# 1 Introduction and Motivation

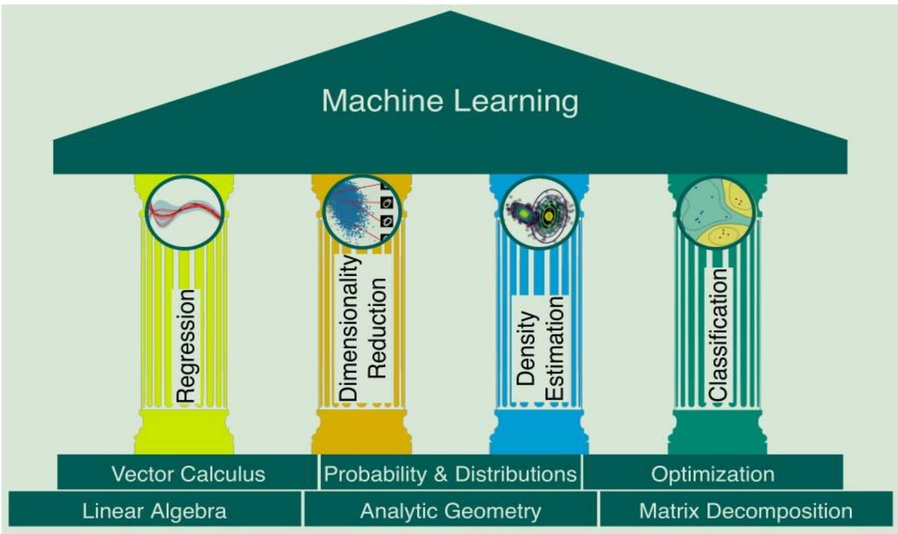
## 1.1 Finding Words for Intuitions

Predictors、
training、
Type checking allows the reader to sanity 5check whether the equation that they are considering contains inputs and outputs of the correct type, and whether they are mixing different types of objects.
data ： we assume that the data has already been data appropriately converted into a numerical representation。
Models are simplified versions of reality, which capture model aspects of the real world that are relevant to the task。

Let us summarize the main concepts of machine learning:
- We use domain knowledge to represent data as vectors.
- We choose an appropriate model, either using the probabilisitic or optimization view.
- We learn from past data by using numerical optimization methods with the aim that it performs well on unseen data.

## 1.2 Two Ways to Read this Book

Two strategies for understanding the mathematics for machine learning:
- Building up the concepts from foundational to more advanced.




- Drilling down from practical needs to more basic requirements.

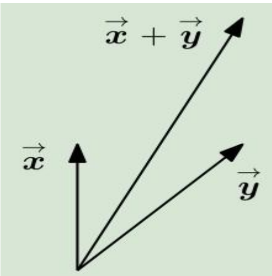
Part I is about Mathematics
Chapter 2 linear algebra
Chapter 3 analytic geometry
Chapter 4 matrix decomposition
Chapter 5 calculus
Chapter 6 probability theory
Chapter 7 optimization
Part II is about Machine Learning

| | Supervised | Unsupervised |
|---|---|---|
| Continuous latent variables | Regression (Chapter 9) | Dimensionality reduction (Chapter 10) |
| Categorical latent variables | Classification (Chapter 12) | Density estimation (Chapter 11) |

1.3 Exercises and Feedback

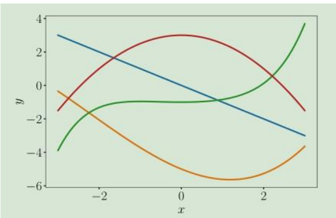
# 2 Linear Algebra

In general, vectors are special objects that can be added together and multiplied by scalars to produce another object of the same kind. Any object that satisfies these two properties can be considered a vector.
some examples of such vector objects:
Geometric vectors




(a) Geometric vectors.

Polynomials are also vectors




(b) Polynomials.

