

Math Basics

(Largely adapted from Stanford CS229 Slides)

Junxian He

Feb 3, 2026

Attendance Quiz APP Download

HKUST iLearn

The Hong Kong University of Science and Technology

10K+
Downloads

E
Everyone ⓘ

Install

Share Add to wishlist



A circular icon with a colorful geometric pattern of dots and shapes in shades of purple, pink, and yellow.	Canvas This will open the 'Canvas Student' app which provides an easy access to the online content of your courses at HKUST - watch videos, post to discussions, submit quizzes, etc.
A dark purple square icon with the letters 'SFQ' in white and a yellow checkmark icon below it.	SFQ Allows you to complete the Student Feedback Questionnaire for all your courses at HKUST on the move.
A red square icon with the letters 'iPRS' in white and a yellow signal-like icon above it.	iPRS Enables you to quickly respond to questions or polls created by your instructor in class.

Linear Independence

A set of vectors $\{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^m$ is said to be **(linearly) dependent** if one vector belonging to the set can be represented as a linear combination of the remaining vectors; that is, if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \dots, \alpha_{n-1} \in \mathbb{R}$; otherwise, the vectors are **(linearly) independent**.

Matrix rank

Linear Independence

Example:

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

are linearly dependent because $x_3 = -2x_1 + x_2$.



Rank of a Matrix

- The **column rank** of a matrix $A \in \mathbb{R}^{m \times n}$ is the largest number of columns of A that constitute a linearly independent set.
- The **row rank** is the largest number of rows of A that constitute a linearly independent set.
- For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A (prove it yourself!), and so both quantities are referred to collectively as the **rank** of A , denoted as $\text{rank}(A)$.

$$\min(m, n)$$

Properties of Rank

Properties of Rank

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be *full rank*.

Properties of Rank

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be **full rank**.

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$. *transpose*

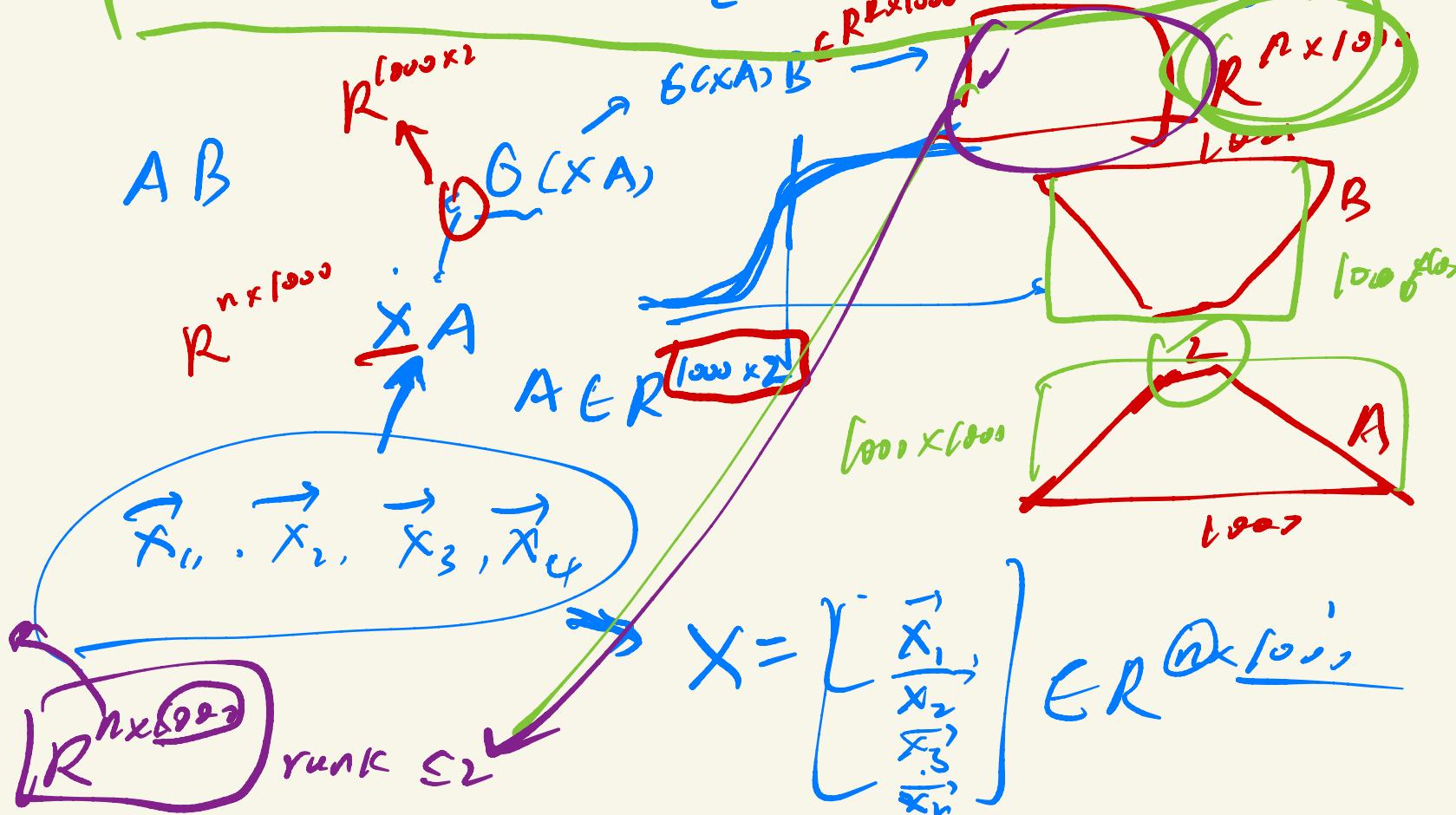
Properties of Rank

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be **full rank**.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{p \times n}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.

Properties of Rank

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be **full rank**.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{p \times n}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.
- For $A, B \in \mathbb{R}^{m \times n}$, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

$$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$$



less

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

$$\text{rank}(C) \leq 2$$

r

$$\text{vectors} = a_1 V_1 + a_2 V_2$$

The Inverse of a Square Matrix

The Inverse of a Square Matrix

- The *inverse* of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted $\overbrace{A^{-1}}$, and is the unique matrix such that

$$\overbrace{A^{-1}A = I = AA^{-1}}.$$

The Inverse of a Square Matrix

- The *inverse* of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1} , and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

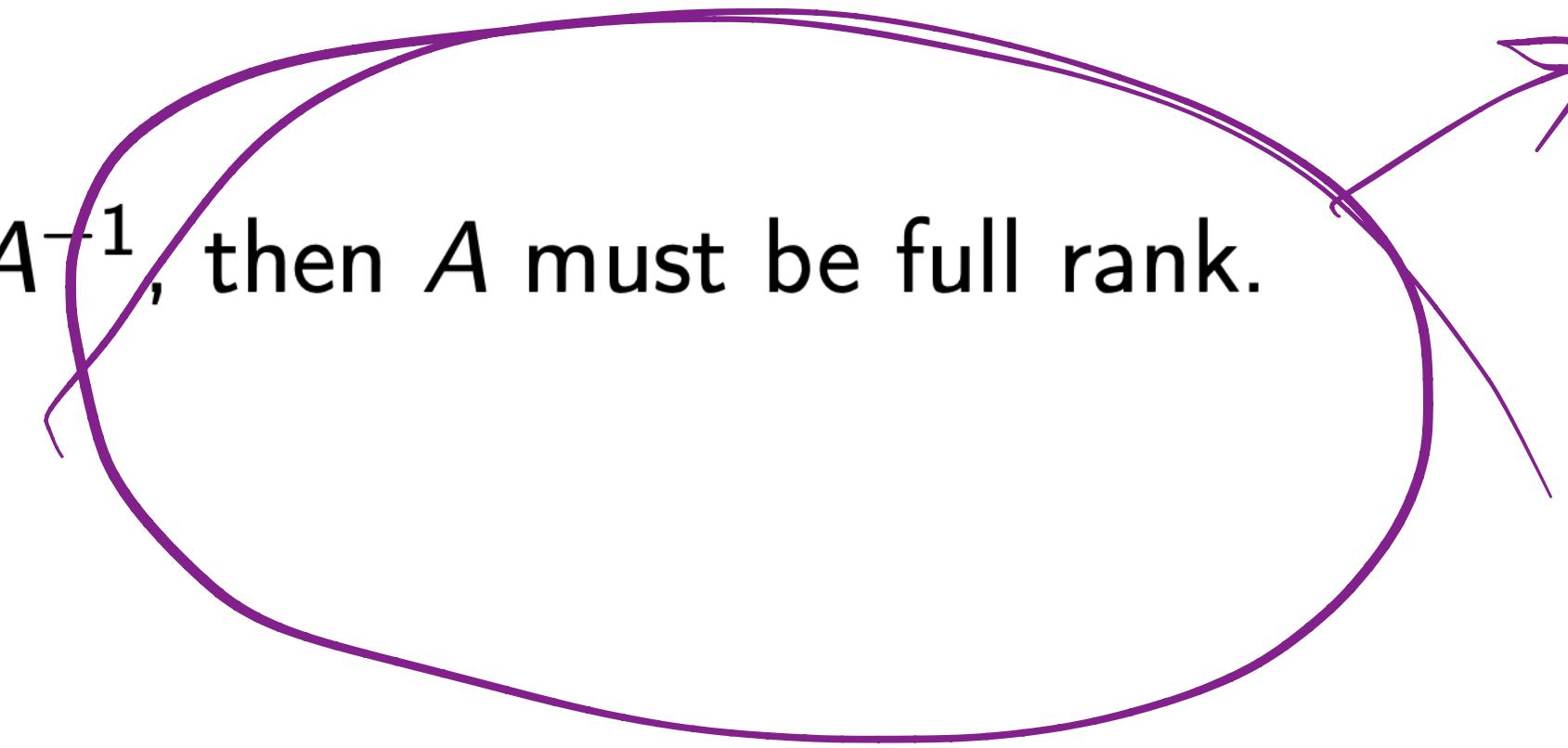
- We say that A is *invertible* or *non-singular* if A^{-1} exists and *non-invertible* or *singular* otherwise.

The Inverse of a Square Matrix

- The *inverse* of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1} , and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

- We say that A is *invertible* or *non-singular* if A^{-1} exists and *non-invertible* or *singular* otherwise.
- In order for a square matrix A to have an inverse A^{-1} , then A must be full rank.



The Inverse of a Square Matrix

- The *inverse* of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1} , and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

- We say that A is *invertible* or *non-singular* if A^{-1} exists and *non-invertible* or *singular* otherwise.

- In order for a square matrix A to have an inverse A^{-1} , then A must be full rank.

- Properties (Assuming $A, B \in \mathbb{R}^{n \times n}$ are non-singular):

- ▶ $(A^{-1})^{-1} = A$
- ▶ $(AB)^{-1} = B^{-1}A^{-1}$
- ▶ $(A^{-1})^T = (A^T)^{-1}$. For this reason this matrix is often denoted A^{-T} .

Orthogonal Matrices

Orthogonal Matrices

- Two vectors $x, y \in \mathbb{R}^n$ are **orthogonal** if $x^T y = 0$.
- A vector $x \in \mathbb{R}^n$ is **normalized** if $\|x\|_2 = 1$.
- A square matrix $U \in \mathbb{R}^{n \times n}$ is **orthogonal** if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being **orthonormal**).

magnitude



inner product

$$x^T y$$

outer product

$$\text{outer}(x, y)$$

inner $\rightarrow R$

outer $\rightarrow R^{n \times m}$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\left[\begin{array}{c} x_1 \\ \vdots \\ x_m \end{array} \right]$$

$n \times m$

$$\begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_m \\ x_2 y_1 & \cdots & \cdots & \end{bmatrix}$$

Orthogonal Matrices

- Two vectors $x, y \in \mathbb{R}^n$ are *orthogonal* if $x^T y = 0$.
- A vector $x \in \mathbb{R}^n$ is *normalized* if $\|x\|_2 = 1$.
- A square matrix $U \in \mathbb{R}^{n \times n}$ is *orthogonal* if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being *orthonormal*).
- **Properties:**
 - ▶ The inverse of an orthogonal matrix is its transpose.

$$U^T U = I = UU^T.$$

Orthogonal Matrices

- Two vectors $x, y \in \mathbb{R}^n$ are *orthogonal* if $x^T y = 0$.
- A vector $x \in \mathbb{R}^n$ is *normalized* if $\|x\|_2 = 1$.
- A square matrix $U \in \mathbb{R}^{n \times n}$ is *orthogonal* if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being *orthonormal*).
- Properties:
 - ▶ The inverse of an orthogonal matrix is its transpose.

$$U^T U = I = UU^T.$$

- ▶ Operating on a vector with an orthogonal matrix will not change its Euclidean norm, i.e.,

$$\|Ux\|_2 = \|x\|_2$$

for any $x \in \mathbb{R}^n$, $U \in \mathbb{R}^{n \times n}$ orthogonal.

Span and Projection

Span and Projection

- The *span* of a set of vectors $\{x_1, x_2, \dots, x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \dots, x_n\}$. That is,

$$\text{span}(\{x_1, \dots, x_n\}) = \left\{ v : v = \sum_{i=1}^n \alpha_i x_i, \quad \alpha_i \in \mathbb{R} \right\}.$$

Linear Combination

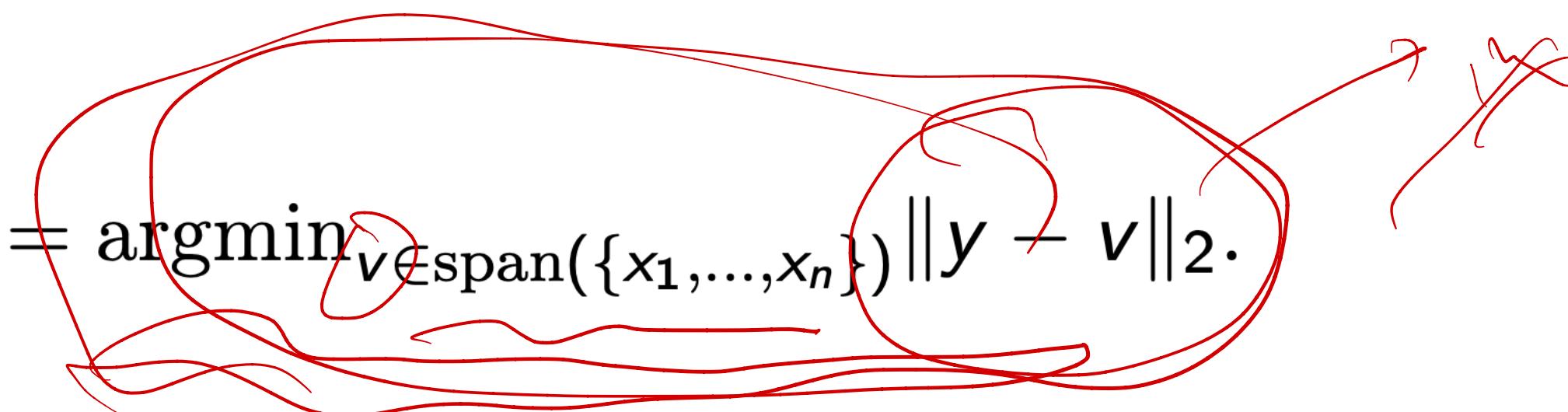
Span and Projection

- The *span* of a set of vectors $\{x_1, x_2, \dots, x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \dots, x_n\}$. That is,

$$\text{span}(\{x_1, \dots, x_n\}) = \left\{ v : v = \sum_{i=1}^n \alpha_i x_i, \quad \alpha_i \in \mathbb{R} \right\}.$$

- The *projection* of a vector $y \in \mathbb{R}^m$ onto the span of $\{x_1, \dots, x_n\}$ is the vector $v \in \text{span}(\{x_1, \dots, x_n\})$, such that v is as close as possible to y , as measured by the Euclidean norm $\|v - y\|_2$.

$$\text{Proj}(y; \{x_1, \dots, x_n\}) = \underset{v \in \text{span}(\{x_1, \dots, x_n\})}{\operatorname{argmin}} \|y - v\|_2.$$



Null Space

The **nullspace** of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(A)$ is the set of all vectors that equal 0 when multiplied by A , i.e.,

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

Determinant

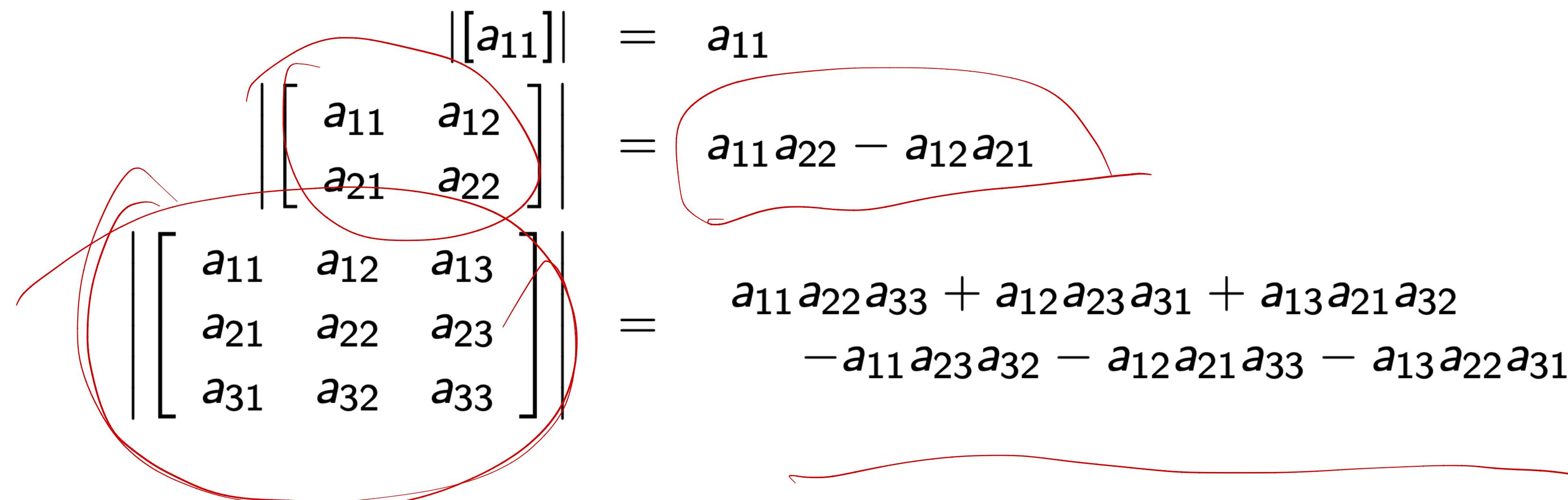
Let $A \in \mathbb{R}^{n \times n}$, $A_{\setminus i, \setminus j} \in \mathbb{R}^{(n-1) \times (n-1)}$ be the *matrix* that results from deleting the i th row and j th column from A .

The general (recursive) formula for the determinant is

$$\begin{aligned}|A| &= \sum_{i=1}^n (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}| \quad (\text{for any } j \in 1, \dots, n) \\ &= \sum_{j=1}^n (-1)^{i+j} a_{ij} |A_{\setminus i, \setminus j}| \quad (\text{for any } i \in 1, \dots, n)\end{aligned}$$

Determinant: Example

However, the equations for determinants of matrices up to size 3×3 are fairly common, and it is good to know them:

$$\begin{aligned} \left| \begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array} \right| &= a_{11} \\ \left| \begin{array}{ccc} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{array} \right| &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\quad - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{aligned}$$


The Determinant

The Determinant

The *determinant* of a square matrix $A \in \mathbb{R}^{n \times n}$, is a function $\det : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$, and is denoted $|A|$ or $\det A$.

The Determinant

The *determinant* of a square matrix $A \in \mathbb{R}^{n \times n}$, is a function $\det : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$, and is denoted $|A|$ or $\det A$.

Given a matrix

$$\begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_n^T & - \end{bmatrix},$$

consider the set of points $S \subset \mathbb{R}^n$ as follows:

$$S = \{v \in \mathbb{R}^n : v = \sum_{i=1}^n \alpha_i a_i \text{ where } 0 \leq \alpha_i \leq 1, i = 1, \dots, n\}.$$

The Determinant

The *determinant* of a square matrix $A \in \mathbb{R}^{n \times n}$, is a function $\det : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$, and is denoted $|A|$ or $\det A$.

Given a matrix

$$\begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_n^T & - \end{bmatrix},$$

consider the set of points $S \subset \mathbb{R}^n$ as follows:

$$S = \{v \in \mathbb{R}^n : v = \sum_{i=1}^n \alpha_i a_i \text{ where } 0 \leq \alpha_i \leq 1, i = 1, \dots, n\}.$$

The absolute value of the determinant of A is a measure of the “volume” of the set S .

