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

COMP 5212
Machine Learning

Math Basics

(Largely adapted from Stanford CS229 Slides)

Junxian He

Sep 5, 2024

Course Quota

Quota has been increased to 82, currently 90+ on waiting list

Attendance Quiz APP Download

HKUST iLearn

The Hong Kong University of Science and Technology

10K+ Downloads | E Everyone

Install | Share | Add to wishlist

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.

SFQ
Allows you to complete the Student Feedback Questionnaire for all your courses at HKUST on the move.

iPRS
Enables you to quickly respond to questions or polls created by your instructor in class.

Linear Independence

A set of vectors $\{x_1, x_2, \ldots x_n\} \subset \mathbb{R}^m$ 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 $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$; otherwise, the vectors are ***(linearly) independent***.

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 $\mathrm{rank}(A)$.

Properties of Rank

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

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

- For $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{p \times n}$, $\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B))$.

- For $A, B \in \mathbb{R}^{m \times n}$, $\mathrm{rank}(A + B) \leq \mathrm{rank}(A) + \mathrm{rank}(B)$.

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

- 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

- 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,

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

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

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

Range

- The ***range*** or the column space of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(A)$, is the the span of the columns of A. In other words,

$$\mathcal{R}(A) = \{v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n\}.$$

- Assuming A is full rank and $n < m$, the projection of a vector $y \in \mathbb{R}^m$ onto the range of A is given by,

$$\mathrm{Proj}(y; A) = \mathrm{argmin}_{v \in \mathcal{R}(A)} \|v - y\|_2.$$

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 ith row and jth column from A.
The general (recursive) formula for the determinant is

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

Determinant: Example

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

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

The Determinant

The ***determinant*** of a square matrix $A \in \mathbb{R}^{n\times n}$, is a function $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$, 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:

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

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

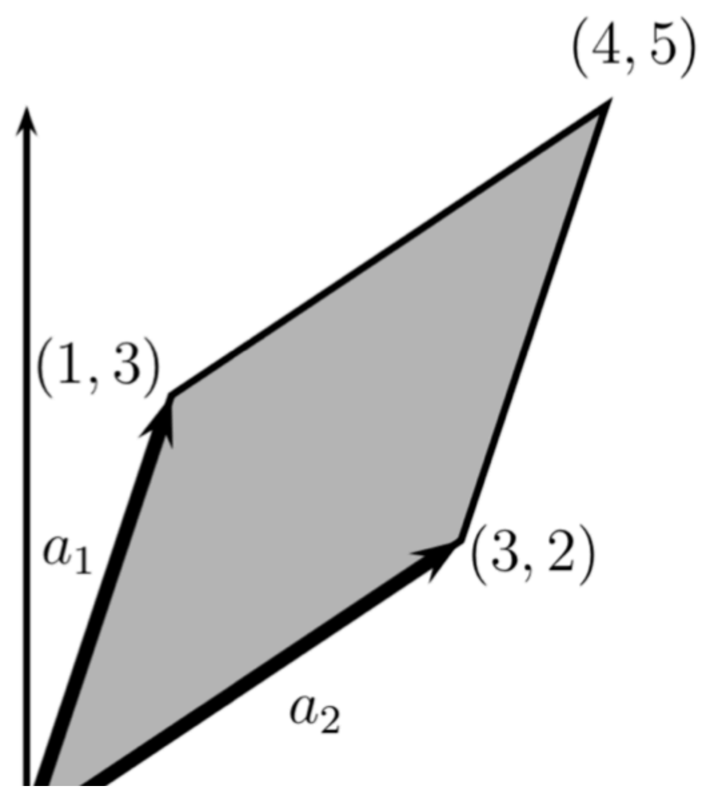
The Determinant

For example, consider the 2×2 matrix,

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

Here, the rows of the matrix are

$$a_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad a_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$



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

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.

Gradient over Matrix

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ 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:

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

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

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

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

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ 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,

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

Note that the Hessian is always symmetric, since

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

Gradients of Linear Functions

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 $\partial/(\partial x)\ ax = a$.

Common Gradient Formula

- $\nabla_x b^T x = b$
- $\nabla_x^2 b^T x = 0$
- $\nabla_x x^T A x = 2Ax$ (if A symmetric)
- $\nabla_x^2 x^T A x = 2A$ (if A symmetric)

Least Squares

- 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 $\|Ax - b\|_2^2$.

Outline

- Linear Algebra Review

- Probability Review

Basic Concepts

- Performing an **experiment** $\rightarrow$ **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

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

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:

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

Equally Likely Outcomes: Define S as a sample space with equally likely outcomes. Then
$P(E) = \dfrac{|E|}{|S|}$

Conditional Probability and Bayes' Rule

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

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

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

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

$$= \boxed{\frac{P(B)P(A \mid B)}{P(A)}}$$

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

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

Law of Total Probability

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

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

We can then write Bayes' Rule as:

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

Chain Rule

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

$$
\begin{aligned}
&P(A_1 \cap A_2 \cap \ldots \cap A_n) \\
&\quad = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_2 \cap A_1) \ldots P(A_n \mid A_{n-1} \cap A_{n-2} \cap \ldots \cap A_1)
\end{aligned}
$$

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 $A \perp B$,

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

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

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

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

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. $\mathbb{R} \to [0,1]$)

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

A CDF must fulfill the following:

- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- 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

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

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

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 $f_1(X), \ldots, f_n(X)$,

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

Example

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?

Variance

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

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

Interpretation: $Var(X)$ is the expected deviation of X from $\mathbb{E}[X]$.
Properties: For any constant $a \in \mathbb{R}$, real-valued function $f(X)$

- $Var[a] = 0$
- $Var[af(X)] = a^2 Var[f(X)]$

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 $k = 0, 1, ..., n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, \ldots$	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 PMF** for discrete RV's X, Y:

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

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

- **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 $X_1, ..., X_n$:

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

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

- **Marginal PDF** of X_1, given joint PDF of $X_1, ..., X_n$:

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

Expectation for multiple random variables

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

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

Covariance

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

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

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

Variance of two variables

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

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 $X_1, \ldots, X_n$:

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

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

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 $g : \mathbb{R}^n \to \mathbb{R}^m$, 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

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.

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

applying linearity of expectation and the fact that $Cov[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$, 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 $\Sigma = diag(Var[X_1], ..., Var[X_n])$

