

Math Basics

(Largely adapted from Stanford CS229 Slides)

Junxian He

Sep 3, 2024

Outline

- Linear Algebra Review

- Probability Review

Outline

- Linear Algebra Review

- Probability Review

Basic Notation

- By $x \in \mathbb{R}^n$, we denote a vector with n entries.

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

- By $A \in \mathbb{R}^{m \times n}$ we denote a matrix with m rows and n columns, where the entries of A are real numbers.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & \vdots & \vdots \\ - & a_m^T & - \end{bmatrix}.$$

Identity Matrix

The *identity matrix*, denoted $I \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

It has the property that for all $A \in \mathbb{R}^{m \times n}$,

$$AI = A = IA.$$

Diagonal Matrix

A *diagonal matrix* is a matrix where all non-diagonal elements are 0. This is typically denoted $D = \text{diag}(d_1, d_2, \dots, d_n)$, with

$$D_{ij} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases}$$

Clearly, $I = \text{diag}(1, 1, \dots, 1)$.

Vector-Vector Product

- *inner product* or *dot product*

$$x^T y \in \mathbb{R} = [\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array}] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i.$$

- *outer product*

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} [\begin{array}{cccc} y_1 & y_2 & \cdots & y_n \end{array}] = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}.$$

Matrix-Vector Product

- If we write A by rows, then we can express Ax as,

$$y = Ax = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_m^T & - \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}.$$

Matrix-Vector Product

- If we write A by columns, then we have:

$$y = Ax = \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \cdots & a^n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a^1 \\ a^2 \\ \vdots \\ a^n \end{bmatrix} x_1 + \begin{bmatrix} a^2 \\ a^2 \\ \vdots \\ a^n \end{bmatrix} x_2 + \dots + \begin{bmatrix} a^n \\ a^n \\ \vdots \\ a^n \end{bmatrix} x_n . \quad (1)$$

y is a *linear combination* of the *columns* of A .

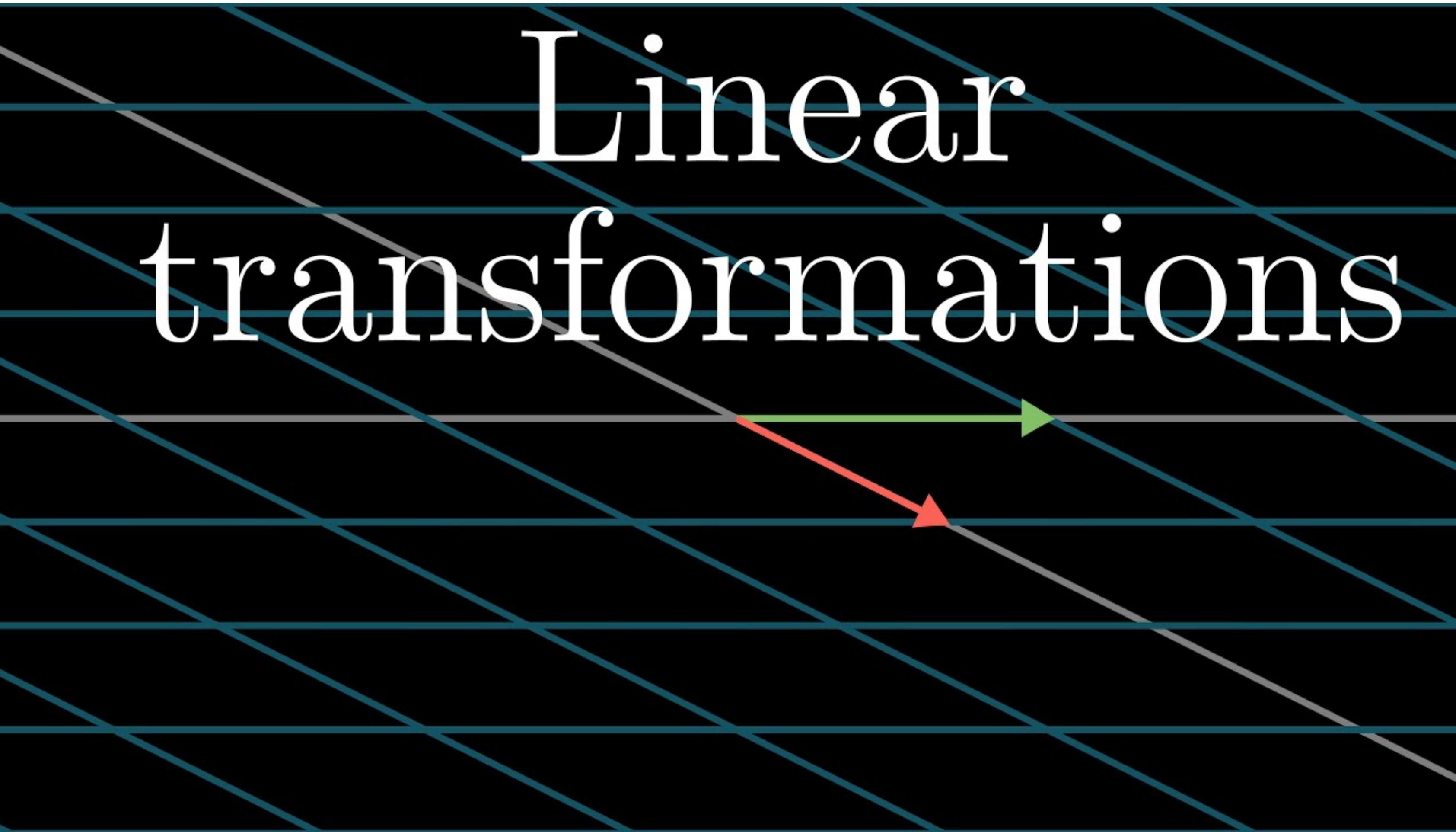
Matrix-Vector Product

It is also possible to multiply on the left by a row vector.