The Determinant

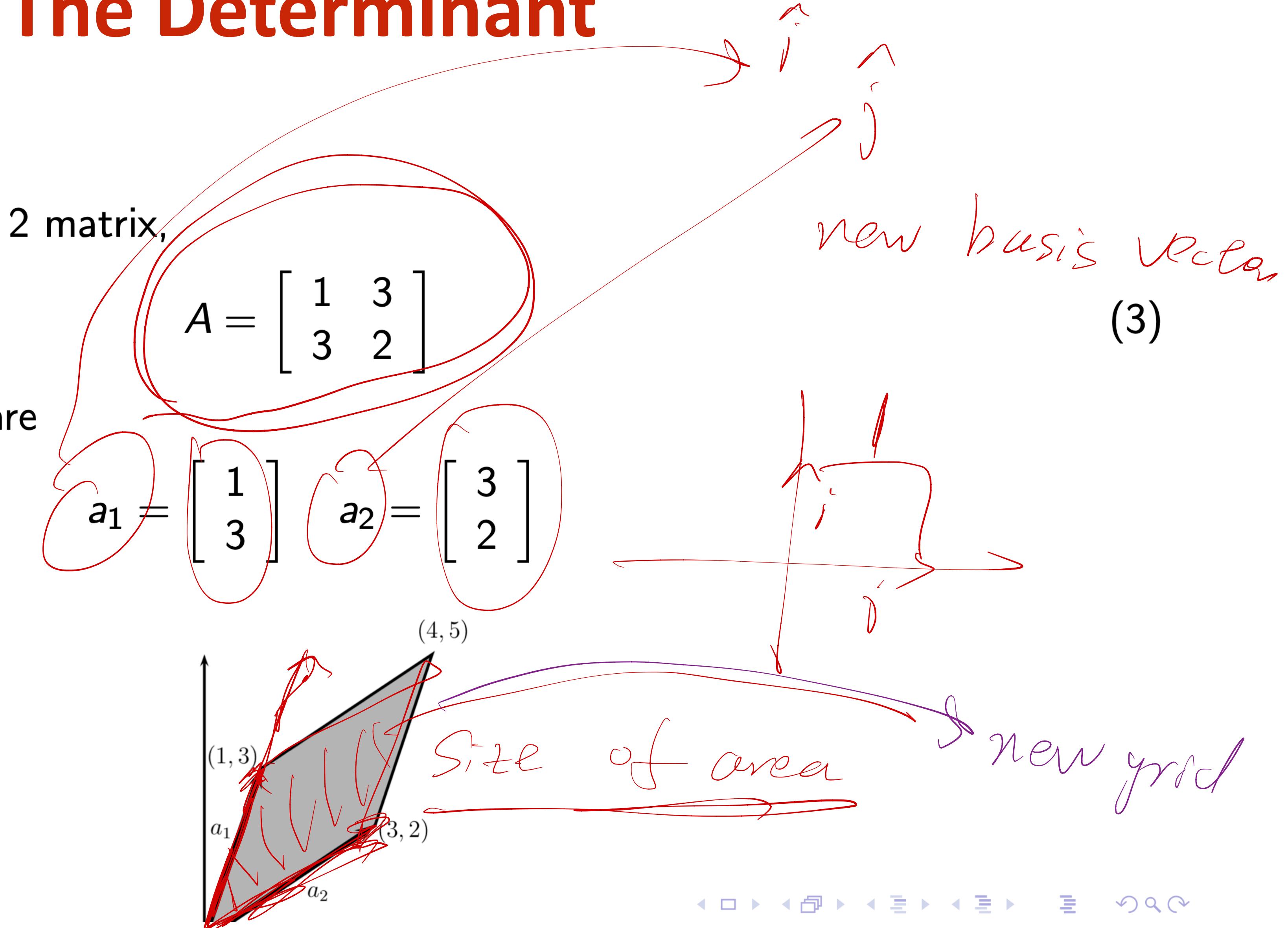
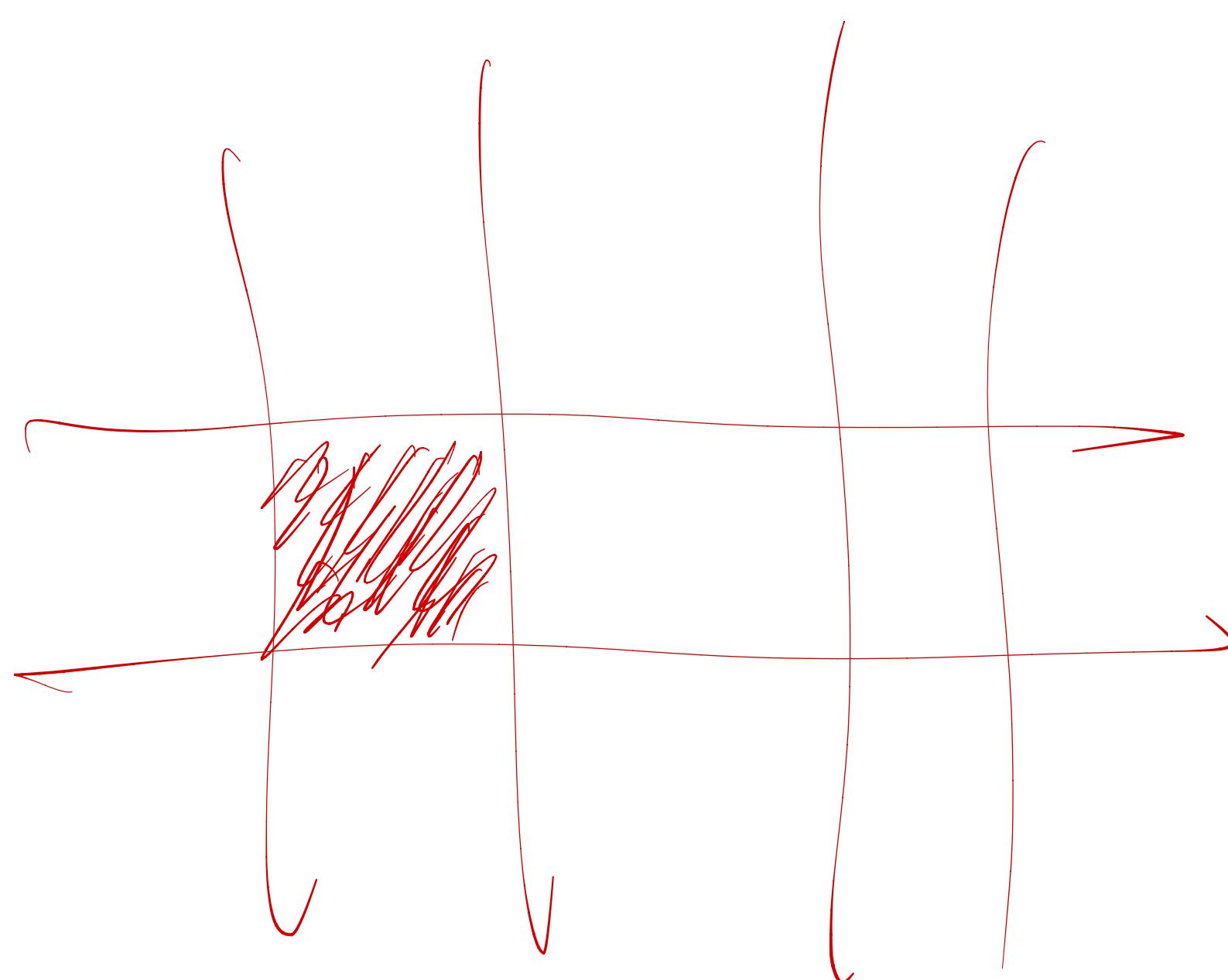
For example, consider the 2×2 matrix,

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}$$

Here, the rows of the matrix are

$$a_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$a_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$



The Determinant: Properties

The Determinant: Properties

Algebraically, the determinant satisfies the following three properties:

1. The determinant of the identity is 1, $|I| = 1$. (Geometrically, the volume of a unit hypercube is 1).

The Determinant: Properties

Algebraically, the determinant satisfies the following three properties:

1. The determinant of the identity is 1, $|I| = 1$. (Geometrically, the volume of a unit hypercube is 1).
2. Given a matrix $A \in \mathbb{R}^{n \times n}$, if we multiply a single row in A by a scalar $t \in \mathbb{R}$, then the determinant of the new matrix is $t|A|$, (Geometrically, multiplying one of the sides of the set S by a factor t causes the volume to increase by a factor t .)

The Determinant: Properties

Algebraically, the determinant satisfies the following three properties:

1. The determinant of the identity is 1, $|I| = 1$. (Geometrically, the volume of a unit hypercube is 1).
2. Given a matrix $A \in \mathbb{R}^{n \times n}$, if we multiply a single row in A by a scalar $t \in \mathbb{R}$, then the determinant of the new matrix is $t|A|$, (Geometrically, multiplying one of the sides of the set S by a factor t causes the volume to increase by a factor t .)
3. If we exchange any two rows a_i^T and a_j^T of A , then the determinant of the new matrix is $-|A|$, for example

The Determinant: Properties

- For $A \in \mathbb{R}^{n \times n}$, $|A| = |A^T|$.
- For $A, B \in \mathbb{R}^{n \times n}$, $|AB| = |A||B|$.
- For $A \in \mathbb{R}^{n \times n}$, $|A| = 0$ if and only if A is singular (i.e., non-invertible). (If A is singular then it does not have full rank, and hence its columns are linearly dependent. In this case, the set S corresponds to a “flat sheet” within the n -dimensional space and hence has zero volume.)
- For $A \in \mathbb{R}^{n \times n}$ and A non-singular, $|A^{-1}| = 1/|A|$.

Eigenvalues and Eigenvectors

Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an *eigenvalue* of A and $x \in \mathbb{C}^n$ is the corresponding *eigenvector* if

$$Ax = \lambda x, \quad x \neq 0.$$

$Ax = \underbrace{\lambda x}_{\text{Scalar}}$

Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an *eigenvalue* of A and $x \in \mathbb{C}^n$ is the corresponding *eigenvector* if

$$Ax = \lambda x, \quad x \neq 0.$$

Intuitively, this definition means that multiplying A by the vector x results in a new vector that points in the same direction as x , but scaled by a factor λ .

$$\frac{dy}{dx}$$

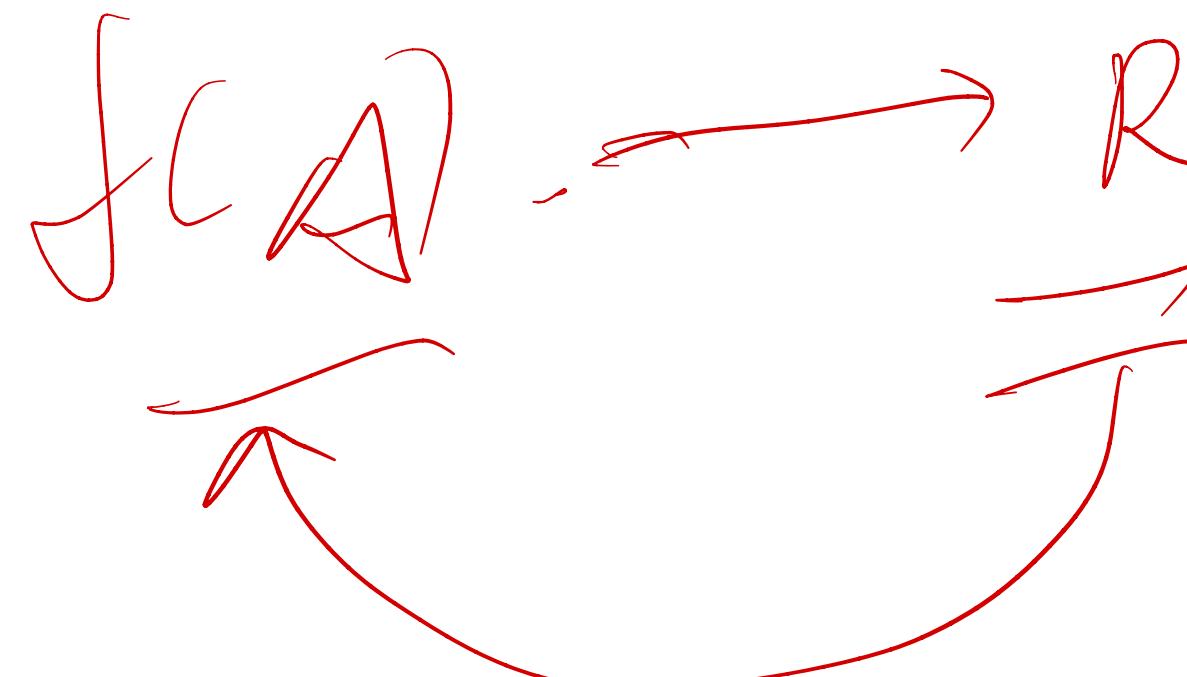
$$y = 2x + 3$$

Gradient over Matrix

$$\frac{dy}{dx}$$

$m \times n$

Suppose that $f: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the **gradient** of f (with respect to $A \in \mathbb{R}^{m \times n}$) is the matrix of partial derivatives, defined as:



$$\nabla_A f(A) \in \mathbb{R}^{m \times n}$$

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \dots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \dots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \dots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

$\nabla_A f(A) \in \mathbb{R}^{m \times n}$

Gradient over Vector

Gradient over Vector

Note that the size of $\nabla_A f(A)$ is always the same as the size of A . So if, in particular, A is just a vector $x \in \mathbb{R}^n$,

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

Gradient over Vector

Note that the size of $\nabla_A f(A)$ is always the same as the size of A . So if, in particular, A is just a vector $x \in \mathbb{R}^n$,

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

It follows directly from the equivalent properties of partial derivatives that:

- $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$.
- For $t \in \mathbb{R}$, $\nabla_x(t f(x)) = t \nabla_x f(x)$.

The Hessian

Second-order derivatives

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a function that takes a vector in \mathbb{R}^n and returns a real number. Then the **Hessian** matrix with respect to x , written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \dots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

$x \in \mathbb{R}^n$

Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}.$$

expensive

Gradients of Linear Functions

$f(x) = b^T x = \sum b_i x_i$

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$. Then

$$f(x) = \sum_{i=1}^n b_i x_i$$

so

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^n b_i x_i = b_k.$$

From this we can easily see that $\nabla_x b^T x = b$. This should be compared to the analogous situation in single variable calculus, where $\partial/(\partial x) ax = a$.

$$\nabla_b b^T x = \underline{\underline{b}}$$

Common Gradient Formula

- $\nabla_x b^T x = b$ $\frac{db^T}{dx} = b$ $x \in \mathbb{R}^n \quad A \in \mathbb{R}^{n \times n}$
 - $\nabla_x^2 b^T x = 0$ $\nabla_b = 0$
 - $\nabla_x x^T A x = 2Ax$ (if A symmetric)
 - $\nabla_x^2 x^T A x = 2A$ (if A symmetric)
- if A is not symmetric:
- $$\nabla_x x^T A x = (\hat{A} + \hat{A}^T)x$$
- $$\nabla_x^2 x^T A x = A + A^T$$

Least Squares

$A^T A$

- Given a full rank matrix $A \in \mathbb{R}^{m \times n}$, and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$, we want to find a vector x such that Ax is as close as possible to b , as measured by the square of the Euclidean norm $\|Ax - b\|_2^2$.

$$\|Ax - b\|^2 = (Ax - b)^T (Ax - b)$$

$$= x^T A^T A x - 2b^T A x + b^T b$$

$$\mathcal{J}_x = \mathcal{J}_x x^T A^T A x - \mathcal{J}_x 2b^T A x + \mathcal{J}_x b^T b$$

$$= 2A^T A x - 2A^T b + 0$$

≥ 0

$$\Rightarrow x^* = [A^T A]^{-1} A^T b$$

Outline

- Linear Algebra Review

- Probability Review

Basic Concepts

Basic Concepts

- Performing an **experiment** → **outcome**
- **Sample Space (S):** set of all possible outcomes of an experiment
- **Event (E):** a subset of S ($E \subseteq S$)



Basic Concepts

- Performing an **experiment** → **outcome**
- **Sample Space (S)**: set of all possible outcomes of an experiment
- **Event (E)**: a subset of S ($E \subseteq S$)
- **Probability (Bayesian definition)**
A number between 0 and 1 to which we ascribe meaning
i.e. our belief that an event E occurs

Basic Concepts

- Performing an **experiment** → **outcome**
- **Sample Space (S)**: set of all possible outcomes of an experiment
- **Event (E)**: a subset of S ($E \subseteq S$)
- **Probability (Bayesian definition)**
A number between 0 and 1 to which we ascribe meaning
i.e. our belief that an event E occurs

- **Frequentist definition of probability**

$$P(E) = \lim_{n \rightarrow \infty} \frac{n(E)}{n}$$


Axiom 1:

$$0 \leq P(E) \leq 1$$

Axiom 2:

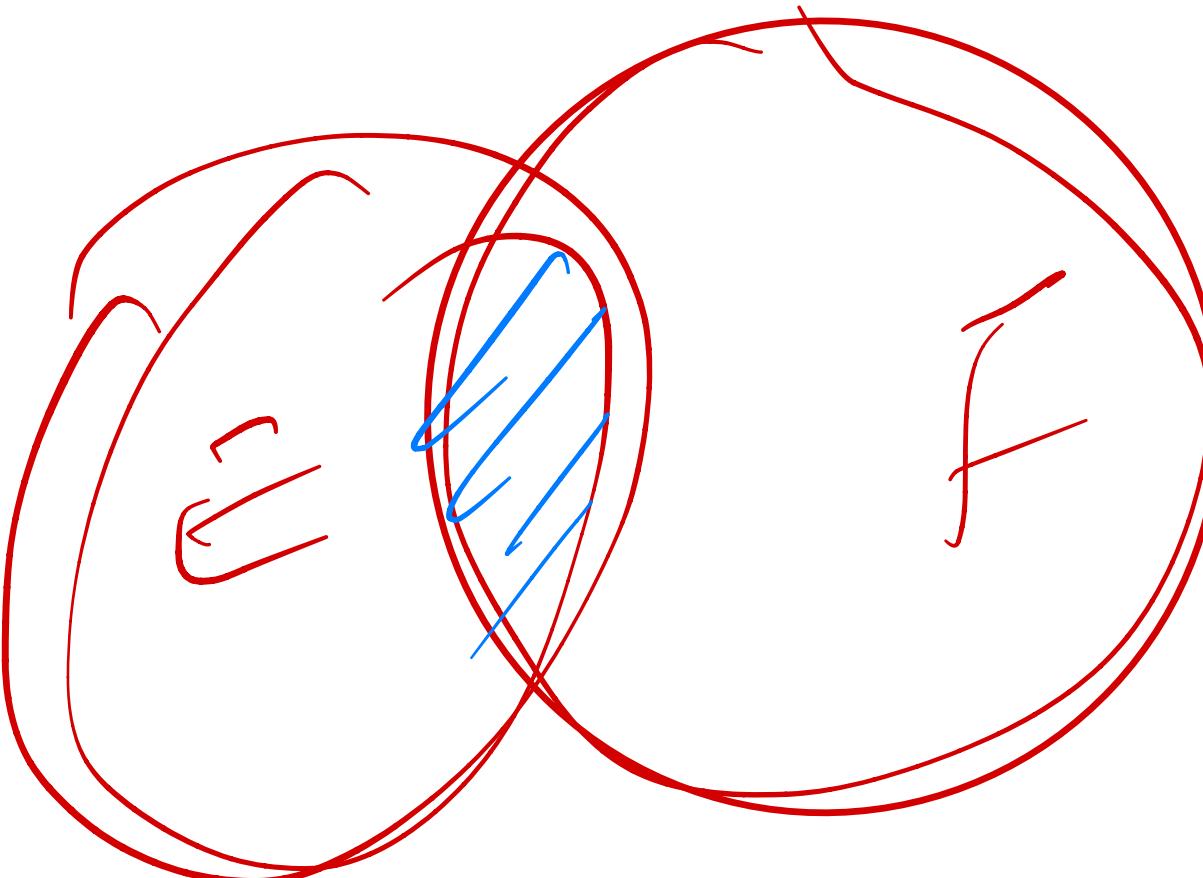
$$P(S) = 1$$

Axiom 1:

$$0 \leq P(E) \leq 1$$

Axiom 2:

$$P(S) = 1$$



$E \subseteq F$, then $P(E) \leq P(F)$

$P(E \cup F) = P(E) + P(F) - P(EF)$ (Inclusion-Exclusion Principle)



$P(E \cup F)$

Axiom 1: $0 \leq P(E) \leq 1$

Axiom 2: $P(S) = 1$

$E \subseteq F$, then $P(E) \leq P(F)$

$P(E \cup F) = P(E) + P(F) - P(EF)$ (Inclusion-Exclusion Principle)

General Inclusion-Exclusion Principle:

$$P\left(\bigcup_{i=1}^n E_i\right) = \sum_{r=1}^n (-1)^{r+1} \sum_{i_1 < \dots < i_r} P(E_{i_1} E_{i_2} \dots E_{i_r})$$

Axiom 1: $0 \leq P(E) \leq 1$

Axiom 2: $P(S) = 1$

$E \subseteq F$, then $P(E) \leq P(F)$

$P(E \cup F) = P(E) + P(F) - P(EF)$ (Inclusion-Exclusion Principle)

General Inclusion-Exclusion Principle:

$$P\left(\bigcup_{i=1}^n E_i\right) = \sum_{r=1}^n (-1)^{r+1} \sum_{i_1 < \dots < i_r} P(E_{i_1} E_{i_2} \dots E_{i_r})$$

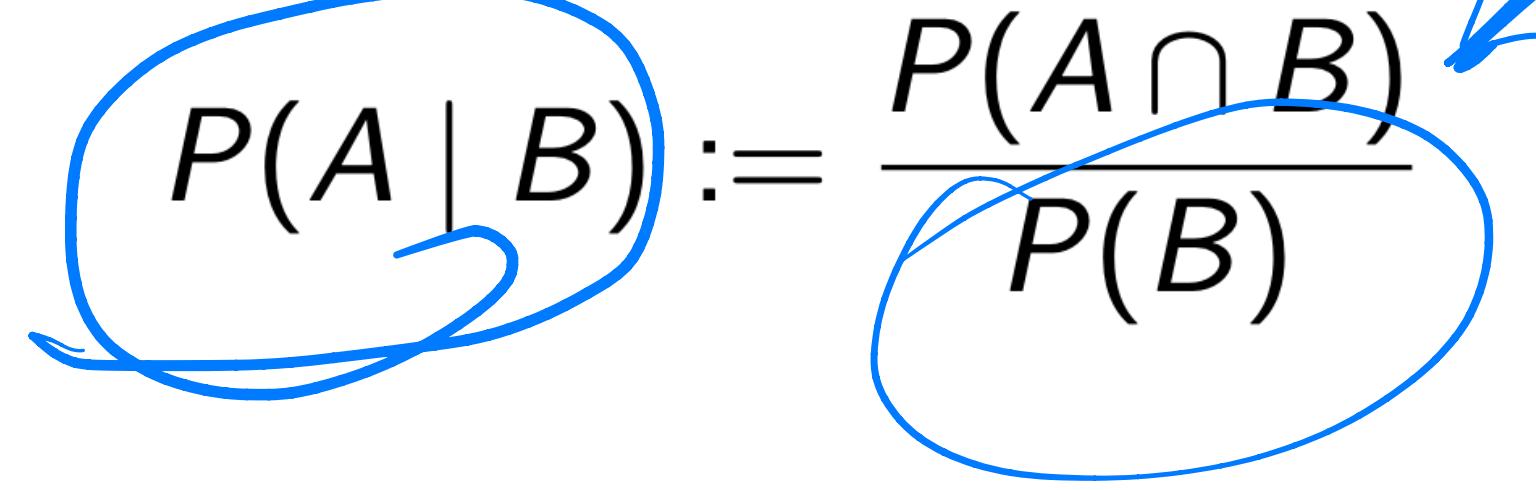
Equally Likely Outcomes: Define S as a sample space with equally likely outcomes. Then

$$P(E) = \frac{|E|}{|S|}$$

Conditional Probability and Bayes' Rule

Conditional Probability and Bayes' Rule

For any events A, B such that $P(B) \neq 0$, we define:

$$P(A | B) := \frac{P(A \cap B)}{P(B)}$$


Conditional Probability and Bayes' Rule

For any events A, B such that $P(B) \neq 0$, we define:

$$P(A | B) := \frac{P(A \cap B)}{P(B)}$$

Let's apply conditional probability to obtain Bayes' Rule!

$$\begin{aligned} P(B | A) &= \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} \\ &= \boxed{\frac{P(B)P(A | B)}{P(A)}} \end{aligned}$$

$P(B | A)$

$P(A | B)$

$P(A | B)$



$P(B | A)$

Conditional Probability and Bayes' Rule

For any events A, B such that $P(B) \neq 0$, we define:

$$P(A | B) := \frac{P(A \cap B)}{P(B)}$$

Let's apply conditional probability to obtain Bayes' Rule!

$$\begin{aligned} P(B | A) &= \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} \\ &= \boxed{\frac{P(B)P(A | B)}{P(A)}} \end{aligned}$$

Conditioned Bayes' Rule: given events A, B, C ,

$$P(A | B, C) = \frac{P(B | A, C)P(A | C)}{P(B | C)}$$

Law of Total Probability

Law of Total Probability

Let B_1, \dots, B_n be n disjoint events whose union is the entire sample space. Then, for any event A ,

$$P(A) = \sum_{i=1}^n P(A \cap B_i)$$

$$= \sum_{i=1}^n P(A | B_i)P(B_i)$$

$$P(A) = \sum_{B_i} P(A, B_i)$$

$$P(X) = \sum_{Z} P(X, Z) = \sum_{Z} P(Z)P(X|Z)$$

Law of Total Probability

Let B_1, \dots, B_n be n disjoint events whose union is the entire sample space. Then, for any event A ,

$$\begin{aligned} P(A) &= \sum_{i=1}^n P(A \cap B_i) \\ &= \sum_{i=1}^n P(A | B_i)P(B_i) \end{aligned}$$

Posterior probability

We can then write Bayes' Rule as:

$$\begin{aligned} P(B_k | A) &= \frac{P(B_k)P(A | B_k)}{P(A)} \\ &= \boxed{\frac{P(B_k)P(A | B_k)}{\sum_{i=1}^n P(A | B_i)P(B_i)}} \end{aligned}$$

Chain Rule

x_n conditioned all previous words

$x_1, x_2, x_3, \dots, x_n$

unprogressive

For any n events A_1, \dots, A_n , the joint probability can be expressed as a product of conditionals:

$$P(A_1 \cap A_2 \cap \dots \cap A_n)$$

$$= P(A_1)P(A_2 | A_1)P(A_3 | A_2 \cap A_1)\dots P(A_n | A_{n-1} \cap A_{n-2} \cap \dots \cap A_1)$$

$$P(x_1, x_2, \dots, x_n, x_5) \rightarrow$$

$$P(x_1) P(x_2 | x_1)$$

$$P(x_3 | x_1, x_2) P(x_4 | x_1, x_2, x_3)$$

$$P(x_1, x_2, x_3) = P(x_1) P(x_2|x_1) P(x_3|x_1, x_2)$$

$$= P(x_3) P(x_2|x_3) P(x_1|x_2, x_3)$$

$$= P(x_2) P(x_1|x_2) P(x_3|x_1, x_2)$$

Independence

Independence

Events A, B are independent if

$$P(AB) = P(A)P(B)$$

$$P(A, B) = P(A) \underbrace{P(B|A)}_{\overline{P(B)}}$$

Independence

Events A, B are independent if

$$P(AB) = P(A)P(B)$$

We denote this as $A \perp B$. From this, we know that if $A \perp B$,

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Independence

Events A, B are independent if

$$P(AB) = P(A)P(B)$$

We denote this as $A \perp B$. From this, we know that if $A \perp B$,

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Implication: If two events are independent, observing one event does not change the probability that the other event occurs.