两个重要网站
https://www.youtube.com/playlist?list=PLlXfTHzgMRUKXD88IdzS14F4NxAZudSmv
https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/

**2.1 Systems of Linear Equations**

**2.2 Matrices**

**2.2.1 Matrix Addition and Multiplication**

**2.2.2 Inverse and Transpose**

**2.2.3 Multiplication by a Scalar**

- Distributivity:
  $(\lambda + \psi)\boldsymbol{C} = \lambda\boldsymbol{C} + \psi\boldsymbol{C}, \quad \boldsymbol{C} \in \mathbb{R}^{m\times n}$
  $\lambda(\boldsymbol{B} + \boldsymbol{C}) = \lambda\boldsymbol{B} + \lambda\boldsymbol{C}, \quad \boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{m\times n}$
- Associativity:
  $(\lambda\psi)\boldsymbol{C} = \lambda(\psi\boldsymbol{C}), \quad \boldsymbol{C} \in \mathbb{R}^{m\times n}$
  $\lambda(\boldsymbol{BC}) = (\lambda\boldsymbol{B})\boldsymbol{C} = \boldsymbol{B}(\lambda\boldsymbol{C}) = (\boldsymbol{BC})\lambda, \quad \boldsymbol{B} \in \mathbb{R}^{m\times n}, \boldsymbol{C} \in \mathbb{R}^{n\times k}$.
  Note that this allows us to move scalar values around.
- $(\lambda\boldsymbol{C})^\top = \boldsymbol{C}^\top\lambda^\top = \boldsymbol{C}^\top\lambda = \lambda\boldsymbol{C}^\top$ since $\lambda = \lambda^\top$ for all $\lambda \in \mathbb{R}$.

2.2.4 Compact Representations of Systems of Linear Equations

2.3 Solving Systems of Linear Equations

2.3.1 Particular and General Solution

2.3.2 Elementary Transformations

2.3.3 The Minus-1 Trick

2.3.4 Algorithms for Solving a System of Linear Equations

$$\boldsymbol{Ax} = \boldsymbol{b} \iff \boldsymbol{A}^\top\boldsymbol{Ax} = \boldsymbol{A}^\top\boldsymbol{b} \iff \boldsymbol{x} = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{b}$$

2.4 Vector Spaces

2.4.1 Groups

Groups play an important role in computer science. Besides providing a fundamental framework for operations on sets, they are heavily used in cryptography, coding theory and graphics.

**Definition 2.6** (Group). Consider a set $\mathcal{G}$ and an operation $\otimes : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ defined on $\mathcal{G}$.

Then $G := (\mathcal{G}, \otimes)$ is called a *group* if the following hold:

1. *Closure* of $\mathcal{G}$ under $\otimes$: $\forall x, y \in \mathcal{G} : x \otimes y \in \mathcal{G}$
2. *Associativity:* $\forall x, y, z \in \mathcal{G} : (x \otimes y) \otimes z = x \otimes (y \otimes z)$
3. *Neutral element:* $\exists e \in \mathcal{G}\, \forall x \in \mathcal{G} : x \otimes e = x$ and $e \otimes x = x$
4. *Inverse element:* $\forall x \in \mathcal{G}\, \exists y \in \mathcal{G} : x \otimes y = e$ and $y \otimes x = e$. We often write $x^{-1}$ to denote the inverse element of $x$.

If additionally $\forall x, y \in \mathcal{G} : x \otimes y = y \otimes x$ then $G = (\mathcal{G}, \otimes)$ is an *Abelian group* (commutative).

2.4.2 Vector Spaces

2.4.3 Vector Subspaces

2.5 Linear Independence

2.6 Basis and Rank

2.6.1 Generating Set and Basis

2.6.2 Rank

2.7 Linear Mappings

2.7.1 Matrix Representation of Linear Mappings

2.7.2 Basis Change

2.7.3 Image and Kernel

The mapping

$$\Phi : \mathbb{R}^4 \to \mathbb{R}^2, \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 + 2x_2 - x_3 \\ x_1 + x_4 \end{bmatrix} \tag{2.124}$$

$$= x_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.125}$$

is linear. To determine $\mathrm{Im}(\Phi)$ we can take the span of the columns of the transformation matrix and obtain

$$\mathrm{Im}(\Phi) = \mathrm{span}[\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}]. \tag{2.126}$$

To compute the kernel (null space) of $\Phi$, we need to solve $\boldsymbol{Ax} = \boldsymbol{0}$, i.e., we need to solve a homogeneous equation system. To do this, we use Gaussian elimination to transform $\boldsymbol{A}$ into reduced row echelon form:

$$\begin{bmatrix} 1 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \rightsquigarrow \cdots \rightsquigarrow \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}. \tag{2.127}$$

2.8 Affine Spaces

2.8.1 Affine Subspaces

2.8.2 Affine Mappings

# 3 Analytic Geometry

3.1 Norms

**Definition 3.1** (Norm)**.** A *norm* on a vector space $V$ is a function

$$\begin{aligned} \|\cdot\| : & V \to \mathbb{R}\,, && (3.1) \\ & \boldsymbol{x} \mapsto \|\boldsymbol{x}\|\,, && (3.2) \end{aligned}$$

which assigns each vector $\boldsymbol{x}$ its *length* $\|\boldsymbol{x}\| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following hold:

- Absolutely homogeneous: $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$
- *Triangle inequality*: $\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
- *Positive definite*: $\|\boldsymbol{x}\| \geqslant 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$.

