Introduction to Stochastic Processes

Pavel Chigansky

E-mail address: pchiga@mscc.huji.ac.il

Preamble

These are lecture notes for the course I have taught at the School of Electrical Engineering of Tel Aviv University during 2003/04. It was intended as a "pre-introduction" to the measure theoretic probability for the graduate students with background in signal processing, information theory, etc. Consequently, most of the theory is told without proofs and much time is allocated to motivating examples. In particular, the classical filtering problem for partially observed processes is revisited on several occasions as the story unfolds. The lecture notes are available at `http://pluto.huji.ac.il/~pchiga/` along with the exercises/solutions files and a number of sample exams.

Please do not hesitate to send any questions, comments, etc. regarding this course to the author's current e-mail `pchiga@mscc.huji.ac.il`.

P.Ch.

30, September, 2007

Contents

CHAPTER 1

The Basics of Mathematical Probability

1. Introduction

What is probability[1] ? Toss a coin n (e.g. 1000) times and let p_n be the ratio of heads outcomes. Intuitively we feel that p_n will be close to $1/2$ if the coin is fair, so it is tempting to say that p_n is the *probability* of heads. Also it is clear that p_n will rarely be equal to $1/2$ exactly and moreover it will depend on n as well as will be different when repeating this experiment (for the same n). Also what is probability then in just one trial ? Clearly the above experiment cannot be used to *define* probability ...

Introduce $\Omega = \{h, t\}$ (h - heads, t - tails), the set of possible outcomes of one coin tossing (i.e. Ω consists of 2 points $\omega_1 = h$ and $\omega_2 = t$). *Define* probability $P(\cdot)$ to be an $\Omega \mapsto [0, 1]$ function, such that $P(\Omega) = P(h) + P(t) = 1$. Let for brevity $p = P(h)$. Now return to n-time tossing of the coin. Let $\Omega = \{(x_1, ..., x_n) : x_i \in \{h, t\}\}$. This set consists of 2^n points ω_i (e.g. $\omega_1 = (hh...h)$ etc.) to which we *assign*, say, equal probabilities, i.e. $P(\omega_i) = 2^{-n}$. Now let us check that this confirms the intuition of the first paragraph [2] (ω_i^ℓ is the value of ℓ-th entry of ω_i)

$$P\Big(\omega_i : \frac{1}{n}\sum_{\ell=1}^{n} I(\omega_i^\ell = h) = 1/2\Big) = 2^{-n}\binom{n}{n/2} = 2^{-n}\frac{n!}{(n/2!)^2} \approx 0 \text{ for large } n$$

i.e. rarely exactly $1/2$ is obtained. Fix some small $\varepsilon > 0$, and let us verify that

$$P\Big(\omega_i : \Big|\frac{1}{n}\sum_{\ell=1}^{n} I(\omega_i^\ell = h) - 1/2\Big| \le \varepsilon\Big)$$

is close to 1 for large n. This can be calculated directly using combinatorics but the rough answer can be obtained in a simpler way by means of Chebyshev inequality (to be elaborated later on in the course).

$$P\Big(\omega_i : \Big|\frac{1}{n}\sum_{\ell=1}^{n} I(\omega_i^\ell = h) - 1/2\Big| > \varepsilon\Big) \le \frac{1}{4n\varepsilon^2}.$$

The above construction works well for any *finite* n. What about an infinite sequence of coin tosses ? I.e. when $\Omega = \big\{\omega = (x_1, x_2, ...) : x_i \in \{h, t\}\big\}$. Now Ω consists of an infinite number of points. If there were countably many points, one could assign to each ω_i probability p_i, such that $\sum p_i = 1$. But this is not the case - indeed, recall that any infinite sequence of 0 and 1 can be considered as a binary expansion of a number in $[0, 1]$ and thus Ω contains as many points as the interval $[0, 1]$, i.e. uncountably many. This implies that the probability of each point $\omega \in \Omega$ should be zero ! At the same time, intuitively we feel that probability of a point to

[1] The (much more) extended treatment of the subject can be found e.g. in (8)

[2] even n are taken for simplicity

be in $[0, 1/2]$ is $1/2$. Clearly some sets, other than points, should be used to define probability in this case.

2. Axioms of probability

The main object of the probability theory is the probability space $(\Omega, \mathcal{F}, P)$, where

(1) Ω is a set of points ω, which is called the sampling space
(2) $\mathcal{F}$ is the σ-algebra of the sets (events) , satisfying the properties
 - $\Omega \in \mathcal{F}$
 - $A \in \mathcal{F} \implies \bar{A} \in \mathcal{F}$
 - if A_n is a sequence of events from $\mathcal{F}$, then

$$\cap_{n=1}^{\infty} A_n \in \mathcal{F} \text{ and } \quad \cup_{n=1}^{\infty} A_n \in \mathcal{F}.$$

(3) P is the probability measure, i.e. positive function $\mathcal{F} \mapsto [0, 1]$, such that $P(\Omega) = 1$ and for any sequence of pairwise nonintersecting sets A_n in $\mathcal{F}$ the σ-additivity property is satisfied, i.e.

$$P\big(\sum_n A_n\big) = \sum_n P(A_n).$$

Example 2.1. In a single coin tossing experiment
- $\Omega = \{h, t\}$
- $\mathcal{F} = \{\emptyset, \{h\}, \{t\}, \Omega\}$ (check that this is algebra)
- $P(\emptyset) = 0$, $P(\Omega) = 1$ and $P(h) = P(t) = 1/2$

Example 2.2. In n coin tosses
- $\Omega = \big\{\omega = (x_1, ..., x_n : x_i \in \{h, t\})\big\}$
- $\mathcal{F} = \{\emptyset, \omega_i \in \Omega \text{ and all possible unions of } \ \omega_i\}$
- for any set $A \in \mathcal{F}$

$$P(A) = \sum_{\omega_i \in A} 2^{-n}$$

The σ-algebra (which is just a finite algebra in this case) $\mathcal{F}$ is the richest (or finest) , i.e. any physical outcome of the experiment belongs to $\mathcal{F}$. In fact many other (more coarse) σ-algebras can be defined on the same Ω. E.g. $\mathcal{F}' = \{\emptyset, E, O, \Omega\}$, where $E = \{\omega_i : \sum_{\ell=1}^{n} \omega_i^{\ell} \text{ is even}\}$ and $O = \{\omega_i : \sum_{\ell=1}^{n} \omega_i^{\ell} \text{ is odd}\}$ (check that this is algebra!). Another probability measure P' is assigned on $\mathcal{F}'$ via $P'(O) = P'(E) = 1/2$.

Exercise 2.3. *Show that the union of two σ-algebras is not necessarily a σ-algebra. Show that an intersection of two σ-algebras is always a σ-algebra.*

Note that in the above examples Ω was always finite and thus $\mathcal{F}$ definition was immediate. What happens e.g. in the infinite coin tossing experiment ?

As before $\Omega = [0, 1]$ can be chosen. Let $\mathcal{I}$ be the collection of sets, obtained by finite unions of nonintersecting intervals of the form $(a, b]$, i.e.

$$A = \sum_i (a_i, b_i], \quad b_i > a_i$$

Clearly $\mathcal{I}$ is algebra (why?). But not a σ-algebra ! E.g. $A_n = (0, 1 - 1/n]$ are in $\mathcal{I}$, but $(0, 1) = \cup A_n$ is not. The minimal (coarsest) σ-algebra, which contains $\mathcal{I}$ is called *Borel* σ-algebra and denoted by $\mathcal{B}$. Clearly such a σ-algebra exists and is obtained by intersection of all σ-algebras containing $\mathcal{I}$ (there is at least one - the finest σ-algebra on Ω).

The Borel σ-algebra is rich enough to contain many events of interest, e.g.

$$\begin{aligned}&\{a\} = \bigcap_n (a-1/n, a] \quad (\textit{points})\\ &(a,b) = \bigcup_n (a, b-1/n] \quad (\textit{open intervals})\\ &\text{etc.}\end{aligned}$$

There are of course sets that do not belong to $\mathcal{B}$.

We can assign the probability P on $\mathcal{I}$ by $P(A) = \sum_i (b_i - a_i)$. Fortunately (and by no means obviously!) this defines a unique probability λ on $\mathcal{B}$, such that for any set $A \in \mathcal{I}$

$$\lambda(A) = P(A),$$

i.e. the restriction of λ on $\mathcal{I}$ coincides with P. In other words, λ is an extension of P on the σ-algebra $\mathcal{B}$, generated by $\mathcal{I}$. Moreover e.g. for any interval (semi closed, closed, open, etc.)

$$\lambda\big((a,b)\big) = b - a.$$

λ is called Lebesgue probability measure.

Now the infinite coin tossing experiment can be studied on the probability space $(\Omega, \mathcal{B}, \lambda)$. E.g. the event (we associate t with 0, and h with 1)

$$A = \{\text{first 17 trials give } t\} = [0, 2^{-17}) \in \mathcal{B}$$

and $\lambda(A) = 2^{-17}$. More fancy events are measurable with respect to $\mathcal{B}$, e.g.

$$\Big\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^n I(\omega^\ell = h) = 1/2\Big\}$$

is $\mathcal{B}$-measurable, since it may be constructed by countable number of union and intersections of more simple $\mathcal{B}$-measurable sets (similar to Section 5.2, p. 12).

3. Probability spaces

As we have seen, the probability space $([0,1], \mathcal{B}, \lambda)$ is sufficient for the coin tossing. More complicate experiments require dealing with other probability spaces.

3.1. $\mathbb{R}$. Consider $\Omega = \mathbb{R}$. The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ can be introduced via semi-open intervals as in the previous section. It turns out that $\mathcal{B}(\mathbb{R})$ coincides with the σ-algebra, generated by the open sets (with the usual distance metric).

A probability measure on $\mathcal{I}(\mathbb{R})$ can be defined by means of a *distribution* function $F(x)$, satisfying the properties

(1) $F(x)$ is a positive nondecreasing function
(2) $F(\infty) = 1$, $F(-\infty) = 0$
(3) $F(x)$ is continuous from the right and has limit from the left for any $x \in \mathbb{R}$

Then for any set $A = \sum_i (a_i, b_i]$ in $\mathcal{I}$ let

$$P(A) = \sum_i \big[F(b_i) - F(a_i)\big].$$

Similarly to Lebesgue measure, this probability measure can be extended to $\mathcal{B}(\mathbb{R})$.

All distribution functions can be one of three kinds (or their combination)

(1) lattice (or purely atomic), if all the increase points of $F(x)$ are isolated (discrete)

(2) absolutely continuous (with respect to extended Lebesgue measure), if there is a nonnegative function $f(x)$ such that

$$F(x) = \int_{-\infty}^{x} f(s)ds$$

where the integral can be understood in the usual Reinman sense

(3) singular, if the increase points of $F(x)$ have Lebesgue measure zero (e.g. Cantor distribution)

Example 3.1.

1. Normal distribution

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2} ds$$

2. Cauchy distribution

$$F(x) = \int_{-\infty}^{x} \frac{1/\pi}{s^2+1} ds$$

3.2. $\mathbb{R}^n$**.** When $\Omega = \mathbb{R}^n$, $\mathcal{B}(\mathbb{R}^n)$ is generated by sets of the form

$$A = \{(-\infty, x_1] \times ... \times (-\infty, x_n]\}.$$

Any probability measure can be constructed by means of n-dimensional distribution function $F(x)$, $x \in \mathbb{R}^n$, satisfying the properties

(1) $\Delta_{a_1,b_1} ... \Delta_{a_n,b_n} F(x) \ge 0$, where Δ_{a_i,b_i} is the difference operator, applied to i-th coordinate

(2) F is continuous from the right (jointly in all the arguments)

(3) $F(\infty, ..., \infty) = 1$

(4) $\lim_{x_i \to -\infty} F(x_1, ..., x_n) = 0$

As in the case of $\mathbb{R}$, F can be atomic, absolutely continuous and singular.

3.3. $\mathbb{R}^\infty$**.** Often we deal with infinite sequences with entries in $\mathbb{R}$ (rather than in $\{0, 1\}$, as in coin tossing case), i.e. with the sampling space $\mathbb{R}^\infty = \{(x_1, x_2, ...) : x_i \in \mathbb{R}\}$. The Borel σ-algebra is generated in this case by *cylindrical* sets of the form

$$I_n = \{x \in \mathbb{R}^\infty : x_1 \in (a_1, b_1], ..., x_n \in (a_n, b_n]\}$$

for $n = 1, 2, ...$ The probability measure P can be defined on any cylindrical set as in the case of $\mathbb{R}^n$. When does this uniquely define a probability measure on $\mathcal{B}(\mathbb{R}^\infty)$? The answer was given by A. Kolmogorov

Theorem 3.2. *Let $P_1, P_2, ...$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$, ..., which satisfy the consistency property, i.e.*

$$P_{n+1}(B \times \mathbb{R}) = P_n(B), \quad B \in \mathcal{B}(\mathbb{R}^n).$$

Then there is a unique probability measure P on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$, such that for any $n = 1, 2, ...$

$$P(B) = P_n(B), \quad B \in \mathcal{B}(\mathbb{R}^n).$$

□

Roughly speaking, the latter means that the measure on $\mathcal{B}(\mathbb{R}^\infty)$ is defined by all the finite dimensional measures.

Example 3.3. Let $F(x)$ be a distribution function on $\mathbb{R}$. Define an n-dimensional distribution function $F_n(x) = \prod_{i=1}^n F(x_i)$ on $\mathbb{R}^n$. Clearly the family of measures P_n, corresponding to $F_n(x)$ is consistent and so there is a probability measure P on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$, such that all its n-marginal distributions coincide with $F_n(x)$.

4. Random variables

Let's fix some probability space $(\Omega, \mathcal{F}, P)$.

Definition 4.1. A real function $\xi : \Omega \mapsto \mathbb{R}$ is called a random variable if

$$\{\omega : \xi(\omega) \in B\} \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Random variables are often used to describe outcomes of the experiments without specifying the underlying probability space. The probabilistic description of ξ is then given by its distribution function

$$F_\xi(x) = P(\{\omega : \xi(\omega) \le x\}),$$

which is the probability measure, *induced* by ξ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Example 4.2. On the space $([0,1], \mathcal{B}, \lambda)$ consider the random variable

$$\xi(\omega) = \mathrm{sign}(\omega - 1/2)$$

Verify that

$$F_\xi(x) = \begin{cases} 0, & x < -1 \\ 1/2, & -1 \le x < 1 \\ 1, & 1 \le x \end{cases}$$

Example 4.3. On $([a,b]^\infty, \mathcal{B}([a,b]^\infty), P)$, $a < b$, let $\xi(\omega) = \overline{\lim}_{n\to\infty} \omega^n$. $\xi(\omega)$ is a random variable (why?) Calculation of $F_\xi(x)$ can be very subtle in this case.

4.1. Expectation and its properties.

A random variable is simple if it has the form

$$X(\omega) = \sum_{i=1}^n x_i I(\omega \in A_i),$$

where A_i are disjoint sets in $\mathcal{F}$. The *expectation* of a simple r.v. is by definition

$$EX = \sum_{i=1}^n x_i P(A_i).$$

An arbitrary random variable (not necessarily simple) is said to be Lebesgue integrable (or to have expectation EX) if there is a sequence of simple random variables X_n, converging to X uniformly, in which case $EX := \lim_{n\to\infty} EX_n$ is defined. Let us verify the correctness of this definition: i.e. (1) that for any uniformly convergent sequence $\lim_{n\to\infty} EX_n$ exists and (2) it does not depend on the choice of the sequence. The first claim holds since EX_n is a Cauchy sequence of numbers

$$\big|EX_n - EX_m\big| \le E|X_n - X_m| \le \sup_\omega |X_n(\omega) - X_m(\omega)| \xrightarrow{n,m\to\infty} 0,$$

while the second claim is true by the following argument. Let $X'_n(\omega)$ be another approximating sequence of simple random variables in the sense $\lim_{n\to\infty} X'_n(\omega) = X(\omega)$ and assume that $\lim_{n\to\infty} EX'_n \neq \lim_{n\to\infty} EX_n$. Let

$$X''_n(\omega) = \begin{cases} X'_n(\omega), & n \text{ is even} \\ X_n(\omega), & n \text{ is odd} \end{cases}$$

Clearly $X''_n(\omega)$ is a sequence of simple r.v. and $\lim_{n\to\infty} X''_n = X$ as well. However EX''_n does not converge, which is a contradiction (to existence!) and hence the Lebesgue integral does not depend of the approximating sequence.

An approximating sequence can be constructed explicitly: for a nonnegative r.v. X let

$$X_n(\omega) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I\left(\frac{k-1}{2^n} \le X(\omega) \le \frac{k}{2^n}\right) + nI\Big(X(\omega) \ge n\Big).$$

Note that $X_n(\omega) \nearrow X(\omega)$. In this case the sequence EX_n does not decrease and hence has a limit (possibly infinite). The limit $EX = \lim_{n\to\infty} EX_n$ is the Lebesgue integral (expectation) of X. For general r.v. X let $X^+ = \max(X, 0)$ and $X^- = -\min(X, 0)$. If $X^+ < \infty$ and $X^- < \infty$, then the expectation is defined as

$$EX = EX^+ - EX^-.$$

The expectation satisfies the following properties:

(1) Let c be a constant and assume that EX exists, then $EcX = cEX$.
(2) $X \le Y$ P-a.s. [3] $\implies$ $EX \le EY$
(3) If $E|X| < \infty$ and $E|Y| < \infty$, then $E(X + Y) = EX + EY$
(4) If $X = Y$ P-a.s. then $EX = EY$
(5) If $X \ge 0$, $EX = 0$, then $X = 0$ P-a.s.
(6) (Chebyshev inequality) if $X \ge 0$ P-a.s. then

$$P(X \ge a) \le \frac{EX}{a}$$

(7) (Cauchy-Schwarz inequality) assume that $EX^2 < \infty$ and $EY^2 < \infty$, then

$$E|XY| \le \sqrt{EX^2 EY^2}$$

(8) (Jensen inequality) assume that $E|X| < \infty$. Then for any convex function $g(x)$

$$g(EX) \le Eg(X)$$

(9) (Lyapunov inequality) assume that $E|X|^p < \infty$. Then for any $0 < q \le p$

$$\Big(E|X|^q\Big)^{1/q} \le \Big(E|X|^p\Big)^{1/p}$$

Proof. Set $r = p/q$, so that the function x^r is convex. Then by Jensen inequality

$$(E|X|^q)^r \le E|X|^{qr} = E|X|^p$$

□

[3] $X < Y$ P-a.s. (P-almost surely) means that $P(\omega : X(\omega) < Y(\omega)) = 1$, which is clearly weaker than $X(\omega) < Y(\omega)$ for any $\omega \in \Omega$. It is customary to neglect sets of zero P measure in probability theory.

Expectation of X can be calculated with respect to the induced measure (read distribution function) rather than the original measure P

$$EX = \int_\Omega X(\omega)dP(\omega) = \int_\mathbb{R} xdF(x).$$

Similarly if $Y = f(X)$ for some measurable function f, then

$$EY = \int_\Omega f(X(\omega))dP(\omega) = \int_\mathbb{R} f(x)dF(x) = \int_\mathbb{R} ydG(y)$$

where $G(y)$ is the distribution of Y.

4.2. Characteristic function. The Fourier transform of the distribution of X is called *characteristic function* of X:

$$\varphi(\lambda) = Ee^{iX\lambda} = \int_\mathbb{R} e^{ix\lambda}dF(x).$$

Due to the properties of the Fourier transform, $\varphi(\lambda)$ uniquely defines $F(x)$.

4.3. Independence. The sets (events) A and B are independent if

$$P(A\cap B) = P(A)P(B),$$

so that the *conditional* probability of A given B ($0 = \frac{0}{0}$ by convention)

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

equals $P(A)$.

Similarly two random variables are said to be independent if their joint distribution function can be factored into

$$F_{XY}(x,y) = F_X(x)F_Y(y).$$

This is equivalent to

(1) $Ef(X)g(Y) = Ef(X)Eg(Y)$ for all bounded functions f and g
(2) $E\exp\{i\lambda X + i\mu Y\} = E\exp\{i\lambda X\}E\exp\{i\mu Y\}$ for any real λ and μ

5. Random processes

The collection of real valued random variables $X = (X_n)_{n\ge 1}$, $n\in\mathbb{N}$, defined on some probability space $(\Omega, \mathcal{F}, P)$ is called random process (sequence) with discrete time.

For any fixed $\omega'\in\Omega$, the sequence $X_n(\omega')$ is called *trajectory* or realization of the random process X. For any fixed n', $X'_n(\omega)$ is a random variable.

Any random process induces a probability measure P_X on the measurable space $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$, which is called the distribution of the process. The projection of this distribution on any fixed finite set of indices $\{n_1, ..., n_d\}$ is called d-dimensional distribution of X, i.e.

$$F_{n_1\dots n_d}(x_1,...,x_d) = P\big(X_{n_1}\le x_1; ...; X_{n_d}\le x_d\big).$$

By Kolmogorov's Theorem 3.2 for any family of consistent finite dimensional distributions there is a probability space on which X can be defined.

5.1. Examples of random processes.

5.1.1. *i.i.d. process.* is the sequence of independent identically distributed r.v. E.g. $\varepsilon = (\varepsilon_n)_{n\ge 1}$ with ε_1 being a standard Gaussian r.v.

5.1.2. *Autoregressive process.* Let ε_n be an i.i.d sequence. The AR process is generated recursively by

$$X_n = \sum_{k=1}^{q} \alpha_k X_{n-k} + \varepsilon_n$$

with a fixed array of numbers $\{\alpha_1, ..., \alpha_q\}$.

5.1.3. *Moving average process.* Let ε_n be and i.i.d. sequence. The MA process is obtained by

$$X_n = \sum_{k=0}^{p} \beta_k \varepsilon_{n-k}$$

where $\{\beta_0, ..., \beta_p\}$ are constants.

5.1.4. *Markov process.* Let $p(x, A)$ be $\mathbb{R} \times \mathcal{B}(\mathbb{R}) \mapsto [0, 1]$ function, such that for any fixed x it is a probability measure on $\mathcal{B}(\mathbb{R})$ and for any fixed $A \in \mathcal{B}(\mathbb{R})$ is a measurable function on $\mathbb{R}$. Let μ be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Define a probability measure on $\mathcal{B}(\mathbb{R}^n)$

$$P_{m_1 ... m_n}(A_{m_1} \times ... \times A_{m_n}) = \int_{\mathbb{R}} \int_{A_{m_1}} \int_{\mathbb{R}} ... \int_{A_{m_n}} \mu(dx_1) \prod_{\ell=1}^{m_n} p(x_{\ell-1}, dx_\ell),$$

where the integration with respect to x_ℓ, $\ell \notin \{m_1, ..., m_n\}$ is on the whole $\mathbb{R}$. Verify that this family of probability measures is consistent. The corresponding random process is called Markov.

5.2. Convergence of random processes. Recall that an infinite sequence of numbers $z = (z_n)_{n \ge 1}$ converges to a limit z ($\lim_{n \to \infty} z_n = z$) if for each $\varepsilon > 0$ there is an index n'_ε, such that for all $n \ge n'_\varepsilon$ $|z_n - z| \le \varepsilon$.

Unlike numeric sequence, the sequence of functions, which any random process is, may converge to a limit (function!) in many different senses.

Definition 5.1. The r.p. X_n converges to a r.v. X pointwise if

$$\lim_{n \to \infty} X_n(\omega) = X(\omega)$$

for all $\omega \in \Omega$.

Definition 5.2. The r.p. X_n converges to a r.v. X in $\mathbb{L}^p$ $p \ge 1$, if $E|X_n|^p < \infty$, $n \ge 1$ and

$$\lim_{n \to \infty} E\big|X - X_n\big|^p = 0$$

The convergence in $\mathbb{L}^1$ and in $\mathbb{L}^2$ is often called convergence in the mean and in the mean square respectively.

Definition 5.3. The r.p. X_n converges to r.v. X in probability if for any $\varepsilon > 0$

$$\lim_{n \to \infty} P(|X_n - X| \ge \varepsilon) = 0.$$

Definition 5.4. The r.p. X_n converges to r.v. X P-a.s. (or with probability one) if

$$P\big(\omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\big) = 1.$$

Definition 5.5. The r.p. X_n converges weakly (or in law) to r.v. X, if

$$\lim_{n\to\infty} Ef(X_n) = Ef(X)$$

for any bounded and continuous function f.

The latter is equivalent to convergence of the sequence of corresponding distribution functions $F_n(x)$ to the distribution function $F(x)$ of X in all x at which $F(x)$ is continuous. Note that weak convergence makes sense even if the random variables are not defined on the same probability space!

Theorem 5.6.

$$\begin{array}{ccccccc} ''\xrightarrow{\mathbb{L}^2}'' & \Longrightarrow & ''\xrightarrow{\mathbb{L}^1}'' & \Longrightarrow & ''\xrightarrow{P}'' & \Longrightarrow & ''\xrightarrow{law}'' \\ & & ''\xrightarrow{P-a.s.}'' & \Longrightarrow & & & \end{array}$$

Proof.

1. Convergence in $\mathbb{L}^2$ implies convergence in $\mathbb{L}^1$. By Cauchy-Schwarz inequality

$$E|X_n - X| \le \sqrt{E|X_n - X|^2} \xrightarrow{n\to\infty} 0 \tag{5.1}$$

More generally convergence in $\mathbb{L}^p$ implies convergence in $\mathbb{L}^q$ for $q \le p$, which similarly to (5.1), follows from the *Lyapunov* inequality

$$E|X_n - X|^q \le \big(E|X_n - X|^p\big)^{q/p} \xrightarrow{n\to\infty} 0$$

2. Convergence in $\mathbb{L}^1$ implies convergence in probability. By Chebyshev inequality for any fixed $\varepsilon > 0$

$$P\big(|X_n - X| \ge \varepsilon\big) \le \frac{E|X_n - X|}{\varepsilon} \xrightarrow{n\to\infty} 0$$

3. P-a.s. convergence implies convergence in probability. For any $\varepsilon > 0$

$$P\big(|X_n - X| \ge \varepsilon\big) \le P\big(\sup_{m\ge n} |X_m - X| \ge \varepsilon\big) = P\big(\cup_{m\ge n}\, \{|X_m - X| \ge \varepsilon\}\big)$$

By continuity of probability measure P

$$\begin{aligned} &\overline{\lim_{n\to\infty}}\, P\big(|X_n - X| \ge \varepsilon\big) \le P\big(\lim_{n\to\infty} \cup_{m\ge n}\{|X_m - X| \ge \varepsilon\}\big) = \\ &P\big(\cap_{n\ge 0} \cup_{m\ge n}\{|X_m - X| \ge \varepsilon\}\big) = P(X_n \not\to X) = 0 \end{aligned}$$

4. Convergence in probability implies weak convergence. Fix an arbitrary continuous function $f(x)$. By definition of continuity, for any given $\varepsilon > 0$ there is a $\delta \ge 0$ such that

$$|x - y| \le \delta \implies |f(x) - f(y)| \le \varepsilon.$$

Then with $\varepsilon > 0$ fixed

$$\begin{aligned} &|Ef(X_n) - Ef(X)| \le E|f(X_n) - f(X)| = \\ &E|f(X_n) - f(X)|I(|X_n - X| > \delta) + E|f(X_n) - f(X)|I(|X_n - X| \le \delta) \le \\ &2\|f\|_\infty P(|X_n - X| > \delta) + \varepsilon \xrightarrow{n\to\infty} \varepsilon. \end{aligned}$$

Since ε is arbitrary, $\lim_{n\to\infty} |Ef(X_n) - Ef(X)| = 0$. □

Theorem 5.7. *The limits (in different senses) of X_n coincide P-a.s.*

PROOF. For example, show that if $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{P-a.s.} X'$, then $P(X \neq X') = 0$, i.e. $P(|X - X'| \geq \varepsilon) = 0$ for any $\varepsilon > 0$. Indeed

$$P(|X - X'| \geq \varepsilon) \leq P(|X - X_n| \geq \varepsilon/2) + P(|X_n - X'| \geq \varepsilon/2) \xrightarrow{n\to\infty} 0$$

since P-a.s. convergence implies convergence in probability. □

The unmentioned implications are false in general.

EXAMPLE 5.8. *($\mathbb{L}^1 \not\Rightarrow \mathbb{L}^2$)* On $([0,1], \mathcal{B}, \lambda)$ define the sequence

$$X_n(\omega) = \sqrt{n} I(\omega \in [0, 1/n])$$

X_n converges to $X \equiv 0$ in $\mathbb{L}^1$

$$EX_n = \sqrt{n}/n = n^{-1/2} \xrightarrow{n\to\infty} 0$$

but not in $\mathbb{L}^2$

$$EX_n^2 \equiv 1 \not\to 0.$$

EXAMPLE 5.9. *(weak convergence does not imply any strong convergence)* E.g. i.i.d. sequence.

