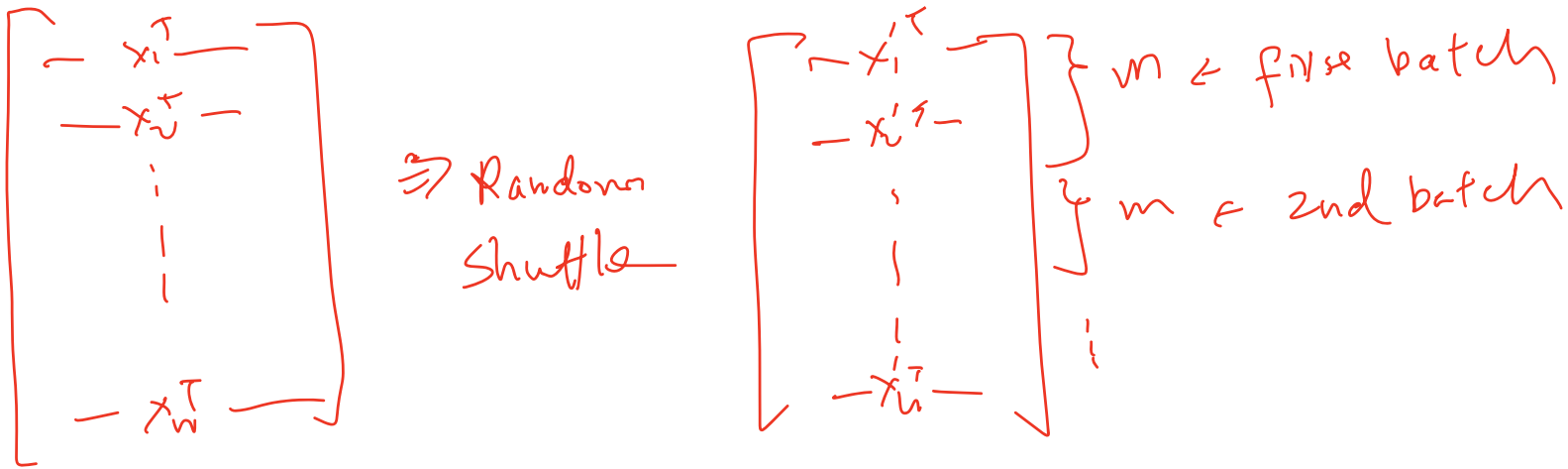
Linear Regression

Recap on Optimization

T/F: for mini-batch SGD, we should always use very large batch

Q: for mini-batch w/ size m, should we sample m points with replacement?



Objective

Learn the first regression algorithm (i.e., predict continuous variable)