In general: events A_1, \dots, A_n are **mutually independent** if

$$P\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} P(A_i)$$

I.I.D

Independent

Identical
distribution

$x_1 \rightarrow x_2 \cdots x_n$

$$\log P(x_1, x_2, \dots, x_n) = \sum_i \log P(x_i)$$

$$\sum_i \log P(x_i)$$

$$\begin{aligned} &= \log P(x_1) + \log P(x_2) \\ &\quad + \log P(x_3 | x_1, x_2) + \dots \end{aligned}$$

Random Variable

A **random variable** X is a variable that probabilistically takes on different values. It maps outcomes to real values

Probability Mass Function (PMF)

Given a **discrete** RV X , a PMF maps values of X to probabilities.

$$p_X(x) := p(x) := P(X = x)$$

For a valid PMF, $\sum_{x \in Val(x)} p_X(x) = 1$.

Cumulative Distribution Function (CDF)

A CDF maps a continuous RV to a probability (i.e. $\mathbb{R} \rightarrow [0, 1]$)

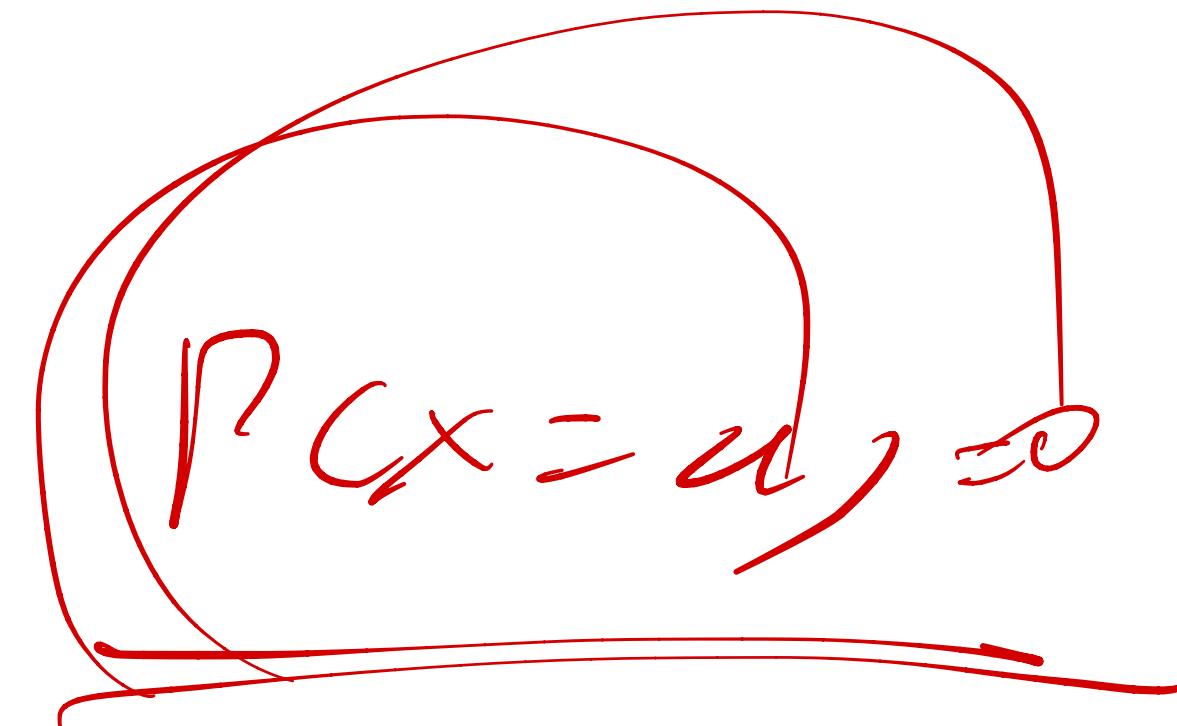
$$F_X(a) := F(a) := P(X \leq a)$$

Cumulative Distribution Function (CDF)

A CDF maps a continuous RV to a probability (i.e. $\mathbb{R} \rightarrow [0, 1]$)

$$F_X(a) := F(a) := P(X \leq a)$$

$$P(X \leq a)$$



A CDF must fulfill the following:

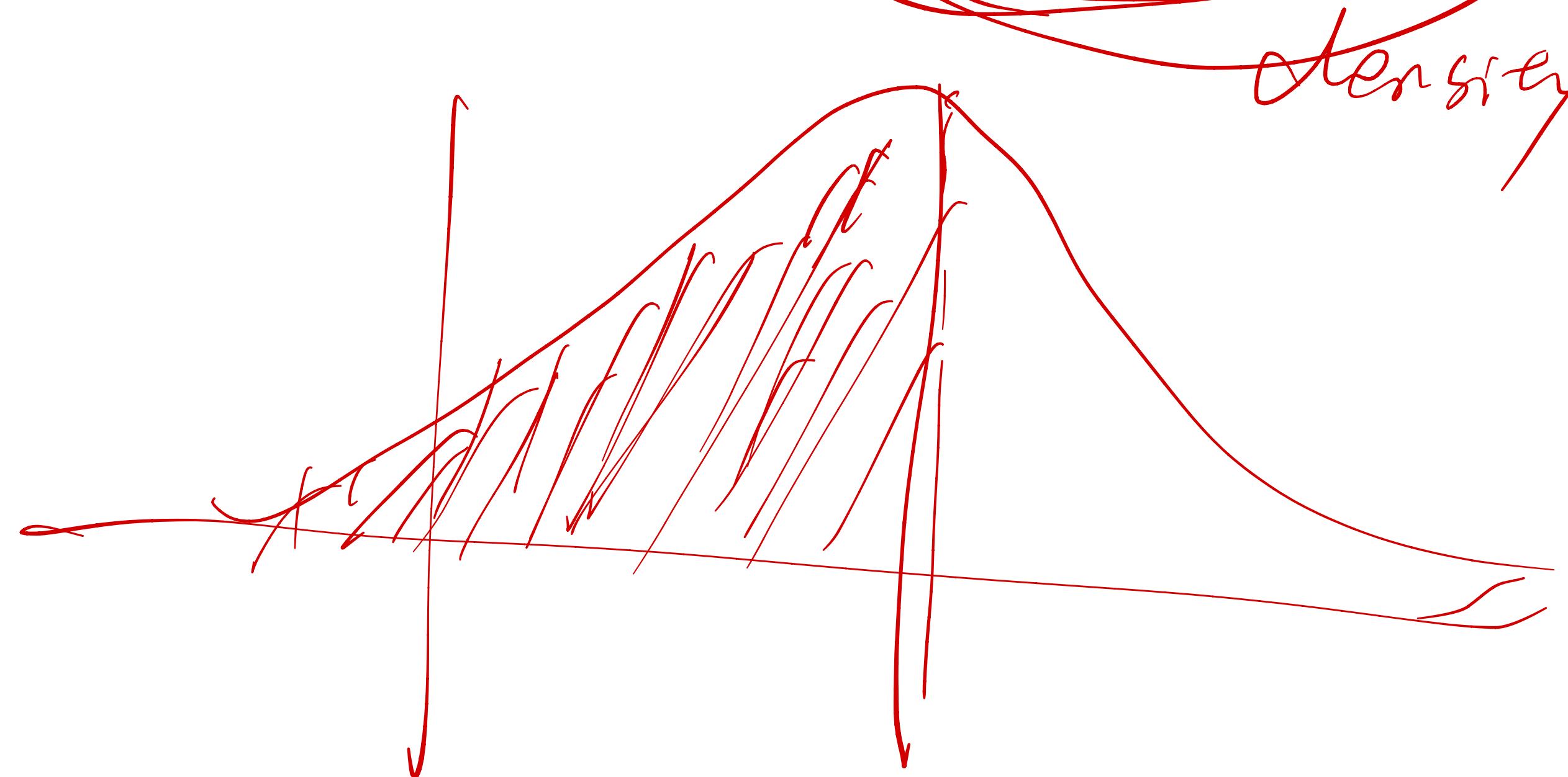
- $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- $\lim_{x \rightarrow \infty} F_X(x) = 1$
- If $a \leq b$, then $F_X(a) \leq F_X(b)$ (i.e. CDF must be nondecreasing)

Also note: $P(a \leq X \leq b) = F_X(b) - F_X(a)$.

Probability Density Function (PDF)

PDF of a continuous RV is simply the derivative of the CDF

$$f_X(x) := f(x) := \frac{dF_X(x)}{dx}$$



Expectation

Expectation

Let g be an arbitrary real-valued function.

- If X is a discrete RV with PMF p_X :

$$\mathbb{E}[g(X)] := \sum_{x \in Val(X)} g(x)p_X(x)$$

Expectation

Let g be an arbitrary real-valued function.

- If X is a discrete RV with PMF p_X :

$$\mathbb{E}[g(X)] := \sum_{x \in Val(X)} g(x)p_X(x)$$

- If X is a continuous RV with PDF f_X :

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x)f_X(x)dx$$


Expectation

Let g be an arbitrary real-valued function.

- If X is a discrete RV with PMF p_X :

$$\mathbb{E}[g(X)] := \sum_{x \in Val(X)} g(x)p_X(x)$$

- If X is a continuous RV with PDF f_X :

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

Intuitively, expectation is a weighted average of the values of $g(x)$, weighted by the probability of x .

Conditional Expectation

$\mathbb{E}[X | Y] = \sum_{x \in Val(x)} x p_{X|Y}(x|y)$ is a function of Y .

Properties of Expectation

Properties of Expectation

For any constant $a \in \mathbb{R}$ and arbitrary real function f :

- $\mathbb{E}[a] = a$
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$

Properties of Expectation

For any constant $a \in \mathbb{R}$ and arbitrary real function f :

- $\mathbb{E}[a] = a$
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$

Linearity of Expectation

Given n real-valued functions $f_1(X), \dots, f_n(X)$,

$$\mathbb{E}\left[\sum_{i=1}^n f_i(X)\right] = \sum_{i=1}^n \mathbb{E}[f_i(X)]$$

$C = A \text{ or } B$

Example

$$P(x) = \sum_c P(x, c)$$

El Goog sources two batteries, A and B , for its phone. A phone with battery A runs on average 12 hours on a single charge, but only 8 hours on average with battery B . El Goog puts battery A in 80% of its phones and battery B in the rest. If you buy a phone from El Goog, how many hours do you expect it to run on a single charge?

how many hours $\rightarrow x$

battery type $\rightarrow c$

$$P(C=c=A) = 80\%$$

$$E(x | C=A) = 12$$

$$= 11.2$$

$$P(A) = 80\%$$

$$\begin{aligned} E(x | C=A) &= 12 \\ E(x) &= \sum_x x \cdot P(x) \\ &= \sum_x x \cdot P(x, C=A) + \sum_x x \cdot P(x, C=B) \\ &= 0.2 \cdot 0.8 \cdot 12 + 0.2 \cdot 8 \\ &= 11.2 \end{aligned}$$

Variance

The **variance** of a RV X measures how concentrated the distribution of X is around its mean.

$$\begin{aligned} \text{Var}(X) &:= \mathbb{E}[(X - \mathbb{E}[X])^2] && \rightarrow \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

Interpretation: $\text{Var}(X)$ is the expected deviation of X from $\mathbb{E}[X]$.

Properties: For any constant $a \in \mathbb{R}$, real-valued function $f(X)$

- $\text{Var}[a] = 0$
- $\text{Var}[af(X)] = a^2 \text{Var}[f(X)]$

Example Distributions

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0. \end{cases}$	p	$p(1 - p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1 - p)^{n-k}$ for $k = 0, 1, \dots, n$	np	$np(1 - p)$
$Geometric(p)$	$p(1 - p)^{k-1}$ for $k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda} \lambda^k}{k!}$ for $k = 0, 1, \dots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Joint and Marginal Distributions

Joint and Marginal Distributions

- Joint PMF for discrete RV's X, Y :

$$p_{XY}(x, y) = P(X = x, Y = y)$$

Note that $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} p_{XY}(x, y) = 1$

Joint and Marginal Distributions

- Joint PMF for discrete RV's X, Y :

$$p_{XY}(x, y) = P(X = x, Y = y)$$

Note that $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} p_{XY}(x, y) = 1$

$$P(X) \leq \sum_{y} p_{XY}$$

- Marginal PMF of X , given joint PMF of X, Y :

$$p_X(x) = \sum_y p_{XY}(x, y)$$

Joint and Marginal Distributions

- Joint PDF for continuous RV's X_1, \dots, X_n :

$$f(x_1, \dots, x_n) = \frac{\delta^n F(x_1, \dots, x_n)}{\delta x_1 \delta x_2 \dots \delta x_n}$$

Note that $\int_{x_1} \int_{x_2} \dots \int_{x_n} f(x_1, \dots, x_n) dx_1 \dots dx_n = 1$

- Marginal PDF of X_1 , given joint PDF of X_1, \dots, X_n :

$$f_{X_1}(x_1) = \int_{x_2} \dots \int_{x_n} f(x_1, \dots, x_n) dx_2 \dots dx_n$$

Expectation for multiple random variables

Expectation for multiple random variables

Given two RV's X, Y and a function $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ of X, Y ,

- for discrete X, Y :

$$\mathbb{E}[g(X, Y)] := \sum_{x \in Val(x)} \sum_{y \in Val(y)} g(x, y) p_{XY}(x, y)$$

Expectation for multiple random variables

Given two RV's X, Y and a function $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ of X, Y ,

- for discrete X, Y :

$$\mathbb{E}[g(X, Y)] := \sum_{x \in Val(x)} \sum_{y \in Val(y)} g(x, y) p_{XY}(x, y)$$

- for continuous X, Y :

$$\mathbb{E}[g(X, Y)] := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy$$

Covariance

Intuitively: measures how much one RV's value tends to move with another RV's value. For RV's X, Y :

$$\begin{aligned} \text{Cov}[X, Y] &:= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

Covariance

Intuitively: measures how much one RV's value tends to move with another RV's value. For RV's X, Y :

$$\begin{aligned} \text{Cov}[X, Y] &:= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

- If $\text{Cov}[X, Y] < 0$, then X and Y are negatively correlated
- If $\text{Cov}[X, Y] > 0$, then X and Y are positively correlated
- If $\text{Cov}[X, Y] = 0$, then X and Y are uncorrelated

Variance of two variables

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$$

Conditional distributions for RVs

Works the same way with *RV's* as with events:

- For discrete X, Y :

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

- For continuous X, Y :

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

- In general, for continuous X_1, \dots, X_n :

$$f_{X_1|X_2, \dots, X_n}(x_1|x_2, \dots, x_n) = \frac{f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)}{f_{X_2, \dots, X_n}(x_2, \dots, x_n)}$$

Bayes' Rule for RVs

Also works the same way for *RV's* as with events:

- For discrete X, Y :

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y' \in Val(Y)} p_{X|Y}(x|y')p_Y(y')}$$

- For continuous X, Y :

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y')dy'}$$

Random Vectors

Random Vectors

Given n RV's X_1, \dots, X_n , we can define a random vector X s.t.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Random Vectors

Given n RV's X_1, \dots, X_n , we can define a random vector X s.t.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Note: all the notions of joint PDF/CDF will apply to X .

Given $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we have:

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}, \mathbb{E}[g(X)] = \begin{bmatrix} \mathbb{E}[g_1(X)] \\ \mathbb{E}[g_2(X)] \\ \vdots \\ \mathbb{E}[g_m(X)] \end{bmatrix}$$

Covariance Matrices

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij -th entry contains the covariance between X_i and X_j .

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij -th entry contains the covariance between X_i and X_j .

$$\Sigma = \begin{bmatrix} \text{Cov}[X_1, X_1] & \dots & \text{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \dots & \text{Cov}[X_n, X_n] \end{bmatrix}$$

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij -th entry contains the covariance between X_i and X_j .

$$\Sigma = \begin{bmatrix} \text{Cov}[X_1, X_1] & \dots & \text{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \dots & \text{Cov}[X_n, X_n] \end{bmatrix}$$

applying linearity of expectation and the fact that $\text{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$, we obtain

$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij -th entry contains the covariance between X_i and X_j .

$$\Sigma = \begin{bmatrix} \text{Cov}[X_1, X_1] & \dots & \text{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \dots & \text{Cov}[X_n, X_n] \end{bmatrix}$$

applying linearity of expectation and the fact that $\text{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$, we obtain

$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

Properties:

- Σ is symmetric and PSD
- If $X_i \perp X_j$ for all i, j , then $\Sigma = \text{diag}(\text{Var}[X_1], \dots, \text{Var}[X_n])$

Multivariate Gaussian

The multivariate Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^n$:

$$p(x; \mu, \Sigma) = \frac{1}{\det(\Sigma)^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

Multivariate Gaussian

The multivariate Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^n$:

$$p(x; \mu, \Sigma) = \frac{1}{\det(\Sigma)^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

Gaussian when $n = 1$.

$$p(x; \mu, \sigma^2) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

Notice that if $\Sigma \in \mathbb{R}^{1 \times 1}$, then $\Sigma = \text{Var}[X_1] = \sigma^2$, and so
 $\Sigma^{-1} = \frac{1}{\sigma^2}$ and $\det(\Sigma)^{\frac{1}{2}} = \sigma$



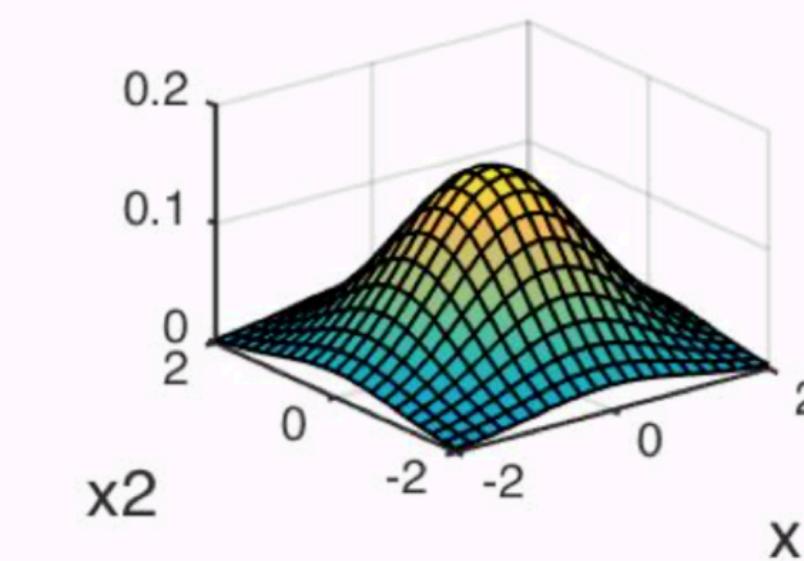
MV Gaussian Visualization

MV Gaussian Visualization

Effect of changing variance

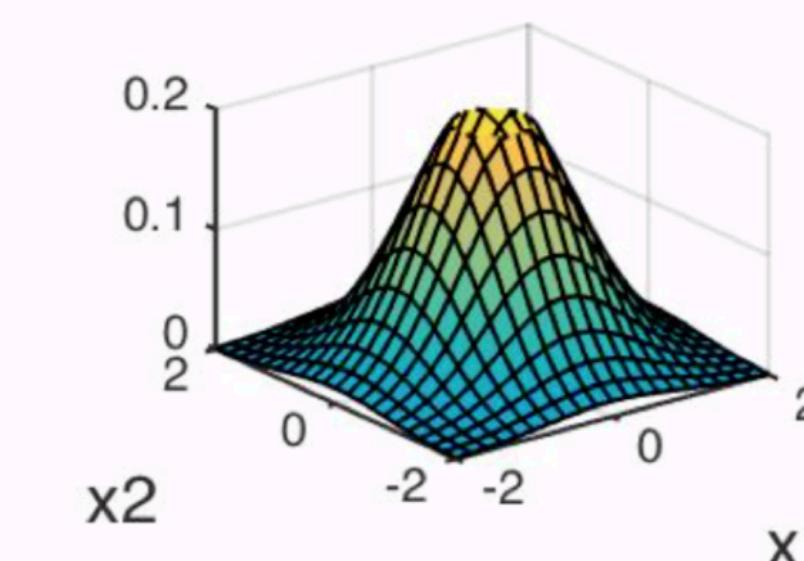
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$