EXAMPLE 5.10. *(P-a.s. convergence is not implied by convergence of any other aforementioned type)* On $([0,1], \mathcal{B}, \lambda)$ define for $n \geq 1$

$$\xi_m^k(\omega) = I\left(\omega \in \left(\frac{k-1}{m}, \frac{k}{m}\right]\right), \quad k = 1, ..., m.$$

and consider the sequence

$$X = \left(\xi_1^1, \xi_2^1, \xi_2^2, \xi_3^1, \xi_3^2, \xi_3^3, ...\right)$$

Clearly X converges to 0 in probability (why?), but not P-a.s., since for any fixed $\omega' \in (0,1]$ and for any index n, there are indices $n_1, n_2, ...$, such that $1 = X_{n_1}(\omega') = X_{n_2}(\omega') = ...$, i.e.

$$P\Big(\omega : 1 = \overline{\lim_{n\to\infty}}\, X_n(\omega) > \underline{\lim_{n\to\infty}}\, X_n(\omega) = 0\Big) = 1.$$

EXAMPLE 5.11. Let ξ_n be a sequence of i.i.d. r.v. with $P(\xi_1 = 0) = P(\xi_1 = 1) = 1/2$ and define

$$U_n = \sum_{m=1}^{n} 2^{-m} \xi_m$$

U_n is a non decreasing sequence bounded by 1 and thus converges for all ω. A fortiori it converges P-a.s, and thus also in probability and in law. With $\varepsilon > 0$ and $\widetilde{U}_n = |U_n - U|$

$$E\widetilde{U}_n = E\widetilde{U}_n I(\widetilde{U}_n > \varepsilon) + E\widetilde{U}_n I(\widetilde{U}_n \leq \varepsilon) \leq 2P(\widetilde{U}_n > \varepsilon) + \varepsilon \xrightarrow{n\to\infty} \varepsilon$$

and thus $\widetilde{U}_n$ converges in $\mathbb{L}^1$ as well by arbitrariness of ε.

Note that the limit is an non degenerate r.v. in this case and its distribution is uniform (why ?).

5.2.1. *Borel-Cantelli lemmas.* It may seem that for studying the almost sure convergence, one always needs the description of the probability space (like e.g. in the Examples 5.10, 5.11 above). Fortunately this is not always the case and often the P-a.s. convergence can be verified by means of Borel-Cantelli Lemmas, given below.

Let A_n be an infinite sequence of sets from $\mathcal{F}$. Assume that A_n decreases, i.e. $A_{n+1} \subseteq A_n$. Then the limit set

$$A = \lim_{n\to\infty} A_n \equiv \cap_{n\ge 1} A_n$$

is well defined. Analogously if $A_n \subseteq A_{n+1}$ the limit set is

$$A = \lim_{n\to\infty} A_n \equiv \cup_{n\ge 1} A_n.$$

Can the limit set be defined for a non-monotonous sequence A_n ? The answer is negative in general, just like it is negative for sequences of numbers: e.g. $x_n = (-1)^n$ has no limit. However, analogously to $\overline{\lim}_{n\to\infty} x_n = \lim_{n\to\infty} \sup_{m\ge n} x_m$ and $\underline{\lim}_{n\to\infty} x_n = \lim_{n\to\infty} \inf_{m\ge n} x_m$ (which are always well defined, though may take infinite values!), we may define

$$\overline{\lim_{n\to\infty}} A_n = \cap_{n\ge 1} \cup_{m\ge n} A_m$$

and

$$\underline{\lim_{n\to\infty}} A_n = \cup_{n\ge 1} \cap_{m\ge n} A_m.$$

Now if $\overline{\lim}_{n\to\infty} A_n$ and $\underline{\lim}_{n\to\infty} A_n$ coincide, we say that A_n has a limit and it equals e.g. $\overline{\lim}_{n\to\infty} A_n$.

Both upper and lower limits have interesting probabilistic interpretations: the set $A_{i.o} = \overline{\lim}_{n\to\infty} A_n$ consists of all the points $\omega \in \Omega$ appearing in the sequence A_n infinitely often (why?); the set $A_e = \underline{\lim}_{n\to\infty} A_n$ contains the points which appear in all the sets A_k, starting $k \ge n'$ for some n', i.e. *eventually* appear in all the sets.

Now we are ready for the first Borel-Cantelli Lemma

Lemma 5.12. *Let $(A_n)_{n\ge 1}$ be a sequence of sets, then*

$$\sum_{n=1}^{\infty} P(A_n) < \infty \qquad \Longrightarrow \qquad P(A_{i.o.}) = 0.$$

Proof.

$$P(A_{i.o}) = P\big(\cap_{n\ge 1} \cup_{m\ge n} A_m\big) \overset{\dagger}{=} \lim_{n\to\infty} P\big(\cup_{m\ge n} A_n\big) \le \lim_{n\to\infty} \sum_{m=n}^{\infty} P(A_m) = 0$$

where the equality † is due to continuity property[4] of the probability measure P. □

Corollary 5.13. *Let $(\xi_n)_{n\ge 1}$ be a sequence of r.v. If for any $\varepsilon > 0$,*

$$\sum_{n=1}^{\infty} P(|\xi_n| \ge \varepsilon) < \infty,$$

ξ_n converges to zero P-a.s.

[4] which follows from the axiomatic σ-additivity property

Proof. Note that

$$\big\{\omega : \xi_n \not\to 0\big\} = \cup_{k\ge 1} \cap_{n\ge 1} \cup_{m\ge n} \{|\xi_m| \ge 1/k\} \equiv \cup_{k\ge 1} A^{1/k}_{i.o}.$$

So

$$P(\xi_n \not\to 0) = P\big(\cup_{k\ge 1} A^{1/k}_{i.o}\big) \le \sum_{k=1}^{\infty} P(A^{1/k}_{i.o})$$

and the desired statement holds, since $P(A^{1/k}_{i.o}) = 0$ for any k. □

Corollary 5.14. *The sequence $(\xi_n)_{n\ge 0}$ converges P-a.s. to zero, if there are constants $C > 0$ and $\rho \in [0, 1)$, such that*

$$E|\xi_n|^p \le C\rho^n, \quad n \ge 1$$

for some $p \ge 1$, i.e. if $(\xi_n)_{n\ge 1}$ converges exponentially in $\mathbb{L}^p$.

Proof. By Chebyshev inequality. □

Is the opposite to Lemma 5.12 true ? I.e. does $\sum_{n=1}^{\infty} P(A_n) = \infty$ imply that $P(A_{i.o.}) = 1$? The answer is generally negative: e.g. take $A_n = I(\omega \le n^{-1})$ on the probability space $([0, 1], \mathcal{B}, \lambda)$. But

Lemma 5.15. *If $(A_n)_{n\ge 1}$ is a sequence of independent sets, then*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \qquad \Longrightarrow \qquad P(A_{i.o.}) = 1.$$

Proof. By the well known set operations rules

$$\overline{A}_{i.o} = \cup_{n\ge 1} \overline{\cup_{m\ge n} A_m} = \cup_{n\ge 1} \cap_{m\ge n} \overline{A}_m$$

and so it suffices to verify that for any fixed n,

$$P\big(\cap_{m\ge n} \overline{A}_m\big) = 0.$$

By independence

$$P\big(\cap_{m\ge n} \overline{A}_m\big) = \prod_{m=n}^{\infty} P(\overline{A}_m) = \prod_{m=n}^{\infty} \big(1 - P(A_m)\big) =$$

$$\exp\Big\{\sum_{m=n}^{\infty} \log\big(1 - P(A_m)\big)\Big\} \le \exp\Big\{-\sum_{m=n}^{\infty} P(A_m)\Big\} = 0$$

where the inequality $\log(1 - x) \le -x$, $x \in [0, 1)$ had been used. □

Example 5.16. *(convergence in probability does not imply P-a.s. convergence)* Let $(\xi_n)_{n\ge 1}$ be a sequence of independent r.v. such that $P(\xi_n = 1) = p_n$ and $P(\xi_n = 0) = 1 - p_n$.

Let $A_n = \{\xi_n = 1\}$, then by the second Borel-Cantelli lemma $\sum_{n=1}^{\infty} p_n = \infty$ (e.g. $p_n = 1/n$) implies $P(A_{i.o}) = 1$, i.e. ξ_n does not converge to zero P-a.s. However it converges to zero in probability.

5.2.2. *Cauchy criteria for convergence.* The sequence of numbers x_n is said to be *fundamental (or Cauchy)* if

$$\sup_{\ell,m\ge n} |x_\ell - x_m| \xrightarrow{n\to\infty} 0.$$

A numerical sequence x_n converges if and only if it is fundamental. The same holds when x_n are vectors in $\mathbb{R}^n$. In an infinite dimensional space, this is not necessarily true: while any convergent sequence is fundamental, the other implication does not hold in general. If it does the space is called *complete.* Completeness is an important property, since it can be used to construct limit objects (as in e.g. Theorem 1.6).

In the next chapter we will extensively use the space of square integrable random variables (i.e. with finite expectation $EX^2 < \infty$), denoted by $\mathbb{L}^2(\Omega, \mathcal{F}, P)$ (or shortly $\mathbb{L}^2$). Clearly $\mathbb{L}^2$ is a linear space ($X \in \mathbb{L}^2$, $Y \in \mathbb{L}^2 \implies \alpha X + \beta Y \in \mathbb{L}^2$ for $\alpha, \beta \in \mathbb{R}$). Endowed with the scalar product [5] $\langle X, Y\rangle = EXY$ it becomes an Euclidian space with the induced norm $\|X\| = \sqrt{\langle X, X\rangle} = \sqrt{EX^2}$. Usually $\mathbb{L}^2$ is *infinite* dimensional, i.e. there is no finite number of elements of $\mathbb{L}^2$, such that any other element is their linear combination or in other words, it has no finite basis. Let us verify completeness of $\mathbb{L}^2$, which will turn it to the *Hilbert* space with all its well developed machinery.

Theorem 5.17. *A sequence of random variables ξ_n from $\mathbb{L}^2$ converges to a random variable ξ in $\mathbb{L}^2$ if and only if it is fundamental (Cauchy) in $\mathbb{L}^2$, i.e.*

$$\lim_{n\to\infty} \sup_{\ell,m\ge n} E(\xi_n - \xi)^2 = 0.$$

Proof. Clearly if $\xi_n \xrightarrow[n\to\infty]{\mathbb{L}^2} \xi$, then it is fundamental:

$$E(\xi_n - \xi_m)^2 \le 2E(\xi_n - \xi)^2 + 2E(\xi_m - \xi)^2 \xrightarrow{n,m\to\infty} 0.$$

The proof of sufficiency is more involved. Suppose that ξ_n is fundamental. Then there is an n_1, such that

$$\|\xi_n - \xi_m\| = \sqrt{E(\xi_n - \xi_m)^2} \le \frac{1}{2^2}, \quad \forall m, n \ge n_1.$$

Similarly there are indices n_k, $k \ge 1$ such that $n_k \ge n_{k-1}$ and

$$\sqrt{E(\xi_n - \xi_m)^2} \le 2^{-2k}, \quad \forall m, n \ge n_k.$$

Let $A_k = \{\omega : |\xi_{n_{k+1}} - \xi_{n_k}| \ge 2^{-k}\}$ then

$$P(A_k) \le 2^k E|\xi_{n_{k+1}} - \xi_{n_k}| \le 2^k \sqrt{E(\xi_{n_{k+1}} - \xi_{n_k})^2} \le 2^{-k}$$

So $\sum_{k=1}^\infty P(A_k) < \infty$ and by the first Borel-Cantelli lemma $P(A_k, i.o.) = 0$, i.e.

$$P\big(|\xi_{n_{k+1}} - \xi_{n_k}| \ge 2^{-k}, i.o.\big) = 0$$

which implies $\sum_{k=1}^\infty |\xi_{n_{k+1}} - \xi_{n_k}| < \infty$ and thus $\xi = \lim_{k\to\infty} \xi_{n_k} = \sum_{k=1}^\infty \big(\xi_{n_{k+1}} - \xi_{n_k}\big)$ exists. Now it is left to show that this ξ is the required limit of ξ_n, that is

[5] Recall that scalar product $\langle x, y\rangle$ is a function of elements pairs from the linear space under consideration (i.e. $\mathbb{L}^2$ here) to $\mathbb{R}$, such that (1) $\langle \alpha X + \beta Y, Z\rangle = \alpha\langle X, Z\rangle + \beta\langle Y, Z\rangle$ for $\alpha, \beta \in \mathbb{R}$; (2) $\langle X, X\rangle \ge 0$; and (3) $\langle X, X\rangle = 0 \implies X = 0$ (in the case $\mathbb{L}$ the latter is understood P-a.s.)

$E(\xi_n - \xi)^2 \to 0$ and that $E\xi^2 < \infty$, i.e. $\xi \in \mathbb{L}^2$. Let us fix $\varepsilon > 0$ and choose N_ε such that $E(\xi_m - \xi_n)^2 \le \varepsilon$ for all $n, m \ge N_\varepsilon$. Then with $n \ge N_\varepsilon$

$$E(\xi_n - \xi)^2 = E(\xi_n - \lim_{k\to\infty} \xi_{n_k})^2 = E \lim_{k\to\infty} (\xi_n - \xi_{n_k})^2 = \\ E \varliminf_{k\to\infty} (\xi_n - \xi_{n_k})^2 \le \varliminf_{k\to\infty} E(\xi_n - \xi_{n_k})^2 \le \varepsilon$$

and hence [6] by arbitrariness of ε, $\lim_{n\to\infty} E(\xi_n - \xi)^2 = 0$. Since $E\xi^2 \le 2E(\xi - \xi_n)^2 + 2E\xi_n^2$, $E\xi^2 < \infty$ as well. □

6. Limit theorems

Limit theorems deal with convergence of sums of r.v. and are one of the central issues of probability theory and stochastic processes. Below we give the simplest versions of classic limit theorems, which originated more than several centuries ago.

6.1. The Weak Law of Large Numbers.

THEOREM 6.1. *Let ξ_n be a sequence of orthogonal random variables with $m = E\xi_n$ and $V = \mathrm{var}(\xi_n) = E(\xi_n - m)^2 < \infty$. Let $S_n = \sum_{k=1}^n \xi_k$, then*

$$\frac{1}{n} S_n \xrightarrow[n\to\infty]{\mathbb{L}^2} m.$$

PROOF.

$$E\Big(\frac{1}{n}S_n - m\Big)^2 = \frac{1}{n^2} E\Big(\sum_{k=1}^n (\xi_k - m)\Big)^2 = \frac{1}{n^2}\sum_{k=1}^n V = V/n \to 0.$$

□

As was already mentioned before the WLLN holds under much more general conditions.

6.2. The Strong Law of Large Numbers. By the strong LLN P-a.s. convergence of the empirical mean is usually meant.

THEOREM 6.2. *(Cantelli) Let ξ_n be a sequence of independent r.v. with finite fourth moment, such that*

$$E|\xi_n - E\xi_n|^4 \le C, \quad n \ge 1$$

for some $C > 0$. Then

$$\lim_{n\to\infty} \frac{S_n - ES_n}{n} = 0, \quad P - a.s.$$

PROOF. Without loss of generality we may consider $E\xi_n = 0$. By the first Borel-Cantelli lemma $S_n/n \to 0$ P-a.s. is implied by

$$\sum P\Big(|S_n/n| \ge \varepsilon\Big) < \infty$$

for any $\varepsilon > 0$. By Chebyshev inequality it is sufficient that

$$\sum E|S_n/n|^4 < \infty.$$

[6] The inequality $E \varliminf_{n\to\infty} X_n \le \varliminf_{n\to\infty} EX_n$ is called the Fatou Lemma. The proof can be found in e.g. (8)

Let us verify the latter condition. First note

$$S_n^4 = \sum_{i=1}^{n} \xi_i^4 + \sum_{i<j} \frac{4!}{2!2!}\xi_i^2\xi_j^2 + \sum_{i\neq j, i\neq k, j<k} \frac{4!}{2!1!1!}\xi_i^2\xi_j\xi_k$$
$$+ \sum_{i<j<k<\ell} 4!\xi_i\xi_j\xi_k\xi_\ell + \sum_{i\neq j}\frac{4!}{3!1!}\xi_i^3\xi_j$$

Since $E\xi_k = 0$

$$ES_n^4 = \sum_{i=1}^{n} E\xi_i^4 + 6\sum_{i<j} E\xi_i^2 E\xi_j^2 \le nC + 6\sum_{i<j}\sqrt{E\xi_i^4 E\xi_j^4} \le$$
$$nC + 6\frac{n(n-1)}{2}C = (3n^2-2n)C \le 3n^2C$$

and thus

$$\sum E(S_n^4/n^4) \le 3C\sum n^{-2} < \infty.$$

□

6.3. Central Limit Theorem. The laws of large numbers state that $S_n = \sum_{i=1}^n \xi_i$ converges "strongly" (e.g. to the expectation of ξ_1), if it is scaled by n^{-1}. What happens if $\sum_{i=1}^n \xi_i$ is examined on a finer scale, say $1/\sqrt{n}$? The Central Limit Theorem states that it converges weakly to a Gaussian r.v. regardless (!) of the precise form of ξ_n distribution.

A useful tool in the proof of weak convergence is

Theorem 6.3. *Let F_n be a sequence of distribution functions on $\mathbb{R}$ and $\varphi_n(t)$ is the corresponding sequence of characteristic functions. Then*

$$F_n(x) \xrightarrow{w} F(x) \quad \Leftrightarrow \quad \varphi_n(t) \to \varphi(t), \forall t \in \mathbb{R}.$$

Theorem 6.4. *Let ξ_n be an i.i.d. sequence with $0 < \mathrm{var}(\xi_1) < \infty$. Then*

$$\frac{S_n - ES_n}{\sqrt{\mathrm{var}(S_n)}} \xrightarrow[n\to\infty]{w} \xi$$

where ξ is a standard Gaussian r.v.

Proof. (Sketch) Denote $m = E\xi_1$ and $\sigma^2 = \mathrm{var}(\xi_1)$ and

$$\varphi(t) = Ee^{it(\xi_1 - m)}.$$

Then

$$\varphi_n(t) \equiv E\exp\Big\{it\frac{S_n - ES_n}{\sqrt{\mathrm{var}(S_n)}}\Big\} = E\exp\Big\{it\frac{\sum^n(\xi_k - m)}{\sqrt{n}\sigma}\Big\} =$$
$$= \prod^n E\exp\Big\{it(\xi_1 - m)/(\sqrt{n}\sigma)\Big\} = \Big[\varphi\Big(\frac{t}{\sqrt{n}\sigma}\Big)\Big]^n.$$

It can be shown (see II.12.14 in (8)) that $\mathrm{var}(\xi) < \infty$ implies

$$\varphi(t) = 1 - \sigma^2t^2/2 + o(t^2), \quad t \to 0.$$

Then

$$\varphi_n(t) = \Big[1 - \frac{\sigma^2t^2}{2n\sigma^2} + o(1/n)\Big]^n \xrightarrow{n\to\infty} e^{-t^2/2}.$$

Theorem 6.3 then implies the statement. □

6.4. The Law of Small Numbers.

Theorem 6.5. *(Poisson) Let for each $n \geq 1$ the i.i.d. r.v. $\xi_n^1, ..., \xi_n^n$ are such that*

$$P(\xi_n^k = 1) = p_n, \quad P(\xi_n^k = 0) = 1 - p_n, \quad 1 \leq k \leq n$$

where $p_n \to 0$ as $n \to \infty$, so that $\lim_{n\to\infty} np_n = \lambda > 0$. Then

$$P(S_n = m) \to \frac{e^{-\lambda}\lambda^m}{m!}$$

for any $m \geq 0$.

Proof. Since $Ee^{it\xi_n^k} = p_n e^{it} + 1 - p_n$, then

$$\varphi_{S_n}(t) = Ee^{itS_n} = \big(1 + p_n(e^{it} - 1)\big)^n = \\ \big(1 + \frac{\lambda}{n}(e^{it} - 1) + o(1/n)\big)^n \to \exp\big\{\lambda(e^{it} - 1)\big\}$$

The claim holds by Theorem 6.3, since $\exp\big\{\lambda(e^{it}-1)\big\}$ is the characteristic function of the Poisson distribution. □

CHAPTER 2

Orthogonal projection and linear estimation

1. Orthogonal projection

1.1. Simple scalar case. Consider a pair of random variables (X,Y) from the Hilbert space $\mathbb{L}^2(\Omega,\mathcal{F},P)$ (see section 5.2.2 for a discussion). Suppose we would like to find the best approximation of X by means of Y and a constant, i.e. to find $\widehat{X}=a_0+a_1Y$ minimizing the *mean square error* $E(X-\widehat{X})^2$. Simple calculations give the desired answer

$$\begin{aligned}&E(X-a_0-a_1Y)^2=E\big(X-EX-a_0+EX-a_1EY-a_1(Y-EY)\big)^2=\\&E\big(X-EX-a_1(Y-EY)\big)^2+E\big(EX-a_0-a_1EY\big)^2\ge\\&E\big(X-EX-a_1(Y-EY)\big)^2=\operatorname{cov}(X)-2a_1\operatorname{cov}(X,Y)+a_1^2\operatorname{cov}(Y)\ge\\&\operatorname{cov}(X)-\operatorname{cov}^2(X,Y)/\operatorname{cov}(Y)\end{aligned}$$

where the equalities are attained at $a_1=\operatorname{cov}(X,Y)/\operatorname{cov}(Y)$ and $a_0=EX-a_1EY$.

In other words

$$\widehat{X}=EX+\operatorname{cov}(X,Y)/\operatorname{cov}(Y)\big(Y-EY\big)$$

gives the best estimate of X from the vector $(1,Y)$ with the estimation error

$$E(X-\widehat{X})^2=\operatorname{cov}(X)-\operatorname{cov}^2(X,Y)/\operatorname{cov}(Y).$$

What happens when $\operatorname{cov}(Y)=0$? In this case $Y=EY$ P-a.s. and so

$$\begin{aligned}E(X-a_0-a_1Y)^2=E(X-a_0-a_1EY)^2=&\\&E(X-EX+EX-a_0-a_1EY)^2\ge\operatorname{cov}(X)\end{aligned}$$

where the equality is attained e.g. $a_0=EX$ and $a_1=0$. So the general formulae

$$\begin{aligned}&\widehat{X}=EX+\operatorname{cov}(X,Y)\operatorname{cov}^{\oplus}(Y)\big(Y-EY\big)\\&E(X-\widehat{X})^2=\operatorname{cov}(X)-\operatorname{cov}^2(X,Y)\operatorname{cov}^{\oplus}(Y)\end{aligned}$$

hold where

$$\operatorname{cov}^{\oplus}(Y)=\begin{cases}\operatorname{cov}^{-1}(Y), & \operatorname{cov}(Y)>0\\ 0, & \operatorname{cov}(Y)=0.\end{cases}$$

The random variable $\widehat{X}$ is called *the orthogonal projection* of X on the linear subspace $\mathcal{M}\subseteq\mathbb{L}^2$ spanned by r.v. 1 and Y, denoted also as $\widehat{E}(X|\mathcal{M})$ and sometimes referred as the conditional expectation in the wide sense of X with respect to $\mathcal{M}$.

Indeed $X=\widehat{X}+(X-\widehat{X})$, where $\widehat{X}$ belongs to $\mathcal{M}$ and $X-\widehat{X}$ is *orthogonal* to $\mathcal{M}$, i.e. for any $Z\in\mathcal{M}$.

$$E(X-\widehat{X})Z=0. \tag{1.1}$$

To see this consider $E(X-\widehat{X}-tZ)^2$, where t is a constant and $Z\in\mathcal{M}$. By optimality

$$E(X-\widehat{X}-tZ)^2 \ge E(X-\widehat{X})^2$$

i.e.

$$-2tE(X-\widehat{X})Z+t^2EZ^2\ge 0.$$

Choose $t=\alpha E(X-\widehat{X})Z$, then

$$\big[E(X-\widehat{X})Z\big]^2\big(-2\alpha+\alpha^2EZ^2\big)\ge 0$$

The latter can hold for small enough α only if (1.1) holds.

The other direction is also true: if $\widetilde{X}\in\mathcal{M}$ and

$$E(X-\widetilde{X})Z=0$$

for all $Z\in\mathcal{M}$, then $\widehat{X}=\widetilde{X}$, P-a.s.

Indeed

$$E(X-Z)^2=E(X-\widetilde{X}+\widetilde{X}-Z)^2=E(X-\widetilde{X})^2+E(\widetilde{X}-Z)^2\ge E(X-\widetilde{X})^2,$$

for any Z. In particular with $Z:=\widehat{X}$, we have $E(\widehat{X}-\widetilde{X})^2=0$ or $\widehat{X}=\widetilde{X}$, P-a.s.

1.2. Vector case. Now let us extend this result to the vector case. Let X be an $\mathbb{L}^2$ r.v. and Y be random vector in $\mathbb{R}^n$ with entries in $\mathbb{L}^2$. Let $\mathcal{M}$ be the linear subspace generated by $1,Y_1,...,Y_n$. We are interested in the optimal estimate $\widehat{X}=a_0+\sum_{i=1}^n a_iY_i$, such that

$$E(X-\widehat{X})^2\le E(X-Z)^2$$

for all $Z\in\mathcal{M}$.

By the very same arguments as in the scalar case, we get

Lemma 1.1. *The estimate $\widehat{X}$ is optimal if and only if $E(X-\widehat{X})Z=0$ for all $Z\in\mathcal{M}$.*

Now let a be the column vector with entries $a_1,...,a_n$. By the above Lemma the optimal estimate is found from

$$E(X-a_0-a^*Y)\cdot 1=0$$
$$E(X-a_0-a^*Y)Y^*=0$$

The first constraint implies $a_0=EX-a^*EY$, so that the second one is rewritten

$$E\big((X-EX)-a^*(Y-EY)\big)(Y-EY)^*=0$$

or in other words

$$\operatorname{cov}(X,Y)-a^*\operatorname{cov}(Y)=0. \tag{1.2}$$

If $\operatorname{cov}(Y)>0$ (positive definite matrix), then

$$a^*=\operatorname{cov}(X,Y)\operatorname{cov}^{-1}(Y)$$

and thus

$$\widehat{X}=EX+\operatorname{cov}(X,Y)\operatorname{cov}^{-1}(Y)(Y-EY).$$

But what if $\operatorname{cov}(Y)$ is singular ?

1.2.1. *Some facts from linear algebra.*

Lemma 1.2. *Any symmetric matrix S ($S = S^*$) is decomposable as $S = U\Lambda U^*$ where U is a real orthogonal matrix ($U^*U = I$) and Λ is a real diagonal matrix.*

Proof. (partial) First show that any eigenvalue of S is real. Let λ be an eigenvalue of S and φ corresponding right eigenvector, i.e. $S\varphi = \lambda\varphi$. Multiply this equation by the conjugate-transposed φ', so that $\varphi' S\varphi = \lambda\|\varphi\|^2$. λ is real since both $\|\varphi\|^2$ and $\varphi' S\varphi$ are real. Indeed, let $\alpha = \varphi' S\varphi$, then $\overline{\alpha} = \varphi^* S\overline{\varphi} = \varphi' S^*\varphi = \varphi' S\varphi = \alpha$. If λ is real, then φ is real as well, being the solution of $(S - \lambda I)\varphi = 0$.

Let us show now that the eigenvectors corresponding to distinct eigenvalues are orthogonal:

$$S\varphi_1 = \lambda_1\varphi_1 \implies \varphi_2^* S\varphi_1 = \lambda_1\varphi_2^*\varphi_1 \implies \lambda_2\varphi_2^*\varphi_1 = \lambda_1\varphi_2^*\varphi_1 \implies \varphi_2^*\varphi_1 = 0.$$

□

Exercise 1.3. *Give a proof without assuming that the eigenvalues are distinct.*

If the diagonal matrix Λ has positive (nonnegative) entries, S is said to be positive (nonnegative) definite matrix, since for any vector v

$$v^* S v = v^* U\Lambda U^* v = \big\|Uv\big\|_\Lambda^2 > 0.$$

Definition 1.4. Let S be a symmetric matrix with $S = U\Lambda U^*$. The pseudo-inverse of S in the sense of Moore-Penrose is $S^\oplus = U\Lambda^\oplus U^*$ where $\Lambda^\oplus$ is the diagonal matrix with entries $\lambda_i^{-1} I(\lambda_i \neq 0)$.

1.2.2. *Back to the estimation problem.* Analogously to the scalar case set

$$\widehat{X} = EX + \mathrm{cov}(X,Y)\,\mathrm{cov}^\oplus(Y)(Y - EY).$$

Let us verify orthogonality i.e. (1.2). First note that if φ is an eigenvector of $\mathrm{cov}(Y)$ corresponding to the zero eigenvalue, then

$$0 = \varphi^* \,\mathrm{cov}(Y)\varphi = E\big((Y - EY)^*\varphi\big)^2 \implies (Y - EY)^*\varphi = 0, P - a.s.$$

and thus

$$\mathrm{cov}(X,Y)\varphi = E(X - EX)(Y - EY)^*\varphi = 0. \tag{1.3}$$

So

$$\mathrm{cov}(X,Y) - \mathrm{cov}(X,Y)\,\mathrm{cov}^\oplus(Y)\,\mathrm{cov}(Y) = \mathrm{cov}(X,Y)U\big(I - \Lambda^\oplus\Lambda\big)U^* = 0$$

where the latter equality is due to (1.3) and the Definition 1.4. Note that orthogonality would be preserved if in Definition 1.4, the diagonal entries of $\Lambda^\oplus$ are defined $\lambda_i^{-1} I(\lambda_i > 0) + \alpha_i I(\lambda_i = 0)$ with any constants α_i and $\alpha_i = 0$ is a convenient choice.

