Mathematical Foundations of Machine Learning

Matias Valdenegro-Toro, University of Groningen

November 13, 2024

This document describes the mathematical fundamentals to effective learn the basics of Machine Learning. These fundamentals are Linear Algebra, Calculus, and Statistics and Probability. These notes are meant to refresh knowledge in specific topics within each subject, not to be a comprehensive reference. Writing is tailored to Bachelor students in Artificial Intelligence.

1 Introduction

1 Introduction **1**
1.1 Notation 2

2 Linear Algebra **2**
2.1 Vector Spaces 3
2.2 Linear Transformations 3
2.3 Matrices 4
2.4 Hyperplanes 5
2.5 Eigenvalues and Eigenvectors 5
2.6 Other Topics 6

3 Calculus **6**
3.1 Functions 6
3.2 Derivatives 7
3.3 Partial Derivatives . . 8
3.4 Gradients 8
3.5 Jacobian 9
3.6 Hessian 9
3.7 Taylor's Decomposition 9

4 Optimization **10**
4.1 Optimization and Derivatives 11
4.2 Gradient Descent . . . 12
4.3 Convexity 12

5 Statistics and Probability **13**
5.1 Summary Statistics . . 13
5.2 Probability 13
5.3 Random Variables . . 14
5.4 Probability Distributions 15

6 Relationship with IML Course **17**

7 Further Reading **18**

Machine learning is often seen as an applied discipline, working with data and models using python libraries, but it is also a field backed by mathematical and statistical concepts. It is important for students learning about machine learning concepts to have the proper mathematical background to understand and apply basic machine learning models.

Linear Algebra, Calculus, and related topics in mathematics are fundamental for proper understanding of machine learning concepts, as these are expressed in mathematical form, and contain the key intuitions and explanations that underpin machine leanring concepts.

The mathematical level required for understanding basic machine learning is not far from standard high school mathematics, two main topics are often covered at the university level, namely linear algebra and calculus, while other topics like optimization and statistics are nice-to-have for working with data and connect with neighboring fields.

These notes are meant for students in the "Introduction to Machine Learning (for AI) at the Department of Artificial Intelligence at the University of Groningen, and it is a refresher on content you should already have from past courses, in particular from Linear Algebra and Multivariable Calculus.

These notes contain summaries and conceptual understanding of the following topics:

Linear Algebra. This is basic understanding of vectors and matrices and operations that apply to them, including their geometric interpretation. This section is a refresher as you probably had a linear algebra course previously.

Calculus. Basic understanding of how functions change in one and multiple variables, covering derivatives and integrals. We also assume that you know basic calculus from a previous course. For machine learning the most important concept is derivatives and gradients, as they are used for optimization.

Optimization. How to obtain an input that minimizes or maximizes a function, and in the context of machine learning, relates to

Symbol	Name	Typical Use / Denotation
μ	Mu	Mean value, mean of probability distribution.
σ	Sigma	Standard deviation, spread of probability distribution.
θ	Theta	Model parameters.
λ	Lambda	Regularization strength coefficient, Lagrange multipliers, Eigenvalues.
$\hat{y}$	y hat	Estimated values, the hat indicates estimation.
$\bar{y}$	y bar	Mean value.
$\tilde{y}$	y tilda	Derivative, alternate value to y.

Table 1: Common mathematical symbols used in Statistics and Machine Learning.

the training process of learning algorithms. We do not assume that you had a course on this topic, and some topics will be covered in the course, but these notes expand on important concepts as well.

Statistics and Probability. Statistics is the science of working with data, and is one of the building blocks of machine learning. Probability is a mathematical language that helps to model uncertainty.

Note that these notes should not be considered mathematically rigorous, we describe concepts using mathematical notation and give layman intuitions about them, but for proofs and strict rigour, we refer the reader to a mathematics books, we provide references in the last section.

We will highlight key concepts that you need for machine learning.

1.1 Notation

Vectors are often noted as bold lowercase letters ($\mathbf{u}, \mathbf{v}, \mathbf{w}$), while matrices are uppercase letters (A, B, C), and scalars are noted with lowercase letters (a, b, c).

Note that notation is a convention between humans, and it is often abused, for example the L^2 norm is not the L norm squared, here the superindex is used to denote the degree, which is an abuse of notation of superindices for numerical powers. Table 1 shows some commonly used symbols, their official names, and what they often denote.

2 Linear Algebra

Linear algebra is everything about linear functions and structures, as vector spaces and matrices.