Outline for Today

1. Intro on Linear Regression

2. Normal equation for linear Regression

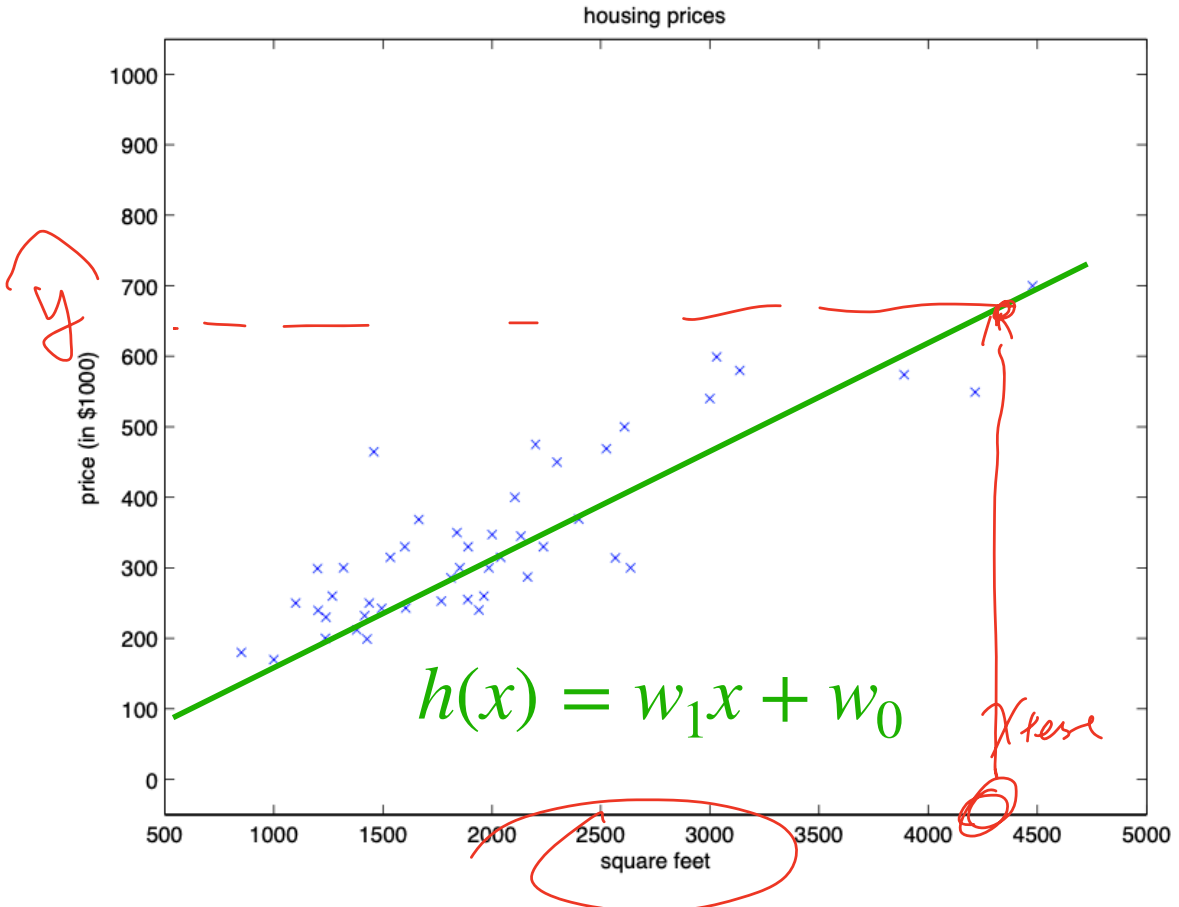
3. Interpretation of Linear Regression using MLE / MAP

Ex: Predicting the house price

Dataset:

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮
x	y

Plot:



(Example from Stanford CS229)

Ex: Predicting the house price (2d case)

Dataset:

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮
$x[1]$	$x[2]$	y

Goal: finding the linear function

$$h(x) = w_1 x[1] + w_2 x[2] + w_0$$

that fits the data well

(Example from Stanford CS229)

Ex: Predicting the house price (2d case)

Dataset:

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮
$x[1]$	$x[2]$	y

As usual, we append 1 to the feature, i.e.,

$$x = \begin{bmatrix} x[1] \\ x[2] \\ 1 \end{bmatrix}$$

So the linear function can be written as:

$$w = \begin{bmatrix} w(1) \\ w(2) \\ w_0 \end{bmatrix} \qquad h(x) = w^\top x$$

Outline for Today

1. Intro on Linear Regression

2. Normal equation for linear Regression

3. Interpretation of Linear Regression using MLE / MAP

Mathematical formulation of linear regression

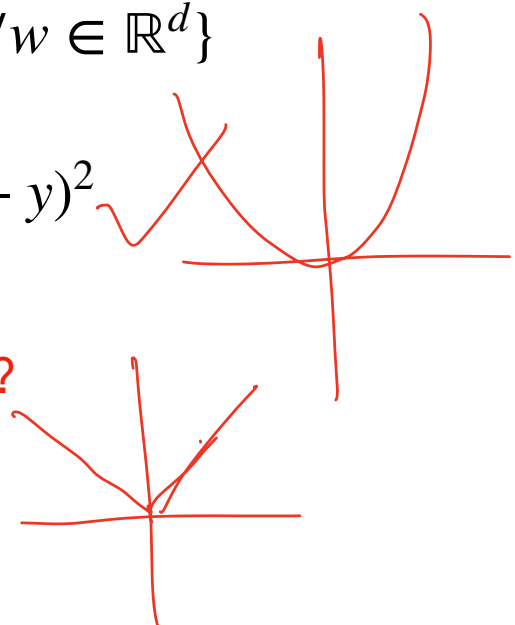
Input: dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Hypothesis: linear function $h(x) = w^\top x$

Hypothesis class: all possible linear functions $\{w^\top x, \forall w \in \mathbb{R}^d\}$

Loss function: squared loss $\ell(w^\top x, y) = (w^\top x - y)^2$

Q: can we use absolute loss, i.e., $|w^\top x - y|$?