Multivariate Gaussian

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

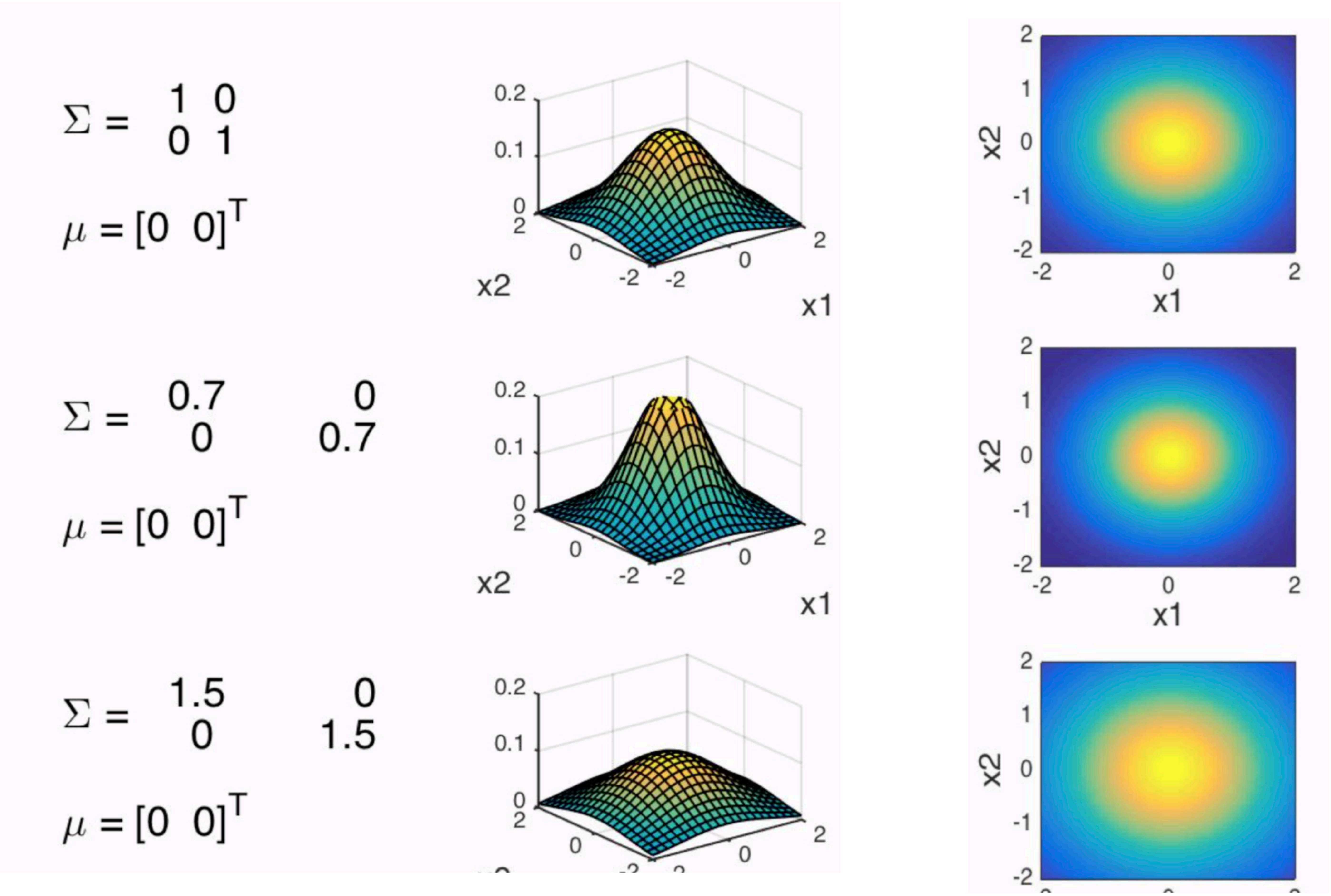
$$p(x; \mu, \Sigma) = \frac{1}{det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

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 $\Sigma = Var[X_1] = \sigma^2$, and so $\Sigma^{-1} = \frac{1}{\sigma^2}$ and $det(\Sigma)^{\frac{1}{2}} = \sigma$