Manhattan Norm

**Example 3.1 (Manhattan Norm)**
The *Manhattan norm* on $\mathbb{R}^n$ is defined for $\boldsymbol{x} \in \mathbb{R}^n$ as

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i|\,, \tag{3.3}$$

where $|\cdot|$ is the absolute value. The left panel of Figure 3.3 indicates all vectors $\boldsymbol{x} \in \mathbb{R}^2$ with $\|\boldsymbol{x}\|_1 = 1$. The Manhattan norm is also called $\ell_1$ *norm*.

Euclidean Norm

**Example 3.2 (Euclidean Norm)**
The length of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is given by

$$\|\boldsymbol{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}\,, \tag{3.4}$$

which computes the *Euclidean distance* of $\boldsymbol{x}$ from the origin. This norm is called the *Euclidean norm*. The right panel of Figure 3.3 shows all vectors $\boldsymbol{x} \in \mathbb{R}^2$ with $\|\boldsymbol{x}\|_2 = 1$. The Euclidean norm is also called $\ell_2$ *norm*.

3.2 Inner Products

3.2.1 Dot Product

**3.2.2 General Inner Products**

**3.2.3 Symmetric, Positive Definite Matrices**

3.3 Lengths and Distances

3.4 Angles and Orthogonality

$$-1 \leqslant \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|} \leqslant 1\,.$$

