

CS5014 Machine Learning

Lecture 2 Maths background review

Lei Fang

School of Computer Science, University of St Andrews

21 Jan 2021



University
of
St Andrews

So why this review session ?

Maths is useful

- rigorous and concise way of communicating results
- help us understand why *and* why not algorithms work
- be able to derive your own model and algorithms!

Refresher on essential concepts

- only a refresher; we expect you have learnt them
 - don't expect to know everything after this lecture
- not complete and not rigorous

Self-assessment for yourself

- identify rusty area
- do self studies afterwards
- maths learning should be never-ending :-)

So why this review session ?

Maths is useful

- rigorous and concise way of communicating results
- help us understand why *and* why not algorithms work
- be able to derive your own model and algorithms!

Refresher on essential concepts

- only a refresher; we expect you have learnt them
 - don't expect to know everything after this lecture
- not complete and not rigorous

Self-assessment for yourself

- identify rusty area
- do self studies afterwards
- maths learning should be never-ending :-)

So why this review session ?

Maths is useful

- rigorous and concise way of communicating results
- help us understand why *and* why not algorithms work
- be able to derive your own model and algorithms!

Refresher on essential concepts

- only a refresher; we expect you have learnt them
 - don't expect to know everything after this lecture
- not complete and not rigorous

Self-assessment for yourself

- identify rusty area
- do self studies afterwards
- maths learning should be never-ending :-)

Mathematics for machine learning

Linear algebra

- leap forward from elementary algebra: 1-d to multi-dimensional
- number line to a number plane (space)

Probability theory and statistics

- study of uncertainty: uncertainty is the norm
 - e.g. rain tomorrow? blood pressure measurement (reading error)?
- how to generalise your results
 - from one sample to the universe: vaccine trial

Calculus

- study of continuous (real-valued) functions (using approximation, say *polynomial*)
 - $y = \sin(x)$ is well approximated by $y = x$ when $x \approx 0$
- useful when we do optimisation

Mathematics for machine learning

Linear algebra

- leap forward from elementary algebra: 1-d to multi-dimensional
- number line to a number plane (space)

Probability theory and statistics

- study of uncertainty: uncertainty is the norm
 - e.g. rain tomorrow? blood pressure measurement (reading error)?
- how to generalise your results
 - from one sample to the universe: vaccine trial

Calculus

- study of continuous (real-valued) functions (using approximation, say *polynomial*)
 - $y = \sin(x)$ is well approximated by $y = x$ when $x \approx 0$
- useful when we do optimisation

Mathematics for machine learning

Linear algebra

- leap forward from elementary algebra: 1-d to multi-dimensional
- number line to a number plane (space)

Probability theory and statistics

- study of uncertainty: uncertainty is the norm
 - e.g. rain tomorrow? blood pressure measurement (reading error)?
- how to generalise your results
 - from one sample to the universe: vaccine trial

Calculus

- study of continuous (real-valued) functions (using approximation, say *polynomial*)
 - $y = \sin(x)$ is well approximated by $y = x$ when $x \approx 0$
- useful when we do optimisation

Useful textbook and references (read the *italic* entries!)

Linear algebra

- *Learning from Data Supplementary Mathematics (Vector and Linear Algebra)* by David Barber;
<https://api.semanticscholar.org/CorpusID:18857001>
- *Chapter 2 of Deep Learning* by Ian Goodfellow, Yoshua Bengio and Aaron Courville
https://www.deeplearningbook.org/contents/linear_algebra.html
- Introduction to Linear Algebra by Gilbert Strang;
<http://math.mit.edu/~gs/linearalgebra/>
- The Matrix Cookbook by Kaare Brandt Petersen, Michael Syskind Pedersen;
<https://www2.imm.dtu.dk/pubdb/pubs/3274-full.html>
 - useful as a reference manual

Probability theory

- *Chapter 2.1-2.3 Information Theory, Inference, and Learning Algorithms by David J.C. MacKay*
<http://www.inference.org.uk/itprnn/book.pdf>
- *Chapter 3.1-3.9 of Deep Learning by Ian Goodfellow, Yoshua Bengio and Aaron Courville*
<https://www.deeplearningbook.org/contents/prob.html>
- Introduction to Probability Models by Sheldon Ross
 - chapter 1; chapter 2.1-2.5, 2.8; chapter 3.1-3.5