Addition Associativity	$u + (v + w) = (u + v) + w$
Addition Commutativity	$u + v = v + u$
Addition Identity Element	There exists vector $\mathbf{0} \in V$, the zero vector, for which $\mathbf{0} + v = v$.
Addition Inverse Element	For $v \in V$ there exists $-v \in V$, the additive inverse, such that $v + (-v) = \mathbf{0}$.
Scalar Distributivity	$a(u + v) = au + av$ $(a + b)v = av + bv$
Scalar Associativity	$a(bv) = (ab)v$
Product Identity Element	$1v = v$.

Table 2: Axioms related to vector spaces, where $u, v, w \in V$ and $a, b \in \mathbb{R}$

2.1 *Vector Spaces*

A vector space is a set $V \in \mathbb{R}^n$ which follows some properties. The elements $v \in V$ are called vectors, and the individual components of v are called scalars. A vector is formed by several scalar components in $\mathbb{R}$. For $v, u \in V$ and c a scalar, then:

Vector Addition $v \in V, u \in V$ implies that $u + v \in V$

Scalar Multiplication $v \in V$ and $a \in \mathbb{R}$ implies that $av \in V$.

Conceptually these properties mean that a vector space is closed under addition, meaning adding two vectors produces a new vector in the same vector space, while a vector space is also closed under scalar multiplication, so the product of a vector and scalar is also a vector in the same vector space.

There are many axioms related to vector spaces, the most important ones are summarized in Table 2.

One very common operations in vector spaces is the dot product, defined for $x, y \in V$:

$$\text{dot}(\mathbf{x}, \mathbf{y}) = x \cdot y = \sum_{i}^{n} \mathbf{x}_i \mathbf{y}_i \tag{1}$$

Norms are defined in vector spaces as a way to measure distance and vector lengths. The p-norm is defined as, where p is the norm order:

$$||x||_p = \left(\sum^{N} |x_i|^p \right)^{1/p} \tag{2}$$

For $p = 2$ we have the standard euclidean distance.

2.2 *Linear Transformations*

An operator or operation $f : V \to W$ is called linear when it fulfills the following properties, for $x, y \in V$ and $a \in \mathbb{R}$ and V, W are vector spaces:

1. $f(x + y) = f(x) + f(y)$
2. $f(ax) = af(x)$

Addition	$(A + B)_{ij} = A_{ij} + B_{ij}$
Scalar Multiplication	$(cA)_{ij} = cA_{ij}$
Substraction	$(A - B)_{ij} = A + (-1)B = A_{ij} - B_{ij}$
Transpose	$(A)^T_{ij} = (A)_{ji}$
Matrix Multiplication	$(C)_{ij} = \sum_k A_{ik} B_{kj}$

Table 3: Common mathematical operations on matrices, where A and B are matrices

Intuitively this means the operation is linear, that is, it can be "distributed" when applied to a sum of vectors, and it can also be "distributed" with respect to scalar product.

Many non-intuitive operations are linear, such as derivatives, integrals, etc.

2.3 *Matrices*

Matrices are rectangular arrays of numbers, in a sense similar to 2D arrays. Elements are arranged into rows and columns, and each element has a 2D index, usually denoted as C_{ij}, where i is the row index, and j is the column index. The size of a matrix is determined by its number of rows N and number of columns M.

There are multiple ways to denote matrices and their elements. Generally uppercase letters are used for a matrix, and the corresponding lowercase letter is used for its elements:

$$A = (a_{ij}) \text{ or } A = [a_{ij}] \tag{3}$$

Which describe the following matrix:

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \tag{4}$$

Square matrices are the ones where $N = M$ holds. The diagonal of a matrix are the elements where row and column indices are equal $i = j$. One specially named matrix is the identity matrix, denoted as $\mathbf{I}$, where all its elements are zero, except for the diagonal which contain only the value 1.0.

Some operations on matrices and their definitions are shown in Table 3. An important operation is matrix multiplication, which has different intuitions than standard multiplication. For matrices A_{nm} and B_{mp}, then $C = AB$ is given by:

$$C_{ij} = \left(\sum_k A_{ik} B_{kj} \right)_{ij} \tag{5}$$

Matrix multiplication is only possible if matrices have compatible sizes, which means the number of columns of the first operand must be equal to the number of rows of the second operand.

Note that in general, matrix multiplication is not commutative, meaning $AB \neq BA$.

Related to matrix multiplication, there is the matrix inverse, denoted as A^{-1}, where:

$$AA^{-1} = A^{-1}A = I \tag{6}$$

Usually only square matrices have the inverse defined for them, but not every square matrix has a defined matrix inverse. If a matrix has a defined inverse, it is called invertible, and if not, it is called singular.

Matrices and Linear Transformations

Matrices are related to linear transformations, where a linear transformation on a finite space can be represented as multiplication with a matrix:

$$f(x) = Ax \tag{7}$$

2.4 Hyperplanes

A hyperplane is defined by a linear equation, given $x, w \in \mathbb{R}^n$ and $b \in \mathbb{R}$:

$$\mathbf{wx} + b = 0 \tag{8}$$

Where x is a point in the hyperplane, w is a vector that defines the hyperplane direction, and b is a scalar. Points x in the hyperplane are defined by equation 8.

