basics blog

basics

Jun 30, 2025

CONTENTS

This material is part of the **basics-books project**.

Contents.

Introduction to statistics

Different approaches to statistics and **descriptive statistics**

Probability theory

Inferential statistics

Inferential and Bayesian statistics

Introduction to Machine Learning: SL, UL, ML

Machine learning (ML) is a branch of artificial intelligence (AI) focused on designing systems that can learn from data to improve their performance on a task. ML frameworks include supervised learning (e.g., regression and classification), unsupervised learning (e.g., clustering, compression, principal component analysis), and reinforcement learning (e.g., planning and control). ML emphasizes practical problem-solving, grounded in statistical methods, numerical optimization, and enabled by advances in computing hardware.

Part I

Introduction to Statistics

CHAPTER

ONE

INTRODUCTION TO STATISTICS

CHAPTER

TWO

DESCRIPTIVE STATISTICS

Part II

Probability Theory

CHAPTER

THREE

INTRODUCTION TO PROBABILITY THEORY

Probability theory is an axiomatic approach to probability, assigning

Stochastic variables

- Definition of stochastic variable
- Discrete and continuous stochastic variables
  - Probability functions, moments (if they exists, see heavy-tailed distribution), and examples
- Multi-dimensional stochastic variables:
  - joint, conditional, marginal probability
  - Bayes' theorem
  - independence
  - moments: covariance, correlation
- Generators…
- I.i.d. variables: law of large numbers, central limit theorem; convergence of statistics (reference to measure in the definition of a sthocastic variable)
- Sampling
- Extra:
  - heavy tails probability functions

Stochastic processes

- Definition of stochastic process
- Time-continuous/time-discrete
- Ergodicity and stationariety:
  - moments, correlation,…
  - analysis in time and Fourier domains of time-signals
- Applications:
  - example of processes:
    - white noise

    * Wiener process (Brownian motion): definition, application, relation with
    * discrete-time Markov process (useful in RL, can be interpreted as a discretized continuous process)
  - response of LTI to random input

Stochastic fields

3.1 Definition of stochastic variable

In this section, first a formal definition of a stochastic (or random) variable is provided and then discussed. While the definition may initially appear abstract or esoteric, but their nothing more than an extension of set theory. Understanding concepts such as *probability space*, *σ-algebra*, *measure*, opens the path to a rigorous approach to probability theory.

Definition 3.1.1 (Random variable)

Given a **probability space** $(\Omega, \mathcal{F}, \nu)$ and a **measurable space** $(E, \mathcal{E})$, a **random variable** is a measurable function, $X : \Omega \ \rightarrow \ E$.

Here, the set Ω is usually defined as the **event set**, $\mathcal{F}$ is a **σ-algebra** on Ω, ν is a **probability measure**; the set E is usually defined as the **set of possible outcomes**, and $\mathcal{E}$ is a **σ-algebra** on it.

Ok, let's explain now every concept appearing in the definition of a random variable.

Definition 3.1.2 (σ-algebra)

A σ-algebra $\mathcal{F}$ on a set Ω, a σ-algebra is a family of subsets of Ω that satisfies some properties:

1. Ω is in the σ-algebra
2. closure under complementation. Taking any subset of Ω that belongs to the σ-algebra, its complent belongs to the σ-algebra as well
3. closure under countable union. A countable (that can be indexed by integer numbers) union of elements in the σ-algebra defines an element that belongs to the σ-algebra

---

Definition 3.1.3 (Measurable space $(\Omega, \mathcal{F})$)

Definition 3.1.4 (Probability measure ν)

Definition 3.1.5 (Probability space $(\Omega, \mathcal{F}, \nu)$)

If $(\Omega, \mathcal{F})$ is a measureble space and ν is a probability measure, then the triplet $(\Omega, \mathcal{F}, \nu)$ is a probability space.

Warning: Event space not coinciding with the power set of events

If event space Ω doesn't coincide with the power set but it's a subset of it, then it's not guaranteed that such an event space can be used to define a σ-algebra.

3.2 Discrete stochastic variables

3.3 Continuous stochastic variables

3.3.1 Examples

Here some common examples of continuous random variables are introduced. Their functional dependence on the value of the r.v. is quite easy to remember, while the normalization factor could look quite "esoteric".

Normal distribution, $\mathcal{N}(\mu, \sigma^2)$

pdf is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \propto \qquad e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Moment	Value
Expected value	μ
Variance	σ^2

Expected value, μ; variance, σ^2.

Unitariety

$$\int_{x=-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \sqrt{2\pi\sigma^2}$$

todo *integral* $\int_{-\infty}^{+\infty} e^{-\alpha x^2} dx$

Expected value

$$\mathbb{E}[X] = \int_{x=-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Variance

$$\mathbb{E}\left[(X-\mu)^2\right] = \dots$$

Chi-square, χ^2_N

$$\chi^2_N := \sum_{n=1}^{N} X_n^2$$

pdf is

$$f(x; n) = \dots \qquad \propto \qquad x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

Student-t distribution, t_ν

$$t_\nu = \frac{Z}{\sqrt{\frac{K}{\nu}}} \ ,$$

with $Z \sim \mathcal{N}(0,1)$, and $K \sim \chi^2_\nu$.

pdf is

$$f(x;n) = ... \quad \propto \quad \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

3.4 Multi-dimensional stochastic variables

- joint distribution

$$p_{XY}(x,y)$$

- marginal distribution. For continuous variables

$$p_X(x) := \int_y p_{XY}(x,y)\,dy$$

  while for discrete variables

$$p_X(x_i) = \sum_j p_{XY}(x_i, y_j)$$

- conditional distribution, $p_{X|Y}(x|y)$. The following holds

$$p_{XY} = p_{X|Y}\,p_Y = p_{Y|X} p_X$$

For continuous r.v., integrating over x the relation $p(x,y) = p(x|y)p(y)$

$$\int_x p(x,y) dx = \int_x p(x|y)\,p(y)\,dx = p(y) \underbrace{\int_x p(x|y)\,dx}_{=1} = p(y) \ ,$$

as the normalization condition holds for conditional distribution $p(x|y)$.