- If we write A by columns, then we can express $x^T A$ as,

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a^1 & a^2 & \dots & a^n \\ | & | & & | \end{bmatrix} = [x^T a^1 \ x^T a^2 \ \dots \ x^T a^n]$$

Linear Transformation



Matrix-Matrix Multiplication

1. As a set of vector-vector products (dot product)

$$C = AB = \begin{bmatrix} & a_1^T & \\ & a_2^T & \\ \vdots & & \\ & a_m^T & \end{bmatrix} \begin{bmatrix} | & | & & | \\ b^1 & b^2 & \dots & b^p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b^1 & a_1^T b^2 & \dots & a_1^T b^p \\ a_2^T b^1 & a_2^T b^2 & \dots & a_2^T b^p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b^1 & a_m^T b^2 & \dots & a_m^T b^p \end{bmatrix}.$$

Matrix-Matrix Multiplication

2. As a sum of outer products

$$C = AB = \begin{bmatrix} & & & \\ | & | & \dots & | \\ a^1 & a^2 & \dots & a^p \\ | & | & & | \end{bmatrix} \begin{bmatrix} & b_1^T & \\ & b_2^T & \\ \vdots & & \\ & b_p^T & \end{bmatrix} = \sum_{i=1}^p a^i b_i^T .$$

Matrix-Matrix Multiplication

3. As a set of matrix-vector products.

$$C = AB = A \begin{bmatrix} | & | & & | \\ b^1 & b^2 & \dots & b^n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ab^1 & Ab^2 & \dots & Ab^n \\ | & | & & | \end{bmatrix}. \quad (2)$$

Here the i th column of C is given by the matrix-vector product with the vector on the right, $c_i = Ab_i$. These matrix-vector products can in turn be interpreted using both viewpoints given in the previous subsection.

Matrix-Matrix Multiplication

4. As a set of vector-matrix products.

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ \vdots & & \\ - & a_m^T & - \end{bmatrix} B = \begin{bmatrix} - & a_1^T B & - \\ - & a_2^T B & - \\ \vdots & & \\ - & a_m^T B & - \end{bmatrix}.$$

Matrix-Matrix Multiplication

- Associative: $(AB)C = A(BC)$.
- Distributive: $A(B + C) = AB + AC$.
- In general, *not* commutative; that is, it can be the case that $AB \neq BA$. (For example, if $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times q}$, the matrix product BA does not even exist if m and q are not equal!)

The Transpose

The *transpose* of a matrix results from “flipping” the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^T \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}.$$

The following properties of transposes are easily verified:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

Trace

The *trace* of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\text{tr}A$, is the sum of diagonal elements in the matrix:

$$\text{tr}A = \sum_{i=1}^n A_{ii}.$$

Norms

A *norm* of a vector $\|x\|$ is informally a measure of the “length” of the vector.

The commonly-used Euclidean or ℓ_2 norm,

$$\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}.$$

The ℓ_1 norm,

$$\|x\|_1 = \sum_{i=1}^n |x_i|$$

Norms

A *norm* of a vector $\|x\|$ is informally a measure of the “length” of the vector.

The ℓ_∞ norm,

$$\|x\|_\infty = \max_i |x_i|.$$

Norms

In fact, all three norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \geq 1$, and defined as

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}.$$

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**.

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 $\text{rank}(A)$.

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

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,

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

- 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 $\|v - y\|_2$.

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

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,

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

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 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:

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

The Determinant

The *determinant* 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 .

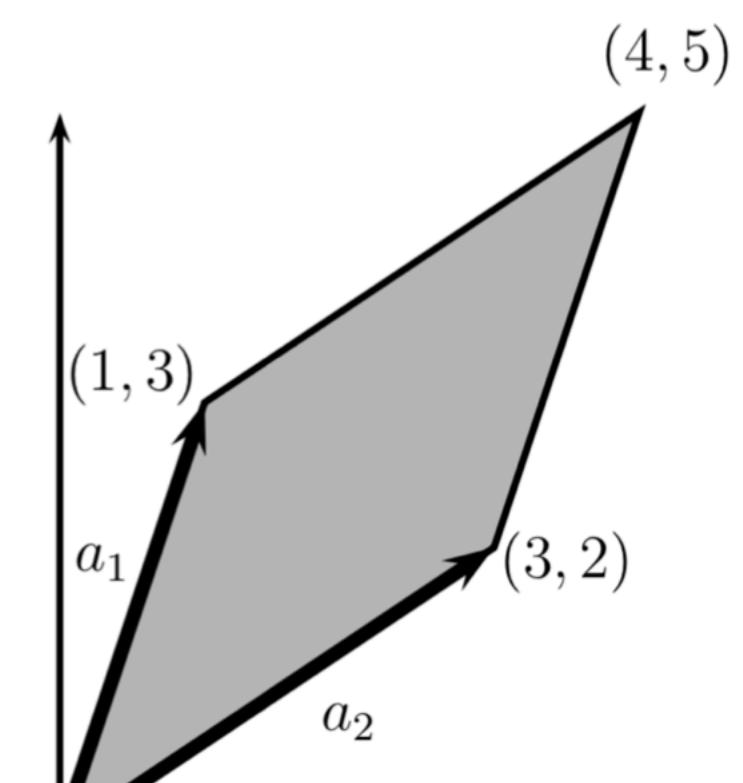
The Determinant

For example, consider the 2×2 matrix,

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

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 $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ 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:

