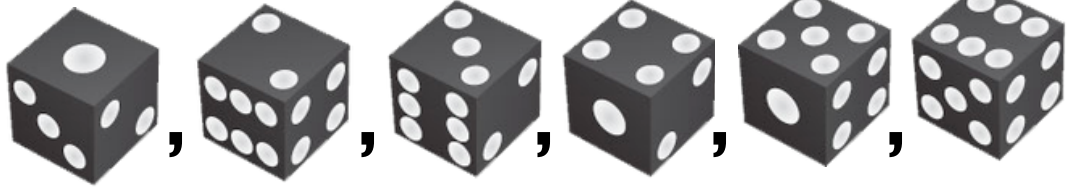
Intro to probability: outcomes

- An **outcome space** specifies the possible outcomes that we would like to reason about, e.g.

$\Omega = \{$  ,  $\}$ Coin toss

$\Omega = \{$  $\}$ Die toss

- We specify a **probability** p(**x**) for each outcome **x** such that

$$p(x) \geq 0, \qquad \sum_{x \in \Omega} p(x) = 1$$

E.g., p() = .6

p() = .4

Intro to probability: events

- An **event** is a subset of the outcome space, e.g.

E = {  } Even die tosses

O = {  } Odd die tosses

- The **probability** of an event is given by the sum of the probabilities of the outcomes it contains,

$$p(E) = \sum_{x \in E} p(x)$$

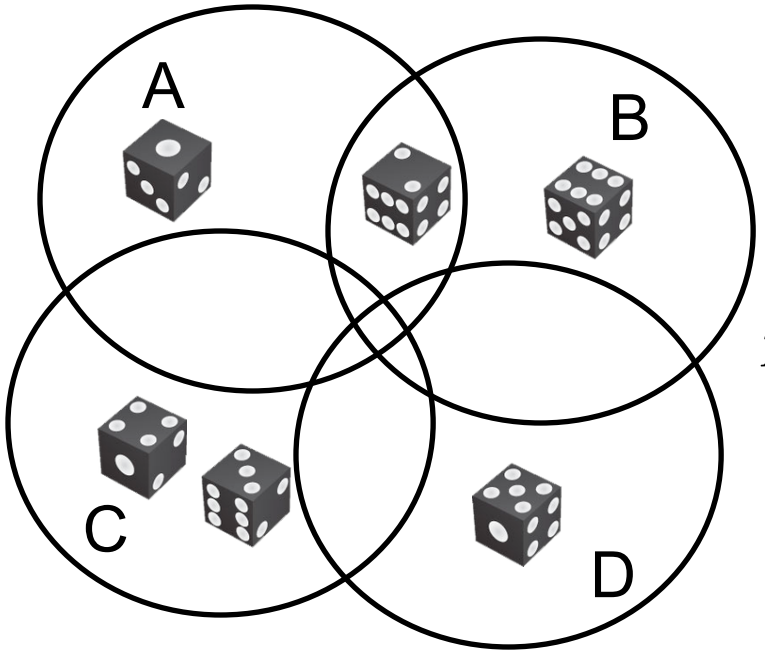
E.g., p(E) = p() + p() + p()

= 1/2, if fair die

Intro to probability: union bound

- P(A or B or C or D or …)

  ≤ P(A) + P(B) + P(C) + P(D) + …



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$
$$\leq p(A) + p(B)$$

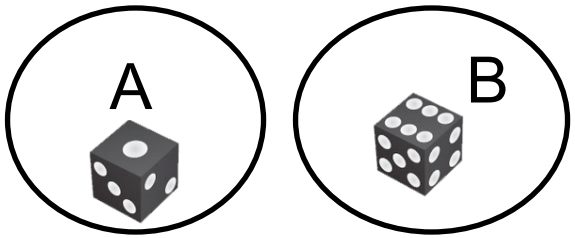
Q: When is this a tight bound?

A: For disjoint events
(i.e., non-overlapping circles)

Intro to probability: independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)p(B)$$



Are these events independent?