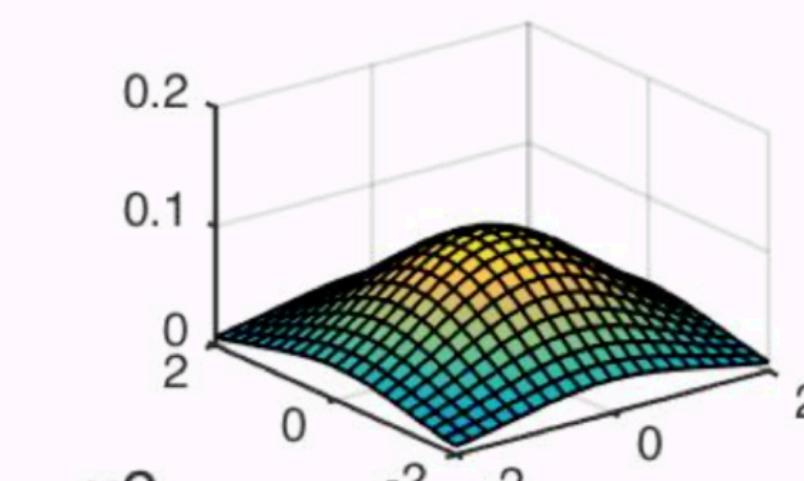
$$\Sigma = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$



$$\Sigma = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$



MV Gaussian Visualization

Effect of changing variance

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

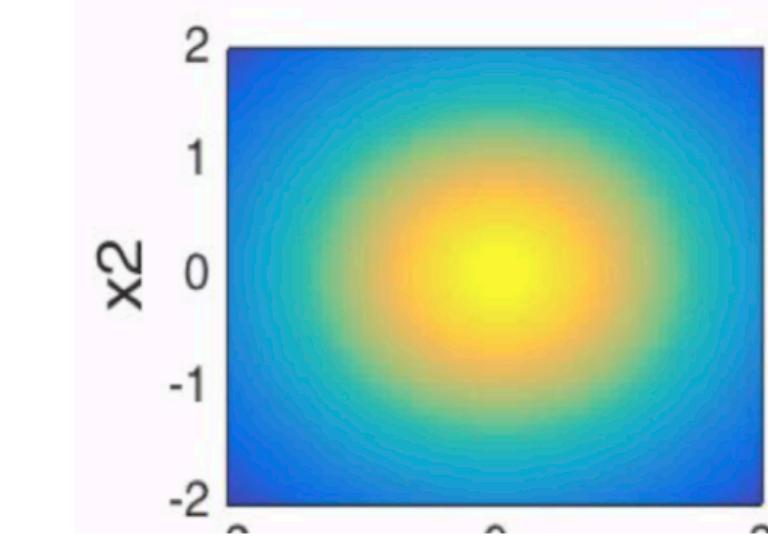
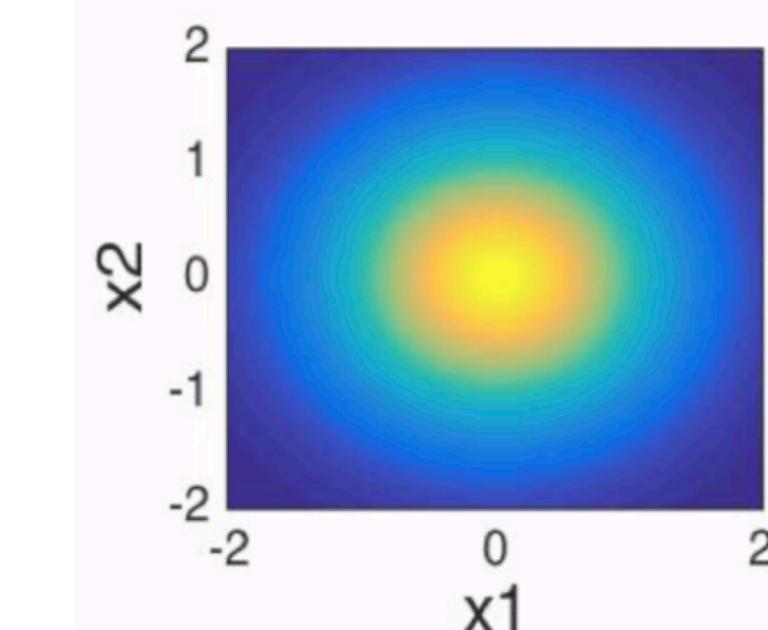
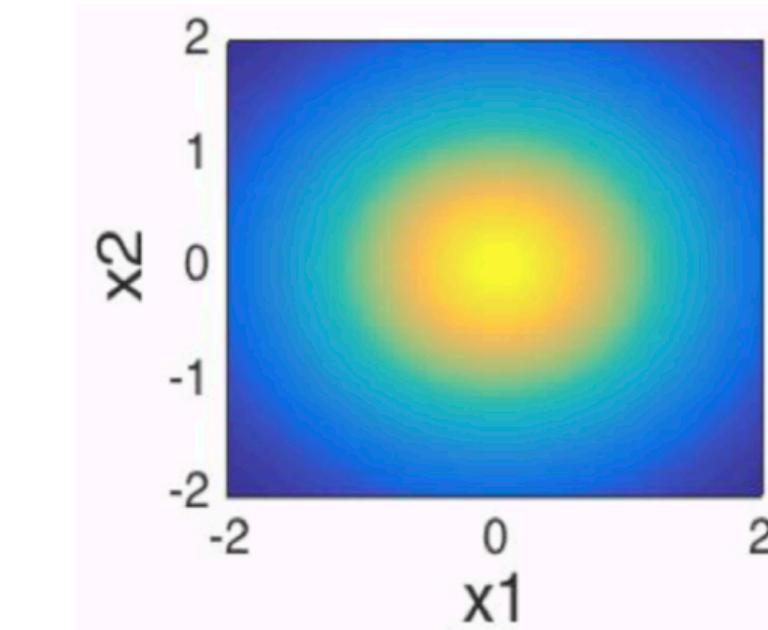
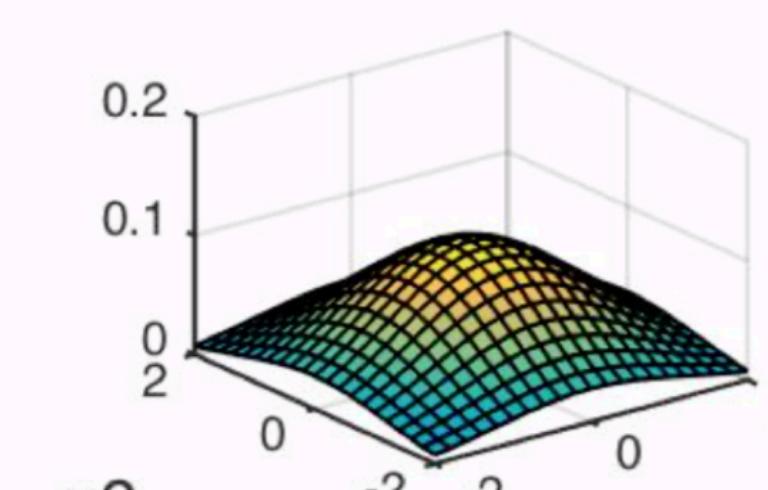
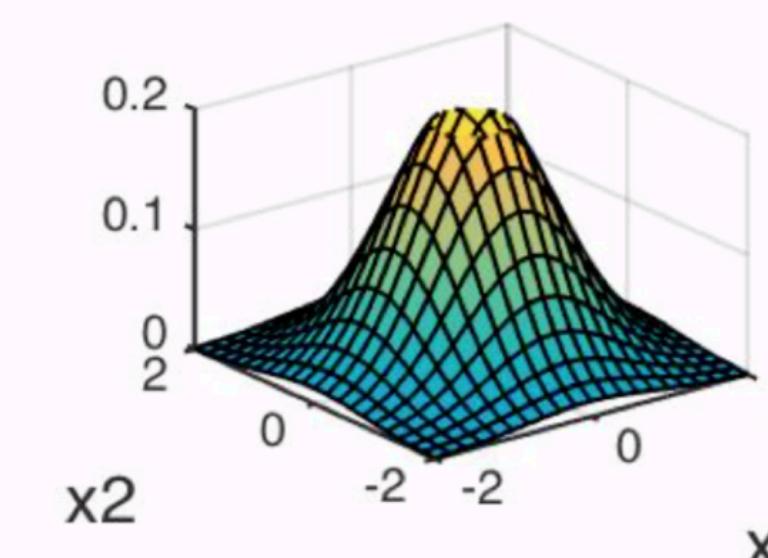
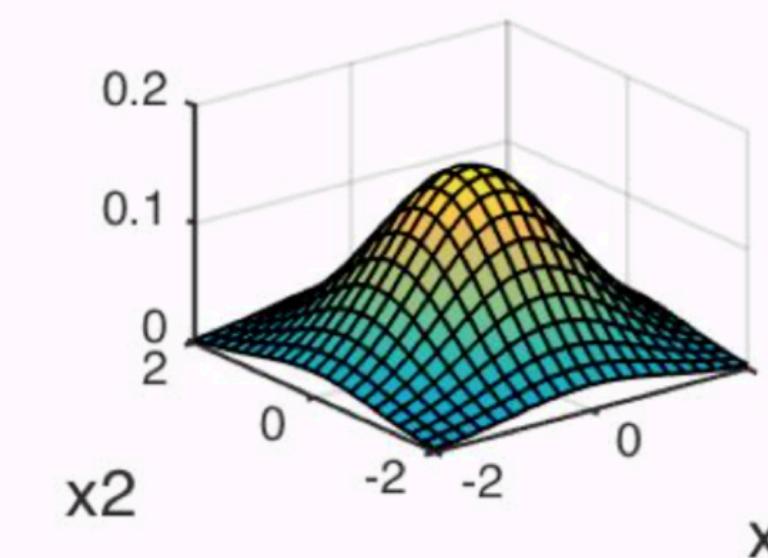
$$\mu = [0 \ 0]^T$$

$$\Sigma = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$

$$\Sigma = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$



MV Gaussian Visualization

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$

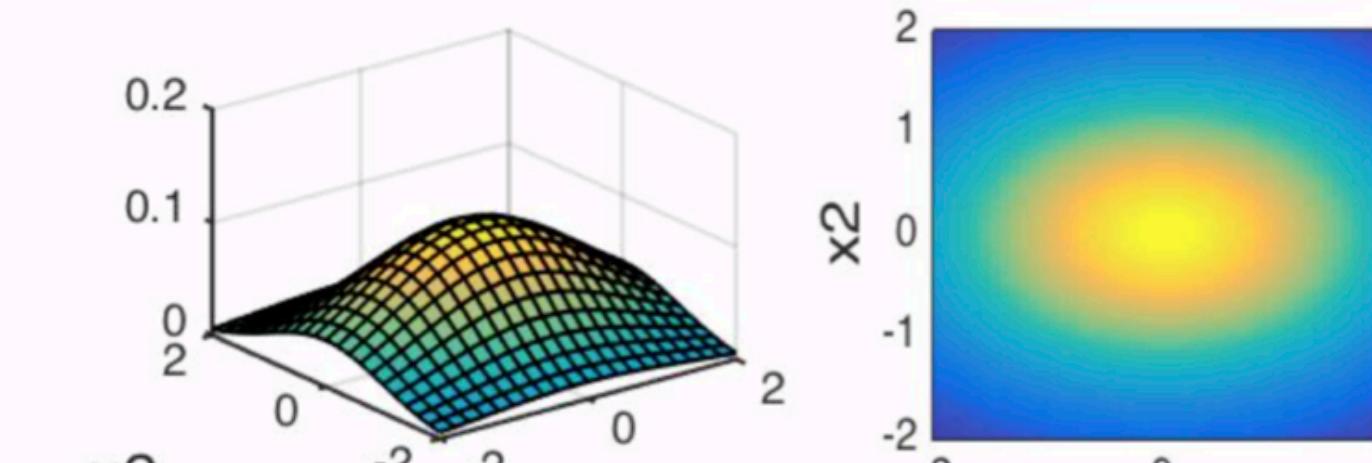
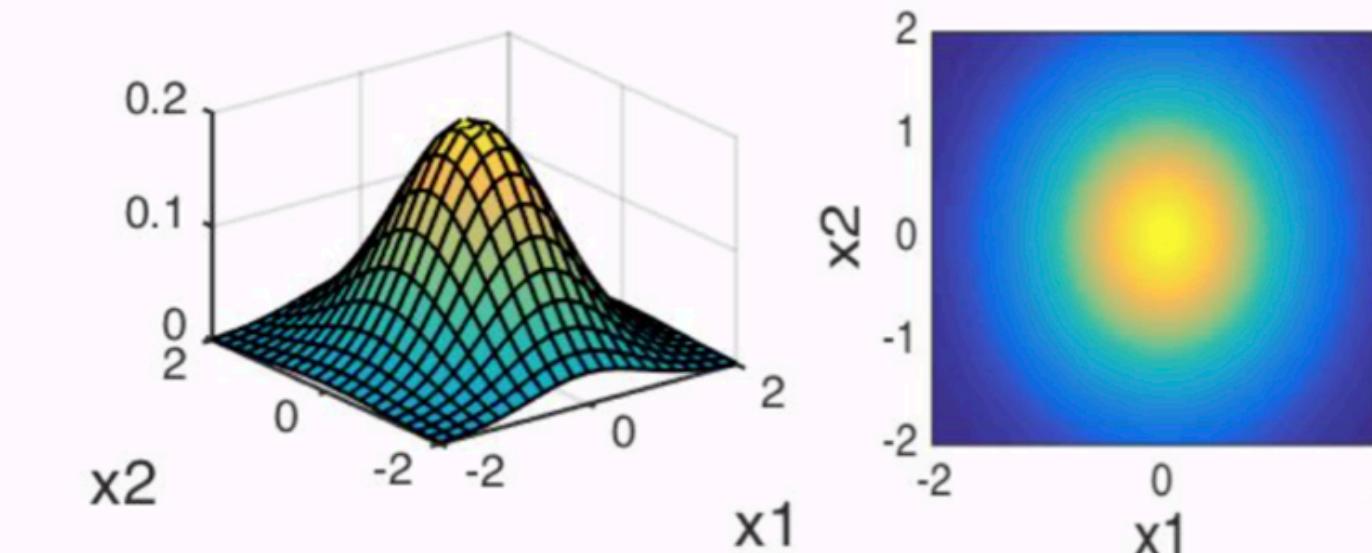
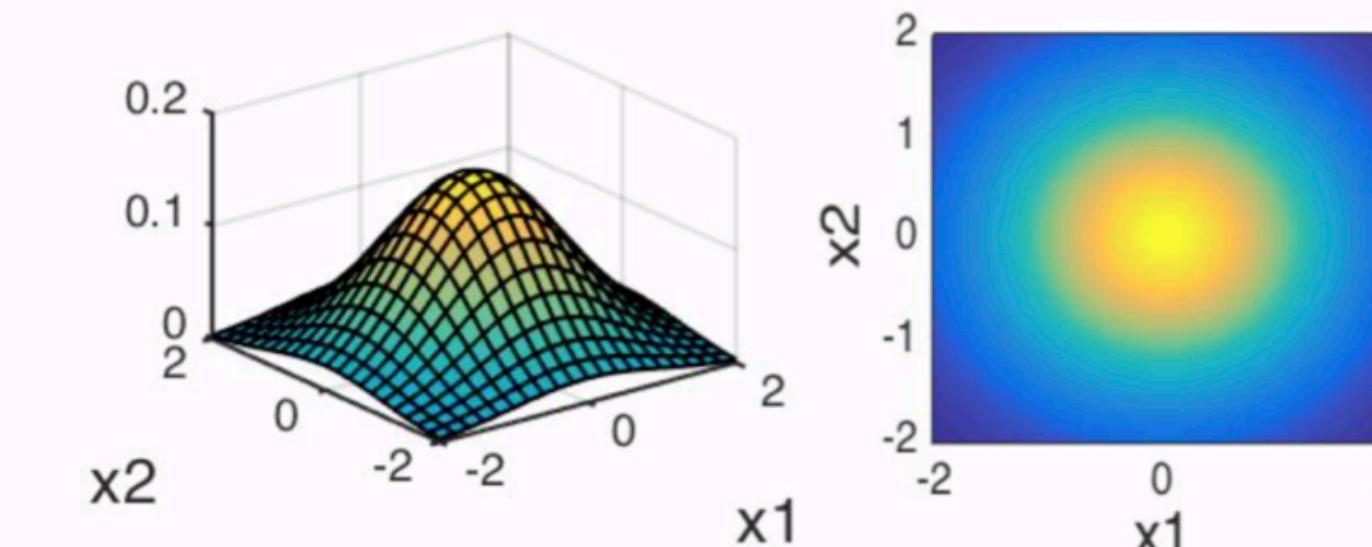
$$\Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$

If $\text{Var}[X_1] \neq \text{Var}[X_2]$:



MV Gaussian Visualization

If X_1 and X_2 are positively correlated:

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

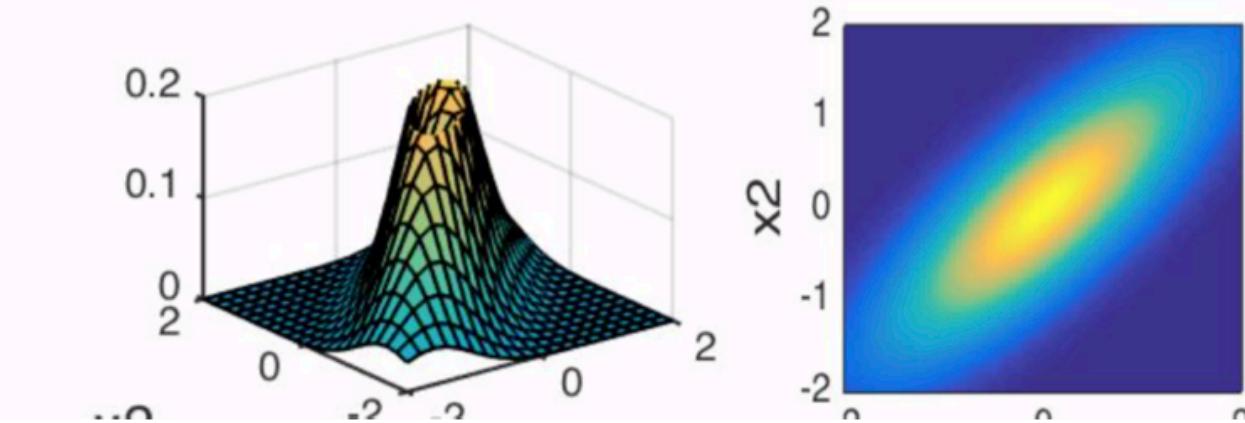
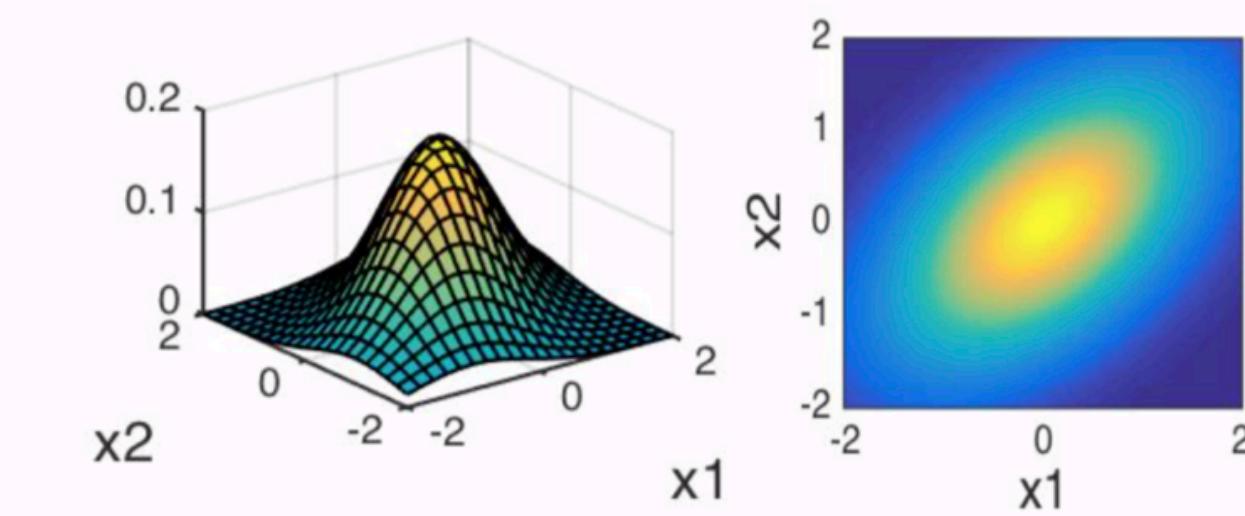
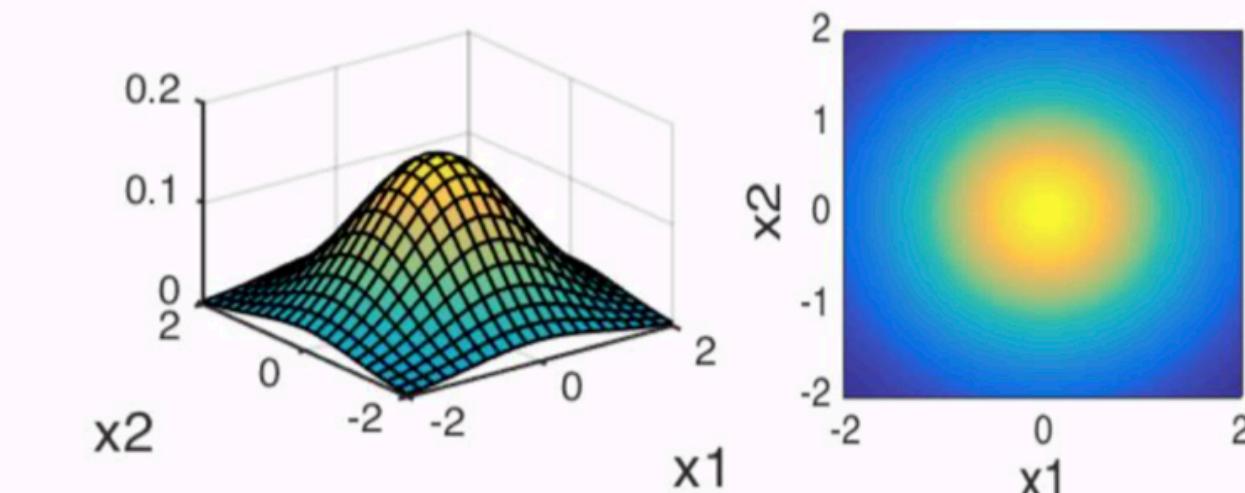
$$\mu = [0 \ 0]^T$$