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

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

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

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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,

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

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** → **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: $P(S) = 1$

$E \subseteq F$, 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

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

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

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

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

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

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

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 $A \perp B$,

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

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

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

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

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

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

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

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

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

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.

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

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	$p(1 - p)$
$Binomial(n, p)$	${n \choose k} p^k (1 - p)^{n-k} \text{ for } k = 0, 1, \dots, n$	np	$np(1 - p)$
$Geometric(p)$	$p(1 - p)^{k-1} \text{ for } k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda} \lambda^k}{k!} \text{ for } k = 0, 1, \dots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a} \text{ for all } x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for all } x \in (-\infty, \infty)$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x} \text{ for all } x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

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, \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

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 :

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

- 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

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

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

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

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

Effect of changing variance

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

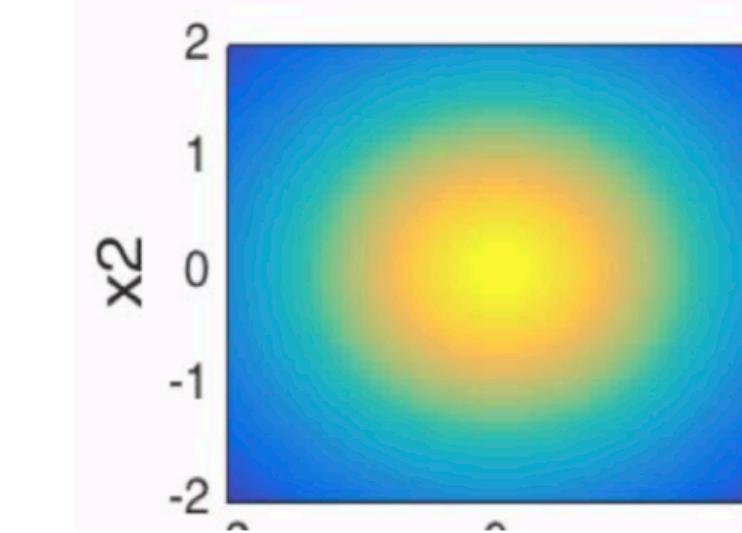
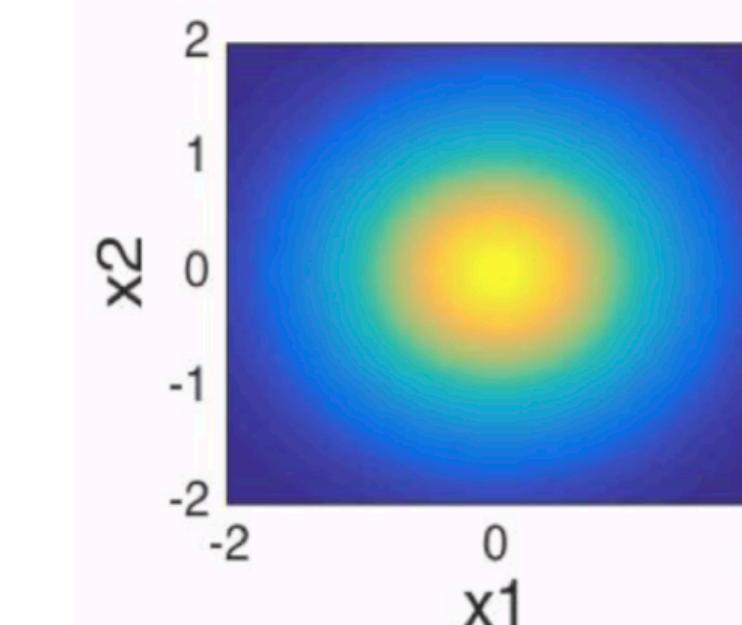
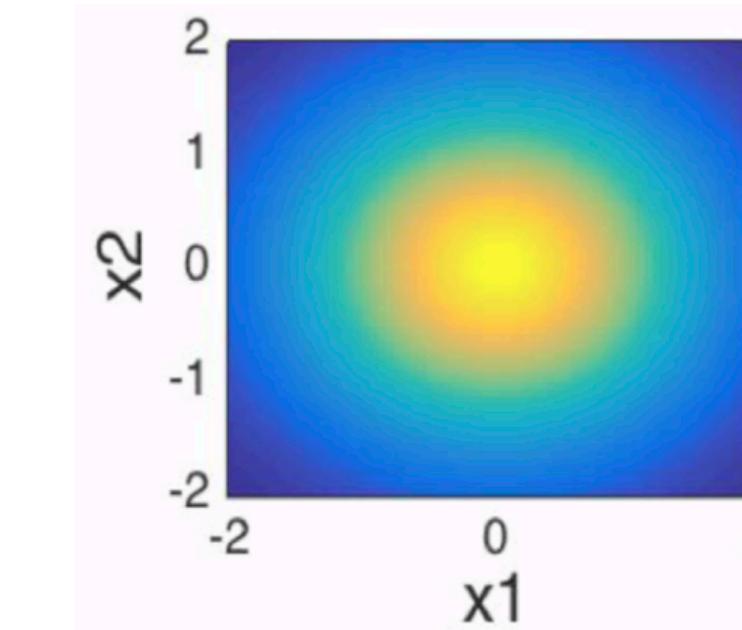
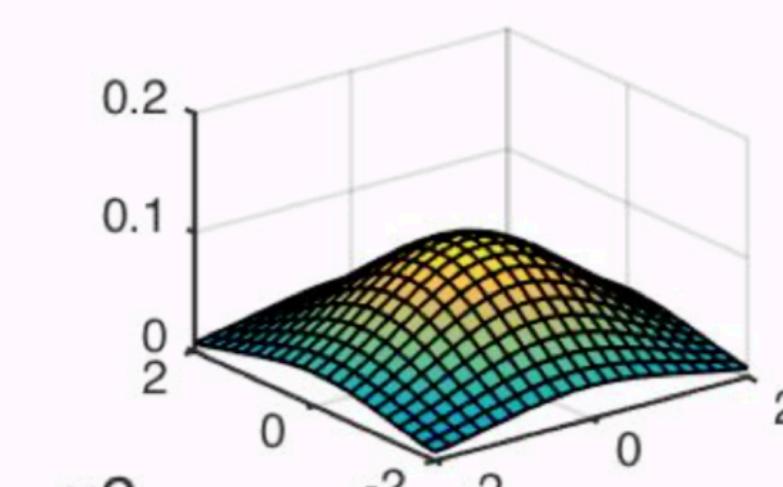
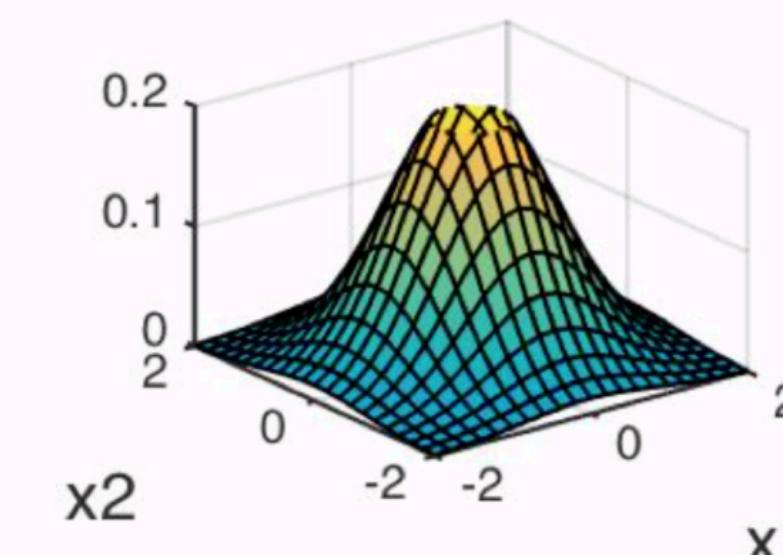
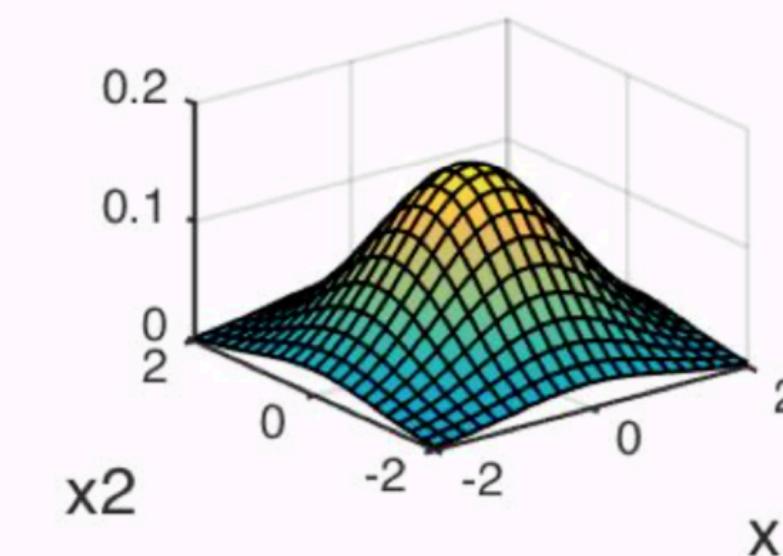
$$\mu = [0 \ 0]^T$$

