

# Applications preview

---

# Maximum Likelihood Estimation

Observe  $X_1, X_2, \dots, X_n$  drawn IID from  $f(x; \theta)$  for some “true”  $\theta = \theta_*$

Likelihood function  $L_n(\theta) = \prod_{i=1}^n f(X_i; \theta)$

Log-Likelihood function  $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^n \log(f(X_i; \theta))$

Maximum Likelihood Estimator (MLE)  $\hat{\theta}_{MLE} = \arg \max_{\theta} L_n(\theta)$

Under benign assumptions, as the number of observations  $n \rightarrow \infty$  we have  $\hat{\theta}_{MLE} \rightarrow \theta_*$

Why is it useful to recover the “true” parameters  $\theta_*$  of a probabilistic model?

- **Estimation** of the parameters  $\theta_*$  is the goal
- Help **interpret** or summarize large datasets
- Make **predictions** about future data
- **Generate** new data  $X \sim f(\cdot; \hat{\theta}_{MLE})$

# Estimation

Observe  $X_1, X_2, \dots, X_n$  drawn IID from  $f(x; \theta)$  for some “true”  $\theta = \theta_*$

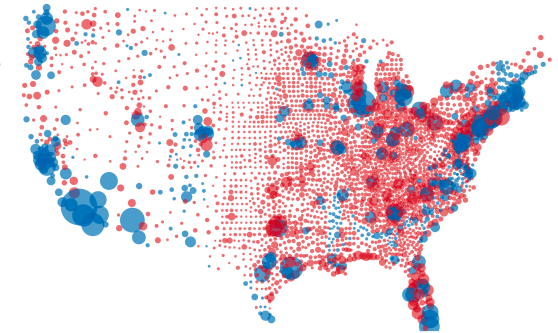
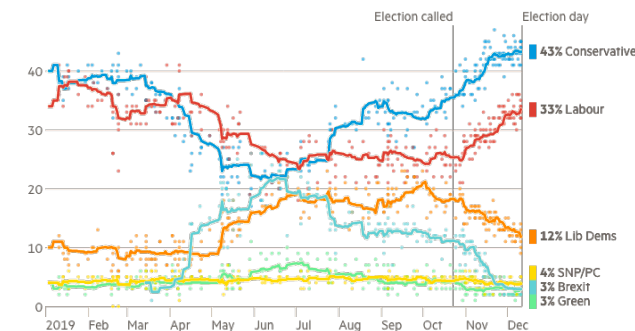
## Opinion polls

How does the greater population feel about an issue?  
Correct for over-sampling?

- $\theta_*$  is “true” average opinion
- $X_1, X_2, \dots$  are sample calls

UK poll tracker

Lines represent weighted averages, points represent polls (%)



## A/B testing

How do we figure out which ad results in more click-through?

- $\theta_*$  are the “true” average rates
- $X_1, X_2, \dots$  are binary “clicks”

Save on prescription drugs - over \$3,637\* a year!

Explore Humana's Medicare plans

Control Treatment

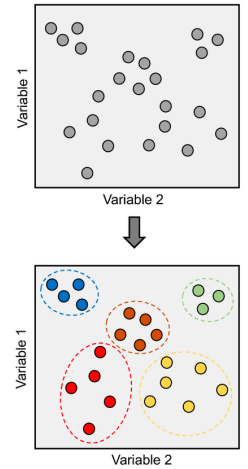
# Interpret

Observe  $X_1, X_2, \dots, X_n$  drawn IID from  $f(x; \theta)$  for some “true”  $\theta = \theta_*$

## Customer segmentation / clustering

Can we identify distinct groups of customers by their behavior?

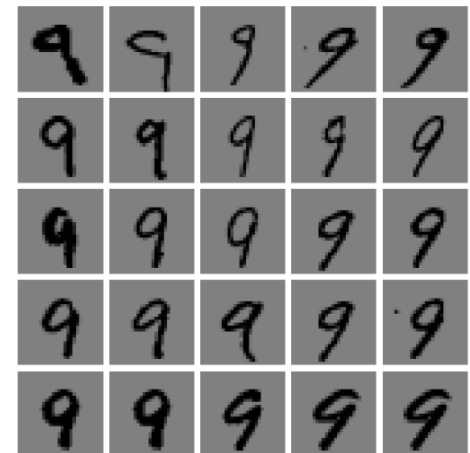
- $\theta_*$  describes “center” of distinct groups
- $X_1, X_2, \dots$  are individual customers



## Data exploration

What are the degrees of freedom of the dataset?