A hyperplane divides a vector space into two subspaces, defined by $wx + b < 0$ and $wx + b > 0$.

Hyperplanes are fully defined by w and b. The geometric interpretation is that w is the normal vector (perpendicular) to the hyperplane, and b defines the hyperplane position on space.

2.5 Eigenvalues and Eigenvectors

An important concept related to matrices are its eigenvalues and eigenvectors, which define some geometrical properties of a linear transformation defined by that matrix.

Consider two vectors $u, v \in V$, these are called parallel or scalar multiples of each other if there is λ such that:

$$u = \lambda v \tag{9}$$

Now consider a linear transformation $f(x) = Ax$ defined by a matrix A. If there exists λ such that the linear transformation produces a scalar multiple of its input:

$$f(v) = Av = \lambda v \tag{10}$$

Then λ is called an eigenvalue of the transformation or matrix A and v is called an eigenvector associated to that eigenvalue. The geometric intepretation is that the transformed vector is parallel and scaled (by λ amount) to the input.

To find eigenvalues for a matrix A, first we must solve the characteristic polynomial, defined as:

$$|A - \lambda I| = \det(A - \lambda I) = 0 \tag{11}$$

Where || or det are the matrix determinant operation. This is a N degree polynomial, with N solutions that each are eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$

Then for each eigenvalue λ, there is an associated eigenvector v. First we should note that there are infinitely many eigenvectors associated to a single eigenvalue, and these are defined as the set E:

$$E = \{v : (A - \lambda I)v = 0\} \tag{12}$$

Usually an eigenvector v associated to a particular eigenvalue λ is represented as a single value that forms the basis for the eigenvector space E, such that other eigenvectors are simply scalar multiples of v.

Generally in Machine Learning applications, the eigenvalues are much more important for several properties than eigenvectors, and often we use computational implementations to compute them.

2.6 Other Topics

Some linear algebra topics were not covered in this summary, namely systems of linear equations, matrix decompositions, linear independence, etc, or detailed computations and properties in many cases. These are often not needed for basic Machine Learning understanding, where concepts are more important than detailed calculations and proofs.

3 Calculus

Calculus is the study of functions and how they change. It is one of the main pillars supporting Machine Learning.

3.1 Functions

Functions are mappings from one set X to another set Y, with the restriction that mapping is only to one element in Y to each element of X. In other words, functions produce a single value per input, and not multiple values.

X is called the domain, and is the set of valid inputs to the function, and Y is called the codomain or range, and is the set of values that can be produced by the function.

Functions are often denoted as $f : X \to Y$, which compactly represents many aspects of the mapping, and explicitly puts a name to the function f.

The mapping is often denoted as $y = f(x)$, where $x \in X$ and $y \in Y$, and x and y represent values for the input and output of the function correspondingly, and other names for x are value of the independent variable, and for y are value of the function or the dependent variable.

Scalar and Vector Functions

The most interesting functions in Machine Learning are scalar and vector functions over the real numbers $\mathbb{R}$. A scalar function is

$f : \mathbb{R}^n \to \mathbb{R}$, which maps a n-dimensional vector of real numbers to a single real number.

A vector function is $f : \mathbb{R}^n \to \mathbb{R}^m$, which maps n-dimensional vectors to m-dimensional vectors.

There are many properties of functions that are not relevant at this point, we will only mention the definition of continuity.

Continuity

A function f is continuous at $x = x_0$ if:

$$\lim_{x \to x_o} f(x) = f(x_0) \tag{13}$$

In conceptual terms, a function is continuous if it can be drawn without lifting a pen, meaning, there are no discontinuities or "jumps" in its values.

3.2 *Derivatives*

The derivative is an important mathematical tool to compute the rate of change of a single variable scalar function $f : \mathbb{R} \to \mathbb{R}$ with respect to its input variable x. Mathematically the derivative is defined as the following limit:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{14}$$

This means that the derivative $f'(x)$ is the infinitesimal rate of change of $f(x)$, denoted as the limit computes the difference between $f(x+h)$ and $f(x)$ and divides it by the infinitesimal h, as the value of h goes to zero.

The derivative is often denoted as $\frac{df}{dx}$, indicating the derivative of f with respect to variable x. The notation $\frac{d}{dx}$ is often used as an operator that operates on functions, by computing the derivative of that function.

Other notations used for the derivative are using the prima ($'$), as f' for derivative of f, and the dot ($\dot{f}$) for derivatives with respect to time[1].