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

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

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

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



MV Gaussian Visualization

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

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

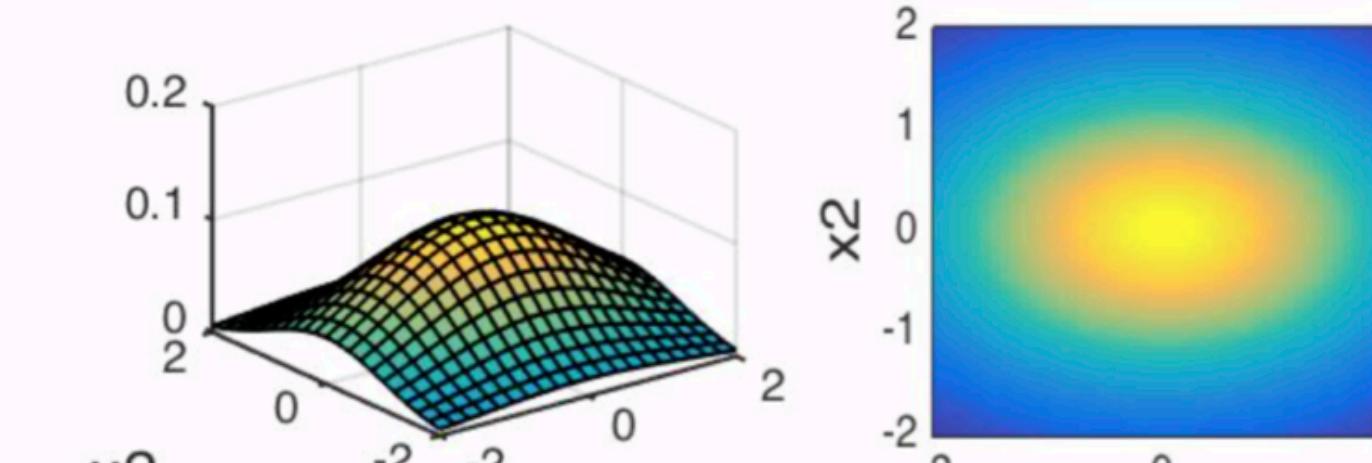
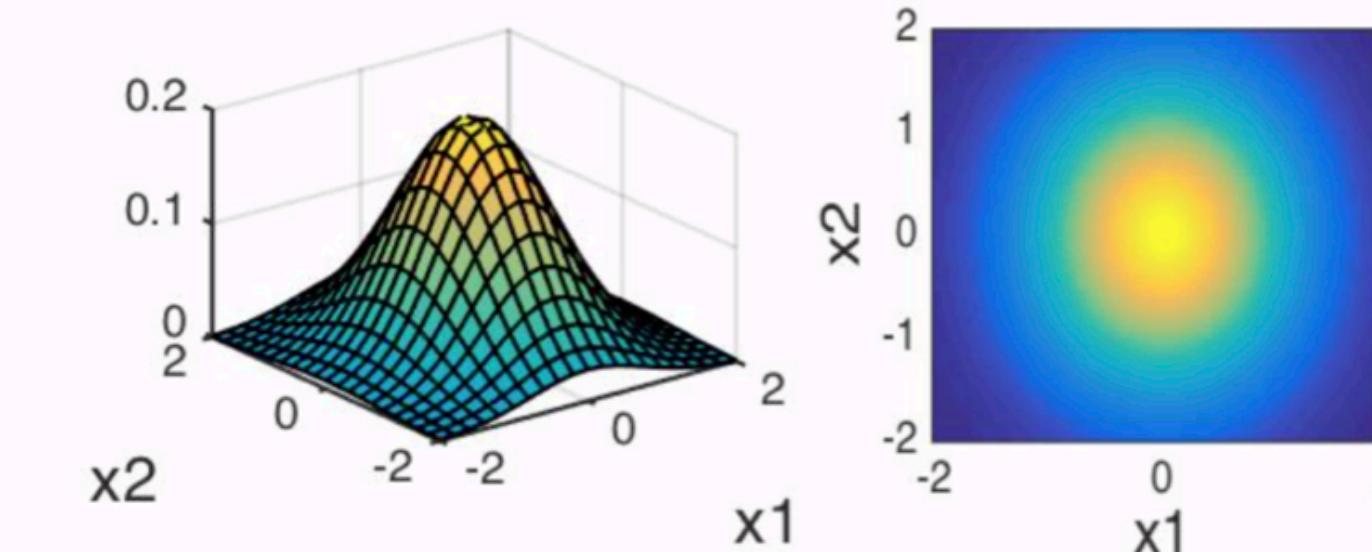
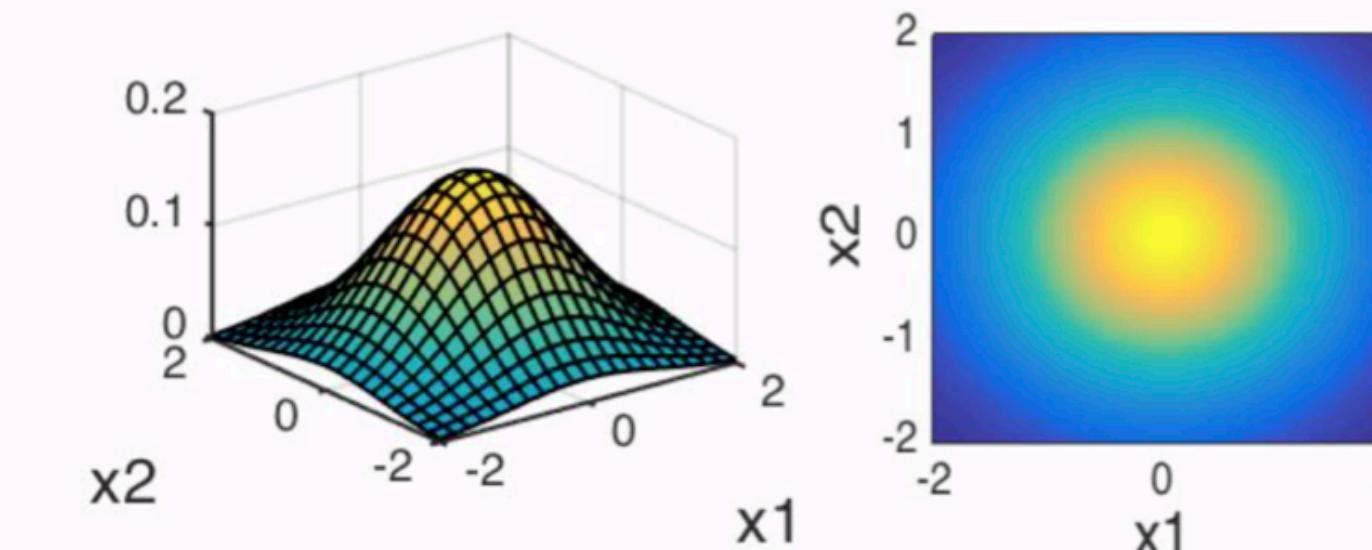
$$\Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$