The following theorem summarizes the results obtained above:

Theorem 1.5. *Let X and Y be random vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ with square integrable entries. The orthogonal projection of X on the linear subspace $\mathcal{M} = \overline{\mathrm{span}}\{1, Y_1, ..., Y_n\}$ is given by*

$$\widehat{E}(X|Y) = EX + \mathrm{cov}(X,Y)\,\mathrm{cov}^\oplus(Y)(Y - EY).$$

Moreover for any $Z \in \mathcal{M}$ $E\big[X - \widehat{E}(X|Y)\big]Z = 0$ and

$$E\big(X - \widehat{E}(X|Y)\big)\big(X - \widehat{E}(X|Y)\big)^* = \mathrm{cov}(X) - \mathrm{cov}(X,Y)\,\mathrm{cov}^\oplus(Y)\,\mathrm{cov}^*(X,Y).$$

1.3. Infinite dimensional case. Suppose we are given a random variable X and a sequence $Y_1, Y_2, ...$ with $X, Y_i \in \mathbb{L}^2$. Let $\mathcal{M}$ be the closed linear subspace generated by $1, Y_1, Y_2, ...$, i.e. all linear combinations of $1, Y_1, Y_2, ...$ and all their mean square limits:

$$\mathcal{M} = \overline{\text{span}}\{1, Y_1, Y_2, ...\}.$$

The formulae of the previous section does not make much sense in this infinite dimensional case. However the orthogonal projection is still well defined:

Theorem 1.6. *There is a unique (P-a.s.) random variable $\widehat{X} := \widehat{E}(X|Y)$, such that*

$$E(X - \widehat{X})^2 = \inf_{Z \in \mathcal{M}} E(X - Z)^2$$

and $E\big(X - \widehat{X}\big)Z = 0$ for any $Z \in \mathcal{M}$.

Proof. Denote $d^2 = \inf_{Z \in \mathcal{M}} E(X - Z)^2$ and let $Z_1, Z_2, ...$ be the sequence such that $\lim_{n\to\infty} E(X - Z_n)^2 = d^2$. Note that

$$E(Z_n - Z_m)^2 = 2E(Z_n - X)^2 + 2E(Z_m - X)^2 - 4E\Big(\frac{Z_n + Z_m}{2} - X\Big)^2.$$

Since $Z_n + Z_m \in \mathcal{M}$, $E\big((Z_n + Z_m)/2 - X\big)^2 \ge d^2$ and so

$$E(Z_n - Z_m)^2 \le 2E(Z_n - X)^2 + 2E(Z_m - X)^2 - 4d^2 \xrightarrow{n\to\infty} 0$$

i.e. Z_n is a fundamental sequence in $\mathbb{L}^2$. The space $\mathbb{L}^2$ is complete, i.e. any fundamental sequence converges to an element in $\mathbb{L}^2$, which is nothing but $\widehat{X}$ P-a.s.

Let us verify the P-a.s. uniqueness: suppose that there is a r.v. $\widetilde{X}$ such that

$$E(X - \widetilde{X})^2 = E(X - \widehat{X})^2 = d^2.$$

Then

$$E(\widehat{X} + \widetilde{X} - 2X)^2 + E(\widehat{X} - \widetilde{X})^2 = 2E(\widetilde{X} - X)^2 + 2E(\widehat{X} - X)^2 = 4d^2.$$

But $E(\widehat{X} + \widetilde{X} - 2X)^2 = 4E\big((\widehat{X} + \widetilde{X})/2 - X\big)^2 \ge 4d^2$ and so $E(\widehat{X} - \widetilde{X})^2 = 0$, i.e. $P(\widehat{X} = \widetilde{X}) = 1$.

Now let us verify orthogonality: since for any $t \in \mathbb{R}$ and $Z \in \mathcal{M}$

$$E(X - \widehat{X} - tZ)^2 \ge E(X - \widehat{X})^2$$

we have

$$t^2 EZ^2 - 2tE(X - \widehat{X})Z \ge 0.$$

Take $t = \alpha E(X - \widehat{X})Z$, $\alpha \in \mathbb{R}$, then

$$\big(E(X - \widehat{X})Z\big)^2\big[\alpha^2 EZ^2 - 2\alpha\big] \ge 0.$$

For sufficiently small $\alpha > 0$, $\big[\alpha^2 EZ^2 - 2\alpha\big] < 0$ and thus $E(X - \widehat{X})Z = 0$. □

1.4. Properties of orthogonal projection. The orthogonal projection satisfies the following properties. Below $\mathcal{M}$ denotes[1] some linear subspace of $\mathbb{L}^2$
1. $E\widehat{E}(X|\mathcal{M}) = EX$

2. $\widehat{E}(X|\mathcal{M}) = \begin{cases} X & X \in \mathcal{M} \\ 0 & X \perp \mathcal{M} \end{cases}$

3. $\widehat{E}(c_1X_1 + c_2X_2|\mathcal{M}) = c_1\widehat{E}(X_1|\mathcal{M}) + c_2\widehat{E}(X_2|\mathcal{M})$

4. If $\mathcal{M}_1 \subseteq \mathcal{M}_2$, then $\widehat{E}\big(\widehat{E}(X|\mathcal{M}_2)|\mathcal{M}_1\big) = \widehat{E}(X|\mathcal{M}_1)$.

5. If $\mathcal{M}_1$ and $\mathcal{M}_2$ are orthogonal subspaces, then

$$\widehat{E}(X|\mathcal{M}_1 \oplus \mathcal{M}_2) = \widehat{E}(X|\mathcal{M}_1) + \widehat{E}(X|\mathcal{M}_2)$$

Proof.
(1) By orthogonality $E(X - \widehat{E}(X|\mathcal{M})) \cdot 1 = 0$
(2) If $X \in \mathcal{M}$, it is the orthogonal projection by definition. If $X \perp \mathcal{M}$, i.e. $EXZ = 0$ for any r.v. $Z \in \mathcal{M}$, then again by definition $\widehat{E}(X|\mathcal{M}) = 0$.
(3) For any r.v. $Z \in \mathcal{M}$

$$E(c_1X_1 + c_2X_2 - c_1\widehat{E}(X_1|\mathcal{M}) - c_2\widehat{E}(X_2|\mathcal{M}))Z = \\ c_1E(X_1 - \widehat{E}(X_1|\mathcal{M}))Z + c_2E(X_2 - \widehat{E}(X_2|\mathcal{M}))Z = 0.$$

(4) For any $Z \in \mathcal{M}_1$

$$0 = E\big[\widehat{E}(X|\mathcal{M}_2) - \widehat{E}\big(\widehat{E}(X|\mathcal{M}_2)|\mathcal{M}_1\big)\big]Z = \\ E\big[\widehat{E}(X|\mathcal{M}_2) - X\big]Z + E\big[X - \widehat{E}\big(\widehat{E}(X|\mathcal{M}_2)|\mathcal{M}_1\big)\big]Z \\ = E\big[X - \widehat{E}\big(\widehat{E}(X|\mathcal{M}_2)|\mathcal{M}_1\big)\big]Z$$

where the latter equality holds since $Z \in \mathcal{M}_1 \implies Z \in \mathcal{M}_2$.
(5) Any $Z \in \mathcal{M}_1 \oplus \mathcal{M}_2$ can be decomposed into $Z = Z_1 + Z_2$ with $Z_1 \in \mathcal{M}_1$ and $Z_2 \in \mathcal{M}_2$. Then

$$E(X - \widehat{E}(X|\mathcal{M}_1) - \widehat{E}(X|\mathcal{M}_2))(Z_1 + Z_2) = \\ E(X - \widehat{E}(X|\mathcal{M}_1))Z_1 + E(X - \widehat{E}(X|\mathcal{M}_2))Z_2 - \\ E\widehat{E}(X|\mathcal{M}_2)Z_1 - E\widehat{E}(X|\mathcal{M}_1)Z_2 = 0$$

where the latter two terms vanish due to $\mathcal{M}_1 \perp \mathcal{M}_2$. □

2. Linear estimation of stationary processes: Kolmogorov-Wiener approach

In many engineering applications it is required to estimate signals from the noisy observations. Within the probabilistic framework both the signal and the observation are assumed to be random processes. Typically the signal X_n is to be estimated from some segment of trajectory of Y, i.e. a functional $\psi_n(Y)$ is to be found to minimize the mean square error criterion. Specifically the following estimation problems are frequently encountered in applications

- *Filtering:* estimate X_n from $Y_0^n := \{Y_0, Y_1, ..., Y_n\}$ for each $n \geq 1$;

[1] $1 \in \mathcal{M}$ is always assumed

- *Prediction:* estimate X_{n+m} with $m > 0$ from $Y_0^n := \{Y_0, Y_1, ..., Y_n\}$ for each $n \geq 1$;
- *Smoothing:* estimate X_{n-m} with $m > 0$ from $Y_0^n := \{Y_0, Y_1, ..., Y_n\}$ for each $n \geq 1$;

In this chapter the linear estimation problems are considered, i.e. the optimal estimator $\psi_n(\cdot)$ is constrained to be a linear functional.

2.1. Stationary processes. The Kolmogorov-Wiener theory deals with estimation of stationary processes.

Definition 2.1. The random process $X = (X_n)_{n\in\mathbb{Z}}$ is stationary if all its finite dimensional distributions are shift independent, i.e.

$$P(X_{n_1} \leq x_1, ..., X_{n_d} \leq x_d) = P(X_{n_1+h} \leq x_1, ..., X_{n_d+h} \leq x_d)$$

for any $h \in \mathbb{Z}$ and any set of indices $n_1, ..., n_d$ and any numbers x_1,...,x_d.

For example the sequence of i.i.d. r.v.'s is a stationary process.

Definition 2.2. The random process $X = (X_n)_{n\in\mathbb{Z}}$ is stationary in the wide sense if its mean sequence is constant and its correlation sequence depends only on the shift, i.e. $EX_n \equiv EX_m$ and $\text{cov}(X_n, X_m) \equiv R(n-m)$ for all n, m.

Clearly if X is a stationary process and $E|X_n|^2 < \infty$, it is also stationary in the wide sense. Hereafter we will abuse the notations, referring wide sense stationary processes as stationary, assuming w.l.o.g. that $EX_n \equiv 0$.

The correlation sequence of a stationary process satisfies a number of important properties

Lemma 2.3. *Let $R(k) = EX_nX_{n+k}$. Then*

(1) *$R(k)$ is a nonnegative definite function, i.e. for any*[2] *complex sequence $z_n \in \mathbb{C}$*

$$\sum_{k,\ell} z_k R(k-\ell)\bar{z}_\ell \geq 0.$$

(2) *$R(k)$ is an even function with maximum at $k = 0$*

Proof.

1.

$$\sum_{k,\ell} z_k R(k-\ell)\bar{z}_\ell = E\big|\sum_{n=1} z_n X_n\big|^2 \geq 0$$

2. $R(k) = EX_nX_{n+k} = R(-k)$. $R(k) = EX_nX_{n+k} \leq \sqrt{EX_n^2 EX_{n+k}^2} = R(0)$. □

If $\sum_k |R(k)| < \infty$, the Fourier transform of $R(k)$

$$f(\lambda) = \sum_k R(k)e^{-i\lambda k}$$

is called spectral density of X. The inverse formula is

$$R(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda k} f(\lambda) d\lambda.$$

For example the spectral density of an i.i.d. sequence (of square integrable r.v.) is constant, which is the reason such sequence is called discrete time *white noise*.

[2] such that the summation makes sense

If $R(k)$ is not summable the spectral density may not exist but nevertheless the spectral distribution (measure) $F(\lambda)$ is well defined so that

$$R(k) = \int_{-\pi}^{\pi} e^{i\lambda k} dF(\lambda).$$

2.2. Smoothing. Consider a pair of (jointly) stationary random processes $(X, Y) = (X_n, Y_n)_{n\in\mathbb{Z}}$, where the signal at time n is to be estimated from the trajectory of $\{Y_n, -\infty \le n \le \infty\}$.

Suppose that the linear estimate of the form

$$\widehat{X}_n = \sum_{k=-\infty}^{\infty} a_k Y_{n-k}$$

is to be constructed to minimize the mean square error, i.e.

$$E(X_n - \widehat{X}_n)^2 \le E(X_n - Z)^2$$

for any $Z \in \mathcal{M} = \overline{\text{span}}\{..., Y_{-1}, Y_0, Y_1, ...\}$. Such an estimate exists (see Theorem 1.6) and is nothing but the orthogonal projection of X on $\mathcal{M}$.

The coefficients of the *optimal linear smoother* a_n can be found from the orthogonality relation

$$E\big(X_n - \sum_k a_k Y_{n-k}\big)Y_\ell = 0, \quad \forall \ell$$

or, assuming stationary processes,

$$R_{XY}(m) = \sum_{k=-\infty}^{\infty} a_k R_Y(m-k), \quad \forall m := n - \ell \tag{2.1}$$

The latter is known as Wiener-Hopf equations and can be solved explicitly in terms of spectral densities, which are assumed to exist. The Fourier transform of the left hand side is $S_{XY}(\lambda)$, while taking the Fourier transform of the right hand side gives[3]

$$\sum_{k=-\infty}^{\infty} a_k \sum_{m=-\infty}^{\infty} R_Y(m-k)e^{-im\lambda} =$$

$$\sum_{k=-\infty}^{\infty} a_k e^{-ik\lambda} \sum_{m'=-\infty}^{\infty} R_Y(m')e^{-im'\lambda} = A(\lambda)S_Y(\lambda).$$

So the coefficients of the optimal smoother can be found by taking the inverse Fourier transform of

$$A(\lambda) = \frac{S_{XY}(\lambda)}{S_Y(\lambda)}$$

where $S_Y(\lambda) > 0$ is assumed. The coefficients of the optimal estimator are obtained via inverse Fourier transform.

[3] and assuming that all the correlations decay fast enough so that the summations can be interchanged

The corresponding minimal mean square error is given by

$$E(X_n - \widehat{X}_n)^2 = E(X_n - \widehat{X}_n)X_n = R_X(0) - \sum_{k=-\infty}^{\infty} a_k R_{XY}(k) =$$
$$\frac{1}{2\pi}\int S_X(\lambda)d\lambda - \frac{1}{2\pi}\int S_{XY}(\lambda)\sum_{k=-\infty}^{\infty} a_k e^{ik\lambda}d\lambda = \tag{2.2}$$
$$\frac{1}{2\pi}\int \left[S_X(\lambda) - \frac{\left|S_{XY}(\lambda)\right|^2}{S_Y(\lambda)}\right]d\lambda.$$

Example 2.4. Suppose that X is a stationary process satisfying the recursion

$$X_n = aX_{n-1} + \varepsilon_n, \quad \forall n$$

where $a \in [0,1)$ is a constant and ε is a standard sequence of i.i.d. r.v. The observation process is $Y_n = X_n + \sigma\xi_n$ with $\sigma > 0$ and ξ_n another standard i.i.d. sequence, independent of X.

The variance $V = EX_n^2$ satisfies the equation $V = a^2V + 1$ and so $V = 1/(1-a^2)$, while the correlation satisfies

$$R_X(k) = aR_X(k-1), \quad k \geq 1.$$

So $R_X(k) = a^{|k|}/(1-a^2)$ for all k. The spectral density is then given by

$$S_{XY}(\lambda) = S_X(\lambda) = 1/(1-a^2)\sum_{k=-\infty}^{\infty} a^{|k|}e^{-ik\lambda} = ... = \frac{1}{1-2a\cos(\lambda)+a^2}$$

The spectral density of Y is given by $S_Y(\lambda) = S_X(\lambda) + \sigma^2$ and so

$$A(\lambda) = \frac{S_X(\lambda)}{S_X(\lambda)+\sigma^2} = \frac{1/\sigma^2}{1-2a\cos(\lambda)+a^2+1/\sigma^2}.$$

The minimal mean square error can be calculated by (2.2).

2.3. Prediction. Let $X = (X_n)_{n\in\mathbb{Z}}$ be a stationary process with positive spectral density $S(\lambda) > 0$, $\forall\lambda \in \mathbb{R}$. The linear prediction problem is to estimate X_m, for some fixed $m \geq 1$ from the observations of $X_{-\infty}^0 = \{..., X_{-1}, X_0\}$, i.e. to find the orthogonal projection $\widehat{E}(X_m|X_{-\infty}^0) = \sum_{n=-\infty}^0 a_nX_n$ such that

$$E\Big(X_m - \sum_{n=-\infty}^{0} a_nX_n\Big)X_\ell = 0, \quad \forall\ell \leq 0.$$

This leads

$$R(m-\ell) - \sum_{n=-\infty}^{0} a_nR(n-\ell) =$$
$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(m-\ell)\lambda}S(\lambda)d\lambda - \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{n=-\infty}^{0} a_ne^{in\lambda}e^{-i\ell\lambda}S(\lambda)d\lambda = \tag{2.3}$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\big[e^{im\lambda} - A(\lambda)\big]S(\lambda)e^{i\ell\lambda}d\lambda = 0, \quad \forall\ell \geq 0$$

The latter holds if the Fourier transform of the function $\big[e^{im\lambda} - A(\lambda)\big]S(\lambda)$ contains only terms with positive exponentials.

Assume that $S(\lambda)$, being a positive real function, can be factored $S(\lambda) = \sigma(\lambda)\bar{\sigma}(\lambda)$ where $\bar{\sigma}(\lambda)$ is a function with Fourier transform of nonnegative exponentials. Then it suffices that the Fourier transform of $h(\lambda) := \big[e^{im\lambda} - A(\lambda)\big]\sigma(\lambda)$ contains only positive exponentials. In other words

$$e^{im\lambda}\sigma(\lambda) = A(\lambda)\sigma(\lambda) + h(\lambda)$$

where $h(\lambda)$ has only positive harmonics. Note $A(\lambda)\sigma(\lambda)$ has only nonpositive harmonics, so if $\sigma(\lambda) = \sum_{k=0}^{\infty} s_k e^{-ik\lambda}$, then

$$h(\lambda) = s_0 e^{im\lambda} + s_1 e^{i(m-1)\lambda} + ... + s_{m-1}e^{i\lambda}$$

and

$$A(\lambda) = \frac{1}{\sigma(\lambda)} \sum_{k=m}^{\infty} s_k e^{-i(k-m)\lambda}.$$

The prediction error is given by

$$V(m) := E\Big(X_m - \sum_{n=-\infty}^{0} a_n X_n\Big)^2 = E\Big(X_m - \sum_{n=-\infty}^{0} a_n X_n\Big)X_m =$$

$$R(0) - \sum_{n=-\infty}^{0} a_n R(n-m) =$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} S(\lambda)d\lambda - \frac{1}{2\pi}\int_{-\pi}^{\pi} \sum_{n=-\infty}^{0} a_n e^{i(n-m)\lambda} S(\lambda)d\lambda =$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \big[1 - A(\lambda)e^{-im\lambda}\big]S(\lambda)d\lambda \overset{\dagger}{=} \frac{1}{2\pi}\int_{-\pi}^{\pi} \big[e^{im\lambda} - A(\lambda)\big]\big[e^{-im\lambda} - \bar{A}(\lambda\big]S(\lambda)d\lambda =$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \big|e^{im\lambda} - A(\lambda)\big|^2 S(\lambda)d\lambda = \frac{1}{2\pi}\int_{-\pi}^{\pi} \big|[e^{im\lambda} - A(\lambda)]\sigma(\lambda)\big|^2 d\lambda =$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \big|h(\lambda)\big|^2 d\lambda = |s_0|^2 + ... + |s_{m-1}|^2$$

where the equality † is due to (2.3). It is worth noting that

$$\lim_{m\to\infty} V(m) = \sum_{\ell=0}^{\infty} |s_\ell|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(\lambda)d\lambda = R(0) = EX_1^2,$$

i.e. the prediction trivializes as $m \to \infty$.

Consider now the case $m = 1$. Suppose the function $\log S(\lambda)$ can be expanded into convergent Fourier series

$$\log S(\lambda) = \sum_{n=-\infty}^{\infty} b_n e^{-in\lambda}.$$

Since $\log S(\lambda)$ is a real function, $b_{-n} = \bar{b}_n$ and so

$$\sigma(\lambda) = \exp\{b_0/2 + b_1 e^{-i\lambda} + b_2 e^{-i2\lambda}...\}.$$

Using the expansion formula $e^z = \sum_{n=0}^{\infty} z^n/n!$, the first Fourier coefficient of $\sigma(\lambda)$ is found

$$s_0 = 1 + b_0/2 + \frac{(b_0/2)^2}{2!} + ... = \exp\{b_0/2\}.$$

which leads to

$$V(1) = \exp\{b_0\} = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi}\log S(\lambda)d\lambda\right\}. \tag{2.4}$$

The latter is known as Szegö-Kolmogorov formula.

Exercise 2.5. *Apply the above formulae to X_n from the Example 2.4.*

2.4. More about stationary processes. Let $\xi = (\xi_n)_{n\in\mathbb{Z}}$ be a stationary (in the wide sense) random sequence and denote by $H_n(\xi) = \overline{\text{span}}\{..., \xi_{n-1}, \xi_n\}$ and $H = \overline{\text{span}}\{..., \xi_{-1}, \xi_0, \xi_1, ...\}$. Introduce[4]

$$S(\xi) = \bigcap_{n\ge 0} H_{-n}(\xi).$$

If $H(\xi) = S(\xi)$, the process is called *singular* and if $S(\xi) = 0$, i.e. contains only random variables equivalent to zero, then the process is called *regular* or *purely nondeterministic*.

For example, the random process $X_n \equiv X$, where X is a square integrable random variable is singular (why?) and an i.i.d. sequence of $\mathbb{L}^2$ random variables is regular. In general a stationary process can be decomposed into singular and regular components $X_n = X_n^s + X_n^r$ (Wold decomposition). In turn, the regular part X_n^r admits the following expansion

$$X_n = \sum_{k=0}^{\infty}\alpha_k\varepsilon_{n-k}, \quad \forall n$$

where α_n's are real numbers and $\varepsilon = (\varepsilon_n)_{n\in\mathbb{Z}}$ is a sequence of orthonormal random variables. The sequence ε is called *innovation* process. It can be shown that a sequence is regular if and only if $\int_{-\pi}^{\pi}\log S(\lambda)d\lambda > -\infty$ (compare to (2.4)). The classic reference on stationary processes is (6).

3. Linear estimation: Kalman's state space approach

The estimation problems that can be considered within the Kolmogorov-Wiener framework are essentially limited to the case of stationary processes. The state space approach proposed by R. Kalman allows to solve many estimation problems in the non-stationary case.

3.1. Recursive orthogonal projection. Suppose that given the signal / observation pair of processes $(X, Y) = (X_n, Y_n)_{n\ge 1}$, the optimal linear estimate of X_n from $Y_1^n = \{Y_1, ..., Y_n\}$ is to be calculated for each $n \ge 1$, i.e. the orthogonal projection $\widehat{E}(X_n|Y_1^n)$ is to be found. The formulae of Theorem 1.5 are not efficient in this case, since each time n is increased the estimate is to be recalculated entirely. The key to a more efficient way of calculating $\widehat{E}(X_n|Y_1^n)$ is given in the next theorem, where the following notations are used

$$\widehat{X}_n = \widehat{E}(X_n|Y_1^n),\ \widehat{X}_{n|n-1} = \widehat{E}(X_n|Y_1^{n-1}),\ \widehat{Y}_{n|n-1} = \widehat{E}(Y_n|Y_1^{n-1})$$
$$P_n^X = E\big(X_n - \widehat{X}_n\big)\big(X_n - \widehat{X}_n\big)^*$$
$$P_{n|n-1}^X = E\big(X_n - \widehat{X}_{n|n-1}\big)\big(X_n - \widehat{X}_{n|n-1}\big)^*$$
$$P_{n|n-1}^Y = E\big(Y_n - \widehat{Y}_{n|n-1}\big)\big(Y_n - \widehat{Y}_{n|n-1}\big)^*$$

[4] recall that intersection of linear subspaces is a linear subspace

$$P^{XY}_{n|n-1} = \big[P^{YX}_{n|n-1}\big]^* = E\big(X_n - \widehat{X}_{n|n-1}\big)\big(Y_n - \widehat{Y}_{n|n-1}\big)^*$$

THEOREM 3.1. *Let $(X,Y) = (X_n, Y_n)_{n\ge 1}$ be a pair of $\mathbb{L}^2$ random processes with values in $\mathbb{R}^k$ and $\mathbb{R}^m$. Denote by Y_1^n the linear subspace spanned by $\{1, Y_1, ..., Y_n\}$. Then for $n \ge 1$*

$$\begin{aligned}\widehat{X}_n &= \widehat{X}_{n|n-1} + P^{XY}_{n|n-1}\big[P^Y_{n|n-1}\big]^{\oplus}\big(Y_n - \widehat{Y}_{n|n-1}\big)\\ P^X_n &= P^X_{n|n-1} - P^{XY}_{n|n-1}\big[P^Y_{n|n-1}\big]^{\oplus}P^{YX}_{n|n-1}\end{aligned}$$

subject to

$$\begin{aligned}&\widehat{X}_{1|0} = EX_1, \quad \widehat{Y}_{1|0} = EY_1\\ &P^X_{1|0} = \mathrm{cov}(X_1), \quad P^{XY}_{1|0} = \mathrm{cov}(X_1, Y_1), \quad P^Y_{1|0} = \mathrm{cov}(Y_1).\end{aligned}$$

PROOF. The random vector

$$\eta := X_n - \widehat{X}_{n|n-1} - P^{XY}_{n|n-1}\big[P^Y_{n|n-1}\big]^{\oplus}\big(Y_n - \widehat{Y}_{n|n-1}\big).$$

is orthogonal to Y_1^{n-1} and thus it suffices to check that it is orthogonal to $Y_n - \widehat{Y}_{n|n-1}$:

$$E\eta\big(Y_n - \widehat{Y}_{n|n-1}\big)^* = P^{XY}_{n|n-1}\Big(I - \big[P^Y_{n|n-1}\big]^{\oplus}P^Y_{n|n-1}\Big) = P^{XY}_{n|n-1}U\Big(I - \Lambda^{\oplus}\Lambda\Big)U^*$$

where $P^Y_{n|n-1} = U\Lambda U^*$. If $P^Y_{n|n-1} > 0$, then the desired property clearly holds. If $P^Y_{n|n-1}$ is only nonnegative definite, then orthogonality follows from the fact that $P^{XY}_{n|n-1}\widetilde{U} = 0$, where $\widetilde{U}$ is the submatrix of the eigenvectors, corresponding to zero eigenvalues.

By the same arguments

$$\begin{aligned}E\eta\eta^* = &P^X_{n|n-1} - P^{XY}_{n|n-1}\big[P^Y_{n|n-1}\big]^{\oplus}P^{YX}_{n|n-1} - P^{XY}_{n|n-1}\big[P^Y_{n|n-1}\big]^{\oplus}P^{YX}_{n|n-1}\\ &+ P^{XY}_{n|n-1}\big[P^Y_{n|n-1}\big]^{\oplus}P^Y_{n|n-1}\big[P^Y_{n|n-1}\big]^{\oplus}P^{YX}_{n|n-1} =\\ &= P^X_{n|n-1} - P^{XY}_{n|n-1}\big[P^Y_{n|n-1}\big]^{\oplus}P^{YX}_{n|n-1}\end{aligned}$$

□

3.2. The Kalman filter. R.Kalman suggested to treat the estimation problem via the state space approach and derived particularly simple and efficient filtering algorithm, assuming that the signal/observation pair is the solution of the linear recursion ($n \ge 1$)

$$\begin{aligned}X_n &= a_0(n) + a_1(n)X_{n-1} + a_2(n)Y_{n-1} + b_1(n)\varepsilon'_n + b_2(n)\varepsilon''_n\\ Y_n &= A_0(n) + A_1(n)X_{n-1} + A_2(n)Y_{n-1} + B_1(n)\varepsilon'_n + B_2(n)\varepsilon''_n\end{aligned} \tag{3.1}$$

where

- $\varepsilon' = (\varepsilon'_n)_{n\ge1}$ and $\varepsilon'' = (\varepsilon''_n)_{n\ge1}$ are orthogonal zero mean sequences with $E\varepsilon'\varepsilon'^* = I$ and $E\varepsilon''\varepsilon''^* = I$
- $a_i(n)$, $A_i(n)$, $b_i(n)$ and $B_i(n)$, $i = 0, 1, 2$ are deterministic matrix (vector) sequences of appropriate dimensions
- the initial conditions X_0 and Y_0 are square integrable random vectors, independent of ε' and ε''

Hereafter the time variable in $a_1(n)$, etc. is omitted for brevity

Theorem 3.2. *(R.Kalman, 1960) The orthogonal projection* $\widehat{X}_n = \widehat{E}(X_n|Y_0^n)$ *and the corresponding error covariance* $P_n = E(X - \widehat{X}_n)(X - \widehat{X}_n)^*$ *satisfy the equations*

$$\widehat{X}_n = a_0 + a_1\widehat{X}_{n-1} + a_2Y_{n-1} + \big(a_1P_{n-1}A_1^* + b \circ B\big)\times \\ \big(A_1P_{n-1}A_1^* + B \circ B\big)^{\oplus}\Big(Y_n - A_0 - A_1\widehat{X}_{n-1} - A_2Y_{n-1}\Big) \quad (3.2)$$

and

$$P_n = a_1P_{n-1}a_1^* + b \circ b - \big(a_1P_{n-1}A_1^* + b \circ B\big)\times \\ \big(A_1P_{n-1}A_1^* + B \circ B\big)^{\oplus}\big(a_1P_{n-1}A_1^* + b \circ B\big)^* \quad (3.3)$$

where $B \circ B := B_1B_1^* + B_2B_2^*$, $b \circ b := b_1b_1^* + b_2b_2^*$ *and* $b \circ B := b_1B_1^* + b_2B_2^*$. *The equations* (3.2) *and* (3.3) *are solved subject to*

$$\widehat{X}_0 = EX_0 + \operatorname{cov}(X_0, Y_0)\operatorname{cov}^{\oplus}(Y_0)(Y - EY)$$
$$P_0 = \operatorname{cov}(X_0) - \operatorname{cov}(X_0, Y_0)\operatorname{cov}^{\oplus}(Y_0)\operatorname{cov}(Y_0, X_0).$$

Proof. Apply Theorem 3.1, finding the expressions for all the building blocks, e.g.

$$\widehat{X}_{n|n-1} = \widehat{E}(X_n|Y_0^{n-1}) = a_0 + a_1\widehat{X}_{n-1} + a_2Y_{n-1}.$$

etc. □

3.2.1. *Some properties of the Kalman filter.*
1. The Kalman filter is the linear recursive equation (3.2), coupled with the Riccati equation (3.3). Note that (3.3) does not depend on the observations and thus can be calculated off-line.