[1] Often used in control theory

The derivative of a function produced another function, which can also be derived, producing higher order derivatives, for example:

$$\frac{d}{dx}\frac{df}{dx} = \frac{d^2 f}{dx^2} \qquad \frac{d}{dx}\frac{d^2 f}{dx^2} = \frac{d^3 f}{dx^3} \tag{15}$$

In practice we do not compute derivatives using the definition but with defined rules derived from the definition, a selection of them is available in Table 4.

The derivative is a linear operation, this means:

$$\frac{d}{dx}\left[af(x) + bg(x)\right] = a\frac{d}{dx}f(x) + b\frac{d}{dx}g(x) \tag{16}$$

Where a and b are scalar constants and f and g are functions.

A special and very important rule is the chain rule, which is the derivative of a function composition $f(g(x))$.

$f(x)$	$f'(x)$	$f(x)$	$f'(x)$
x	1	c	0
f g	f′g + f g′	$\frac{f}{g}$	$\frac{f'g-fg'}{g^2}$
$\log(x)$	$\frac{1}{x}$	e^x	e^x

Table 4: Summary of most rules for computing derivatives of single variable functions, where f and g are functions of x.

$$\frac{d}{dx}f(g(x)) = f'(g(x))g'(x) \tag{17}$$

The chain rule means, to compute the derivative of a function composition $f(g(x))$, compute the derivative of f.

Differentiability

An important concept is that derivatives do not always exist, there are constraints introduced into functions to be able to be differentiated, that is, differentiable.

A function $f(x)$ is differentiable at $x = x_0$, if and only if it is continuous at $x = x_0$. In the context of machine learning, a function is called differentiable if its differentiable on its whole domain.

This means that functions that are not continuous, are not differentiable.

Examples of functions that are not differentiable are: the indicator function ($f(x) = \mathbb{1}[x = 0]$), the step function, etc. Usually functions that have discontinuous sections are not differentiable, like functions defined for discrete data. A clear example of this is the function that computes accuracy, which is not differentiable.

3.3 *Partial Derivatives*

Partial derivatives are the generalization of derivatives to scalar functions of multiple variables, like $f(x_1, x_2, ..., x_m)$. They are denoted using a different symbol: ∂.

A partial derivative is computed on a specific variable, denoted as $\frac{\partial f}{\partial x_1}$, or $\frac{\partial f}{\partial x_2}$, or $\frac{\partial f}{\partial x_m}$, etc, and the partial derivative computation is done by treating all other variables as constant, only computing the derivative on a single selected variable.

The interpretation of a partial derivative $\frac{\partial f}{\partial x}$ is the rate of change for f given x and other variables are kept constant.

The rules for computing partial derivatives are the same as with standard derivatives, plus the extra rules of computing derivatives with respect to one variable only and keeping the rest constant.

It is also possible to compute higher level partial derivatives,

$$\frac{\partial}{\partial x}\frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial x^2} \qquad \frac{\partial}{\partial y}\frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial x \partial y} \tag{18}$$

3.4 *Gradients*

Gradient is the generalization of derivatives for scalar functions on multiple variables $f(x_1, x_2, ..., x_m)$, based on partial derivatives. Note that partial derivatives are computed for a single input variable at a time, so a function of m variables has m partial derivatives,

one for each input variable. The gradient generalizes the concept of partial derivative, computing all partial derivatives of a scalar function f.

The gradient is the vector formed with the partial derivatives of f with respect to each of its variables:

$$\nabla f = \left[\frac{\partial f}{x_1}, \frac{\partial f}{x_2}, ..., \frac{\partial f}{x_m}\right] \tag{19}$$

An important property of the gradient is that ∇f it points in the direction of maximum function increase, while $-\nabla f$ is the direction of maximum function decrease. These properties are very useful for optimization and training machine learning models.

3.5 Jacobian

What about vector functions? The generalization of gradients exists for this kind of function, and it is called the Jacobian.

Given a vector function $f : \mathbb{R}^n \to \mathbb{R}^m$, which we can decompose as $f(x_1, x_2, ..., x_n) = [f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_m(\mathbf{x})]$ [2], then the Jacobian is defined as:

[2] Note here that $\mathbf{x} = [x_1, x_2, ..., x_m]$

$$J(f)_{ij} = \frac{\partial f_i}{\partial x_j} \qquad i = 1...m, j = 1...n \tag{20}$$

The Jacobian is a $m \times n$ matrix that contains all the partial derivatives of the function, with respect to all variables and all sub-functions (the f_i's).