No! $p(A \cap B) = 0$

$$p(A)p(B) = \left(\frac{1}{6}\right)^2$$

Intro to probability: independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)p(B)$$

Analogy: outcome space defines all possible sequences of e-mails in training set

- Suppose our outcome space had two different die:

$\Omega = \{$  $\}$ 2 die tosses

$6^2 = 36$ outcomes

and the probability of each outcome is defined as

p() = a_1 b_1 p() = a_1 b_2 ⋯

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

b_1	b_2	b_3	b_4	b_5	b_6
.19	.11	.1	.22	.18	.2

$$\sum_{i=1}^{6} a_i = 1$$

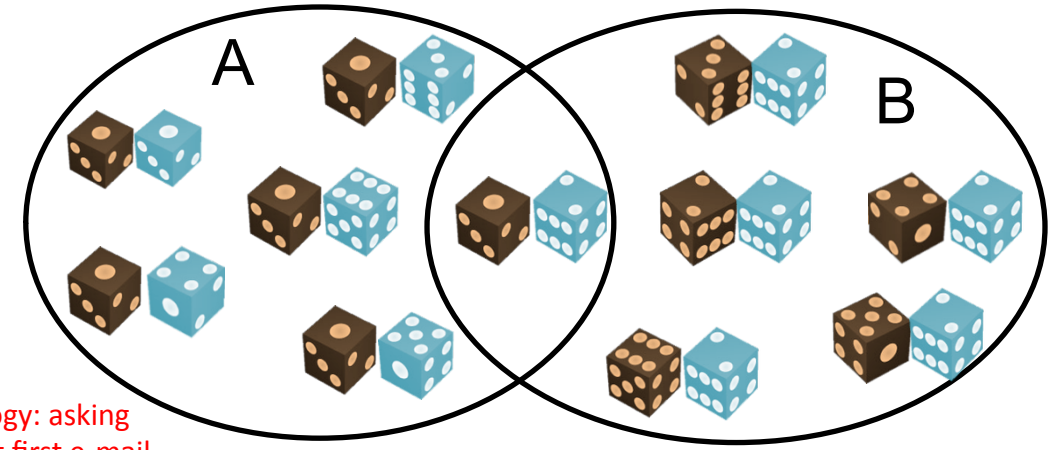
$$\sum_{j=1}^{6} b_j = 1$$

Intro to probability: independence

- Two events A and B are **independent** if

$$p(A \cap B) = p(A)p(B)$$

- Are these events independent?



Analogy: asking about first e-mail in training set

p(A) = p( )

p(B) = p( ) $= b_2$

$$= \sum_{j=1}^{6} a_1 b_j = a_1 \sum_{j=1}^{6} b_j = a_1$$

Analogy: asking about second e-mail in training set

Yes! $p(A \cap B) =$ p( )

$p(A)p(B) =$ p( ) p( )

Intro to probability: discrete random variables

- A **random variable** X is a mapping $X : \Omega \to D$
  - D is some set (e.g., the integers)
  - Induces a partition of all outcomes Ω
- For some $x \in D$, we say

$$p(X = x) = p(\{\omega \in \Omega : X(\omega) = x\})$$

  "probability that variable X assumes state x"
- Notation: $\text{Val}(X)$ = set D of all values assumed by X
  (will interchangeably call these the "values" or "states" of variable X)

$\Omega = \{$  $\}$ 2 die tosses

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - D = How long will it take to drive to work?
  - L = Where am I?

- We denote random variables with capital letters

- Random variables have domains
  - R in {true, false} (sometimes write as {+r, ¬r})
  - D in [0, ∞)
  - L in possible locations, maybe {(0,0), (0,1), …}

Intro to probability: discrete random variables

- $p(X)$ is a distribution: $\sum_{x \in \mathrm{Val}(X)} p(X = x) = 1$
- E.g. X_1 may refer to the value of the first dice, and X_2 to the value of the second dice
- We call two random variables X and Y *identically distributed* if Val(X) = Val(Y) and p(X=s) = p(Y=s) for all s in Val(X)

p() = a_1 b_1 p() = a_1 b_2 ⋯

X_1 and X_2 NOT identically distributed

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

b_1	b_2	b_3	b_4	b_5	b_6
.19	.11	.1	.22	.18	.2

$$\sum_{j=1}^{6} b_j = 1$$

Ω = {  }

2 die tosses

Intro to probability: discrete random variables

- $p(X)$ is a distribution: $\sum_{x \in \mathrm{Val}(X)} p(X = x) = 1$
- E.g. X_1 may refer to the value of the first dice, and X_2 to the value of the second dice
- We call two random variables X and Y *identically distributed* if Val(X) = Val(Y) and p(X=s) = p(Y=s) for all s in Val(X)

p() = a_1 a_1 p() = a_1 a_2 ⋯

X_1 and X_2 identically distributed

a_1	a_2	a_3	a_4	a_5	a_6
.1	.12	.18	.2	.1	.3

$$\sum_{i=1}^{6} a_i = 1$$

Ω = {  } 2 die tosses

Intro to probability: discrete random variables

- X=x is simply an event, so can apply union bound, etc.
- Two random variables **X** and **Y** are **independent** if:

$$p(X = x, Y = y) = p(X = x)p(Y = y) \quad \forall x \in \text{Val}(X), y \in \text{Val}(Y)$$