- $\theta_*$  describes the principle directions of variation
- $X_1, X_2, \dots$  are the individual images



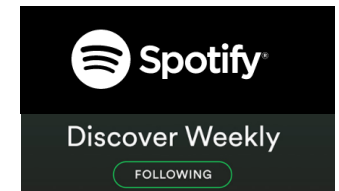
# Predict

Observe  $X_1, X_2, \dots, X_n$  drawn IID from  $f(x; \theta)$  for some “true”  $\theta = \theta_*$

## Content recommendation

Can we predict how much someone will like a movie based on past ratings?

- $\theta_*$  describes user’s preferences
- $X_1, X_2, \dots$  are (movie, rating) pairs



## Object recognition / classification

Identify a flower given just its picture?

- $\theta_*$  describes the characteristics of each kind of flower
- $X_1, X_2, \dots$  are the (image, label) pairs



(a)



(b)



(c)

Figure 1.1: Three types of Iris flowers: Setosa, Versicolor and Virginica. Used with kind permission of Dennis Krumb and SIGNA.

index	sl	sw	pl	pw	label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
...					
50	7.0	3.2	4.7	1.4	Versicolor
...					
149	5.9	3.0	5.1	1.8	Virginica

# Generate

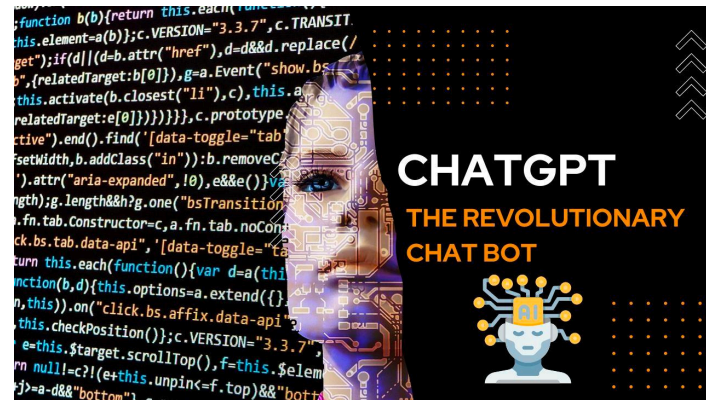
Observe  $X_1, X_2, \dots, X_n$  drawn IID from  $f(x; \theta)$  for some “true”  $\theta = \theta_*$

## Text generation

Can AI generate text that could have been written like a human?

- $\theta_*$  describes language structure
- $X_1, X_2, \dots$  are text snippets found online

“Kaia the dog wasn't a natural pick to go to mars.  
No one could have predicted she would...”



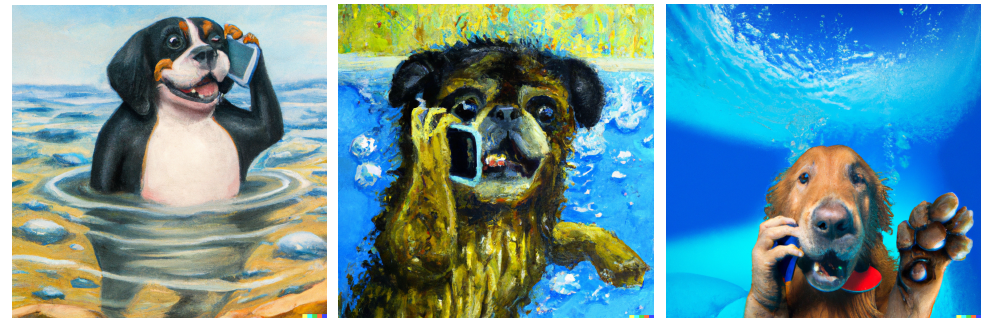
<https://chat.openai.com/chat>

## Image to text generation

Can AI generate an image from a prompt?

- $\theta_*$  describes the coupled structure of images and text
- $X_1, X_2, \dots$  are the (image, caption) pairs found online

“dog talking on cell phone under water, oil painting”