3.6 Hessian

A generalization of the second order derivatives to scalar functions is the Hessian. Given a scalar function $f : \mathbb{R}^n \to R$, the Hessian is defined as:

$$H(f)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \tag{21}$$

The Hessian is a matrix with dimensions $n \times m$ and each entry contains the corresponding cross partial derivatives. The Hessian interpretation is that it describes the local curvature of the function, just like the second derivative ($\frac{d^2 f}{dx^2}$) describes the function curvature in a single variable function.

The Hessian and Jacobian are related, the Hessian can be computed as the Jacobian of the gradient of a function.

$$H(f) = J(\nabla f)^T \tag{22}$$

3.7 Taylor's Decomposition

An approximation that is often used in machine learning, and helps understand more complex models, is the Taylor approximation, deriving from the Taylor series.

For The Taylor series is defined for a continuous function f as:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^n(a)}{n!}(x-a)^n \tag{23}$$

Where a is a point in the domain of f where derivatives are evaluated, and this assumes f is infinitely differentiable at a. If $a = 0$, this is called a Maclaurin series. Then the Taylor approximation corresponds to truncating the Taylor series to k terms:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + ... + \frac{f^k(a)}{k!}(x-a)^k + h_k(x) \tag{24}$$

Where $h_k(x)$ is the remainder function, representing the approximation error due to truncation of the Taylor series, and $\lim_{k\to\infty} h_k(x) = 0$

4 *Optimization*

Optimization is the art of finding variables that optimize (minimize or maximize) a certain function, subject to some constraints from a set of possible choices. Training a machine learning model involves an optimization problem, and in this section we describe the very basic theory of mathematical optimization for continuous variables that is useful in ML.

An optimization problem often has the form:

$$\min_{x \in D} f(x) \text{ subject to } C(x) > 0 \tag{25}$$

Where f is the function to be optimized, often called the objective function, D is the domain for x, meaning the set of valid values that x can take, and C is a function that represents constraints. Many constraints can be represented in this form.

The objective of an optimization algorithm when solving an optimization problem, is to find an optimal value x^*, also known as optima, that fulfils the constraints and minimizes or maximizes the objective function.

Note that an optimization problem might involve minimizing or maximizing a function, and these problems are equivalent as $\min_x f(x)$ is the same as $\max_x -f(x)$, meaning they have the same optima.

There are several types of optima:

Local Optima An optima that is only better (bigger or smaller) in a region around the optima, but its not the best value that the objective function can obtain. Mathematically $f(x^*) < f(x)$ for $\forall x$ subject to $||x^* - x|| < \delta$, with $\delta > 0$.

Global Optima An optima that obtains the best value of the objective function. Mathematically $f(x^*) < f(x)$ for $\forall x \in D$

The difficulty with optimization problems is that depending on the objective function's geometry, there can be multiple local min-

ima, which can "trap" the execution of an optimization algorithm and prevent the global optima from being found.

There are many methods to solve optimization problems with different domains and constraints, in these section we only cover two methods based on derivatives and gradients, as they are the most useful for machine learning, and what you will use

4.1 Optimization and Derivatives

It is possible to solve optimization problems on continuous variables using the objective function derivatives.

Single Variable Functions

A critical point of a single variable scalar function f are points x_0 where the derivative is zero, that is, $f'(x_0) = 0$ or where the function is not differentiable.

What is the relevance of critical points to optimization? Fermat's theorem says that all local optima of a continuous function happen at critical points.

Second Derivative Test

Critical points do not tell you if they are minima or maxima, to determine that we need to explore the curvature of the function at the critical point value . The second derivative ($\frac{d^2f}{dx^2}$ or just $f''(x)$) contains information about the curvature that can be used to infer minima or maxima:

- If $f''(x_0) < 0$, then x_0 is a local minima.
- If $f''(x_0) > 0$, then x_0 is a local maxima.
- If $f''(x_0) = 0$, then this method cannot make a conclusion.

A simple method to find optima is then:

1. Find all the critical points of the function by solving $f'(x) = 0$.
2. For each critical point, use the second derivative test to determine if they are local minima or maxima.
3. If the domain of the function is not $\mathbb{R}$, that means it is a bounded set D, then consider the extrema/bounds of that set as well.
4. To determine a global optima, test all the critical points and the extrema, the optimal value (highest/lowest) will be the global minima.

Multivariable Functions

These insights also apply for multivariable functions , in particular for scalar functions of multiple variables, critical points are defined as points x_0 where the gradient is the zero vector, that is $\nabla f = \mathbf{0}$. As a second derivative test, the eigenvalues of the Hessian matrix describe the local curvature. If all the eigenvalues of the Hessian matrix at x_0 are positive, then x_0 is a local minima. If all eigenvalues are negative, then it is a local maxima, and if eigenvalues are a mix of positive and negative values, then x_0 is a saddle point. In the case that the Hessian matrix is singular, the method is not conclusive.