Property 3.4.1

$$p(i,j) = p(i|j) p(j)$$
$$\sum_i p(i,j) = \underbrace{\sum_i p(i|j)}_{=1} p(j) = p(j)$$

3.4.1 Moments

- expected value

$$\mu_{\mathbf{X}} := \mathbb{E}\left[\mathbf{X}\right] = \int_{\mathbf{x}} p(\mathbf{x})\,\mathbf{x}\,d\mathbf{x}$$

- covariance

$$\sigma^2_{\mathbf{X}} := \mathbb{E}\left[\Delta\mathbf{X}\,\Delta\mathbf{X}^T\right] = \int_{\mathbf{x}} p(\mathbf{x})\,\Delta\mathbf{x}\Delta\mathbf{x}^T\,d\mathbf{x}\ ,$$

with $\Delta\mathbf{X} := \mathbf{X} - \mu_{\mathbf{X}}$, and $\Delta\mathbf{x} = \mathbf{x} - \mu_{\mathbf{X}}$.

Taking a pair of components X_i, X_j of the random vector $\mathbf{X}$, their covariance is the ij component of the array σ^2,

$$\sigma^2_{ij} := \mathbb{E}\left[\Delta X_i\,\Delta X_j\right] =: \rho_{ij}\sigma_i\sigma_j\ ,$$

having introduced **(Pearson) correlation**, ρ_{ij}, between random variable X_i and X_j, and being σ_i the standard deviation of variable X_i, square root of its variance σ^2_i,

$$\begin{aligned}\sigma^2_i &= \mathbb{E}\left[(X_i - \mu_i)^2\right] = \\ &= \int_{\mathbf{x}} (x_i - \mu_i)^2 p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \\ &= \int_{x_i} (x_i - \mu_i)^2 p_i(x_i)\,dx_i\end{aligned}$$

Here the integrals read

$$\begin{aligned}\mu_i &= \int_{\mathbf{x}} x_i\,p_{\mathbf{X}}(\mathbf{x}) x_i\,d\mathbf{x} = \\ &= \int_{\mathbf{x}} x_i\,p(x_1, x_2, \dots, x_i, \dots, x_n) dx_1 dx_2 \dots dx_i \dots dx_n = \\ &= \int_{\mathbf{x}} x_i\,p(x_i) p(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n | x_i) dx_1 dx_2 \dots dx_i \dots dx_n = \\ &= \int_{x_i} x_i\,p(x_i) \underbrace{\int_{x_1} \dots \int_{x_n} p(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n | x_i) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n}_{=1\ \forall x_i}\,dx_i = \\ &= \int_{x_i} x_i\,p(x_i)\,dx_i\ .\end{aligned}$$

Property of correlation. $|\rho_{XY}| \le 1$. Proof with Cauchy-Schwartz inequality **todo**

Notation

Here, covariance is indicated as σ^2. This is not a power 2, but just a symbol, at most recalling that covariance matrix is **semi-definite positive**.

Properties of covariance.

- symmetric
- semi-definite positive
- spectrum…

3.4.2 Bayes' theorem

Theorem 3.4.1 (Bayes' theorem)

Where $p_Y(y) \neq 0$,

$$p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

3.4.3 Statistical independence

Definition 3.4.1 (Independent random variables)

Given two random variables X, Y with joint distribution, the random variable X is independent from Y if its conditional probability equals its marginal probability,

$$p_{X|Y} = p_X \ ,$$

i.e. the probability of X doesn't depend on Y.

Independence implies no correlation

Given two random variables X, Y are independent if $p(x|y) = p(x)$ and thus $p(x,y) = p(x)p(y)$. Covariance of two random variable reads

$$\sigma^2_{xy} = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \ ,$$

and if they're independent, it immediately follows that their covariance σ^2_{XY} is zero (and so their correlation ρ_{XY})

$$\sigma^2_{xy} = \underbrace{\mathbb{E}\left[X-\mu_X\right]}_{=0} \underbrace{\mathbb{E}\left[Y-\mu_Y\right]}_{=0} = 0 \ ,$$

as the expected value of the deviation from the expected value is zero, $\mathbb{E}\left[X - \mathbb{E}[X]\right] = 0$.

Proof for continuous r.v.

$$\begin{aligned}
\sigma^2_{xy} &= \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \\
&= \int_{x,y} (x-\mu_X)(y-\mu_Y)p(x,y)\, dxdy = && (1)\\
&= \int_{x,y} (x-\mu_X)(y-\mu_Y)p(x)p(y)\, dxdy = \\
&= \int_x (x-\mu_X)p(x)dx \int_y (y-\mu_Y)p(y)dy = && (2)
\end{aligned}$$

having used here the common notation abuse $p_X(x) = p(x)$ and (1) statistical independence, $p(x,y) = p(x)p(y)$, and (2) $\mathbb{E}\left[X - \mathbb{E}[X]\right] = 0$.

Proof for discrete r.v.

Repeat the proof for continuous r.v. using summations instead of integrals.

3.5 Transformations of probability functions

3.6 Characteristic functions

Characteristic function of a random variable X is defined as

$$\varphi_X(t) := \mathbb{E}\left[e^{itX}\right] \ .$$

Characteristic function of a continuous random variable with proability density function $f(x)$ thus reads

$$\varphi_X(t) := \mathbb{E}\left[e^{itX}\right] = \int_{x \in D_x} f(x) e^{itx} \, dx \ ,$$

i.e. its the **Fourier transform** of its pdf.