2. Even if the coefficients in (3.1) are constant, the solution of (3.3) is a still nonconstant sequence of matrices. Under certain conditions (observability and controllability) P_n converges to a positive definite limit. In general the latter is not guaranteed.

3. The Kalman filter recursions can be seen as two stage algorithm: at the *propagation* stage the estimate $\widehat{X}_{n-1}$ and the error covariance P_{n-1} are extrapolated according to the dynamic equations to the time n and in the *update* stage the new estimate and error covariance is calculated on the basis of the interpolated estimate and the new observation. The sequence $\xi_n = Y_n - A_0 - A_1\widehat{X}_{n-1} - A_2Y_{n-1}$ consists of orthogonal random vectors and is called *innovations* to emphasize the fact that it carries all the sufficient information contained in the observations.

4. Clearly the Kalman filter is applicable to estimation problems with nonstationary signals. In the case of stationary processes with rational spectral densities, a model of (3.1) type can be found. The Kalman filter then leads to the estimate equations which asymptotically reduce to those obtained by Kolmogorov-Wiener theory.

3.3. Examples and applications.

3.3.1. *Simple scalar model.* Consider the simple model ($n \ge 1$)

$$\begin{aligned} X_n &= aX_{n-1} + b\varepsilon_n \\ Y_n &= AX_{n-1} + B\xi_n \end{aligned}$$

subject to a r.v. X_0 with zero mean and unit variance. $(\varepsilon_n)_{n\ge1}$ and $(\xi_n)_{n\ge1}$ are independent standard i.i.d sequences.

The Kalman filter equations read

$$\begin{aligned} \widehat{X}_n &= a\widehat{X}_{n-1} + aAP_{n-1}/(A^2P_{n-1}+B^2)(Y_n - A\widehat{X}_{n-1}) \\ P_n &= a^2P_{n-1} + b^2 - \big[aAP_{n-1}\big]^2/(A^2P_{n-1}+B^2) \end{aligned}$$

subject to $\widehat{X}_0 = 0$ and $P_0 = 1$.

It can be shown that in this case the limit $P_\infty = \lim_{n\to\infty} P_n$ exists and is the unique positive solution of

$$P_\infty = a^2P_\infty + b^2 - \big[aAP_\infty\big]^2/(A^2P_\infty+B^2).$$

Note that $P_\infty < \infty$ even if $|a| > 1$, i.e. when the system is unstable.

3.3.2. *Phase Locking Loop.* The signal $A(t) > 0$ modulates the sinusoidal carrier of a known frequency $\omega = 1$, so that the transmitted wave is given by

$$x(t) = A(t)\cos(t+\varphi)$$

where φ is the initial phase, chosen by the transmitter.

The receiver's antenna picks up this signal and passes the transmission to the A/D converter so that the sequence

$$Y_{n+1} = x(\Delta n) + \xi_{n+1} = A(\Delta n)\cos(\Delta n + \varphi) + \xi_{n+1}, \quad n = 0, 1, ...$$

is available to the receiver's processor, where $\Delta > 0$ is the sampling step and $(\xi_n)_{n\ge0}$ is an i.i.d. sequence of standard Gaussian r.v.

Suppose that $A(t)$ is constant (practically slowly varying signal), i.e. $A(t) \equiv A$. Both A and φ are necessary for the transmission decoding and are unknown to the receiver. Assume that both are r.v. with uniform distribution on $[a,b]$ ($b > a > 0$) and $[0, 2\pi]$ respectively. The following questions are of interest

(1) Find a recursion for calculation of $\widehat{x}(t) = \widehat{E}\big(x(t)|Y_k, k\Delta \le t\big)$.

(2) Does the limit $\lim_{t\to\infty} E\big(x(t) - \widehat{x}(t)\big)^2$ exist ? If yes, can perfect synchronization be achieved as $t \to \infty$?

1. Introduce

$$Z(t) := \begin{bmatrix} \zeta_1(t) \\ \zeta_2(t) \end{bmatrix} = \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix}$$

and define[5] a sequence $Z_n = Z(\Delta n)$, $n = 0, 1, 2, ...$ This sequence [6] satisfies the recursion

[5] Compare to

$$\dot{Z}_t = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} Z_t, \quad \text{s.t. } Z_0 = \begin{bmatrix} A\cos\varphi \\ -A\sin\varphi \end{bmatrix}$$

[6] Indeed

$$\begin{aligned} &\zeta_1(\Delta(n+1)) = A\cos(\Delta n + \varphi + \Delta) = A\cos(\Delta n + \varphi)\cos\Delta - A\sin(\Delta n+\varphi)\sin\Delta = \\ &= \zeta_1(\Delta n)\cos(\Delta) + \zeta_2(\Delta n)\sin(\Delta) \end{aligned}$$

$$Z_{n+1} = \begin{bmatrix} \cos\Delta & \sin\Delta \\ -\sin\Delta & \cos\Delta \end{bmatrix} Z_n := \theta(\Delta) Z_n, \quad \text{s.t. } Z_0 = \begin{bmatrix} A\cos\varphi \\ -A\sin\varphi \end{bmatrix} \tag{3.4}$$

The matrix $\theta(\Delta)$ is known as *direct cosine matrix* or *rotation matrix.*

Recall, that the observation sequence is

$$Y_n = [1\ 0]Z_{n-1} + \xi_n := u^* Z_{n-1} + \xi_n, \quad n \ge 1 \tag{3.5}$$

Note that for any $t \in [\Delta n, \Delta n + \Delta)$, $x(t) = u^* Z(t)$ and

$$\begin{aligned} \widehat{Z}(t) &= \widehat{E}\big(Z(t)|Y_k, k\Delta \le t\big) = \theta(t-\Delta n)\widehat{E}\big(Z(\Delta n)|Y_k, k\Delta \le t\big) = \\ &= \theta(t-\Delta n)\widehat{Z}(\Delta n) \end{aligned} \tag{3.6}$$

so it suffices to determine only the sequence $\widehat{Z}_n := \widehat{Z}(\Delta n)$ and thus also only $\widehat{x}(\Delta n)$. The equations (3.4) and (3.5) are in the form of the Kalman filter model, so $\widehat{x}_n = u^* \widehat{Z}_n$, $n = 1, 2, ...$ where

$$\begin{aligned} \widehat{Z}_n &= \theta(\Delta)\widehat{Z}_{n-1} + \frac{\theta(\Delta)P_{n-1}u}{u^*P_{n-1}u + 1}\big(Y_n - u^*\widehat{Z}_{n-1}\big) \\ P_n &= \theta(\Delta)P_{n-1}\theta^*(\Delta) - \frac{\theta(\Delta)P_{n-1}uu^*P_{n-1}\theta^*(\Delta)}{u^*P_{n-1}u + 1} \end{aligned} \tag{3.7}$$

How to choose the initial conditions? Define $Y_0 \equiv 0$. Clearly $\widehat{E}\big(Z_n|Y_1^n\big) \equiv \widehat{E}\big(Z_n|Y_0^n\big)$, so one can solve (3.7) starting with index $n = 1$, with

$$\begin{aligned} \widehat{Z}_0 &= \widehat{E}\big(Z_0|Y_0\big) = EZ_0 \\ P_0 &= EZ_0Z_0^* \end{aligned}$$

Now

$$EZ_0 = E\begin{bmatrix} A\cos\varphi \\ -A\sin\varphi \end{bmatrix} = 0$$

and

$$P_0 = \begin{bmatrix} EA^2E\cos^2\varphi & -EA^2E\cos\varphi\sin\varphi \\ -EA^2E\cos\varphi\sin\varphi & EA^2E\sin^2\varphi \end{bmatrix} = \frac{b^3 - a^3}{3(b-a)}\frac{1}{2}I \equiv C \cdot I$$

2. The existence of the limit $\lim_{n\to\infty} P_n$ can be established as a special case of a more general theory of Riccati equations. In this case however, P_n can be found explicitly and some additional insight can be gained. We make use of the *matrix inversion lemma*

Lemma 3.3. $A = B^{-1} + CD^{-1}C^* \Leftrightarrow A^{-1} = B - BC\big(D + C^*BC\big)^{-1}C^*B$

and similarly

$$\begin{aligned} \zeta_2(\Delta(n+1)) &= -A\sin(\Delta n + \varphi + \Delta) = -A\sin(\Delta n + \varphi)\cos(\Delta) - A\cos(\Delta n + \varphi)\sin(\Delta) = \\ &= \zeta_2(\Delta n)\cos(\Delta) - \zeta_1(\Delta n)\sin(\Delta) \end{aligned}$$

and (3.4) follows.

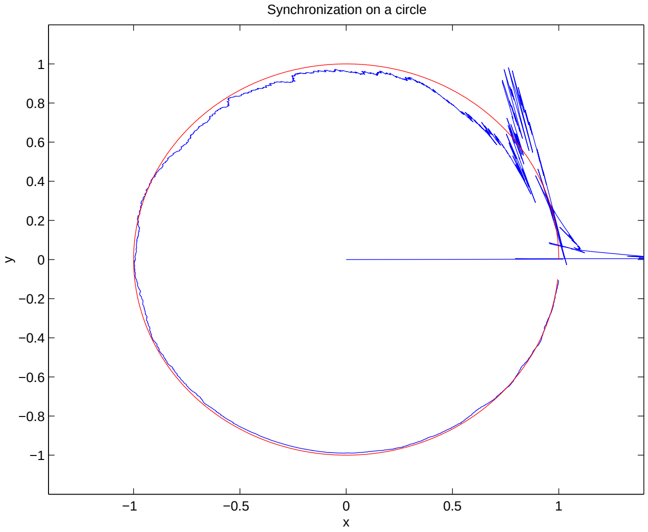

FIGURE 1. The phase locking process $\Delta = 2$msec.

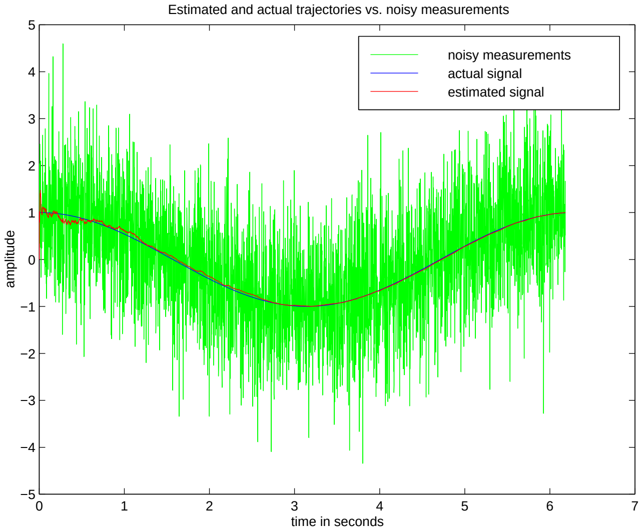

FIGURE 2. Actual trajectory, measured data, estimate ...

Put $Q_n = P_n^{-1}$ and apply the lemma with $B^{-1} := P_{n-1}$, $C = P_{n-1}u$, $D = -\big(u^*P_{n-1}u+1\big)$ to the Riccati equation [7]from (3.7):

$$Q_n \quad = \quad \theta(\Delta)\Big(P_{n-1}^{-1} - \frac{P_{n-1}^{-1}P_{n-1}uu^*P_{n-1}P_{n-1}^{-1}}{-u^*P_{n-1}u - 1 + u^*P_{n-1}P_{n-1}^{-1}P_{n-1}u}\Big)\theta^*(\Delta) =$$

$$= \theta(\Delta)\Big(Q_{n-1} + uu^*\Big)\theta^*(\Delta)$$

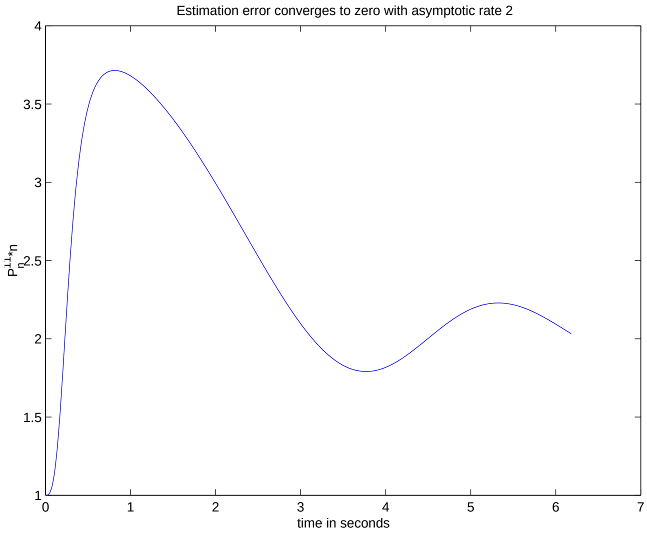

FIGURE 3. Mean square error of x-coordinate, multiplied by n, i.e. nP_n^{11}.

This is a linear recursion and it can be solved explicitly, $n \geq 1$

$$Q_n = \theta^n(\Delta)Q_0\theta^{n*}(\Delta) + \sum_{k=1}^{n} \theta^{n-k}(\Delta)\theta(\Delta)uu^*\theta^*(\Delta)\theta^{(n-k)*}(\Delta) =$$

$$= C^{-1}I + \sum_{k=0}^{n-1} \theta^{k+1}(\Delta)uu^*\theta^{(k+1)*}(\Delta) = C^{-1}I + \sum_{k=1}^{n} \theta(\Delta k)uu^*\theta^*(\Delta k)$$

where the property $\theta^k(\Delta) = \theta(\Delta k)$ had been used.

Now

$$P_n^{11} = E\big(x(\Delta n) - \widehat{x}(\Delta n)\big)^2 = \frac{Q_n^{22}}{Q_n^{11}Q_n^{22} - Q_n^{12}Q_n^{21}}$$

The elements of Q_n are readily found, e.g.

$$Q_n^{11} = 1/C + \sum_{k=1}^{n} \cos^2(\Delta k)$$

so

$$P_n^{11} =$$

$$\frac{1/C + \sum_{k=1}^n \sin^2(\Delta k)}{\big[1/C + \sum_{k=1}^n \cos^2(\Delta k)\big]\big[1/C + \sum_{k=1}^n \sin^2(\Delta k)\big] - \big[\sum_{k=1}^n \sin(\Delta k)\cos(\Delta k)\big]^2}$$

Let us investigate the asymptotic of P_n^{11} as $n \to \infty$. Recall that [8]

$$\frac{1}{n}\sum_{k=1}^{n} \cos(\Delta k) \xrightarrow{n\to\infty} 0, \qquad \frac{1}{n}\sum_{k=1}^{n} \sin(\Delta k) \xrightarrow{n\to\infty} 0$$

[7] The property $\theta\theta^* = I$ had been used here

for $\Delta \neq 2\pi$. So e.g.

$$\lim_{n\to\infty} \frac{1}{n}\Big(1/C + \sum_{k=1}^{n} \sin^2(\Delta k)\Big) = \lim_{n\to\infty} \frac{1}{n}\Big(1/C + \sum_{k=1}^{n} \frac{1}{2} - \sum_{k=1}^{n} \frac{1}{2}\cos(2\Delta k)\Big) = \frac{1}{2}$$

$$\lim_{n\to\infty} \frac{1}{n}\Big(1/C + \sum_{k=1}^{n} \cos^2(\Delta k)\Big) = \lim_{n\to\infty} \frac{1}{n}\Big(1/C + \sum_{k=1}^{n} \frac{1}{2} + \sum_{k=1}^{n} \frac{1}{2}\cos(2\Delta k)\Big) = \frac{1}{2}$$

$$\lim_{n\to\infty} \frac{1}{n}\Big(\sum_{k=1}^{n} \cos(\Delta k)\sin(\Delta k)\Big) = \lim_{n\to\infty} \frac{1}{n}\Big(\sum_{k=1}^{n} \frac{1}{2}\sin(2\Delta k)\Big) = 0$$

from which follows

$$\lim_{n\to\infty} nP_n^{11} = \frac{1/2}{1/2 \cdot 1/2 + 0} = 2$$

which suggests that the estimation error decays to zero as $1/n$ at rate 2 (irrespectively to Δ). I.e the finer Δ is chosen for fixed interval the better convergence is obtained. The sampling rate is then can be chosen to achieve essential convergence during the interval on which $A(t)$ remains effectively constant.

Note also that the estimate of $y(t) = A\sin(t+\varphi) = -\zeta_2(t)$ is obtained as a byproduct and can be used by the receiver.

3.3.3. *Tracking a particle motion.* This problem is taken from the original paper by R.Kalman (12). A number of particles leaves the origin at time $n = 0$ with random velocities; after $n = 0$, each particle moves with a constant (unknown velocity). Suppose that the position of one of these particles is measured, the data being contaminated by stationary, additive, correlated noise. What is the optimal estimate of the position and velocity of the particle at the time of the last measurement ?

Let $x_1(n)$ be the position and $x_2(n)$ the velocity of the particle; $x_3(n)$ is the noise. The problem is then represented by the model:

$$\begin{aligned} x_1(n+1) &= x_1(n) + x_2(n) \\ x_2(n+1) &= x_2(n) \\ x_3(n+1) &= \varphi x_3(n) + u_3(n) \\ y(n) &= x_1(n) + x_3(n) \end{aligned} \tag{3.8}$$

and the additional conditions

(1) $Ex_1^2(0) = Ex_2(0) = 0$, $Ex_2^2(0) = a^2 > 0$
(2) $Eu_3(n) = 0$, $Eu_3^2(n) = b^2$

The objective is to find $\widehat{x}_i(n) = \widehat{E}\big[x_i(n)|y_0^{n-1}\big]$, $i = 1, 2, 3$. Several solution versions are possible in this problem.

[8] this can be verified directly:

$$\frac{1}{n}\sum_{k=1}^{n} \cos(\Delta k) = \frac{1}{n}\sum_{k=1}^{n} 1/2\big(e^{j\Delta k} + e^{-j\Delta k}\big)$$

where the latter sum is split into pair of geometrical sequences.

1. The equivalent problem is formulated in the vector form:

$$X_n = \begin{pmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{pmatrix}$$

Then for $n \geq 1$

$$X_{n+1} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \varphi \end{pmatrix}}_{:=A} X_n + \underbrace{\begin{pmatrix} 0 \\ 0 \\ b \end{pmatrix}}_{:=B} u_n \tag{3.9}$$

where $u_n = u_3(n)/b$ is an i.i.d. Gaussian white noise.

$$Z_{n+1} := y_n = \underbrace{\begin{pmatrix} 1 & 0 & 1 \end{pmatrix}}_{:=C^*} X_n \tag{3.10}$$

Let $\widehat{X}_n := \widehat{E}(X_n|y_0^{n-1}) = \widehat{E}(X_n|Z_1^n)$. Then $n \geq 0$:

$$\widehat{X}_{n+1} = A\widehat{X}_n + AP_nC(C^*P_nC)^{\oplus}\big[Z_{n+1} - C^*A\widehat{X}_n\big] \tag{3.11}$$

$$P_{n+1} = AP_nA^* + BB^* - AP_nC(C^*P_nC)^{\oplus}C^*P_nA^* \tag{3.12}$$

with

$$\widehat{X}_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad P_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & a^2 & 0 \\ 0 & 0 & b^2 \end{pmatrix}$$

2. Note that $x_1(n) = nx_2(n)$. Redefine

$$X_n = \begin{pmatrix} x_2(n) \\ x_3(n) \end{pmatrix}$$

consequently

$$X_{n+1} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & \varphi \end{pmatrix}}_{:=A} X_n + \underbrace{\begin{pmatrix} 0 \\ b \end{pmatrix}}_{:=B} u_n$$

and

$$Z_{n+1} := y(n) = \underbrace{\begin{pmatrix} n & 1 \end{pmatrix}}_{:=C_n^*} X_n$$

The Kalman filter is again given by (3.11) and (3.12) with the newly defined terms (A, B, C_n and Z_n). Note that it is of lower dimension this time and

$$\widehat{X}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad P_0 = \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix}$$

Let us solve the Ricatti equation (3.12). Set:

$$P_n := \begin{pmatrix} \alpha_n & \gamma_n \\ \gamma_n & \beta_n \end{pmatrix}$$

Then:

$$C_n^*P_nC_n = n^2\alpha_n + 2n\gamma_n + \beta_n$$

and

$$P_n C_n C_n^* P_n = \begin{pmatrix} n\alpha_n + \gamma_n \\ n\gamma_n + \beta_n \end{pmatrix} \begin{pmatrix} n\alpha_n + \gamma_n, & n\gamma_n + \beta_n \end{pmatrix} =$$

$$= \begin{pmatrix} n^2\alpha_n^2 + 2n\alpha_n\gamma_n + \gamma_n^2, & n^2\alpha_n\gamma_n + n(\alpha_n\beta_n + \gamma_n^2) + \gamma_n\beta_n \\ n^2\alpha_n\gamma_n + n(\alpha_n\beta_n + \gamma_n^2) + \gamma_n\beta_n & n^2\gamma_n^2 + 2n\gamma_n\beta_n + \beta_n^2 \end{pmatrix}$$

and finally:

$$\begin{aligned} \alpha_{n+1} &= \alpha_n - \frac{n^2\alpha_n^2 + 2n\alpha_n\gamma_n + \gamma_n^2}{n^2\alpha_n + 2n\gamma_n + \beta_n} = \\ &= \frac{\alpha_n\beta_n - \gamma_n^2}{n^2\alpha_n + 2n\gamma_n + \beta_n} \end{aligned} \tag{3.13}$$

$$\begin{aligned} \gamma_{n+1} &= \varphi\gamma_n - \varphi\frac{n^2\alpha_n\gamma_n + n(\alpha_n\beta_n + \gamma_n^2) + \gamma_n\beta_n}{n^2\alpha_n + 2n\gamma_n + \beta_n} = \\ &= \varphi n\frac{\gamma_n^2 - \alpha_n\beta_n}{n^2\alpha_n + 2n\gamma_n + \beta_n} \end{aligned} \tag{3.14}$$

$$\begin{aligned} \beta_{n+1} &= \varphi^2\beta_n - \varphi^2\frac{n^2\gamma_n^2 + 2n\gamma_n\beta_n + \beta_n^2}{n^2\alpha_n + 2n\gamma_n + \beta_n} + b^2 = \\ &= \varphi^2 n^2\frac{\beta_n\alpha_n - \gamma_n^2}{n^2\alpha_n + 2n\gamma_n + \beta_n} + b^2 \end{aligned} \tag{3.15}$$

from which follows ($n \geq 0$):

$$\gamma_{n+1} = -\varphi n\alpha_{n+1} \tag{3.16}$$

$$\beta_{n+1} = b^2 + \varphi^2 n^2\alpha_{n+1} \tag{3.17}$$

so that ($n \geq 1$)

$$\begin{aligned} \alpha_{n+1} &= \frac{\alpha_n(b^2 + \varphi^2(n-1)^2\alpha_n) - \varphi^2(n-1)^2\alpha_n^2}{n^2\alpha_n - 2n\varphi(n-1)\alpha_n + (n-1)^2\varphi^2\alpha_n + b^2} = \\ &= \frac{\alpha_n b^2}{n^2\alpha_n - 2n\varphi(n-1)\alpha_n + (n-1)^2\varphi^2\alpha_n + b^2} \end{aligned}$$

and from (3.13), $\alpha_1 = a^2$. Set $\rho_n := \alpha_n^{-1}$. Then ($n \geq 1$):

$$\begin{aligned} \rho_{n+1} &= b^{-2}\big(n^2 - 2n\varphi(n-1) + (n-1)^2\varphi^2\big) + \rho_n = \\ & \rho_n + b^{-2}\big(n - (n-1)\varphi\big)^2 \end{aligned}$$

and ($n \geq 2$)

$$\rho_n = a^{-2} + b^{-2}\sum_{k=1}^{n-1}\big(k - (k-1)\varphi\big)^2$$

So that ($n \geq 2$)

$$\alpha_n = \rho_n^{-1} = \frac{a^2b^2}{b^2 + a^2\sum_{k=1}^{n-1}\big(k - (k-1)\varphi\big)^2}$$

and

$$\gamma_n = \frac{-\varphi(n-1)b^2a^2}{b^2 + a^2\sum_{k=1}^{n-1}\big(k - (k-1)\varphi\big)^2}$$

$$\beta_n = b^2 + \frac{\varphi^2(n-1)^2 b^2 a^2}{b^2 + a^2 \sum_{k=1}^{n-1}\big(k-(k-1)\varphi\big)^2}$$

Several interesting observations are worth noting.

(a) Consider the $\underline{\text{case } \varphi \neq 1}$. Note that $\alpha_n \to 0$, $\gamma_n \to 0$ and $\beta_n \to b^2$ as $n \to \infty$, since [9]

$$\begin{aligned}|\beta_n - b^2| &\le b^2 \frac{\varphi^2(n-1)^2}{\sum_{k=1}^{n-1}\big(k(1-\varphi)+\varphi\big)^2} \le \\ &\le \frac{b^2}{(1-\varphi)^2}\frac{\varphi^2(n-1)^2}{\sum_{k=1}^{n-1}k^2} \to 0, \quad n \to \infty\end{aligned}$$

This means that the estimate of the velocity converges (in the mean square) to the real velocity and ceases to update ($\widehat{x}_2(n+1) = \widehat{x}_2(n)$). However the estimate of the current sample of the noise remains uncertain with error b^2 and is generated by:

$$\widehat{x}_3(n+1) = \varphi\widehat{x}_3(n) + \varphi(y(n) - n\widehat{x}_2(n) - \widehat{x}_3(n)) = \varphi(y(n) - n\widehat{x}_2(n))$$

(b) Surprisingly, when $\varphi = 1$, it follows from the formulae above that as $n \to \infty$

$$\begin{aligned}\alpha_n &= \frac{a^2b^2}{b^2 + a^2(n-1)} \propto \frac{b^2}{n} \\ \gamma_n &\propto b^2 \\ \beta_n &\propto b^2 n\end{aligned}$$

i.e. the noise sequence $x_3(n)$ can not be estimated efficiently, though the velocity of the particle $x_2(n)$ is inferred perfectly as $n \to \infty$! Moreover when $|\varphi| > 1$, the noise is "exponentially unstable", while with $\varphi = 1$ its trajectories diverge to infinity with a linear rate. On the first glance it may seem that if the position estimate converges in the former case, it should converge in the latter case a fortiori!