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



MV Gaussian Visualization

If X_1 and X_2 are negatively correlated:

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

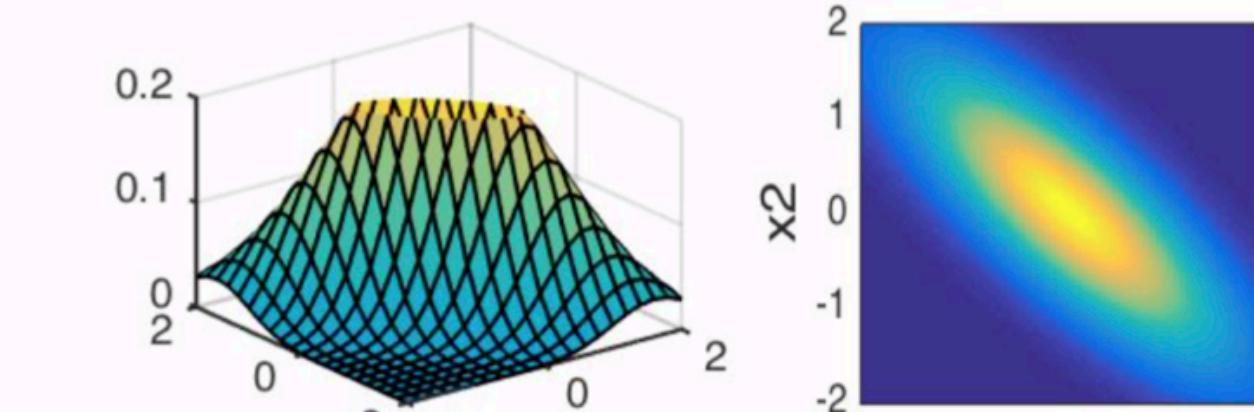
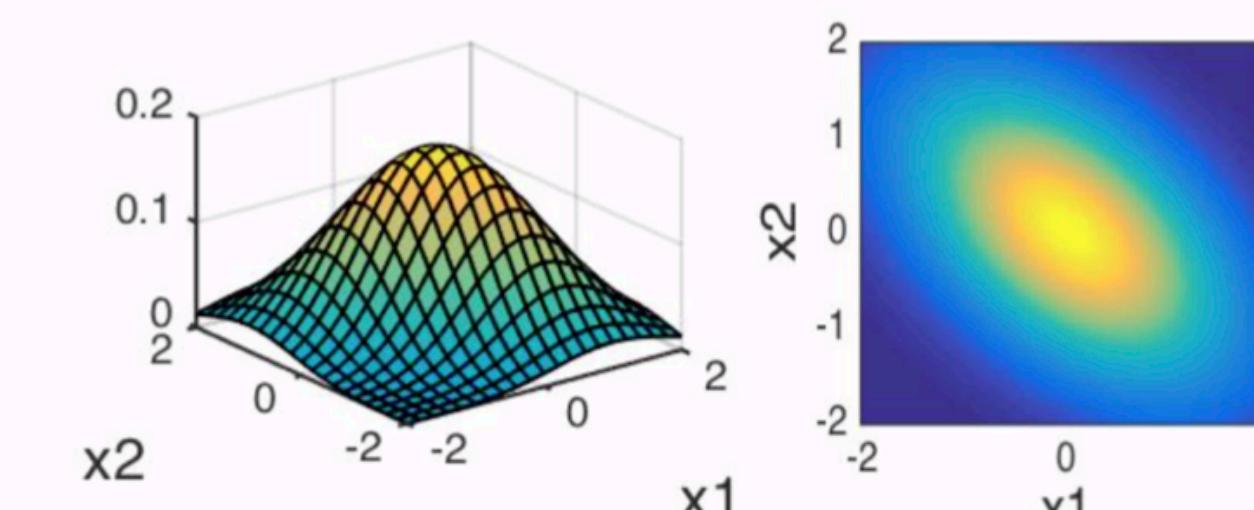
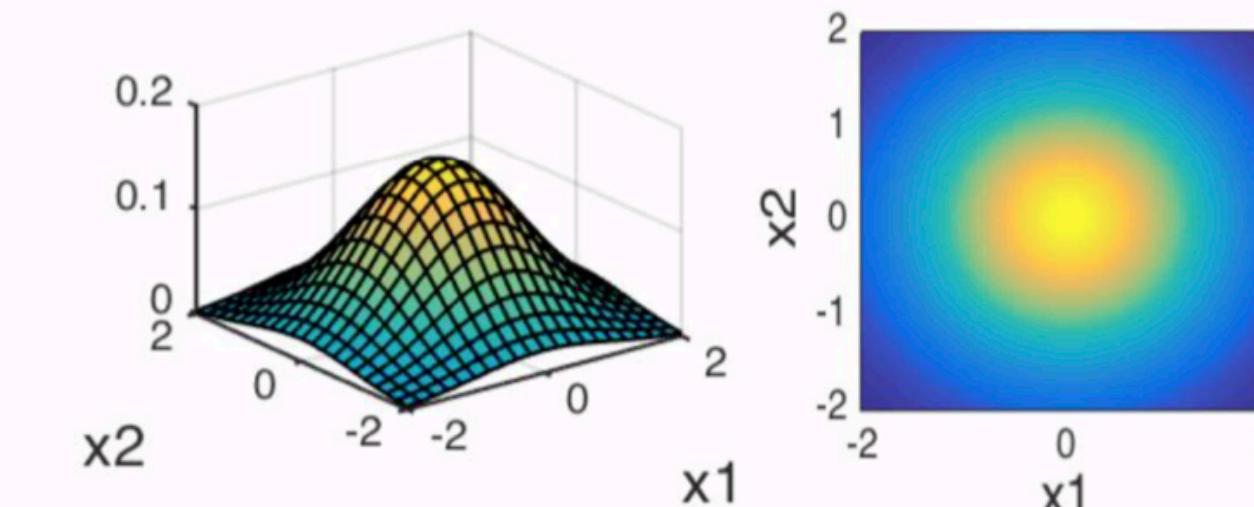
$$\mu = [0 \ 0]^T$$

$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$

$$\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

$$\mu = [0 \ 0]^T$$

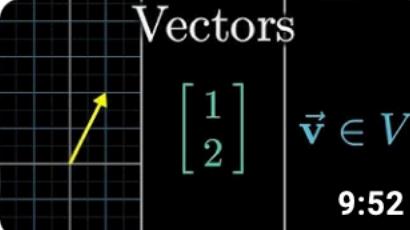
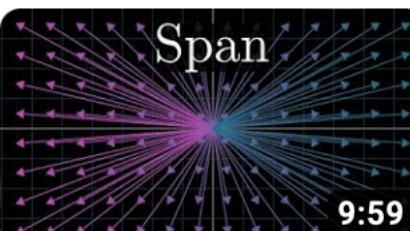
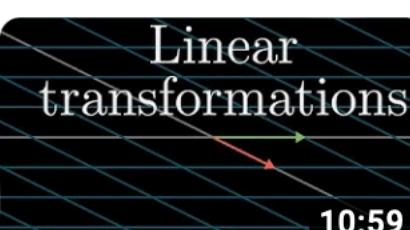
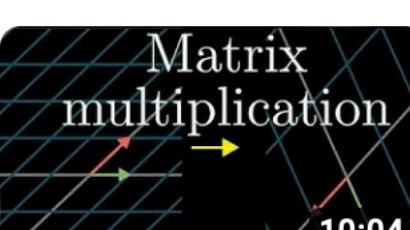
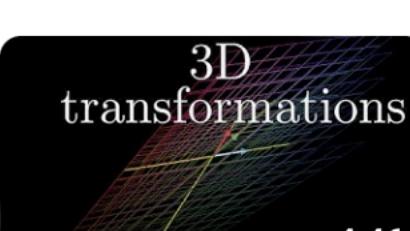
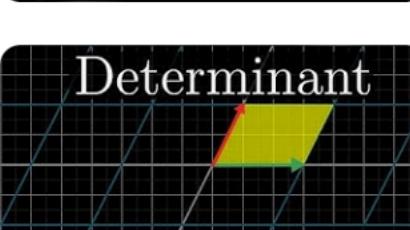
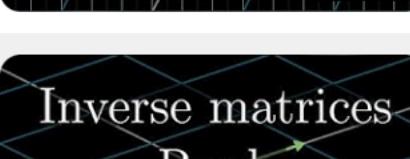


The purpose of computation is
insight, not numbers.

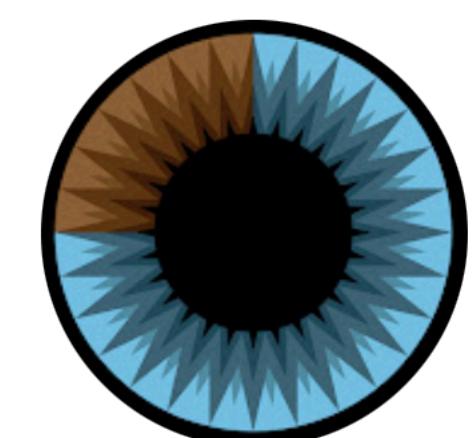
- Richard Hamming

The purpose of computation is insight, not numbers.

- Richard Hamming

- 1 Vectors | Chapter 1, Essence of linear algebra

[1]
 $\vec{v} \in V$
9:52
- 2 Linear combinations, span, and basis vectors | Chapter 2, Essence of linear algebra

Span
9:59
- 3 Linear transformations and matrices | Chapter 3, Essence of linear algebra

Linear transformations
10:59
- 4 Matrix multiplication as composition | Chapter 4, Essence of linear algebra

Matrix multiplication
10:04
- 5 Three-dimensional linear transformations | Chapter 5, Essence of linear algebra

3D transformations
4:46
- 6 Determinant | Chapter 6, Essence of linear algebra

Determinant
10:03
- 7 Inverse matrices, column space and null space | Chapter 7, Essence of linear algebra

Inverse matrices
Rank

<https://www.youtube.com/@3blue1brown/courses>



3Blue1Brown 

@3blue1brown · 5.88M subscribers · 172 videos

My name is Grant Sanderson. Videos here cover a variety of topics in math, or adjacent fiel... >

3blue1brown.com and 7 more links

 Subscribed 



香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 2

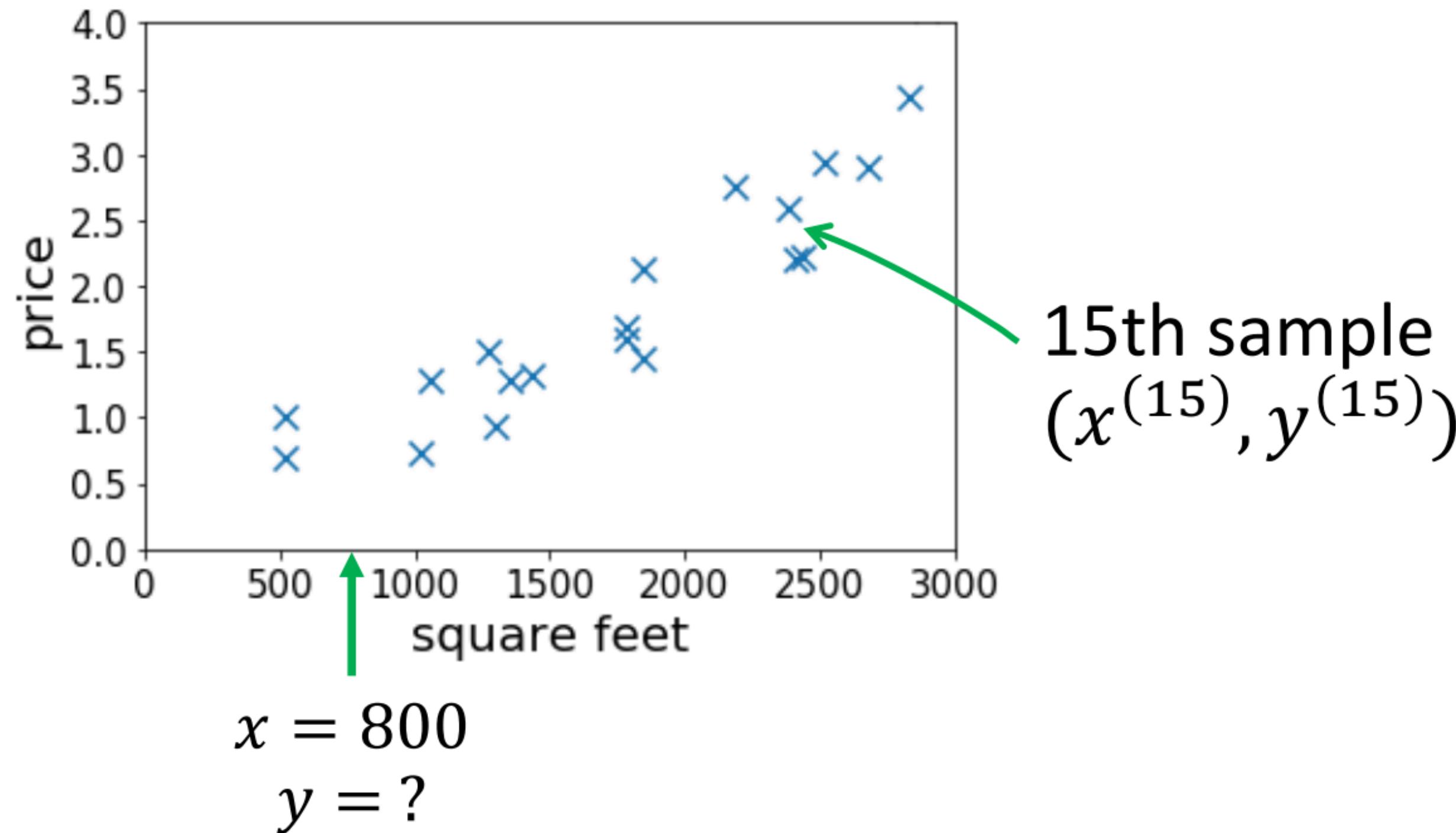
Supervised Learning: Regression

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$

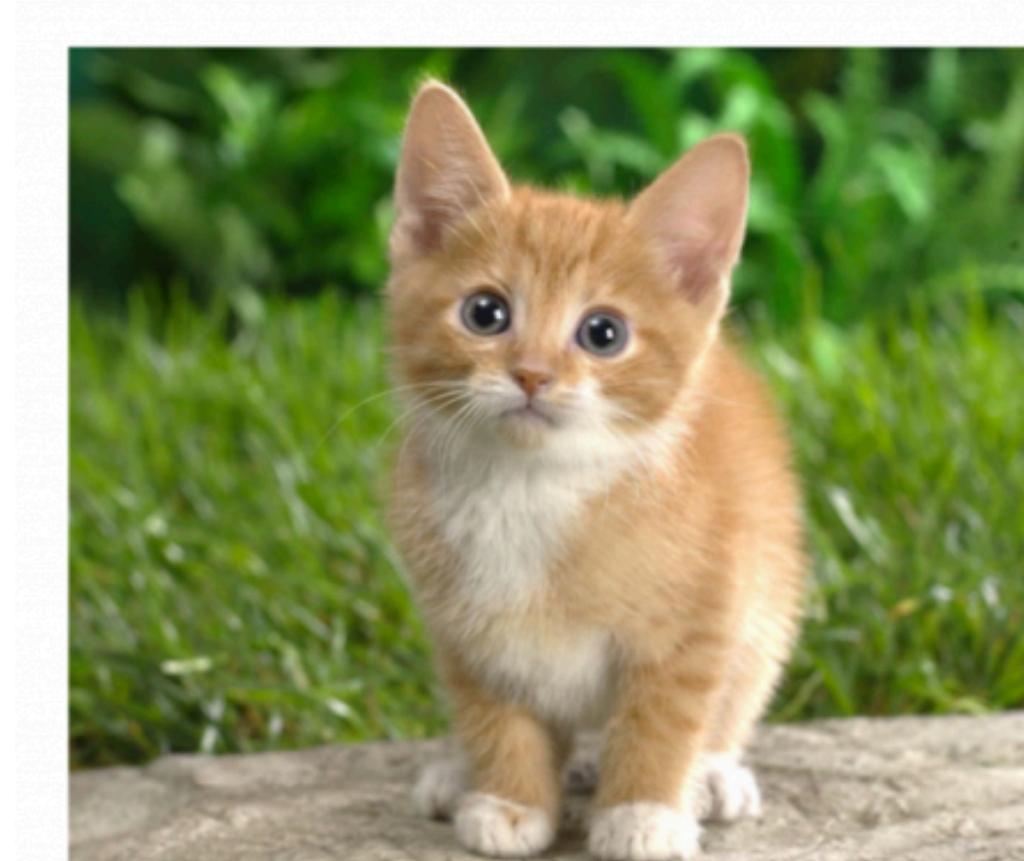
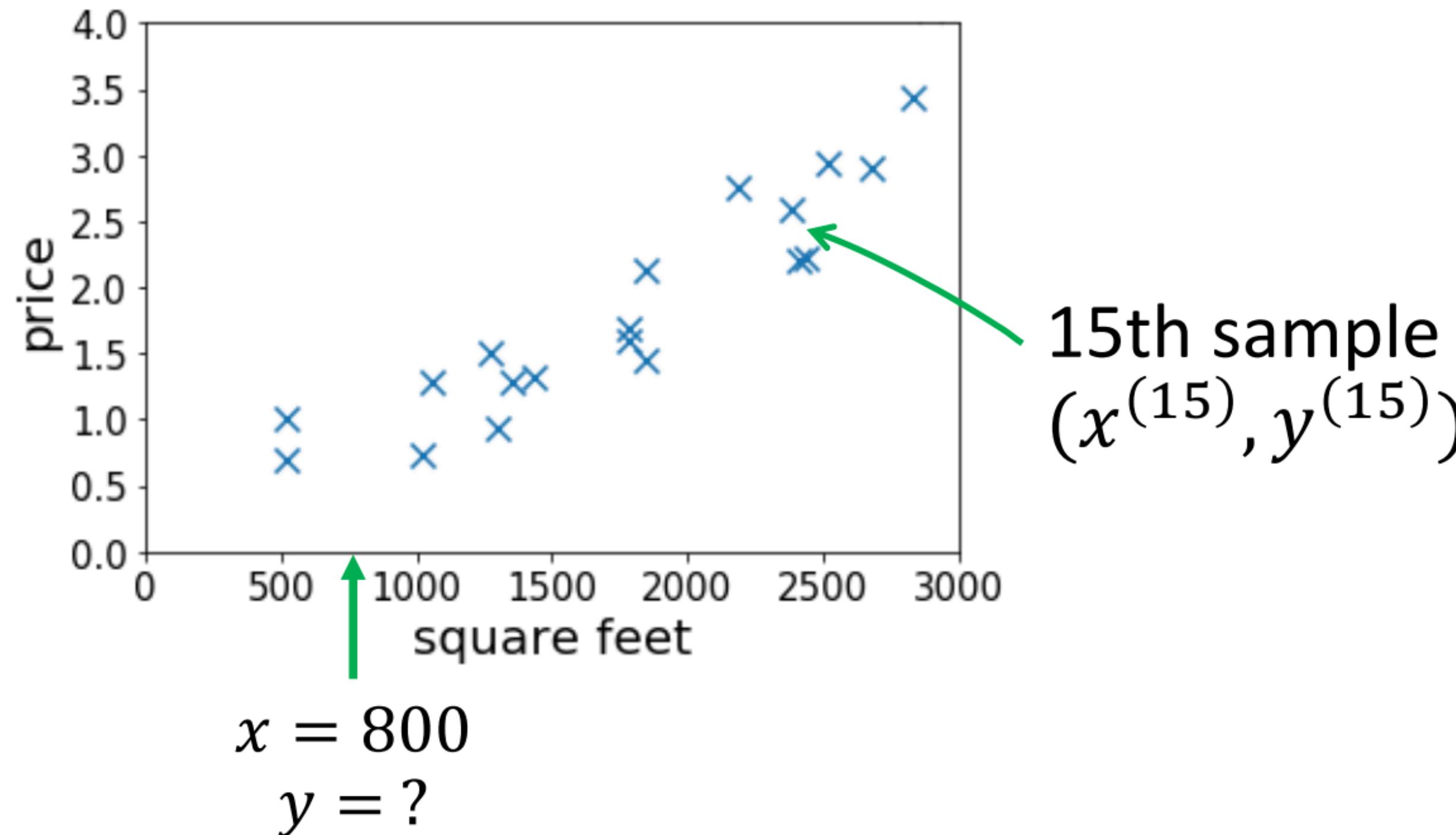
Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$