Example 3.6.1 (Characteristic function of a multi-dimensional variable)

$$Z(\mathbf{Y})$$

$$\varphi_{Z(\mathbf{Y})} := \mathbb{E}\left[e^{itZ(\mathbf{Y})}\right] = \int_{\mathbf{y}} e^{itZ(\mathbf{y})} f(\mathbf{y}) \, d\mathbf{y}$$

Example 3.6.2 (Characteristic function of a linear combination of independent variables)

$$Z(\mathbf{Y}) = a_1 Y_1 + \dots a_n Y_n \ ,$$

with

$$f(\mathbf{y}) = f(y_1, \dots, y_n) = f_1(y_1) \dots f_n(y_n) \ .$$

$$\begin{aligned}
\varphi_{Z(\mathbf{Y})} & := \mathbb{E}\left[e^{itZ(\mathbf{Y})}\right] = \\
& = \int_{\mathbf{y}} e^{it\left(\sum_k a_k y_k\right)} f(\mathbf{y}) \, d\mathbf{y} = \\
& = \int_{y_1} e^{ita_1 y_1} f_1(y_1) \, dy_1 \dots \int_{y_n} e^{ita_n y_n} f_n(y_n) \, dy_n = \\
& = \varphi_{Y_1}(a_1 t) \dots \varphi_{Y_n}(a_n t) \ .
\end{aligned}$$

Example 3.6.3 (Taylor expansion of characteristic function)

For "small" values of t, an approximation of the characteristic function is provided by Taylor expansion around $t = 0$,

$$\begin{aligned}\int e^{iyt} f(y)\,dy &= \int \left[1 + iyt - \frac{1}{2}(yt)^2 + o(t^2)\right] f(y) dy = && (1)\\ &= 1 + i\mu t - \frac{1}{2} t^2 \left(\sigma^2 + \mu^2\right) + o(t^2)\end{aligned}$$

as (1) $\sigma^2 = \mathbb{E}[(y-\mu)^2] = \mathbb{E}[y^2] - \mu^2$

Example 3.6.4 (Characteristic function of a normal distribution $\mathcal{N}(0,1)$)

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

$$\begin{aligned}\int_{x=-\infty}^{+\infty} e^{ixt} f(x)\,dx &= \frac{1}{\sqrt{2\pi}} \int_{x=-\infty}^{+\infty} e^{ixt - \frac{x^2}{2}}\,dx = && (1)\\ &= \frac{1}{\sqrt{2\pi}} \int_{x=-\infty}^{+\infty} e^{-\frac{(x-it)^2}{2}}\,dx\, e^{-\frac{t^2}{2}} = && (2)\\ &= \frac{1}{\sqrt{2\pi}} \sqrt{2\pi}\, e^{-\frac{t^2}{2}} = e^{-\frac{t^2}{2}}\end{aligned}$$

having (1) completed the square $(x - it)^2 = x^2 - i2xt - t^2$, and evaluated the integral **todo** (it's similar to the standard result $\int_{-\infty}^{+\infty} e^{x^2}\,dx = \sqrt{2\pi}$, but with complex variable. Link to math material, complex calculus).

3.7 Convergence in statistics

3.7.1 Convergence in distribution - weak convergence

A sequence of X_i of real-valued random variables, cumulative distribution functions F_i, converges in distribution to a random variable X with cumulative distribution F is

$$\lim_{n \rightarrow +\infty} F_n(x) = F(x) \ ,$$

for $\forall x \in \mathbb{R}$ where $F(x)$ is continuous.

For multi-valued random variables, the condition reads

$$\lim_{n \rightarrow +\infty} P(X_n \in A) = P(X \in A) \ ,$$

for every $A \subset \mathbb{R}^n$ …**todo**

3.7.2 Convergence in probability

$$\lim_{n \rightarrow +\infty} P\left(|X_n - X| > \varepsilon\right) = 0$$

Warning: Convergence in probability and convergence in distribution

Convergence in probability $\rightarrow$ convergence in distribution, but not viceversa.

Example taken from wikipedia

3.7.3 Almost sure convergence - strong convergence

$$P\left(\lim_{n\rightarrow+\infty} X_n = X\right) = 1$$

i.e. events for which X_n doesn't converge to X has probability 0,

$$P\left(\omega \in \Omega : \lim_{n\rightarrow+\infty} X_n(\omega) = X(\omega)\right) = 1$$

3.7.4 Sure convergence - pointwise convergence

$$\left\{\omega \in \Omega : \lim_{n\rightarrow+\infty} X_n(\omega) = X(\omega)\right\} = \Omega \ .$$

The same definition of almost sure convergence, without allowing the existance of sets with zero probability where convergence is not satisfied. Thus, it's likely there is no point in using sure converence instead of almost sure convergence in proability theory.

3.7.5 Convergence in absolute moments: mean,...

$$\lim_{n\rightarrow+\infty} \mathbb{E}\left(|X_n - X|^r\right) = 0$$

3.8 Independent identically distributed random variables

Definition 3.8.1 (Independent identically distributed (iid) random variables)

3.8.1 Law of the large numbers

Weak form

todo

Strong form

todo

3.8.2 Central Limit Theorem

Theorem 3.8.1 (CLT)

Let $\{X_k\}_{k=1:n}$ a sequence of iid random variables with average value $\mathbb{E}[X_k] = \mu$ and **finite**[1] variance $\mathbb{E}[(X_k - \mu)^2] = \sigma^2 < \infty$, then the **sample average**

$$\overline{X}_n := \frac{1}{n}\sum_{k=1}^n X_k \ ,$$

converges in distribution - or weakly converges - to the normal distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$,

$$\overline{X}_n \quad \to^d \quad \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \ .$$

Proof of CLT

Let $\{X_k\}_{k=1:n}$ the sequence of iid random variables. Thus, $\sum_{k=1}^n X_k$ has expected value $n\mu$ and variance $n\sigma^2$. Let

$$Z_n := \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n\sigma^2}} = \sum_{k=1}^n \frac{X_k - \mu}{\sqrt{n\sigma^2}} =: \sum_{k=1}^n \frac{Y_k}{\sqrt{n}} \ .$$

Expeceted value and variance of variables Y_k are respectively $\mathbb{E}[Y_k] = 0$ and $\mathbb{E}[Y_k^2] = 1$. The *characteristic function* of Z_n, see *Example 3.6.2* for the linear combination of independent variables, reads

$$\begin{aligned}\varphi_{Z_n}(t) = \varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) \dots \varphi_{Y_n}\left(\frac{t}{\sqrt{n}}\right) = & \quad (1) \\ = \left[\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n \ , &\end{aligned}$$

as (1) the variables are not only independent but identically distributed: as they have the same pdf, they also have the same characteristic function. Expanding in Taylor series, see example *Example 3.6.3* for $\frac{t}{\sqrt{n}} \to 0$, the approximation of the characteristic function reads (remembering that Y_n have zero expected value and unit variance),

$$\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) \sim 1 - \frac{t^2}{2n} \ ,$$

while

$$\varphi_{Z_n}(t) = \left[\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n \sim \left[1 - \frac{t^2}{2n}\right]^n \sim e^{-\frac{t^2}{2}} \ ,$$

i.e. it converges to the characteristic function of a normal distribution $\mathcal{N}(0,1)$, see *Example 3.6.4*.