Joint probability. Formally, given by the event $X = x \cap Y = y$

- The **expectation** of **X** is defined as: $E[X] = \sum_{x \in \text{Val}(X)} p(X = x)x$

- If **X** is binary valued, i.e. x is either 0 or 1, then:

$$\begin{aligned} E[X] &= p(X = 0) \cdot 0 + p(X = 1) \cdot 1 \\ &= p(X = 1) \end{aligned}$$

- Linearity of expectations: $E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$

PAC bound and Bias-Variance tradeoff

for all h, with probability at least 1-δ:

$$\text{error}_{true}(h) \leq \underbrace{\text{error}_D(h)}_{\text{“bias”}} + \underbrace{\sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}}_{\text{“variance”}}$$

- For large |H|
  - low bias (assuming we can find a good h)
  - high variance (because bound is looser)
- For small |H|
  - high bias (is there a good h?)
  - low variance (tighter bound)

Probability Distributions

- Discrete random variables have distributions

$P(T)$

T	P
warm	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

- A discrete distribution is a TABLE of probabilities of values
- The probability of a state (lower case) is a single number

$$P(W = rain) = 0.1 \qquad P(rain) = 0.1$$

- Must have:

$$\forall x \, P(x) \geq 0 \qquad \sum_x P(x) = 1$$

Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment:

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - How many assignments if *n* variables with domain sizes *d*?

  - Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- For all but the smallest distributions, impractical to write out or estimate
  - Instead, we make additional assumptions about the distribution

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_w P(t, w)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(w) = \sum_t P(t, w)$$

$P(W)$

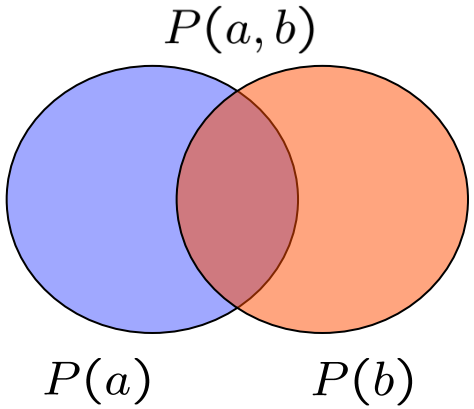
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = r|T = c) = ???$$

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

The Product Rule

- Sometimes have conditional distributions but want the joint

$$P(x|y) = \frac{P(x,y)}{P(y)} \quad \Leftrightarrow \quad P(x,y) = P(x|y)P(y)$$

- Example:

$P(W)$

W	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\Leftrightarrow$

$P(D,W)$

D	W	P
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Let's us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many practical systems (e.g. ASR, MT)

- In the running for most important ML equation!

Linear Algebra

- Linear Algebra Basics
- Matrix Calculus
- Singular Value Decomposition (SVD)
- Eigenvalue Decomposition
- Low-rank Matrix Inversion
- Matlab essentials (recitation)

Scalars

- A scalar is a single number
- Integers, real numbers, rational numbers, etc.
- We denote it with italic font:

$$a, n, x$$

Vectors

- A vector is a 1-D array of numbers:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{2.1}$$

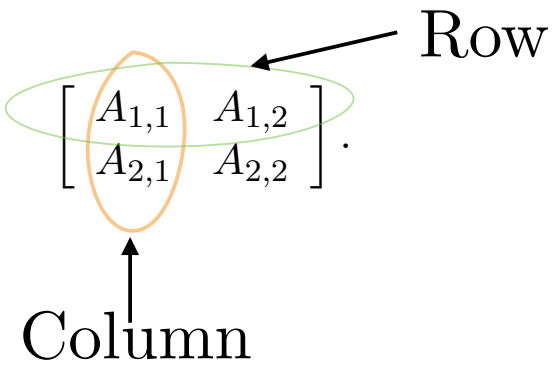
- Can be real, binary, integer, etc.