4.2 Gradient Descent

The method we covered previously is analytical, which means it produces closed form equations for the optima, which is not possible when the objective function is very complex, like the ones used in neural networks.

An alternative method that is very popular is gradient descent, which allows to minimize (descent) or maximize (ascent) an objective function f only based on its gradient ∇f, using the following relation:

$$x_{n+1} = x_n - \alpha \nabla f(x_n) \tag{26}$$

This recurrent relation starts from a initial point x_0, and the next iteration x_1 moves in the negative direction of the gradient evaluated on x_0, which effectively moves in the direction of decreasing the objective function, but since the gradient only gives a direction, step size α must be assumed. Then the recurrent relation is iterated until a stopping criteria is met, such a certain number of iterations M, a minimum value for the objective function, or a computational budget.

If you evaluate the objective function on the sequence $x_0, x_1, ..., x_n$, then its value should be consistently decreasing.

A critical parameter is α, also called step size (mathematics) or learning rate (in machine learning), which determines how much to move along the gradient direction in every iteration. This parameter should be tuned to the specific geometry of the objective function, but common values to start are $0 < \alpha < 1$, like $\alpha \in [0.1, 0.01]$. To tune this parameter, use a value that makes the objective function to consistently decrease.

The difference between Gradient Descent, and Gradient Ascent, is the gradient step direction. To descent the negative gradient is used (as shown in Eq 26), and for ascent the positive gradient direction is used, for which the relation would be $x_{n+1} = x_n + \alpha \nabla f(x_n)$.

A major disadvantage of gradient descent is that there are no guarantees on reaching a global optima. When the gradient is zero, the recurrent relation basically stops and produces the same value, indicating that this method can get stuck on local minima.

Another disadvantage is that gradient descent requires a starting point x_0, which is usually randomly selected, meaning that multiple runs of the whole iterative process might end with different optima.

4.3 Convexity

Convexity is an important concept in optimization. As we saw in the previous section, gradient descent has problems getting stuck in local optima due to the gradient vanishing (value equal to zero), but then what are some nice-to-have properties in an objective function so its easy to optimize?

The mathematical definition of a convex function is, a function

$f : D \to \mathbf{R}$, where the following holds $\forall t \in (0,1)$ and $\forall x_1, x_2 \in D$:

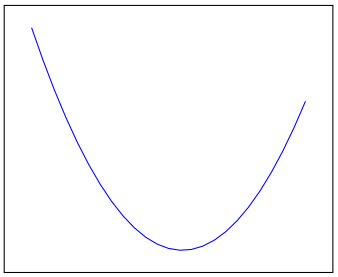
$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \tag{27}$$

Of course this is a fully mathematical definition, a simpler definition is a function f where the second derivative is always non-negative. Note that a function might also be convex on a subset $D' \subset D$ of its domain instead of its full domain D.

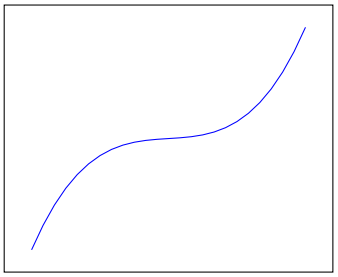
A convex objective function has many implications for optimization:

- Convex functions have a single global optima.
- Convex functions have no local optima.

These properties make convex functions ideal for optimization. Concave functions have very similar definitions that mirror the ones for convex functions, but in the opposite direction. Also if f is convex, the $-f$ is concave.



(a) Example of a convex function



(b) Example of a non-convex function

5 *Statistics and Probability*

Statistics is the science of working with data and probability is about modelling uncertainty. In statistical modelling, one often assumes that for a given study, there is a population that contains quantities of interest, but in practice one does not have access to the full population, only to a sample of that population (a subset), and the job of statistics is to answer questions of interest (inference) about the population, only from the samples.

5.1 *Summary Statistics*

Given N data observations $x_1, x_2, ..., x_N$, we usually want to build summaries, quantities that offer a simplified view of the data, and there are two main ones. First the mean which measures central tendency:

$$\bar{x} = N^{-1} \sum x_i \tag{28}$$

And the variance, which measures spread:

$$\text{var}(x) = N^{-1} \sum (x_i - \bar{x})^2 \tag{29}$$

And the standard deviation being the square root of the variance: $\text{std}(x) = \sqrt{\text{var}(x)}$. The standard deviation is in the same scale as the original data points, while the variance is in squared scale, so the standard deviation is often preferred.

5.2 *Probability*

Before defining probability, these are defined over events. An event is a set of outcomes from an experiment that involves randomness, that is, the experimental outcomes are not always the same.