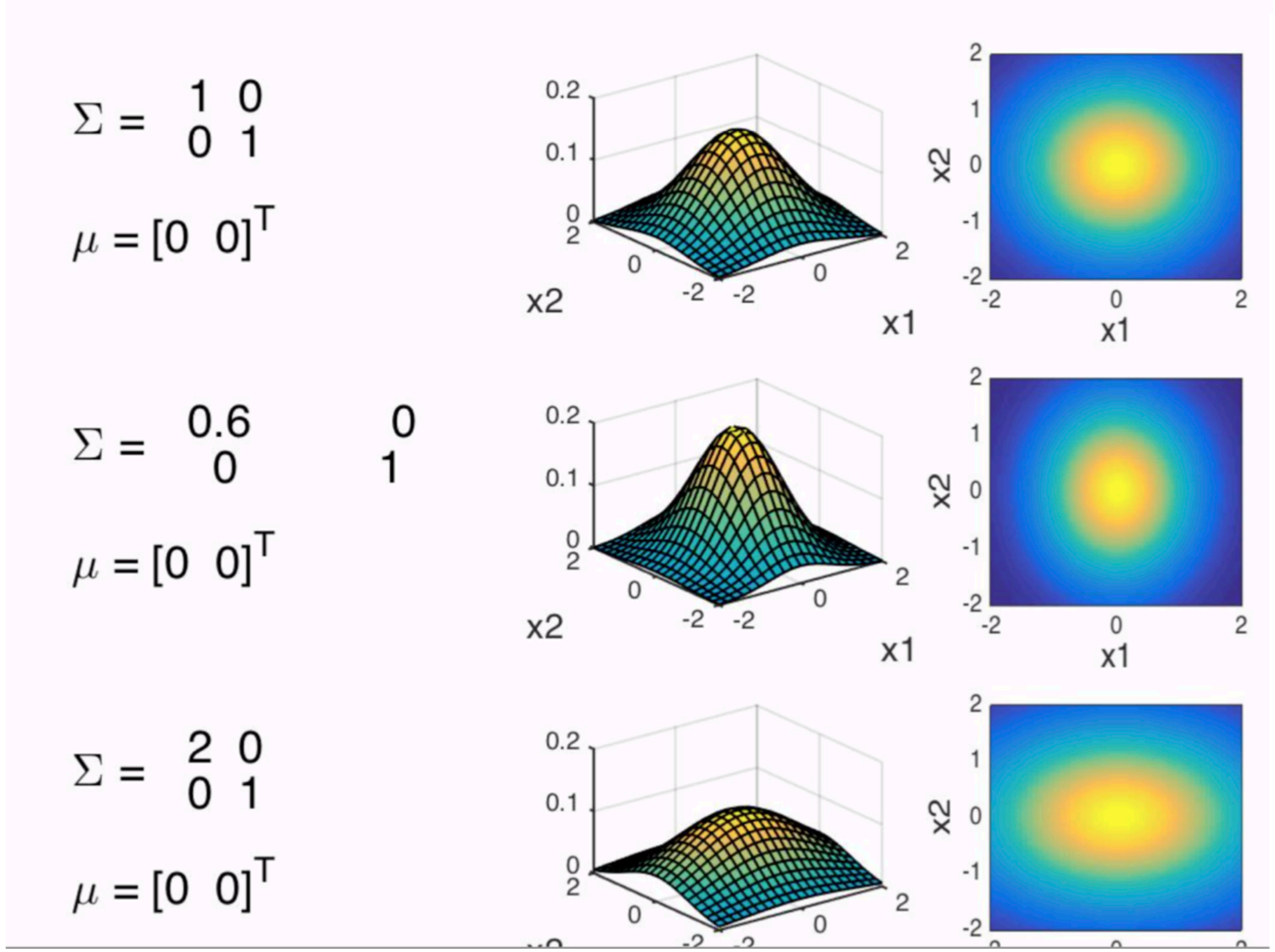
MV Gaussian Visualization



Effect of changing variance

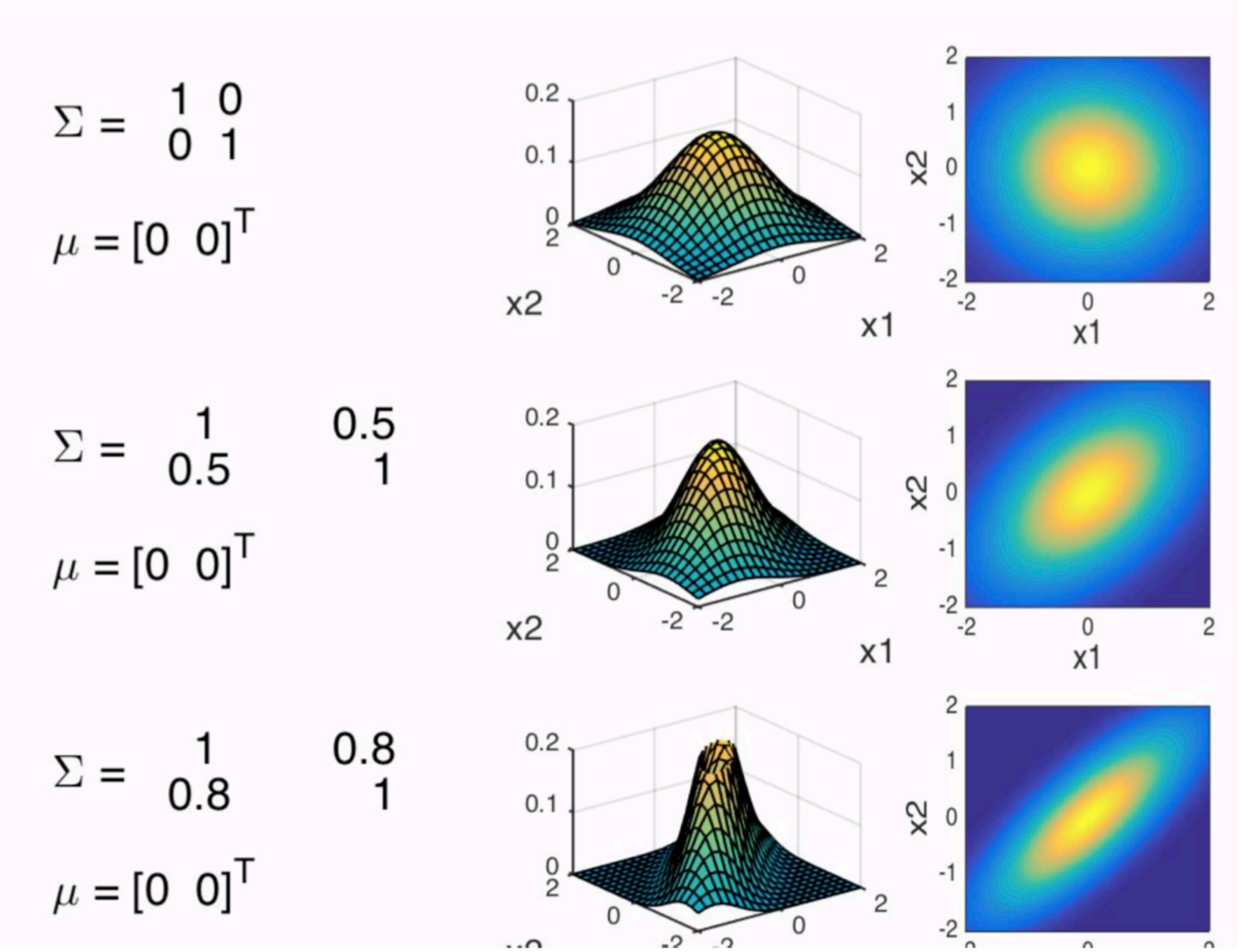
MV Gaussian Visualization

If $Var[X_1] \neq Var[X_2]$:



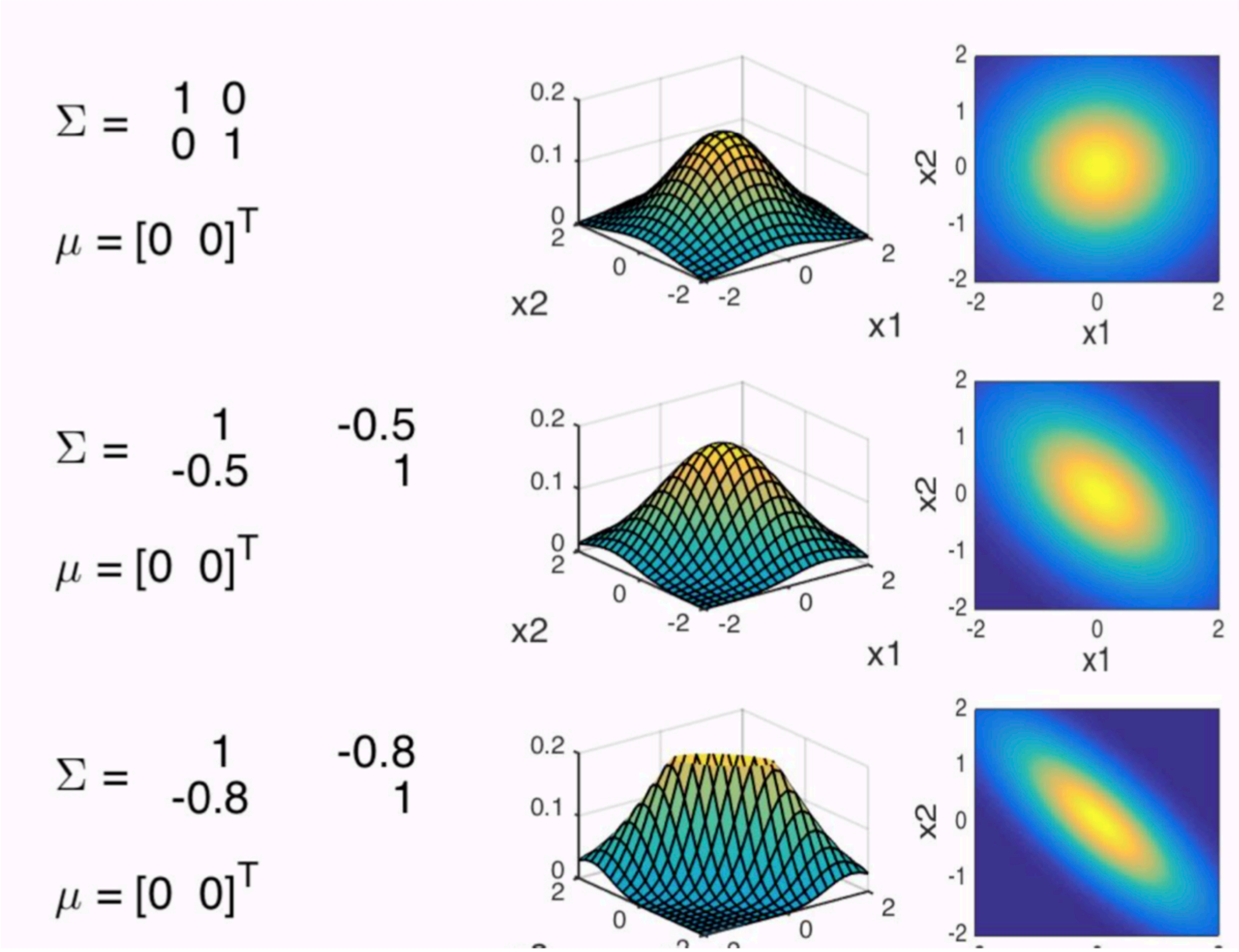
MV Gaussian Visualization

If X_1 and X_2 are positively correlated:



