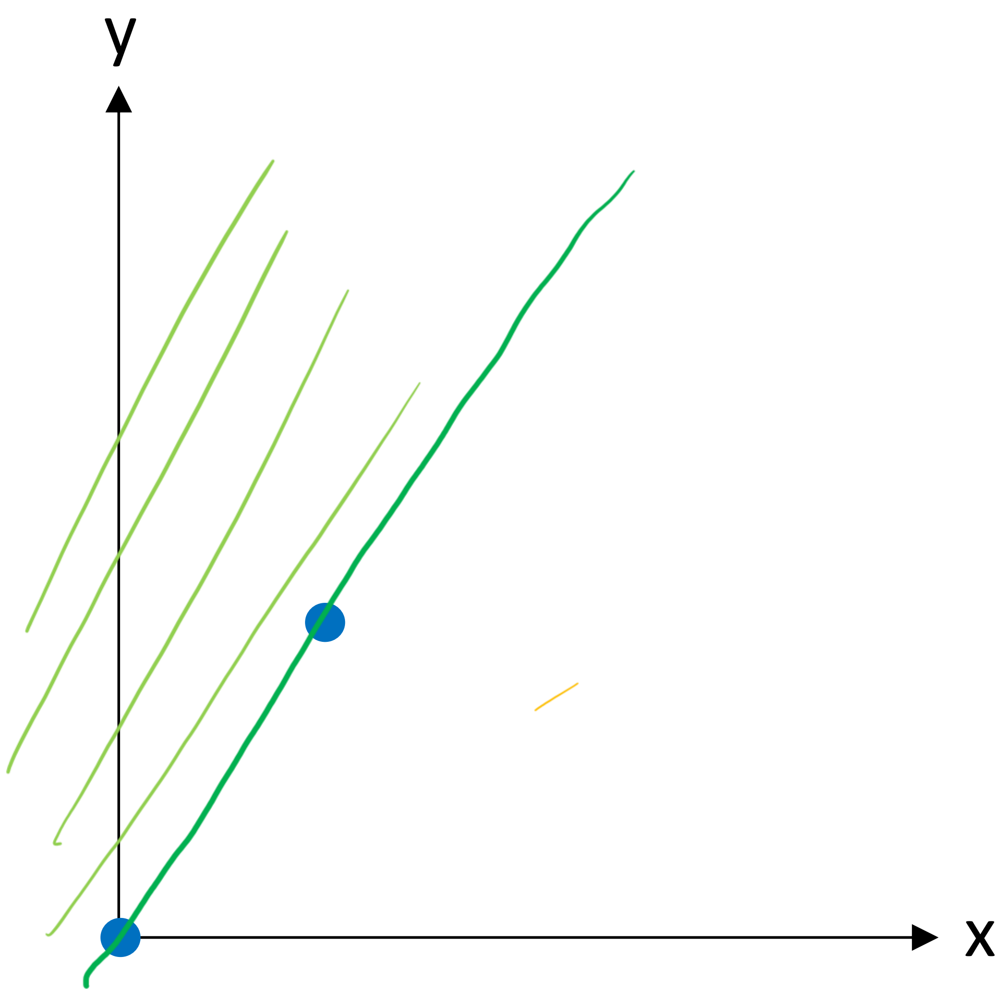


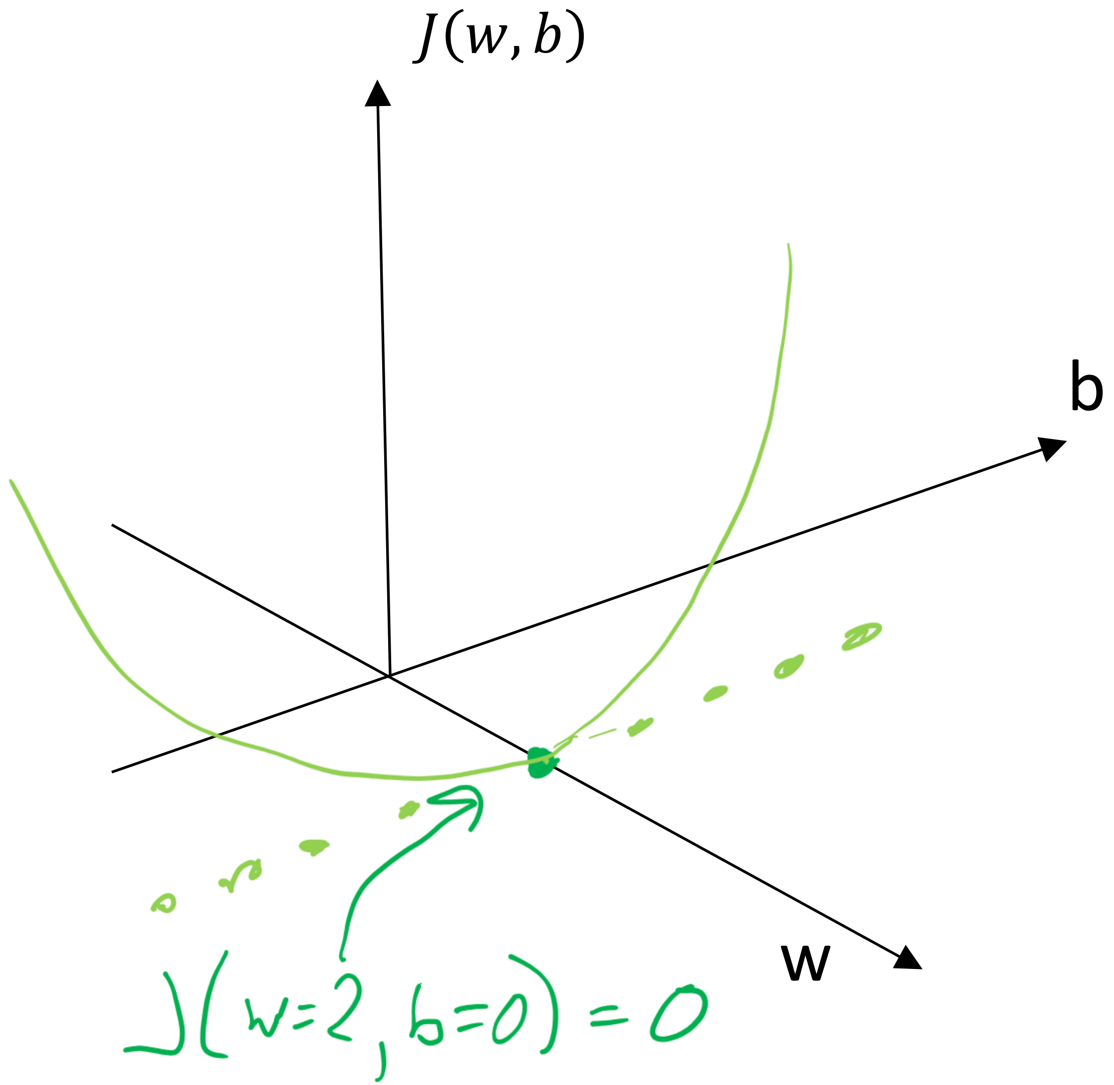


Linear Regression

$$J(w,b) = \frac{1}{2}\left[\left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2 + \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2\right]$$

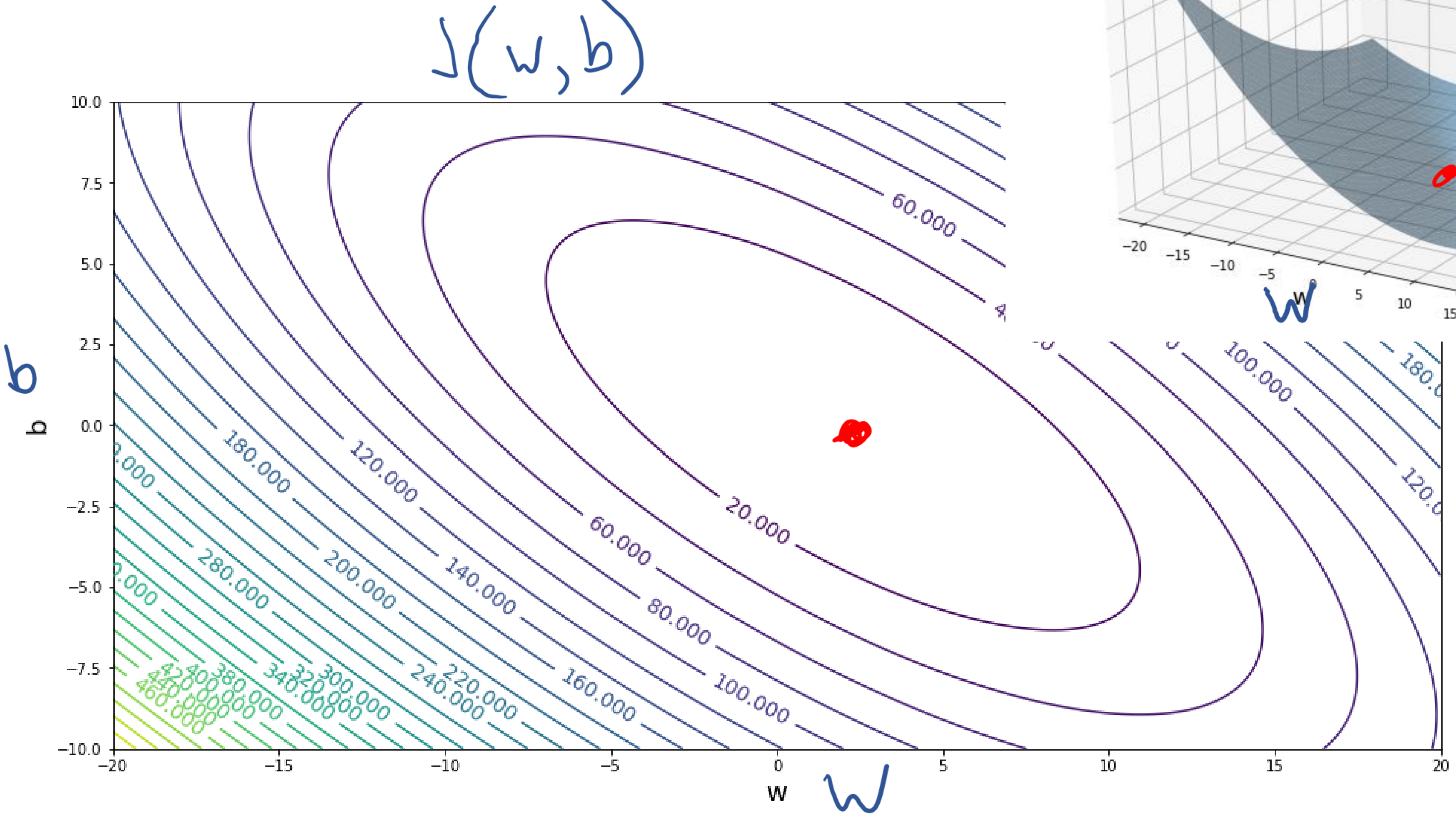
What shape is the obejective function when we consider both parameters w and b?

