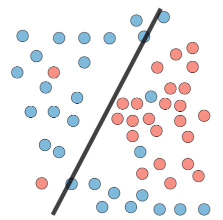
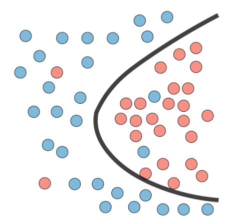
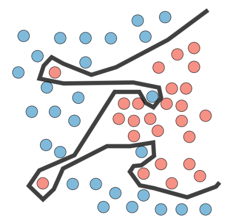
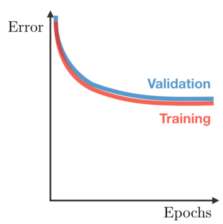
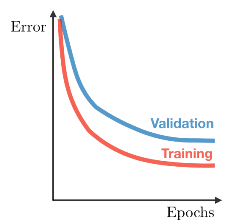
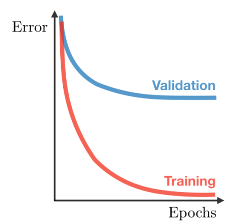
Classification			
Deep learning			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data

❒ **Error analysis** – Error analysis is analyzing the root cause of the difference in performance between the current and the perfect models.

❒ **Ablative analysis** – Ablative analysis is analyzing the root cause of the difference in performance between the current and the baseline models.

5 Refreshers

5.1 Probabilities and Statistics

5.1.1 Introduction to Probability and Combinatorics

❒ **Sample space** – The set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by S.

❒ **Event** – Any subset E of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in E, then we say that E has occurred.

❒ **Axioms of probability** – For each event E, we denote $P(E)$ as the probability of event E occuring. By noting $E_1,...,E_n$ mutually exclusive events, we have the 3 following axioms:

$$(1)\quad \boxed{0\leqslant P(E)\leqslant 1}\qquad (2)\quad \boxed{P(S)=1}\qquad (3)\quad \boxed{P\left(\bigcup_{i=1}^n E_i\right)=\sum_{i=1}^n P(E_i)}$$

❒ **Permutation** – A permutation is an arrangement of r objects from a pool of n objects, in a given order. The number of such arrangements is given by $P(n, r)$, defined as:

$$\boxed{P(n,r)=\frac{n!}{(n-r)!}}$$

❒ **Combination** – A combination is an arrangement of r objects from a pool of n objects, where the order does not matter. The number of such arrangements is given by $C(n, r)$, defined as:

$$\boxed{C(n,r)=\frac{P(n,r)}{r!}=\frac{n!}{r!(n-r)!}}$$

Remark: we note that for $0 \leqslant r \leqslant n$, we have $P(n,r) \geqslant C(n,r)$.

5.1.2 Conditional Probability

❒ **Bayes' rule** – For events A and B such that $P(B) > 0$, we have:

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

Remark: we have $P(A\cap B)=P(A)P(B|A)=P(A|B)P(B)$.

❒ **Partition** – Let $\{A_i, i\in[\![1,n]\!]\}$ be such that for all i, $A_i\neq\varnothing$. We say that $\{A_i\}$ is a partition if we have:

$$\boxed{\forall i\neq j, A_i\cap A_j=\emptyset\quad\textrm{and}\quad\bigcup_{i=1}^n A_i=S}$$

Remark: for any event B in the sample space, we have $P(B)=\displaystyle\sum_{i=1}^n P(B|A_i)P(A_i)$.

❒ **Extended form of Bayes' rule** – Let $\{A_i,i\in[\![1,n]\!]\}$ be a partition of the sample space. We have:

$$\boxed{P(A_k|B)=\frac{P(B|A_k)P(A_k)}{\displaystyle\sum_{i=1}^nP(B|A_i)P(A_i)}}$$

❒ **Independence** – Two events A and B are independent if and only if we have:

$$\boxed{P(A\cap B)=P(A)P(B)}$$

5.1.3 Random Variables

❒ **Random variable** – A random variable, often noted X, is a function that maps every element in a sample space to a real line.