- Example notation for type and size:

$$\mathbb{R}^n$$

Matrices

- A matrix is a 2-D array of numbers:



(2.2)

- Example notation for type and shape:

$$\boldsymbol{A} \in \mathbb{R}^{m \times n}$$

Tensors

- A tensor is an array of numbers, that may have
  - zero dimensions, and be a scalar
  - one dimension, and be a vector
  - two dimensions, and be a matrix
  - or more dimensions.

Basic concepts

- Vector in R^n is an ordered set of n real numbers.
  - e.g. v = (1,6,3,4) is in R^4
  - A column vector:
  - A row vector:

$$\begin{pmatrix} 1 \\ 6 \\ 3 \\ 4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 6 & 3 & 4 \end{pmatrix}$$

- m-by-n matrix is an object in R^{mxn} with m rows and n columns, each entry filled with a (typically) real number:

$$\begin{pmatrix} 1 & 2 & 8 \\ 4 & 78 & 6 \\ 9 & 3 & 2 \end{pmatrix}$$

Basic concepts

Vector norms: A norm of a vector ||x|| is informally a measure of the "length" of the vector.

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

- Common norms: L_1, L_2 (Euclidean)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- L_∞ $\|x\|_\infty = \max_i |x_i|$

Basic concepts

We will use lower case letters for vectors The elements are referred by x_i.

- Vector dot (inner) product:

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

- Vector outer product:

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

Basic concepts

We will use upper case letters for matrices. The elements are referred by $A_{i,j}$.

- Matrix product:

$$A \in \mathbb{R}^{m \times n} \qquad B \in \mathbb{R}^{n \times p}$$

$$C = AB \in \mathbb{R}^{m \times p}$$

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

e.g.

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

Special matrices

$$\begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$$ diagonal $$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}$$ upper-triangular

$$\begin{pmatrix} a & b & 0 & 0 \\ c & d & e & 0 \\ 0 & f & g & h \\ 0 & 0 & i & j \end{pmatrix}$$ tri-diagonal $$\begin{pmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{pmatrix}$$ lower-triangular

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$ I (identity matrix)

Matrix Transpose

$$(\boldsymbol{A}^\top)_{i,j} = A_{j,i}. \tag{2.3}$$

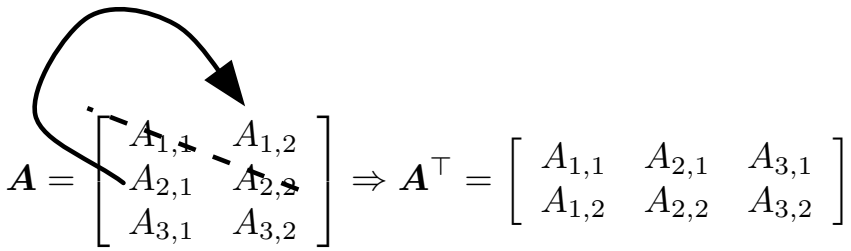


Figure 2.1: The transpose of the matrix can be thought of as a mirror image across the main diagonal.

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top. \tag{2.9}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Systems of Equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2.11}$$

expands to

$$\boldsymbol{A}_{1,:}\boldsymbol{x} = b_1 \tag{2.12}$$

$$\boldsymbol{A}_{2,:}\boldsymbol{x} = b_2 \tag{2.13}$$

$$\dots \tag{2.14}$$

$$\boldsymbol{A}_{m,:}\boldsymbol{x} = b_m \tag{2.15}$$

Solving Systems of Equations

- A linear system of equations can have:
  - No solution
  - Many solutions
  - Exactly one solution: this means multiplication by the matrix is an invertible function

Matrix Inversion

- Matrix inverse:

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n. \tag{2.21}$$

- Solving a system using an inverse:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2.22}$$

$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.23}$$

$$\boldsymbol{I}_n\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.24}$$

- Numerically unstable, but useful for abstract analysis

Invertibility

- Matrix can't be inverted if...
  - More rows than columns
  - More columns than rows
  - Redundant rows/columns ("linearly dependent", "low rank")

Norms

- Functions that measure how "large" a vector is
- Similar to a distance between zero and the point represented by the vector
  - $f(\boldsymbol{x}) = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{0}$
  - $f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ (the *triangle inequality*)
  - $\forall \alpha \in \mathbb{R}, f(\alpha \boldsymbol{x}) = |\alpha| f(\boldsymbol{x})$