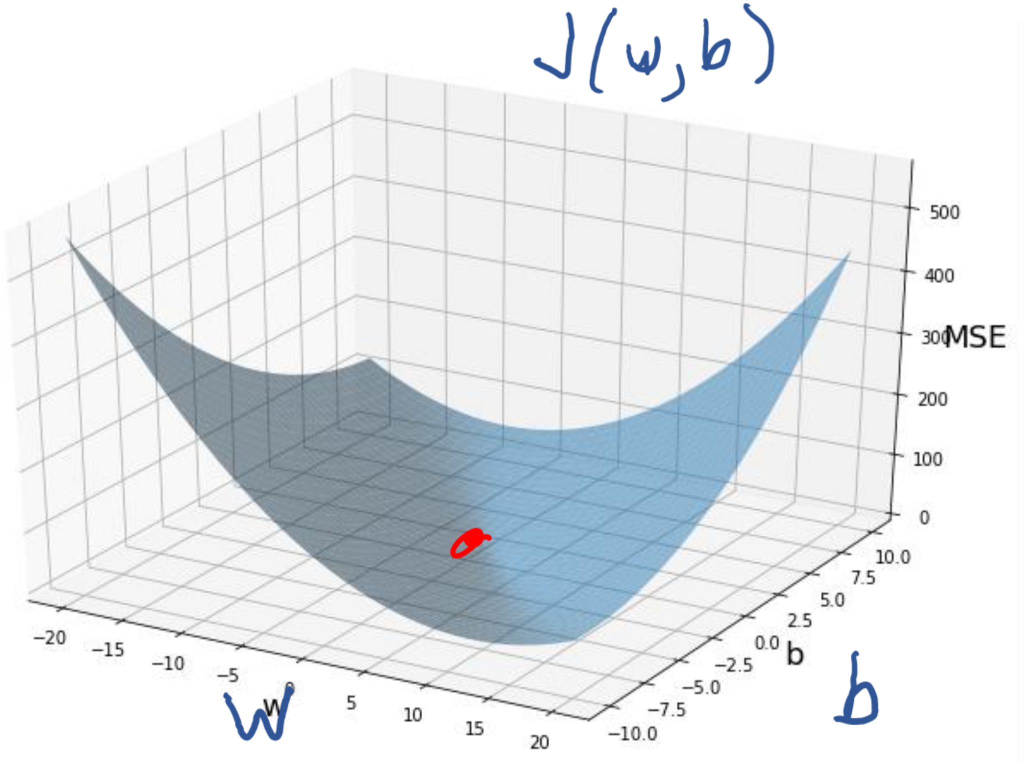




Linear Regression

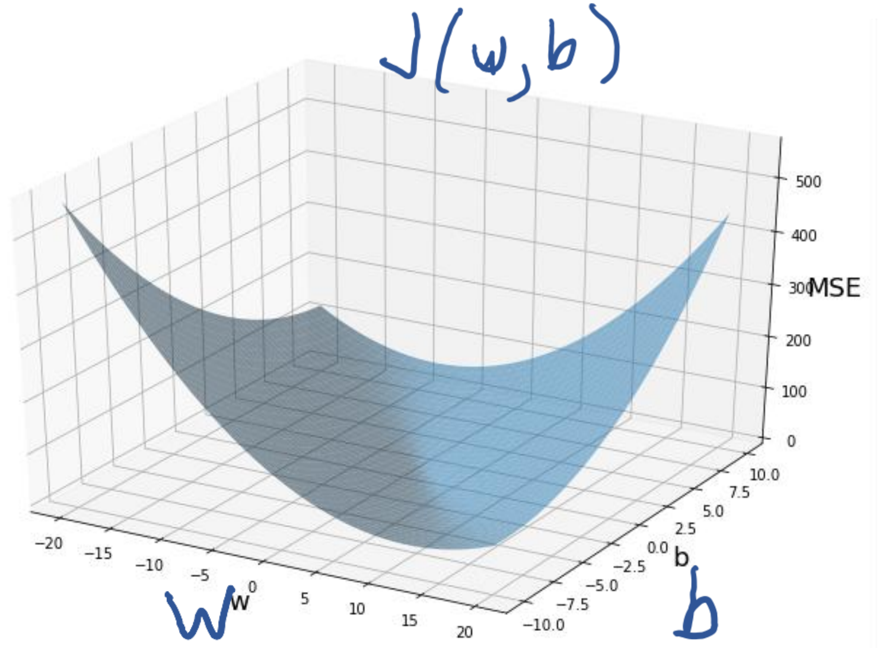
Optimizing the objective

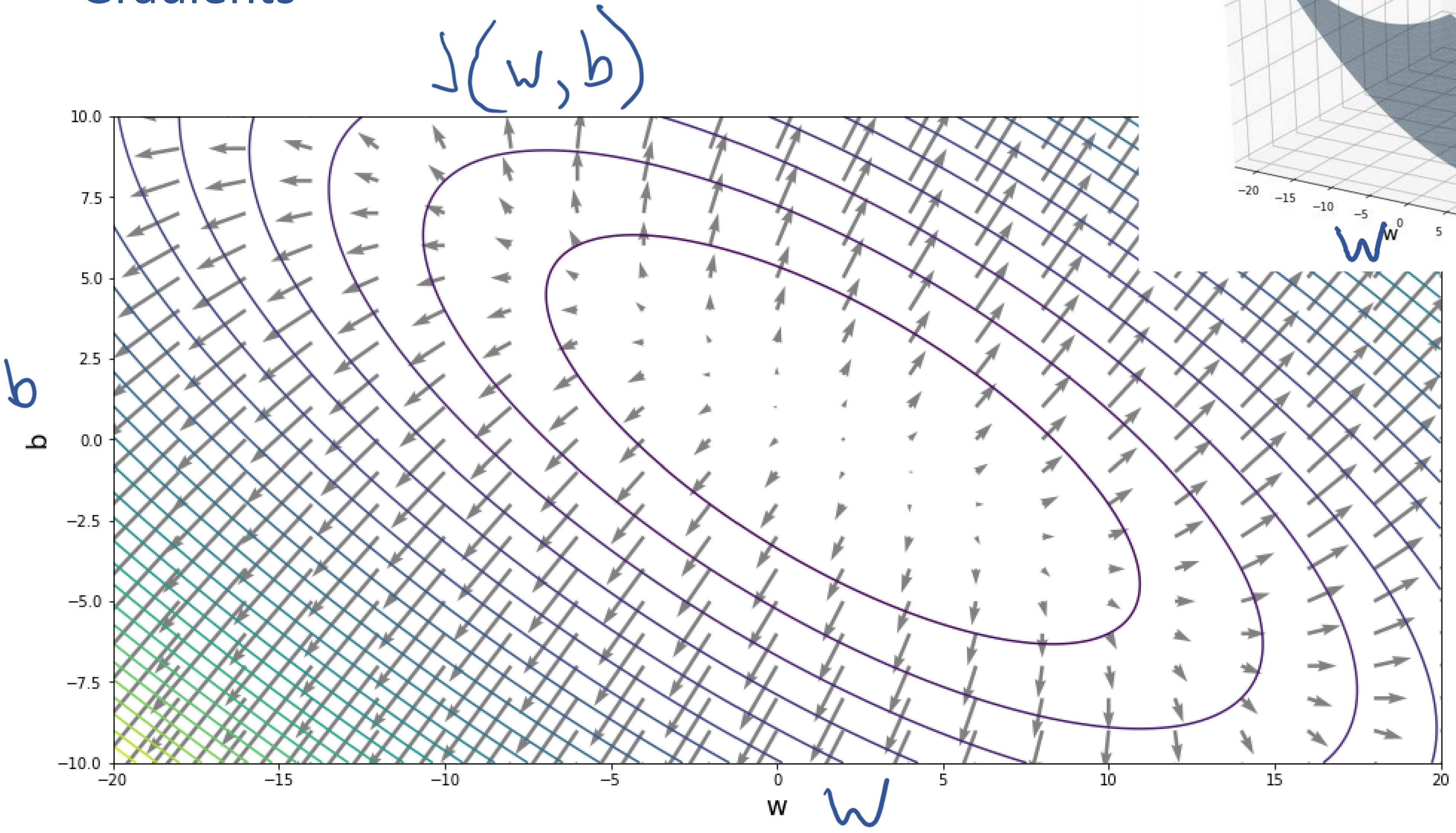




Optimization

Gradients

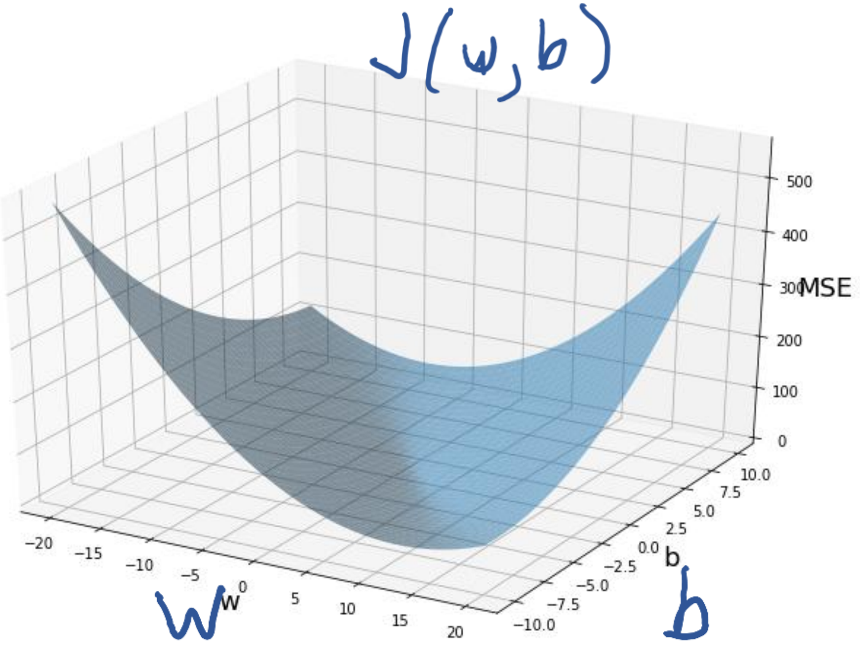


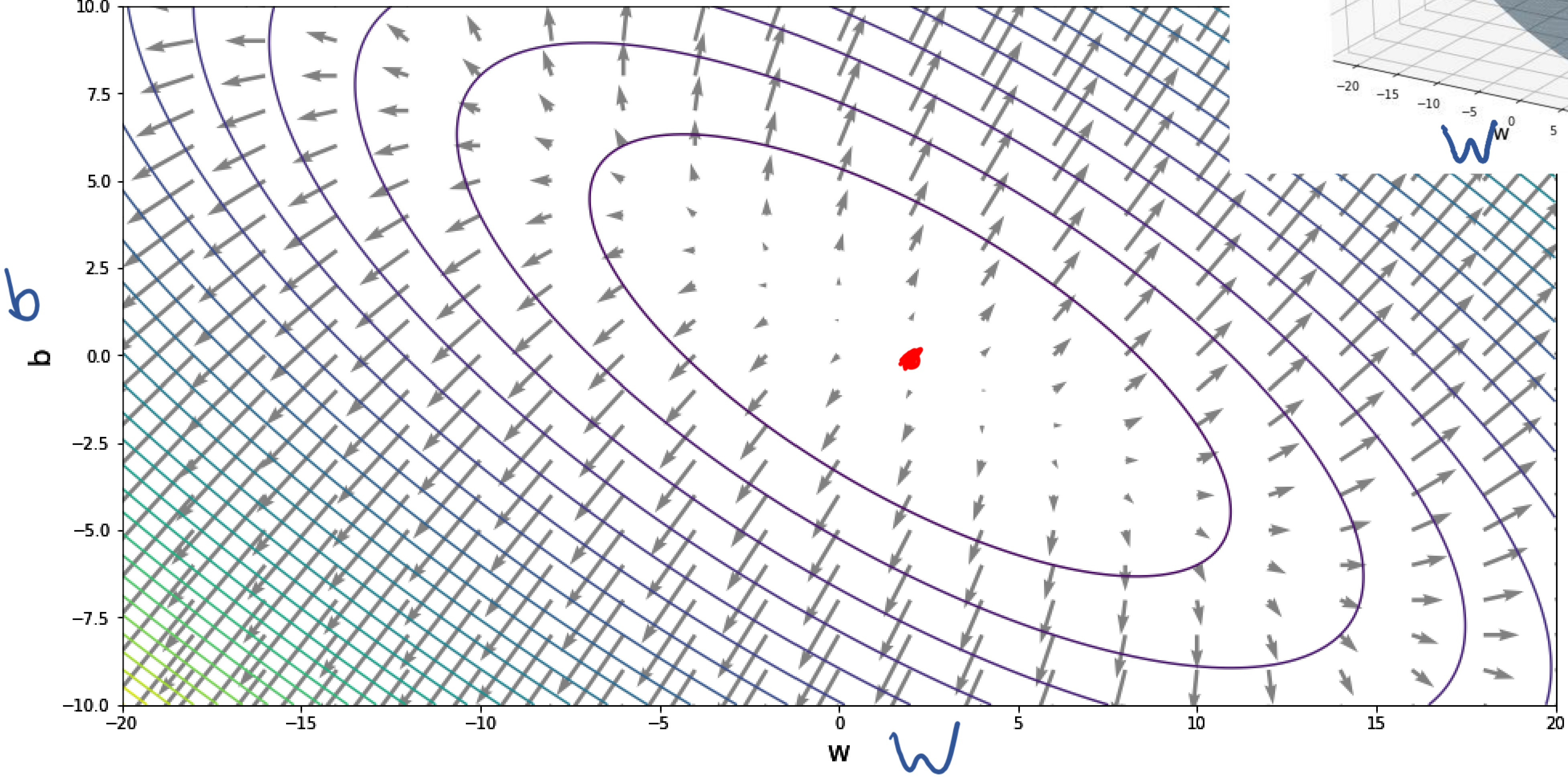


Optimization

Linear regression has a closed-form solution!!