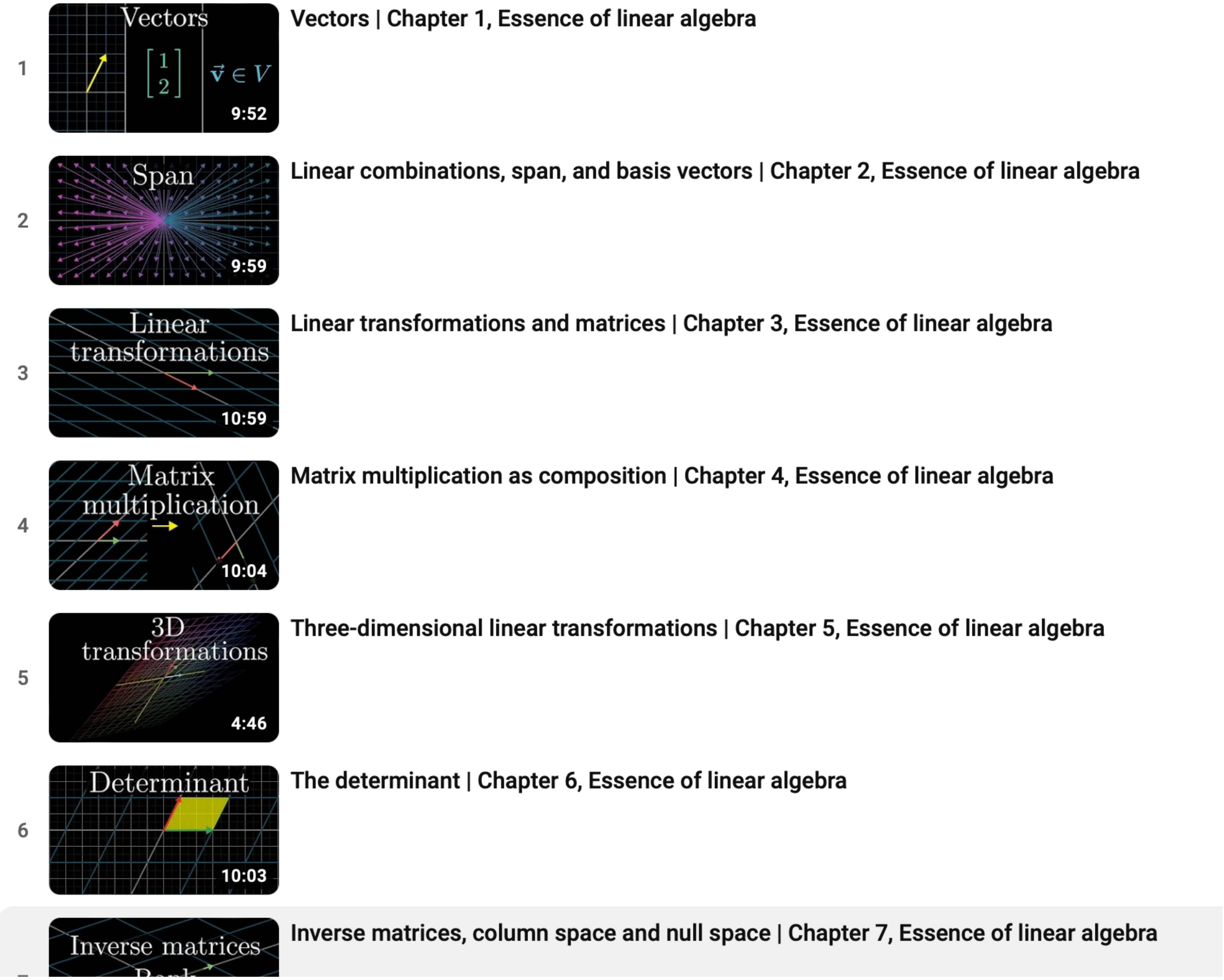
MV Gaussian Visualization

If X_1 and X_2 are negatively correlated:

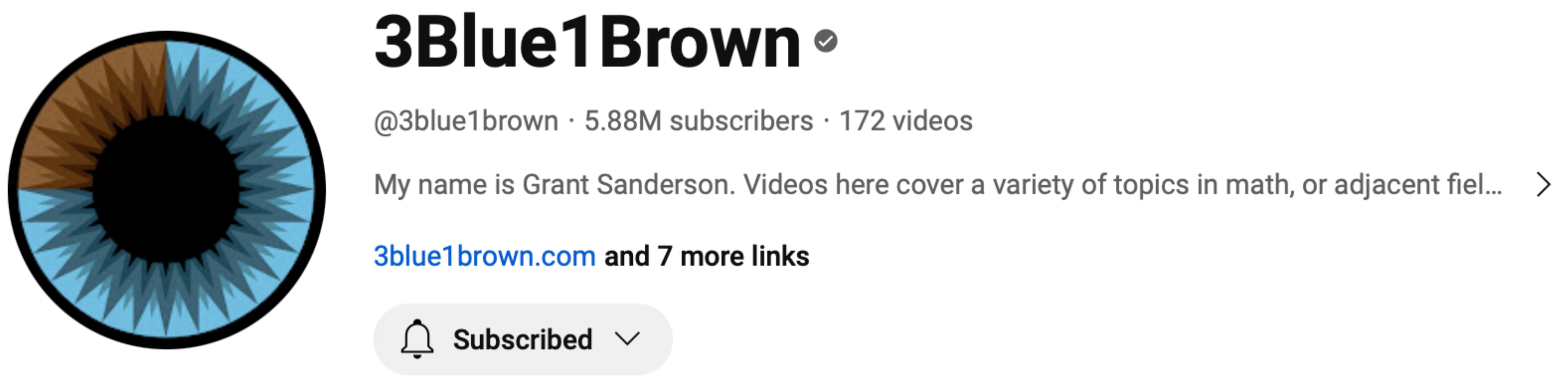


The purpose of computation is insight, not numbers.

- Richard Hamming



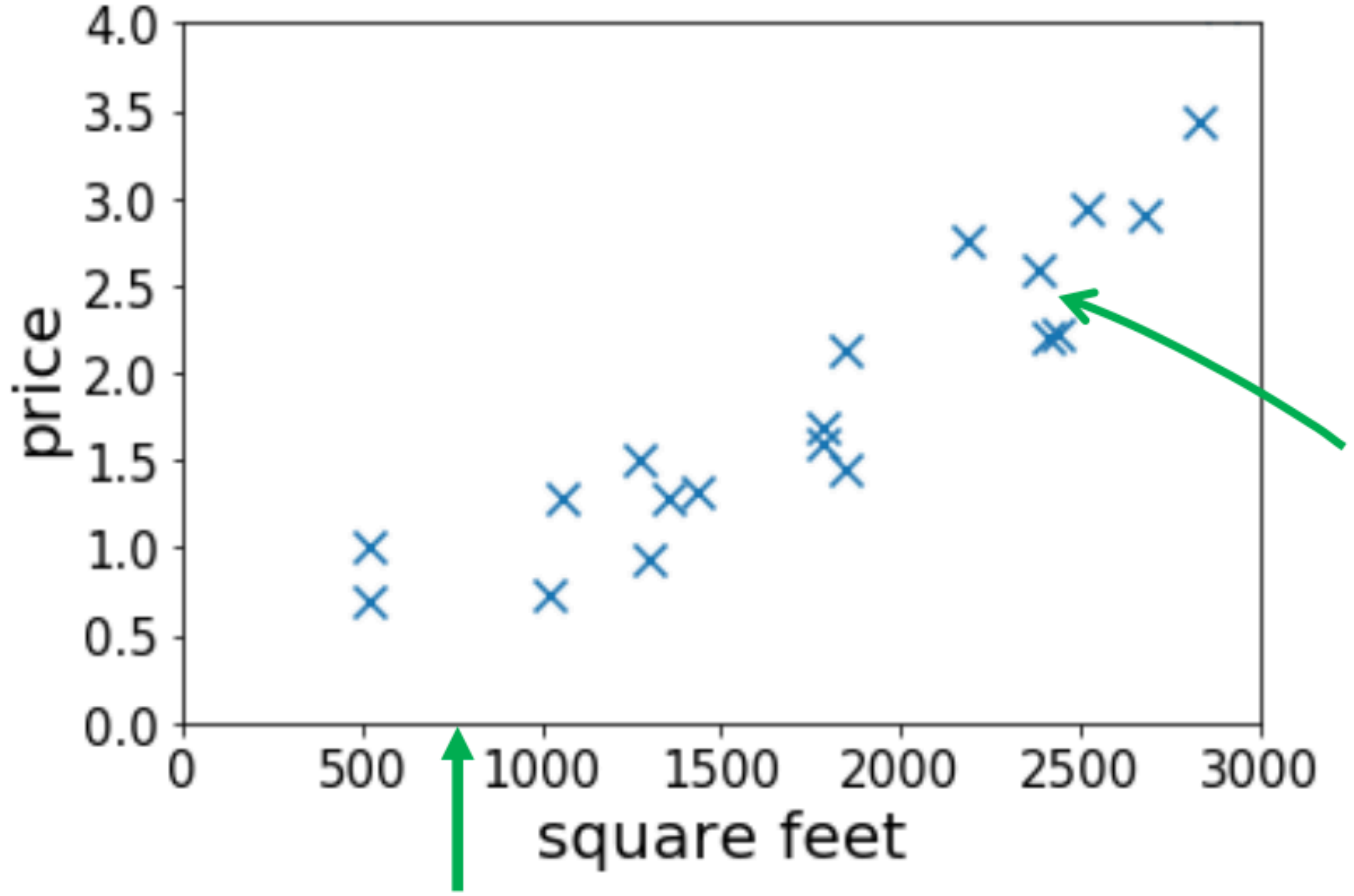
https://www.youtube.com/@3blue1brown/courses