Supervised Learning

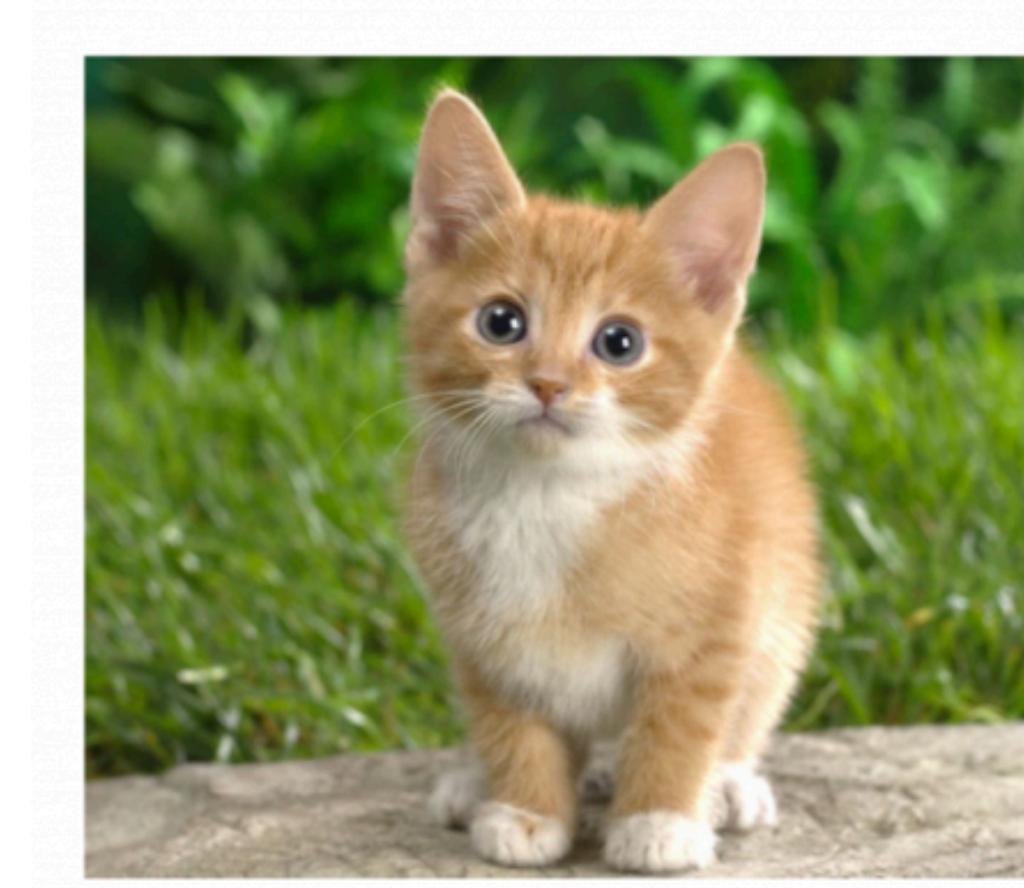
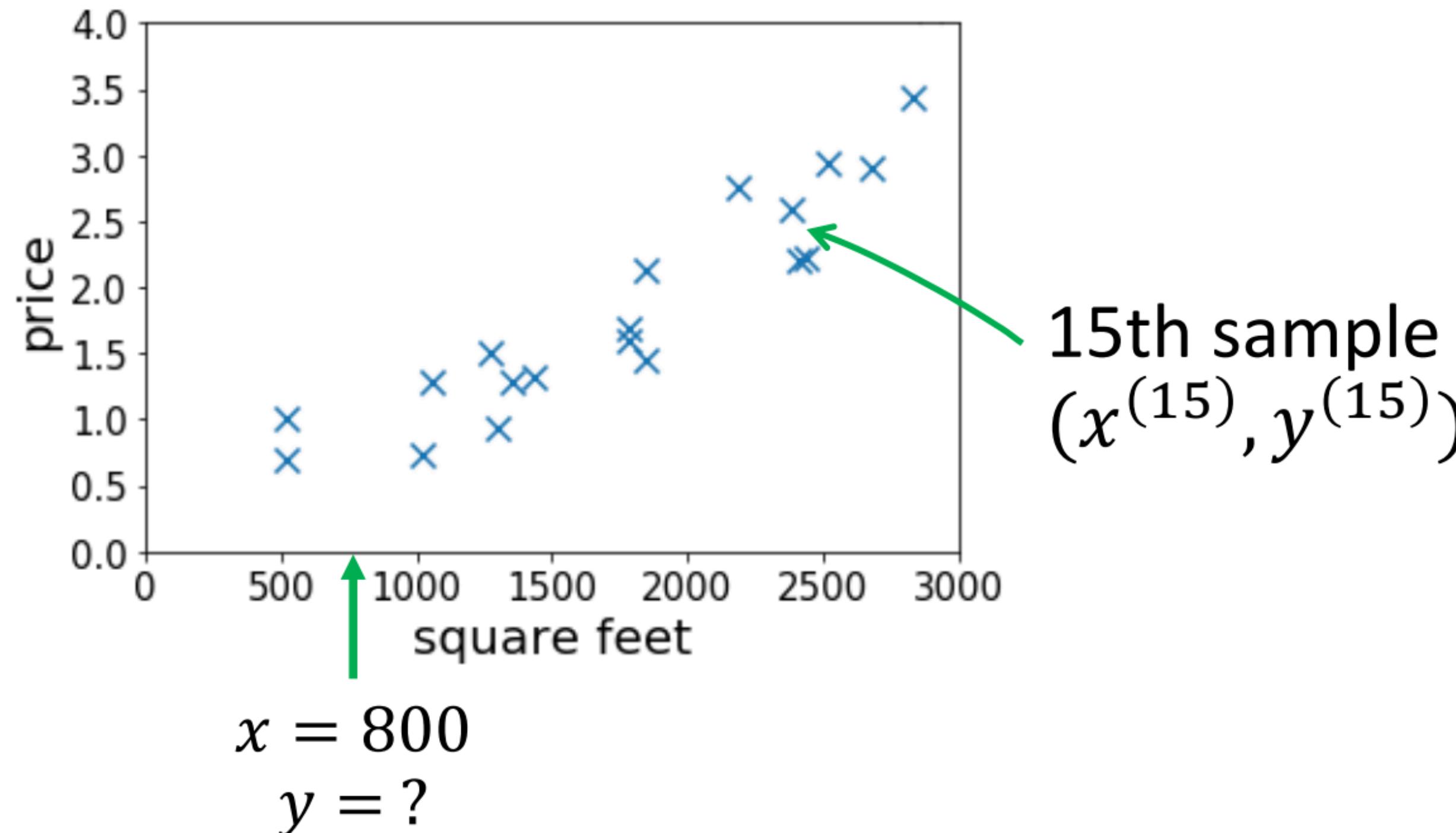
- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$



CAT

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$



Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$
- A training set is set of pairs $\{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$
s.t. $x^{(i)} \in \mathcal{X}$ and $y^{(i)} \in \mathcal{Y}$ for $i = 1, \dots, n$.

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$
- A training set is set of pairs $\{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$
s.t. $x^{(i)} \in \mathcal{X}$ and $y^{(i)} \in \mathcal{Y}$ for $i = 1, \dots, n$.
- Given a training set our goal is to produce a good prediction function h

Supervised Learning

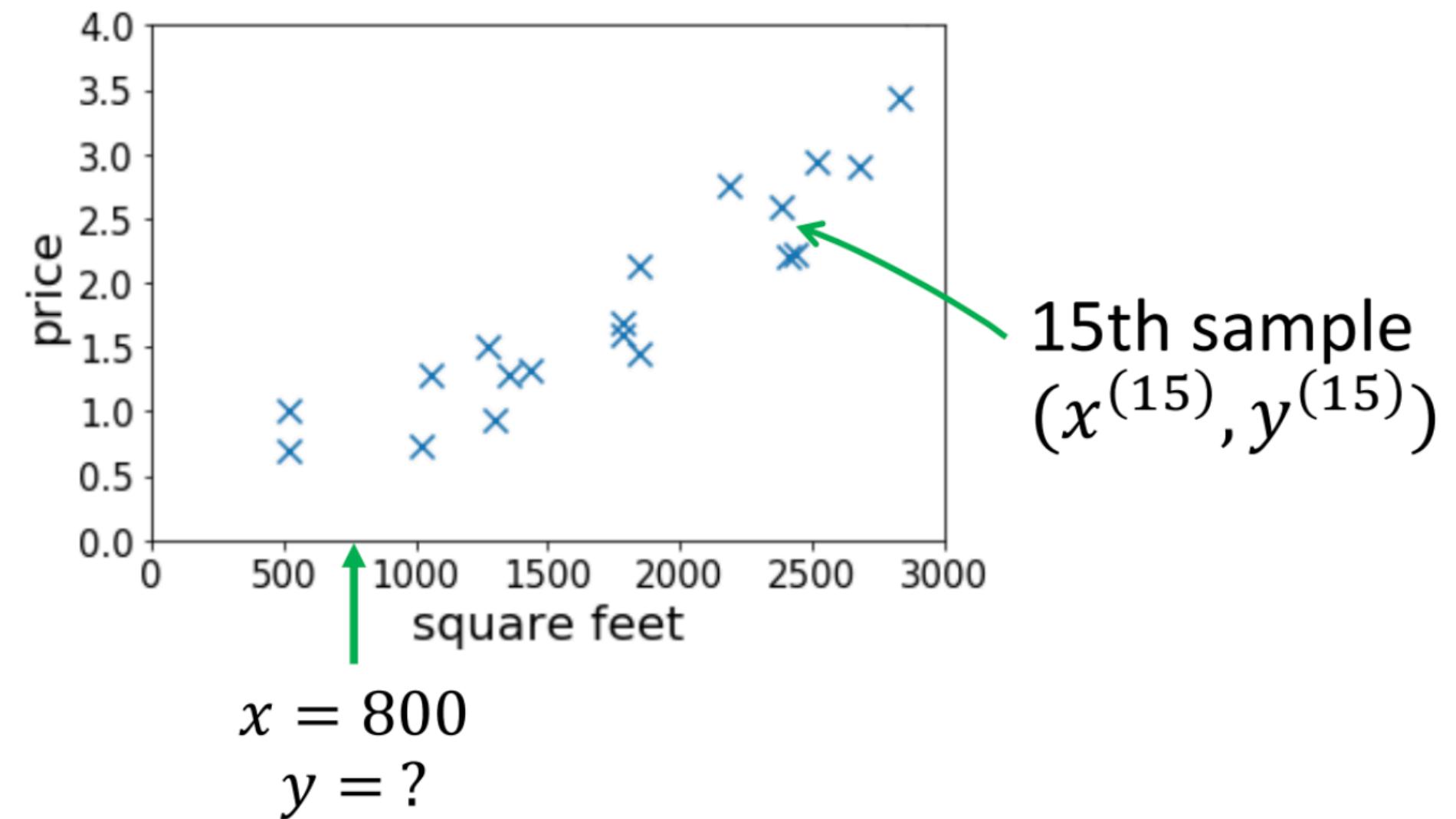
- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$
- A training set is set of pairs $\{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$
s.t. $x^{(i)} \in \mathcal{X}$ and $y^{(i)} \in \mathcal{Y}$ for $i = 1, \dots, n$.
- Given a training set our goal is to produce a good prediction function h
- If \mathcal{Y} is continuous, then called a regression problem
- If \mathcal{Y} is discrete, then called a classification problem

Supervised Learning

- How to define “good” for a prediction function?
 - Metrics / performance

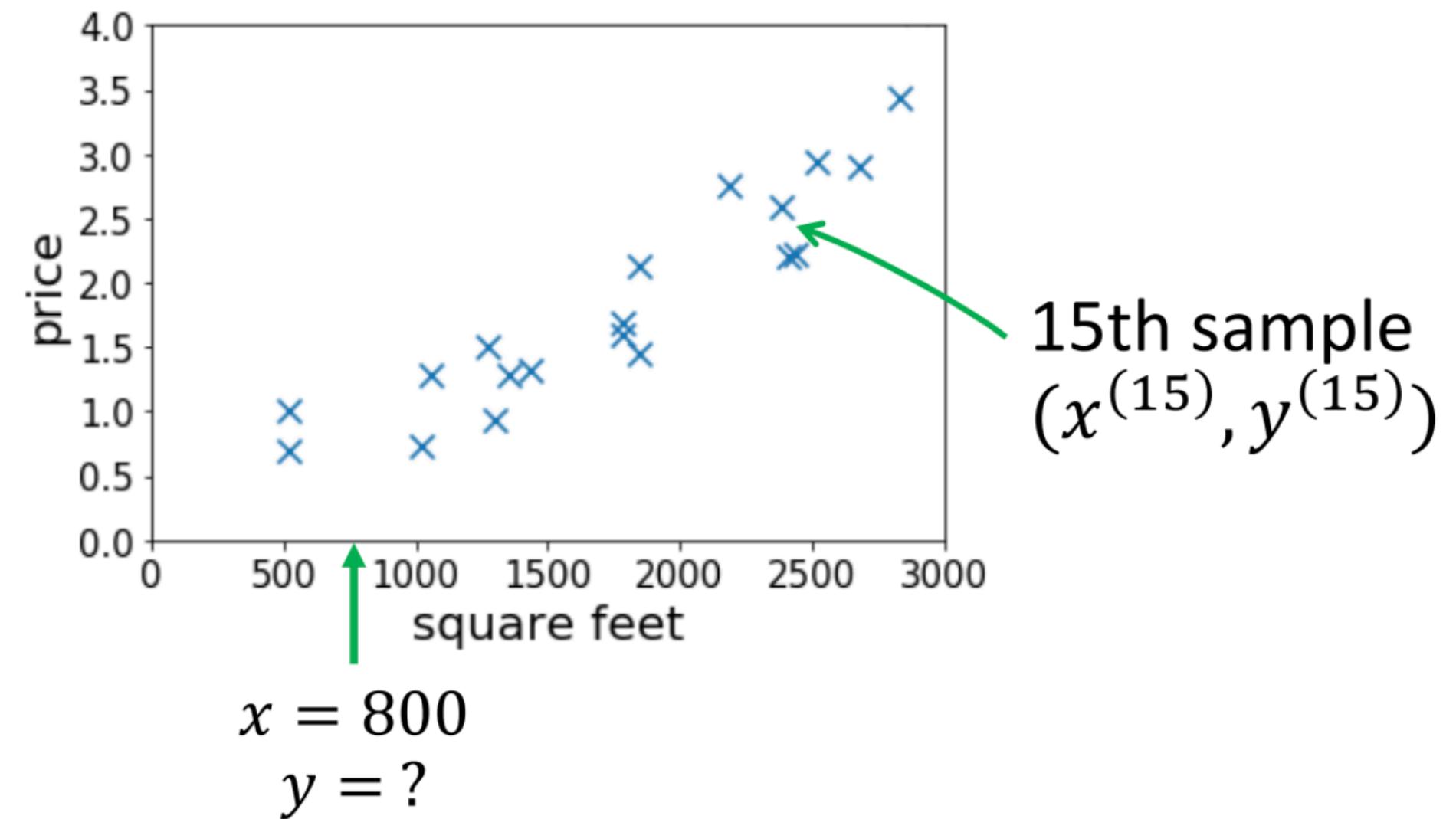
Supervised Learning

- How to define “good” for a prediction function?
 - Metrics / performance



Supervised Learning

- How to define “good” for a prediction function?
 - Metrics / performance

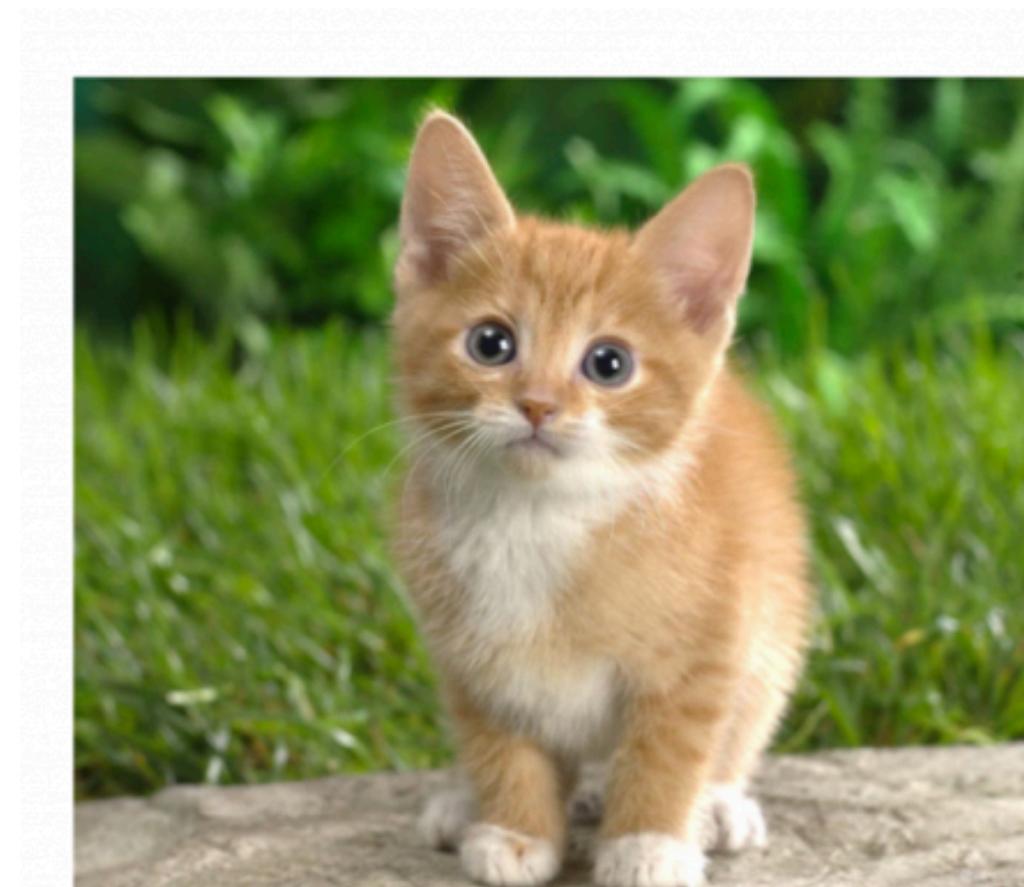
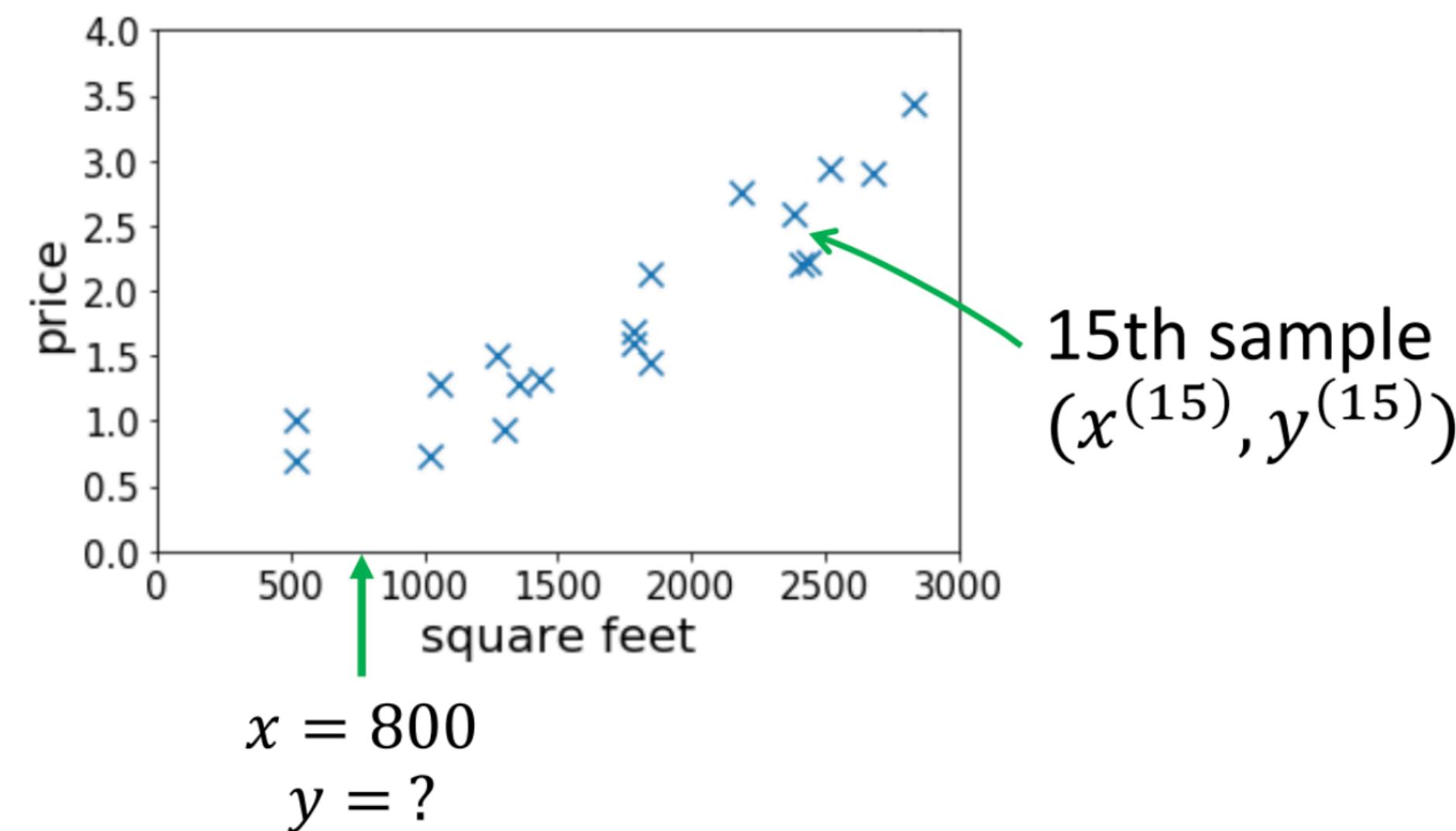