<https://labs.openai.com/>

# Linear Regression

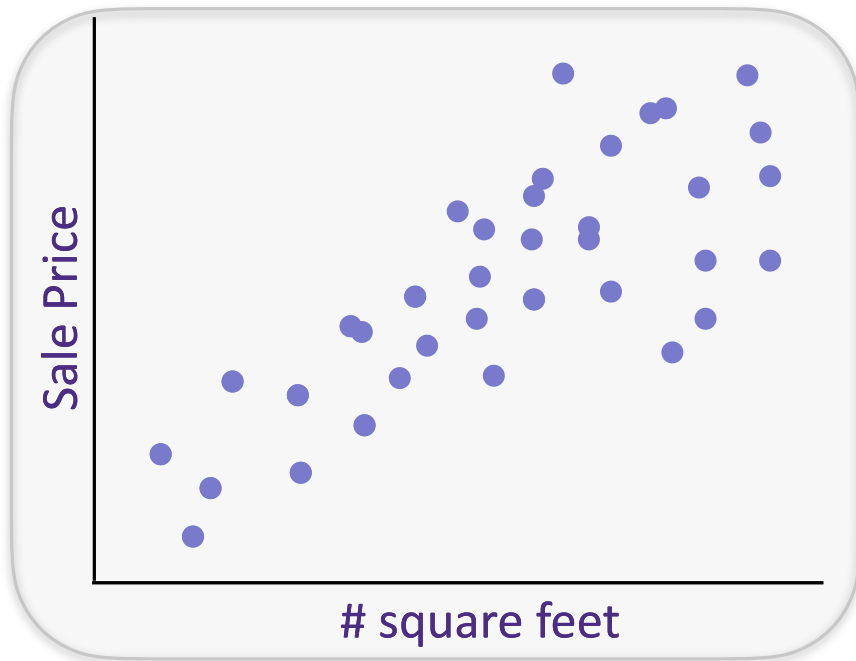
---

# The regression problem, 1-dimensional

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$  House sale price from

$x = \{\# \text{ sq. ft.}\}$



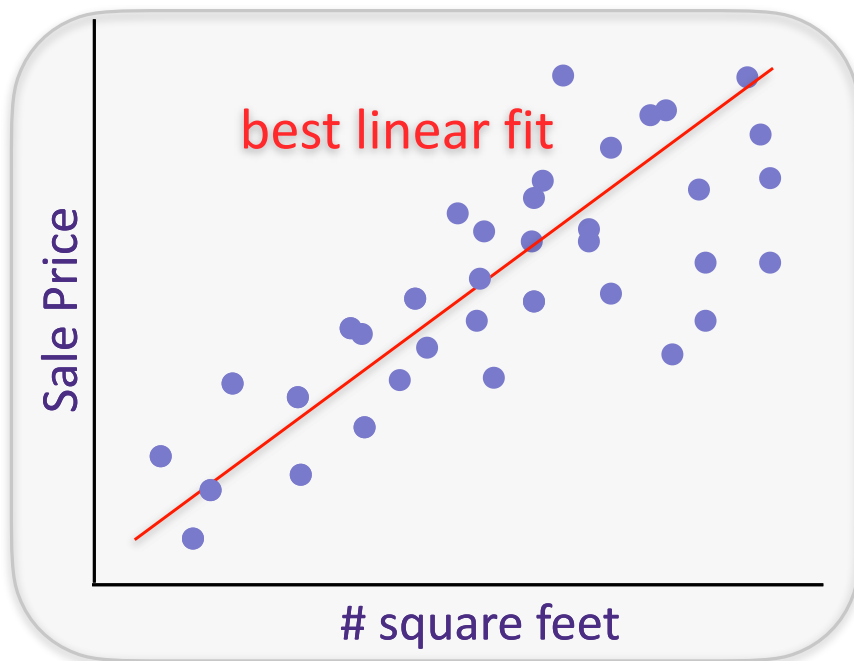
Training Data:  $x_i \in \mathbb{R}$   
 $\{(x_i, y_i)\}_{i=1}^n$   $y_i \in \mathbb{R}$

# Fit a function to our data, 1-d

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$  House sale price from

$x = \{\# \text{ sq. ft.}\}$



Training Data:  $x_i \in \mathbb{R}$   
 $y_i \in \mathbb{R}$   
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis/Model: linear

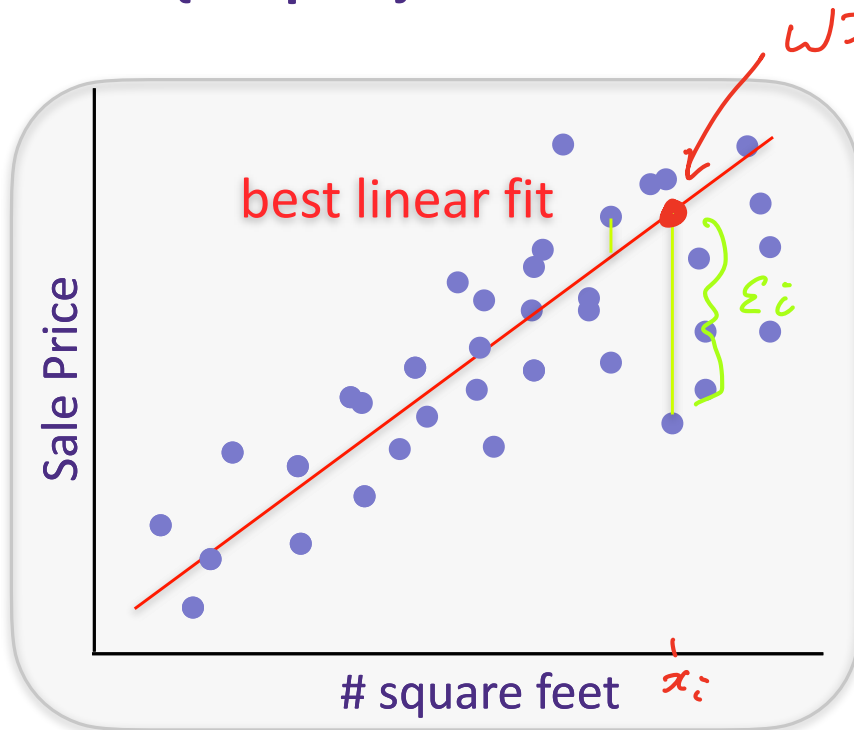
$$y_i = x_i w + \epsilon_i \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