Let us try to understand this phenomenon. Set $x_2(n) \equiv \xi$ for brevity and assume first that $\varphi \neq 1$. Define $\delta y(0) = y(0)$ and for $n \ge 1$

$$\begin{aligned}\delta y(n) &:= y(n) - \varphi y(n-1) = (1-\varphi)\sum_{k=1}^{n-1}\xi + \xi + x_3(n) - \varphi x_3(n-1) = \\ &= \big[n(1-\varphi)+\varphi\big]\xi + u_3(n-1) := \varrho_n\xi + bu_{n-1}\end{aligned}$$

Note that the 'signal-to-noise' ratio increases with n. Set $\widehat{\xi}_n := \widehat{E}(\xi|\delta y_0^n) = \widehat{E}(\xi|y_0^n) = \widehat{E}(x_2(n)|y_0^n)$ and $P_n = E(\xi - \widehat{\xi}_n)^2$. Then:

$$P_n = P_{n-1} - \frac{P_{n-1}^2\varrho_n^2}{\varrho_n^2 P_{n-1} + b^2} = \frac{P_{n-1}b^2}{\varrho_n^2 P_{n-1} + b^2}$$

[9] It can be easily seen that $\sum_{k=1}^n k^2$ is of $O(n^3)$:

$$n^3 = \sum_{k=1}^n\big[k^3 - (k-1)^3\big] = \sum_{k=1}^n\big[3k^2 - 3k + 1\big] = 3\sum_{k=1}^n k^2 + O(n^2)$$

Set $Q_n = P_n^{-1}$:

$$Q_n = Q_{n-1} + 1/b^2 \varrho_n^2 \propto 1/b^2 \sum_{k=1}^{n} \varrho_k^2 \propto n^3$$

i.e. P_n is $O(1/n^3)$. Now form the estimate of $x_3(n)$:

$$\widetilde{x}_3(n+1) \quad = \quad \varphi\big(y(n) - n\widehat{\xi}_n\big)$$

Introduce $\Delta_3(n) = x_3(n) - \widetilde{x}_3(n)$ and $\Delta_2(n) = \xi - \widehat{\xi}_n$, then:

$$\begin{aligned} \Delta_3(n+1) &= \varphi x_3(n) + u_3(n) - \varphi\big(y(n) - n\widehat{\xi}_n\big) = \\ &= u_3(n) - n\varphi\Delta_2(n) \end{aligned}$$

so that

$$E\Delta_3^2(n) = b^2 + \varphi^2 n^2 E\Delta_2^2(n) \to b^2, \quad n \to \infty$$

Now let us consider the case $\varphi = 1$. Verify that $x_3(n)$ can not be estimated with the mean square error growth rate better than $O(n)$. Note that $\delta y(n) := y(n) - \varphi y(n-1) = \xi + u_3(n-1)$ so that $\varrho_n \equiv 1$ and hence $P_n \propto 1/n$ for n large. Recall that $x_1(n) = n\xi$, so that $\widehat{x}_1(n) = n\widehat{\xi}_n$ and $E(x_1(n) - \widehat{x}_1(n))^2 = n^2 P_n \propto n$. Since $y(n) = x_1(n) + x_3(n)$, we conclude [10] that $x_3(n)$ can not be estimated with the error growth rate better that $O(n)$.

3.3.4. *Solution of linear equations by means of the Kalman filter.*

Consider the system of linear equations $Ax = b$, where A is an $n \times m$ real matrix and b is an $n \times 1$ real vector. Let $r = \operatorname{rank}(A) \le \min(n, m)$. Depending on r these equations may have the unique solution, an infinite number of solutions or no solutions at all. If $n = m$ and A is nonsingular, then $x^\circ = A^{-1}b$ is the unique solution. As shown below, in the case of infinitely many solutions, there is one with the minimal Euclidian norm, while in the case of no solutions there is always a unique vector x°, which minimizes the norm $\|Ax - b\|_2$ among $x \in \mathbb{R}^m$. There is one formula to calculate any of the aforementioned vectors: $x^\circ = A^\oplus y$, where $A^\oplus$ is the *Moore-Penrose pseudoinverse of* A, which is briefly described below.

Theorem 3.4. *(Singular Values Decomposition)*
If A is a real $m \times n$ matrix , then there exist orthogonal matrices

$$U = \big[u_1, ..., u_m\big] \in \mathbb{R}^{m\times m}, \quad V = \big[v_1, ..., v_n\big] \in \mathbb{R}^{n\times n}$$

such that

$$U^* A V = \operatorname{diag}(\sigma_1, ..., \sigma_p) \in \mathbb{R}^{m\times n}, \quad p = \min\{m, n\} \tag{3.18}$$

where $\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_p \ge 0$.

Proof. (Version I) Assume e.g. that $m \le n$ and let $r := \operatorname{rank}(A) \le m$. Set $Q := AA^* \in \mathbb{R}^{m\times m}$. Since Q is symmetric and non-negative definite there exists an orthogonal matrix $U \in \mathbb{R}^{m\times m}$, such that $U^* Q U = \operatorname{diag}(\sigma_1^2, ..., \sigma_r^2, 0, ..., 0)$, such that $\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_r > 0$. Let U' and U'' be the columnwise partition of U, such that $U' = \big[u_1, ..., u_r\big] \in \mathbb{R}^{m\times r}$ and $U'' = \big[u_{r+1}, ..., u_m\big] \in \mathbb{R}^{m\times(m-r)}$, where u_i are columns of U. Clearly $U'^* Q U' = \operatorname{diag}(\sigma_1^2, ..., \sigma_r^2) := \Gamma \in \mathbb{R}^{r\times r}$,

[10] Use the following fact. If $Z = X + Y$ and $\widehat{X} = E(X|Z)$ and $P = E(X - \widehat{X}_n)^2$, then $\widehat{Y} = E(Y|Z) = E(Z - X|Z) = Z - \widehat{X}$ and $Q := E(Y - \widehat{Y})^2 = P$

$U''^*QU'' = 0 \in \mathbb{R}^{(m-r)\times(m-r)}$, etc. Introduce matrix $V' := A^*U'\Gamma^{-1/2} \in \mathbb{R}^{n\times r}$. Note that

$$V'^*V' = \Gamma^{-1/2}U'^*AA^*U'\Gamma^{-1/2} = I_r \in \mathbb{R}^{r\times r}$$

where I_r stands for identity matrix of size r. The latter implies that the columns of V' are orthogonal. Choose $V'' \in \mathbb{R}^{n\times(n-r)}$ with orthogonal columns which span the orthogonal complement to the subspace spanned by columns of V' (this is always possible!). Form $V = \left[V'\ V''\right] \in \mathbb{R}^{n\times n}$, which is orthogonal matrix by construction. Then the statement of the theorem follows from

$$U^*AV = \begin{pmatrix} U'^* \\ U''^* \end{pmatrix} A \begin{pmatrix} V' & V'' \end{pmatrix} = \begin{pmatrix} U'^*AV' & U'^*AV'' \\ U''^*AV' & U''^*AV'' \end{pmatrix} =$$

$$= \begin{pmatrix} U'^*AA^*U'\Gamma^{-1/2} & U'^*AV'' \\ U''^*AV' & U''^*AV'' \end{pmatrix} = \begin{pmatrix} \Gamma^{1/2} & 0 \\ 0 & 0 \end{pmatrix}$$

where the latter equality follows from the facts: (i) $U''^*AA^*U'' = 0 \implies U''^*A = 0$ and (ii) $U'^*AV'' = \Gamma^{1/2}\Gamma^{-1/2}U'^*AV'' = \Gamma^{1/2}V'^*V'' = 0$. □

Proof. (Version II) Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ a pair of unit vectors (i.e. e.g. $\|x\| = \sqrt{\sum_{i=1}^m x_i^2} = 1$) such that

$$Ax = \sigma y$$

where $\sigma = \|A\|_2$ (recall that $\|A\|_2^2 = \max_{x\neq 0}\|Ax\|^2/\|x\|^2$). It is always possible to choose such vectors: e.g. $x := \arg\max_{\xi\in\mathbb{R}^n:\|\xi\|=1}\|A\xi\|$ and $y := Ax/\sigma$.

Since any orthonormal set of vectors can be completed to an orthonormal basis, there exist $U' \in \mathbb{R}^{m\times(m-1)}$ and $V' \in \mathbb{R}^{n\times(n-1)}$, such that $U = [y\ U'] \in \mathbb{R}^{m\times m}$ and $V = [x\ V'] \in \mathbb{R}^{n\times n}$ are orthogonal. Note that U^*AV is of the following structure

$$U^*AV = \begin{pmatrix} \sigma & w^* \\ 0 & B \end{pmatrix} := A'$$

where $w \in \mathbb{R}^{(n-1)}$ and $B \in \mathbb{R}^{(m-1)\times(n-1)}$. Since

$$\left\| A' \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 \geq (\sigma^2 + w^*w)^2$$

so that $\|A'\|_2^2 \geq (\sigma^2 + w^*w)$. But $\sigma^2 = \|A\|_2^2 = \|A'\|_2^2$ (since $\|\cdot\|_2$-norm is invariant under orthogonal transformations), which implies that $w^*w = 0$ or $w = 0$. Now the statement of the theorem is verified by induction. □

Theorem 3.5. *Let $U^*AV = \Sigma$ be the SVD of matrix $A \in \mathbb{R}^{m\times n}$ with $r =$ rank(A). If $U = [u_1, ..., u_m]$ and $V = [v_1, ..., v_n]$ and $b \in \mathbb{R}^m$, then*

$$x^\circ = \sum_{i=1}^r \frac{u_i^*b}{\sigma_i} v_i \tag{3.19}$$

has the least $\|\cdot\|_2$ norm among all such vectors x, that bring $\|Ax - b\|$ to minimum. Moreover

$$\|Ax^\circ - b\|_2^2 = \sum_{i=r+1}^m \left(u_i^*b\right)^2$$

Proof. First let us show that there exists a unique vector of least norm that brings $\|Ax-b\|_2$ to minimum. (Note that if $r<n$, then there are a lot of vectors which minimize this norm, since if x is such a vector then $x+z$, where $z \in \text{null}(A)$, also is minimizing). Let

$$\chi = \{x \in \mathbb{R}^n : \|Ax-b\|_2 \le \|Ay-b\|_2, \forall y \neq x\}$$

χ is convex. Indeed fix $x_1, x_2 \in \chi$, $\lambda \in [0,1]$, then

$$\|A\big(\lambda x_1 + (1-\lambda)x_2\big) - b\|_2 \le \lambda\|Ax_1 - b\|_2 + (1-\lambda)\|Ax_2 - b\|_2 = \min_{x\in\mathbb{R}^n} \|Ax-b\|_2$$

so that $\lambda x_1 + (1-\lambda_2)x_2 \in \chi$. Since χ is convex, there is a unique vector in χ with the least norm.

For any $x \in \mathbb{R}^n$ (recall that orthogonal transformation does not change $\|\cdot\|_2$ norm of a vector)

$$\begin{aligned}\|Ax-b\|_2^2 &= \|(U^*AV)(V^*x) - U^*b\|_2^2 = \|\Sigma\alpha - U^*b\|_2^2 = \\ &= \Sigma_{i=1}^r(\sigma_i\alpha_i - u_i^*b)^2 + \Sigma_{i=r+1}^m(u_i^*b)^2 \ge \Sigma_{i=r+1}^m(u_i^*b)^2 \qquad (3.20)\end{aligned}$$

where $\alpha := V^*x$. The lower bound in (3.20) is attained if $\alpha_i = u_i^*b/\sigma_i$, $i=1,...,r$ are chosen. Since $x = V\alpha$, the least norm of x is obtained if α_i, $i=r+1,...,m$ are set to zero. This completes the proof. □

Definition 3.6. Moore-Penrose inverse of A is a matrix in $\mathbb{R}^{n\times m}$ defined by

$$A^\oplus = V\Sigma^\oplus U^*$$

where

$$\Sigma^\oplus = \text{diag}\left(\frac{1}{\sigma_1},...,\frac{1}{\sigma_r},0,...,0\right) \in \mathbb{R}^{n\times m}$$

Remark 3.7. Note that

$$x^\circ = A^\oplus b$$

Corollary 3.8. *(i) If $n=m=rank(A)$, then $A^\oplus = A^{-1}$; (ii) if* $\text{rank}(A)=n$, *then $A^\oplus = (A^*A)^{-1}A^*$*

Proof. It is easy to verify that in both cases the matrices are unique solutions to minimization of $\|Ax-b\|$ so the statement of the Corollary is implied by uniqueness of $A^\oplus$. The latter can be verified also directly, taking into account that Σ is full rank. □

Theorem 3.9. *$A^\oplus$ is the unique matrix satisfying:*

$$AA^\oplus A = A \qquad (3.21)$$

and there exists a pair of matrices Q and P such that

$$A^\oplus = PA^* = A^*Q \qquad (3.22)$$

i.e. rows and columns of $A^\oplus$ are linear combinations of the rows and columns of A^*.

Proof. First we show that there exist a unique matrix satisfying the (3.21) and (3.22). Assume that there are two matrices satisfying these equations: $A_1^\oplus$ and $A_2^\oplus$. Then

$$AA_1^\oplus A = A, \quad A_1^\oplus = P_1A^* = A^*Q_1$$

and

$$AA_2^\oplus A = A, \quad A_2^\oplus = P_2A^* = A^*Q_2$$

for some matrices P_1, P_2, Q_1, Q_2. Let $D = A_1^{\oplus} - A_2^{\oplus}$, $P = P_1 - P_2$, $Q = Q_1 - Q_2$, then

$$ADA = 0 \in \mathbb{R}^{m\times n}, \quad D = PA^* = A^*Q$$

and (by $D^* = Q^*A$)

$$(DA)^*(DA) = A^*D^*DA = A^*Q^*ADA = 0$$

that is $DA = 0$. Making use of $D^* = AP^*$ we find that

$$DD^* = DAP^* = 0$$

so $D = A_1^{\oplus} - A_2^{\oplus} = 0$.

Now it is left to check that $A^{\oplus}$ satisfies (3.21) and (3.22).

$$AA^{\oplus}A = (U\Sigma V^*)(V\Sigma^{\oplus}U^*)(U\Sigma V^*) = U\Sigma\Sigma^{\oplus}\Sigma V^* = U\Sigma V^* = A$$

To check (3.22) we use the relation

$$\Sigma^{\oplus} = (\Sigma^*\Sigma)^{\oplus}\Sigma^* = \Sigma^*(\Sigma\Sigma^*)^{\oplus}$$

which is verified by simple manipulations with diagonal matrices.

$$\begin{aligned} A^{\oplus} &= V\Sigma^{\oplus}U^* = V(\Sigma^*\Sigma)^{\oplus}\Sigma^*U^* = V(\Sigma^*\Sigma)^{\oplus}V^*V\Sigma^*U^* := PA^* \\ A^{\oplus} &= V\Sigma^{\oplus}U^* = V\Sigma^*(\Sigma\Sigma^*)^{\oplus}U^* = V\Sigma^*U^*U(\Sigma\Sigma^*)^{\oplus}U^* := A^*Q \end{aligned}$$

□

Remark 3.10. Since $A^{\oplus}$ is a unique matrix, Theorem 3.9 can be used as its definition.

Lemma 3.11. *$A^{\oplus}$ obeys the following properties*

(1) $AA^{\oplus}A = A$, $A^{\oplus}AA^{\oplus} = A^{\oplus}$
(2) $(A^*)^{\oplus} = (A^{\oplus})^*$
(3) $(A^{\oplus})^{\oplus} = A$
(4) $(A^{\oplus}A)^2 = A^{\oplus}A$, $(A^{\oplus}A)^* = A^{\oplus}A$, $(AA^{\oplus})^2 = AA^{\oplus}$, $(AA^{\oplus})^* = AA^{\oplus}$
(5) $(A^*A)^{\oplus} = A^{\oplus}(A^*)^{\oplus} = A^{\oplus}(A^{\oplus})^*$
(6) $A^{\oplus} = (A^*A)^{\oplus}A^* = A^*(AA^*)^{\oplus}$
(7) $A^{\oplus}AA^* = A^*AA^{\oplus} = A^*$
(8) *if S is an orthogonal matrix, then $(SAS^*)^{\oplus} = SA^{\oplus}S^*$*
(9) *if A is a symmetric nonnegative definite matrix of order $n \times n$ of rank $r < n$, then*

$$A^{\oplus} = T^*(TT^*)^{-2}T$$

where $T \in \mathbb{R}^{r\times n}$ of rank r is defined by the decomposition

$$A = T^*T$$

Proof. Properties (1)-(7) can be verified directly for diagonal matrix Σ and then extended due to orthogonality of U and V to A:

(1) $A^{\oplus}AA^{\oplus} = V\Sigma^{\oplus}U^*U\Sigma V^*V\Sigma^{\oplus}U^* = V\Sigma^{\oplus}U^* = A^{\oplus}$
(2) $A^* = V\Sigma^*U^*$, so that by definition $(A^*)^{\oplus} = U(\Sigma^{\oplus})^*V^* = (A^{\oplus})^*$
(3) $A^{\oplus} = V\Sigma^{\oplus}U^*$. By definition $(A^{\oplus})^{\oplus} = U(\Sigma^{\oplus})^{\oplus}V^* = U\Sigma V^* = A$

(4)
$$\begin{aligned}&(A^{\oplus}A)^2 = A^{\oplus}AA^{\oplus}A = A^{\oplus}A\\ &(A^{\oplus}A)^* = A^*(A^*)^{\oplus} = V\Sigma^*U^*U(\Sigma^*)^{\oplus}V^* = V(\Sigma^{\oplus}\Sigma)^*V^* =\\ &= V\Sigma^{\oplus}\Sigma V^* = V\Sigma^{\oplus}U^*U\Sigma V^* = A^{\oplus}A, \quad \text{etc.}\end{aligned}$$

(5)
$$\begin{aligned}&(A^*A)^{\oplus} = (V\Sigma^*U^*U\Sigma V^*)^{\oplus} = (V\Sigma^*\Sigma V^*)^{\oplus} = V(\Sigma^*\Sigma)^{\oplus}V^* =\\ &= V\Sigma^{\oplus}(\Sigma^{\oplus})^*V^* = V\Sigma^{\oplus}U^*U(\Sigma^{\oplus})^*V^* = A^{\oplus}(A^*)^{\oplus}, \quad \text{etc.}\end{aligned}$$

(6) follows from (5), (4), (1)

(7) follows from (1) and (4)

(8) If S is orthogonal, then $SUU^*S^* = I$, i.e. SU is orthogonal as well. So is SV. It follows
$$(SAS^*)^{\oplus} = (SU\Sigma V^*S^*)^{\oplus} = SV\Sigma^{\oplus}U^*S^* = SA^{\oplus}S^*$$

(9) Any matrix $A \in \mathbb{R}^{m\times n}$ of rank r can be decomposed into
$$A = BC$$
where $B \in \mathbb{R}^{m\times r}$ and $C \in \mathbb{R}^{r\times n}$. Indeed form B from r independent columns of A, then choose rows of C so that $A = BC$. Let us show that $A^{\oplus} = C^{\oplus}B^{\oplus}$. Since C and B are full rank we have
$$C^{\oplus}B^{\oplus} = C^*(CC^*)^{-1}(B^*B)^{-1}B^*$$
and so
$$BC(C^{\oplus}B^{\oplus})BC = BCC^*(CC^*)^{-1}(B^*B)^{-1}B^*BC = BC$$
i.e. (3.21) holds. Similarly it can be shown that (3.22). Since in the case of symmetric A, $B = T^*$ and $C = T$, the desired formula result follows.

□

Consider a set of linear algebraic equations $Ax = b$, where A is an $m \times n$ matrix, b is a vector in $\mathbb{R}^m$ and $x \in \mathbb{R}^n$ is the vector to be found to minimize the Euclidian norm $\|Ax-b\|$. If A is square and nonsingular, then the solution is unique $x^\circ = A^{-1}b$, while in general $x^\circ = A^{\oplus}b$ as discussed above. The calculation of $A^{\oplus}$ is essentially equivalent to the problem of finding the eigenvalues of the nonnegative definite matrix AA^* (see Theorem 3.4) and for large matrices may be cumbersome. Moreover if only x° is required the calculation of $A^{\oplus}$ is unnecessary.

Consider the following auxiliary estimation problem: let X be a zero mean Gaussian random vector with unit covariance matrix and $Y = AX$, where A. The optimal linear estimate of X from Y is

$$\begin{aligned}\widehat{X} = \widehat{E}(X|Y) = \mathrm{cov}(X,Y)\,\mathrm{cov}^{\oplus}(Y)Y =\\ \mathrm{cov}(X)A^*\big[A\,\mathrm{cov}(X)A^*\big]^{\oplus}Y = A^*\big[AA^*\big]^{\oplus}Y = A^{\oplus}Y, \quad (3.23)\end{aligned}$$

where the latter equality is nothing but the property (6) of the pseudoinverse.

Notice that (3.23) holds only for the observation vector Y, compatible with the model $Y = AX$ and hence coincides with the linear function $x = A^{\oplus}y$ only on the n-dimensional subspace $\{y \in \mathbb{R}^m : y = Ax, x \in \mathbb{R}^n\}$. Hence e.g. when $m \le n$ and A is full rank, the orthogonal projection formula obtained in the auxiliary stochastic problem coincides with the solution of the linear set of equations $Ax = y$. If y is not in the range of Ax, it can still be substituted into the formula of the orthogonal projection, but the result will typically differ from $A^{\oplus}y$.

The auxiliary problem (and thus, the system of linear equations $Ax = b$ with $b \in \{y \in \mathbb{R}^m : y = Ax, x \in \mathbb{R}^n\}$) can be solved recursively by the Kalman filter. Denote the rows of A by a^i, $i = 1, ..., m$ and consider the signal satisfying the recursion

$$\begin{aligned} X_i &= X_{i-1} \\ X_0 &= X, \end{aligned}$$

The observation is generated by

$$Y_i = a^i X_{i-1}.$$

The Kalman filter equations read

$$\begin{aligned} \widehat{X}_i &= \widehat{X}_{i-1} + P_{i-1}a^{i*}\big[a^i P_{i-1}a^{i*}\big]^{\oplus}\big(Y_i - a^i\widehat{X}_{i-1}\big) \\ P_i &= P_{i-1} - P_{i-1}a^{i*}\big[a^i P_{i-1}a^{i*}\big]^{\oplus} a^i P_{i-1}. \end{aligned}$$

Clearly $\widehat{X}_i = A_i^{\oplus} Y^i$, where A_i is the matrix of the first i rows of A and Y^i is the sub-vector of the first entries of Y and $\widehat{X}_m = A^{\oplus}Y$. In other words, the Kalman filter can be applied to y (in the range of Ax) to generate $\widehat{x}_i = A_i^{\oplus} y^i$ for each $i = 1, ..., m$ and $\hat{x}_m = A^{\oplus}y$.

Notice, however, that if $m > n$, and the first n rows of A are linearly independent, then the algorithm ignores the rest of the rows and hence will not yield $x = A^{\oplus}y$ in this case (as we have already stressed before).

Since

$$\big[a^i P_{i-1}a^{i*}\big]^{\oplus} = \begin{cases} \big[a^i P_{i-1}a^{i*}\big]^{-1}, & a^i P_{i-1}a^{i*} > 0 \\ 0 & \text{otherwise} \end{cases}.$$

the natural question is what happens when $a^i P_{i-1}a_i^*$ vanishes ?

It turns out that

$$a^i P_{i-1}a^{i*} = \min_{c_1,\dots,c_{i-1}} \|a^i - \sum_{\ell} c_\ell a^\ell\|^2, \tag{3.24}$$

which means that when $a^i P_{i-1}a^{i*} = 0$ is encountered, the row a_i is the linear combination of the previous rows and thus the matrix is not full rank.

To verify (3.24), use the properties of pseudoinverse

$$\begin{aligned} &\min_{c_1,\dots,c_{i-1}} \|a^i - \sum_{\ell} c_\ell a^\ell\|^2 = \min_{c\in\mathbb{R}^{i-1}} \|a^i - c^* A_{i-1}\|^2 = \|a^i - a^i A_{i-1}^{\oplus} A_{i-1}\|^2 = \\ &a^i\big(I - A_{i-1}^{\oplus}A_{i-1}\big)\big(I - A_{i-1}^{\oplus}A_{i-1}\big)^* a^{i*} = a^i\big(I - A_{i-1}^{\oplus}A_{i-1}\big)^2 a^{i*} = \\ &a^i\big(I - A_{i-1}^{\oplus}A_{i-1}\big)a^{i*} = a^i P_{i-1}a^{i*}. \end{aligned}$$

It is also worth noting that in practice small values of $a^i P_{i-1}a^{i*}$ may indicate that the matrix is close to singular and numerical problems are possible.

CHAPTER 3

Conditional expectation and nonlinear estimation

1. Conditional expectation

Let $(\Omega, \mathcal{F}, P)$ be a fixed probability space. The *conditional probability* of event A, given event B is defined

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.1}$$

where $0/0 = 0$ is understood for definiteness. The probabilistic interpretation of this definition is that once B happens (i.e. $\omega' \in B$ is obtained), the probability of A changes: it becomes the probability of being in the part of A, which is also in B, normalized by probability of B. For example, intuitively one expects that if A and B are mutually exclusive, then probability of A should be zero, given B happened. Or if B is a subset of A, i.e. if B happens then A happens too, the probability of A given B should be one.

Consider now the events $\mathcal{D} = \{D_1, ..., D_n\}$, such that $D_i \cap D_j = \emptyset$ and $\Omega = \sum_{i=1}^n D_i$. The conditional probability of A given the partition $\mathcal{D}$ is a random variable (!)

$$P(A|\mathcal{D})(\omega) = \sum_{i=1}^n \frac{P(A \cap D_i)}{P(D_i)} I_{D_i}(\omega). \tag{1.2}$$

Finally the *conditional expectation* of a r.v. X with values in $\{x_1, ..., x_m\}$, given the partition $\mathcal{D}$ is a r.v.

$$E(X|\mathcal{D}) = \sum_{j=1}^m x_j \sum_{i=1}^n \frac{P(\{X(\omega) = x_j\} \cap D_i)}{P(D_i)} I_{D_i}(\omega). \tag{1.3}$$

However in many cases the conditioning with respect to events of zero probability is required. E.g. suppose one chooses at random a point ω from $([0,1], \mathcal{B}, \lambda)$ and then tosses a coin with probability of heads $p = \omega$. Intuitively one would expect that $P(h|\omega = 1/3) = 1/3$, though the definitions above cannot provide this answer.

The following definition of the conditional expectation is much more general

Definition 1.1. Let $X(\omega)$ be a real random variable with $E|X| < \infty$, defined on a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}'$ be a sub σ-algebra of $\mathcal{F}$. The conditional expectation $E(X|\mathcal{F}')(\omega)$ is a random variable, such that

(1) $\big\{E(X|\mathcal{F}')(\omega) \le x\big\} \in \mathcal{F}'$ for any $x \in \mathbb{R}$ (in other words, $E(X|\mathcal{F}')(\omega)$ is $\mathcal{F}'$-measurable)
(2) $E\big(X - E(X|\mathcal{F}')\big)I(A) = 0$ for any $A \in \mathcal{F}'$.

The proof of correctness of this definition relies on certain auxiliary facts from measure theory and can be found in (8).

Let us see how the aforementioned "elementary" definitions coincide with Definition 1.1.

Let e.g. $\mathcal{F}' = \{B, \bar{B}, \Omega, \emptyset\}$ and verify that

$$P(A|\mathcal{F}')(\omega) := E\big(I_A(\omega)|F'\big)(\omega) = \frac{P(A\cap B)}{P(B)} I_B(\omega) + \frac{P(A\cap \bar{B})}{P(\bar{B})} I_{\bar{B}}(\omega).$$

First let us check that this random variable is $\mathcal{F}'$ measurable. In fact any r.v. of the form $\xi(\omega) = \beta_1 I_B(\omega) + \beta_2 I_{\bar{B}}(\omega)$ is measurable w.r.t. $\mathcal{F}'$. Indeed (if e.g. $\beta_2 > \beta_1$)

$$\{\omega : \xi(\omega) \le x\} = \begin{cases} \Omega, & \max(\beta_1,\beta_2) < x \\ B & \min(\beta_1,\beta_2) \le x < \max(\beta_1,\beta_2) \\ \emptyset & x < \min(\beta_1,\beta_2) \end{cases} \in \mathcal{F}'.$$

Further

$$E\Big(I_A(\omega) - \frac{P(A\cap B)}{P(B)} I_B(\omega) - \frac{P(A\cap \bar{B})}{P(\bar{B})} I_{\bar{B}}(\omega)\Big) I_B(\omega) =$$

$$E\Big(I_A(\omega) I_B(\omega) - \frac{P(A\cap B)}{P(B)} I_B(\omega)\Big) = P(A\cap B) - \frac{P(A\cap B)}{P(B)} P(B) = 0$$

and similarly for the other atoms of $\mathcal{F}'$. Repeating these calculations for n atoms one obtains (1.2).