Solve for $\nabla J(w, b) = 0$ for $w,^* \ b^*$





$$\nabla J(\theta) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Optimization

$$J(w,b) = \frac{1}{2}\left[\left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2 + \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2\right]$$

Gradients

Linear Regression

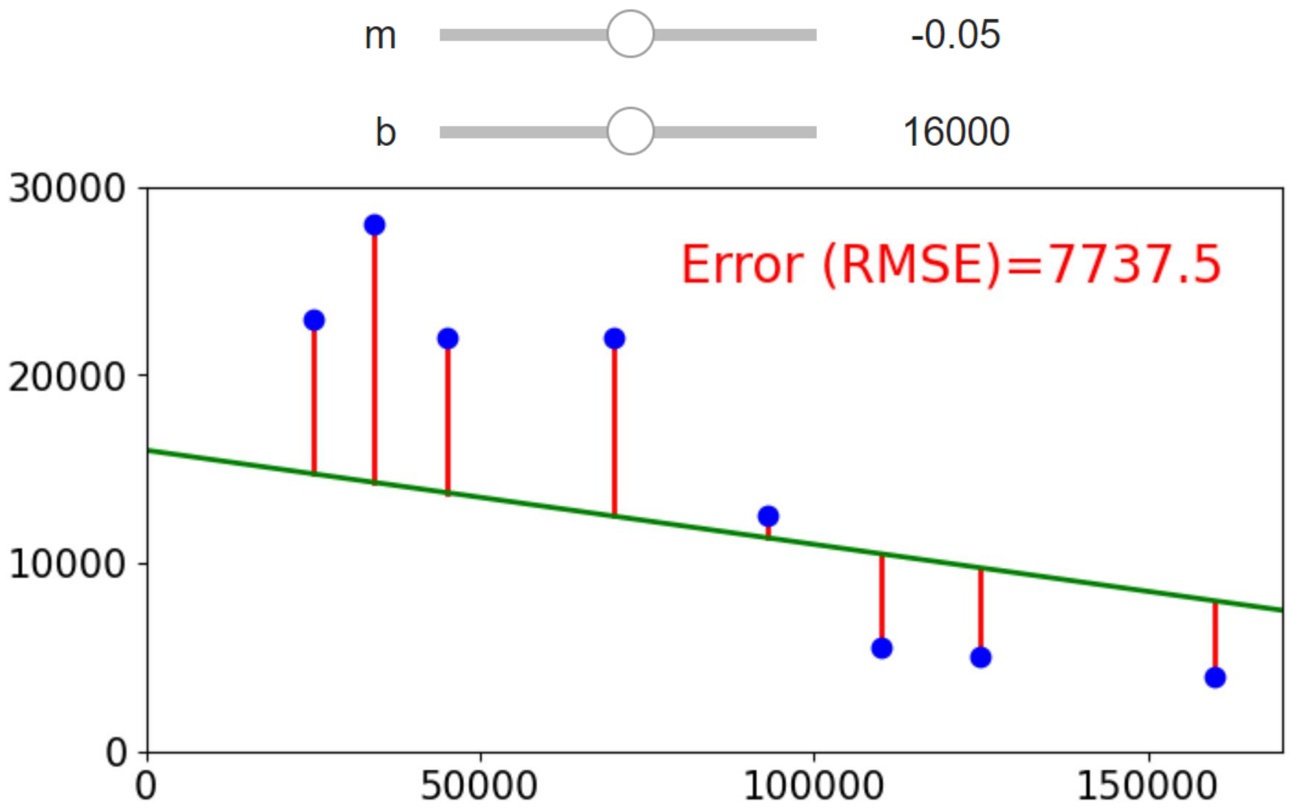
Gradient Descent

Linear Regression

Demos to optimize the objective

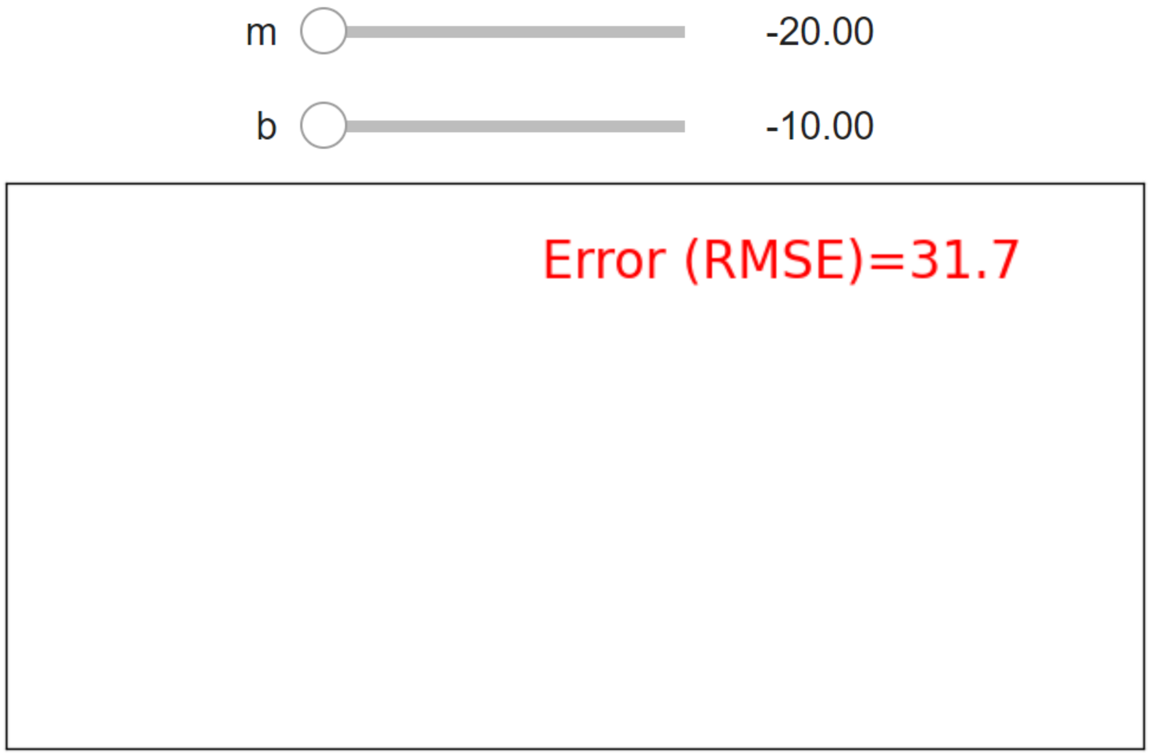
Linear model optimization

regression_interactive.ipynb



Blind version

regression_blind_interactive.ipynb



Poll 6

Open blind optimization on course website

Mess with sliders to find best model

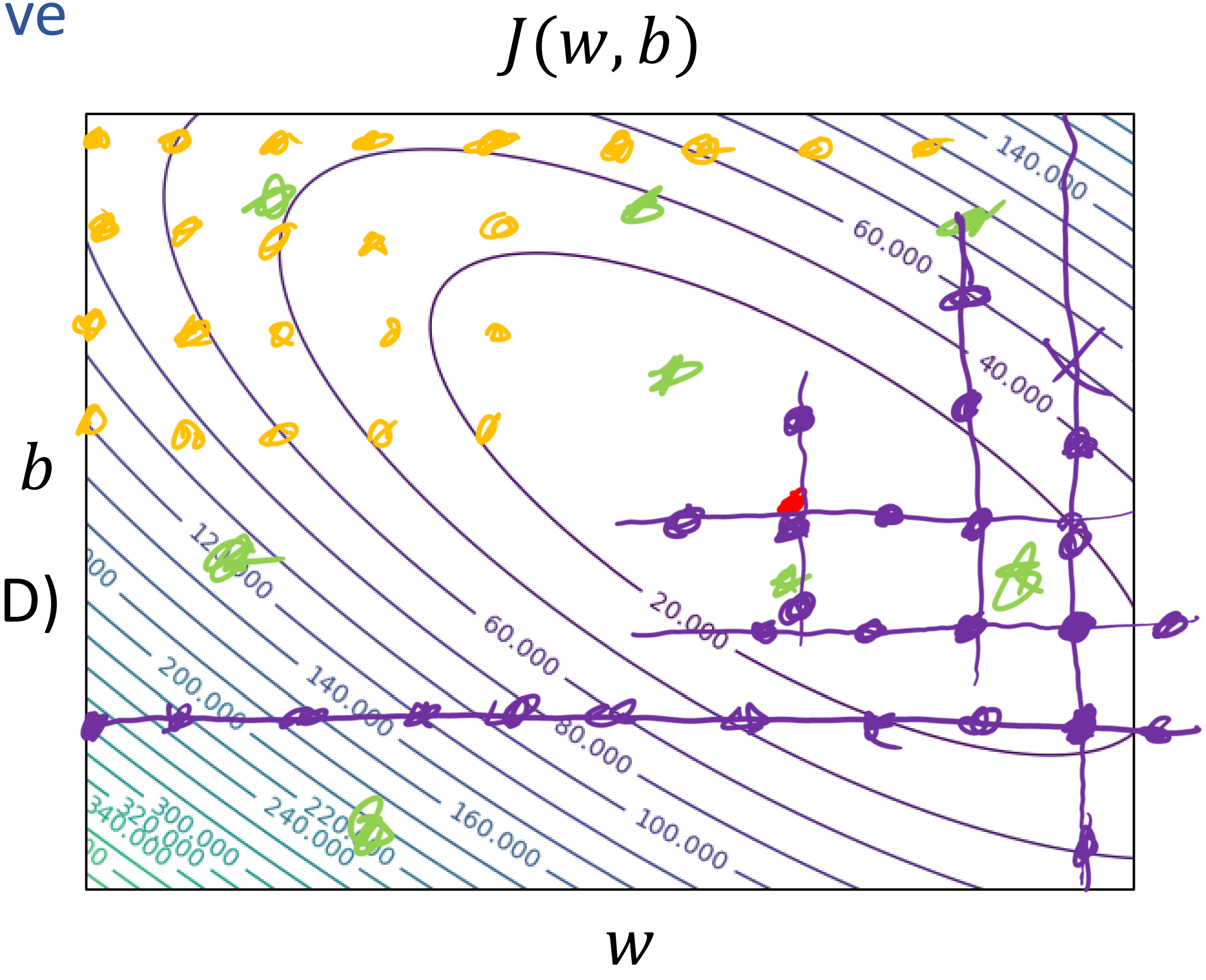
What's the best RMSE that you got?

What's are the corresponding best parameters?