❒ **Cumulative distribution function (CDF)** – The cumulative distribution function F, which is monotonically non-decreasing and is such that $\underset{x\rightarrow-\infty}{\lim}F(x)=0$ and $\underset{x\rightarrow+\infty}{\lim}F(x)=1$, is defined as:

$$\boxed{F(x)=P(X\leqslant x)}$$

Remark: we have $P(a<X\leqslant B)=F(b)-F(a)$.

❒ **Probability density function (PDF)** – The probability density function f is the probability that X takes on values between two adjacent realizations of the random variable.

❒ **Relationships involving the PDF and CDF** – Here are the important properties to know in the discrete (D) and the continuous (C) cases.

Case	**CDF F**	**PDF f**	**Properties of PDF**
(D)	$F(x)=\displaystyle\sum_{x_i\leqslant x}P(X=x_i)$	$f(x_j)=P(X=x_j)$	$0\leqslant f(x_j)\leqslant1$ and $\displaystyle\sum_jf(x_j)=1$
(C)	$F(x)=\displaystyle\int_{-\infty}^xf(y)dy$	$f(x)=\displaystyle\frac{dF}{dx}$	$f(x)\geqslant0$ and $\displaystyle\int_{-\infty}^{+\infty}f(x)dx=1$

❒ **Variance** – The variance of a random variable, often noted Var(X) or σ^2, is a measure of the spread of its distribution function. It is determined as follows:

$$\boxed{\textrm{Var}(X)=E[(X-E[X])^2]=E[X^2]-E[X]^2}$$

❒ **Standard deviation** – The standard deviation of a random variable, often noted σ, is a measure of the spread of its distribution function which is compatible with the units of the actual random variable. It is determined as follows:

$$\boxed{\sigma=\sqrt{\textrm{Var}(X)}}$$

❒ **Expectation and Moments of the Distribution** – Here are the expressions of the expected value $E[X]$, generalized expected value $E[g(X)]$, k^{th} moment $E[X^k]$ and characteristic function $\psi(\omega)$ for the discrete and continuous cases:

Case	$E[X]$	$E[g(X)]$	$E[X^k]$	$\psi(\omega)$
(D)	$\displaystyle\sum_{i=1}^nx_if(x_i)$	$\displaystyle\sum_{i=1}^ng(x_i)f(x_i)$	$\displaystyle\sum_{i=1}^nx_i^kf(x_i)$	$\displaystyle\sum_{i=1}^nf(x_i)e^{i\omega x_i}$
(C)	$\displaystyle\int_{-\infty}^{+\infty}xf(x)dx$	$\displaystyle\int_{-\infty}^{+\infty}g(x)f(x)dx$	$\displaystyle\int_{-\infty}^{+\infty}x^kf(x)dx$	$\displaystyle\int_{-\infty}^{+\infty}f(x)e^{i\omega x}dx$

Remark: we have $e^{i\omega x}=\cos(\omega x)+i\sin(\omega x)$.

❒ **Revisiting the k^{th} moment** – The k^{th} moment can also be computed with the characteristic function as follows:

$$\boxed{E[X^k]=\frac{1}{i^k}\left[\frac{\partial^k\psi}{\partial\omega^k}\right]_{\omega=0}}$$

❒ **Transformation of random variables** – Let the variables X and Y be linked by some function. By noting f_X and f_Y the distribution function of X and Y respectively, we have:

$$\boxed{f_Y(y)=f_X(x)\left|\frac{dx}{dy}\right|}$$

❒ **Leibniz integral rule** – Let g be a function of x and potentially c, and a,b boundaries that may depend on c. We have:

$$\boxed{\frac{\partial}{\partial c}\left(\int_a^bg(x)dx\right)=\frac{\partial b}{\partial c}\cdot g(b)-\frac{\partial a}{\partial c}\cdot g(a)+\int_a^b\frac{\partial g}{\partial c}(x)dx}$$