Calculus

- Use your book of choice; read multivariate calculus part as well
- Appendix of Bayesian Reasoning and Machine Learning by David Barber
<http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/200620.pdf>

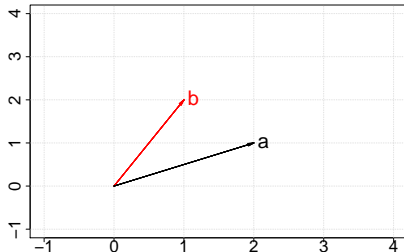
Linear algebra: Basic concepts

- vectors
- norms and distances
- linear independence, span, subspace
- matrices, linear transformation
- matrix operations
- rank

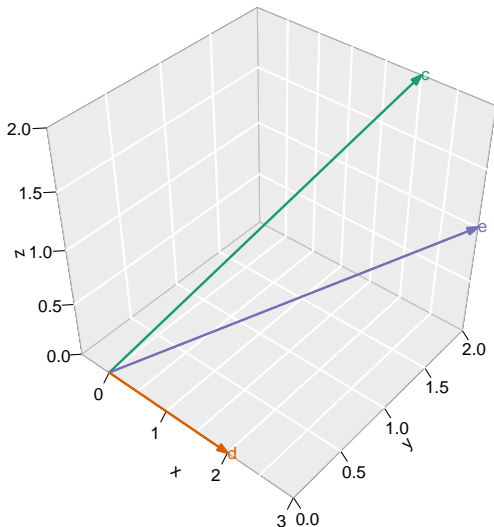
Vector

A vector is a collection of n scalars

- $\mathbf{a} \in R^n$, default option is column vector i.e. $n \times 1$
- represents a **displacement** in R^n
- e.g. $\mathbf{a} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ (or $\mathbf{a} = [2, 1]^T$ to save space)



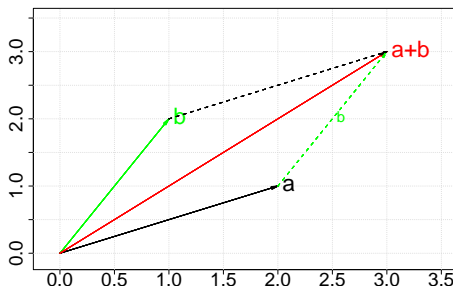
Some 3-d vectors $\mathbf{c} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}$, $\mathbf{d} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$, $\mathbf{e} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$



Vector addition

$$\mathbf{a} + \mathbf{b} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{bmatrix} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_d + b_d \end{bmatrix}$$

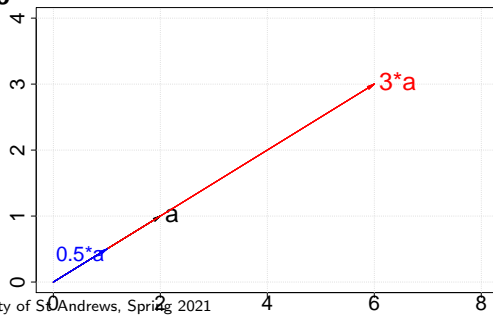
- generalisation from scalar arithmetics; remember $2+1$ on a number axis ?
- parallelogram rule



Vector scaling/multiplication

$$k \cdot \mathbf{a} = k \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} k \times a_1 \\ k \times a_2 \\ \vdots \\ k \times a_d \end{bmatrix}, k \in \mathbb{R} \text{ or a scalar}$$

- geometrically, scaling means shrinking or stretching a vector
 - the direction does not change but length changes
- and obviously $n \cdot \mathbf{a} = \mathbf{a} + \dots + \mathbf{a} = \sum_n \mathbf{a}$
- $0 \cdot \mathbf{a} = \mathbf{0}$



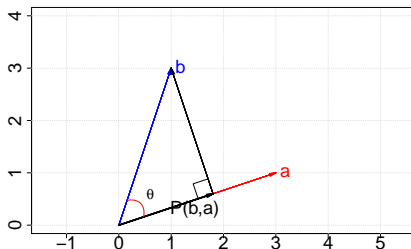
Inner product

$$\mathbf{a}^T \mathbf{b} = [a_1, a_2, \dots, a_d] \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{bmatrix} = \sum_{i=1}^d a_i \times b_i$$