Mathematical formulation of linear regression

Formulating the optimization problem:

$$\arg\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Q: how to solve this?

Linear regression solution

$$\arg\min_w \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Let's compute the closed-form solution:

Define $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$

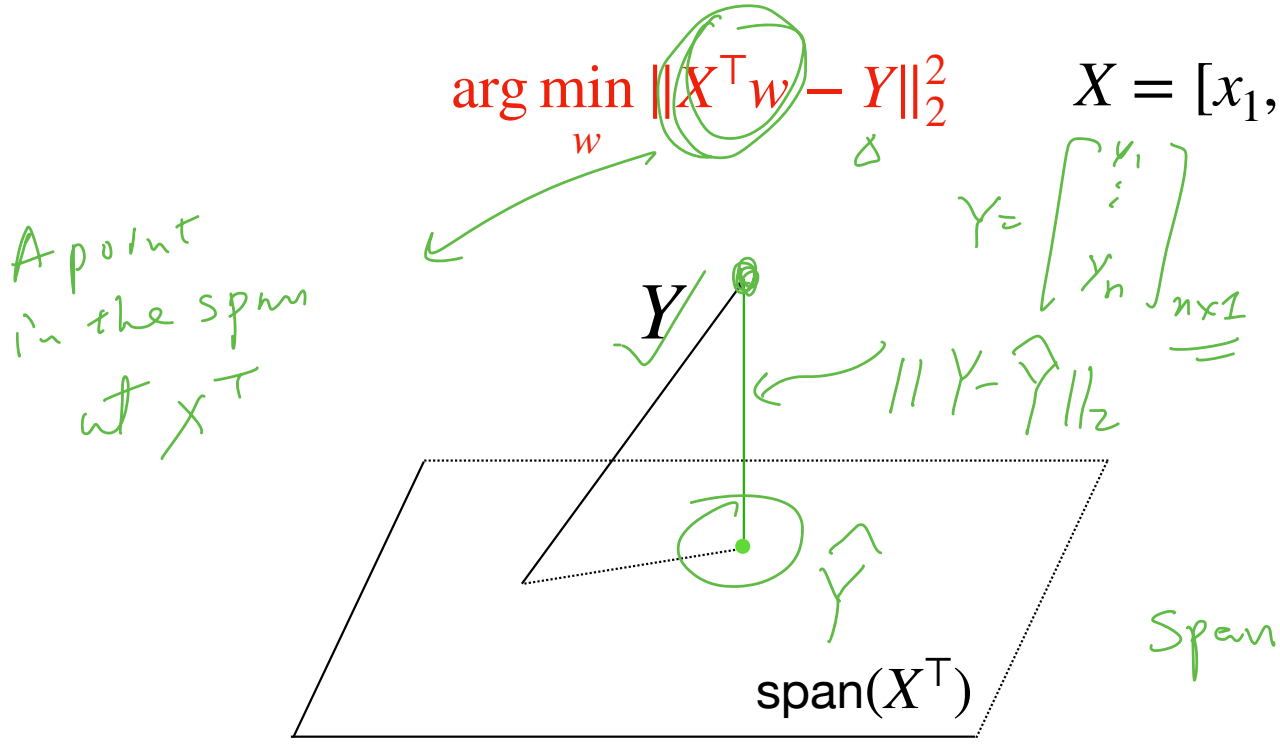
$$\sum_{i=1}^{n} (w^\top x_i - y_i)^2 = \|X^\top w - Y\|_2^2$$