❒ **Chebyshev's inequality** – Let X be a random variable with expected value μ and standard deviation σ. For $k,\sigma>0$, we have the following inequality:

$$\boxed{P(|X-\mu|\geqslant k\sigma)\leqslant\frac{1}{k^2}}$$

5.1.4 Jointly Distributed Random Variables

❒ **Conditional density** – The conditional density of X with respect to Y, often noted $f_{X|Y}$, is defined as follows:

$$\boxed{f_{X|Y}(x)=\frac{f_{XY}(x,y)}{f_Y(y)}}$$

❒ **Independence** – Two random variables X and Y are said to be independent if we have:

$$\boxed{f_{XY}(x,y)=f_X(x)f_Y(y)}$$

❒ **Marginal density and cumulative distribution** – From the joint density probability function f_{XY}, we have:

Case	Marginal density	Cumulative function
(D)	$f_X(x_i)=\displaystyle\sum_j f_{XY}(x_i,y_j)$	$F_{XY}(x,y)=\displaystyle\sum_{x_i\leqslant x}\sum_{y_j\leqslant y}f_{XY}(x_i,y_j)$
(C)	$f_X(x)=\displaystyle\int_{-\infty}^{+\infty}f_{XY}(x,y)dy$	$F_{XY}(x,y)=\displaystyle\int_{-\infty}^{x}\int_{-\infty}^{y}f_{XY}(x',y')dx'dy'$

❒ **Distribution of a sum of independent random variables** – Let $Y=X_1+...+X_n$ with $X_1,...,X_n$ independent. We have:

$$\boxed{\psi_Y(\omega)=\prod_{k=1}^n\psi_{X_k}(\omega)}$$

❒ **Covariance** – We define the covariance of two random variables X and Y, that we note σ_{XY}^2 or more commonly $\mathrm{Cov}(X,Y)$, as follows:

$$\boxed{\mathrm{Cov}(X,Y)\triangleq\sigma_{XY}^2=E[(X-\mu_X)(Y-\mu_Y)]=E[XY]-\mu_X\mu_Y}$$

❒ **Correlation** – By noting σ_X,σ_Y the standard deviations of X and Y, we define the correlation between the random variables X and Y, noted ρ_{XY}, as follows:

$$\boxed{\rho_{XY}=\frac{\sigma_{XY}^2}{\sigma_X\sigma_Y}}$$

Remarks: For any X,Y, we have $\rho_{XY}\in[-1,1]$. If X and Y are independent, then $\rho_{XY}=0$.

❒ **Main distributions** – Here are the main distributions to have in mind:

Type	Distribution	PDF	$\psi(\omega)$	$E[X]$	$\mathrm{Var}(X)$
(D)	$X\sim\mathcal{B}(n,p)$ Binomial	$P(X=x)=\binom{n}{x}p^xq^{n-x}$ $x\in[\![0,n]\!]$	$(pe^{i\omega}+q)^n$	np	npq
	$X\sim\mathrm{Po}(\mu)$ Poisson	$P(X=x)=\frac{\mu^x}{x!}e^{-\mu}$ $x\in\mathbb{N}$	$e^{\mu(e^{i\omega}-1)}$	μ	μ
(C)	$X\sim\mathcal{U}(a,b)$ Uniform	$f(x)=\frac{1}{b-a}$ $x\in[a,b]$	$\frac{e^{i\omega b}-e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$X\sim\mathcal{N}(\mu,\sigma)$ Gaussian	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $x\in\mathbb{R}$	$e^{i\omega\mu-\frac{1}{2}\omega^2\sigma^2}$	μ	σ^2
	$X\sim\mathrm{Exp}(\lambda)$ Exponential	$f(x)=\lambda e^{-\lambda x}$ $x\in\mathbb{R}_+$	$\frac{1}{1-\frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

5.1.5 Parameter estimation

❒ **Random sample** – A random sample is a collection of n random variables $X_1,...,X_n$ that are independent and identically distributed with X.