Event	Probability
A	$\mathbb{P}(A) \in [0,1]$
not A	$\mathbb{P}(\text{not } A) = 1 - \mathbb{P}(A)$
A or B	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
	$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if A and B are mut. exclusive.
A and B	$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$
	$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ if A and B are independent.
A given B	$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$ (Bayes Rule)

Table 5: Some basic rules and identities for probabilities, given events A, B.

A probability is defined for an event E as the ratio of possible ways it can happen to the total number of outcomes.

$$\mathbb{P}(E) = \frac{\text{\# of ways } E \text{ can happen}}{\text{\# of total outcomes}} \tag{30}$$

Probabilities are real numbers in the $[0,1]$ range, and are interpreted as likelihood or chance that the event will happen, with a higher probability value indicating higher chance of an event happening.

Sum Rule

Events A and B are mutually exclusive if one happens, the other cannot happen. This is equivalent to:

$$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) \tag{31}$$

Product Rule

Events A and B are independent if one happening does not affect the other. This is equivalent to:

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A)\mathbb{P}(B) \tag{32}$$

Conditional Probability

A conditional probability is the probability of an event A given that another event B has already occurred, which takes the assumption that both events have some sort of relationship. This is denoted by $\mathbb{P}(A \mid B)$ and defined as:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{33}$$

Note that in general $\mathbb{P}(A) \neq \mathbb{P}(A \mid B) \neq \mathbb{P}(B \mid A) \neq \mathbb{P}(B)$. Only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ and/or $\mathbb{P}(B \mid A) = \mathbb{P}(B)$, this indicates that A and B are independent.

5.3 *Random Variables*

A random variable is a mathematical concept used to model randomness, where it is a object which depend on a random event, that consists of two main components:

Sample Space A random value has a set of allowable values, usually called sample space and denoted by $\mathcal{A}$.

Mapping A random variable is a mapping $X : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A}$ are possible outcomes in the sample space, mapped to real numbers (their probability).

The possible values of a random variable are events, which are subset of a *sample space*, which is any predefined set (of any kind, real numbers, matrices, vectors, etc).

Random variables can be discrete or continuous, which defines their sample space. Random variables are a generalization of probability, not in discrete events but on sample spaces, for example random variables on the real numbers or in subsets of $\mathbb{R}$.

5.4 Probability Distributions

A probability distribution is a function that maps from a random variable to probability (discrete) or probability density (continuous) values, as a way to characterize how probability is distributed on the variable.

Probability Function

A probability function is a function that usually defines a probability distribution, and is defined as:

$$f_X(x) = \mathbb{P}(X = x) \tag{34}$$

This definition works well for discrete random variables, but it is problematic for continuous ones, because $\mathbb{P}(X = x) = 0$ for all x. For discrete distributions, this is usually called the probability mass function.

Probability Density Function

For continuous probability distributions , we define the probability density function (PDF), which is a function that defines a continuous probability distribution, and is defined as $f_X(x)$ that follows:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx \tag{35}$$

For any $a \leq b$. PDFs model probability density instead of plain probability, so they can (and will be) bigger than one.

Note that if $a = b$, then:

$$\mathbb{P}(a \leq X \leq a) = \int_a^a f_X(x)dx = 0 \tag{36}$$

Kolmogorov Axioms

All probability distributions need to follow these axioms, called Kolmogorov's axioms:

Probability values are between 0 and 1

$0 \leq \mathbb{P}(X \in E) \leq 1 \qquad \forall E \in \mathcal{A}$

Sum of all events is 1

$\sum_{X \in \mathcal{A}} \mathbb{P}(X) = 1 \qquad \int_{\mathcal{A}} P(X)dX = 1$

Disjoint Family of Sets

$\mathbb{P}(X \in \cup_i E_i) = \sum_i \mathbb{P}(X \in E_i)$

For any disjoint family of sets $E_i \in \mathcal{A}$

Cumulative Distribution Function

The Cumulative Distribution Function (CDF) is a function defined by:

$$F_X(x) = \mathbb{P}(X \leq x) \tag{37}$$

This function gives the probability that a random variable X is less than a value x. A special property is:

$$\mathbb{P}(a < X \leq b) = F(b) - F(a) \tag{38}$$

For continuous random variables, there are some additional dualities that relate the PDF and CDF:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{39}$$

$$f_X(x) = \frac{d}{dx}F_X(x) \tag{40}$$

The CDF and PDF are related through the integral/derivative of each other.

Inverse CDF/Quantile Function

The Inverse CDF , also known as the Quantile function, is defined as:

$$F^{-1}(q) = \inf\{x : F(x) > q\} \tag{41}$$

Where $q \in [0, 1]$ is a specific quantile. The inf method can be conceptually thought as taking the minimum.
If $F(x)$ is a strictly increasing function, then $F^{-1}(q)$ is the unique value x so $F(x) = q$ holds.
Quantiles are equal divisions of the probability space, some have special names, for example percentiles divide in 100 divisions, or quartiles over 4 divisions. The percentile 50 is always the median ($F^{-1}(0.5)$).