Supervised Learning: Regression

Supervised Learning

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



15th sample
$(x^{(15)}, y^{(15)})$

$x = 800$
$y = ?$

X

CAT Y

Supervised Learning

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

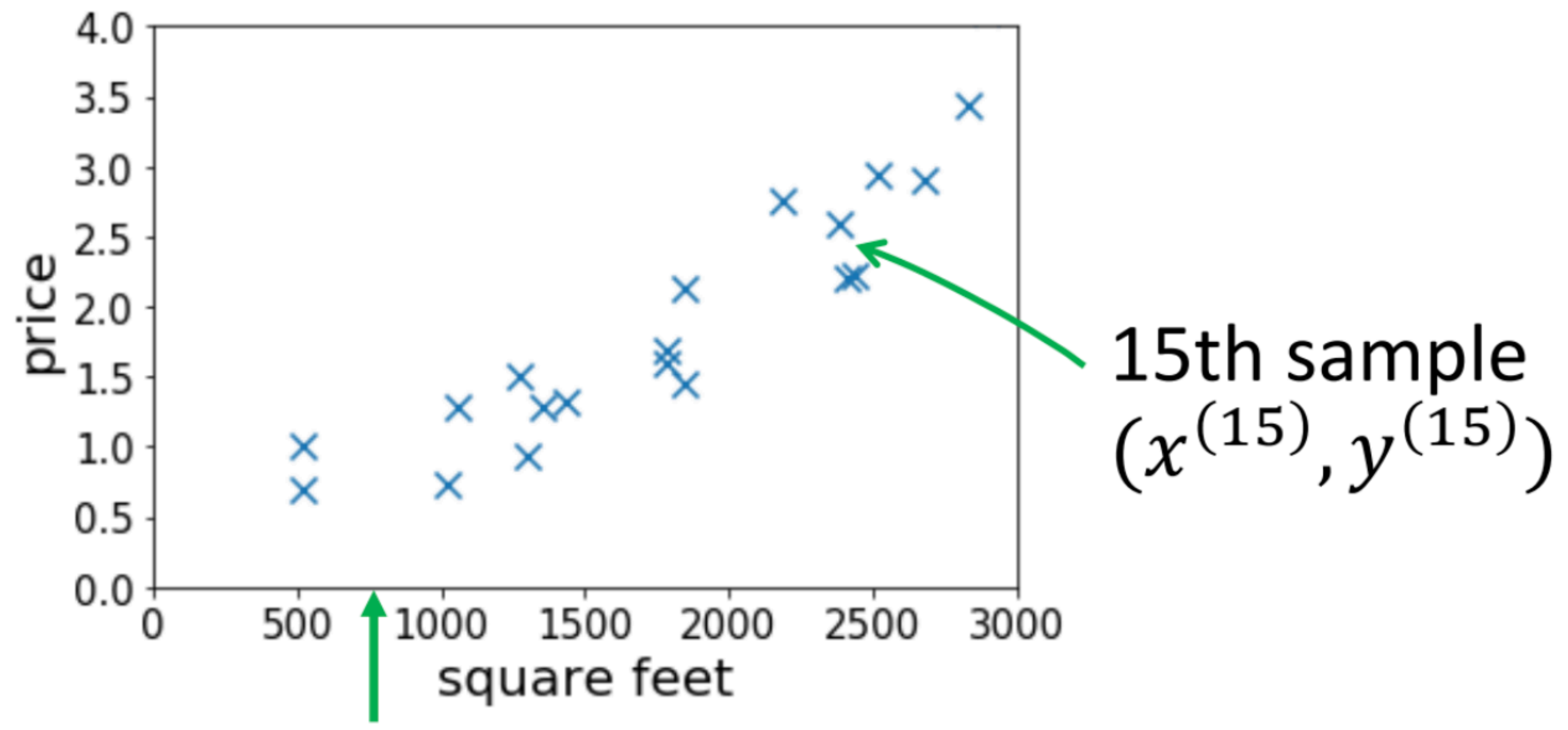
- A training set is set of pairs $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$
  s.t. $x^{(i)} \in \mathcal{X}$ and $y^{(i)} \in \mathcal{Y}$ for $i = 1, \ldots, n$.