Levy's continuity theorem completes the proof. **todo**

[1] Does the CLT hold for ***heavy-tailed*** *distributions*?

3.9 Heavy-tailed distributions

- Does the *CLT* hold for heavy-tailed distributions?

3.9.1 References

- Clauset, Aaron and Woodard, Ryan, "Estimating the historical and future probabilities of large terrorist events." Annals of Applied Statistics 7(4), 1838-1865 (2013).

CHAPTER

FOUR

STOCHASTIC PROCESSES

Definition of stochastic process.

Examples.

- White noise, $\xi(t)$, is a zero-mean process with no correlation between its values at different times

$$\mathbb{E}\left[\xi(t)\,\xi(s)\right] = \delta(t-s)$$

- Wiener process (Brownian motion), $W(t)$

$$\begin{aligned}
&W(0) = 0 \\
&W(t) \text{ has independent increments} \\
&W(t) - W(s) \sim N(0, t-s) \text{ for } t > s \\
&W(t) \text{ are continuous but nowhere differentiable}
\end{aligned}$$

  Informal relation between Wiener process and white noise signal

$$\begin{aligned}
W(t) - W(s) &= \int_s^t \xi(\tau)\,d\tau \\
``\frac{dW(t)}{dt} &= \xi(t)"
\end{aligned}$$

  where the derivative relation doesn't hold in the classical sense, as $W(t)$ is nowhere differentiable

- time-discrete Markov processes

Applications

- LTI
- Stochastic differential equations…

$$dX(t) = \mu(t)\,dt + \sigma(t)\,dW(t)$$

Assumptions.

- Stationariety
- Ergodicity

$$k_{xy}(\tau) := \mathbb{E}[x(t)y(t-\tau)] = \lim_{T\to+\infty}\left\{\frac{1}{2T}\int_{t=-T}^{T} x(t)\,y(t-\tau)\,dt\right\}$$

4.1 Wiener process - Brownian motion

- Introduction: history and relation with other problems (diffusion?)
- Definition and some theory
- Simulation of Wiener process, demonstration of properties shown in theory section

4.1.1 Definition

Definition 4.1.1 (Wiener process - Brownian motion)

A Wiener process is a random process $W(t)$ with

1. initial condition, almost surely

$$W(0) = 0$$

2. increments with zero-mean *normal distribution*

$$W(t) - W(s) \sim \mathcal{N}(0, |t-s|)$$

3. W has independent increments: $W(t) - W(t+u)$ is independent from W_s, $s < t$
4. in $W(t)$ is almost surely continuous in t

Almost sure convergence in statistics

"Almost surely" here means *almost sure converngece* and it is explained in the section dealing with *convergence in statistics*, and used below to prove some properties of a Wiener process.

4.1.2 Properties

Property 4.1.1 (Covariance of increments)

Covariance of an increment follows the definition of Wiener process and the definition of *normal distribution*,

$$\mathbb{E}\left[(W(t) - W(s))^2\right] = \mathbb{E}\left[\mathcal{N}(0, |t-s|)\right] = |t-s| \ . \tag{4.1}$$

Covariance of independent increments - on non-overlapping ranges - is zero, as *independence implies no correlation*, i.e. zero covariance. Thus, if $a \le b \le c \le d$, $W(b)-W(a)$ and $W(d)-W(c)$ are independent by property (3) in *Definition 4.1.1* of Wiener process, and thus their covariance - and correlation - is zero,

$$\mathbb{E}\left[(W(b) - W(a))(W(d) - W(c))\right] = 0 \tag{4.2}$$

Covariance of two generic increments reads

$$\mathbb{E}\left[\left(W(t_1) - W(s_1)\right)\left(W(t_2) - W(s_2)\right)\right] = \left|[s_1, t_1] \cap [s_2, t_2]\right| \tag{4.3}$$

as it's proved below.

Proof of the covariance of two generic increments

If $s_1 \le s_2 \le t_2 \le t_1$,

$$\begin{aligned}
&\mathbb{E}\left[\left(W(t_1) - W(s_1)\right)\left(W(t_2) - W(s_2)\right)\right] = \\
&\quad = \mathbb{E}\left[\left(W(t_1) - W(t_2) + W(t_2) - W(s_2) + W(s_2) - W(s_1)\right)\left(W(t_2) - W(s_2)\right)\right] = \\
&\quad = \underbrace{\mathbb{E}\left[\left(W(t_1) - W(t_2)\right)\left(W(t_2) - W(s_2)\right)\right]}_{=0} + \underbrace{\mathbb{E}\left[\left(W(t_2) - W(s_2)\right)\left(W(t_2) - W(s_2)\right)\right]}_{=|t_2 - s_2|} + \\
&\quad + \underbrace{\mathbb{E}\left[\left(W(s_2) - W(s_1)\right)\left(W(t_2) - W(s_2)\right)\right]}_{=0} = \\
&\quad = 0 + |t_2 - s_2| + 0 = \left|[s_1, t_1] \cap [s_2, t_2]\right| .
\end{aligned}$$

Similarly, if $s_1 \le s_2 \le t_1 \le t_2$,

$$\begin{aligned}
&\mathbb{E}\left[\left(W(t_1) - W(s_1)\right)\left(W(t_2) - W(s_2)\right)\right] = \\
&\quad = \mathbb{E}\left[\left(W(t_1) - W(s_2) + W(s_2) - W(s_1)\right)\left(W(t_2) - W(t_1) + W(t_1) - W(s_2)\right)\right] = \\
&\quad = \underbrace{\mathbb{E}\left[\left(W(t_1) - W(s_2)\right)\left(W(t_2) - W(t_1)\right)\right]}_{=0} + \underbrace{\mathbb{E}\left[\left(W(t_1) - W(s_2)\right)\left(W(t_1) - W(s_2)\right)\right]}_{=|t_1 - s_2|} + \\
&\quad + \underbrace{\mathbb{E}\left[\left(W(s_2) - W(s_1)\right)\left(W(t_2) - W(t_1)\right)\right]}_{=0} + \underbrace{\mathbb{E}\left[\left(W(s_2) - W(s_1)\right)\left(W(t_1) - W(s_2)\right)\right]}_{=0} = \\
&\quad = 0 + |t_1 - s_2| + 0 = \left|[s_1, t_1] \cap [s_2, t_2]\right| .
\end{aligned}$$

All the other situations can be proved in the same way.