# Fit a function to our data, 1-d

Given past sales data on zillow.com, predict:

$y =$  House sale price from

$x = \{\# \text{ sq. ft.}\}$



Training Data:  $x_i \in \mathbb{R}$   
 $y_i \in \mathbb{R}$   
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis/Model: linear

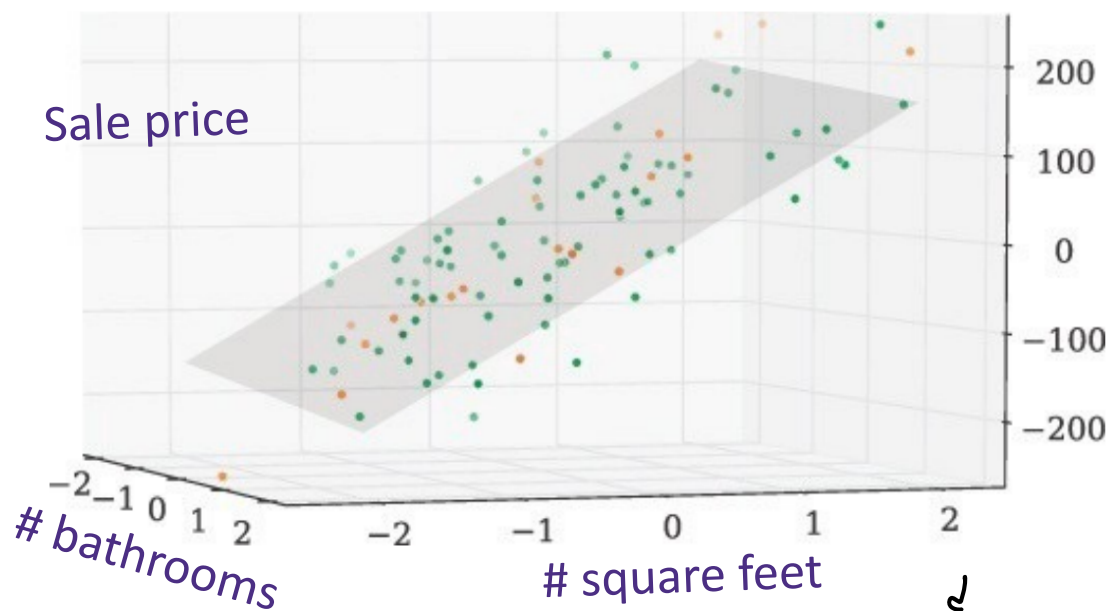
$$y_i = x_i w + \epsilon_i \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

# The regression problem, d-dim

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$  House sale price from

$x = \{\# \text{ sq. ft.}, \text{zip code}, \text{date of sale}, \text{etc.}\}$



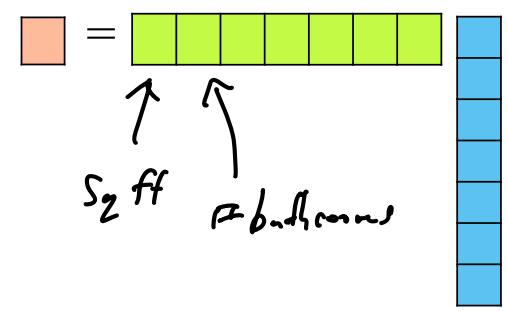
Training Data:  $x_i \in \mathbb{R}^d$   
 $y_i \in \mathbb{R}$   
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis/Model: linear

$$y_i = x_i^T w + \epsilon_i \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$w \in \mathbb{R}^d$

$$x_i^T w = \langle x_i, w \rangle = \sum_{j=1}^d x_i(j) w(j)$$



# The regression problem, d-dim

$$z \sim \mathcal{N}(\mu, \sigma^2)$$

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$  House sale price from

$x = \{\# \text{ sq. ft.}, \text{ zip code}, \text{ date of sale}, \text{ etc.}\}$