- 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



$x = 800$
$y = ?$

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

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

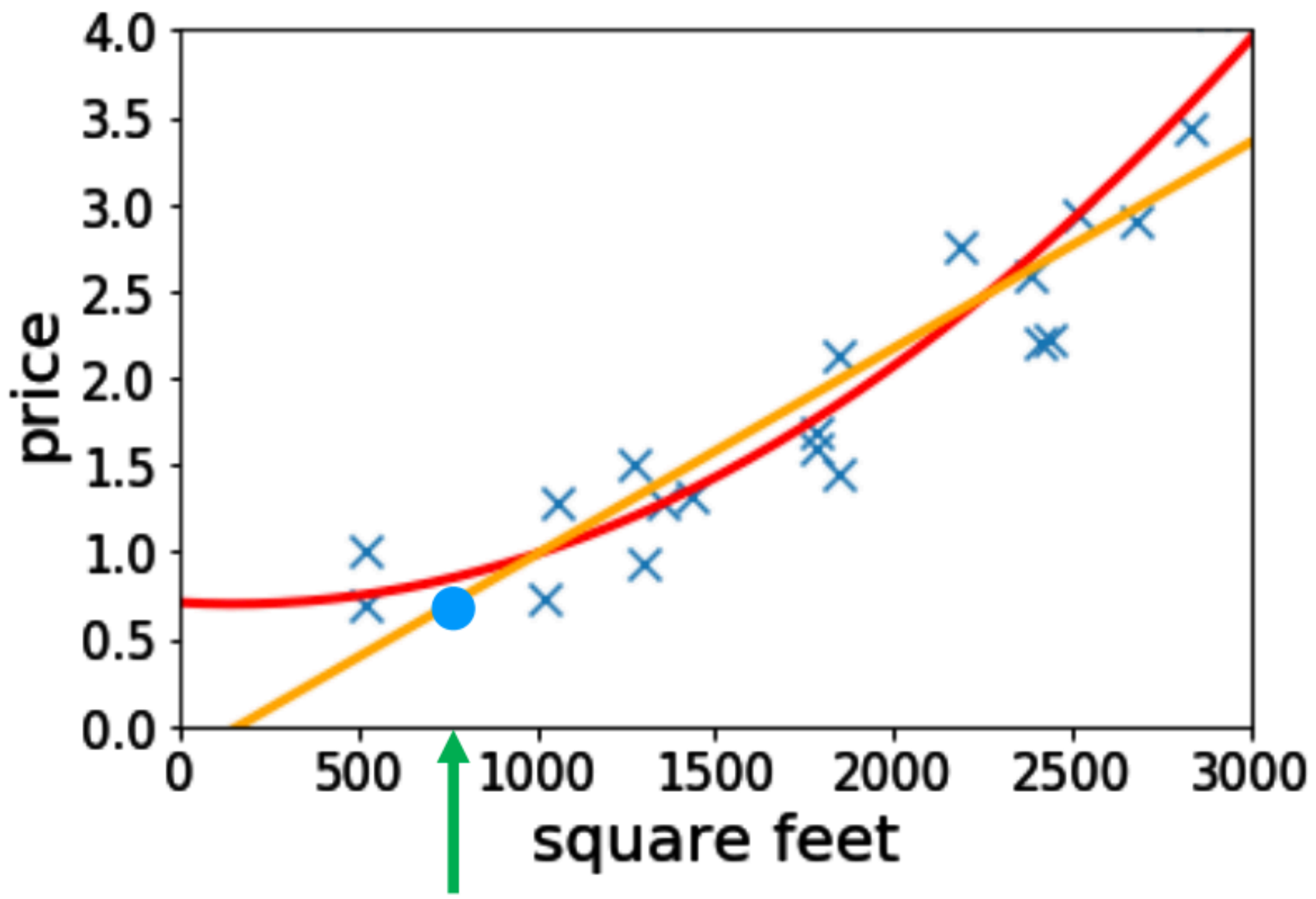
$$\mathbb{I}(\hat{y} = y*) = \begin{cases} 1, & \hat{y} = y* \\ 0 & otherwise \end{cases}$$

Supervised Learning

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

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

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 $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \cdots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset

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

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

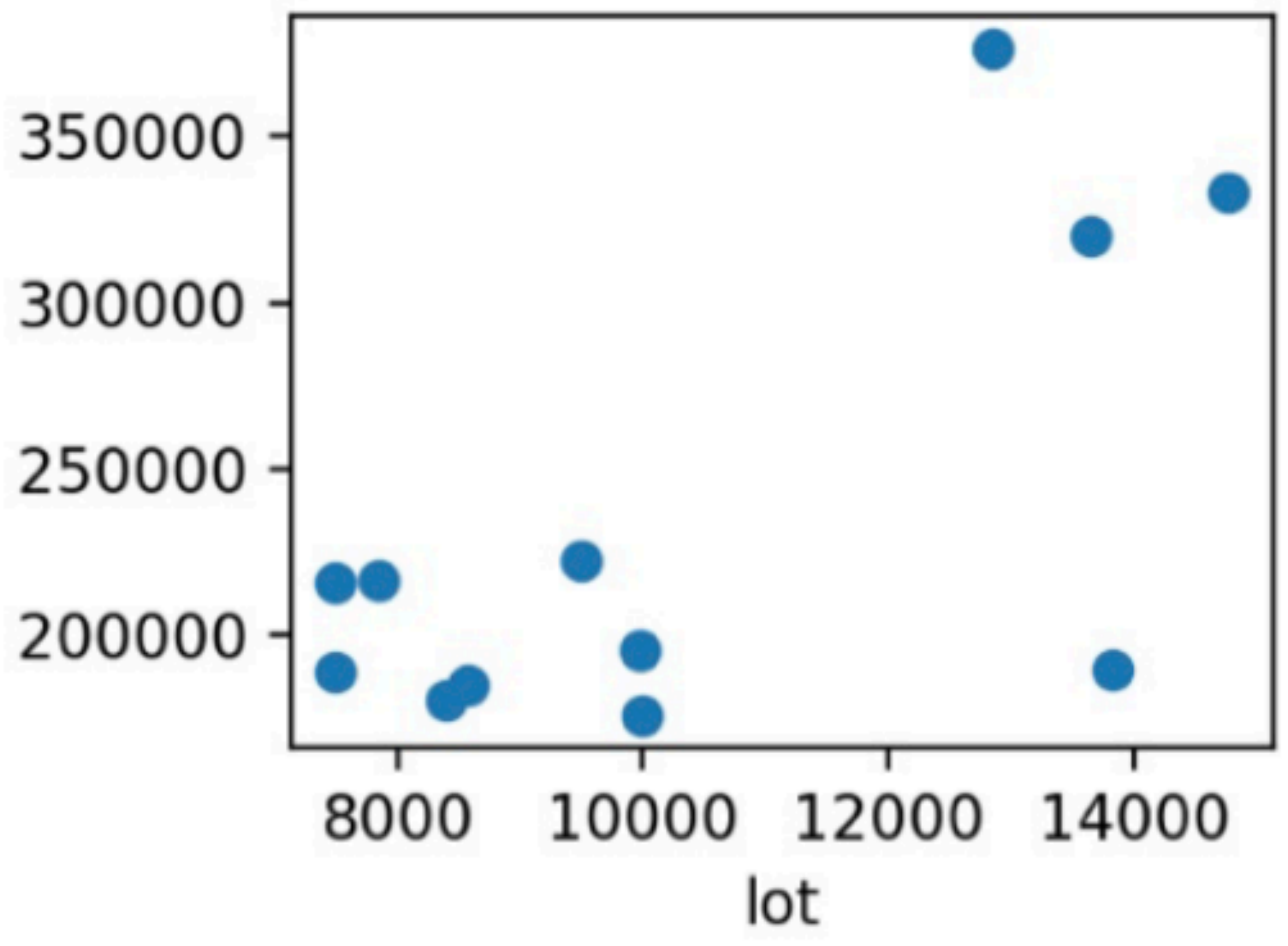
Not only for supervised learning

Example: Regression using Housing Data

Example from Stanford CS229

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



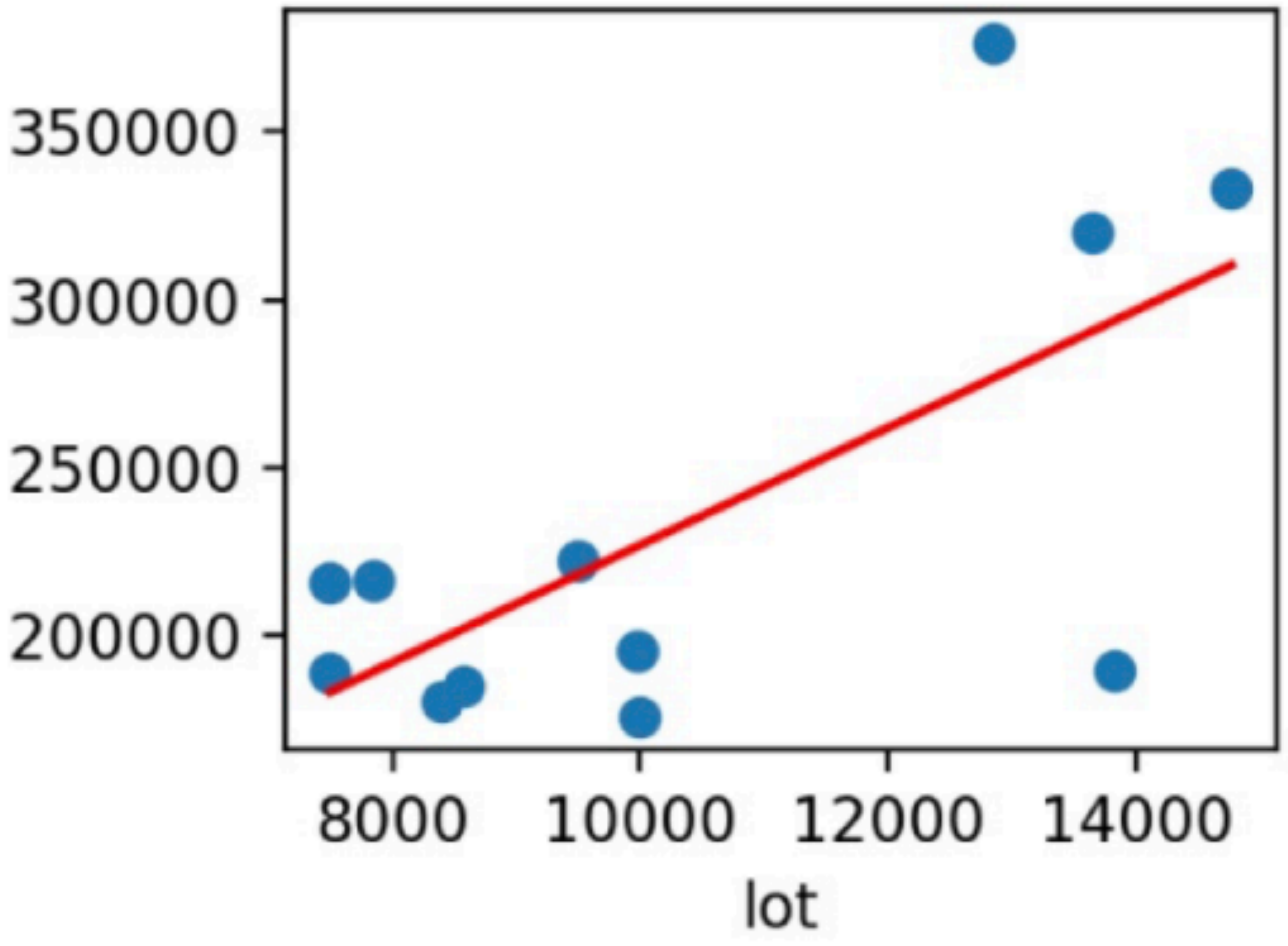
Represent h as a Linear Function

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

Popular choice

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

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

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

	Price
$y^{(1)}$	400
$y^{(2)}$	900

We write the vectors as (important notation)

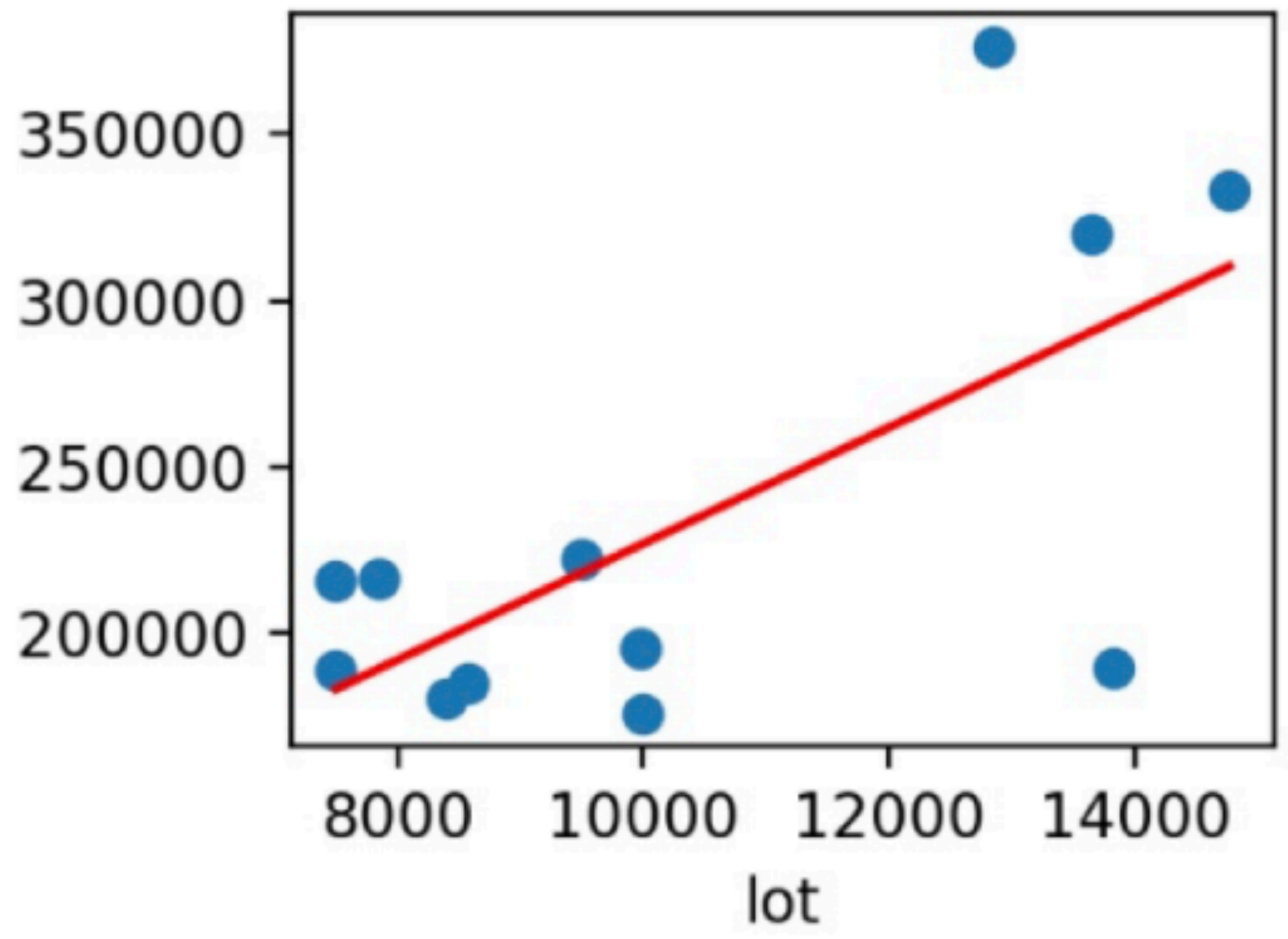
$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \text{ and } x^{(1)} = \begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 2104 \\ 4 \\ 45 \end{pmatrix} \text{ and } y^{(1)} = 400$$

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 $x_0 = 1$)