- $\mathbf{a}^T \mathbf{b} = \mathbf{b}^T \mathbf{a}$ and the result is a scalar
- $\mathbf{a}^T (\mathbf{b} + \mathbf{c}) = \mathbf{a}^T \mathbf{b} + \mathbf{a}^T \mathbf{c}$
- $(k\mathbf{a})^T \mathbf{b} = \mathbf{a}^T (k\mathbf{b}) = k(\mathbf{a}^T \mathbf{b})$
- $\mathbf{a}^T \mathbf{a} = \sum_{i=1}^d a_i^2$ is squared Euclidean distance between \mathbf{a} and $\mathbf{0}$
- $\mathbf{a}^T \mathbf{a} \geq 0$ and $\mathbf{a} = \mathbf{0}$ if and only if $\mathbf{a}^T \mathbf{a} = 0$

Another interpretation:

$$\mathbf{a}^T \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta$$



- θ is the angle between \mathbf{a}, \mathbf{b}
 - $\mathbf{a}^T \mathbf{b} = 0$ if and only if $\mathbf{a} \perp \mathbf{b}$
- $\|\mathbf{a}\| \cos \theta$ is the projected length of \mathbf{a} on \mathbf{b}
- $\|\mathbf{b}\| \cos \theta$ is the projected length of \mathbf{b} on \mathbf{a}
- $P(\mathbf{b}, \mathbf{a})$ denotes the projected vector of \mathbf{b} to \mathbf{a}
 - so $\|\mathbf{b}\| \cos \theta = \|P(\mathbf{b}, \mathbf{a})\|$
- and (prove it or convince yourself!)

$$P(\mathbf{b}, \mathbf{a}) = \|\mathbf{b}\| \cos \theta * \frac{\mathbf{a}}{\|\mathbf{a}\|} = \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}} \mathbf{a}$$

Matrix

A rectangular array of real numbers $A \in R^{m \times n}$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} \text{---} & \tilde{\mathbf{a}}_1 & \text{---} \\ \vdots & \vdots & \vdots \\ \text{---} & \tilde{\mathbf{a}}_m & \text{---} \end{bmatrix}$$

- can be viewed as a collection of n column vectors

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n];$$

- or row vectors $\mathbf{A} = [\tilde{\mathbf{a}}_1^T, \tilde{\mathbf{a}}_2^T, \dots, \tilde{\mathbf{a}}_m^T]^T$
- sometimes written as $\mathbf{A} = (a_{ij}) \ i = 1, \dots, m, j = 1, \dots, n$

Matrix operations

- addition: $\mathbf{A} + \mathbf{B} = \mathbf{C} = (c_{ij})$ where $c_{ij} = a_{ij} + b_{ij}$
- scaling: $k\mathbf{A} = \mathbf{C}$ where $c_{ij} = k * a_{ij}$
- transpose: $\mathbf{A}^T = \mathbf{C}$ where $c_{ij} = a_{ji}$
- multiplication: Let $\mathbf{A} \in R^{m \times s}$, $\mathbf{B} \in R^{s \times n}$

$$\mathbf{AB} = \mathbf{C}, \mathbf{C} \in R^{m \times n}$$

where

$$c_{ij} = \sum_{k=1}^s a_{ik} b_{jk}$$

or $c_{ij} = \tilde{\mathbf{a}}_i^T \mathbf{b}_j$

- $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$
- $\mathbf{AB} \neq \mathbf{BA}$
- $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$
- \mathbf{I} identity matrix: $\mathbf{IA} = \mathbf{A}$ or $\mathbf{AI} = \mathbf{A}$
- inverse (only applies to square matrix): $\mathbf{A}^{-1}\mathbf{A} = \mathbf{AA}^{-1} = \mathbf{I}$

Examples

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 5 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix} = ?$$

it is not allowed as the dimensions do not match

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix}^T = \begin{bmatrix} 2 & 6 & 1 \\ 3 & 4 & 1 \end{bmatrix}$$

Examples

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 5 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix} = ?$$

it is not allowed as the dimensions do not match

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix}^T = \begin{bmatrix} 2 & 6 & 1 \\ 3 & 4 & 1 \end{bmatrix}$$

Example

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 5 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} \boxed{2 \times 5 + 3 \times 1} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 5 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 13 & 7 & 10 \\ 34 & 16 & 20 \\ 6 & 3 & 4 \end{bmatrix}$$

$$\begin{bmatrix} 5 & 2 & 2 \\ 1 & 1 & 2 \end{bmatrix} \times \begin{bmatrix} 2 & 3 \\ 6 & 4 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 24 & 25 \\ 10 & 9 \end{bmatrix}$$