nique $\omega \in [0, \pi]$ with

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|}\,,$$

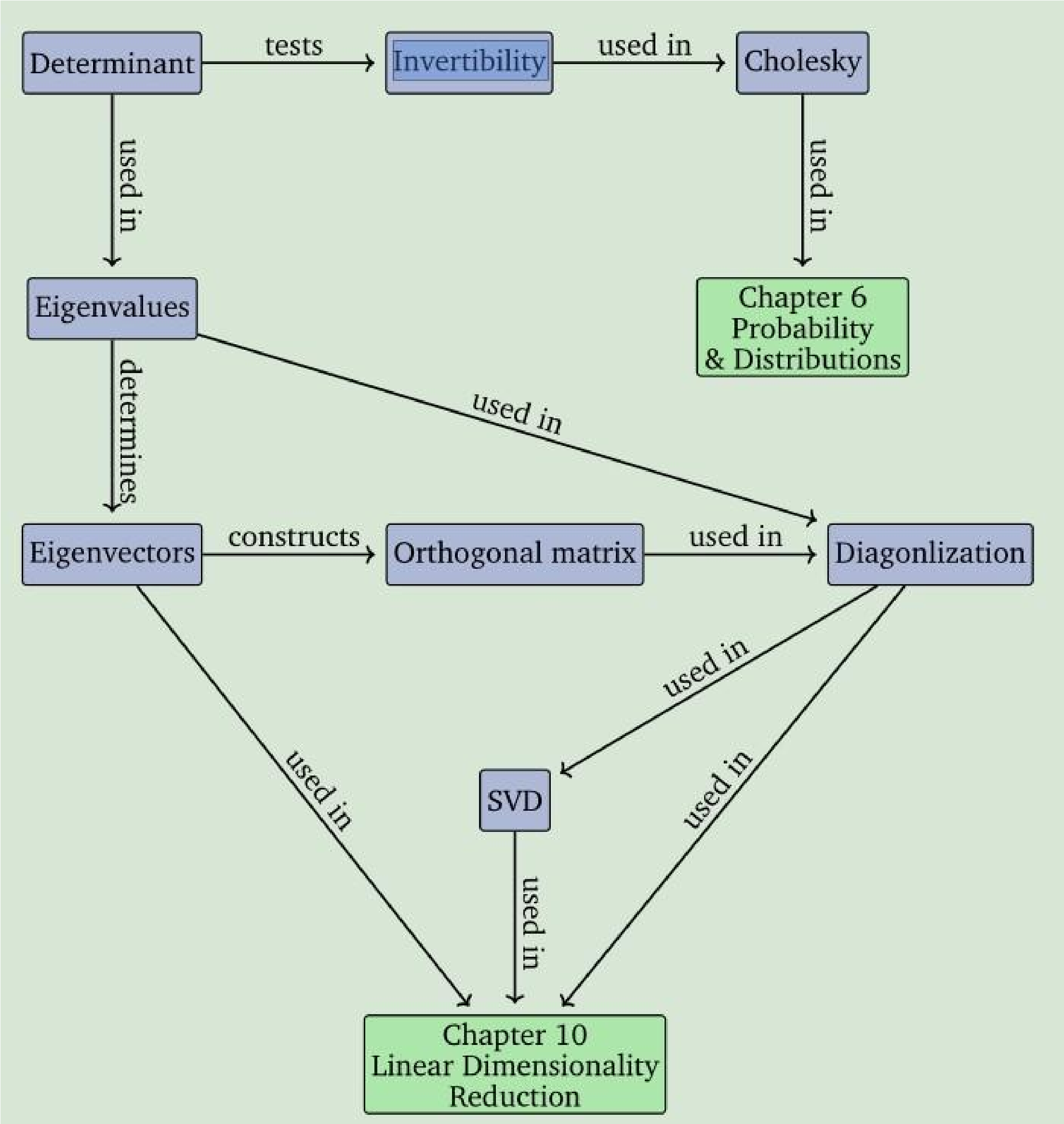
3.5 Orthonormal Basis

**3.6 Inner Product of Functions**

**3.7 Orthogonal Projections**

3.8 Rotations *******

# 4 Matrix Decompositions




4.1 Determinant and Trace

4.2 Eigenvalues and Eigenvectors

4.3 Cholesky Decomposition

**Theorem 4.17.** *Cholesky Decomposition: A symmetric positive definite matrix $\boldsymbol{A}$ can be factorized into a product $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a lower triangular matrix with positive diagonal elements:*

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & \cdots & l_{n1} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix}. \tag{4.43}$$

*$\boldsymbol{L}$ is called the Cholesky factor of $\boldsymbol{A}$.*

4.4 Eigendecomposition and Diagonalization

4.5 Singular Value Decomposition

4.6 Matrix Approximation

**4.7 Matrix Phylogeny**

# Vector Calculus

5.1 Differentiation of Univariate Functions

5.1.1 Taylor Series

5.1.2 Differentiation Rules

$$\text{Product Rule:} \quad (f(x)g(x))' = f'(x)g(x) + f(x)g'(x) \tag{5.37}$$

$$\text{Quotient Rule:} \quad \left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2} \tag{5.38}$$

$$\text{Sum Rule:} \quad (f(x) + g(x))' = f'(x) + g'(x) \tag{5.39}$$

$$\text{Chain Rule:} \quad \big(g(f(x))\big)' = (g \circ f)'(x) = g'(f(x))f'(x) \tag{5.40}$$

Here, $g \circ f$ is a function composition $x \mapsto f(x) \mapsto g(f(x))$.

5.2 Partial Differentiation and Gradients

5.2.1 Basic Rules of Partial Differentiation

$$\text{Product Rule:} \quad \frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x})g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} g(\boldsymbol{x}) + f(\boldsymbol{x}) \frac{\partial g}{\partial \boldsymbol{x}} \tag{5.54}$$

$$\text{Sum Rule:} \quad \frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}} \tag{5.55}$$

$$\text{Chain Rule:} \quad \frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}\big(g(f(\boldsymbol{x}))\big) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \boldsymbol{x}} \tag{5.56}$$

5.2.2 Chain Rule

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

5.3 Gradients of Vector-Valued Functions

For a function $f : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $\boldsymbol{x} = [x_1, \dots, x_n]^\top \in \mathbb{R}^n$, the corresponding vector of function values is given as

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m . \tag{5.64}$$

**Definition 5.6** (Jacobian)**.** The collection of all first-order partial derivatives of a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is called the *Jacobian*. The Jacobian $\boldsymbol{J}$ is an $m \times n$ matrix, which we define and arrange as follows:

$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \tag{5.67}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} , \tag{5.68}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} , \quad J(i,j) = \frac{\partial f_i}{\partial x_j} . \tag{5.69}$$

5.4 Gradients of Matrices

**Example 5.11 (Gradient of Vectors with Respect to Matrices)**
Let us consider the following example, where

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}\,, \quad \boldsymbol{f} \in \mathbb{R}^M, \boldsymbol{A} \in \mathbb{R}^{M\times N}, \boldsymbol{x} \in \mathbb{R}^N \tag{5.97}$$

and where we seek the gradient $d\boldsymbol{f}/d\boldsymbol{A}$. Let us start again by determining the dimension of the gradient as