Linear Regression

Methods for optimizing the objective

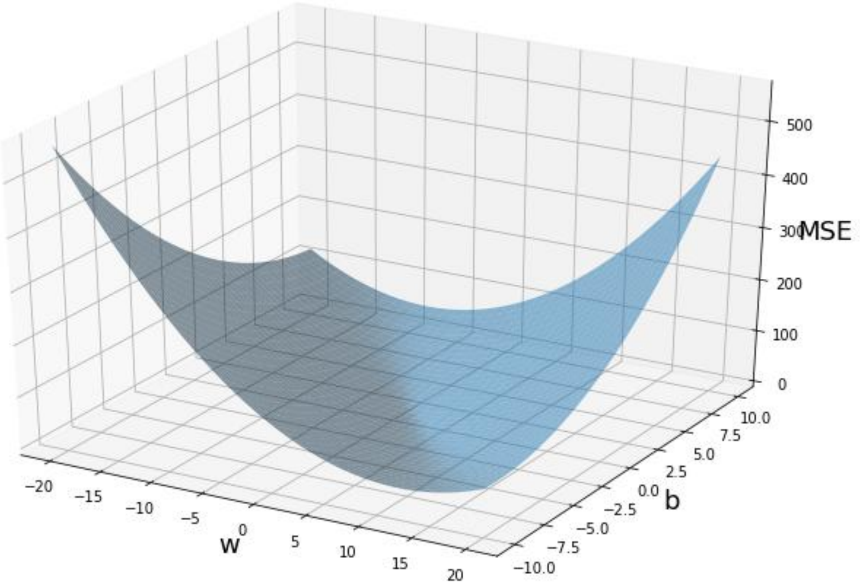
- Closed-form solution
- Grid search
- Random search
- Gradient descent
- Stochastic gradient descent (SGD)
- Minibatch SGD
- Coord descent (slider) grid

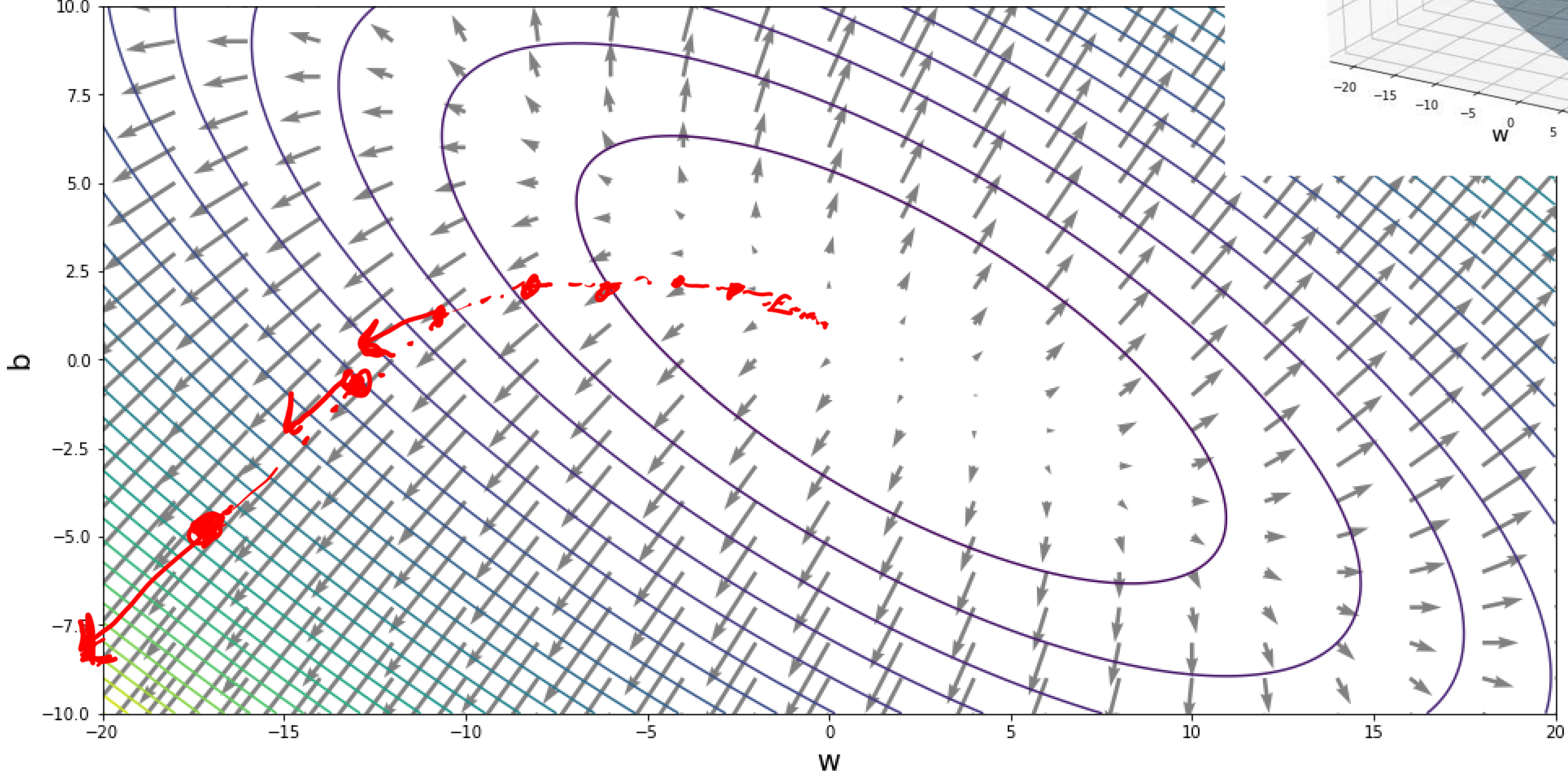


Optimization

Gradients

$$\nabla_\theta J = \frac{\partial J}{\partial \theta} = \begin{bmatrix} \partial J / \partial w \\ \partial J / \partial b \end{bmatrix}$$





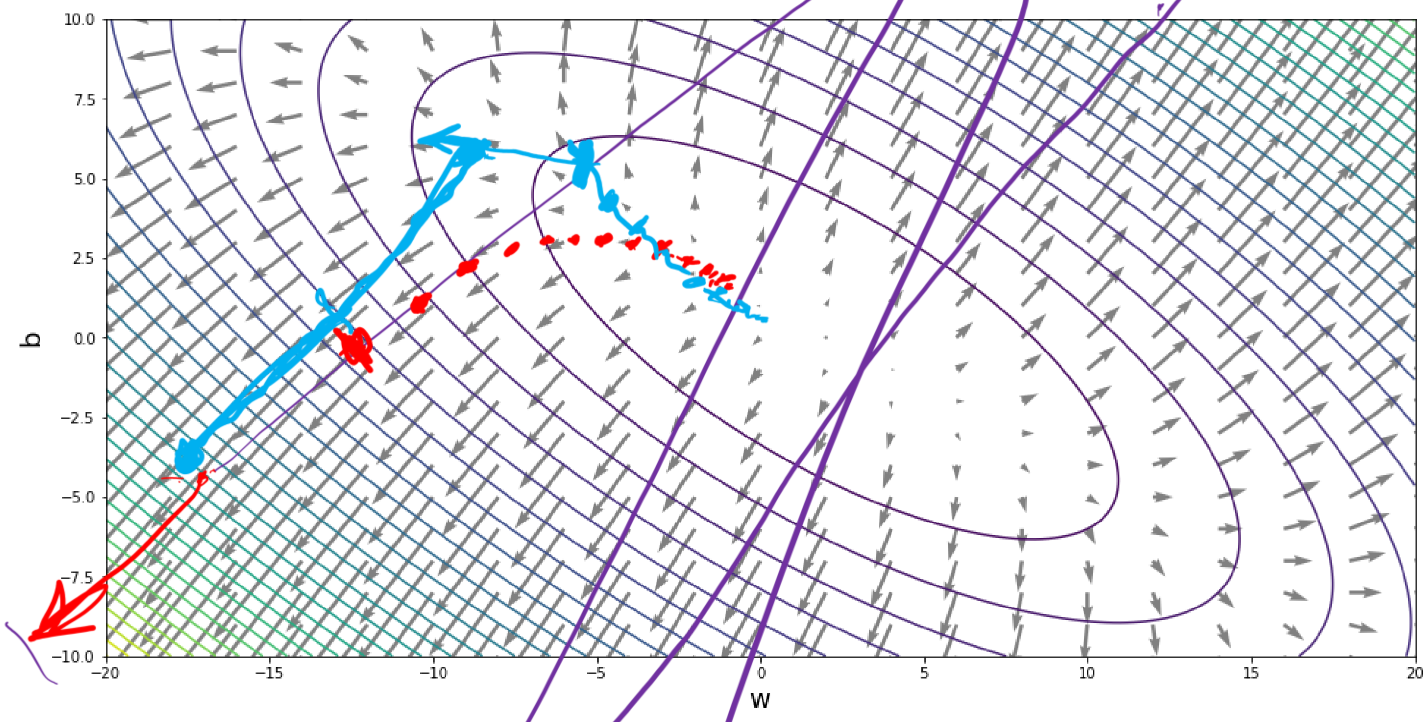
Optimization

Gradient descent

Hyperparameter:

learning rate α (or sometimes γ, η)

$\alpha = 10$
$\alpha = 2$
$\alpha = 1$



While not converged

Compute gradient: $\nabla J = \left[\frac{\partial J}{\partial w}, \frac{\partial J}{\partial b}\right]^{\mathsf{T}}$

Update parameters

$$w \leftarrow w - \alpha \frac{\partial J}{\partial w}$$

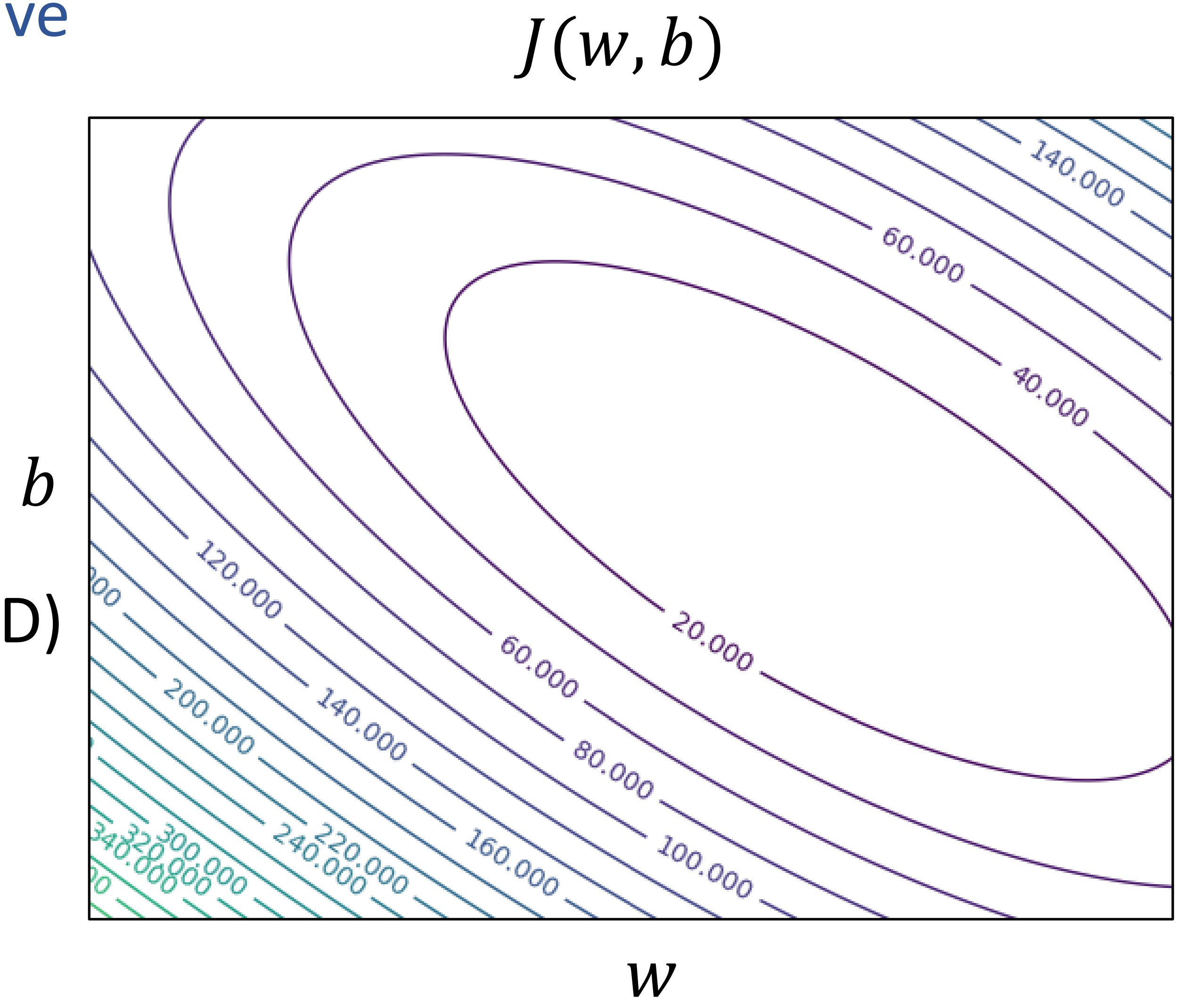
$$b \leftarrow b - \alpha \frac{\partial J}{\partial b}$$

$$\vec{\theta}^{(t+1)} \leftarrow \vec{\theta}^{(t)} - \alpha \frac{dJ}{d\vec{\theta}}$$