❒ **Estimator** – An estimator $\hat{\theta}$ is a function of the data that is used to infer the value of an unknown parameter θ in a statistical model.

❒ **Bias** – The bias of an estimator $\hat{\theta}$ is defined as being the difference between the expected value of the distribution of $\hat{\theta}$ and the true value, i.e.:

$$\boxed{\mathrm{Bias}(\hat{\theta})=E[\hat{\theta}]-\theta}$$

Remark: an estimator is said to be unbiased when we have $E[\hat{\theta}]=\theta$.

❒ **Sample mean and variance** – The sample mean and the sample variance of a random sample are used to estimate the true mean μ and the true variance σ^2 of a distribution, are noted $\overline{X}$ and s^2 respectively, and are such that:

$$\boxed{\overline{X}=\frac{1}{n}\sum_{i=1}^nX_i}\quad\textrm{and}\quad\boxed{s^2=\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\overline{X})^2}$$

❒ **Central Limit Theorem** – Let us have a random sample $X_1,...,X_n$ following a given distribution with mean μ and variance σ^2, then we have:

$$\boxed{\overline{X}\underset{n\rightarrow+\infty}{\sim}\mathcal{N}\left(\mu,\frac{\sigma}{\sqrt{n}}\right)}$$

5.2 Linear Algebra and Calculus

5.2.1 General notations

❒ **Vector** – We note $x\in\mathbb{R}^n$ a vector with n entries, where $x_i\in\mathbb{R}$ is the i^{th} entry:

$$x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}\in\mathbb{R}^n$$

❒ **Matrix** – We note $A\in\mathbb{R}^{m\times n}$ a matrix with m rows and n columns, where $A_{i,j}\in\mathbb{R}$ is the entry located in the i^{th} row and j^{th} column:

$$A=\begin{pmatrix}A_{1,1}&\cdots&A_{1,n}\\\vdots&&\vdots\\A_{m,1}&\cdots&A_{m,n}\end{pmatrix}\in\mathbb{R}^{m\times n}$$

Remark: the vector x defined above can be viewed as a $n\times1$ matrix and is more particularly called a column-vector.

❒ **Identity matrix** – The identity matrix $I\in\mathbb{R}^{n\times n}$ is a square matrix with ones in its diagonal and zero everywhere else:

$$I=\begin{pmatrix}1&0&\cdots&0\\0&\ddots&\ddots&\vdots\\\vdots&\ddots&\ddots&0\\0&\cdots&0&1\end{pmatrix}$$

Remark: for all matrices $A\in\mathbb{R}^{n\times n}$, we have $A\times I=I\times A=A$.

❒ **Diagonal matrix** – A diagonal matrix $D\in\mathbb{R}^{n\times n}$ is a square matrix with nonzero values in its diagonal and zero everywhere else:

$$D=\left(\begin{array}{cccc}d_1 & 0 & \cdots & 0\\ 0 & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & d_n\end{array}\right)$$

Remark: we also note D as $diag(d_1,...,d_n)$.

5.2.2 Matrix operations

❒ **Vector-vector multiplication** – There are two types of vector-vector products:

- inner product: for $x,y\in\mathbb{R}^n$, we have:

$$\boxed{x^Ty=\sum_{i=1}^n x_iy_i\in\mathbb{R}}$$

- outer product: for $x\in\mathbb{R}^m, y\in\mathbb{R}^n$, we have:

$$\boxed{xy^T=\left(\begin{array}{ccc}x_1y_1 & \cdots & x_1y_n\\ \vdots & & \vdots\\ x_my_1 & \cdots & x_my_n\end{array}\right)\in\mathbb{R}^{m\times n}}$$

❒ **Matrix-vector multiplication** – The product of matrix $A\in\mathbb{R}^{m\times n}$ and vector $x\in\mathbb{R}^n$ is a vector of size $\mathbb{R}^m$, such that:

$$\boxed{Ax=\left(\begin{array}{c}a_{r,1}^Tx\\ \vdots\\ a_{r,m}^Tx\end{array}\right)=\sum_{i=1}^n a_{c,i}x_i\in\mathbb{R}^m}$$

where $a_{r,i}^T$ are the vector rows and $a_{c,j}$ are the vector columns of A, and x_i are the entries of x.