$$\Rightarrow \arg\min_w \|X^\top w - Y\|_2^2$$

Linear regression solution

$$\arg\min_{w} \|X^\top w - Y\|_2^2$$

$$X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$$



column i

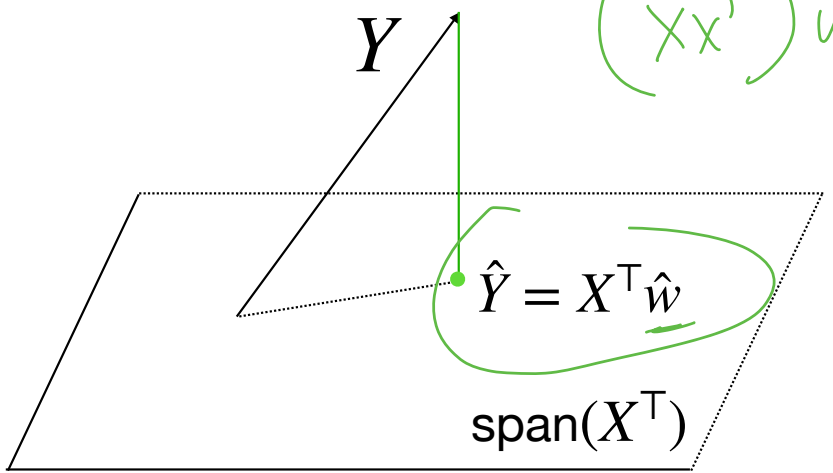
$\text{Span}(A) = \left\{ \sum_i a_i A_i \right\}$

Linear regression solution

$$\arg\min_w \|X^\top w - Y\|_2^2$$

$$\nabla_w \|X^\top w - Y\|_2^2 = XX^\top w - XY$$

if $XX^\top$ is full rank, then $\hat{w} = (XX^\top)^{-1}XY$

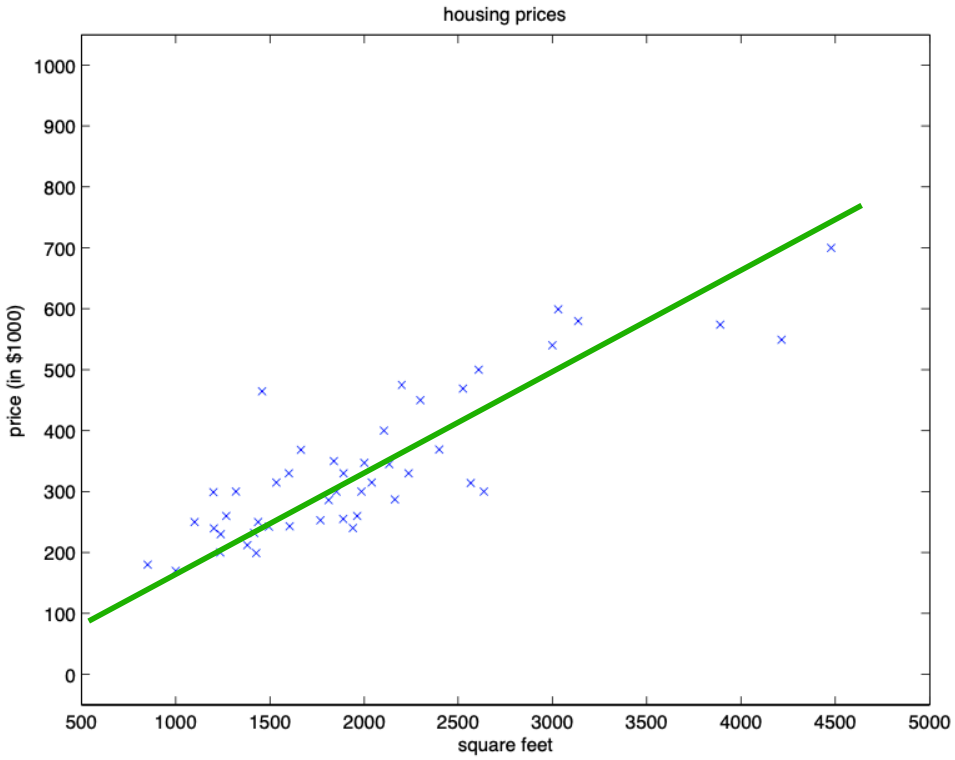


What if $XX^\top$ is not full rank?