Linear Regression

Methods for optimizing the objective

- Closed-form solution
- Grid search
- Random search
- Gradient descent (Batch)
- **Stochastic gradient descent (SGD)**
- Minibatch SGD



Linear Regression Gradient Descent

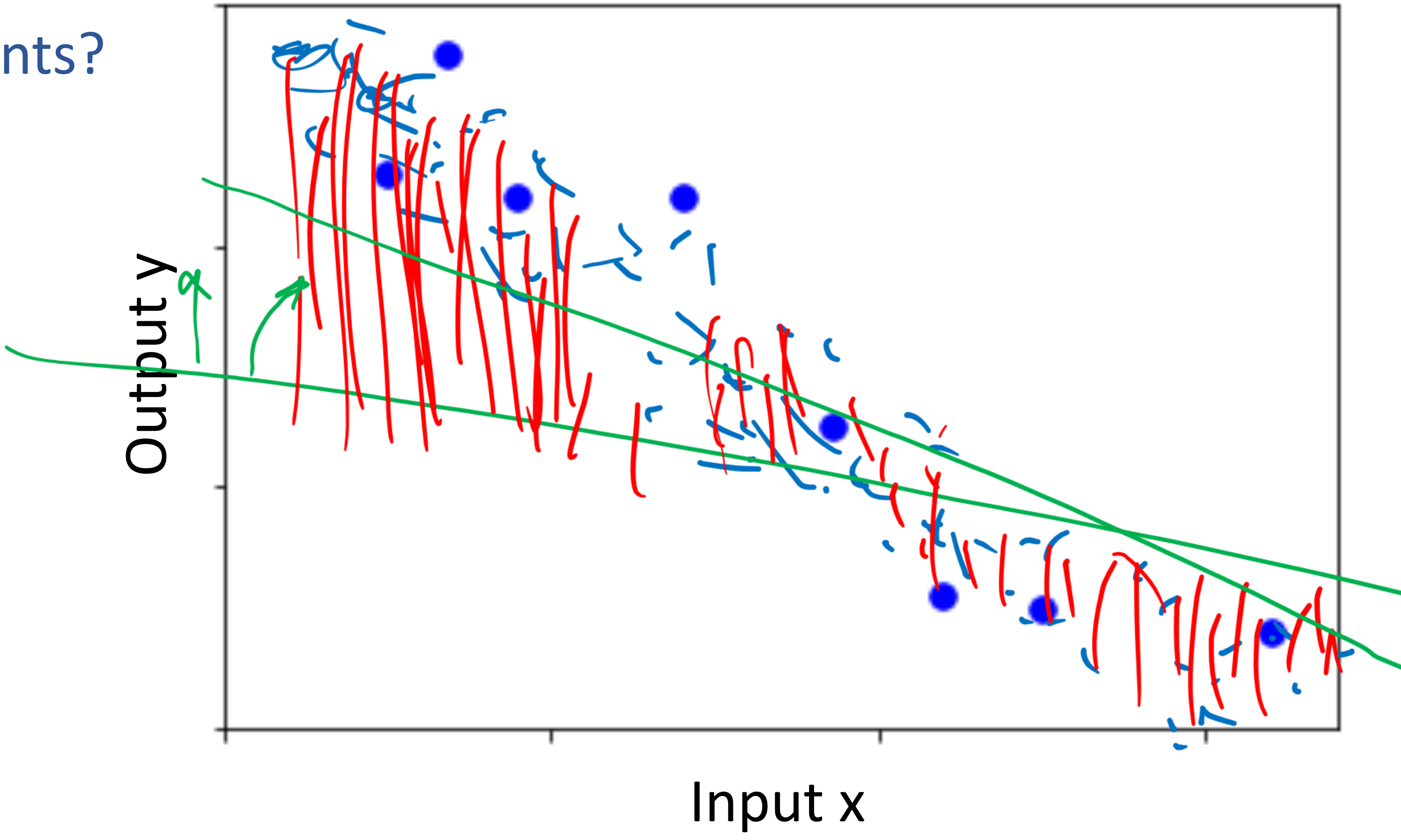
$$\theta \leftarrow \theta - \alpha \frac{\partial J}{\partial \theta}$$

What happens in gradient descent when we have N=1,000,000 training points?

Y 1000000 x 1

w, b

X



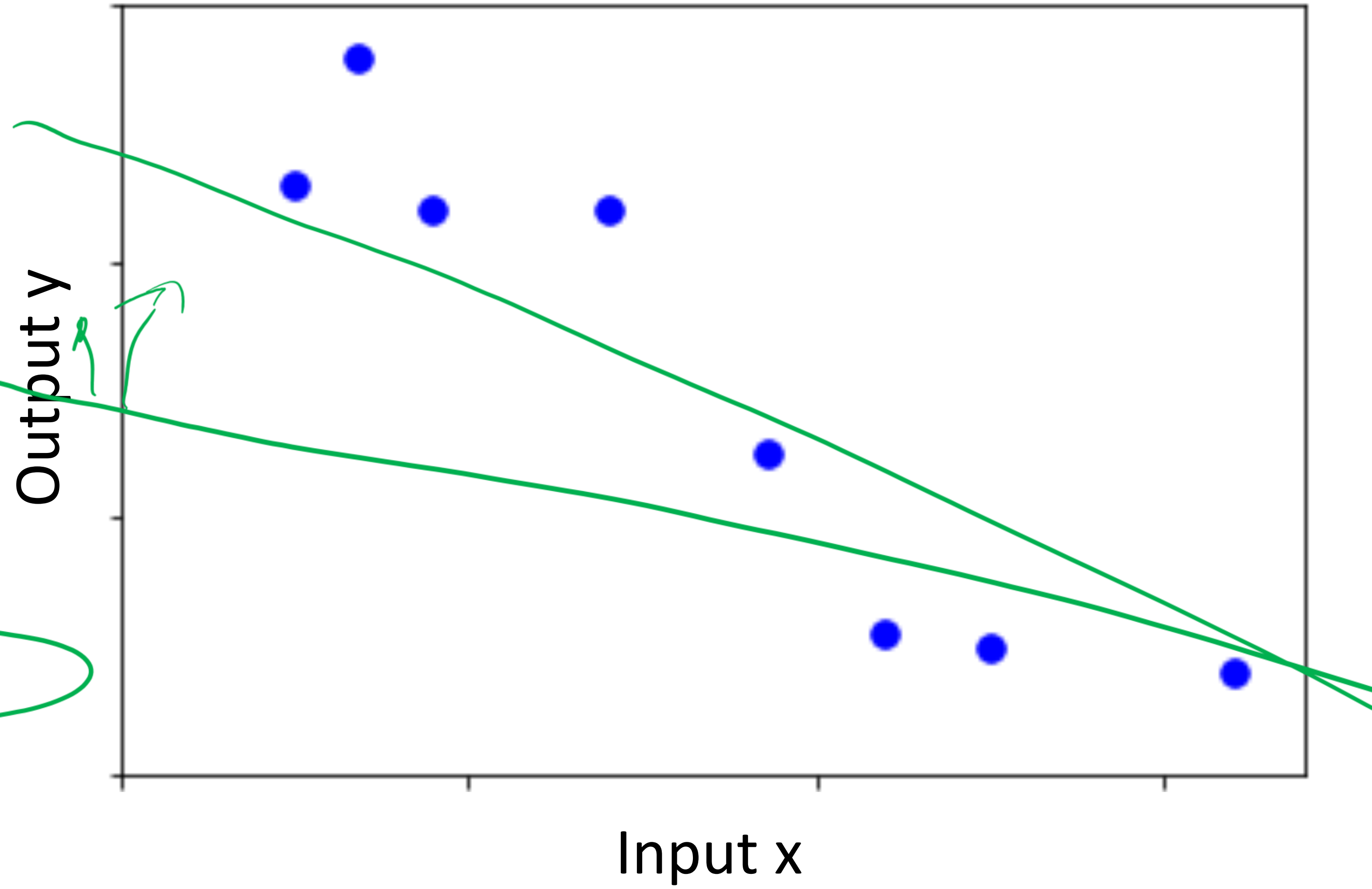
Linear Regression Gradient Descent

$$\theta \leftarrow \theta - \alpha \frac{dJ^{(b)}}{d\theta}$$

What happens in gradient descent when we have N=1,000,000 training points?

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N=1000000}$$

$$J^{(b)}(\theta) = \frac{1}{B} \sum_{i=1}^{B=8}$$



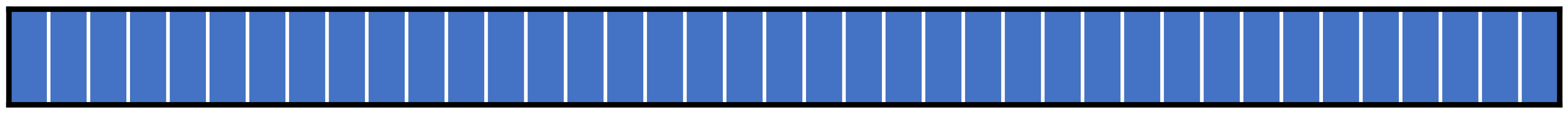
Gradient Descent

Epoch
one pass through dataset

Batch -> Mini-batch -> Stochastic

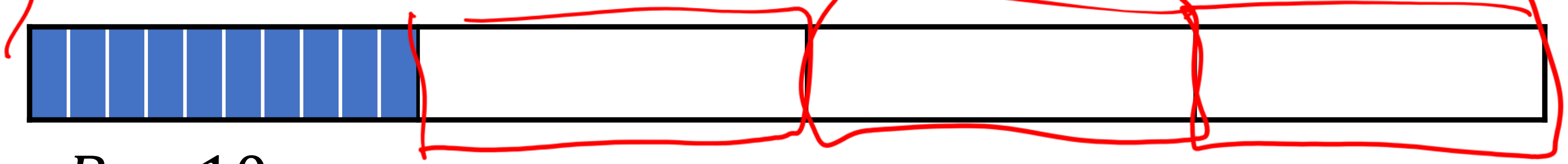
Basically, just a difference in how many training points are each time we compute a gradient (and then take a step with that gradient