Example

The inverse of $\mathbf{A} = \begin{bmatrix} 3 & 0 \\ 0 & 5 \end{bmatrix}$ is $\mathbf{A}^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/5 \end{bmatrix}$ as $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$

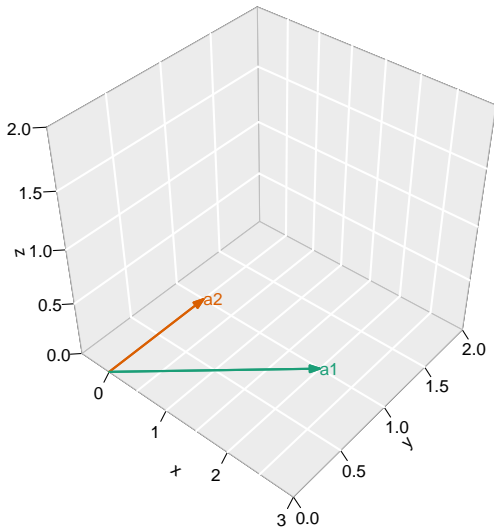
The inverse of \mathbf{I} is itself $\mathbf{I}^{-1} = \mathbf{I}$

Span, linear independence

- **linear combination** is just sum of some scaled vectors
 - $\lambda_1 \cdot \mathbf{a}_1 + \lambda_2 \cdot \mathbf{a}_2 + \dots + \lambda_n \mathbf{a}_n$, $\mathbf{a}_i \in R^m$ for $i = 1, \dots, n$
 - \mathbf{a}_i are vectors (of the same length) and λ_i are the scalars
- **span** is the set of all possible linear combination

$$\text{Span}(\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n\}) = \left\{ \sum_{i=1}^n \lambda_i \mathbf{a}_i \mid \lambda_i \in R, i = 1, \dots, n \right\}$$

- what is the span of $\{[1, 0]^T, [0, 1]^T\}$?
- how about $\{[2, 1]^T, [0, 1]^T\}$?
- how about $\{[2, 1]^T, [4, 2]^T\}$?
- how about $\{[2, 1, 0]^T, [0, 1, 0]^T\}$?
 - ▶ it is a **subspace** (bottom plane) in R^3



- **linear independence:** $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$ is linear independent if there exist no $\lambda_1, \dots, \lambda_n$ (except all being 0) such that

$$\lambda_1 \cdot \mathbf{a}_1 + \lambda_2 \cdot \mathbf{a}_2 + \dots + \lambda_n \mathbf{a}_n = \mathbf{0}$$

- how about $\{[2, 3]^T, [4, 6]^T\}$?
 - are $\{[1, 0]^T, [0, 1]^T\}$ LI?
 - essentially a way to tell whether there is any *redundant* vectors in the set
- **rank** of a matrix is defined as the maximum number of linearly independent column vectors

Example

The column vectors of the matrix

$$[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

are not linearly independent, as

$$\lambda_1 \mathbf{a}_1 + \lambda_2 \mathbf{a}_2 + \lambda_3 \mathbf{a}_3 = \mathbf{0}$$

holds for $\lambda_1 = \lambda_2 = 1, \lambda_3 = -2$. In other words, one of them is redundant. And $\text{rank}(\mathbf{A}) = 2$

Matrix vector multiplication

$$\mathbf{Ax} = \begin{bmatrix} | & | & \dots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & \dots & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} | \\ \mathbf{a}_1 \\ | \end{bmatrix} + x_2 \begin{bmatrix} | \\ \mathbf{a}_2 \\ | \end{bmatrix} + \dots + x_n \begin{bmatrix} | \\ \mathbf{a}_n \\ | \end{bmatrix}$$

- another view of the multiplication
- *linear combination* of the column vectors of \mathbf{A}
 - $x_1 \cdot \mathbf{a}_1 + x_2 \cdot \mathbf{a}_2 + \dots + x_n \mathbf{a}_n$
 - \mathbf{a}_i are the column vectors and x are the scalars
- so ... $\mathbf{Ax} = \mathbf{y}$ essentially solves for what ?