(We will talk about regularization soon)

Prediction using linear regression

Once we learned $\hat{w}$, we can use it to make prediction on any new feature x



$$\hat{w} = (XX^T)^{-1}XY$$

Given x_{test}, our prediction is:

$$\hat{y} = x_{test}^\top \hat{w}$$

$$XY = \sum_{i=1}^{n} x_i y_i$$

$$= x_{test}^\top (XX^\top)^{-1} XY$$

$$= \sum_i \left(x_{test}^\top (XX^\top)^{-1} x_i \right) \cdot y_i$$

$\beta_i \in \mathbb{R}$

Outline for Today

1. Intro on Linear Regression

2. Normal equation for linear Regression

3. Interpretation of Linear Regression using MLE / MAP

Derive Linear regression via Maximum Likelihood Estimation

Assume $P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2/\sigma^2\right)$, i.e., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Let's maximize the log-likelihood of the data, i.e.,

$$\arg\max_w \sum_{i=1}^{n} \ln P(y_i \mid x_i; w)$$

$$= \arg\max_w \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(w^\top x_i - y_i)^2 - \ln(Z)$$

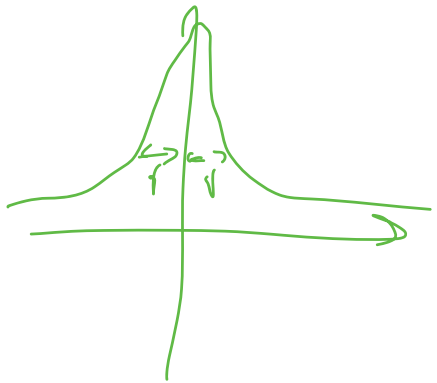
$$= \arg\min_w \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Derive Linear regression via MAP

Assume $P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2/\sigma^2\right)$, i.e., $y = w^\top x + \epsilon$, $\epsilon \sim \mathcal{N}(0,\sigma^2)$

To use MAP, we need to define a prior over w, we use Gaussian as well here:

$w \sim \mathcal{N}(0, r^2 I)$ prior



$p(w) = \frac{1}{Z} \exp\left(-\frac{1}{2} w^T w / r^2\right) \Rightarrow \mathcal{N}(0, r^2 I)$

Derive Linear regression via MAP

$$w \sim \mathcal{N}(0, r^2 I) \qquad P(y \mid x; w) = \frac{1}{Z} \exp\left(-\frac{1}{2}(y - x^\top w)^2 / \sigma^2\right)$$

MAP:

$P(w \mid D) \propto P(w) \cdot P(D \mid w)$

Bayes Rule

$$\arg\max_w \ln P(w \mid \mathcal{D})$$

$$= \arg\max_w \ln P(w) + \ln P(\mathcal{D} \mid w)$$

same in MLE

$$= \arg\max_w \frac{-w^\top w}{2r^2} + \sum_{i=1}^{n} \frac{-1}{2\sigma^2}(w^\top x_i - y_i)^2 \quad) \cdot -2\sigma^2$$

$$= \arg\min_w \frac{\sigma^2}{r^2} w^\top w + \sum_{i=1}^{n} (w^\top x_i - y_i)^2 \qquad = \arg\min_w \lambda \|w\|_2^2 + \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

$\lambda = ?$

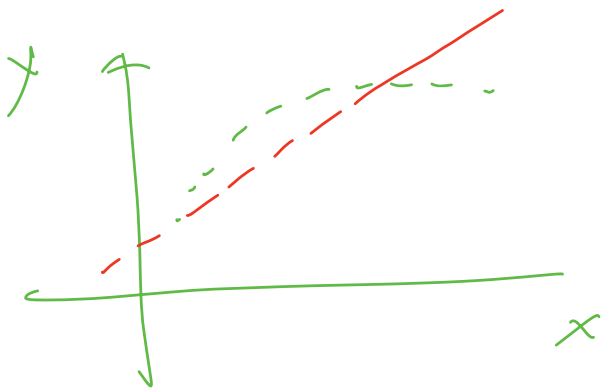
Ridge Linear Regression

$$\arg\min_{w} \lambda \|w\|_2^2 + \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

In this case, we can derive a closed-form solution as well:

$$\hat{w} = (XX^\top + \lambda I)^{-1} XY$$

Note that it works even $XX^\top$ is not full rank



Summary for today

1. Linear regression, Normal equation, and MLE / MAP interpretation

2. Your take-home question: what is the SGD update rule for Linear regression? Is the update rule intuitively explainable?

3. Next Tue: Support Vector Machine!