Property 4.1.2 (Statistics of maximum)

For $a \ge 0$,

$$P\left(M(t) \ge a\right) = 2P\left(W(t) \ge a\right) = 2 - 2\,\phi\left(\frac{a}{\sqrt{t}}\right) ,$$

with

$$M(t) = \max_{0 \le \tau \le t} W(\tau)$$

and

$$\phi(x) = \int_{y=-\infty}^{x} p_{\mathcal{N}(0,1)}(x)\,dx$$

is the cumulative probability function of a normal distribution $\mathcal{N}(0, 1)$.

Proof.

The second inequality immediately follows from the very definition of Wiener process with initial conditions $W(0)$,

$$\begin{aligned}
P\left(W(t) - W(0) \ge a\right) &= P\left(\mathcal{N}(0, t) \ge a\right) = && (1)\\
&= P\left(\mathcal{N}\left(0, 1\right) > \frac{a}{\sqrt{t}}\right) = \\
&= \int_{x=\frac{a}{\sqrt{t}}}^{+\infty} p(y)\,dy = && (2)\\
&= 1 - \int_{x=-\infty}^{\frac{a}{\sqrt{t}}} p(y)\,dy = 1 - \phi\left(\frac{a}{\sqrt{t}}\right)
\end{aligned}$$

having used (1) scaling rule for *transformation of probability functions*... **todo**, and (2) the normalization condition of the probability density $1 = \int_{x=-\infty}^{+\infty} p(x)\, dx$, and the definition of cumulative probability function.

First inequality. In order to prove the first inequality, it could be useful to introduce the definition of **stepping time**, τ_a, as the random variable defined as

$$\tau_a = \min_s \{ s : W(s) = a \} \ .$$

Using *reflection principle*, it follows

$$\begin{aligned}
P(M(t) \ge a) = & & (1)\\
& = P(M(t) \ge a, W(t) \ge a) + P(M(t) \ge a, W(t) < a) = & (2)\\
& = P(W(t) \ge a) + P(M(t) \ge a, W(t) - W(\tau_a) < 0) = & (3)\\
& = P(W(t) \ge a) + P(M(t) \ge a, W'(t - \tau_a) < 0) = & (4)\\
& = P(W(t) \ge a) + P(M(t) \ge a) P(W'(t - \tau_a) < 0) = & (5)\\
& = P(W(t) \ge a) + \frac{1}{2} P(M(t) \ge a) \ .
\end{aligned}$$

haing (1) used “marginalization” to write $P(A) = P(A, B) + P(A, \overline{B})$, (2) recognized that if $B : W(t) \ge a$ then $A : M(t) \ge a$ or $B \subseteq A$, and thus $P(A, B) = P(B)$, and that $a = W(\tau_a)$, (3) defined the Wiener process $W'(t - \tau_a) := W(t) - W(\tau_a)$, independent from $W(s)$, $0 \le s \le \tau_a$, (4) exploited the independence of the two conditions (**todo** *be more explicit, proof needed?*), (5) and the symmetry of Wiener process to get $P(W'(t - \tau_a) < 0) = \frac{1}{2}$.

Thus, it follows the requied relation

$$P(M(t) \ge a) = 2 P(W(t) \ge a) \ .$$

Property 4.1.3 ($W(t)$ is almost surely not differentiable)

For all time t, a Wiener process is almost surely not differentiable, i.e. ...**todo**

Proof.

todo *check details*

Wiener process is differentiable in t if the limit

$$\lim_{h \to 0} \frac{W(t+h) - W(t)}{h} = \ell$$

exists finite. Definition of limit reads,

$$\forall \varepsilon > 0 \quad \exists U_{0,\delta} \quad \text{s.t.} \quad \left| \frac{W(t+h) - W(t)}{h} - \ell \right| < \varepsilon \qquad \forall h \in U_{0,\delta} \backslash \{0\}$$

todo *how to go from this definition to the following one?*

Let E_{ε, A, t_0} be the event s.t. for a given t_0, $W(t)$ is differentiable in t_0, i.e. $\exists\ A$, ε_0 const. s.t. $\forall \varepsilon$ s.t. $0 < \varepsilon < \varepsilon_0$, $W(t) - W(t_0) \le A\varepsilon$ holds for $\forall \varepsilon$, $0 < t - t_0 \le \varepsilon$.

Let $E_{A,t_0} = \cap_\varepsilon E_{\varepsilon,A,t_0}$. Then

$$\begin{aligned}
P\left(E_{\varepsilon,A,t_0}\right) &= P\left(|W(t) - W(t_0)| \le A\varepsilon \text{ for } \forall t - t_0 \text{ s.t. } 0 < t - t_0 \le \varepsilon\right) = && (1)\\
&= P\left(M(t - t_0)| \le A\varepsilon\right) = && (2)\\
&= 1 - P\left(M(t - t_0)| \ge A\varepsilon\right) = && (3)\\
&= 1 - \left[2 - 2\phi\left(\frac{A\varepsilon}{\sqrt{\Delta t}}\right)\right] = \\
&= -1 + 2\phi\left(\frac{A\varepsilon}{\sqrt{\Delta t}}\right) \ ,
\end{aligned}$$

having used $(1)\dots$, $(2)\dots$, $(3)\dots$

Now, being $\varepsilon \le \Delta t$, it follows that $\frac{\varepsilon}{\sqrt{\Delta t}} \le \sqrt{\Delta t}$. As $\varepsilon \to 0$, then $\frac{\varepsilon}{\sqrt{\Delta t}} \to 0$, and $\phi\left(\frac{A\varepsilon}{\sqrt{\Delta t}}\right) \to \frac{1}{2}$, and $P(E_{\varepsilon,A,t_0}) \to 0$

4.2 White noise

Definition 4.2.1 (White noise - properties)