The formula (1.3) is checked in the same way: as before the right hand side is measurable w.r.t $\mathcal{F}'$ and

$$E\Big(X - E(X|\mathcal{D})\Big) I_{D_\ell}(\omega) = E\Big(X - \sum_{j=1}^m x_j \sum_{i=1}^n \frac{P(\{X(\omega) = x_j\}\cap D_i)}{P(D_i)} I_{D_i}(\omega)\Big) I_{D_\ell}(\omega)$$

$$E\Big(X I_{D_\ell}(\omega) - \sum_{j=1}^m x_j \frac{P(\{X(\omega) = x_j\}\cap D_\ell)}{P(D_\ell)} I_{D_\ell}(\omega)\Big) =$$

$$E\Big(\sum_{k=1}^m x_k I(X = x_k) I_{D_\ell}(\omega) - \sum_{j=1}^m x_j \frac{P(\{X(\omega) = x_j\}\cap D_\ell)}{P(D_\ell)} I_{D_\ell}(\omega)\Big) = 0.$$

The special important case is conditioning with respect to σ-algebra, generated by a r.v. i.e. minimal σ-algebra with respect to which this r.v. is measurable. For example if $X = I(\omega \le 1/2)$ is defined on $([0,1], \mathcal{B}, \lambda)$, then the σ-algebra, generated by X is atomic (P-a.s.) $\mathcal{F}_X = \{\Omega, \emptyset, A, \bar{A}\}$, where $A = (0, 1/2]$. The following definition is the special case of Definition 1.1

Definition 1.2. Let X and Y be a pair of r.v. and $E|X| < \infty$. The conditional expectation of X with respect to Y (or σ-algebra $\mathcal{F}_Y$ generated by Y), denoted by $E(X|Y)$ is an $\mathcal{F}_Y$-measurable r.v. such that

$$E\big(X - E(X|Y)\big)Z = 0$$

for any bounded $\mathcal{F}_Y$-measurable r.v. Z.

It turns out (see (8)) that any r.v. Z measurable w.r.t. $\mathcal{F}_Y$ has the form $Z = f(Y)$, for some function $f(x)$. So the Definition 1.2 is equivalent to

Proposition 1.3. *Let X and Y be a pair of r.v. and $E|X| < \infty$. The conditional expectation $E(X|Y)$ is given by $E(X|Y) = g(Y)$, where $g(x)$ is such that*

$$E\big(X - g(Y)\big)f(Y) = 0 \tag{1.4}$$

for any bounded function f.

In the view of the latter proposition, we will write $E(X|Y=y)$ to denote any function $g(y)$, satisfying (1.4)

Now let us check that Definition 1.2 (or 1.1) gives the the right answer in the coin tossing experiment with random probability for heads. On some probability space $(\Omega, \mathcal{F}, P)$ let π (the "random" heads probability) and ξ be a pair of i.i.d. r.v. with uniform distributions $U(0,1)$ For any $p \in [0,1]$ define $Z_p = I(\xi \le p)$. Clearly Z_p is a binary random variable with $P(Z=1)=p$. Consider $\zeta = Z_\pi$ and find $P(\zeta = 1|\pi) = E\big[I(\zeta=1)|\pi\big]$. For an arbitrary bounded function $f: \mathbb{R} \mapsto \mathbb{R}$

$$E\big[I(\zeta=1)-\pi\big]f(\pi) = E\big[I(\xi \le \pi)-\pi\big]f(\pi) = \int_0^1\int_0^1 \big[I(x \le y) - y\big]f(y)dydx =$$
$$\int_0^1 \Big[\int_0^1 I(x\le y)dx - y\big]f(y)dy = 0$$

and thus $E\big[I(\zeta=1)|\pi\big]=\pi$.

As mentioned before the Definition 1.1 is general and leads to familiar formulae under special setups. For example

Corollary 1.4. *Assume that two real r.v. X and Y have joint probability density*

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} P\big(X \le x, Y \le y\big).$$

Then for any bounded function $h: \mathbb{R} \mapsto \mathbb{R}$

$$E\big(h(X)|Y\big) = \frac{\int_{\mathbb{R}} h(x) f\big(x, Y(\omega)\big)dx}{\int_{\mathbb{R}} f\big(x, Y(\omega)\big)dx}, \tag{1.5}$$

or equivalently, the conditional probability distribution

$$dF(x;Y) := dP\big(X \le x|Y\big) = \frac{f\big(x,Y(\omega)\big)dx}{\int_{\mathbb{R}} f\big(x, Y(\omega)\big)dx}.$$

Proof. Clearly the right hand side of (1.5) is a function of Y and so only orthogonality should be checked: fix an arbitrary bounded function $\psi: \mathbb{R} \mapsto \mathbb{R}$

$$E\left(h(X) - \frac{\int_{\mathbb{R}} h(x) f\big(x, Y(\omega)\big)dx}{\int_{\mathbb{R}} f\big(x, Y(\omega)\big)dx}\right)\psi(Y) =$$
$$\iint_{\mathbb{R}^2}\left(h(s) - \frac{\int_{\mathbb{R}} h(x) f\big(x,t\big)dx}{\int_{\mathbb{R}} f\big(x,t\big)dx}\right)\psi(t) f(s,t)dsdt =$$
$$\iint_{\mathbb{R}^2} h(s)\psi(t)f(s,t)dsdt - \int_{\mathbb{R}} \frac{\int_{\mathbb{R}} h(x) f\big(x,t\big)dx}{\int_{\mathbb{R}} f\big(x,t\big)dx}\psi(t)\left[\int_{\mathbb{R}} f(s,t)ds\right]dt = 0$$

□

The formula (1.5) is a form of *the Bayes rule*, which is the central tool in calculation of conditional expectations. The Bayes rule is valid for the setups, much more general than in Corollary 1.4.

EXAMPLE 1.5. Suppose that a system malfunctions at τ seconds after it is powered, where τ is an exponential r.v., i.e.

$$P(\tau \le t) = \begin{cases} 1 - e^{-\lambda t}, & t \ge 0 \\ 0 & t < 0 \end{cases}$$

(1) What is the expected period of proper operation ?
(2) What is the probability of additional t seconds of operation, if the system already operates properly for s seconds ?
(3) What is the expected period of proper operation if the system already operates properly for s seconds ?

The expected period till failure is[1]

$$E\tau = E\int_0^\tau ds = E\int_0^\infty I(s \le \tau)ds = \int_0^\infty P(\tau \ge s)ds = \int_0^\infty \big[1 - P(\tau \le s)\big]ds = \int_0^\infty e^{-\lambda s}ds = \frac{1}{\lambda}$$

Further for $t \ge 0$

$$P(\tau \ge s + t|\tau \ge s) = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(\tau \ge t).$$

The latter is referred as *memoryless* property of the r.v. τ: the fact that the system works properly for s seconds already does not change the probability of proper operation during next t seconds ! It turns out that this property characterizes exponential distribution, i.e. only the latter is memoryless among all distributions with density.

Also

$$E(\tau|\tau \ge s) = E\left(\int_0^\infty I(\tau \ge u)du\Big|\tau \ge s\right) = \int_0^s 1du + \int_s^\infty P(\tau \ge u|\tau \ge s)du = s + \int_0^\infty P(\tau \ge u')du' = s + 1/\lambda$$

i.e. the additional expected time till the first failure is $1/\lambda$.

1.1. Properties of conditional expectation. The conditional expectation satisfies the following properties [2] (below X, Y and Z are random variables, for which the expectations are assumed to exists, when needed; generic bounded functions are denoted by g, f and h)

a. $E(C|Y) = C$ for constant C

PROOF. Constant r.v. C is certainly a function of Y and thus the claim is true by definition of the conditional expectation. □

b. If $X \le Y$ (P-a.s.) then $E\big(X|Z\big) \le E\big(Y|Z\big)$ P-a.s.

[1] The same answer is obtained by integrating vs. probability density $\lambda e^{-\lambda t}$.

[2] The conditioning w.r.t. random variables is considered here - conditioning in the general case (i.e. w.r.t. σ-algebras) is treated similarly

PROOF. For a positive function f

$$Y \geq X \quad \Longrightarrow \quad E\big(Y - X\big)f(Z) \geq 0$$

By definition

$$0 \leq E\big(Y - X\big)f(Z) = EE(Y|Z)f(Z) - EE(X|Z)f(Z) = \\ E\big(E(Y|Z) - E(X|Z)\big)f(Z)$$

In particular this holds with $f(Z) = I\big(E(Y|Z) - E(X|Z) \leq 0\big)$, i.e.

$$E\big(E(Y|Z) - E(X|Z)\big)I\big(E(Y|Z) - E(X|Z) \leq 0\big) \geq 0$$

which implies $P\big(E(Y|Z) - E(X|Z) \leq 0\big) = 0$. □

c. $\big|E(X|Y)\big| \leq E\big(|X|\big|Y\big)$

PROOF. Since $-|X| \leq X \leq |X|$, by **(b)**

$$-E\big(|X|\big|Y\big) \leq E\big(X\big|Y\big) \leq E\big(|X|\big|Y\big)$$

that is

$$\big|E\big(X\big|Y\big)\big| \leq E\big(|X|\big|Y\big).$$

□

d. For constants a and b, $E(aX + bY|Z) = aE(X|Z) + bE(Y|Z)$.

PROOF.

$$E\big(aX + bY - aE(X|Z) - bE(Y|Z)\big)f(Z) = \\ aE\big(X - E(X|Z)\big)f(Z) + bE\big(Y - E(Y|Z)\big)f(Z) = 0.$$

□

e. $E(g(X)|X) = g(X)$ P-a.s. for any g

PROOF. Holds by definition since for any f

$$E(g(X) - g(X))f(X) = 0.$$

□

f. $E\big(E(X|Y)\big) = EX$

PROOF. By definition

$$E\big(X - E(X|Y)\big) \cdot 1 = 0.$$

□

g. $E\big(E(X|Y,Z)\big|Z\big) = E(X|Z)$ (*"smoothing" property*)

PROOF. For any f

$$0 = E\Big[E(X|Y,Z) - E\big(E(X|Y,Z)\big|Z\big)\Big]f(Z) =$$

$$E\Big[E(X|Y,Z) - X\Big]f(Z) + E\Big[X - E\big(E(X|Y,Z)\big|Z\big)\Big]f(Z)$$

$$= E\Big[X - E\big(E(X|Y,Z)\big|Z\big)\Big]f(Z)$$

where the latter equality holds since $f(Z)$ is in particular a function of (Y,Z). □

REMARK 1.6. In σ-algebras terms, the above property reads: if[3] $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then $E(X|\mathcal{G}_1) = E\big(E(X|\mathcal{G}_2)\big|\mathcal{G}_1\big)$ P-a.s. In particular this is valid with $\mathcal{G}_1$ generated by $\{Y_1, ..., Y_m\}$ and $\mathcal{G}_2$ by $\{Y_1, ..., Y_m, Y_{m+1}, ..., Y_n\}$.

h. $E\big(E(X|Y)\big|Y,Z\big) = E(X|Y)$ P-a.s.

PROOF. particular case of **(e).** □

i. If X and Y are independent, then $E(X|Y) = EX$.

PROOF. Being a constant, EX is a trivial function of Y. Moreover

$$E\big(X - EX\big)f(Y) = EXEf(Y) - EXEf(Y) = 0.$$

□

j. $E(Xg(Y)|Y) = g(Y)E(X|Y)$ for bounded g

PROOF.

$$E\big(Xg(Y) - g(Y)E(X|Y)\big)f(Y) = E\big(X - E(X|Y)\big)g(Y)f(Y) = 0$$

□

REMARK 1.7. The latter property holds also when only $E|g(Y)| < \infty$ and $E|Xg(Y)| < \infty$.

k. If $g(x)$ is convex, then $E\big(g(X)|Y\big) \geq g\big(E(X|Y)\big)$ P-a.s. (Jensen inequality)

PROOF. For any $x_0 \in \mathbb{R}$, $g(x) \geq g(x_0) + \alpha(x - x_0)$ for some $\alpha \in \mathbb{R}$. Then

$$g(X) \geq g\big(E(X|Y)\big) + \alpha\big(X - E(X|Y)\big)$$

and by properties **(b)** and **(e)**

$$E\big(g(X)|Y\big) \geq g\big(E(X|Y)\big).$$

□

l. Assume that $EX^2 < \infty$, then $E(X|Y)$ is the *optimal* estimate of X given Y, in the sense

$$E\big(X - E(X|Y)\big)^2 = \inf_{\psi} E\big(X - \psi(Y)\big)^2$$

where the infimum is taken among all functions such that $E\psi^2(Y) < \infty$.

[3] Note that inclusion in $\mathcal{G}_1 \subseteq \mathcal{G}_2$ means that any event from $\mathcal{G}_1$ is also an event from $\mathcal{G}_2$

Proof. By virtue of property **(j)** (see Remark 1.7) and the definition of conditional expectation

$$E\big(X-\psi(Y)\big)^2 = E\big(X-E(X|Y)+E(X|Y)-\psi(Y)\big)^2 = \\ E\big(X-E(X|Y)\big)^2 + E\big(E(X|Y)-\psi(Y)\big)^2 \ge E\big(X-E(X|Y)\big)^2$$

□

1.2. Additional results, examples and applications.

1.2.1. *Gaussian processes.* The random vector X in $\mathbb{R}^n$ is Gaussian if its characteristic function has the form

$$\varphi(\lambda) = Ee^{i\lambda^* X} = \exp\big\{i\lambda^* m - \frac{1}{2}\lambda^*\Gamma\lambda\big\}, \quad \forall \lambda \in \mathbb{R}^n \tag{1.6}$$

where m and Γ are a vector and a nonnegative definite matrix.

It is not hard to check that $m = EX$ and $\Gamma = \operatorname{cov}(X)$. Moreover if Γ is nonsingular the distribution of X has the density

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det\Gamma}} \exp\big\{-1/2(x-m)^*\Gamma^{-1}(x-m)\big\}.$$

If Γ is singular the density does not exist: e.g. according to the above definition a constant r.v. is Gaussian as well.

The next theorem gives a number of important properties of the Gaussian r.v.

Theorem 1.8. *Let $(X,Y)\in\mathbb{R}^{m+n}$ be a Gaussian random vector, then*

(1) *X and Y are independent if and only if they are orthogonal (uncorrelated)*
(2) *Any linear combination $Z = AX + b$ is a Gaussian random vector with mean $EZ = AEX + b$ and covariance $\operatorname{cov}(Z) = A\operatorname{cov}(X)A^*$.*
(3) *$E(X|Y) = \widehat{E}(X|Y)$ and moreover the conditional distribution $P(X\in B|Y)$, $B\subset\mathcal{B}(\mathbb{R}^d)$ is P-a.s. Gaussian with the mean*

$$E(X|Y) = EX + \operatorname{cov}(X,Y)\operatorname{cov}^{\oplus}(Y)\big(Y-EY\big) \tag{1.7}$$

and (deterministic!) covariance

$$\operatorname{cov}(X|Y) = \operatorname{cov}(X) - \operatorname{cov}(X,Y)\operatorname{cov}^{\oplus}(Y)\operatorname{cov}(Y,X). \tag{1.8}$$

Proof.

(1) For Gaussian vectors $E\|X\|^p<\infty$ and $E\|Y\|^p<\infty$ for any p, so independence implies orthogonality. Conversely, if X and Y are orthogonal, i.e. $\operatorname{cov}(X,Y)=0$, then Γ is block diagonal with block matrices $\Gamma_{11}=\operatorname{cov}(X)$ and $\Gamma_{22}=\operatorname{cov}(Y)$. Then

$$\varphi_{n+m}(\lambda) = \exp\{i\lambda^*\mu - 1/2\lambda^*\Gamma\lambda\} = \\ \exp\{i\lambda_1^*\mu_1 - 1/2\lambda_1^*\Gamma_{11}\lambda_1\}\exp\{i\lambda_2^*\mu_2 - 1/2\lambda_2^*\Gamma_{22}\lambda_2\} = \varphi_m(\lambda_1)\varphi_n(\lambda_2)$$

where $\mu_1 = EX$, $\mu_2 = EY$ and $\varphi_n(\cdot)$ denotes Gaussian characteristic function of an $n\times 1$ vector and λ_1 and λ_2 are $m\times 1$ and $n\times 1$ vectors.

(2)

$$\psi(\lambda) = E\Big(e^{i\lambda^* Z}\Big) = E\exp\{i\lambda^* b + i\lambda^* AX\} = \exp\{i\lambda^* b\}E\exp\{i\lambda^* AX\} = \\ \exp\{i\lambda^*\mu - 1/2\lambda^* G\lambda\}$$

where $\mu = b + AEX$ and $G = A\,\mathrm{cov}(X)A^*$. The latter is a nonnegative definite matrix and thus the characteristic function of Z corresponds to a Gaussian distribution with appropriate parameters.
(3) It is to be shown that

$$\varphi(\lambda; Y) = E\Big(e^{i\lambda^* X}\big|Y\Big) = \exp\{i\lambda^*\mu(Y) - 1/2\lambda^*\Gamma(Y)\lambda\}$$

where $\mu(Y)$ is given by (1.7) and $\Gamma(Y)$ by (1.8).

Recall that X can be decomposed into

$$X = \widehat{E}(X|Y) + \big(X - \widehat{E}(X|Y)\big)$$

where $\big(X - \widehat{E}(X|Y)\big)$ is orthogonal to Y (and $\widehat{E}(X|Y)$). Since $\widehat{E}(X|Y)$ is a linear map of Y, the vector $(X, Y, \widehat{E}(X|Y)$ is Gaussian as well by virtue of (2). Hence $(\big(X - \widehat{E}(X|Y)\big), Y)$ is Gaussian and since $\big(X - \widehat{E}(X|Y)\big)$ and Y are orthogonal, they are also independent. So for an arbitrary bounded function h

$$E\big(X - \widehat{E}(X|Y)\big)h(Y) = E\big(X - \widehat{E}(X|Y)\big)Eh(Y) = 0$$

i.e. $\widehat{E}(X|Y) = E(X|Y)$. The equation 1.7 follows immediately.

Since $X - \widehat{E}(X|Y)$ and Y are independent we have

$$E\Big(\exp\{i\lambda^* X\}|Y\Big) = \exp\{i\lambda^*\widehat{E}(X|Y)\}E\Big(\exp\{i\lambda^*(X - \widehat{E}(X|Y)\}|Y\Big) =$$
$$\exp\{i\lambda^*\widehat{E}(X|Y)\}E\Big(\exp\{i\lambda^*(X - \widehat{E}(X|Y)\}\Big) =$$
$$\exp\{i\lambda^*\widehat{E}(X|Y)\}\exp\{-1/2\lambda^*\,\mathrm{cov}(X - \widehat{E}(X|Y))\lambda\}$$

and the desired result follows. □

The r.p. $X = (X_n)_{n\ge 0}$ is Gaussian, if all its finite dimensional distributions are Gaussian and consistent. For example, the sequence generated by the recursion $(n \ge 1)$

$$X_n = a_n X_{n-1} + b_n\varepsilon_n \tag{1.9}$$

with constants a_n and b_n and standard i.i.d. Gaussian sequence $\varepsilon = (\varepsilon_n)_{n\ge 0}$ is a Gaussian Markov process.

It turns out that any zero mean Gaussian Markov process has the structure of (1.9):

Lemma 1.9. *Let $X = (X_n)_{n\ge 0}$ be a Gaussian Markov process. Then there exist deterministic sequences $(a_n)_{n\ge 1}$ and $(b_n)_{n\ge 1}$, such that for $n \ge 1$*

$$X_n = a_n X_{n-1} + b_n\varepsilon_n$$

where ε is a standard Gaussian i.i.d. sequence.

Proof. Let $X_0^n = \sigma\{X_0, X_1, ..., X_n\}$ and consider the decomposition

$$X_n = E(X_n|X_0^{n-1}) + \big(X_n - E(X_n|X_0^{n-1})\big)$$

Since the process is Markov $E(X_n|X_0^{n-1}) = E(X_n|X_{n-1})$ and since it is Gaussian $E(X_n|X_{n-1}) = a_n X_{n-1}$ for some constant a_n. Now set

$$b_n = \sqrt{E(X_n - E(X_n|X_{n-1}))^2}$$

and $\varepsilon_n = \big(X_n - E(X_n|X_0^{n-1})\big)/b_n$, so that

$$X_n = a_n X_{n-1} + b_n\varepsilon_n.$$

It is left to check that ε_n is an i.i.d. Gaussian sequence. First note that being a linear transformation of Gaussian r.v. $\varepsilon = (\varepsilon_n)_{n\ge 1}$ is a Gaussian process. Also e.g. for $n > m$

$$E\varepsilon_n\varepsilon_m = b_n^{-1}b_m^{-1}E(X_n - E(X_n|X_0^{n-1}))(X_m - E(X_m|X_0^{m-1})) = 0$$

and thus ε_n and ε_m are independent. Similarly independence can be verified for any number of entries of ε. Since $\mathrm{cov}(\varepsilon_n) = 1$ and the process is Gaussian, the sequence ε is a standard i.i.d. Gaussian sequence. □

COROLLARY 1.10. *Any Gaussian Markov stationary process has the correlation function of the form*

$$R(n) \propto a^{|k|}$$

with some $|a| < 1$.

Indeed, due to representation of Lemma (1.9) for any $n \ge 1$

$$R(n-1,n) = EX_nX_{n-1} = a_nEX_{n-1}^2 = a_nR(n-1,n-1).$$

Since the process is stationary $R(1) = R(n-1,n) = a_nR(n-1,n-1) = a_nR(0)$. Assuming $R(0) > 0$ (otherwise the claim is trivial!) we obtain $a_n = R(1)/R(0) =$ const for all $n \ge 1$. This implies that X_n satisfies

$$X_n = aX_{n-1} + b_n\varepsilon_n,$$

which implies that $\mathrm{cov}(X_n) = a^2\,\mathrm{cov}(X_{n-1}) + b_n^2$ for any $n \ge 1$ and thus $b_n \equiv b$, i.e. b_n is constant as well. The correlation function of the stationary process satisfying the recursion with constant coefficients is $b^2/(1-a^2)a^{|k|}$ if $|a| < 1$. If $|a| \ge 1$ the process can not be stationary (e.g. its variance grows).

2. Nonlinear estimation

In chapter 2 the linear estimation problem was addressed: the optimal estimate was constrained to be a linear transformation of the observation process. This constraint may be quite restrictive in many problems, i.e. the accuracy can be significantly improved if the nonlinear estimates are considered. In the next section we address the simplest nonlinear filtering problem.

2.1. Filtering of Markov chains.

2.1.1. *Finite state Markov chains.* The notion of Markov process was introduced in the section 5.1.4. Markov processes in discrete time are often called Markov chains. The simplest Markov chain can be constructed in the following way: let $\mathbb{S} = \{a_1, ..., a_d\}$ be a finite set of real numbers (symbols); p_0 be some probability distribution on $\mathbb{S}$ (which can be identified with a vector from the simplex $\mathcal{S} = \{x \in \mathbb{R}^d : x_i \ge 0, \sum_{i=1}^d x_i = 1\}$) and *transition probabilities* $\lambda_{ij} \in [0,1]$, $1 \le i,j \le d$, $\sum_{j=1}^d \lambda_{ij} = 1$, $i = 1, ..., d$.

Now let X_0 be a random variable with values in $\mathbb{S}$ and distribution p_0. Define X_n recursively as a random variable

$$X_n = \sum_{i=1}^d \varepsilon_n(i)I(X_{n-1} = a_i)$$

where $(\varepsilon_n)_{n\ge 1}$ is an i.i.d. vector sequence with $P(\varepsilon_1(i) = a_j) = \lambda_{ij}$, $j = 1, ..., d$.

By construction the process $X = (X_n)_{n\ge 0}$ is Markov. Denote by p_n its marginal distribution at time n, i.e. $p_n(i) = P(X_n = a_i)$. and let Λ be the matrix with entries λ_{ij}. Then (why?)

$$p_n = \Lambda^* p_{n-1} = \Lambda^{*n} p_0. \tag{2.1}$$

The chain X is said to be *ergodic* if it converges weakly (in law) to a random variable with positive distribution on $\mathbb{S}$, i.e. if $p_n \xrightarrow{n\to\infty} \mu$ with $\mu(i) > 0$. It can be shown that X is ergodic if and only if its transition matrix Λ is q-primitive, i.e. Λ^q has positive entries for some $q \ge 1$. Finite state Markov chains has been the subject of research since 30's. For further reading see (8) and the references therein.

2.1.2. *Filtering in the Hidden Markov Models (HMM).* Consider the signal / observation pair (X, Y), where $X = (X_n)_{n\ge 0}$ is a finite state Markov chain with $(\mathbb{S}, \Lambda, p_0)$ and $Y = (Y_n)_{n\ge 1}$ is generated by

$$Y_n = h(X_n) + \xi_n$$

where h is $\mathbb{S} \mapsto \mathbb{R}$ function and $\xi = (\xi_n)_{n\ge 1}$ is an i.i.d. sequence with ξ_1 having density $f(x)$. This setting is often referred as Hidden Markov Model (HMM). One of the basic problems in HMM is to estimate the state X_n from the observations $Y_1^n = \sigma\{Y_1, ..., Y_n\}$. Similarly to the Kalman filter, the solution can be given in the efficient recursive form

Theorem 2.1. *For any $\varphi : \mathbb{S} \mapsto \mathbb{R}$, $E(\varphi(X_n)|Y_1^n) = \sum_{i=1}^d \varphi(a_i)\pi_n(i)$, where $\pi_n(i) = P(X_n = a_i|Y_1^n)$. The vector process $\pi = (\pi_n)_{n\ge 1}$ is generated by*

$$\pi_n = \frac{D(Y_n)\Lambda^* \pi_{n-1}}{\langle 1, D(Y_n)\Lambda^* \pi_{n-1}\rangle}, \quad n \ge 1 \tag{2.2}$$

subject to $\pi_0 = p_0$, where $\langle 1, x\rangle = \sum_{i=1}^d x_i$, $x \in \mathbb{R}^d$ and $D(y)$, $y \in \mathbb{R}$ is a diagonal matrix with entries $f(y - h(a_i))$, $i = 1, ..., d$.

Proof. First note that

$$E(\varphi(X_n)|Y_1^n) = E\big(\sum_{i=1}^d \varphi(a_i) I(X_n = a_i)|Y_1^n\big) = \sum_{i=1}^d \varphi(a_i) P(X_n = a_i|Y_1^n).$$

Let $G^i : \mathbb{R} \times \mathbb{R}^{n-1} \mapsto \mathbb{R}$ be a function such that $P(X_n = a_i|Y_1^n) = G^i(Y_n; Y_1^{n-1})$. Then G^i should satisfy

$$E\big(I(X_n = a_i) - G^i(Y_n; Y_1^{n-1})\big)\psi(Y_n)\Psi(Y_1^{n-1}) = 0 \tag{2.3}$$

for any bounded $\psi : \mathbb{R} \mapsto \mathbb{R}$ and $\Psi : \mathbb{R}^{n-1} \mapsto \mathbb{R}$. Note that no generality is lost if only the functions of the form $\psi(y)\Psi(y_1^{n-1})$ are considered, since any bounded function of n variables can be approximated by series of the products of functions of single variable (e.g. indicators). Note also that (2.3) holds if

$$E\Big(\big(I(X_n = a_i) - G^i(Y_n; Y_1^{n-1})\big)\psi(Y_n)|Y_1^{n-1}\Big) = 0, \quad P - a.s.$$

is satisfied. Further

$$
\begin{aligned}
&E\big(I(X_n = a_i)\psi(Y_n)|Y_1^{n-1}\big) = E\big(I(X_n = a_i)\psi(h(a_i) + \xi_n)|Y_1^{n-1}\big) = \\
&\int_{\mathbb{R}} \psi(h(a_i) + s)f(s)dsE\big(I(X_n = a_i)|Y_1^{n-1}\big) = \\
&\int_{\mathbb{R}} \psi(s)f(s - h(a_i))dsE\Big(E\big(I(X_n = a_i)|X_{n-1}, Y_1^{n-1}\big)|Y_1^{n-1}\Big) = \\
&\int_{\mathbb{R}} \psi(s)f(s - h(a_i))dsE\Big(E\big(I(X_n = a_i)|X_{n-1}\big)|Y_1^{n-1}\Big) = \\
&\int_{\mathbb{R}} \psi(s)f(s - h(a_i))dsE\Big(\sum_{j=1}^{d} \lambda_{ji} I(X_{n-1} = a_j)|Y_1^{n-1}\Big) = \\
&\int_{\mathbb{R}} \psi(s)f(s - h(a_i))ds\sum_{j=1}^{d} \lambda_{ji}\pi_{n-1}(j)
\end{aligned}
$$

and similarly

$$
\begin{aligned}
&E\big(G^i(Y_n; Y_1^{n-1})\psi(Y_n)|Y_1^{n-1}\big) = \\
&E\big(\sum_{\ell=1}^{d} I(X_n = a_\ell)G^i(h(a_\ell) + \xi_n; Y_1^{n-1})\psi(h(a_\ell) + \xi_n)|Y_1^{n-1}\big) = \\
&\sum_{\ell=1}^{d} \int_{\mathbb{R}} G^i(h(a_\ell) + s; Y_1^{n-1})\psi(h(a_\ell) + s)f(s)dsE\Big(I(X_n = a_\ell)|Y_1^{n-1}\Big) = \\
&\sum_{\ell=1}^{d} \int_{\mathbb{R}} G^i(s; Y_1^{n-1})\psi(s)f(s - h(a_\ell))dsE\Big(I(X_n = a_\ell)|Y_1^{n-1}\Big) = \\
&\sum_{\ell=1}^{d} \int_{\mathbb{R}} G^i(s; Y_1^{n-1})\psi(s)f(s - h(a_\ell))ds\sum_{j=1}^{d} \lambda_{j\ell}\pi_{n-1}(j).
\end{aligned}
$$

By arbitrariness of ψ

$$
G^i(s; Y_1^{n-1}) = \frac{f(s - h(a_i))\sum_{j=1}^{d} \lambda_{ji}\pi_{n-1}(j)}{\sum_{\ell=1}^{d} f(s - h(a_\ell))\sum_{j=1}^{d} \lambda_{j\ell}\pi_{n-1}(j)}
$$

which proves the recursion (2.2). □

EXAMPLE 2.2. *Binary Symmetric Channel*

The message X_n is a binary Markov chain with

$$
P\{X_n = 1|X_{n-1} = 0\} = P\{X_n = 0|X_{n-1} = 1\} = \lambda
$$

with the initial distribution $P\{X_0 = 1\} = p$. It is transmitted via symmetric channel, so that the observed sequence is ($n \geq 1$)

$$
Y_n = X_n \oplus \xi_n = (X_n + \xi_n) \mod 2
$$

where ξ_n is an i.i.d. binary noise sequence $P\{\xi_n = 1\} = \varepsilon$. Find the recursive filtering estimate of X_n from Y_1^n.