Matrix vector multiplication

$$\mathbf{Ax} = \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} | \\ \mathbf{a}_1 \\ | \end{bmatrix} + x_2 \begin{bmatrix} | \\ \mathbf{a}_2 \\ | \end{bmatrix} + \dots + x_n \begin{bmatrix} | \\ \mathbf{a}_n \\ | \end{bmatrix}$$

- another view of the multiplication
- *linear combination* of the column vectors of \mathbf{A}
 - $x_1 \cdot \mathbf{a}_1 + x_2 \cdot \mathbf{a}_2 + \dots + x_n \mathbf{a}_n$
 - \mathbf{a}_i are the column vectors and x are the scalars
- so ... $\mathbf{Ax} = \mathbf{y}$ essentially solves for what ?
 - \mathbf{y} is in the column space of \mathbf{A} or not ...
 - if not, then there is no solution
 - if yes, there will be some solution(s)? unique solution or ?

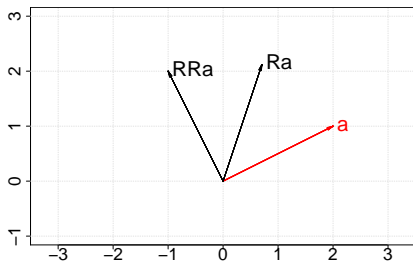
Matrix vector multiplication (some interpretations)

So \mathbf{Ax} is a linear transformation: $\mathbf{x} \rightarrow \mathbf{y}$

- Rotation: rotate \mathbf{x} anti-clockwise by θ

$$R\mathbf{x} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

say $\theta = \pi/4$



- \mathbf{R} is a rotation or orthogonal matrix if $\mathbf{R}^T = \mathbf{R}^{-1}$ (what does it imply?)
 - $\mathbf{R}^T \mathbf{R} = \mathbf{R} \mathbf{R}^T = \mathbf{I}$

► preserves length $(R\mathbf{x})^T (R\mathbf{x}) = \mathbf{x}^T \mathbf{R}^T \mathbf{R} \mathbf{x} = \mathbf{x}^T \mathbf{x}$

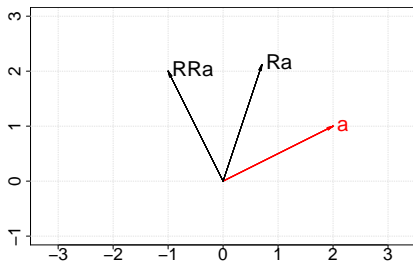
Matrix vector multiplication (some interpretations)

So \mathbf{Ax} is a linear transformation: $\mathbf{x} \rightarrow \mathbf{y}$

- Rotation: rotate \mathbf{x} anti-clockwise by θ

$$R\mathbf{x} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

say $\theta = \pi/4$



- \mathbf{R} is a rotation or orthogonal matrix if $\mathbf{R}^T = \mathbf{R}^{-1}$ (what does it imply?)
 - $\mathbf{R}^T \mathbf{R} = \mathbf{R} \mathbf{R}^T = \mathbf{I}$
 - ▶ preserves length $(\mathbf{R}\mathbf{x})^T (\mathbf{R}\mathbf{x}) = \mathbf{x}^T \mathbf{R}^T \mathbf{R} \mathbf{x} = \mathbf{x}^T \mathbf{x}$

- Projection (an example): project \mathbf{x} to \mathbf{a}

$$\begin{aligned} P(\mathbf{x}, \mathbf{a}) &= \|\mathbf{x}\| \cos\theta * \frac{\mathbf{a}}{\|\mathbf{a}\|} = \frac{\mathbf{a}^T \mathbf{x}}{\mathbf{a}^T \mathbf{a}} \mathbf{a} \\ &= \frac{\mathbf{a} \cdot \mathbf{a}^T \mathbf{x}}{\mathbf{a}^T \mathbf{a}} = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} \mathbf{x} \end{aligned}$$

- $\mathbf{P} = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}}$ is a projection matrix (what is the shape of \mathbf{P} ?);
- it transforms \mathbf{x} to its projection
- what if we project it again (and again and again ...) ? i.e.