A white noise is a random process with

- zero expected value

$$\mathbb{E}[\xi(t)] = 0$$

- Dirac delta correlation

$$\mathbb{E}[\xi(t)\xi(s)] = \delta(t - s)$$

todo *link to math:functional-analysis:distributions*

Definition 4.2.2 (White noise - time derivative of Wiener process $W(t)$ in the sense of distributions)

4.3 Stochastic calculus

4.3.1 Ito's lemma

It allows to find the differential of a time-dependent function of a stochastic process. Let $f(t, x)$ be a twice-differentiable scalar function. Its Taylor series gives

$$\Delta f = \frac{\partial f}{\partial t}\Delta t + \frac{\partial f}{\partial x}\Delta x + \frac{1}{2}\frac{\partial^2 f}{\partial t^2}\Delta t^2 + \frac{\partial^2 f}{\partial t \partial x}\Delta t\, \Delta x + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\Delta x^2$$

If the argument x of the function f is chosen to be a random process X_t satisfying *Ito drift-diffusion process*,

$$dX_t = \mu_t\, dt + \sigma_t\, dW_t \ ,$$

the differential of function $f(t, X_t)$ results from the limit of Taylor series

$$\begin{aligned}
df &= \lim_{dt \to 0, dW_t \to 0} \{\Delta f\} = \\
&= \lim_{dt \to 0, dW_t \to 0} \left\{ \partial_t f dt + \partial_x f dX_t + \frac{1}{2} \left[\partial_{tt} f \, dt^2 + 2 \partial_{xt} \, dt dX_t + \partial_{xx} f dX_t^2 \right] \right\} = \\
&= \lim_{dt \to 0, dW_t \to 0} \left\{ \partial_t f dt + \partial_x f \left(\mu_t dt + \sigma_t dW_t \right) + \frac{1}{2} \left[\partial_{tt} f \, dt^2 + 2 \partial_{xt} \, dt \left(\mu_t dt + \sigma_t dW_t \right) + \partial_{xx} f \left(\mu_t dt + \sigma_t dW_t \right)^2 \right] \right\} =
\end{aligned}$$

For $dt \to 0$, $(dW_t)^2 = O(dt)$; keeping only terms of order lower than or equal to $O(dt)$, the differential becomes,

$$df = \left(\partial_t f + \mu_t \partial_x f \right) dt + \sigma_t \partial_x f \, dW_t + \frac{\sigma_t^2}{2} \partial_{xx} f \, dW_t^2 \ .$$

Replacing dW_t^2 with dt **todo** *why?*, and recalling the SDE of the Ito drift-diffusion process,

$$\begin{aligned}
df &= \left(\partial_t f + \mu_t \partial_x f + \frac{\sigma_t^2}{2} \partial_{xx} f \right) dt + \sigma_t \partial_x f \, dW_t = \\
&= \left(\partial_t f + \frac{\sigma_t^2}{2} \partial_{xx} f \right) dt + \partial_x f \left(\mu \, dt + \sigma_t \, dW_t \right) = \\
&= \left(\partial_t f + \frac{\sigma_t^2}{2} \partial_{xx} f \right) dt + \partial_x f \, dX_t \ .
\end{aligned}$$

4.3.2 Ito's calculus

Integegration w.r.t. Browinan motion produces a random variable that can be defined as

$$\int_0^t F \, dW := \lim_{n \to +\infty} \sum_{[t_{i-1}, t_i] \in \pi_n} F_{t_{i-1}} \left(W_{t_i} - W_{t_{i-1}} \right) \ ,$$

being π_n a partition of interval $[0, t]$, and H a random proces **todo** *with some characteristics...*

Example 4.3.1 (Integral of a Brownian motion w.r.t. itself)

$$Y(t) = \int_{s=0}^t W_s \, dW_s = \frac{1}{2} W_t^2 - \frac{t}{2} \ .$$

The expected value for each t of the random process Y_t is zero for all t, $\mathbb{E}\left[W_t^2\right] = 0$, as the expected value of W_t^2 is the variance of W_t, and thus t by definition of the Wiener process.

Evaluation of the integral

Let $f(t, x) = x^2$. Let's find the differential df evaluated for $x = W_t$ using *Ito's lemma*, retaining only terms with order up to $O(dt)$. Since $\partial_t f \equiv 0$,

$$df = \partial_x f|_{x = W_t} dW_t + \frac{1}{2} \partial_{xx} f|_{x = W_t} dW_t^2$$

and thus, replacing $dW_t^2 = dt$,

$$dW_t^2 = 2 W_t \, dW_t + dt \ .$$

or

$$W_t\,dW_t = d\left(\frac{W_t^2}{2}\right) - \frac{dt}{2}\ .$$

Thus (**todo** *add details if needed. A bit too much freedom in using differentials over stochastic processes here*),

$$\begin{aligned}
Y(t) &= \int_{s=0}^{t} W_s\,dW_s\,ds = \\
&= \int_{s=0}^{t} \left(\frac{W_s}{2}\right)\,ds - \int_{s=0}^{t} \frac{1}{2}\,ds = \\
&= \frac{1}{2}\left(W_t^2 - W_0^2\right) - \frac{t}{2}\ .
\end{aligned}$$

4.3.3 Ito processes

Ito drift-diffusion process

An Ito drift-diffusion process is a stochastic process satisfying the stochastic differential equation (SDE)

$$dX_t = \mu_t\,dt + \sigma_t\,dW_t\ , \tag{4.4}$$

with W_t a Wiener process. If $\mu_t = \mu$, $\sigma_t = \sigma$ are constant a closed-form solution can be found using *Ito's lemma*, for $f(t,x) = x$, or by direct (stochastic) integration of the SDE (4.4), as

$$\begin{aligned}
\int_{s=0}^{t} dX_s &= \int_{s=0}^{t} \mu\,ds + \int_{s=0}^{t} \sigma\,dW_s \\
X_t - X_0 &= \mu t + \sigma\left(W_t - W_0\right)\ ,
\end{aligned}$$

so that $X_t - X_0 \sim \mathcal{N}\left(\mu t, \sigma^2 t\right)$.

Scaling of a Wiener process

Term σW_t represents a scaling of a Wiener process $W_t \sim \mathcal{N}(0,t)$ with zero expected value and variance t. Multiplication by factor σ results in a multiplication of the expected value by σ and variance by σ^2.