First note that the transition matrix of X_n is

$$
\Lambda = \begin{pmatrix} 1 - \lambda & \lambda \\ \lambda & 1 - \lambda \end{pmatrix}
$$

and the conditional distribution of Y_n, given X_n is

$$f(i,j) := P\{Y_n = i | X_n = j\} = \begin{cases} 1-\varepsilon, & i=j \\ \varepsilon, & i \neq j \end{cases}, \quad i,j \in \{0,1\}$$

so that $f(Y_n,1) = (1-2\varepsilon)Y_n + \varepsilon$ and $f(Y_n,0) = (1-2\varepsilon)(1-Y_n)+\varepsilon$. Let $\pi_n = P\{X_n = 1|Y_1^n\}$. Slightly modifying the proof of Theorem 2.1, the recursion is obtained:

$$\pi_n = \frac{\big[(1-2\varepsilon)Y_n+\varepsilon\big]\pi_{n|n-1}}{\big[(1-2\varepsilon)Y_n+\varepsilon\big]\pi_{n|n-1} + \big[(1-2\varepsilon)(1-Y_n)+\varepsilon\big]\big(1-\pi_{n|n-1}\big)} \tag{2.4}$$

where

$$\pi_{n|n-1} = \lambda(1-\pi_{n-1}) + (1-\lambda)\pi_{n-1}.$$

The recursion (2.4) can be rewritten as:

$$\begin{aligned} \pi_n \quad = \quad & \frac{(1-\varepsilon)\pi_{n|n-1}}{(1-\varepsilon)\pi_{n|n-1} + \varepsilon\big(1-\pi_{n|n-1}\big)} Y_n + \\ & + \frac{\varepsilon\pi_{n|n-1}}{\varepsilon\pi_{n|n-1} + (1-\varepsilon)\big(1-\pi_{n|n-1}\big)}(1-Y_n) \end{aligned} \tag{2.5}$$

□

The optimal *linear* estimate $\widehat{E}(\varphi(X_n)|Y_1^n)$ can also be efficiently calculated via the Kalman filter:

THEOREM 2.3. *Assume that ξ is an i.i.d. sequence with $\sigma^2 = E\xi_1^2 < \infty$ and $E\xi_1 = 0$. Then for any $\varphi : \mathbb{S} \mapsto \mathbb{R}$, $\widehat{E}(\varphi(X_n)|Y_1^n) = \sum_{i=1}^d \varphi(a_i)\eta_n(i)$, where $\eta_n(i) = \widehat{E}(I(X_n = a_i)|Y_1^n)$. The vector process $\eta = (\eta_n)_{n\ge1}$ is generated by*

$$\eta_n = \Lambda^*\eta_{n-1} + \frac{\big(\Lambda^* P_{n-1}\Lambda + V_n\big)h}{h^*\Lambda^*P_{n-1}\Lambda h + h^*V_n h + \sigma^2}\big(Y_n - h^*\Lambda^*\eta_{n-1}\big) \tag{2.6}$$

$$P_n = \Lambda^*P_{n-1}\Lambda + V_n - \frac{\big(\Lambda^*P_{n-1}\Lambda + V_n\big)hh^*\big(\Lambda^*P_{n-1}\Lambda + V_n\big)}{h^*\Lambda^*P_{n-1}\Lambda h + h^*V_n h + \sigma^2} \tag{2.7}$$

$$V_n = \operatorname{diag}(p_n) - \Lambda^* \operatorname{diag}(p_{n-1})\Lambda \tag{2.8}$$

subject to $\eta_0 = p_0$ and $V_0 = P_0 = \operatorname{diag}(p_0) - p_0 p_0^*$, where h is a vector with entries $h(i)$ and p_n satisfies (2.1).

PROOF. Introduce the random vector process

$$I_n = \begin{pmatrix} I(X_n = a_1) \\ \vdots \\ I(X_n = a_d) \end{pmatrix}$$

and let $p_n = EI_n$, i.e. $p_n(i) = P(X_n = a_i)$.

Note that $\widehat{E}(\varphi(X_n)|Y_1^n) = \widehat{E}(\langle\varphi, I_n\rangle|Y_1^n) = \langle\varphi, \eta_n\rangle$, where $\langle x,y\rangle = \sum_{i=1}^d x_i y_i$, $x,y \in \mathbb{R}^d$ and the φ is vector in $\mathbb{R}^d$ with entries $\varphi(a_i)$.

Let $\varepsilon = (\varepsilon_n)_{n\ge1}$ be the vector random process, satisfying

$$\varepsilon_n = I_n - \Lambda^* I_{n-1}.$$

Then $E\varepsilon_n = p_n - \Lambda^* p_{n-1} = 0$ (see (2.1)) and

$$\begin{aligned}
&\operatorname{cov}(\varepsilon_n) = E\big(I_n - \Lambda^* I_{n-1}\big)\big(I_n - \Lambda^* I_{n-1}\big)^* = \\
&E\big(I_n - \Lambda^* I_{n-1}\big)I_n^* - E\big(I_n - \Lambda^* I_{n-1}\big)I_{n-1}^*\Lambda = \\
&E\,\mathrm{diag}(I_n) - E\Lambda^* I_{n-1}E(I_n^*|I_{n-1}) - E\big(E(I_n|I_{n-1}) - \Lambda^* I_{n-1}\big)I_{n-1}^*\Lambda = \\
&\mathrm{diag}(p_n) - E\Lambda^* I_{n-1}I_{n-1}^*\Lambda = \mathrm{diag}(p_n) - \Lambda^*\,\mathrm{diag}(p_{n-1})\Lambda := V_n
\end{aligned}$$

Moreover for $m \neq n$ (say $n > m$)

$$\begin{aligned}
E\varepsilon_n\varepsilon_m^* &= E(I_n - \Lambda^* I_{n-1})(I_m - \Lambda^* I_{m-1})^* = \\
&\qquad E(E(I_n|X_0^{n-1}) - \Lambda^* I_{n-1})(I_m - \Lambda^* I_{m-1})^* = 0
\end{aligned}$$

The equations (2.6)-(2.7) are nothing but the Kalman filter, corresponding to the system

$$\begin{aligned}
&I_n = \Lambda^* I_{n-1} + \varepsilon_n \\
&Y_n = h^* I_n + \xi_n = h^*\Lambda^* I_{n-1} + h^*\varepsilon_n + \xi_n.
\end{aligned}$$

□

EXERCISE 2.4. *Apply the equations of Theorem 2.3 to the model in Example 2.2.*

In general the mean square error of the nonlinear filter (2.2) is strictly less than of the filter (2.6). However the equations (2.6)-(2.7) have their pros: the corresponding filtering error is obtained via P_n, whereas the performance of (2.2) can not be assessed in a simple way. Also their stability properties are better known.

2.1.3. *Filtering of the occupation times and transitions counters.* In many HMM applications it is required to calculate the estimates of certain processes related to the signal X. For example, the efficient calculation of the transition probabilities estimates is based (see (10) for details) on the filtering estimates of the *occupation times*

$$\theta_n^i = \sum_{k=0}^n I(X_k = a_i), \quad i = 1, ..., d$$

and the *number of i-to-j transitions*

$$N_n^{ij} = \sum_{k=1}^n I(X_k = a_j)I(X_{k-1} = a_i).$$

The particularly efficient algorithms for calculation of $\widetilde{\theta}_n^i = E\big(\theta_n^i|Y_1^n\big)$ and $\widetilde{N}_n^{ij} = E\big(N_n^{ij}|Y_1^n\big)$ were derived by the authors of (10).

Occupation time

First note that θ_n^r obeys the recursion:

$$\theta_n^r = \theta_{n-1}^r + I(X_n = a_r)$$

Introduce the vector $Z_n := \theta_n^r I_n$, where I_n is the vector of indicators as in the proof of Theorem 2.3. Clearly

$$\begin{aligned}
Z_n &:= \theta_n^r I_n = \big[\theta_{n-1}^r + I_n(r)\big]I_n = \theta_{n-1}^r I_n + e_r I_n(r) = \\
&= \theta_{n-1}^r\big[\Lambda^* I_{n-1} + I_n - \Lambda^* I_{n-1}\big] + e_r I_n(r) = \\
&= \Lambda^* Z_{n-1} + \theta_{n-1}^r\varepsilon_n + e_r I_n(r) \qquad (2.9)
\end{aligned}$$

where e_r is the r-th vector from the standard basis of $\mathbb{R}^d$ and $\varepsilon_n := I_n - \Lambda^* I_{n-1}$.

Conditioning both sides of (2.9) with respect to Y_1^{n-1} we arrive at

$$\widetilde{Z}_{n|n-1} := E(Z_n|Y_1^{n-1}) = \Lambda^* \widetilde{Z}_{n-1} + E\big(\theta^r_{n-1}\varepsilon_n|Y_1^{n-1}\big) + e_r \pi_{n|n-1}(r) \qquad (2.10)$$

The mid term in (2.10) vanishes:

$$\begin{aligned} E\big(\theta^r_{n-1}\varepsilon_n|Y_1^{n-1}\big) &= E\Big\{E\big(\theta^r_{n-1}\varepsilon_n|I_0^{n-1}, Y_1^{n-1}\big)\big|Y_1^{n-1}\Big\} = \\ &= E\Big\{\theta^r_{n-1}E\big(I_n - \Lambda^* I_{n-1})|I_{n-1}\big)\big|Y_1^{n-1}\Big\} \equiv 0 \end{aligned}$$

so that

$$\widetilde{Z}_{n|n-1} = \Lambda^* \widetilde{Z}_{n-1} + e_r\langle e_r, \Lambda^*\pi_{n-1}\rangle \qquad (2.11)$$

Let us find $\widetilde{Z}_n = E(Z_n|Y_1^n)$ in terms of Y_n and $\widetilde{Z}_{n|n-1}$. Set

$$\widetilde{Z}_n = G(Y_n; Y_1^{n-1})$$

For any bounded $\psi : \mathbb{R} \mapsto \mathbb{R}$

$$E\Big\{\big(Z_n - G(Y_n; Y_1^{n-1})\big)\psi(Y_n)\big|Y_1^{n-1}\Big\} = 0$$

Calculating the first term componentwise:

$$\begin{aligned} E(Z_n(i)\psi(Y_n)|Y_1^{n-1}) &= E\big(\theta^r_n I(X_n = a_i)\psi(h(a_i) + \xi_n)|Y_1^{n-1}\big) = \\ &= E\big(\theta^r_n I(X_n = a_i)|Y_1^{n-1}\big)\int \psi(h(a_i) + x)f(x)dx = \\ &= \widetilde{Z}_{n|n-1}(i)\int \psi(x)f(x - h(a_i))dx \qquad (2.12) \end{aligned}$$

Analogously:

$$E\big(G(Y_n; Y_1^{n-1})\psi(Y_n)|Y_1^{n-1}\big) = \\ \sum_i \pi_{n|n-1}(i)\int G(x; Y_1^{n-1})\psi(x)f(x - h(a_i))dx \qquad (2.13)$$

The eq. (2.12), (2.13) and $\pi_{n|n-1} = \Lambda^*\pi_{n-1}$ lead to:

$$\widetilde{Z}_n = \frac{D_n \widetilde{Z}_{n|n-1}}{\langle 1, D_n\Lambda^*\pi_{n-1}\rangle} \qquad (2.14)$$

where D_n is a diagonal matrix with the elements $f(Y_n - h(a_i))$.

Finally combining (2.14) and (2.11) we arrive at the filtering equations:

$$\begin{aligned} \widetilde{Z}_n &= \frac{D_n\Lambda^*\widetilde{Z}_{n-1}}{\langle 1, D_n\Lambda^*\pi_{n-1}\rangle} + \frac{D_n e_r\langle e_r, \Lambda^*\pi_{n-1}\rangle}{\langle 1, D_n\Lambda^*\pi_{n-1}\rangle}, \quad t \ge 1 \qquad (2.15) \\ \widetilde{Z}_0 &= E\theta^r_0 I_0 = EI(X_0 = a_r)I_0 = e_r p_0(r) \end{aligned}$$

where π_n is generated by (2.2).

To recover the optimal estimate of θ^r_n recall that $\langle 1, I_n\rangle \equiv 1$, so that:

$$\widetilde{\theta}^r_n = E(\theta^r_n\langle 1, I_n\rangle|Y_1^n) = \langle 1, E(\theta^r_n I_n|Y_1^n)\rangle = \langle 1, \widetilde{Z}_n\rangle$$

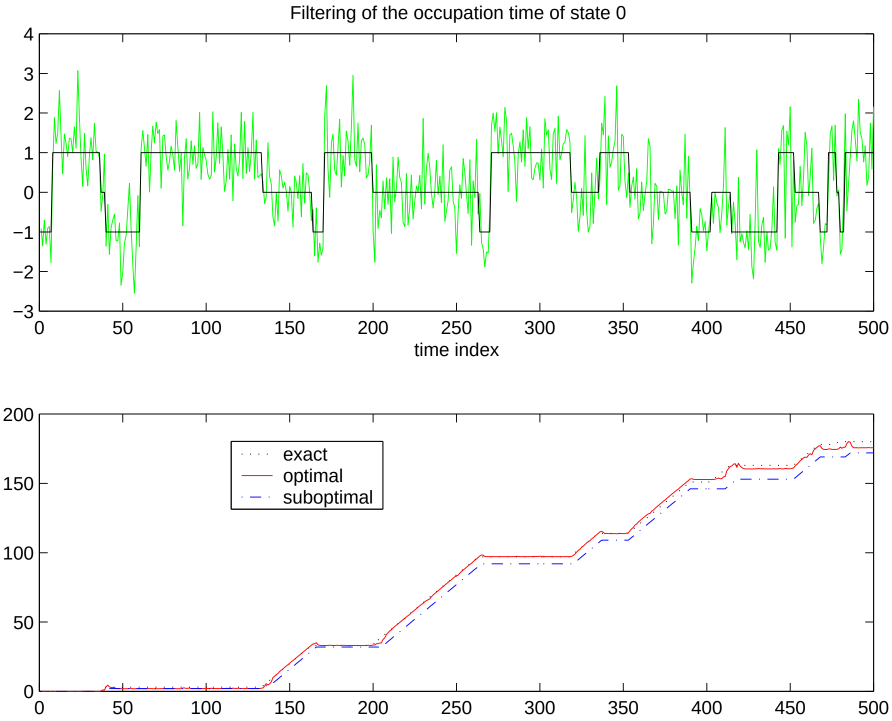

FIGURE 1. Filtering of the occupation time of the 0 state of a Markov chain. See the noisy path versus the signal realization at the upper plot. The optimal and suboptimal estimates are drawn versus the exact signal path.

2.1.4. *Simulation example.* Consider the following example: a slowly switching Markov chain with the state space $\{-1, 0, 1\}$ is observed in zero mean Gaussian noise. The optimal estimate for the 0 state occupation time has been calculated by the recursion (2.15). In figure 1 it is compared to the exact signal path and the following *suboptimal* estimate:

$$\widetilde{\vartheta}^0_n = \widetilde{\vartheta}^0_{n-1} + I(\widetilde{X}_n = 0)$$

where $\widetilde{X}_n = \text{argmax}_{1 \le i \le d}\, \pi_n(i)$ is the MAP (maximum a posteriori probability) estimate of X_n.

<u>Number of r-to-s transitions</u>

$$N^{rs}_n = \sum_{j=1}^{n} I(X_{j-1} = a_r) I(X_j = a_s) = N^{rs}_{n-1} + \langle e_r, I_{n-1}\rangle \langle e_s, I_n\rangle \tag{2.16}$$

Define vectors $Q_n := N^{rs}_n I_n$ and $\varepsilon_n := I_n - \Lambda^* I_{n-1}$. Then:

$$\begin{aligned} Q_n &:= N^{rs}_n I_n = \big[N^{rs}_{n-1} + \langle e_r, I_{n-1}\rangle\langle e_s, I_n\rangle\big]\big[\Lambda^* I_{n-1} + \varepsilon_n\big] = \qquad (2.17) \\ &= N^{rs}_{n-1}\Lambda^* I_{n-1} + N^{rs}_{n-1}\varepsilon_n + \langle e_r, I_{n-1}\rangle\langle e_s, \Lambda^* I_{n-1} + \varepsilon_n\rangle\big[\Lambda^* I_{n-1} + \varepsilon_n\big] \end{aligned}$$

Denote $\widetilde{Q}_{n|n-1} = E\big[Q_n | Y_1^{n-1}\big]$ and $\widetilde{Q}_n = E\big[Q_n | Y_1^n\big]$.

Conditioning (2.17) on Y_1^{n-1} gives:

$$\widetilde{Q}_{n|n-1} = \Lambda^* \widetilde{Q}_{n-1} + \alpha_{n-1} + \beta_{n-1} \tag{2.18}$$

where

$$\alpha_{n-1} = E\big[\langle e_r, I_{n-1}\rangle\langle e_s, \Lambda^* I_{n-1}\rangle \Lambda^* I_{n-1}\big|Y_1^{n-1}\big]$$

and

$$\beta_{n-1} = E\big[\langle e_r, I_{n-1}\rangle\langle e_s, \varepsilon_n\rangle\varepsilon_n|Y_1^{n-1}\big].$$

The above objects can be simplified:

$$\begin{aligned} \alpha_{n-1} &= E\big[\langle e_r, I_{n-1}\rangle\langle e_s, \Lambda^* e_r\rangle\Lambda^* e_r\big|Y_1^{n-1}\big] = \\ &= \langle e_r, \pi_{n-1}\rangle\langle e_s, \Lambda^* e_r\rangle\Lambda^* e_r \end{aligned}$$

Also:

$$\begin{aligned} \beta_{n-1} &= E\big[\langle e_r, I_{n-1}\rangle\langle e_s, \varepsilon_n\rangle\varepsilon_n\big|Y_1^{n-1}\big] = \\ &= E\big[\langle e_r, I_{n-1}\rangle\langle e_s, I_n - \Lambda^* I_{n-1}\rangle\varepsilon_n\big|Y_1^{n-1}\big] = \\ &= E\big[\langle e_r, I_{n-1}\rangle\langle e_s, I_n\rangle\varepsilon_n\big|Y_1^{n-1}\big] = \\ &= E\big[\langle e_r, I_{n-1}\rangle\langle e_s, I_n\rangle(I_n - \Lambda^* I_{n-1})\big|Y_1^{n-1}\big] = \\ &= E\big[\langle e_r, I_{n-1}\rangle\langle e_s, I_n\rangle(e_s - \Lambda^* e_r)\big|Y_1^{n-1}\big] = \\ &= E\big[\langle e_r, I_{n-1}\rangle\langle e_s, \Lambda^* I_{n-1}\rangle(e_s - \Lambda^* e_r)\big|Y_1^{n-1}\big] = \\ &= E\big[\langle e_r, I_{n-1}\rangle\langle e_s, \Lambda^* e_r\rangle(e_s - \Lambda^* e_r)\big|Y_1^{n-1}\big] = \\ &= \langle e_r, \pi_{n-1}\rangle\langle e_s, \Lambda^* e_r\rangle(e_s - \Lambda^* e_r) \end{aligned}$$

So that the eq. (2.18) reads:

$$\begin{aligned} \widetilde{Q}_{n|n-1} &= \Lambda^* \widetilde{Q}_{n-1} + \langle e_r, \pi_{n-1}\rangle\langle e_s, \Lambda^* e_r\rangle\Lambda^* e_r + \\ &\quad + \langle e_r, \pi_{n-1}\rangle\langle e_s, \Lambda^* e_r\rangle(e_s - \Lambda^* e_r) = \\ &= \Lambda^* \widetilde{Q}_{n-1} + \langle e_r, \pi_{n-1}\rangle\langle e_s, \Lambda^* e_r\rangle e_s \end{aligned} \tag{2.19}$$

Set $\widetilde{Q}_n := G(Y_n; Y_1^{n-1})$. Then for any bounded $\psi(x)$

$$E\Big[\big(Q_n - G(Y_n; Y_1^{n-1})\big)\psi(Y_n)|Y_1^{n-1}\Big] = 0 \tag{2.20}$$

The first term is calculated explicitly ($a_i \in \mathcal{S}$)

$$\begin{aligned} E\big[Q_n(i)\psi(Y_n)|Y_1^{n-1}\big] &= E\big[N_n^{rs} I_n(i)\psi(h(a_i) + \xi_n)|Y_1^{n-1}\big] = \\ = \quad \widetilde{Q}_{n|n-1}(i)\int \psi(x) f(x - h(a_i))dx \end{aligned}$$

Expanding similarly the second term, we arrive at:

$$\widetilde{Q}_n = \frac{D_n \widetilde{Q}_{n|n-1}}{\langle 1, D_n \Lambda^* \pi_{n-1}\rangle} \tag{2.21}$$

The equations (2.19) and (2.21) lead to a finite dimensional filter for Q_n:

$$\widetilde{Q}_n = \frac{D_n \Lambda^* \widetilde{Q}_{n-1} + D_n \langle e_r, \pi_{n-1}\rangle\langle e_s, \Lambda^* e_r\rangle e_s}{\langle 1, D_n \Lambda^* \pi_{n-1}\rangle}, \quad t \geq 1 \tag{2.22}$$

This recursion should be solved subject to $\widetilde{Q}_0 = 0$. The estimate of N_n^{rs} is recovered from $\widetilde{Q}_n$ by

$$\widetilde{N}_n^{rs} = \langle 1, \widetilde{Q}_n\rangle$$

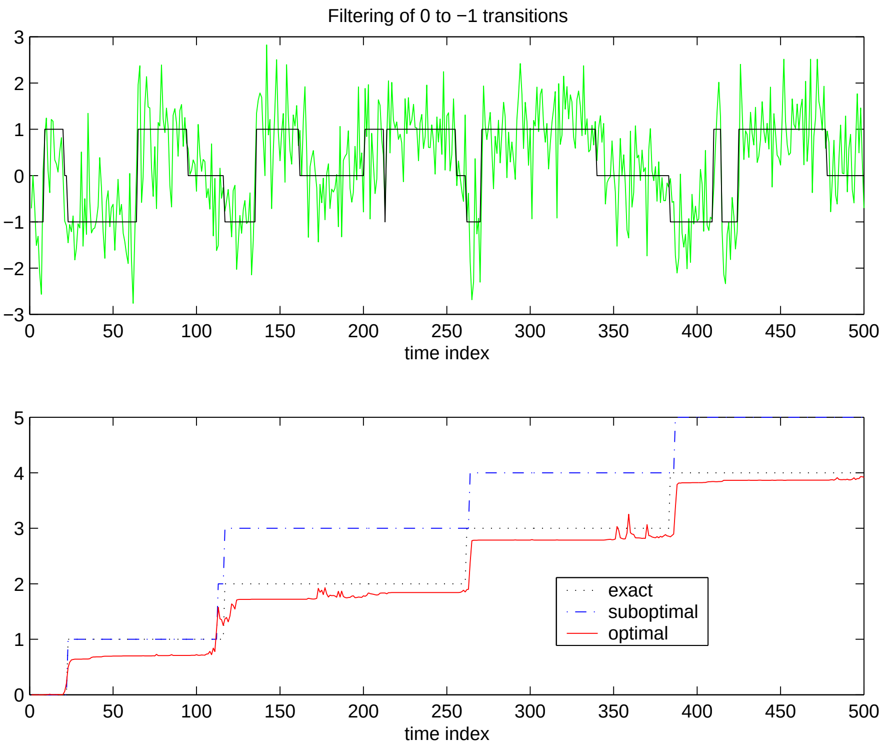

FIGURE 2. Filtering of the number of transition from the 0 to -1. See the noisy path versus the signal realization at the upper plot. The optimal and suboptimal estimates are drawn versus the exact signal path at the lower plot.

2.1.5. *Simulation example.* A slowly switching Markov chain with states in $\{-1, 0, 1\}$ is observed in Gaussian noise. The optimal filtering estimate of the number of transition form state 0 to -1 is drawn at Figure 2, along with suboptimal estimate

$$\widetilde{N}^{0,-1}_n = \widetilde{N}^{0,-1}_{n-1} + I(\widetilde{X}_{n-1} = 0)I(\widetilde{X}_n = -1)$$

where $\widetilde{X}_n$ is the MAP estimate of X_n.

2.2. The general filtering problem. Let $(X_n)_{n\ge 0}$ be a Markov process with the state space $\mathbb{S} \subseteq \mathbb{R}$, the transition kernel $\lambda(x, dy)$, i.e. for any

$$P(X_n \in A|X_0^{n-1}) = \int_A \lambda(X_{n-1}, dy), \quad A \subset \mathcal{B}(\mathbb{S})$$

and initial distribution $p(dx)$, i.e.

$$P(X_0 \in A) = \int_A p(dx).$$

Typical example is the process X generated by a nonlinear recursion

$$X_n = a(n, X_{n-1}) + \varepsilon_n, \quad n \ge 1$$

subject to a random variable X_0, where $a(n,x)$ is $\mathbb{Z}^+ \times \mathbb{R} \mapsto \mathbb{R}$ function and $\varepsilon = (\varepsilon_n)_{n\ge 0}$ is an i.i.d. sequence with ε_1 having probability density $\varphi(x)$. In this case

$$\lambda(x,dy) = \varphi\big(y - a(n,x)\big)dy. \tag{2.23}$$

Suppose that the observation process is given by $(n \ge 1)$

$$Y_n = A(n,X_n) + \xi_n$$

where $A(n,x)$ is $\mathbb{Z}^+ \times \mathbb{R} \mapsto \mathbb{R}$ function and $\xi = (\xi_n)_{n\ge 0}$ is an i.i.d. sequence, independent of ε, with ξ_1 having probability density $\psi(x)$.

THEOREM 2.5. *Assume that $\lambda(x,dy)$ satisfies* (2.23) *and X_0 has probability density $p(x)$. Then the conditional distribution $P(X_n \le x|Y_1^n)$ has density $\pi_n(x)$, satisfying the recursion*

$$\pi_n(x) = \frac{\psi\big(Y_n - A(n,x)\big)\int_{\mathbb{S}}\varphi\big(x - a(n,s)\big)\pi_{n-1}(s)ds}{\int_{\mathbb{R}}\psi\big(Y_n - A(n,x)\big)\int_{\mathbb{S}}\varphi\big(x - a(n,s)\big)\pi_{n-1}(s)dsdx} \tag{2.24}$$

subject to $\pi_0(x) = p(x)$.

PROOF. The proof is similar to Theorem 2.1. □

The equation (2.24) is a recursion propagates densities and its implementation generally requires two integrations at each time step, which makes the optimal filtering equations practically useless. Usually certain approximations are used in the applications in general.

In certain special cases, the conditional density $\pi_n(x)$ can be parameterized by a finite sufficient statistics, i.e. there is a vector process $\theta = (\theta_n(Y))_{n\ge 1}$, such that $\pi_n(x) = f(x,\theta_n(Y))$, and θ can be generated by recursion, not involving integration. For example, for the linear Gaussian systems, i.e. if with $a(n,x) = a_n x$ and $A(n,x) = A_n x$, $\varphi(x)$ and $\psi(x)$ are Gaussian densities and X_0 is a Gaussian r.v., the conditional density $\pi_n(x)$ is Gaussian[4] as well. Its sufficient statistic is two dimensional: its mean $\widehat{X}_n$ and covariance P_n are generated by the Kalman filter, so that

$$E\big(f(X_n)|Y_1^n\big) = \frac{1}{\sqrt{2\pi P_n}}\int_{\mathbb{R}}\exp\left\{-\frac{(x-\widehat{X}_n)^2}{2P_n}\right\}f(x)dx.$$

Though the latter calculation involves integration as well, for many functions f is can be tabulated (or computed exactly: e.g. polynomials, etc.) off-line.

When such parametrization is possible, the filtering equation (2.24) is said to be finite dimensional, since its finite dimensional *realization* exists.

Another example of finite dimensional filter is the equation (2.2), where the conditional distribution is parameterized by d-dimensional vector π_n. Finite dimensional filters exist in the filtering problem of the occupation times and transition counters, eq. (2.15) and (2.22). In general no constructive way to verify existence of a finite dimensional filter in a specific problem is currently known.