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

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

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

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



MV Gaussian Visualization

If X_1 and X_2 are positively correlated:

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

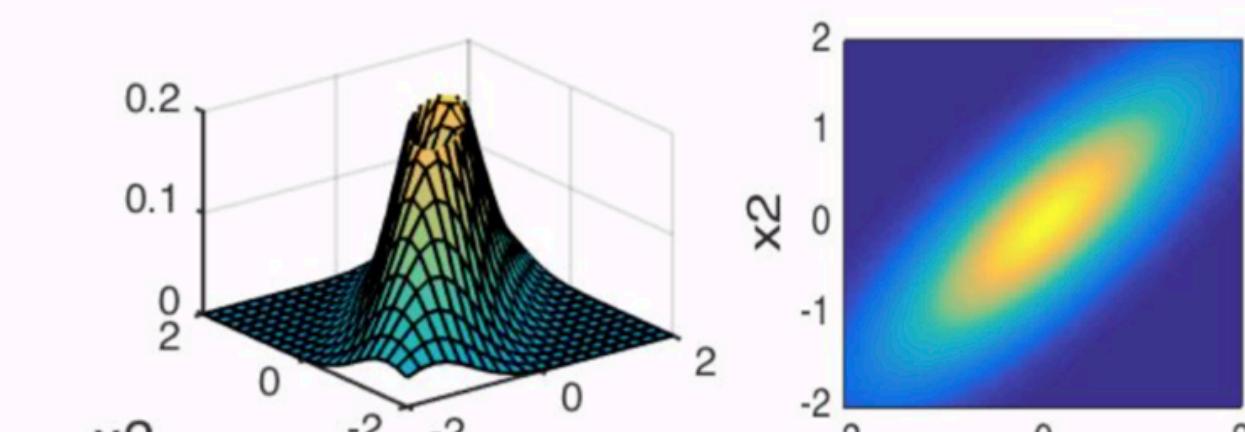
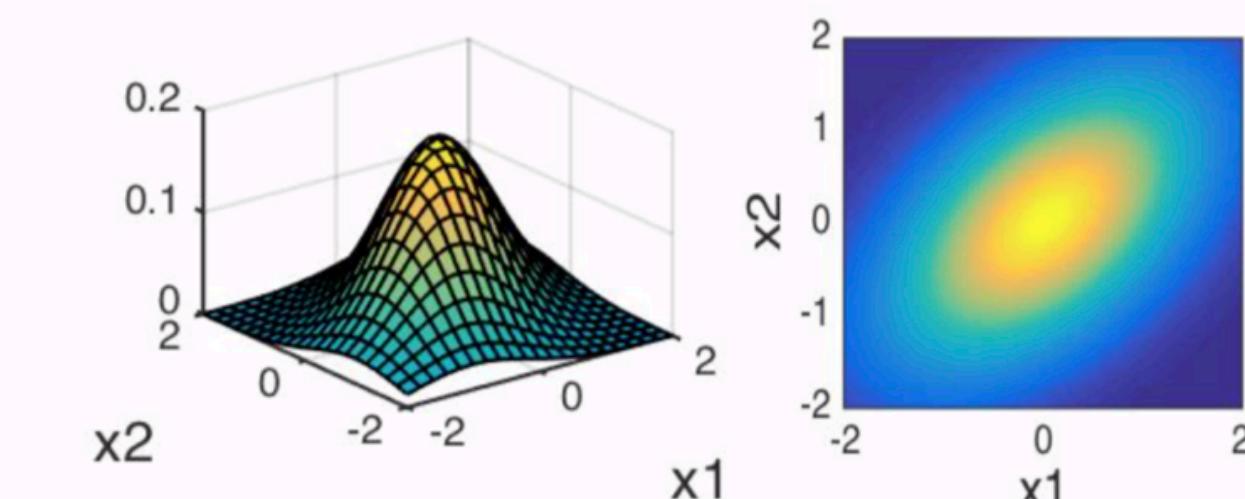
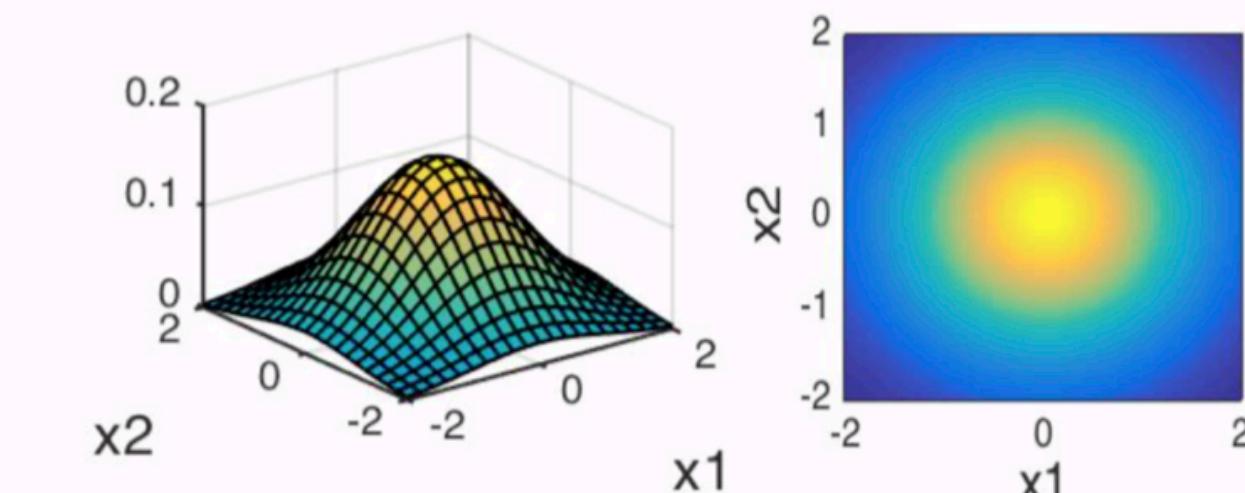
$$\mu = [0 \ 0]^T$$