$\mathbf{P}(\mathbf{P}\mathbf{x})$

▶ it remains unchanged, $\mathbf{P}\mathbf{P}\mathbf{x} = \mathbf{P}\mathbf{x}$

▶ or $\mathbf{P}\mathbf{P} = \mathbf{P}$

$$\frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} = \frac{\mathbf{a} \mathbf{a}^T \mathbf{a} \mathbf{a}^T}{(\mathbf{a}^T \mathbf{a})^2} = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}}$$

▶ mathematics is the subject of making sense :-)

- Projection (an example): project \mathbf{x} to \mathbf{a}

$$\begin{aligned} P(\mathbf{x}, \mathbf{a}) &= \|\mathbf{x}\| \cos\theta * \frac{\mathbf{a}}{\|\mathbf{a}\|} = \frac{\mathbf{a}^T \mathbf{x}}{\mathbf{a}^T \mathbf{a}} \mathbf{a} \\ &= \frac{\mathbf{a} \cdot \mathbf{a}^T \mathbf{x}}{\mathbf{a}^T \mathbf{a}} = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} \mathbf{x} \end{aligned}$$

- $\mathbf{P} = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}}$ is a projection matrix (what is the shape of \mathbf{P} ?);
- it transforms \mathbf{x} to its projection
- what if we project it again (and again and again ...) ? i.e.

$\mathbf{P}(\mathbf{P}\mathbf{x})$

- ▶ it remains unchanged, $\mathbf{P}\mathbf{P}\mathbf{x} = \mathbf{P}\mathbf{x}$
- ▶ or $\mathbf{P}\mathbf{P} = \mathbf{P}$

$$\frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}} = \frac{\mathbf{a} \mathbf{a}^T \mathbf{a} \mathbf{a}^T}{(\mathbf{a}^T \mathbf{a})^2} = \frac{\mathbf{a} \mathbf{a}^T}{\mathbf{a}^T \mathbf{a}}$$

- ▶ mathematics is the subject of making sense :-)

Probability theory

- Random variable
- Probability distribution
- Probability mass function and density function
- Probability rules
- Expectation, variance, covariance
- Conditional expectation

Random variable and probability distribution

Random variable X associates with a **probability distribution** $P(X)$

- formally, a r.v. is a mapping from sample space Ω to a target space \mathcal{T}
- e.g. toss a fair coin twice, r.v. X is the number of heads turned up
 - the sample space is $\Omega = \{HH, TT, HT, TH\}$
 - target space is $\mathcal{T} = \{0, 1, 2\}$
 - the probability distribution is

$$P(X) = \begin{cases} 0.25 & X = 0 \\ 0.5 & X = 1 \\ 0.25 & X = 2 \end{cases}$$

- the distribution P must satisfy

$$P(X = x) > 0, \text{ and } \sum_{x \in \mathcal{T}} P(X = x) = 1$$

Random variable - discrete r.v.

If r.v. X 's target space \mathcal{T} is discrete

- X is called **discrete random variable**
- the probability distribution P is called **probability mass function** (p.m.f.)
- and

$$0 \leq P(X = x) \leq 1, \text{ and } \sum_{x \in \mathcal{T}} P(X = x) = 1$$

Example - discrete r.v.

Bernoulli distribution Tossing a coin, $\mathcal{T} = H, T$,

$$P(X = H) = p, P(X = T) = 1 - p, 0 \leq p \leq 1$$

Binomial distribution Tossing a coin N times, the r.v. X that the number of head shows up is

$$P(X = k) = \binom{N}{k} \cdot p^k (1 - p)^{N-k}$$

(convince yourself why)

Multinoulli distribution Throw a fair 6-facet die, $\mathcal{T} = 1, 2, \dots, 6$, the distribution is

$$P(X = i) = 1/6$$

Verify the above P s satisfy the requirements of p.m.f.

Random variable - continuous r.v.

If r.v. X 's target space \mathcal{T} is continuous

- X is called **continuous random variable**
- the probability distribution p is called **probability density function** (p.d.f.)
- and satisfies

$$p(x) \geq 0, \text{ and } \int_{x \in T} p(x) dx = 1$$

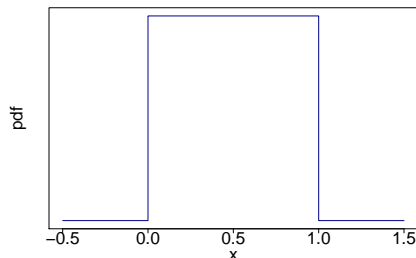
- pdf is not probability as $p(x)$ can be greater 1;
- for $\forall x \ P(X = x) = 0$
- calculate probability over an interval: e.g.

$$P(X \in [a, b]) = \int_a^b p(x) dx$$

Example - continuous r.v.