(Batch) GD



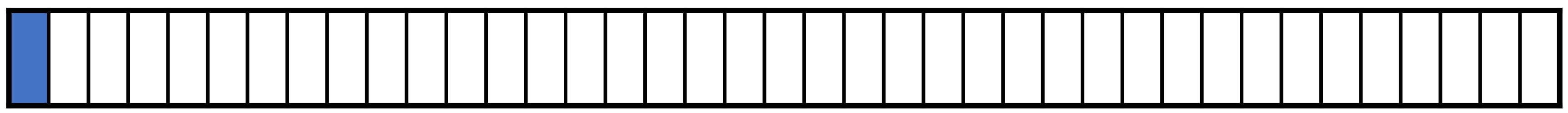
Effective batch size = N

Mini-batch SGD



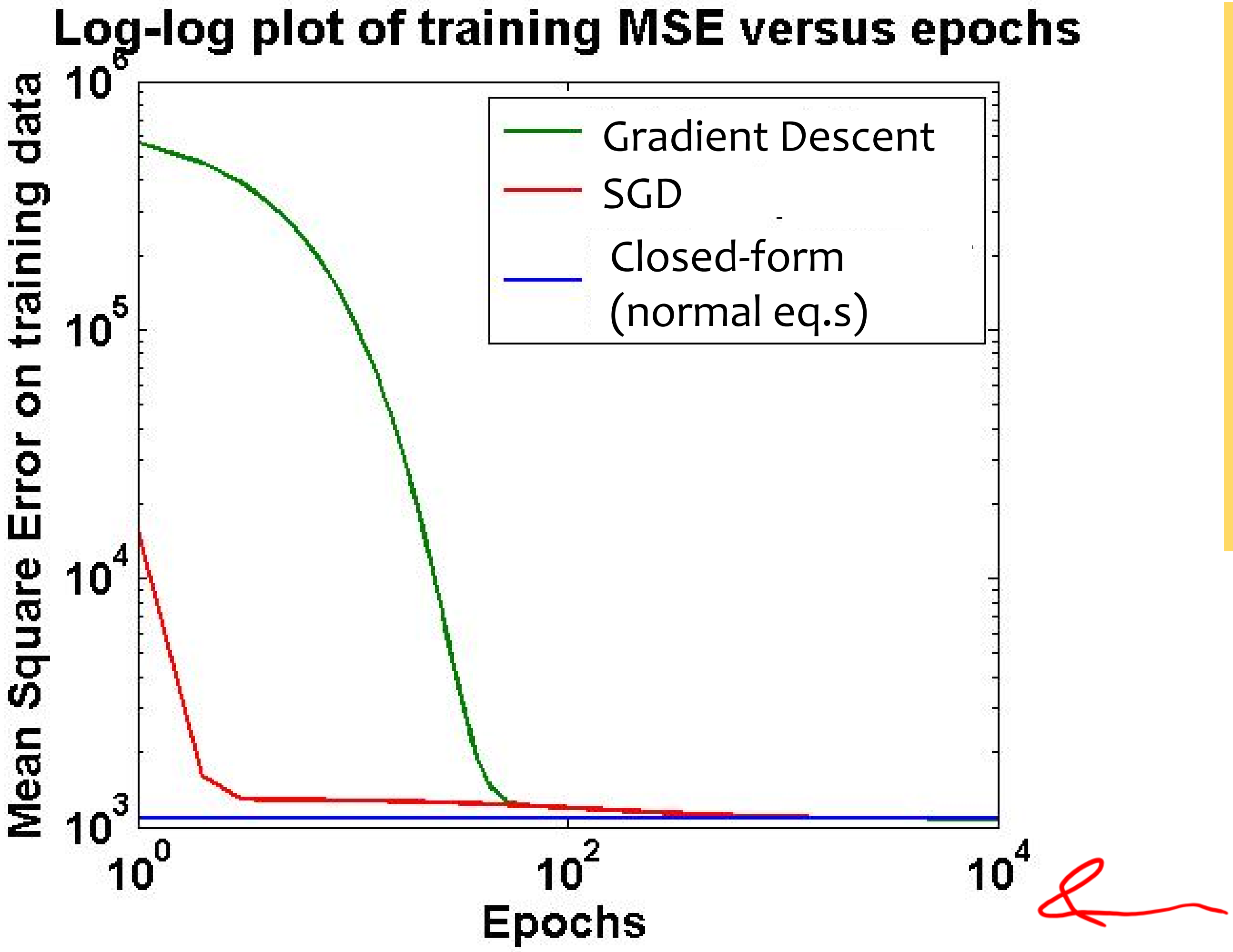
Batch size = B, e.g., $B = 10$

SGD



Effective batch size = 1

Convergence Curves



- *Def*: an **epoch** is a single pass through the training data

1. For GD, only **one update** per epoch
2. For SGD, ***N* updates** per epoch
   N = (# train examples)

- SGD reduces MSE much more rapidly than GD
- For GD / SGD, training MSE is initially large due to uninformed initialization

Additional Slides

(Batch) Gradient Descent and SGD

(Batch) Gradient Descent

Algorithm 1 Gradient Descent

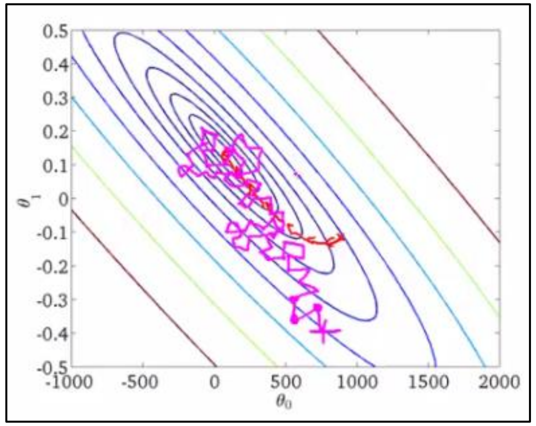
procedure GD($\mathcal{D}$, $\boldsymbol{\theta}^{(0)}$)
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$
  while not converged **do**
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
  return $\boldsymbol{\theta}$

Slide credit: CMU MLD Matt Gormley

Stochastic Gradient Descent (SGD)

Algorithm 2 Stochastic Gradient Descent (SGD)

procedure SGD($\mathcal{D}, \boldsymbol{\theta}^{(0)}$)
 $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$
 while not converged **do**
  $i \sim$ Uniform($\{1, 2, \ldots, N\}$)
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J^{(i)}(\boldsymbol{\theta})$
 return θ



We need a per-example objective:

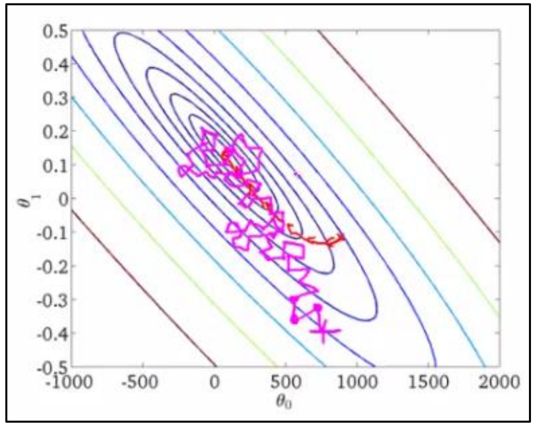
$$\text{Let } J(\boldsymbol{\theta}) = \textstyle\sum_{i=1}^{N} J^{(i)}(\boldsymbol{\theta})$$

Slide credit: CMU MLD Matt Gormley

Stochastic Gradient Descent (SGD)

Algorithm 2 Stochastic Gradient Descent (SGD)

procedure SGD($\mathcal{D}$, $\boldsymbol{\theta}^{(0)}$)
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)}$
    while not converged **do**
        for $i \in$ shuffle($\{1, 2, \ldots, N\}$) **do**
            $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J^{(i)}(\boldsymbol{\theta})$
    return $\boldsymbol{\theta}$



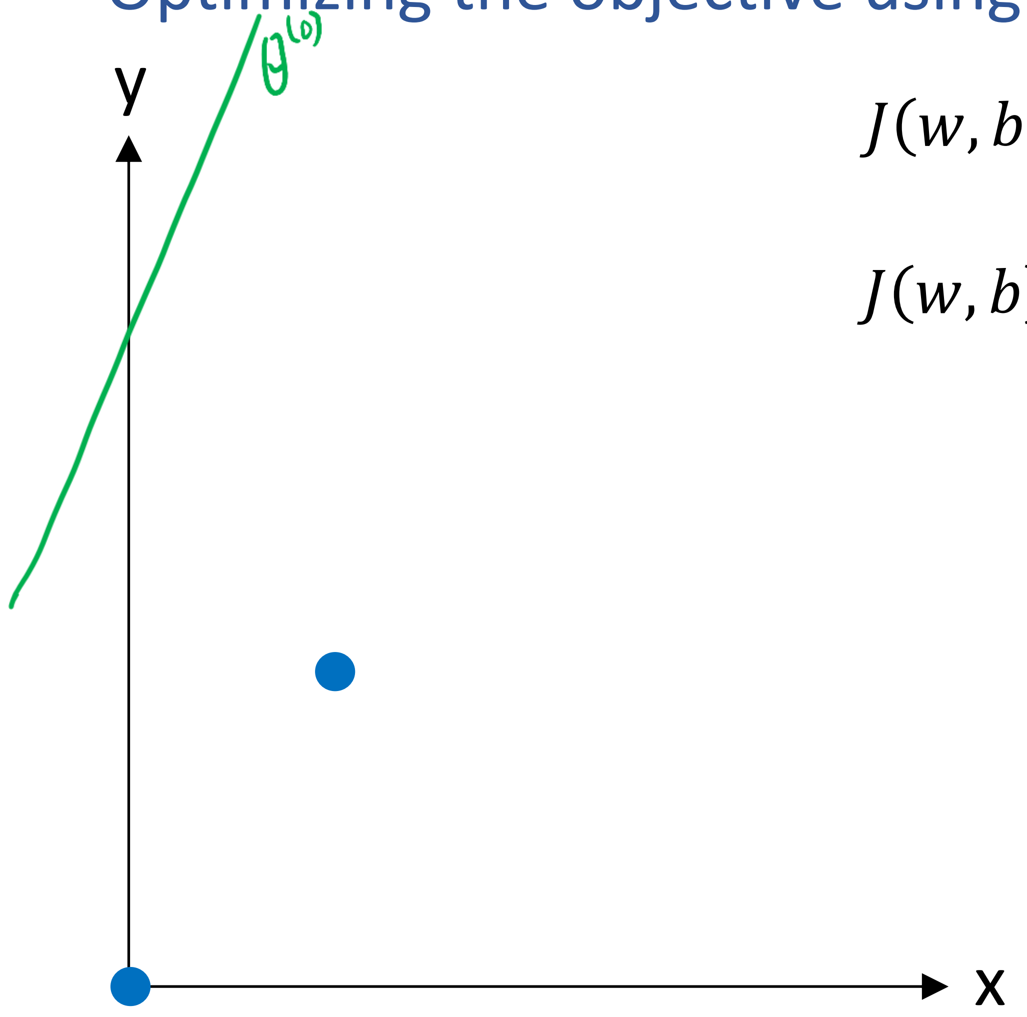
In practice, it is common to implement SGD using sampling **without** replacement (i.e. shuffle({1,2,… N}), even though most of the theory is for sampling **with** replacement (i.e. Uniform({1,2,… N}).

We need a per-example objective:

$$\text{Let } J(\boldsymbol{\theta}) = \sum_{i=1}^{N} J^{(i)}(\boldsymbol{\theta})$$

(Batch) Gradient Descent

Optimizing the objective using all points (just two in this example)



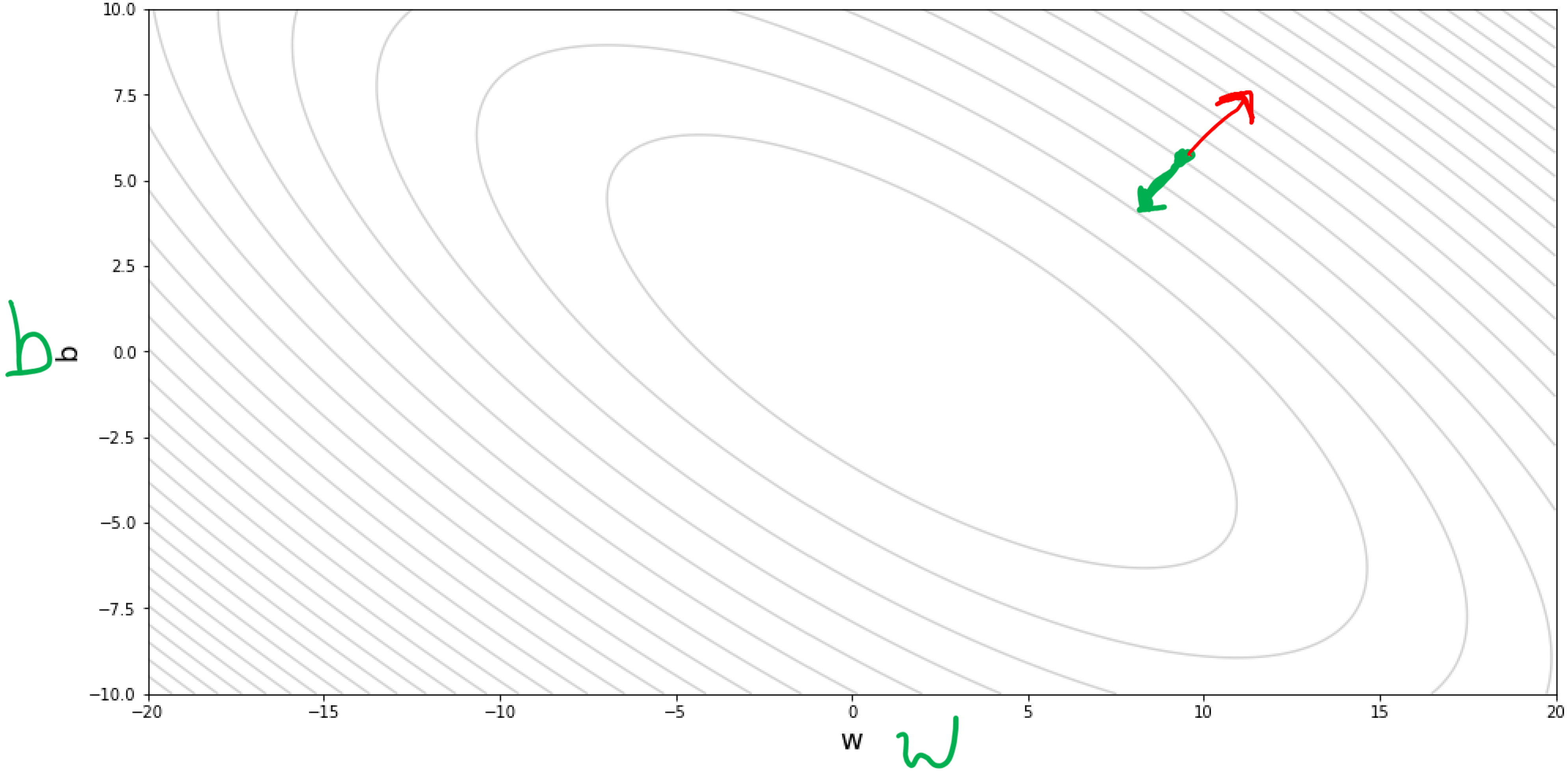
$$J(w,b) = \frac{1}{2}\left[\left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2 + \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2\right]$$