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

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

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

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



MV Gaussian Visualization

If X_1 and X_2 are negatively correlated:

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

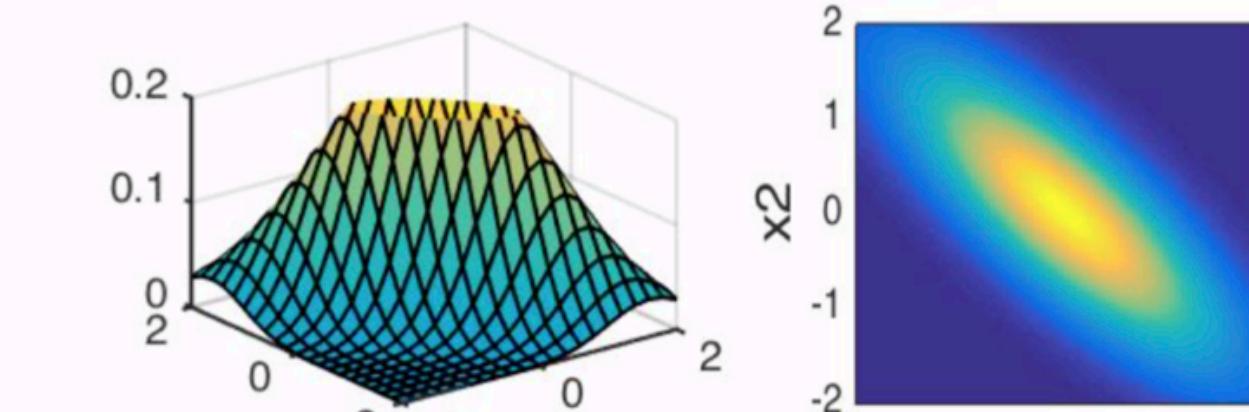
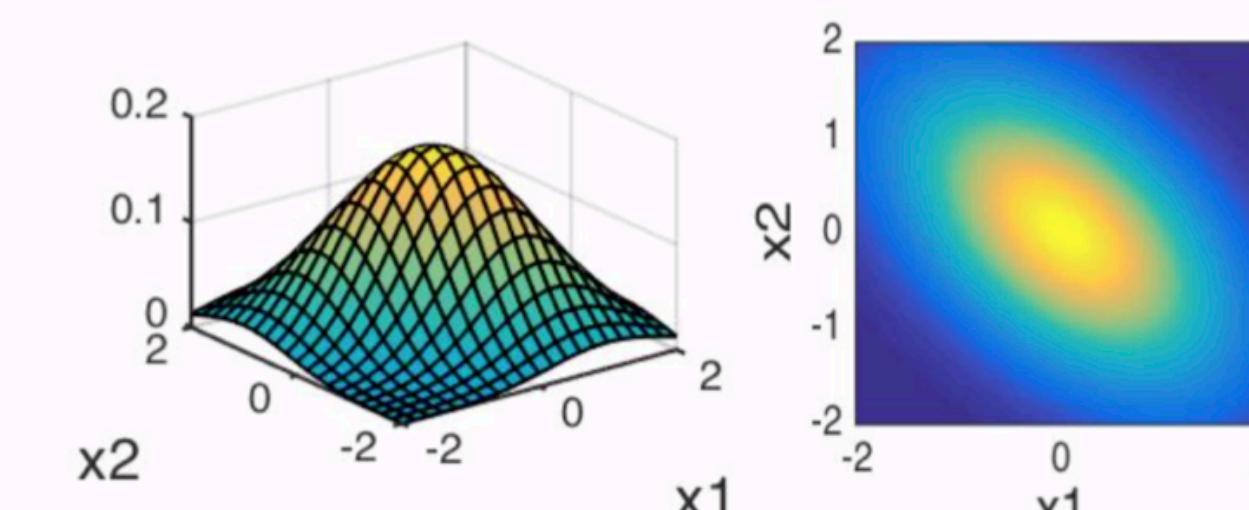
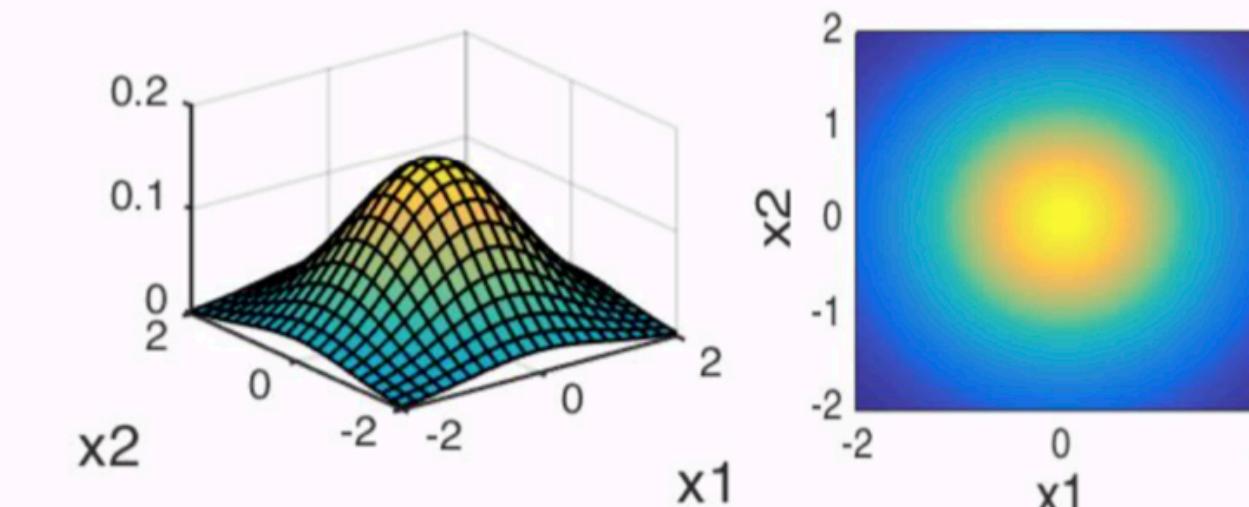
$$\mu = [0 \ 0]^T$$