$$p(z | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

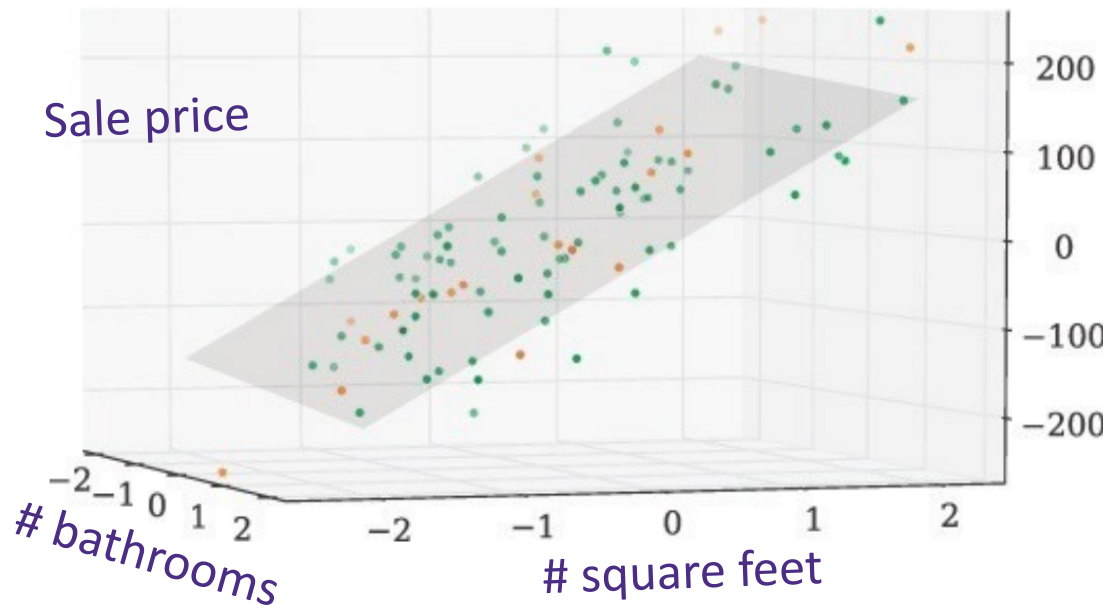
$$\mu = x_i^T w$$

Training Data:  $x_i \in \mathbb{R}^d$   
 $y_i \in \mathbb{R}$   
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis/Model: linear

$$y_i = x_i^T w + \epsilon_i \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i^T w - y_i)^2}{2\sigma^2}}$$

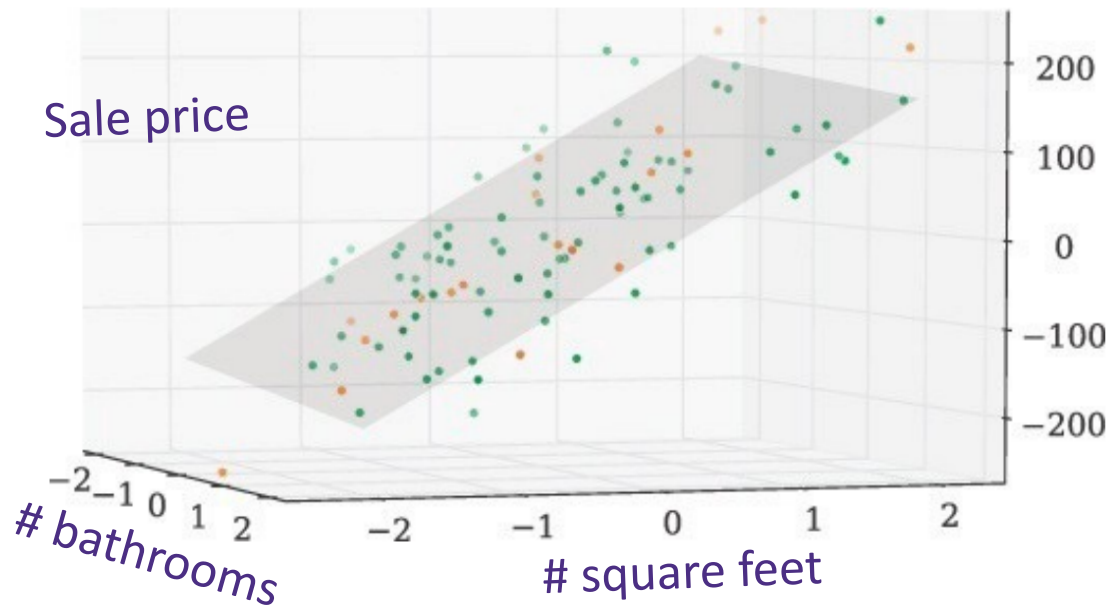


# The regression problem, d-dim

Given past sales data on [zillow.com](https://www.zillow.com), predict:

$y =$  House sale price *from*

$x =$  {# sq. ft., zip code, date of sale, etc.}



Training Data:  $x_i \in \mathbb{R}^d$   
 $y_i \in \mathbb{R}$   
 $\{(x_i, y_i)\}_{i=1}^n$