Norms

- L^p norm

$$||\boldsymbol{x}||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$

- Most popular norm: L2 norm, p=2
- L1 norm, p=1: $||\boldsymbol{x}||_1 = \sum_i |x_i|$. (2.31)
- Max norm, infinite p: $||\boldsymbol{x}||_\infty = \max_i |x_i|$. (2.32)

Special Matrices and Vectors

- Unit vector:

$$||\boldsymbol{x}||_2 = 1. \tag{2.36}$$

- Symmetric Matrix:

$$\boldsymbol{A} = \boldsymbol{A}^\top. \tag{2.35}$$

- Orthogonal matrix:

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}. \tag{2.37}$$
$$\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$$

Linear independence

- A set of vectors is linearly independent if none of them can be written as a linear combination of the others.
- Vectors $v_1,\ldots,v_k$ are linearly independent if $c_1v_1+\ldots+c_kv_k = 0$ implies $c_1=\ldots=c_k=0$

$$\begin{pmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

e.g. $$\begin{pmatrix} 1 & 0 \\ 2 & 3 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

(u,v)=(0,0), i.e. the columns are linearly independent.

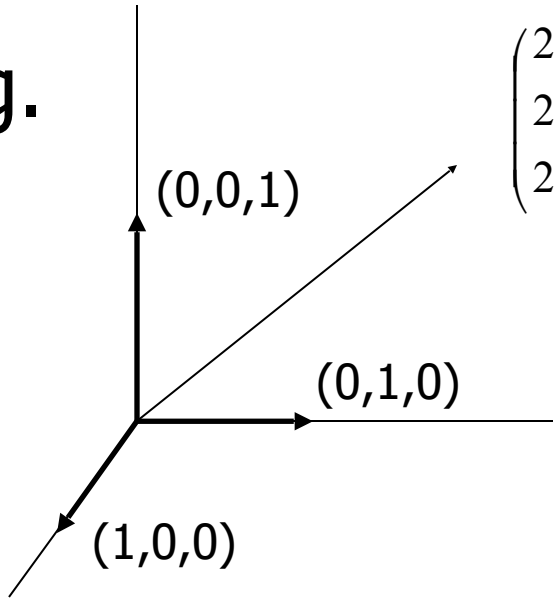
$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

x3 = −2x1 + x2

Span of a vector space

- If all vectors in a vector space may be expressed as linear combinations of a set of vectors $v_1,\ldots,v_k$, then $v_1,\ldots,v_k$ spans the space.
- The cardinality of this set is the dimension of the vector space.

e.g.



$$\begin{pmatrix}2\\2\\2\end{pmatrix} = 2\begin{pmatrix}1\\0\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\\0\end{pmatrix} + 2\begin{pmatrix}0\\0\\1\end{pmatrix}$$

- A basis is a maximal set of linearly independent vectors and a minimal set of spanning vectors of a vector space

Identity Matrix

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

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Rank of a Matrix

- rank(A) (the rank of a m-by-n matrix A) is

  The maximal number of linearly independent columns

  =The maximal number of linearly independent rows

  =The dimension of col(A)

  =The dimension of row(A)

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$$

- If A is n by m, then
  - rank(A)<= min(m,n)
  - If n=rank(A), then A has full row rank
  - If m=rank(A), then A has full column rank

Inverse of a matrix

- Inverse of a square matrix A, denoted by A^{-1} is the unique matrix s.t.
  - $AA^{-1} = A^{-1}A = I$ (identity matrix)
- If A^{-1} and B^{-1} exist, then
  - $(AB)^{-1} = B^{-1}A^{-1}$,
  - $(A^T)^{-1} = (A^{-1})^T$
- For orthonormal matrices

$$\mathbf{A}^{-1} = \mathbf{A}^{\mathsf{T}}$$

- For diagonal matrices

$$\mathbf{D}^{-1} = \mathrm{diag}\{d_1^{-1}, \ldots, d_n^{-1}\}$$

Dimensions

	Scalar	Vector	Matrix
Scalar	$\frac{\mathrm{d}y}{\mathrm{d}x}$	$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}x} = \left[\frac{\partial y_i}{\partial x}\right]$	$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}x} = \left[\frac{\partial y_{ij}}{\partial x}\right]$
Vector	$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial y}{\partial x_j}\right]$	$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{x}} = \left[\frac{\partial y_i}{\partial x_j}\right]$	
Matrix	$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{X}} = \left[\frac{\partial y}{\partial x_{ji}}\right]$		