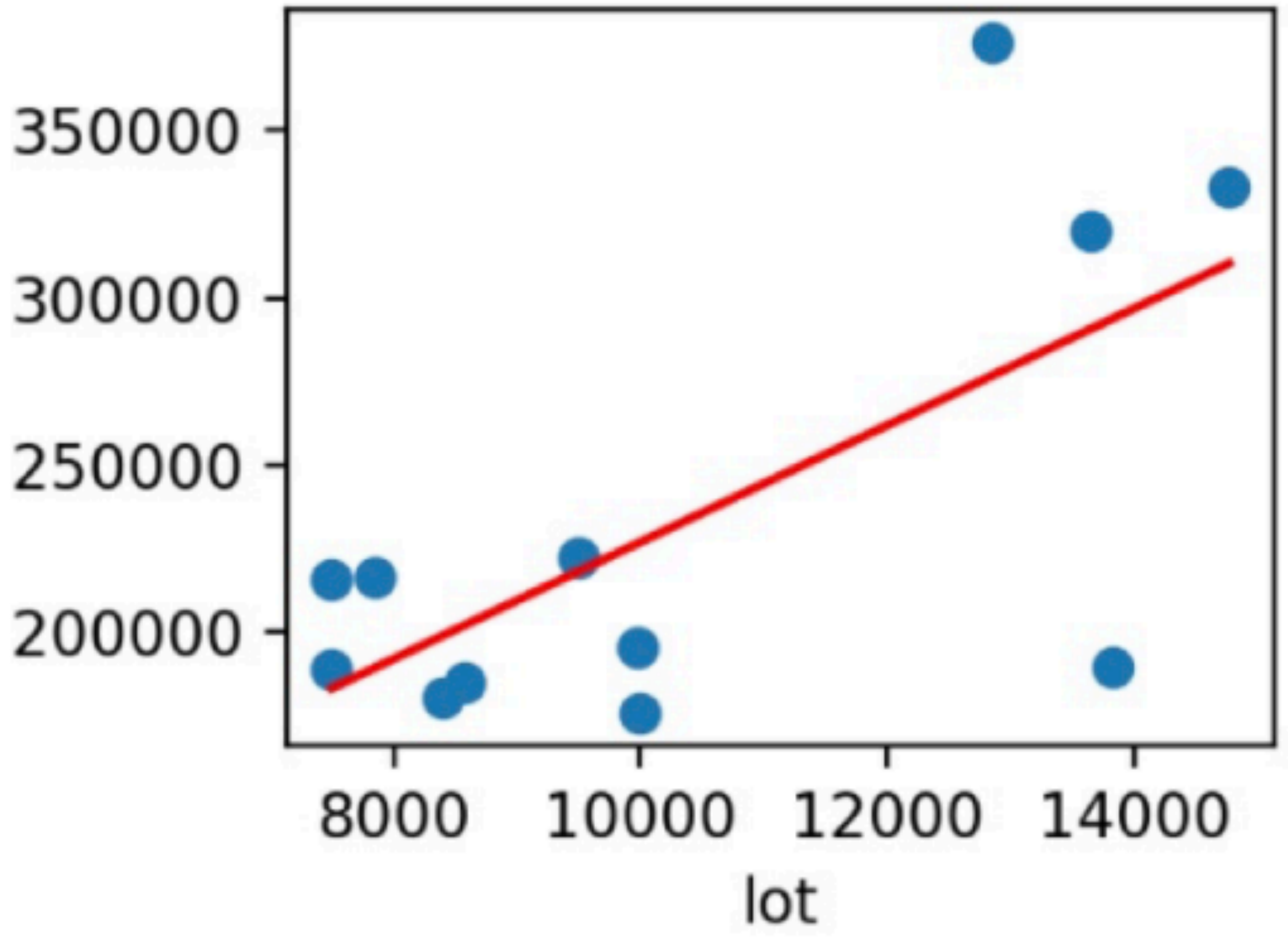
Vector Notation of Prediction



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

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

Loss Function

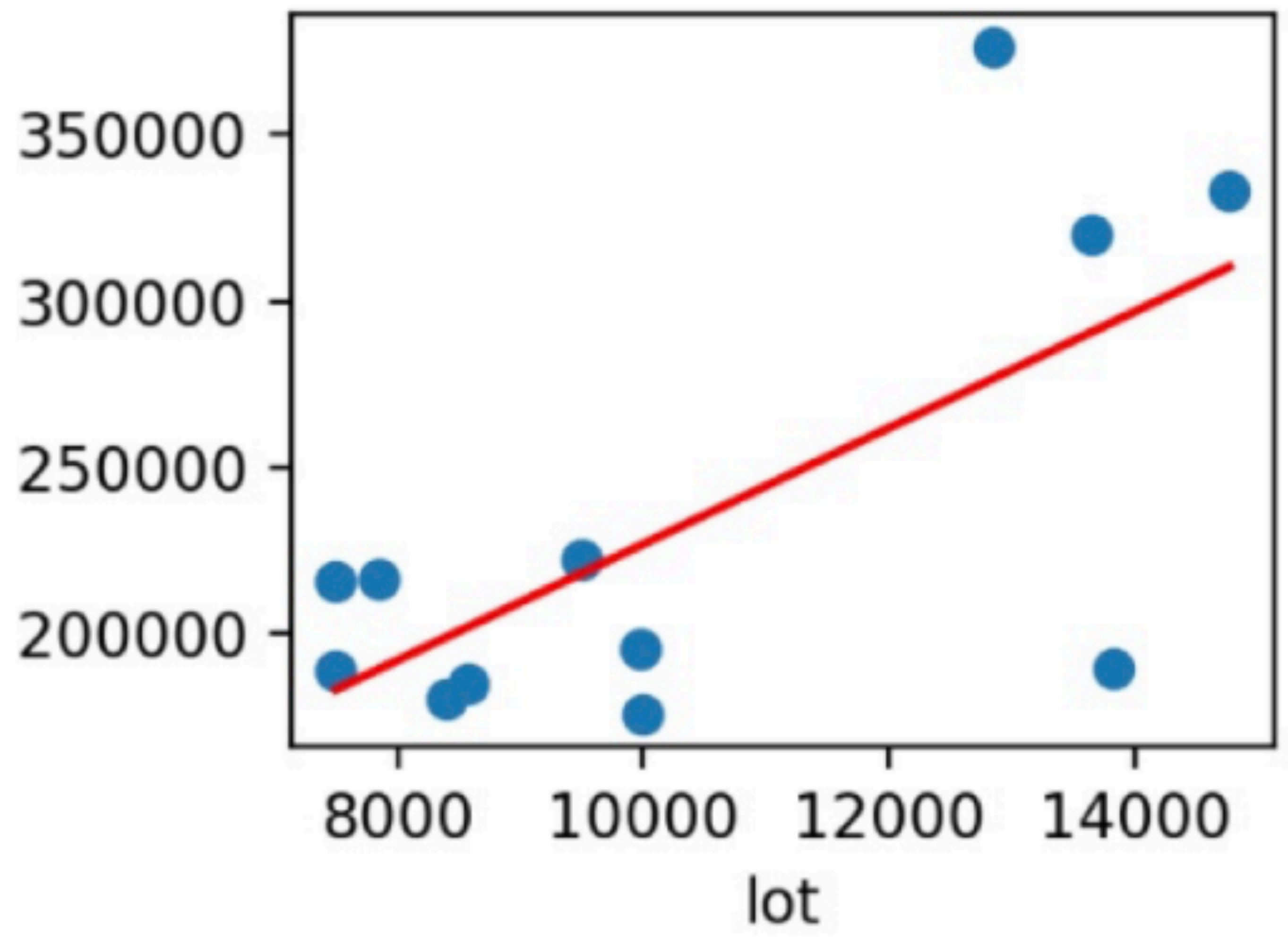


$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T\theta$$

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



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

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

Choose

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

Solving Least Square Problem

Direct Minimization

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

Choose

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

Solving Least Square Problem

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

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

When is X^TX invertible? What if it is not invertible?

Thank You!

Q & A