Hypothesis/Model: linear

$$y_i = x_i^T w + \epsilon_i \quad \epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^T w)^2/2\sigma^2}$$

# Maximizing log-likelihood

**Training Data:**  $x_i \in \mathbb{R}^d$   
 $y_i \in \mathbb{R}$   
 $\{(x_i, y_i)\}_{i=1}^n$

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^\top w)^2/2\sigma^2}$$

**Likelihood:**  $P(\mathcal{D}|w, \sigma) = \prod_{i=1}^n p(y_i|x_i, w, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2}$

$$\log(P(\mathcal{D}|w, \sigma)) = \sum_{i=1}^n \left( -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i^\top w - y_i)^2}{2\sigma^2} \right)$$

# Maximum Likelihood Estimation

Observe  $X_1, X_2, \dots, X_n$  drawn IID from  $f(x; \theta)$  for some “true”  $\theta = \theta_*$

**Likelihood function**  $L_n(\theta) = \prod_{i=1}^n f(X_i; \theta)$

**Log-Likelihood function**  $l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^n \log(f(X_i; \theta))$

**Maximum Likelihood Estimator (MLE)**  $\hat{\theta}_{MLE} = \arg \max_{\theta} L_n(\theta)$

Under benign assumptions, as the number of observations  $n \rightarrow \infty$  we have  $\hat{\theta}_{MLE} \rightarrow \theta_*$

Why is it useful to recover the “true” parameters  $\theta_*$  of a probabilistic model?

- **Estimation** of the parameters  $\theta_*$  is the goal
- Help **interpret** or summarize large datasets
- Make **predictions** about future data
- **Generate** new data  $X \sim f(\cdot; \hat{\theta}_{MLE})$

# Maximizing log-likelihood

$$f: \mathbb{R}^d \rightarrow \mathbb{R}$$

$$\nabla_{\omega} f(\omega) = \begin{bmatrix} \frac{\partial f(\omega)}{\partial \omega_1} \\ \vdots \\ \frac{\partial f(\omega)}{\partial \omega_d} \end{bmatrix} \in \mathbb{R}^d$$

Training Data:  $x_i \in \mathbb{R}^d$   
 $y_i \in \mathbb{R}$   
 $\{(x_i, y_i)\}_{i=1}^n$

$$p(y|x, \omega, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^T \omega)^2 / 2\sigma^2}$$

**Likelihood:**  $P(D|\omega, \sigma) = \prod_{i=1}^n p(y_i|x_i, \omega, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^T \omega)^2 / 2\sigma^2}$

**Maximize (wrt  $\omega$ ):**  $-\log P(D|\omega, \sigma) = \log \left( \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^T \omega)^2 / 2\sigma^2} \right)$

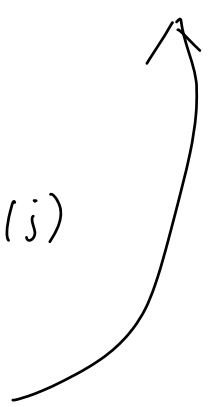
$x_i x_i^T = \begin{bmatrix} | & \text{---} \\ \hline & \end{bmatrix} = \begin{bmatrix} (k,l) \\ \square \end{bmatrix} x_{i(k)} x_{i(l)}$

$$= \sum_{i=1}^n \left( -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i^T \omega - y_i)^2}{2\sigma^2} \right)$$

$$\nabla_{\omega} (*) = \nabla_{\omega} - \sum_{i=1}^n \frac{(x_i^T \omega - y_i)^2}{2\sigma^2} = - \sum_{i=1}^n \frac{2(x_i^T \omega - y_i)}{2\sigma^2} \cdot x_i = 0$$

$$\left( \sum_{i=1}^n x_i x_i^T \right) \omega = \sum_{i=1}^n x_i x_i^T \omega = \left( \sum_{i=1}^n x_i y_i \right) \rightarrow \hat{\omega}_{MLE} = \left( \sum_{i=1}^n x_i x_i^T \right)^{-1} \left( \sum_{i=1}^n x_i y_i \right)$$

$$\left[ \nabla_w \sum_{i=1}^n \frac{(x_i^T w - y_i)^2}{2\sigma^2} \right]_j = \frac{\partial}{\partial w_j} \sum_{i=1}^n \frac{(x_i^T w - y_i)^2}{2\sigma^2} = - \sum_{i=1}^n \frac{(x_i^T w - y_i)}{\sigma^2} \cdot x_i(j)$$

$$\begin{aligned} \frac{\partial}{\partial w(j)} \frac{(x_i^T w - y_i)^2}{2\sigma^2} &= \frac{\partial}{\partial w(j)} \frac{\left( \sum_{k=1}^L x_i(k) w(k) - y_i \right)^2}{2\sigma^2} \\ &= \frac{2 \left( \sum_{k=1}^L x_i(k) w(k) - y_i \right)}{2\sigma^2} \cdot x_i(j) \\ &= \frac{(x_i^T w - y_i)}{\sigma^2} \cdot x_i(j) \end{aligned}$$