$$J(w,b) = \frac{1}{2}\left[\ J^{(1)}(w,b) \ + \ J^{(2)}(w,b)\ \right]$$

$$J^{(1)}(w,b) = \left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2$$

$$J^{(2)}(w,b) = \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2$$

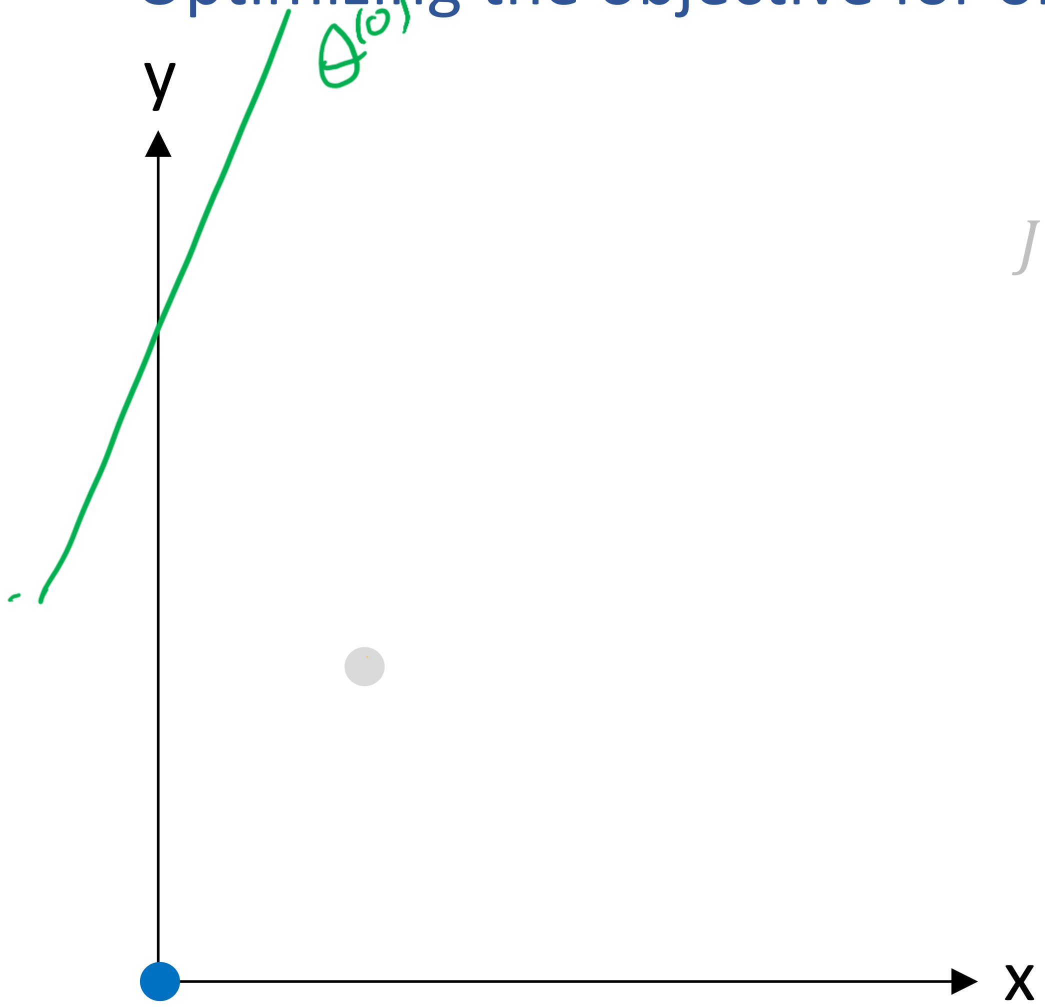
(Batch) Gradient Descent



$$J(w,b) = \frac{1}{2}\left[\ J^{(1)}(w,b)\ +\ J^{(2)}(w,b)\ \right]$$

Stochastic Gradient Descent

Optimizing the objective for one point at a time

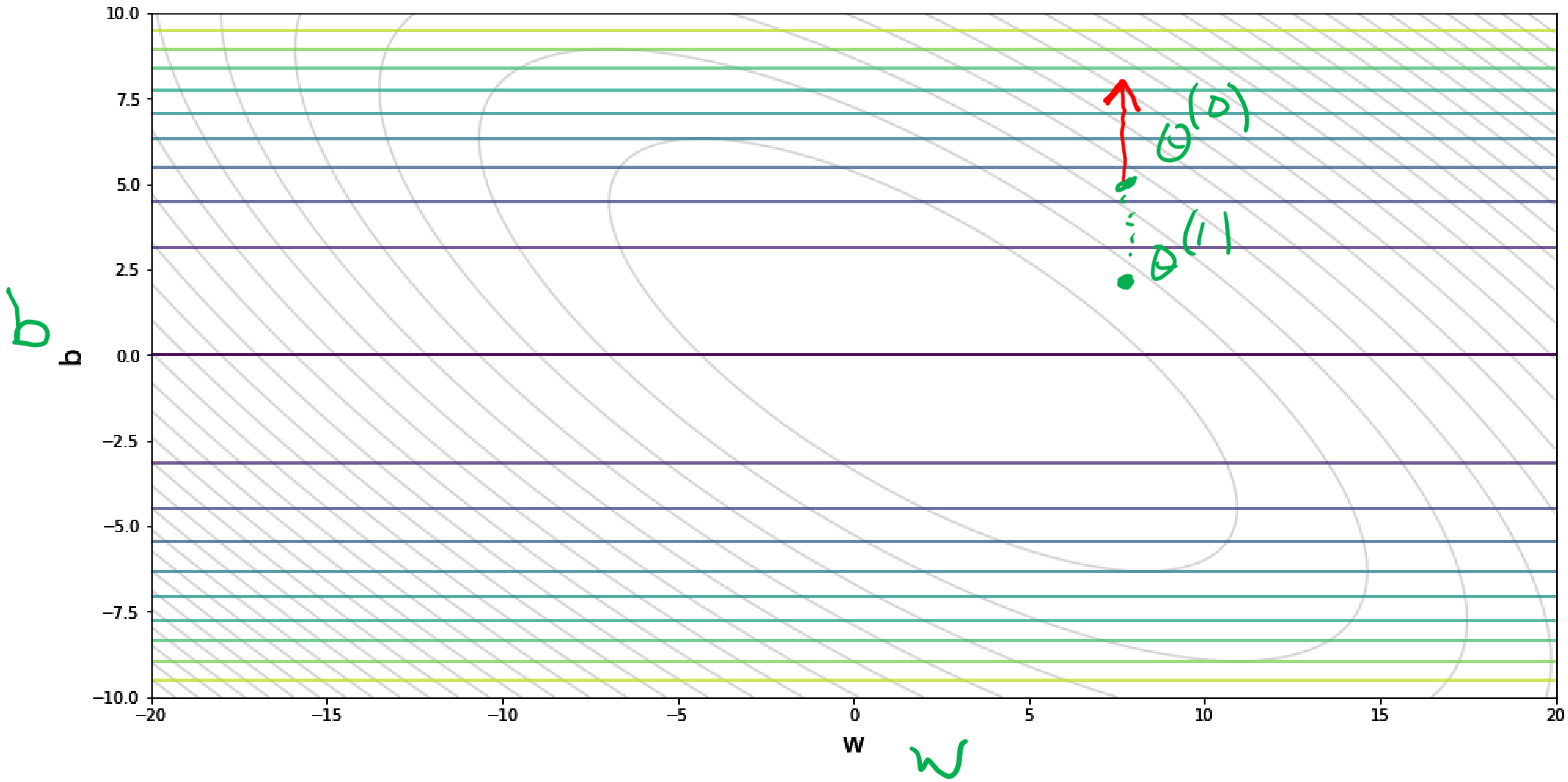


$$J(w,b) = \frac{1}{2}\left[\; J^{(1)}(w,b) \;+\; J^{(2)}(w,b) \;\right]$$

$$J^{(1)}(w,b) = \left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2$$

$$J^{(2)}(w,b) = \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2$$

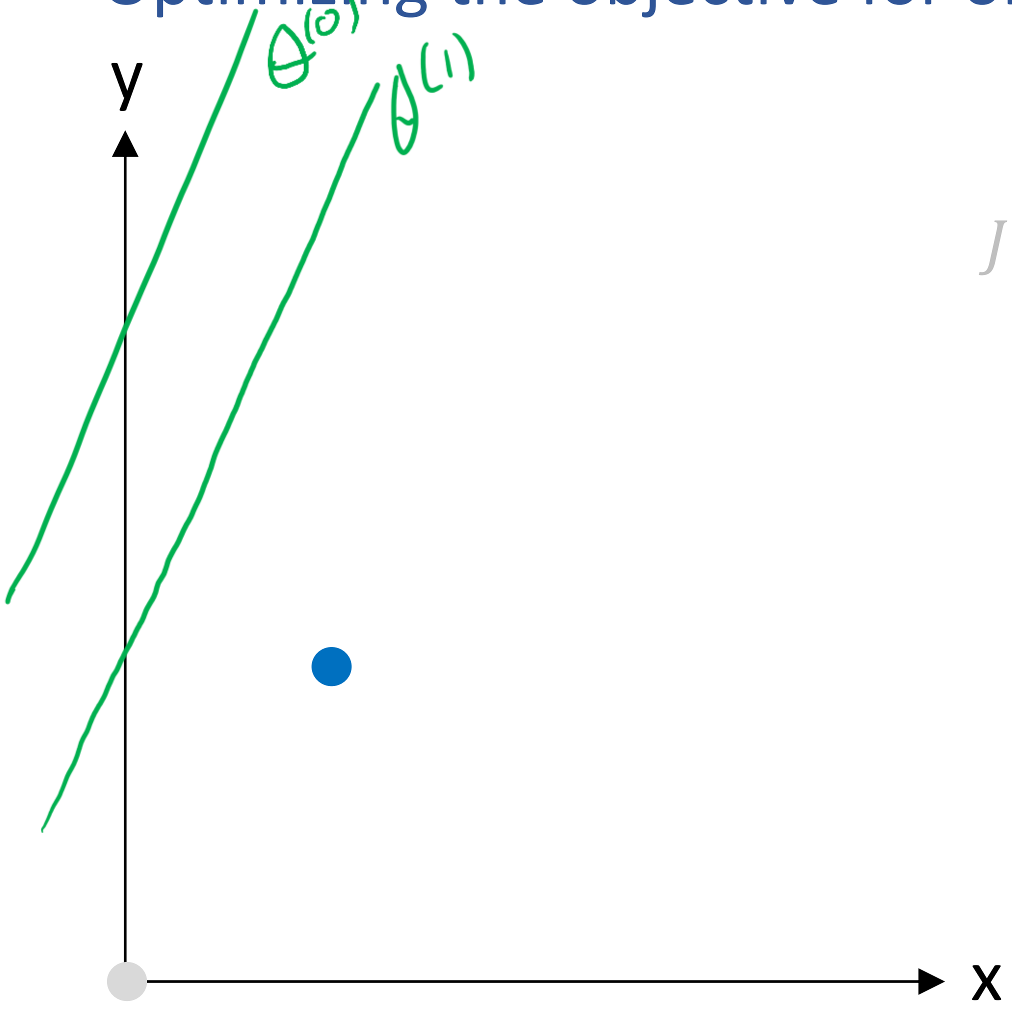
Stochastic Gradient Descent



$$J^{(1)}(w,b) = \left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2$$

Stochastic Gradient Descent