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

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

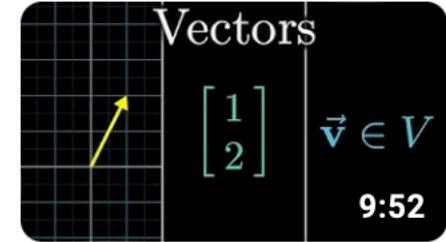
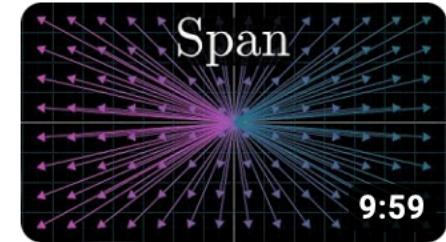
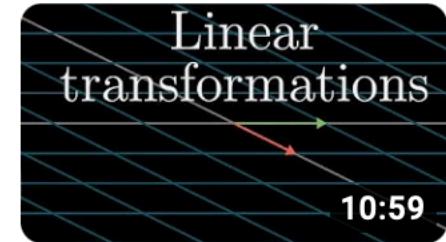
$$\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

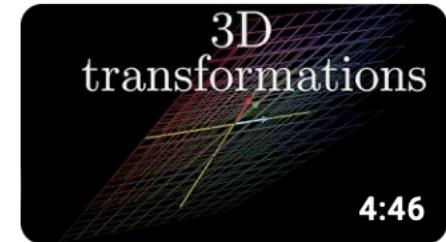
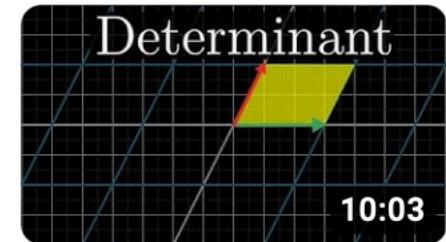
$$\mu = [0 \ 0]^T$$



The purpose of computation is insight, not numbers.

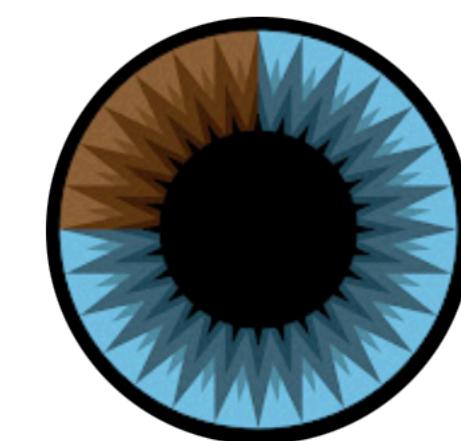
- Richard Hamming

- 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

**Thank You!
Questions?**