# Maximizing log-likelihood

**Training Data:**  $x_i \in \mathbb{R}^d$   
 $\{ (x_i, y_i) \}_{i=1}^n$   $y_i \in \mathbb{R}$

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^\top w)^2/2\sigma^2}$$

**Likelihood:**  $P(\mathcal{D}|w, \sigma) = \prod_{i=1}^n p(y_i|x_i, w, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2}$

**Maximize (wrt  $w$ ):**  $\log P(\mathcal{D}|w, \sigma) = \log \left( \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2} \right)$

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

# Maximizing log-likelihood

---

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

Set derivate=0, solve for w

# Maximizing log-likelihood

---

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

Set derivate=0, solve for w

$$\hat{w}_{MLE} = \left( \sum_{i=1}^n x_i x_i^\top \right)^{-1} \sum_{i=1}^n x_i y_i$$

# The regression problem in matrix notation

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

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n \times 1} \quad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$$

$d$  : # of features

$n$  : # of examples/datapoints

# The regression problem in matrix notation

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

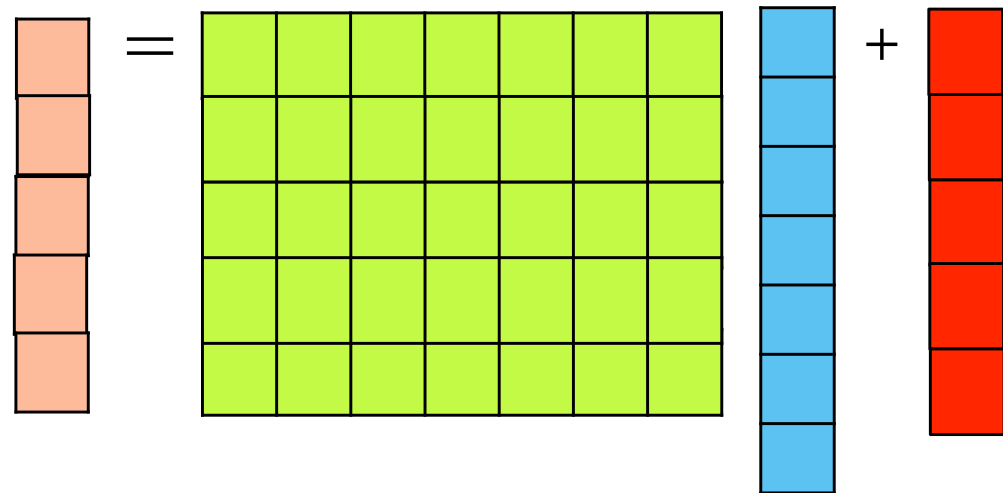
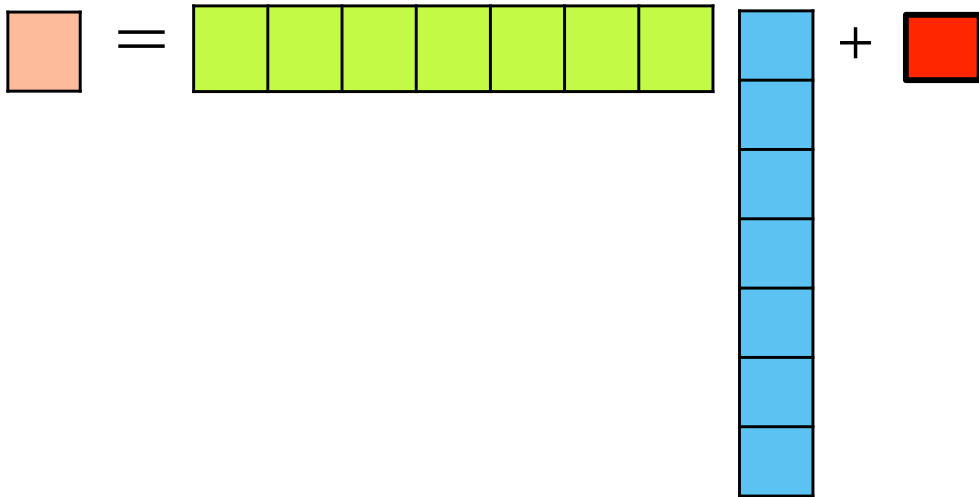
$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}$$