By Thomas Minka. Old and New Matrix Algebra Useful for Statistics

Examples

$$\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T$$

$$\frac{\partial \mathbf{a}^T \mathbf{X}^T \mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T$$

$$\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T \mathbf{X}^T \mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T$$

$$\frac{\partial \mathbf{x}^T \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B} + \mathbf{B}^T)\mathbf{x}$$

http://matrixcookbook.com/

Singular Value Decomposition (SVD)

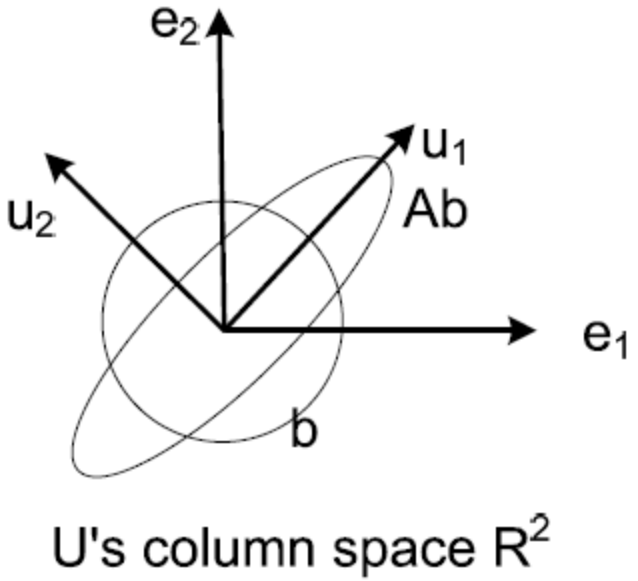
- Any matrix A can be decomposed as $A=UDV^T$, where
  - D is a diagonal matrix, with d=rank(A) non-zero elements
  - The fist d rows of U are orthogonal basis for col(A)
  - The fist d rows of V are orthogonal basis for row(A)

- Applications of the SVD
  - Matrix Pseudoinverse
  - Low-rank matrix approximation

Eigen Value Decomposition

- Any symmetric matrix A can be decomposed as $A=UDU^T$, where
  - D is diagonal, with d=rank(A) non-zero elements
  - The fist d rows of U are orthogonal basis for col(A)=row(A)

- Re-interpreting Ab
  - First stretch b along the direction of u_1 by d_1 times
  - Then further stretch it along the direction of u_2 by d_2 times



Low-rank Matrix Inversion

- In many applications (e.g. linear regression, Gaussian model) we need to calculate the inverse of covariance matrix X^TX (each row of n-by-m matrix X is a data sample)
- If the number of features is huge (e.g. each sample is an image, #sample n<<#feature m) inverting the m-by-m X^TX matrix becomes an problem
- Complexity of matrix inversion is generally $O(n^3)$
- Matlab can comfortably solve matrix inversion with m=thousands, but not much more than that

Low-rank Matrix Inversion

- With the help of SVD, we actually do NOT need to explicitly invert X^TX
  - Decompose $X=UDV^T$
  - Then $X^TX = VDU^TUDV^T = VD^2V^T$
  - Since $V(D^2)V^TV(D^2)^{-1}V^T=I$
  - We know that $(X^TX)^{-1}= V(D^2)^{-1}V^T$
  - Inverting a diagonal matrix D^2 is trivial

Eigendecomposition

- Eigenvector and eigenvalue:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{2.39}$$