❒ **Matrix-matrix multiplication** – The product of matrices $A\in\mathbb{R}^{m\times n}$ and $B\in\mathbb{R}^{n\times p}$ is a matrix of size $\mathbb{R}^{n\times p}$, such that:

$$\boxed{AB=\left(\begin{array}{ccc}a_{r,1}^Tb_{c,1} & \cdots & a_{r,1}^Tb_{c,p}\\ \vdots & & \vdots\\ a_{r,m}^Tb_{c,1} & \cdots & a_{r,m}^Tb_{c,p}\end{array}\right)=\sum_{i=1}^n a_{c,i}b_{r,i}^T\in\mathbb{R}^{n\times p}}$$

where $a_{r,i}^T, b_{r,i}^T$ are the vector rows and $a_{c,j}, b_{c,j}$ are the vector columns of A and B respectively.

❒ **Transpose** – The transpose of a matrix $A\in\mathbb{R}^{m\times n}$, noted A^T, is such that its entries are flipped:

$$\boxed{\forall i,j,\qquad A_{i,j}^T=A_{j,i}}$$

Remark: for matrices A,B, we have $(AB)^T=B^TA^T$.

❒ **Inverse** – The inverse of an invertible square matrix A is noted A^{-1} and is the only matrix such that:

$$\boxed{AA^{-1}=A^{-1}A=I}$$

Remark: not all square matrices are invertible. Also, for matrices A,B, we have $(AB)^{-1}=B^{-1}A^{-1}$

❒ **Trace** – The trace of a square matrix A, noted $\textrm{tr}(A)$, is the sum of its diagonal entries:

$$\boxed{\textrm{tr}(A)=\sum_{i=1}^n A_{i,i}}$$

Remark: for matrices A,B, we have $tr(A^T)=tr(A)$ and $tr(AB)=tr(BA)$

❒ **Determinant** – The determinant of a square matrix $A\in\mathbb{R}^{n\times n}$, noted $|A|$ or $\det(A)$ is expressed recursively in terms of $A_{\backslash i,\backslash j}$, which is the matrix A without its i^{th} row and j^{th} column, as follows:

$$\boxed{\det(A)=|A|=\sum_{j=1}^n(-1)^{i+j}A_{i,j}|A_{\backslash i,\backslash j}|}$$

Remark: A is invertible if and only if $|A|\neq 0$. Also, $|AB|=|A||B|$ and $|A^T|=|A|$.

5.2.3 Matrix properties

❒ **Symmetric decomposition** – A given matrix A can be expressed in terms of its symmetric and antisymmetric parts as follows:

$$\boxed{A=\underbrace{\frac{A+A^T}{2}}_{\textrm{Symmetric}}+\underbrace{\frac{A-A^T}{2}}_{\textrm{Antisymmetric}}}$$

❒ **Norm** – A norm is a function $N:V\longrightarrow[0,+\infty[$ where V is a vector space, and such that for all $x,y\in V$, we have:

- $N(x+y)\leqslant N(x)+N(y)$
- $N(ax)=|a|N(x)$ for a scalar
- if $N(x)=0$, then $x=0$

For $x\in V$, the most commonly used norms are summed up in the table below:

Norm	Notation	Definition	Use case
Manhattan, L^1	$\|\|x\|\|_1$	$\displaystyle\sum_{i=1}^n \|x_i\|$	LASSO regularization
Euclidean, L^2	$\|\|x\|\|_2$	$\displaystyle\sqrt{\sum_{i=1}^n x_i^2}$	Ridge regularization
p-norm, L^p	$\|\|x\|\|_p$	$\displaystyle\left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}}$	Hölder inequality
Infinity, L^∞	$\|\|x\|\|_\infty$	$\underset{i}{\max}\ \|x_i\|$	Uniform convergence

❒ **Linearly dependence** – A set of vectors is said to be linearly dependent if one of the vectors in the set can be defined as a linear combination of the others.
Remark: if no vector can be written this way, then the vectors are said to be linearly independent.