$d$  : # of features

$n$  : # of examples/datapoints

$$y_i = x_i^\top w + \epsilon_i$$

$$\mathbf{y} = \mathbf{X}w + \epsilon$$



# The regression problem in matrix notation

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2 = \sum_{i=1}^n \underbrace{(y_i - x_i^T w)}_{z_i} \underbrace{(y_i - x_i^T w)}_{z_i} = \sum_i z_i^2$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$d$  : # of features

$n$  : # of examples/datapoints

"Least squares"

$$y_i = x_i^T w + \epsilon_i$$

$$\mathbf{y} = \mathbf{X}w + \epsilon$$

$$\begin{aligned} \hat{w}_{LS} &= \arg \min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= \arg \min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w) \end{aligned}$$

$$\ell_2 \text{ norm: } \|z\|_2 = \sqrt{\sum_{i=1}^n z_i^2} = \sqrt{z^T z}$$

# The regression problem in matrix notation

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$d$  : # of features

$n$  : # of examples/datapoints

$$y_i = x_i^T w + \epsilon_i$$

$$\mathbf{y} = \mathbf{X}w + \epsilon$$

$$\begin{aligned} \hat{w}_{LS} &= \arg \min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= \arg \min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w) \end{aligned}$$

$$\begin{aligned} \mathbf{X} &\in \mathbb{R}^{n \times d} \\ \mathbf{X}^T \mathbf{X} &\in \mathbb{R}^{d \times d} \end{aligned}$$

$$\nabla_w \|\mathbf{X}w - \mathbf{y}\|_2^2 = 2\mathbf{X}^T (\mathbf{X}w - \mathbf{y}) = 0 \quad \Rightarrow \quad \mathbf{X}^T \mathbf{X}w = \mathbf{X}^T \mathbf{y}$$

$$\hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# The regression problem in matrix notation

$$\hat{w}_{MLE} = \arg \min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$d$  : # of features

$n$  : # of examples/datapoints

$$y_i = x_i^T w + \epsilon_i$$

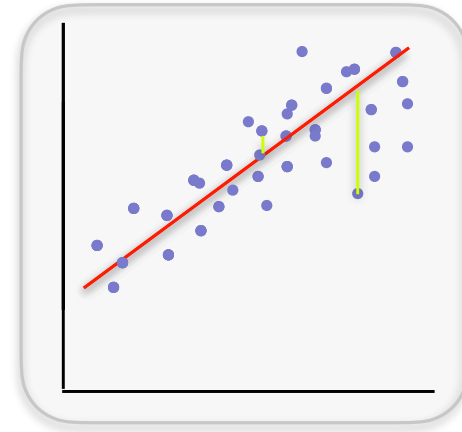
$$\mathbf{y} = \mathbf{X}w + \epsilon$$

$$\begin{aligned} \hat{w}_{LS} &= \arg \min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= \arg \min_w (\mathbf{y} - \mathbf{X}w)^T (\mathbf{y} - \mathbf{X}w) \end{aligned}$$

$$\hat{w}_{LS} = \hat{w}_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

# The regression problem in matrix notation

$$\begin{aligned}\hat{w}_{LS} &= \arg \min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\end{aligned}$$



What about an offset?

$$\begin{aligned}\hat{w}_{LS}, \hat{b}_{LS} &= \arg \min_{w,b} \sum_{i=1}^n (y_i - (x_i^T w + b))^2 \\ &= \arg \min_{w,b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2\end{aligned}$$

# Dealing with an offset

---

$$\hat{w}_{LS}, \hat{b}_{LS} = \arg \min_{w, b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2$$

# Dealing with an offset

---

$$\hat{w}_{LS}, \hat{b}_{LS} = \arg \min_{w, b} \|\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)\|_2^2$$

$$\mathbf{X}^T \mathbf{X} \hat{w}_{LS} + \hat{b}_{LS} \mathbf{X}^T \mathbf{1} = \mathbf{X}^T \mathbf{y}$$

$$\mathbf{1}^T \mathbf{X} \hat{w}_{LS} + \hat{b}_{LS} \mathbf{1}^T \mathbf{1} = \mathbf{1}^T \mathbf{y}$$

If  $\mathbf{X}^T \mathbf{1} = 0$  (i.e., if each feature is mean-zero) then

$$\hat{w}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\hat{b}_{LS} = \frac{1}{n} \sum_{i=1}^n y_i$$

# Make Predictions

---

$$\hat{\mathbf{w}}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\hat{b}_{LS} = \frac{1}{n} \sum_{i=1}^n y_i$$

A new house is about to be listed. What should it sell for?

$$\hat{y}_{\text{new}} = x_{\text{new}}^T \hat{\mathbf{w}}_{LS} + \hat{b}_{LS}$$

# Process

---

Decide on a **model** for the likelihood function  $f(x; \theta)$

Find the function which fits the data best

**Choose a loss function- least squares**

**Pick the function which minimizes loss on data**

Use function to make prediction on new examples