Geometric Brownian Motion, GBM

A geometric Brownian motion is a stochastic process satisfying the SDE

$$dX_t = \mu X_t\,dt + \sigma X_t\,dW_t\ .$$

Example 4.3.2 (GBM in Finance)

GBM can be used as a model of the price of an asset with constant expected return and variance of returns with normal distribution.

Let $f(x) = \ln x$ be evaluated for $x = X_t$. Ito's lemma, with $\partial_t f \equiv 0$, provides the expression of the differential

$$\begin{aligned}
df &= \partial_x f|_{X_t} dX_t + \frac{1}{2}\partial_{xx} f|_{X_t} dX_t^2 = \\
&= \partial_x f|_{X_t} \left(\mu X_t \, dt + \sigma X_t \, dW_t\right) + \frac{1}{2}\partial_{xx} f|_{X_t} \left(\mu X_t \, dt + \sigma X_t \, dW_t\right)^2 = \\
&= \frac{1}{X_t}\left(\mu X_t \, dt + \sigma X_t \, dW_t\right) - \frac{1}{2}\frac{1}{X_t^2}\sigma^2 X_t^2 \, dW_t^2 = \\
d\left(\ln X_t\right) &= \left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma \, dW_t \ ,
\end{aligned}$$

whose solution after integration reads

$$\ln X_t = \ln X_0 + \left(\mu - \frac{\sigma^2}{2}\right) t + \sigma \, W_t \ ,$$

or

$$X_t = X_0 \, e^{\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t} \ .$$

Geometric Brownian Motion with drift

A geometric Brownian motion is a stochastic process satisfying the SDE

$$dX_t = \mu X_t \, dt - C \, dt + \sigma X_t \, dW_t \ .$$

Example 4.3.3 (GBM with constant withdrawal in finance)

GBM with drift can be used in finance as a model to represent DCA strategy and pension withdrawal, and to show and discuss **sequence risk**.

The solution reads

$$X_t = X_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)(t - t_0) + \sigma(W_t - W_0)} + \int_{s=0}^{t} C e^{\left(\mu - \frac{\sigma^2}{2}\right)(t-s) + \sigma(W_t - W_s)} \, ds \ .$$

Integration factor method for linear SDEs

Integration factor method for linear SDEs

$$dX_t = a \, dt + b \, dW_t \ ,$$

with $a(X_t, t, W_t)$, $b(X_t, t, W_t)$

aims at finding an exponential factor $e^{\alpha t + \beta W_t}$ that allows to get an integrable expression of the differential

$$d\left(e^{\alpha t + \beta W_t} X_t\right) = d \, f\left(t, W_t, X_t\right) \ .$$

Taylor expansion of this expression up to terms of order $dt \sim dW_t^2$ reads

$$\begin{aligned}
d\,f(t, W_t, X_t) &= \partial_t f\,dt + \partial_w f\,dW_t + \partial_x f \underbrace{dX_t}_{a dt + b dW_t} + \\
&+ \frac{1}{2}\left(\underbrace{\partial_{tt} f\,dt^2}_{o(dt)} + \partial_{ww} f \underbrace{dW_t^2}_{dt} + \partial_{xx} f \underbrace{dX_t^2}_{b^2\,dW_t^2 = b^2\,dt} + \underbrace{2\partial_{tw} f\,dt\,dW_t + 2\partial_{tx} f\,dt\,dX_t}_{o(dt)} + 2\partial_{xw} f \underbrace{dW_t\,dX_t}_{b dW_t^2 = b dt} \right) = \\
&= dt\left[\partial_t f + a\partial_x f + \frac{1}{2}\partial_{ww} f + \frac{1}{2} b^2 \partial_{xx} f + 2b\partial_{xw} f \right] + dW_t\left[\partial_w f + b\partial_x f \right]
\end{aligned}$$

Proof (with integration factor method, for linear SDEs)

GBM motion with drift and constant coefficients is governed by SDE

$$dX_t = \mu X_t\,dt - C\,dt + \sigma X_t\,dW_t\;.$$

Referring to the general expression of SDEs, coefficients a, b of the GBM with drift read

$$\begin{aligned}
a &= \mu\,X_t + C \\
b &= \sigma\,X_t\;.
\end{aligned}$$

Partial derivatives of function $f = e^{\alpha t + \beta w}\,x$ appearing in the solution of SDEs through integration factor method read

$$\begin{aligned}
\partial_t f &= e^{\alpha t + \beta w}\,x\,\alpha \\
\partial_w f &= e^{\alpha t + \beta w}\,x\,\beta \\
\partial_x f &= e^{\alpha t + \beta w} \\
\partial_{xx} f &= 0 \\
\partial_{wx} f &= e^{\alpha t + \beta w}\,\beta \\
\partial_{ww} f &= e^{\alpha t + \beta w}\,x\,\beta^2
\end{aligned}$$

It's now possible to simplify the RHS of the the expression of the differential df. Namely, it's possible to choose values of α, β in order to get simpler expressions of the factors of the differentials dt and dW_t

$$\begin{aligned}
dW_t: \quad \partial_{W_t} f &= \left.\left(\partial_w f + b\partial_x f \right)\right|_{t, W_t, X_t} = \\
&= e^{\alpha t + \beta W_t}\,\left(X_t \beta + b \right) = \\
&= e^{\alpha t + \beta W_t}\,X_t\left(\beta + \sigma \right) \\
dt: \quad \partial_t f &= \left.\left[\partial_t f + a\partial_x f + \frac{1}{2}\partial_{ww} f + \frac{1}{2} b^2 \partial_{xx} f + b\partial_{xw} f \right]\right|_{t, W_t, X_t} \\
&= e^{\alpha t + \beta W_t}\left[X_t \alpha + a + \frac{1}{2} X_t \beta^2 + 0 + b\beta \right] = \\
&= e^{\alpha t + \beta W_t}\left[X_t \alpha + \mu X_t + C + \frac{1}{2} X_t \beta^2 + \sigma X_t \beta \right] = \\
&= e^{\alpha t + \beta W_t}\left[X_t\left(\alpha + \mu + \frac{1}{2}\beta^2 + \sigma\beta \right) + C \right]\;.
\end{aligned}$$