Optimizing the objective for one point at a time

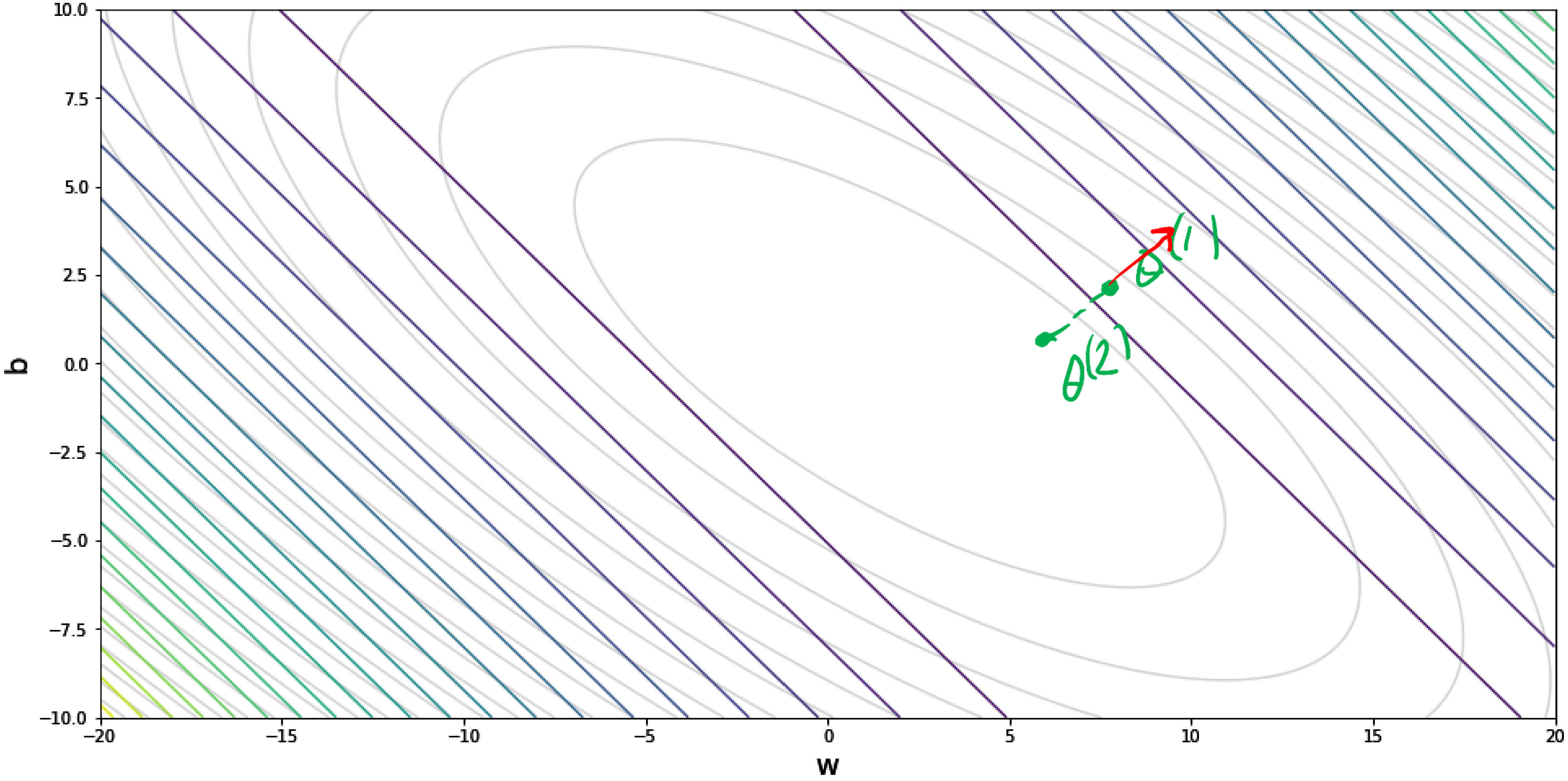


$$J(w,b) = \frac{1}{2}\left[\ J^{(1)}(w,b)\ +\ J^{(2)}(w,b)\ \right]$$

$$J^{(1)}(w,b) = \left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2$$

$$J^{(2)}(w,b) = \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2$$

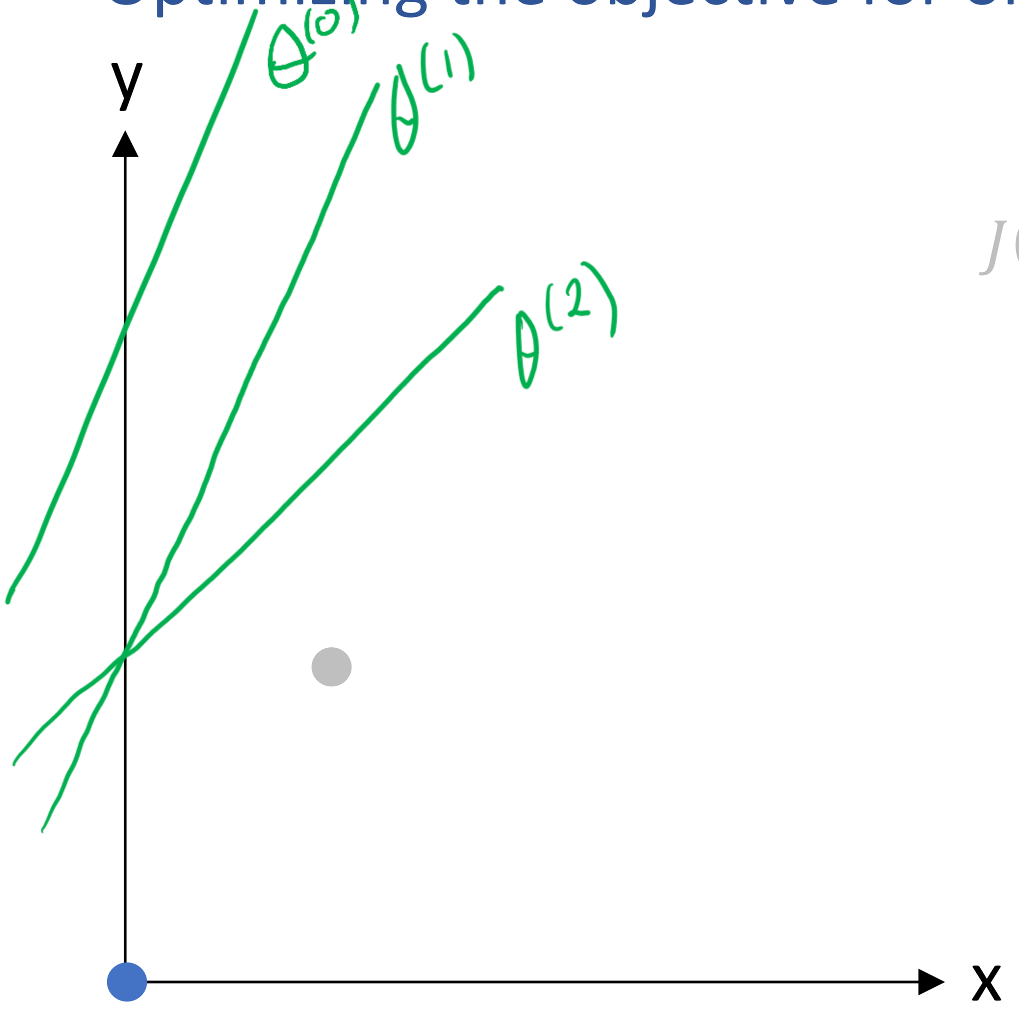
Stochastic Gradient Descent



$$J^{(2)}(w,b) = \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2$$

Stochastic Gradient Descent

Optimizing the objective for one point at a time

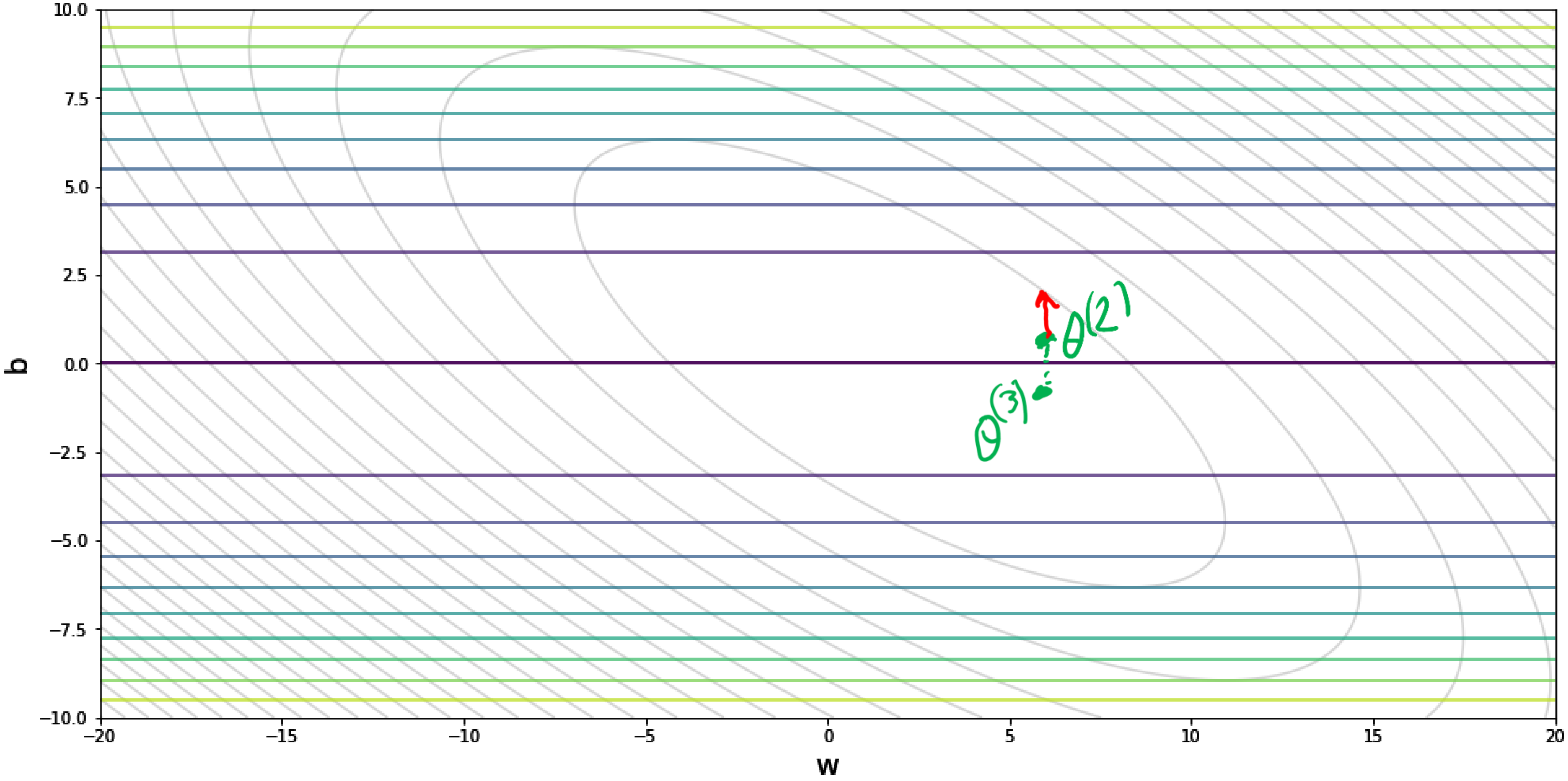


$$J(w,b) = \frac{1}{2}\left[\ J^{(1)}(w,b)\ +\ J^{(2)}(w,b)\ \right]$$

$$J^{(1)}(w,b) = \left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2$$

$$J^{(2)}(w,b) = \left(y^{(2)} - \left(wx^{(2)} + b\right)\right)^2$$

Stochastic Gradient Descent



$$J^{(1)}(w,b) = \left(y^{(1)} - \left(wx^{(1)} + b\right)\right)^2$$