$$|\hat{y} - y^*|$$

\hat{y} is the prediction, y^* is the truth

Supervised Learning

- How to define “good” for a prediction function?
 - Metrics / performance



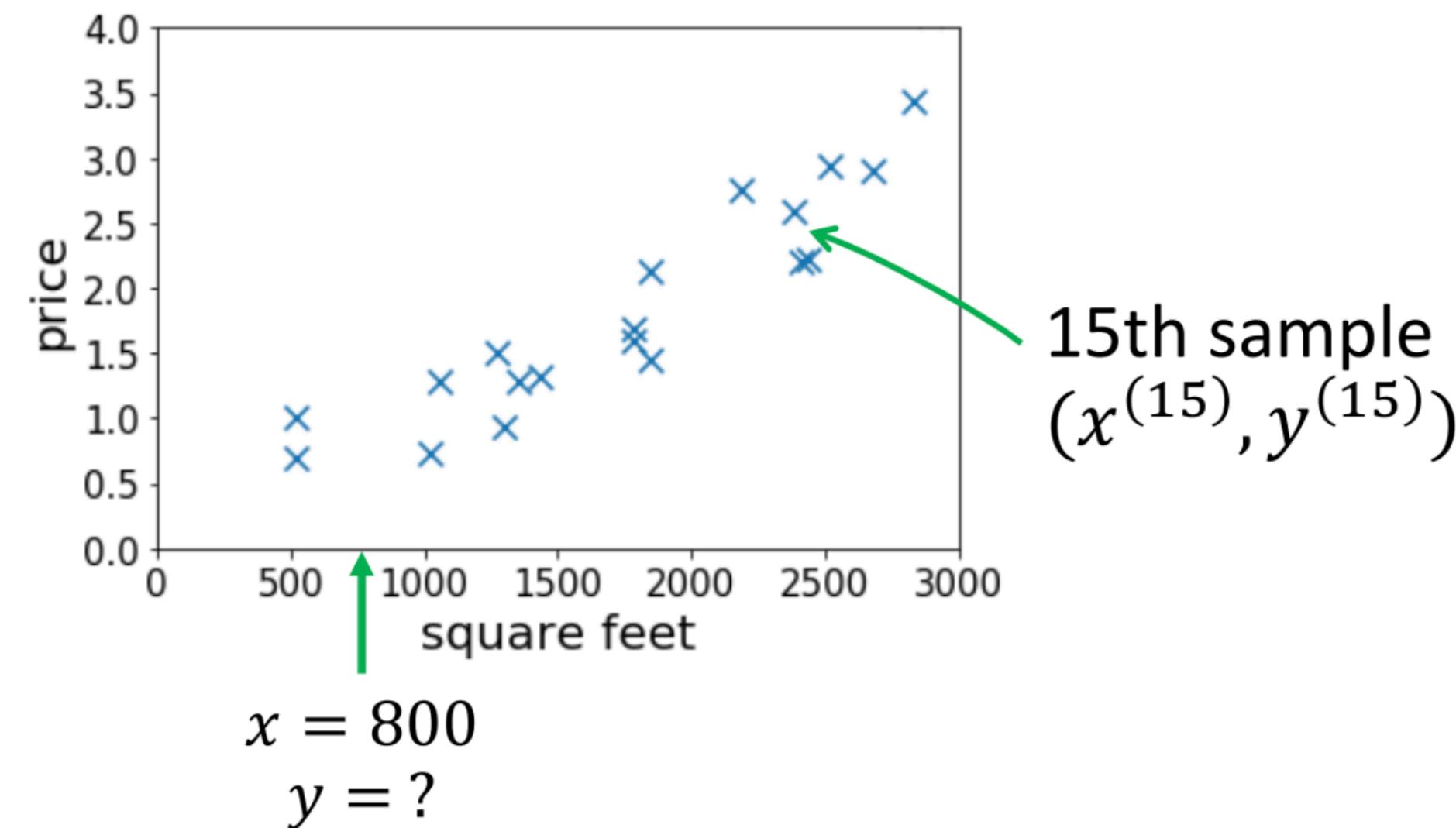
CAT

$$|\hat{y} - y^*|$$

\hat{y} is the prediction, y^* is the truth

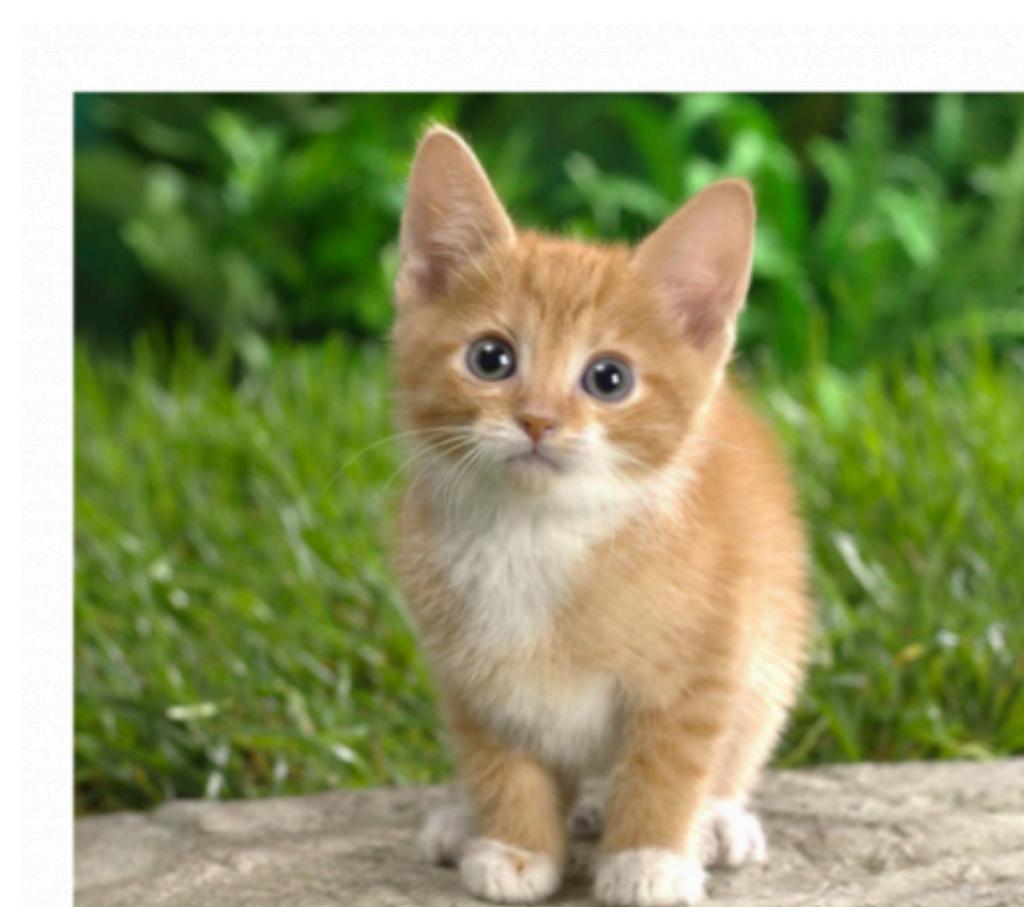
Supervised Learning

- How to define “good” for a prediction function?
 - Metrics / performance



$$|\hat{y} - y^*|$$

\hat{y} is the prediction, y^* is the truth



$$\mathbb{I}(\hat{y} = y^*) = \begin{cases} 1, & \hat{y} = y^* \\ 0 & \text{otherwise} \end{cases}$$

Supervised Learning

- How to define “good” for a prediction function?
 - Metrics / performance
 - Good on unseen data

Supervised Learning

- How to define “good” for a prediction function?

- Metrics / performance
- Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \dots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Supervised Learning

- How to define “good” for a prediction function?

- Metrics / performance
- Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \dots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset

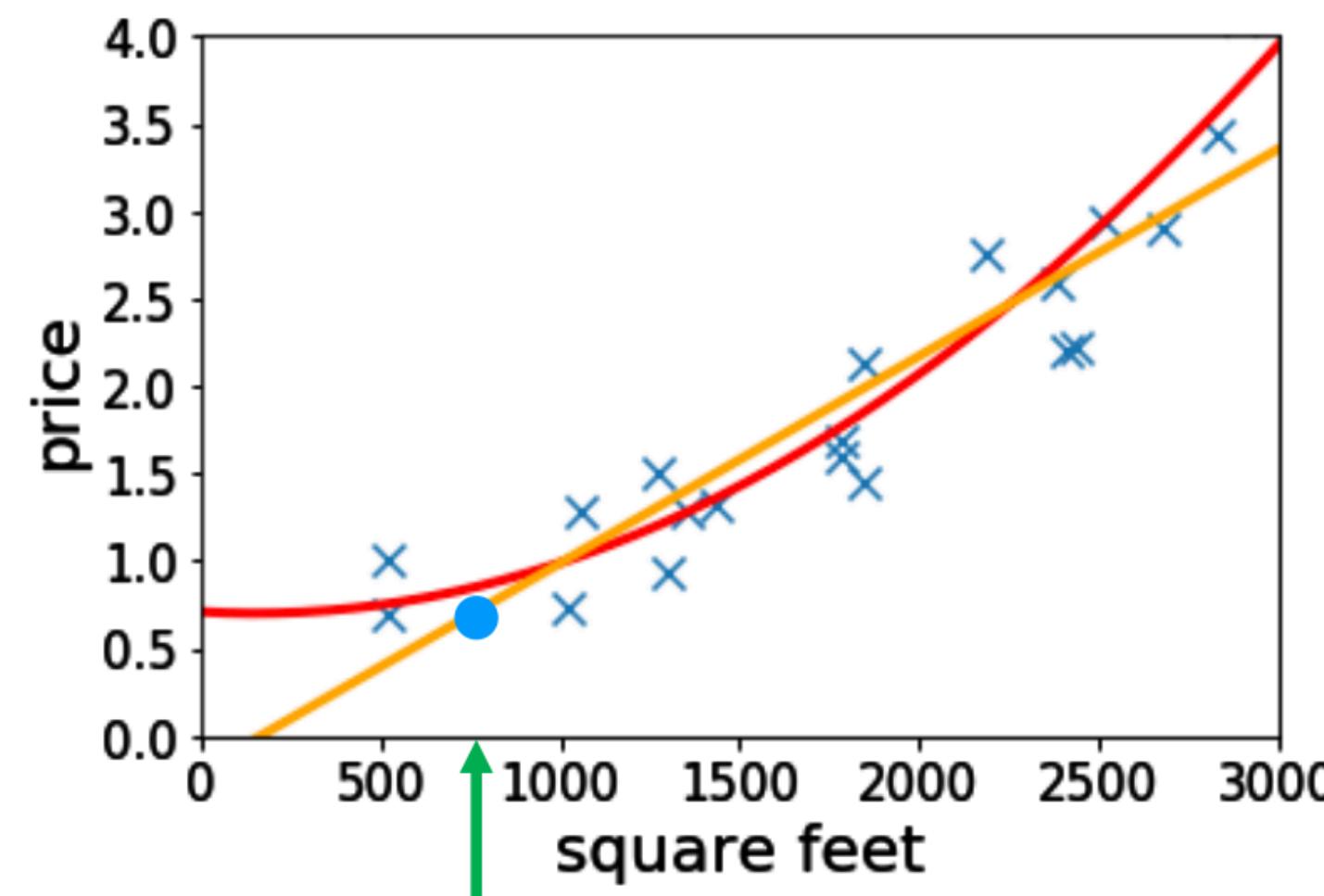
Supervised Learning

- How to define “good” for a prediction function?

- Metrics / performance
- Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \dots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset



Which curve to choose?

Supervised Learning

- How to define “good” for a prediction function?

- Metrics / performance
- Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \dots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset

Supervised Learning

- How to define “good” for a prediction function?

- Metrics / performance
- Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \dots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset

Test dataset is another set of pairs $\{(\tilde{x}^{(1)}, \tilde{y}^{(1)}), \dots, (\tilde{x}^{(L)}, \tilde{y}^{(L)})\}$

Does not overlap with training and validation dataset

Supervised Learning

- How to define “good” for a prediction function?

- Metrics / performance
- Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \dots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset

Test dataset is another set of pairs $\{(\tilde{x}^{(1)}, \tilde{y}^{(1)}), \dots, (\tilde{x}^{(L)}, \tilde{y}^{(L)})\}$

Does not overlap with training and validation dataset

Completely unseen before deployment

Realistic setting

Supervised Learning

- How to define “good” for a prediction function?

- Metrics / performance
- Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \dots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset

Test dataset is another set of pairs $\{(\tilde{x}^{(1)}, \tilde{y}^{(1)}), \dots, (\tilde{x}^{(L)}, \tilde{y}^{(L)})\}$

Does not overlap with training and validation dataset

Completely unseen before deployment

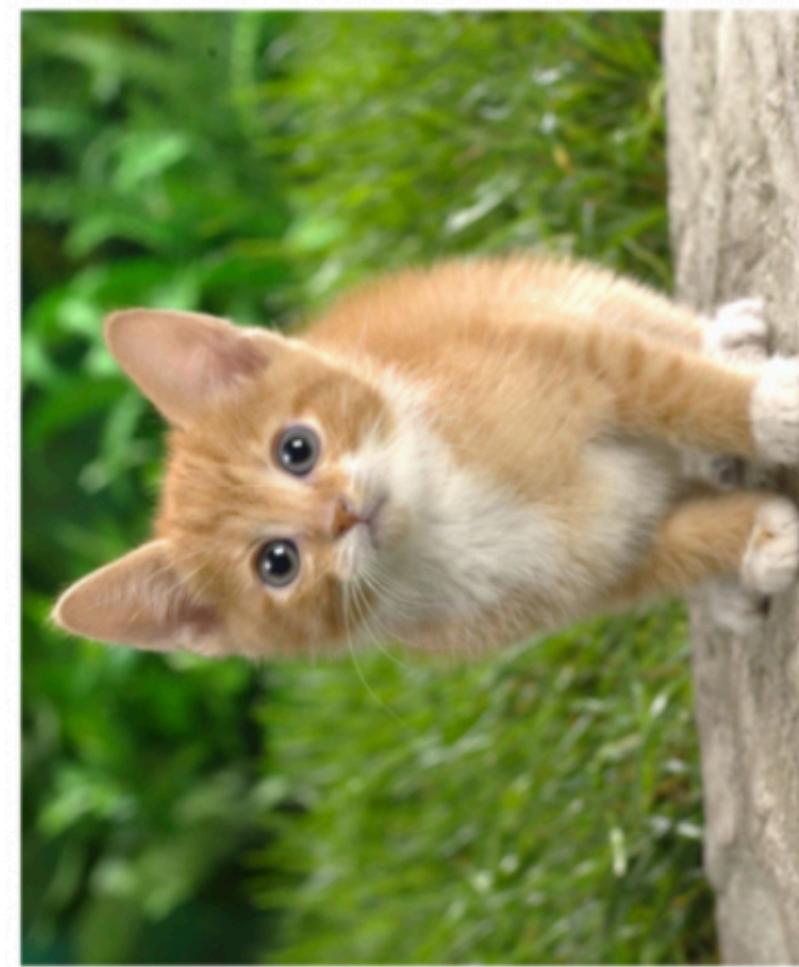
Realistic setting

Hyperparameter tuning is a form of training

Supervised Training



Train



Validation



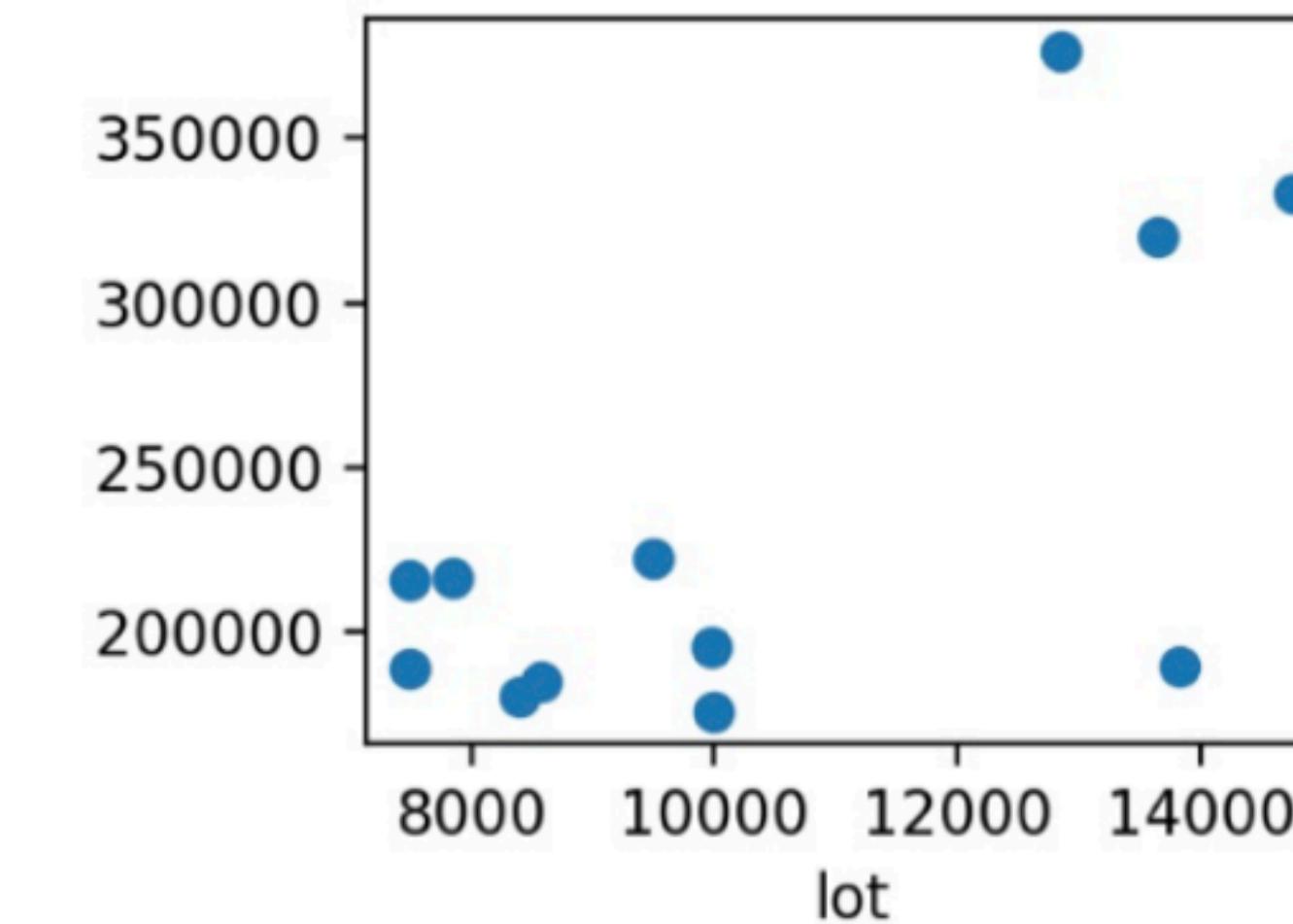
Test

Not only for supervised learning

Example: Regression using Housing Data

Example Housing Data

	SalePrice	Lot.Area
4	189900	13830
5	195500	9978
9	189000	7500
10	175900	10000
12	180400	8402
22	216000	7500
36	376162	12858
47	320000	13650
55	216500	7851
56	185088	8577



Represent h as a Linear Function

$h(x) = \theta_0 + \theta_1 x_1$ is an *affine function*

Popular choice

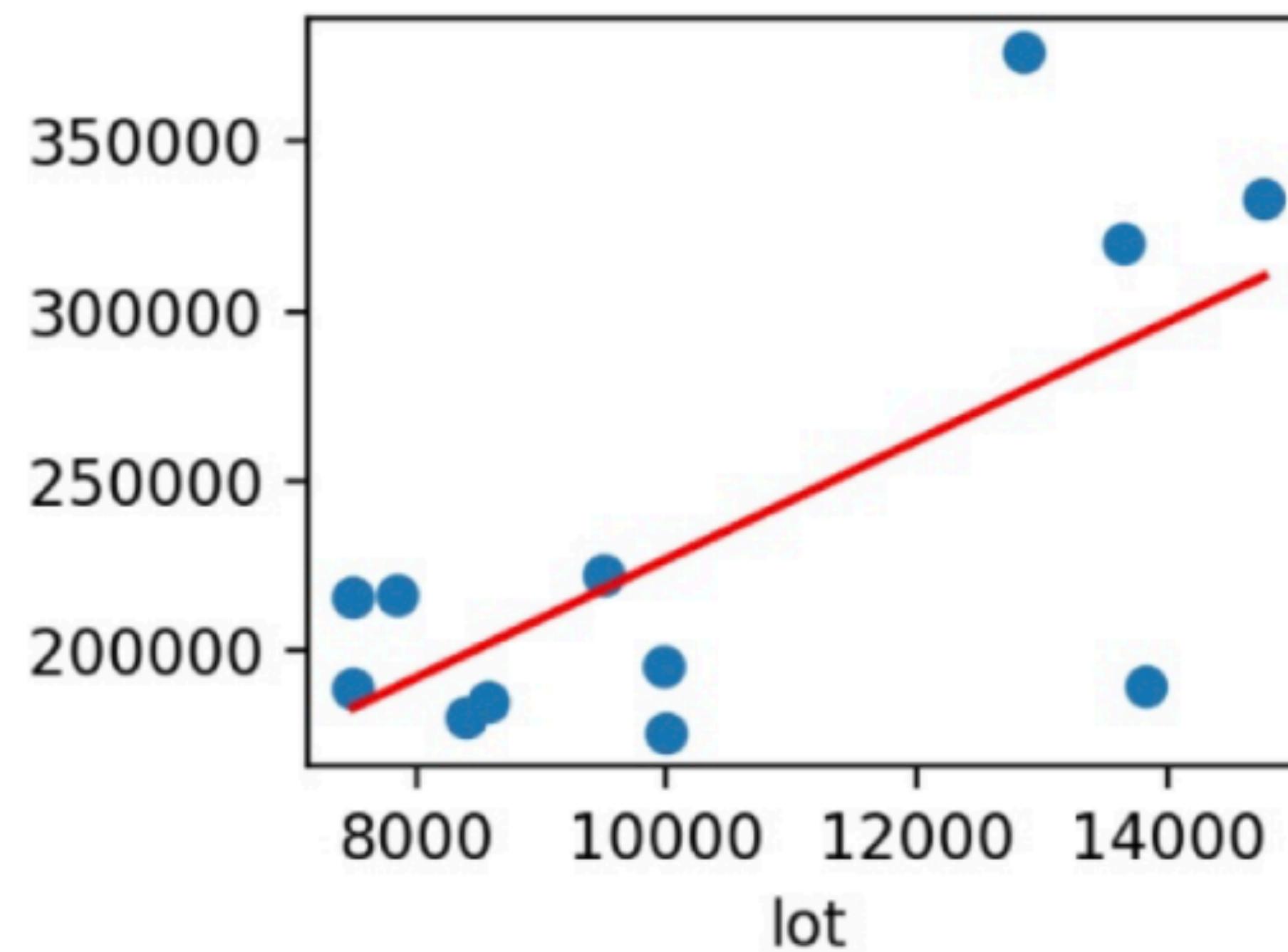
Represent h as a Linear Function

$h(x) = \theta_0 + \theta_1 x_1$ is an *affine function*

Popular choice

The function is defined by **parameters** θ_0 and θ_1 , the function space is greatly reduced

Simple Line Fit



More Features

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

More Features

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

What's a prediction here?

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3.$$

More Features

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

What's a prediction here?

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3.$$

With the convention that $x_0 = 1$ we can write:

$$h(x) = \sum_{j=0}^3 \theta_j x_j$$

Vector Notations

Vector Notations

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

We write the vectors as (important notation)

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \text{ and } x^{(1)} = \begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 2104 \\ 4 \\ 45 \end{pmatrix} \text{ and } y^{(1)} = 400$$

Vector Notations

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

We write the vectors as (important notation)

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \text{ and } x^{(1)} = \begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 2104 \\ 4 \\ 45 \end{pmatrix} \text{ and } y^{(1)} = 400$$

We call θ **parameters**, $x^{(i)}$ is the input or the **features**, and the output or **target** is $y^{(i)}$. To be clear,

(x, y) is a training example and $(x^{(i)}, y^{(i)})$ is the i^{th} example.