Uniform distribution $\mathcal{T} = [0, 1]$, X has equal chance to take any value between 0 and 1; the pdf is

$$p(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$



Easy to verify $\int_0^1 p(x) dx = \int_0^1 dx = 1$

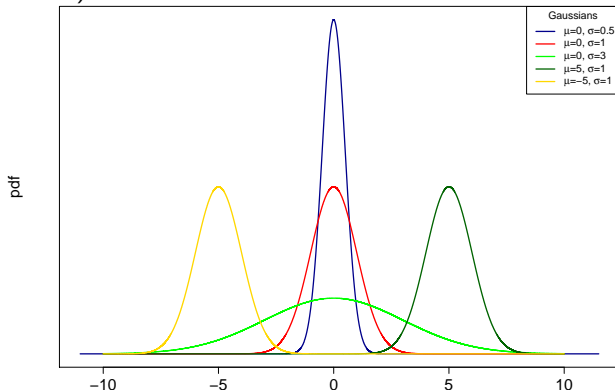
What's the probability that $0 < X < 0.5$?

Example - continuous r.v.

Gaussian distribution $\mathcal{T} = R$, or $X \in R$ the pdf is

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

$\left(\frac{x-\mu}{\sigma}\right)^2$ is a distance measure: how far x is away from μ (measured by σ as a unit)



Question

Calculate quickly:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = ?$$

For $X \sim \mathcal{N}(\mu, \sigma)$, what is $P(X < \mu) = ?$

Joint distribution

- r.v. $\mathbf{X} = [X_1, X_2, \dots, X_n]^T$ can be multidimensional (each X_i is r.v.)
 - essentially a *random vector*
- Still satisfies the same requirements

$$\forall \mathbf{x}, 0 < P(\mathbf{X} = \mathbf{x}) < 1, \sum_{x_1} \sum_{x_2} \dots \sum_{x_n} P(\mathbf{X} = [x_1, x_2, \dots, x_n]) = 1$$

or

$$\forall \mathbf{x}, p(\mathbf{X} = \mathbf{x}) > 0, \int \int \dots \int p(\mathbf{X} = \mathbf{x}) dx_1 dx_2 \dots dx_n = 1$$

- for bivariate case, i.e. $n = 2$, X_1, X_2 are **independent** if $P(\mathbf{X}) = P(X_1)P(X_2)$ (e.g. rolling two dice independently)

Example: discrete joint distribution

The joint distribution of X snow or not, $Y \in \{\text{spring, summer, autumn, winter}\}$ represents the season that x belongs to :

	$y = \text{Spring}$	$y = \text{Summer}$	$y = \text{Autumn}$	$y = \text{winter}$
$x = F$	0.05	0.25	0.075	0
$x = T$	0.2	0	0.175	0.25

It is easy to verify that

$$\sum_x \sum_y p(x, y) = 1$$

Example: continuous joint distribution

If X, Y 's joint p.d.f is

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}$$

X, Y are bivariate Gaussian distributed (X, Y are *independent*).

Probability rules

There are only two probability rules (use integration instead of sum for continuous r.v.):

1. Product rule:

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

2. Sum rule (marginalisation):

$$p(x) = \sum_y p(x, y), \quad p(y) = \sum_x p(x, y)$$

Conditional probability

Conditional probability (distribution) by product rule:

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

- probability distribution of x conditional on the value of y

	$y = \text{Spring}$	$y = \text{Summer}$	$y = \text{Autumn}$	$y = \text{winter}$
$x = F$	0.05	0.25	0.075	0
$x = T$	0.2	0	0.175	0.25

- $P(Y = \text{Spring})$? use sum rule $P(Y = \text{Spring}) = \sum_{x=\{T,F\}} P(X = x, Y = \text{Spring}) = 0.05 + 0.2 = 0.25$
- $P(X = T | Y = \text{Spring})$?
 $P(x = T | y = \text{Spring}) = \frac{P(x=T, y=\text{Spring})}{P(y=\text{Spring})} = \frac{0.05}{0.25} = 0.2$

Expectation and variance

Expection of a r.v. is defined as

$$E[X] = \sum_x xP(x) \text{ or } E[X] = \int xP(x)dx$$

Variance of a r.v. is defined as

$$\text{var}[X] = \sum_x x^2P(x) - (E[X])^2 \text{ or } \text{var}[X] = \int x^2P(x)dx - (E[X])^2$$

Reference