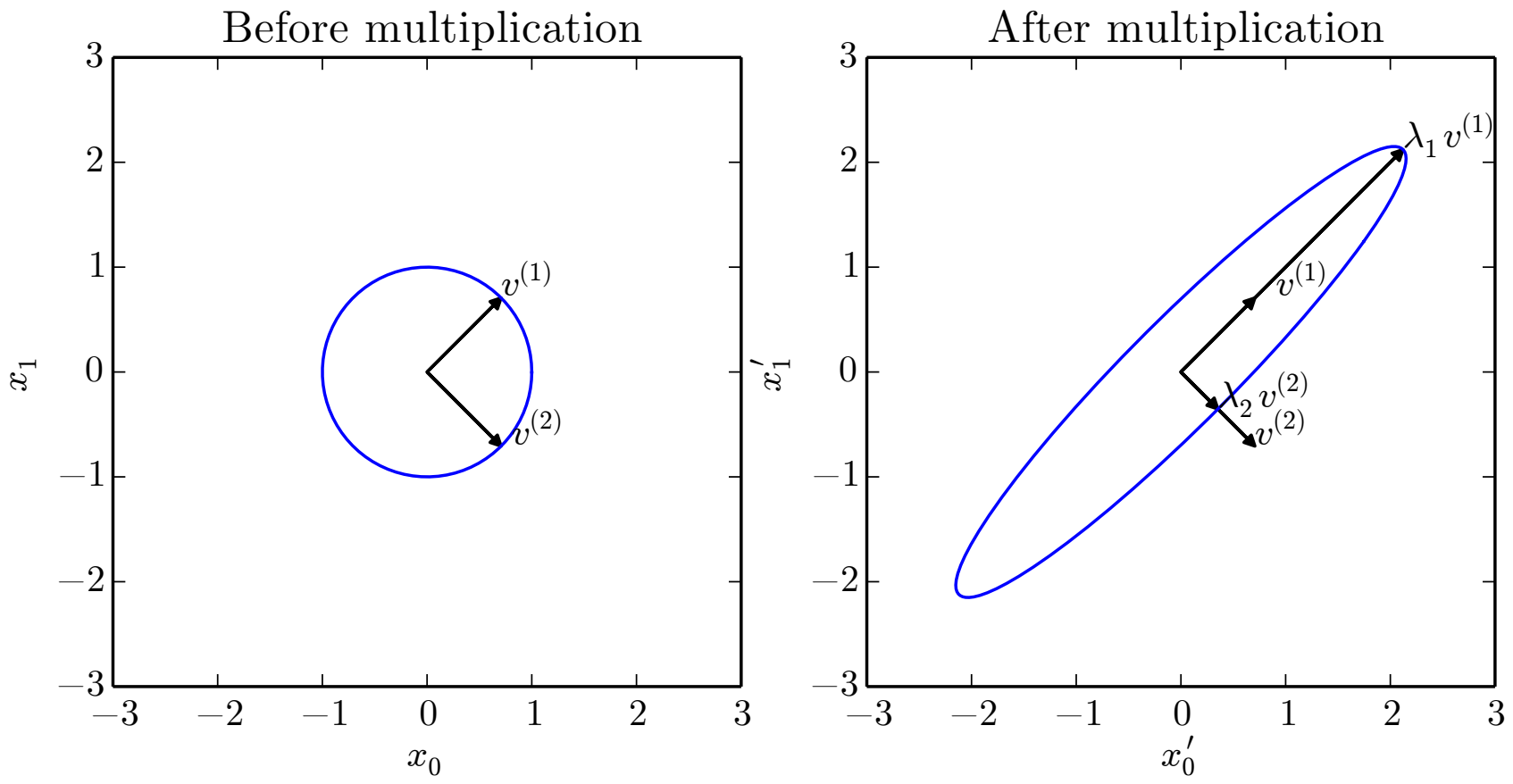
- Eigendecomposition of a diagonalizable matrix:

$$\boldsymbol{A} = \boldsymbol{V}\text{diag}(\boldsymbol{\lambda})\boldsymbol{V}^{-1}. \tag{2.40}$$

- Every real symmetric matrix has a real, orthogonal eigendecomposition:

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\top} \tag{2.41}$$

Effect of Eigenvalues



Singular Value Decomposition

- Similar to eigendecomposition
- More general; matrix need not be square

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}. \tag{2.43}$$

Moore-Penrose Pseudoinverse

$$\boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{y}$$

- If the equation has:
  - Exactly one solution: this is the same as the inverse.
  - No solution: this gives us the solution with the smallest error $||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}||_2$.
  - Many solutions: this gives us the solution with the smallest norm of $\boldsymbol{x}$.

Computing the Pseudoinverse

The SVD allows the computation of the pseudoinverse:

$$\boldsymbol{A}^{+} = \boldsymbol{V}\boldsymbol{D}^{+}\boldsymbol{U}^{\top}, \tag{2.47}$$

Take reciprocal of non-zero entries

Trace

$$\mathrm{Tr}(\boldsymbol{A}) = \sum_i \boldsymbol{A}_{i,i}. \tag{2.48}$$

$$\mathrm{Tr}(\boldsymbol{ABC}) = \mathrm{Tr}(\boldsymbol{CAB}) = \mathrm{Tr}(\boldsymbol{BCA}) \tag{2.51}$$