Vector Notations

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

We write the vectors as (important notation)

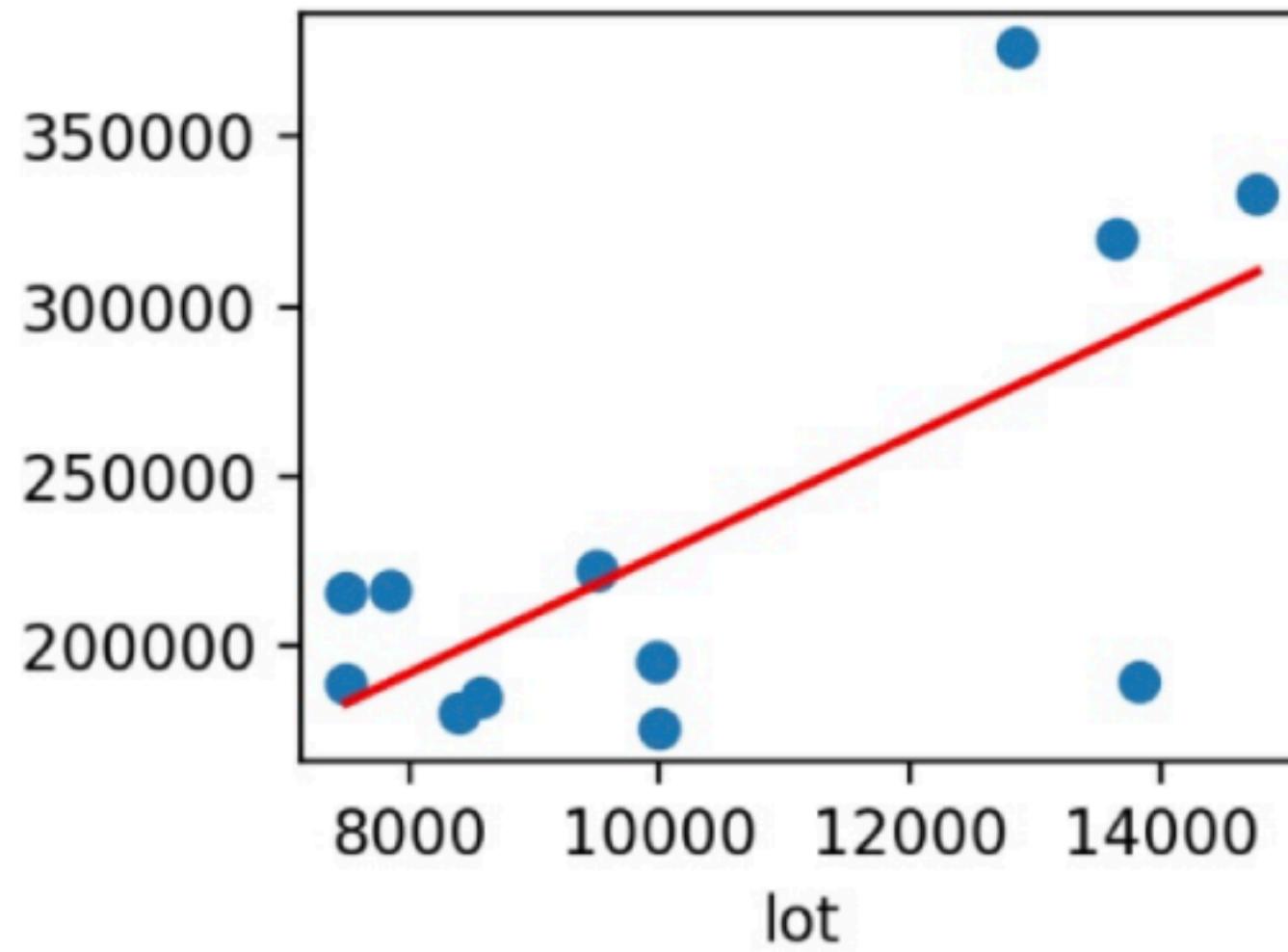
$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \text{ and } x^{(1)} = \begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 2104 \\ 4 \\ 45 \end{pmatrix} \text{ and } y^{(1)} = 400$$

We call θ **parameters**, $x^{(i)}$ is the input or the **features**, and the output or **target** is $y^{(i)}$. To be clear,

(x, y) is a training example and $(x^{(i)}, y^{(i)})$ is the i^{th} example.

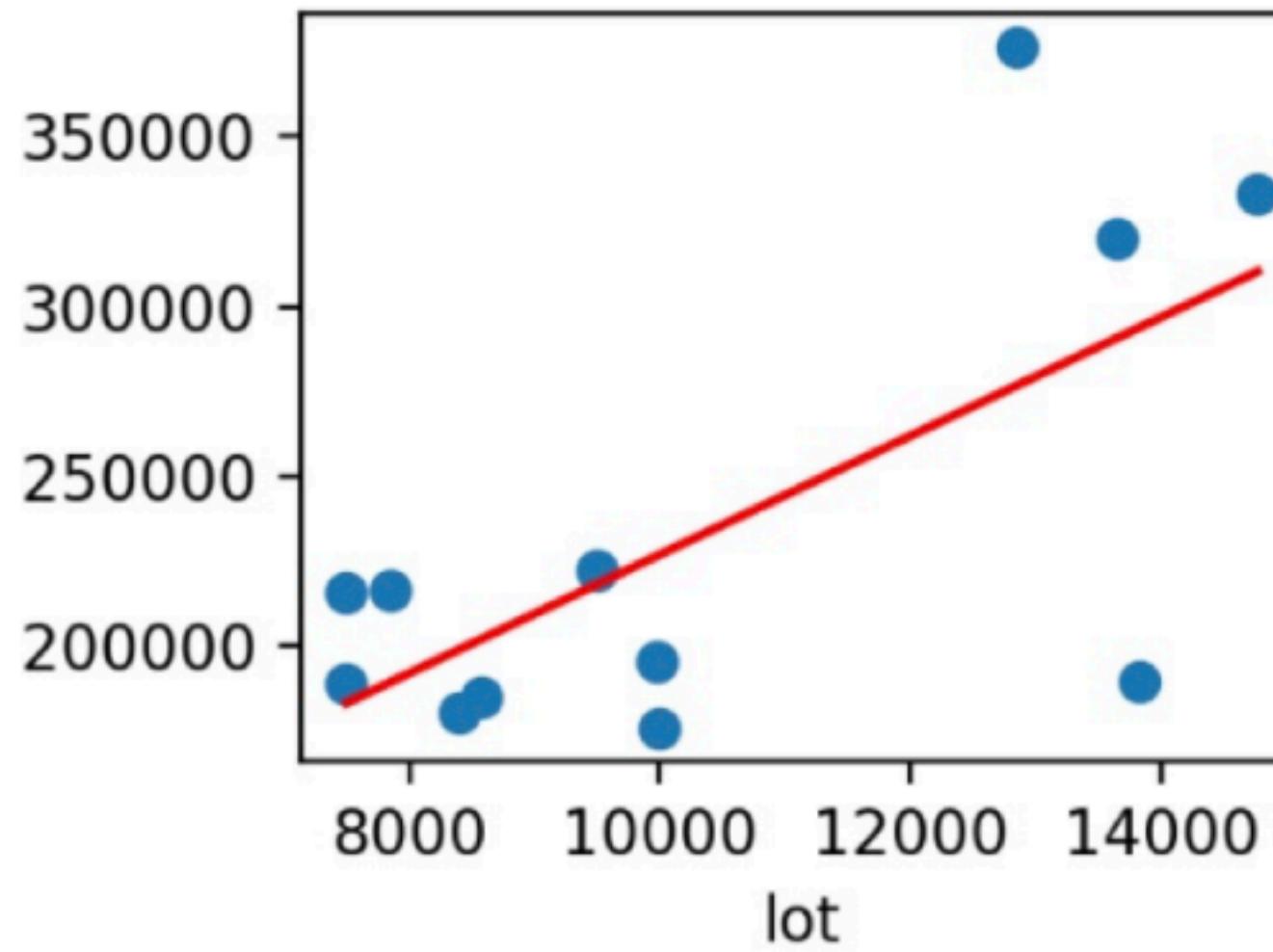
We have n examples. There are d features. $x^{(i)}$ and θ are $d+1$ dimensional (since $x_0 = 1$)

Vector Notation of Prediction



$$h_{\theta}(x) = \sum_{j=0}^d \theta_j x_j = x^T \theta$$

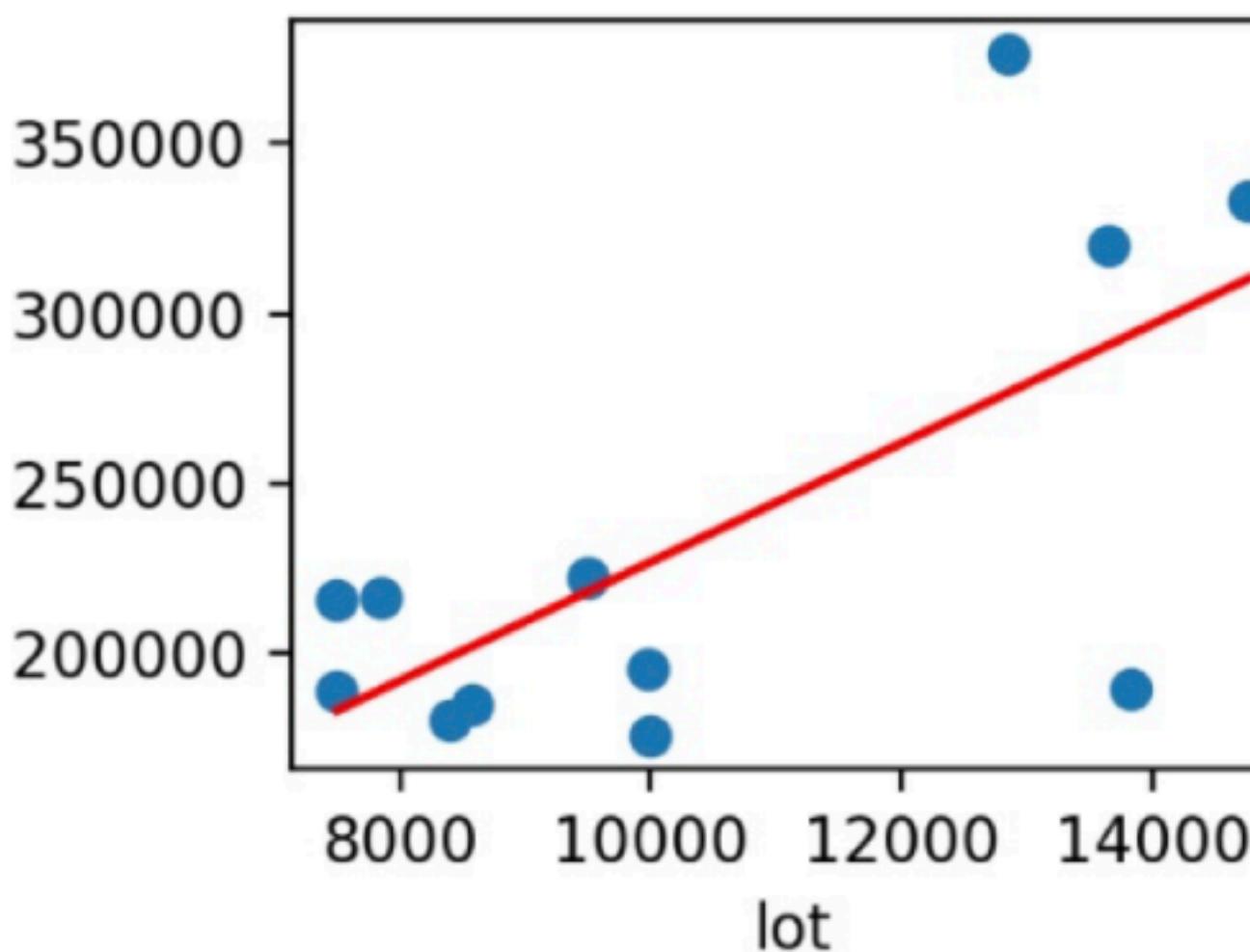
Vector Notation of Prediction



$$h_{\theta}(x) = \sum_{j=0}^d \theta_j x_j = x^T \theta$$

We want to choose θ so that $h_{\theta}(x) \approx y$

Loss Function



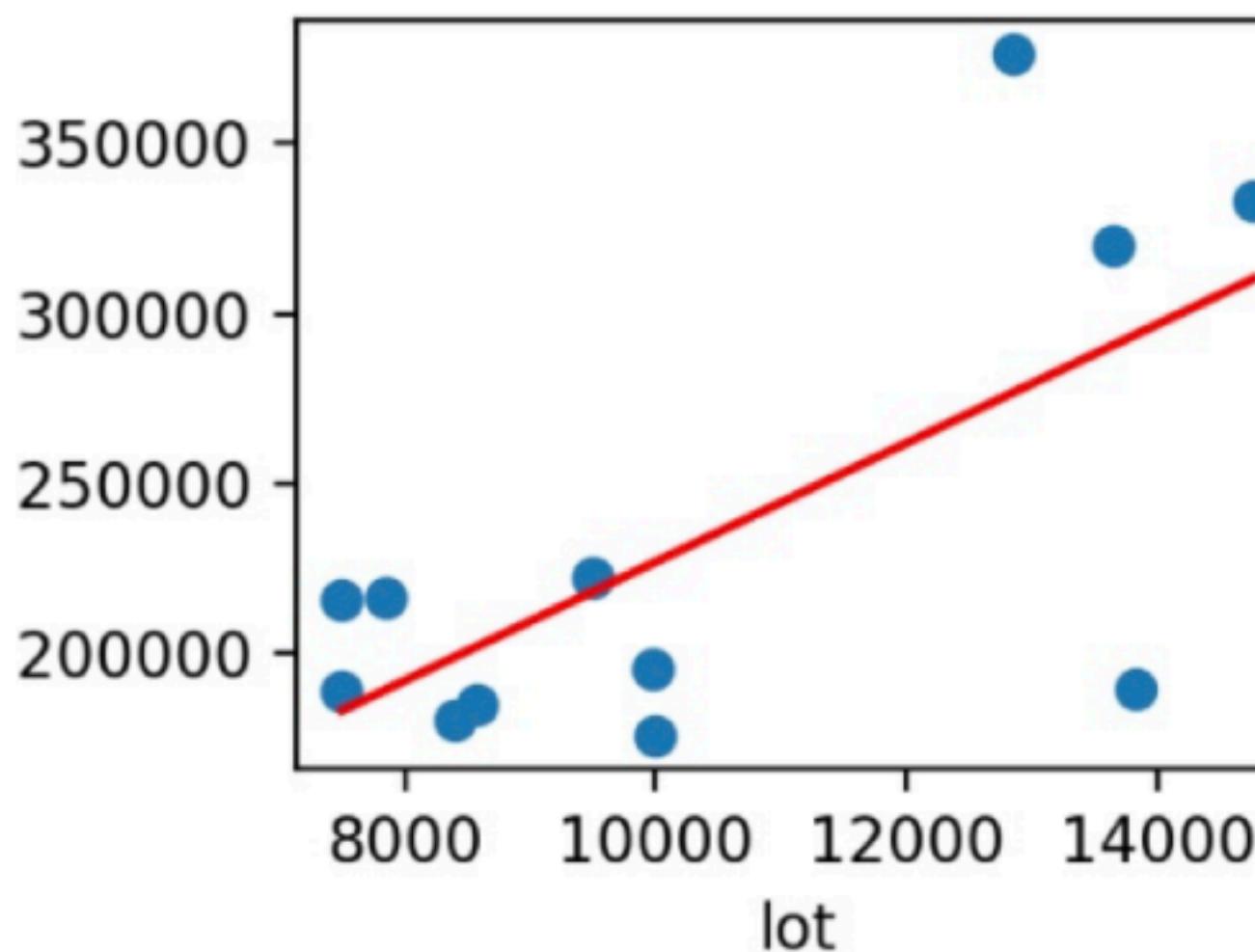
$$h_{\theta}(x) = \sum_{j=0}^d \theta_j x_j = x^T \theta$$

We want to choose θ so that $h_{\theta}(x) \approx y$



How to quantify the deviation of $h_{\theta}(x)$ from y

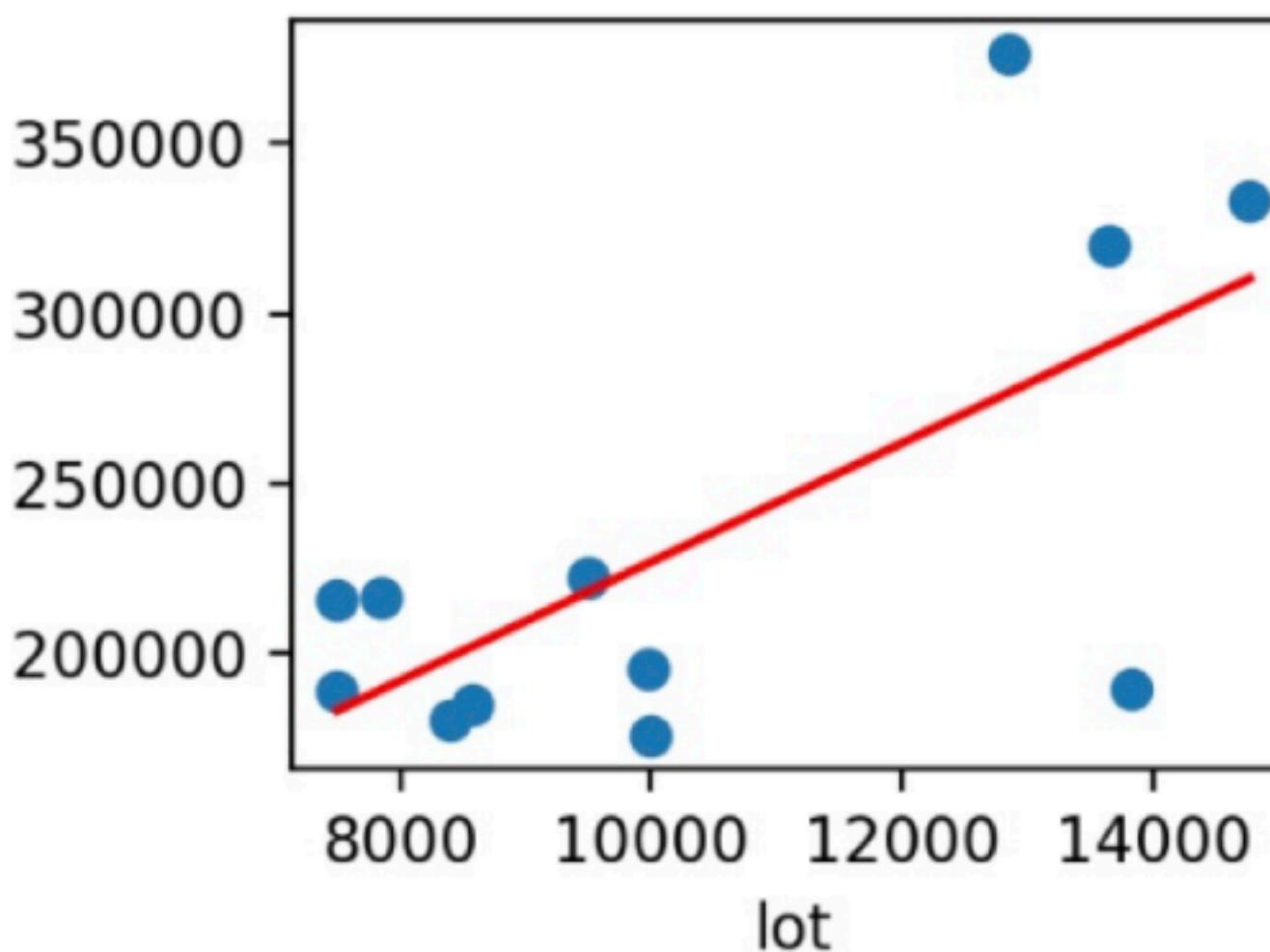
Least Squares



$$h_{\theta}(x) = \sum_{j=0}^d \theta_j x_j = x^T \theta$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n \left(h_{\theta}(x^{(i)}) - y^{(i)} \right)^2$$

Least Squares



$$h_{\theta}(x) = \sum_{j=0}^d \theta_j x_j = x^T \theta$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n \left(h_{\theta}(x^{(i)}) - y^{(i)} \right)^2$$

Choose

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

Solving Least Square Problem

Direct Minimization

$$h_{\theta}(x) = \sum_{j=0}^d \theta_j x_j = x^T \theta$$
$$J(\theta) = \frac{1}{2} \sum_{i=1}^n \left(h_{\theta}(x^{(i)}) - y^{(i)} \right)^2$$

Choose

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

Solving Least Square Problem

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \nabla_{\theta} \frac{1}{2} (\vec{X}\theta - \vec{y})^T (\vec{X}\theta - \vec{y}) \\&= \frac{1}{2} \nabla_{\theta} \left((\vec{X}\theta)^T \vec{X}\theta - (\vec{X}\theta)^T \vec{y} - \vec{y}^T (\vec{X}\theta) + \vec{y}^T \vec{y} \right) \\&= \frac{1}{2} \nabla_{\theta} \left(\theta^T (\vec{X}^T \vec{X}) \theta - \vec{y}^T (\vec{X}\theta) - \vec{y}^T (\vec{X}\theta) \right) \\&= \frac{1}{2} \nabla_{\theta} \left(\theta^T (\vec{X}^T \vec{X}) \theta - 2(\vec{X}^T \vec{y})^T \theta \right) \\&= \frac{1}{2} (2\vec{X}^T \vec{X}\theta - 2\vec{X}^T \vec{y}) \\&= \vec{X}^T \vec{X}\theta - \vec{X}^T \vec{y}\end{aligned}$$

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

Solving Least Square Problem

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \nabla_{\theta} \frac{1}{2} (\vec{X}\theta - \vec{y})^T (\vec{X}\theta - \vec{y}) \\&= \frac{1}{2} \nabla_{\theta} \left((\vec{X}\theta)^T \vec{X}\theta - (\vec{X}\theta)^T \vec{y} - \vec{y}^T (\vec{X}\theta) + \vec{y}^T \vec{y} \right) \\&= \frac{1}{2} \nabla_{\theta} \left(\theta^T (\vec{X}^T \vec{X}) \theta - \vec{y}^T (\vec{X}\theta) - \vec{y}^T (\vec{X}\theta) \right) \\&= \frac{1}{2} \nabla_{\theta} \left(\theta^T (\vec{X}^T \vec{X}) \theta - 2(\vec{X}^T \vec{y})^T \theta \right) \\&= \frac{1}{2} (2\vec{X}^T \vec{X}\theta - 2\vec{X}^T \vec{y}) \\&= \vec{X}^T \vec{X}\theta - \vec{X}^T \vec{y}\end{aligned}$$

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

Solving Least Square Problem

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \nabla_{\theta} \frac{1}{2} (\vec{X}\theta - \vec{y})^T (\vec{X}\theta - \vec{y}) \\&= \frac{1}{2} \nabla_{\theta} \left((\vec{X}\theta)^T \vec{X}\theta - (\vec{X}\theta)^T \vec{y} - \vec{y}^T (\vec{X}\theta) + \vec{y}^T \vec{y} \right) \\&= \frac{1}{2} \nabla_{\theta} \left(\theta^T (\vec{X}^T \vec{X}) \theta - \vec{y}^T (\vec{X}\theta) - \vec{y}^T (\vec{X}\theta) \right) \\&= \frac{1}{2} \nabla_{\theta} \left(\theta^T (\vec{X}^T \vec{X}) \theta - 2(\vec{X}^T \vec{y})^T \theta \right) \\&= \frac{1}{2} (2\vec{X}^T \vec{X}\theta - 2\vec{X}^T \vec{y}) \\&= \vec{X}^T \vec{X}\theta - \vec{X}^T \vec{y}\end{aligned}$$

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

When is $\vec{X}^T \vec{X}$ invertible? What if it is not invertible?

Thank You!
Q & A