❒ **Matrix rank** – The rank of a given matrix A is noted $\textrm{rank}(A)$ and is the dimension of the vector space generated by its columns. This is equivalent to the maximum number of linearly independent columns of A.

❒ **Positive semi-definite matrix** – A matrix $A\in\mathbb{R}^{n\times n}$ is positive semi-definite (PSD) and is noted $A\succeq 0$ if we have:

$$\boxed{A=A^T}\quad\textrm{and}\quad\boxed{\forall x\in\mathbb{R}^n,\quad x^TAx\geqslant0}$$

Remark: similarly, a matrix A is said to be positive definite, and is noted $A\succ0$, if it is a PSD matrix which satisfies for all non-zero vector x, $x^TAx>0$.

❒ **Eigenvalue, eigenvector** – Given a matrix $A\in\mathbb{R}^{n\times n}$, λ is said to be an eigenvalue of A if there exists a vector $z\in\mathbb{R}^n\backslash\{0\}$, called eigenvector, such that we have:

$$\boxed{Az=\lambda z}$$

❒ **Spectral theorem** – Let $A\in\mathbb{R}^{n\times n}$. If A is symmetric, then A is diagonalizable by a real orthogonal matrix $U\in\mathbb{R}^{n\times n}$. By noting $\Lambda=\textrm{diag}(\lambda_1,...,\lambda_n)$, we have:

$$\boxed{\exists\Lambda\textrm{ diagonal},\quad A=U\Lambda U^T}$$

❒ **Singular-value decomposition** – For a given matrix A of dimensions $m\times n$, the singular-value decomposition (SVD) is a factorization technique that guarantees the existence of U $m\times m$ unitary, Σ $m\times n$ diagonal and V $n\times n$ unitary matrices, such that:

$$\boxed{A=U\Sigma V^T}$$

5.2.4 Matrix calculus

❒ **Gradient** – Let $f:\mathbb{R}^{m\times n}\rightarrow\mathbb{R}$ be a function and $A\in\mathbb{R}^{m\times n}$ be a matrix. The gradient of f with respect to A is a $m\times n$ matrix, noted $\nabla_A f(A)$, such that:

$$\boxed{\Big(\nabla_A f(A)\Big)_{i,j}=\frac{\partial f(A)}{\partial A_{i,j}}}$$

Remark: the gradient of f is only defined when f is a function that returns a scalar.

❒ **Hessian** – Let $f:\mathbb{R}^n\rightarrow\mathbb{R}$ be a function and $x\in\mathbb{R}^n$ be a vector. The hessian of f with respect to x is a $n\times n$ symmetric matrix, noted $\nabla_x^2f(x)$, such that:

$$\boxed{\Big(\nabla_x^2f(x)\Big)_{i,j}=\frac{\partial^2f(x)}{\partial x_i\partial x_j}}$$

Remark: the hessian of f is only defined when f is a function that returns a scalar.

❒ **Gradient operations** – For matrices A,B,C, the following gradient properties are worth having in mind:

$$\boxed{\nabla_A\textrm{tr}(AB)=B^T}\quad\boxed{\nabla_{A^T}f(A)=\left(\nabla_Af(A)\right)^T}$$

$$\boxed{\nabla_A\textrm{tr}(ABA^TC)=CAB+C^TAB^T}\quad\boxed{\nabla_A|A|=|A|(A^{-1})^T}$$