Setting

$$\begin{aligned}
\beta &= -\sigma \\
\alpha &= -\mu - \frac{1}{2}\beta^2 - \sigma\beta = -\mu + \frac{\sigma^2}{2}\;,
\end{aligned}$$

the differential df becomes

$$d\left(e^{\left(-\mu+\frac{\sigma^2}{2}\right)t-\sigma W_t}\,X_t\right) = Ce^{\left(-\mu+\frac{\sigma^2}{2}\right)t-\sigma W_t}$$

and integration gives

$$X_t = X_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)(t-t_0)+\sigma(W_t-W_0)} + \int_{s=0}^{t} Ce^{\left(\mu-\frac{\sigma^2}{2}\right)(t-s)+\sigma(W_t-W_s)}\,ds\ .$$

Part III

Statistical Inference

CHAPTER

FIVE

INTRODUCTION TO STATISTICAL INFERENCE

Part IV

Introduction to Machine Learning

CHAPTER

SIX

INTRODUCTION TO MACHINE LEARNING

Artificial intelligence can be broadly defined as a field dealing with making machines perform tasks that require intelligence, when performed by humans, like: reasoning, perception, representation, language processing, planning, learning

Machine learning is a branch of AI focused on statistical algoritms that can **learn from data** and **generalize to unseen data** and perform tasks, without explicit instructions.[1]

Three core paradigms. Algorithms in machine learning can be divided into three paradigms:

- *Supervised Learning, SL*: algorithm learns from labelled data; many applications can be reduced to 2 main tasks: **regression** (or function approximation) and **classification**.
- *Unsupervised Learning, UL*: algorithm learns pattern from un-labelled data; examples of taks in UL are clustering, dimensionality reduction (and recognition of *main* components in data), compression (retaining only relevant components in data). Some historical algorithms and linear algebra decompositions can be interpreted or generalized as unsupervised learning.
- *Reinforcement Learning, RL*: an algorithm (**agent**) learns a **policy** - i.e. the way to behave - interacting with an **environment**, and maximizing some performance to efficiently perform required tasks. Applications of RL includes **planning** and **control**.

Goals and methodology. ML is mainly a engineering-oriented and an application-focused discipline, relying on statistical inference (**todo** *be more explicit*). A *ML model* usually takes an input $\mathbf{u}$, and produces an output $\mathbf{y}$, depending on its own structure and a set of parameters θ and hyper-parameters μ. Learning usually relies on **optimization** of an objective function

$$L(\theta; \mu) \ ,$$

w.r.t. parameters θ, whose value is learned/adjusted towards an optimal solution θ^* that makes $L(\theta^*; \mu)$ extreme. The choice of hyper-parameters μ instead influences the training process and model behavior. Optimization usually relies on gradient methods, updating the parameters in the direction of the gradient of the objective function w.r.t. the parameters,

$$\theta \ \leftarrow \ \theta + \alpha \nabla_\theta L(\theta; \mu) \ .$$

Optimization of model parameters is made fast by the use of **back-propagation** and **automatic differentiation** (AD), which efficiently compute gradients of the cost function with respect to the model's parameters, and technically feasible for large-dimensional models - as the ones used in multi-layered neural networks, in deep learning[2] - by recent hardware improvement. These algorithms are not only feasible but also particularly well-suited (being a major driver for new designs) to modern processing architectures, such as **GPUs** and **TPUs**, that accelerate the large-scale matrix and tensor computations involved in both the forward and backward passes of training.

todo *Show NVIDIA, TSMC revenues*

[1] "Without explicit instructions" means that a systems has no user-coded behavior, but learns it usually via **optimization**, usually either involving minimization of an error function or maximization of an objective function or energy/information content.

[2] Deep learning can be roughly defined as that branch of machine learning using multi-layered neural networks, indeed.

---

todo Add references: Bishop,…

6.1 Models in Machine Learning

Linear models

Kernel methods

Decision trees and ensembles

Neural networks

…probabilitstic models, clustering models, dimensionality reduction models,…

Reinforcement learning models: Q-learning (tabular, and DQN), Policy gradient, Actor-Critic, Proximal Policy Optimization,…

6.2 Good Practices in Machine Learning

CHAPTER

SEVEN

SUPERVISED LEARNING

Theory.

Examples.

7.1 SL: theory

Supervised learning can be thought as a **function approximation problem**. Given a set of data

$$\{(x_i, y_i)\}_{i=1:N} \ ,$$

supervised learning can be formulated as the evaluation of a function $\hat{y}(x, \theta)$, or a **model**, that *approximates well* the relation between input x_i and output y_i,

$$y_i \simeq \hat{y}\left(x_i; \theta\right) \ .$$

Two main tasks of SL can be distinguished on the output of the function: **regression** can be formulated as function approximation with continuous output, while in **classification** the function maps inputs to discrete output/**labels**

Learning process aims at finding values of the parameters θ (and hyper-parameters μ), that minimize a "prediction" error function, e.g. for a scalar output function,

$$E(\theta) = \frac{1}{2} \sum_{i \in D_{Tr}} |\hat{y}(x_i; \theta) - y_i|^2 \ ,$$

being D_{Tr} the set of indices belonging to the *training set*. Minimization usually relies on gradient methods of the error function w.r.t. the parameters θ,

$$\begin{aligned}
\nabla_\theta E(\theta, \mathbf{x}_{Tr}, \mathbf{y}_{Tr}) &= \sum_{i \in D_{Tr}} \left(\hat{y}(x_i; \theta) - y_i\right) \nabla_\theta \hat{y}(x_i; \theta) \\
\theta &\leftarrow \theta - \alpha \nabla_\theta E(\theta, \mathbf{x}_{Tr}, \mathbf{y}_{Tr}) \ ,
\end{aligned}$$

with α an hyper-parameter called *learning rate*, governing the "length" of the update step. Other objective functions to be maximised or minimized can be used. Slight variations to objective functions allow for regularization (e.g. parameter weighting)

Dataset. Available data $\{x_i, y_i\}_i$ is divided in different sets:

- training set: for learning/tuning model parameters, minimizing an error function
- validation set: for early stopping, and hyper-parameter tuning (e.g. to avoid
- test set: to evaluate model performance

CHAPTER

EIGHT

UNSUPERVISED LEARNING

CHAPTER

NINE

REINFORCEMENT LEARNING

PROOF INDEX