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} \in \mathbb{R}^{M\times(M\times N)}\,. \tag{5.98}$$

By definition, the gradient is the collection of the partial derivatives:

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}, \quad \frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1\times(M\times N)}\,. \tag{5.99}$$

To compute the partial derivatives, it will be helpful to explicitly write out the matrix vector multiplication:

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \ldots, M\,, \tag{5.100}$$

and the partial derivatives are then given as

$$\frac{\partial f_i}{\partial A_{iq}} = x_q\,. \tag{5.101}$$

This allows us to compute the partial derivatives of $f_i$ with respect to a row of $\boldsymbol{A}$, which is given as

$$\frac{\partial f_i}{\partial A_{i,:}} = \boldsymbol{x}^\top \in \mathbb{R}^{1\times 1\times N}\,, \tag{5.102}$$

$$\frac{\partial f_i}{\partial A_{k\neq i,:}} = \boldsymbol{0}^\top \in \mathbb{R}^{1\times 1\times N} \tag{5.103}$$

5.5 Useful Identities for Computing Gradients

In the following, we list some useful gradients that are frequently required in a machine learning context (Petersen and Pedersen, 2012):

$$\frac{\partial}{\partial \boldsymbol{X}} f(\boldsymbol{X})^\top = \left(\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}\right)^\top \tag{5.112}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \operatorname{tr}(f(\boldsymbol{X})) = \operatorname{tr}\left(\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) \tag{5.113}$$

$$\frac{\partial}{\partial \boldsymbol{X}} \det(f(\boldsymbol{X})) = \det(f(\boldsymbol{X})) \operatorname{tr}\left(f^{-1}(\boldsymbol{X})\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) \tag{5.114}$$

$$\frac{\partial}{\partial \boldsymbol{X}} f^{-1}(\boldsymbol{X}) = -f^{-1}(\boldsymbol{X})\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} f^{-1}(\boldsymbol{X}) \tag{5.115}$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{X}^{-1} \boldsymbol{b}}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1})^\top \boldsymbol{a} \boldsymbol{b}^\top (\boldsymbol{X}^{-1})^\top \tag{5.116}$$

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}^\top \tag{5.117}$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}^\top \tag{5.118}$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a} \boldsymbol{b}^\top \tag{5.119}$$

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{B} \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^\top (\boldsymbol{B} + \boldsymbol{B}^\top) \tag{5.120}$$

$$\frac{\partial}{\partial \boldsymbol{s}} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W} \boldsymbol{A} \quad \text{for symmetric } \boldsymbol{W} \tag{5.121}$$

Here, we use tr as the trace operator (see Definition 4.3) and det is the determinant (see Section 4.1).

5.6 Backpropagation and Automatic Differentiation

5.6.1 Gradients in a Deep Network

5.6.2 Automatic Differentiation

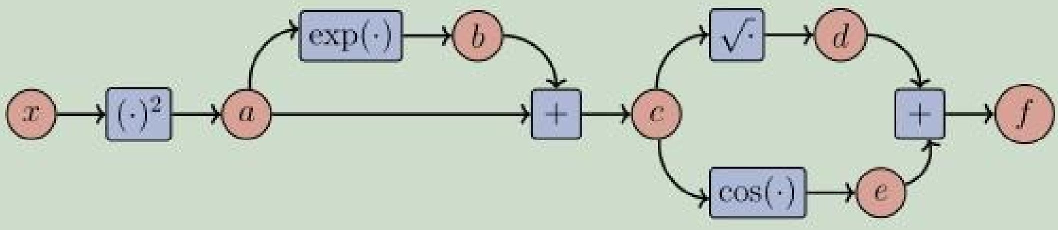
**Example 5.13**

Consider the function

$$f(x) = \sqrt{x^2 + \exp(x^2)} + \cos\left(x^2 + \exp(x^2)\right) \tag{5.135}$$

from (5.122). If we were to implement a function $f$ on a computer, we would be able to save some computation by using *intermediate variables*:

$$a = x^2\,, \tag{5.136}$$
$$b = \exp(a)\,, \tag{5.137}$$
$$c = a + b\,, \tag{5.138}$$
$$d = \sqrt{c}\,, \tag{5.139}$$
$$e = \cos(c)\,, \tag{5.140}$$
$$f = d + e\,. \tag{5.141}$$




5.7 Higher-order Derivatives

5.8 Linearization and Multivariate Taylor Series

# 6 Probability and Distributions

6.1 Construction of a Probability Space

6.1.1 Philosophical Issues

6.1.2 Probability and Random Variables