Expectation

Expectation or expected value is a linear operation defined for continuous distributions with PDF $X \sim f_X$ as:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)dx \tag{42}$$

And for discrete distributions as:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} x_i\, f_X(x_i) \tag{43}$$

The expected value $\mathbb{E}[X]$ is associated with the mean, while the variance can be computed as $\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Covariance

Covariance is a measure of the joint variation between two related variables X and Y, and is defined as:

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{44}$$

Not to be confused with the covariance matrix. If covariance is normalized with the standard deviation of each variable (σ_X, σ_Y), you obtain the correlation:

$$\text{CORR}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sigma_X \sigma_Y} \tag{45}$$

Now that we have made most definitions about probability distributions, we can define the most important probability distributions. The most important distribution is the Gaussian distribution.

Gaussian Distribution

The Gaussian or Normal distribution a continuous distribution

defined for $x \in \mathbb{R}$ with two parameters, mean μ and variance σ^2, and with PDF given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5(\frac{x-\mu}{\sigma})^2} \tag{46}$$

The expectation and variance of the Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ is:

$$\mathbb{E}[X] = \mu$$
$$\text{Var}[X] = \sigma^2$$

Uniform Distribution

Another important distribution is the uniform distribution , where all values have equal probability. It is a continuous or Discrete distribution where all values in a range $[a, b]$ are equally likely. It is parameterized by a and b and has PDF/mass function defined as:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

It is usually denoted as $X \sim U(a, b)$. The mean is $\mathbb{E}[X] = 0.5(a + b)$ and variance is $\text{Var}[X] = \frac{1}{12}(b - a)^2$.

The importante of the Gaussian distribution is due to the central limit theorem.

Given a sequence of random variables $X_1, X_2, ..., X_n$ that are independent and identically distributed with population mean μ and variance σ^2. First let's define the sample mean or average:

$$\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} \tag{48}$$

Central Limit Theorem

The central limit theorem states that as $n \to \infty$, then the distribution of $\bar{X}_n$ is a Gaussian distribution:

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{49}$$

That is the variance of the sample mean decreases with $\frac{1}{n}$, while the standard deviation decreases with $\frac{1}{\sqrt{n}}$.

6 *Relationship with IML Course*

Minimum Knowledge. The minimum mathematical knowledge to successfully pass the course would be Linear Algebra and Calculus, in particular knowledge on matrices, hyperplanes, eigenvalues and eigenvectors, and gradients.

Recommended Knowledge. In addition to the minimum knowledge, it is recommended to be proficient with linear transformations, and basic statistics and probability, in particular probability, random variables, and probability distributions.

Advanced Knowledge. For a full theoretical understanding of the course, the whole content of this document is required, in particular optimization concepts, taylor's approximation, and vector calculus (Jacobian and Hessians).

Math Concept	Usage in Machine Learning
Linear Algebra	Model equations are built using vectors and matrices.
	Linear models are related to hyperplanes.
	Linear models relate to the concept of linear transformations.
	Eigendecomposition is related to PCA and normalization.
Calculus	Computation of gradients for use in gradient descent.
	Optimization intuitions related to first and second derivatives.
Optimization	Gradient-based optimization is the main training method of ML models.
	Basic optimization concepts also apply to training ML models.
Statistics	Statistics is working with data, concepts like losses and model-fitting come from Statistics.
	There is significant overlap between Statistics and Machine Learning concepts.
	Data is modeled as independent and identically distributed data points.
Probability	Classifiers output probabilities of correctness.
	Probabilistic modeling is used to model uncertainty.

Table 6: Summary of now Machine Learning and Mathematical concepts relate to each other.

A mapping between mathematical concepts and machine learning ones is shown in Table 6, from where a clear understanding of where each mathematical concept is used in machine learning contexts.

7 *Further Reading*

This document does not aim to be a comprehensive resource on mathematics for basic machine learning, but being more of a refresher or handout. For further details you can consult the following books:

Linear Algebra Klein PN. Coding the matrix: Linear algebra through computer science applications. Newtonian Press; 2013.

Calculus Simmons GF. Calculus with analytic geometry. McGraw Hill; 1996.

Optimization Kochenderfer MJ. Algorithms for Optimization. The MIT Press; 2019.

Statistics Spiegelhalter D. The Art of Statistics: Learning from Data. Penguin UK; 2019.

Reinhart A. Statistics done wrong: The woefully complete guide. No starch press; 2015.

Probability Bertsekas D, Tsitsiklis JN. Introduction to Probability. Athena Scientific; 2008.