[4] recall that conditional expectation and orthogonal projection coincide in the Gaussian case

CHAPTER 4

Stochastic processes in continuous time

The continuous time stochastic process can be viewed as a parametric family of random variables $X = (X_t)_{t\ge 0}$ with $t \in \mathbb{R}^+$. The theory of these processes is significantly more delicate than in the discrete time case. This chapter gives a non-rigorous introductory treatment of several important topics in this field. For further reading the reader is referred to (1), (2), (13), (14), (7), (5).

1. Poisson process

Let $(\tau_n)_{n\ge 1}$ be a sequence of i.i.d. exponential random variables with parameter λ (see Example 1.5 in chapter 3)

$$P(\tau_1 \le t) = \begin{cases} 1 - e^{-\lambda t}, & t \ge 0 \\ 0 & t < 0 \end{cases}$$

The process $\Pi = (\Pi_t)_{t\ge 0}$

$$\Pi_t = \max\{k : \tau_1 + ... + \tau_k \le t\}$$

is called[1] Poisson process with intensity λ. By definition the trajectories of Π_t are piecewise constant functions (since Π_t takes integer values), continuous from the right and having limits from the left; $\Pi_0 = 0$ with probability one.

Theorem 1.1.

(1) *The process* Π *is Markov*

(2) Π_t *has Poisson distribution with parameter* λt*, i.e.*

$$P(\Pi_t = k|\Pi_0^s) = \frac{(\lambda(t-s))^{(k-\Pi_s)} e^{-\lambda(t-s)}}{(k-\Pi_s)!} I(k \ge \Pi_s), \quad \forall t \ge s \ge 0,$$

and so in particular

$$E\Pi_t = \lambda t, \quad E(\Pi_t - \lambda t)^2 = \lambda t.$$

(3) Π *has stationary independent increments, i.e. for any* $v > u > t > s$ *and* $k \ge \ell \ge 0$

$$P(\Pi_v - \Pi_u = k, \Pi_t - \Pi_s = \ell) = P(\Pi_v - \Pi_u = k)P(\Pi_t - \Pi_s = \ell) = \frac{(\lambda(v-u))^k e^{-\lambda(v-u)}}{k!} \frac{(\lambda(t-s))^\ell e^{-\lambda(t-s)}}{\ell!}. \quad (1.1)$$

[1] the convention $\max\{\emptyset\} = 0$ is understood

Proof. Introduce $\sigma_k = \sum_{i=1}^k \tau_i$. Then

$$P(\Pi_t = k|\Pi_0^s) = \sum_{\ell=1}^k P(\Pi_t = k|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell) I(\Pi_s = \ell)$$

and thus

$$P(\Pi_t = k|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell) = \frac{(\lambda(t-s))^{(k-\ell)} e^{-\lambda(t-s)}}{(k-\ell)!}$$

is to be verified. Further

$$P(\Pi_t = k|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell) = P(\sigma_k \le t < \sigma_{k+1}|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell) =$$
$$E\Big(P(\sigma_k \le t < \sigma_{k+1}|\tau_1, ..., \tau_{\ell+1})\Big|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell\Big) =$$
$$E\Big(P(\tau_{\ell+2} + ... + \tau_k \le t - \sigma_\ell - \tau_{\ell+1} < \tau_{\ell+2} + ... + \tau_{k+1}|\sigma_\ell, \tau_{\ell+1})\Big|\sigma_\ell, \tau_{\ell+1} > s - \sigma_\ell\Big)$$
$$= P\big(\tau_{\ell+2} + ... + \tau_k \le t - \sigma_\ell - \tau_{\ell+1} < \tau_{\ell+2} + ... + \tau_{k+1}\big|\sigma_\ell, \tau_{\ell+1} > s - \sigma_\ell\big) =$$
$$e^{\lambda(s-\sigma_\ell)} \int_{s-\sigma_\ell}^\infty P\big(\tau_{\ell+2} + ... + \tau_k \le t - \sigma_\ell - u < \tau_{\ell+2} + ... + \tau_{k+1}\big|\sigma_\ell\big)\lambda e^{-\lambda u} du =$$
$$= \int_0^\infty P\big(\tau_{\ell+2} + ... + \tau_k \le t - s - u' < \tau_{\ell+2} + ... + \tau_{k+1}\big)\lambda e^{-\lambda u'} du' =$$
$$= P\big(\tau_{\ell+1} + \tau_{\ell+2} + ... + \tau_k \le t - s < \tau_{\ell+1} + \tau_{\ell+2} + ... + \tau_{k+1}\big) =$$
$$= P\big(\tau_1 + ... + \tau_{k-\ell} \le t - s < \tau_1 + ... + \tau_{k-\ell+1}\big) = P\big(\Pi_{t-s} = k - \ell\big).$$

Now it is left to show that

$$P(\Pi_t = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \quad k \ge 0. \tag{1.2}$$

Note that

$$P(\Pi_t = k) = P(\sigma_k \le t < \sigma_k + \tau_{k+1}) = EI(\sigma_k \le t)I(\tau_{k+1} > t - \sigma_k) =$$
$$EI(\sigma_k \le t)e^{-\lambda(t-\sigma_k)} = \int_0^t e^{-\lambda(t-s)} dP(\sigma_k \le s). \tag{1.3}$$

and

$$P(\sigma_k \le s) = P(\tau_k \le s - \sigma_{k-1}) = EP(\tau_k \le s - \sigma_{k-1}|\sigma_{k-1}) =$$
$$EI(s - \sigma_{k-1} \ge 0)(1 - e^{-\lambda(s-\sigma_{k-1})}) = \int_0^s (1 - e^{-\lambda(s-u)}) dP(\sigma_{k-1} \le u) \tag{1.4}$$

Clearly

$$P(\sigma_1 \le s) = P(\tau_1 \le s) = 1 - e^{-\lambda s}$$

and so by induction $P(\sigma_k \le s)$ has density, which by (1.4) satisfies

$$\frac{dP(\sigma_k \le s)}{ds} = \lambda \int_0^s e^{-\lambda(s-u)} \frac{dP(\sigma_{k-1} \le u)}{du} du.$$

By induction

$$\frac{dP(\sigma_k \le s)}{ds} = \lambda \frac{(\lambda s)^{k-1} e^{-\lambda s}}{(k-1)!}.$$

The latter is known as Erlang distribution. The equation (1.2) and thus also the statement of (2) now follow from (1.3).

The claim (3) follows directly from (2)

$$P(\Pi_v - \Pi_u = k, \Pi_t - \Pi_s = \ell) = EI(\Pi_t - \Pi_s = \ell)P(\Pi_v = k + \Pi_u|\Pi_0^u) =$$
$$EI(\Pi_t - \Pi_s = \ell)\frac{(\lambda(v-u))^k e^{-\lambda(v-u)}}{k!} = \frac{(\lambda(t-s))^\ell e^{-\lambda(t-s)}}{\ell!}\frac{(\lambda(v-u))^k e^{-\lambda(v-u)}}{k!}.$$

□

It can be shown that any process with piecewise constant trajectories with unit jumps, stationary independent increments and covariance process λt is the Poisson process, i.e. the times between the jumps are necessarily exponential random variables.

Poisson process is one of the fundamental building blocks in the theory of jump processes: including continuous time Markov chains, etc. It is also popular in many applications such as queueing theory. The typical setup consists of a number of servers and a queue. The arriving clients join the queue and get the service according to certain regime from one of the servers (e.g. FIFO). Each service times are usually assumed to be independent random variables (either exponential or not). The queueing system types have conventional names: e.g. M(λ)/G/1/∞/FIFO stands for one server system with infinite FIFO queue length, where the customers arrivals stream is Markov (i.e. Poissonian with intensity λ) and the service times are of general distribution. Usually FIFO and infinite queue are the defaults and then the same system is called M/G/1. The typical questions are: what the expected queue length is; or what is the distribution of the idle times, etc. For example, the expected waiting time in the queue of the stationary M(λ)/G/1 system is given by the Khinchin-Pollaczek formula

$$W_q = \frac{\lambda ES^2}{2(1-\rho)}, \quad \rho = \lambda ES$$

where S is the random service time.

2. Wiener process

Let $\xi = (\xi_n)_{n\ge 0}$ be a sequence of i.i.d. random variables with zero mean and unit variance. For $0 \le t \le 1$ define the scaled *random walk* process

$$W_t^n = \frac{1}{\sqrt{n}}\sum_{j=1}^{\lfloor tn \rfloor} \xi_j.$$

For any fixed t by the Central Limit Theorem W_t^n converges weakly (in distribution) to a Gaussian random variable with zero mean and variance t. Moreover it is not hard to check that for any set of points $0 \le t_1 < t_2 < ... < t_m \le 1$, the random vector $[W_{t_1}^n, ..., W_{t_m}^n]$ converges weakly to a Gaussian vector with zero mean and the covariance matrix Γ with entries[2] $\gamma_{ij} = t_i \wedge t_j$.

It turns out that $W^n = (W_t^n)_{0\le t\le 1}$ has also a weak limit as a continuous time random process, i.e. for any bounded and continuous functional ψ on the space[3] of functions continuous from the right and with limits from the left

$$\lim_{n\to\infty} E\psi(W^n) = E\psi(W),$$

[2] $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$

[3] with appropriate metric

where W is a Gaussian process with continuous trajectories, independent and stationary increments, zero mean and covariance function $EW_tW_s = s \wedge t$. The limit process is called the Wiener process. It is the mathematical model for the Brownian motion, i.e. the motion of a particle driven by interactions with other particles in a solution, which was first described by botanist Brown.

It turns out that any process X satisfying the properties

(1) X has continuous trajectories
(2) For any $t > s$, $E(X_t|\mathcal{F}_s^X) = X_s$, $E\big[(X_t - X_s)^2|\mathcal{F}_s^X\big] = t - s$, where $\mathcal{F}_s^X = \sigma\{X_u, u \le s\}$

is necessarily the Wiener process, i.e. it is Gaussian!

The trajectories of the Wiener process are extremely irregular. For example, they are nowhere differentiable and even have unbounded variation, i.e.

$$\bigvee_a^b W_t(\omega) = \sup \sum_j |W_{t_{j+1}} - W_{t_j}| = \infty, \quad P - a.s.$$

where the supremum is taken with respect to all partitions $a \le t_1 < t_2 < ... < t_n \le b$.

The Wiener process is the main ingredient in many probabilistic and statistical models in physics and engineering. Due to the unusual properties of its trajectories it is also a fascinating object for mathematical research.

3. Ito stochastic integral with respect to Wiener process

In many engineering applications the systems are subject to random perturbations (noise). In particular, it is customary to deal with the differential equations

$$\dot{X}_t = a(t, X_t) + b(t, X_t)\nu_t, \quad X_0 = x \tag{3.1}$$

where ν is so called "white" noise process with intensity N_0, i.e. a process with zero mean and Dirac δ correlation function: $E\nu_t\nu_s = N_0\delta(t - s)$. The spectral density of ν_t, being the Fourier transform of δ, is constant, which is the origin for the term "white".

One can try to construct such process by formal differentiation of the Wiener process:

$$E\dot{W}_t\dot{W}_s = \frac{\partial^2}{\partial t\partial s}EW_sW_t = \frac{\partial^2}{\partial t\partial s}t \wedge s = \delta(t - s) \tag{3.2}$$

Exercise 3.1. *Show that for e.g. $t > s > 0$ and continuous f*

$$\int_0^\infty \frac{\partial^2}{\partial t\partial s}(s \wedge t)f(s)ds = f(t)$$

i.e. the latter equality in (3.2) *formally holds.*

However, as it was mentioned above, the trajectories of the Wiener process are nowhere differentiable and thus such definition is incorrect. Note that ν does not make sense physically as well: it has infinite variance, which contradicts the engineering intuition/practice.

The next step is to interpret (3.1) as an integral equation, i.e.

$$X_t = X_0 + \int_0^t a(s, X_s)ds + \int_0^t b(s, X_s)dW_s. \tag{3.3}$$

This construction also fails, because the last term cannot be defined as Lebesgue - Stieltjes integral due to unbounded variation of the trajectories of W.

Example 3.2. Let $0 = t_0 < t_1 < ... < t_n = T$ be a partition of $[0, T]$ and set

$$I^n = \sum_{i=0}^{n-1} W_{t_i}\big(W_{t_{i+1}} - W_{t_i}\big), \quad S^n = \sum_{i=0}^{n-1} W_{t_{i+1}}\big(W_{t_{i+1}} - W_{t_i}\big)$$

The only difference between these random variables is the point at which the integrand W_t is sampled in each interval $(t_i, t_{j+1}]$. While for Stieltjes integrable functions this would not influence the limit as $n \to \infty$, for the highly irregular trajectories of W this turns to be a crucial matter. Clearly for any $n \geq 1$

$$EI^n = 0, \quad ES^n = T$$

and thus the limits (if exist in some sense) would not coincide!

Nevertheless (3.3) makes sense if the integral with respect to the Wiener process is well defined for certain class of random processes. The ultimate construction of such integral was given by K.Ito.

All the random objects below are defined on some probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}^W_t$ be the sub-σ-algebras of $\mathcal{F}$ generated by the process W, i.e. $\mathcal{F}^W_t = \sigma\{W_s, s \leq t\}$. Clearly $\mathcal{F}^W_t$ is an increasing family of σ-algebras, which is called *filtration*. The random process $X = (X_t)_{0\leq t \leq T}$ is said to be *adapted* with respect to filtration $\mathcal{F}^W_t$ if for any fixed t, X_t is $\mathcal{F}^W_t$-measurable.

The idea of construction is to define the integral for a class of simpler random processes, which is sufficiently rich to approximate more complex processes of interest. It turns out that any adapted random process X, satisfying[4]

$$E\int_0^T X_s^2 ds < \infty \tag{3.4}$$

can be approximated by the piecewise constant (or *simple*) processes of the form [5]

$$X^n_t = \sum_{j=1}^{n} \alpha_j 1_{(t_j, t_{j+1}]}(t), \tag{3.5}$$

where $(t_j)_{j\leq n}$ is some partition (depending on n) and $(\alpha_j)_{j\leq n}$ is a sequence of r.v., such that α_j is F^W_t-measurable if $t_j \leq t$. The approximation holds in the sense that

$$\lim_{n\to\infty} E\int_0^T (X_s - X^n_s)^2 ds = 0.$$

Since in $\mathbb{L}^2$ any converging sequence is also fundamental it follows

$$\lim_{n,m\to\infty} E\int_0^T (X^m_s - X^n_s)^2 ds = 0. \tag{3.6}$$

[4] the Ito integral can also be defined under much weaker assumption

$$P\Big(\int_0^T X_s^2 ds < \infty\Big) = 1.$$

[5] For example, if X_t has continuous pathes, then $X^n_{t_i} = X_{t_j}$ can be taken - see e.g. (5) for a more detailed account

For a simple function the Ito integral is defined as

$$I_T(X^n) := \sum_{j=1}^{n} \alpha_j \big[W_{t_{j+1}} - W_{t_j}\big]$$

and is denoted by $\int_0^T X^n_s dW_s$.

Note that

$$\begin{aligned}
&E\big(I_T(X^n)\big)^2 = E\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_j\alpha_i \big[W_{t_{j+1}} - W_{t_j}\big]\big[W_{t_{i+1}} - W_{t_i}\big] = \\
&E\sum_{j=1}^{n} \alpha_j^2 E\Big(\big[W_{t_{j+1}} - W_{t_j}\big]^2 \big| \mathcal{F}^W_{t_j}\Big) + \\
&E\sum_{i\neq j} \alpha_i\alpha_j E\Big(\big[W_{t_{j+1}} - W_{t_j}\big]\big[W_{t_{i+1}} - W_{t_i}\big] \big| \mathcal{F}^W_{t_i\vee t_j}\Big) = \\
&\sum_{j=1}^{n} E\alpha_j^2 [t_{j+1} - t_j] = \int_0^T E\big(X^n_s\big)^2 ds
\end{aligned}$$

This property and (3.6) give (note that $X^n_s - X^m_s$ is again a simple function of the form (3.5))

$$E(I_T(X^n) - I_T(X^m))^2 = E\big(I_T(X^n - X^m)\big)^2 = \int_0^T E(X^n_s - X^m_s)^2 ds \xrightarrow{n,m\to\infty} 0,$$

which means that the sequence $I_T(X^n)$ is fundamental in $\mathbb{L}^2$ and thus converges [6] in $\mathbb{L}^2$ to a limit $I_T(X)$. This limit is defined to be the Ito integral of X and denoted by

$$I_T(X) = \int_0^T X_s dW_s.$$

The Ito integral can be considered as a stochastic process with time parameter t:

$$I_t(X) = \int_0^t X_s dW_s := \int_0^T 1_{[0,t]}(s) X_s dW_s$$

REMARK 3.3. This construction assumed that X_t is $\mathcal{F}^W_t$-adapted, i.e. for any X_t is $\mathcal{F}^W_t$ measurable for any t. Clearly all the arguments can be replicated if X_t is $\mathcal{F}_t$-adapted for some larger filtration $\mathcal{F}_t \supseteq \mathcal{F}^W_t$. This extends the definition of the stochastic integrals to more general integrands, e.g. $\int_0^t W_t dV_t$ where V and W are independent Wiener processes.

3.1. Properties of the Ito integral. Consider the following basic properties of the Ito integral

1. $I_t(aX + bY) = aI_t(X) + bI_t(Y)$
2. $E(I_t(X)|\mathcal{F}^W_s) = I_s(X)$ for $t > s$ and in particular $EI_t(X) = 0$
3. $EI_s(X)I_t(Y) = \int_0^{t\wedge s} EX_uY_u du$ and in particular $EI^2_t(X) = \int_0^t EX^2_s ds$
4. $I_t(X)$ has continuous trajectories for $0 \le t \le T$

[6] the P-a.s. convergence can be verified as well

The first three properties are verified for the simple processes and then their validity for more complex processes is established by passing to the limit. For example, let X^n and Y^n be the approximations of X and Y. The process $aX^n + bY^n$ is clearly simple (and adapted) and thus $I_t(aX^n + bY^n) = aI_t(X^n) + bI_t(Y^n)$. So (1) holds, since $aI_t(X^n) + bI_t(Y^n) \xrightarrow[n\to\infty]{\mathbb{L}^2} aI_t(X) + bI_t(Y)$. The fourth property will not be proved here (note that it also holds for simple functions!).

4. Stochastic differential equations

Consider the integral equation

$$X_t = X_0 + \int_0^t a(s, X_s)ds + \int_0^t b(s, X_s)dW_s, \quad 0 \le t \le T \tag{4.1}$$

where X_0 is a random variable, $a(s,x)$ and $b(s,x)$ are $\mathbb{R}_+ \times \mathbb{R} \mapsto \mathbb{R}$ functions and W is the Wiener process. Assume that X_0 and W are independent. The integration w.r.t "ds" is understood in the usual sense (i.e. Riemann or Lebesgue).

Definition 4.1. A non $\mathcal{F}_t^W$-adapted process X is a (strong) solution of (4.1), if

$$P\left(\int_0^T |a(s,X_s)|ds < \infty\right) = 1, \quad P\left(\int_0^T [b(s,X_s)]^2 ds < \infty\right) = 1$$

and (4.1) holds P-a.s.

It can be shown e.g. that if $a(t,x)$ and $b(t,x)$ satisfy the Lipschitz condition:

$$[a(t,y) - a(t,y')]^2 + [b(t,y) - b(t,y')]^2 \le L[y-y']^2, \quad t \in [0,T]$$

with some constant L and increase not faster than linearly

$$a^2(t,y) + b^2(t,y) \le L(1+y^2)$$

then (4.1) has the unique solution. Usually the equation (4.1) is written in the differential form in the spirit of the regular calculus

$$dX_t = a(t,X_t)dt + b(t,X_t)dW_t$$

subject to $X_0 = x$. The process generated by this stochastic differential equation is called *diffusions* with *drift* $a(t,x)$ and diffusion coefficient $b(t,x)$.

4.1. The Ito formula.

4.1.1. *Scalar case.* Consider the random process $\xi = (\xi_t)_{0\le t\le T}$, which has the Ito differential

$$d\xi_t = a(t,\xi_t)dt + b(t,\xi_t)dW_t \tag{4.2}$$

where a and b are functions, satisfying the appropriate properties. It turns out that the process $\zeta_t = f(t,\xi_t)$ also has Ito differential, if f is sufficiently smooth.

Theorem 4.2. *(Ito formula) Let the function $f(t,x)$ be continuous and has the continuous partial derivatives $f'_t(t,x)$, $f'_x(t,x)$ and $f''_{xx}(t,x)$. Assume that the random process ξ has the stochastic differential* (4.2). *Then the process $f(t,\xi_t)$ also has a stochastic differential and*

$$df(t,\xi_t) = \big[f'_t(t,\xi_t) + f'_x(t,\xi_t)a(t,\xi_t) + \frac{1}{2}f''_{xx}(t,\xi_t)b^2(t,\xi_t)\big]dt + \\ f'_x(t,\xi_t)b(t,\xi_t)dW_t \tag{4.3}$$

Proof. (heuristic sketch) Since f is twice differentiable, for any partition $\{t_j\}$

$$f(T,\xi_T) - f(0,\xi_0) = \sum_{j=1}^{n} \left[f(t_{j+1},\xi_{t_{j+1}}) - f(t_j,\xi_{t_j}) \right] = \sum_{j=1}^{n} f'_t(t_j,\xi_{t_j})[t_{j+1}-t_j] +$$

$$\sum_{j=1}^{n} f'_x(t_j,\xi_{t_j})[\xi_{t_{j+1}} - \xi_{t_j}] + \frac{1}{2}\sum_{j=1}^{n} f''_{xx}(t_j,\xi_{t_j})[\xi_{t_{j+1}} - \xi_{t_j}]^2 + r^n \quad (4.4)$$

where the residual terms r^n contain summation of the higher powers of $[t_{j+1}-t_j]$ and $[\xi_{t_{j+1}} - \xi_{t_j}]$. Denote $f_j = f(t_j,\xi_{t_j})$, etc. for brevity. The first term in (4.4) converges to $\int_0^T f'_t(s,\xi_s)ds$, while the second term gives

$$\sum_{j=1}^{n} f'_{xj} a_j [t_{j+1}-t_j] + \sum_{j=1}^{n} f'_{xj} b_j [W_{t_{j+1}} - W_{t_j}] \xrightarrow{n\to\infty}$$

$$\int_0^t a(s,\xi_s) f'_x(s,\xi_s) ds + \int_0^t b(s,\xi_s) f'_x(s,\xi_s) dW_s$$

where the latter integral is in the sense of Ito. The last term in (4.4) gives

$$\frac{1}{2}\sum_{j=1}^{n} f''_{xxj}[\xi_{t_{j+1}} - \xi_{t_j}]^2 = \frac{1}{2}\sum_{j=1}^{n} f''_{xxj} a_j^2 [t_{j+1}-t_j]^2 +$$

$$\sum_{j=1}^{n} f''_{xxj} a_j b_j [W_{t_{j+1}} - W_{t_j}][t_{j+1}-t_j] +$$

$$\frac{1}{2}\sum_{j=1}^{n} f''_{xxj} b_j^2 [W_{t_{j+1}} - W_{t_j}]^2 := J_1 + J_2 + J_3$$

The term J_1 converges to zero as $\max_j \left|t_{j+1}-t_j\right| \to 0$

$$\Big|\sum_{j=1}^{n} f''_{xxj} a_j^2 [t_{j+1}-t_j]^2\Big| \le \max_j \left|t_{j+1}-t_j\right| \sum_{j=1}^{n} f''_{xxj} a_j^2 [t_{j+1}-t_j]$$

since $\sum_{j=1}^{n} f''_{xxj} a_j^2 [t_{j+1}-t_j]$ converges to the integral $\int_0^T f''_{xx}(s,\xi_s) a^2(s,\xi_s) ds$.

The term J_2 also converges to zero in $\mathbb{L}^2$

$$E\Big(\sum_{j=1}^{n} f''_{xxj} a_j b_j [W_{t_{j+1}} - W_{t_j}][t_{j+1}-t_j]\Big)^2 =$$

$$E\sum_{\ell=1}^{n}\sum_{j=1}^{n} f''_{xx\ell} a_\ell b_\ell [W_{t_{\ell+1}} - W_{t_\ell}][t_{\ell+1}-t_\ell] f''_{xxj} a_j b_j [W_{t_{j+1}} - W_{t_j}][t_{j+1}-t_j] =$$

$$E\sum_{\ell=1}^{n}\sum_{j=1}^{n} f''_{xx\ell} f''_{xxj} a_j b_j a_\ell b_\ell [t_{\ell+1}-t_\ell][t_{j+1}-t_j] E\big([W_{t_{\ell+1}} - W_{t_\ell}][W_{t_{j+1}} - W_{t_j}] | \mathcal{F}^W_{t_j \vee t_\ell}\big)$$

$$= E\sum_{j=1}^{n} \big(f''_{xxj} a_j b_j\big)^2 [t_{j+1}-t_j]^3 \xrightarrow{n\to\infty} 0$$

The third term J_3 converges to a nonzero limit. Indeed

$$E\left(\sum_{j=1}^n f''_{xxj} b_j^2 [W_{t_{j+1}} - W_{t_j}]^2 - \sum_{j=1}^n f''_{xxj} b_j^2 [t_{j+1} - t_j]\right)^2 =$$

$$E\left(\sum_{j=1}^n f''_{xxj} b_j^2 \big\{[W_{t_{j+1}} - W_{t_j}]^2 - [t_{j+1} - t_j]\big\}\right)^2 =$$

$$E\sum_{j=1}^n \sum_{\ell=1}^n f''_{xxj} b_j^2 f''_{xx\ell} b_\ell^2 \big\{[W_{t_{j+1}} - W_{t_j}]^2 - [t_{j+1} - t_j]\big\}\big\{[W_{t_{\ell+1}} - W_{t_\ell}]^2 - [t_{\ell+1} - t_\ell]\big\} =$$

$$E\sum_{j=1}^n \big(f''_{xxj}\big)^2 b_j^4 \big\{[W_{t_{j+1}} - W_{t_j}]^2 - [t_{j+1} - t_j]\big\}^2 = E\sum_{j=1}^n \big(f''_{xxj}\big)^2 b_j^4 2[t_{j+1} - t_j]^2 \xrightarrow{n\to\infty} 0$$

whereas

$$\sum_{j=1}^n f''_{xxj} b_j^2 [t_{j+1} - t_j] \xrightarrow{n\to\infty} \int_0^T f''_{xx}(s,\xi_s) b^2(s,\xi_s) ds.$$

Similarly the residual terms r^n can be shown to vanish as $n \to \infty$ and thus the right hand side of (4.4) gets the integral form of the required formula as $n \to \infty$. □

Remark 4.3. Note that if a function W_t had finite variation then by the chain rules of the classic calculus one would obtain

$$df(t,X_t) = f'_t(t,X_t)dt + f'_x(t,X_t)dX_t = \big[f'_t(t,X_t) + f'_x(t,X_t)a(t,X_t)\big]dt + f'_x(t,X_t)b(t,X_t)dW_t,$$

i.e. the Ito formula has a "non classical" appendix $1/2 f''_{xx} b^2 dt$!

Example 4.4. Let $f(t,x) = x^2$ and $X_t \equiv W_t$, i.e. $a \equiv 0$ and $b \equiv 1$. Then by Ito formula

$$d(W_t^2) = 2W_t dW_t + \frac{1}{2} 2dt$$

which actually means that

$$W_t^2 = 2\int_0^t W_s dW_s + t.$$

Example 4.5. Consider the Ornstein-Uhlenbeck process X satisfying the SDE

$$dX_t = a_t X_t dt + b_t dW_t$$

subject to random initial condition $X_0 = \eta$, where a_t and b_t are deterministic functions. Let $m_t = EX_t$ and $V_t = E(X_t - m_t)^2$. Recall that the differential equation is nothing but notation for

$$X_t = \eta + \int_0^t a_s X_s ds + \int_0^t b_s dW_s.$$

Applying the expectation to the latter equality, obtain

$$m_t = E\eta + \int_0^t a_s m_s ds$$

or in other words

$$\dot m_t = a_t m_t, \quad m_0 = E\eta.$$

Let $\Delta_t = X_t - m_t$, then

$$\Delta_t = \eta - E\eta + \int_0^t a_s \Delta_s ds + \int_0^t b_s dW_s.$$

Now apply the Ito formula to Δ_t^2:

$$d(\Delta_t^2) = 2\Delta_t^2 a_t dt + 2\Delta_t b_t dW_t + b_t^2 dt$$

which stands for

$$\Delta_t^2 = \Delta_0^2 + \int_0^t (2a_s\Delta_s^2 + b_s^2)ds + \int_0^t 2\Delta_s b_s dW_s.$$

Taking the expectation from both sides obtain

$$V_t = E(\eta - E\eta)^2 + \int_0^t (2a_s V_s + b_s^2)ds$$

or

$$\dot{V}_t = 2a_t V_t + b_t^2, \quad V_0 = \mathrm{var}(\eta).$$

Note that if $a_t \equiv -a < 0$ and $b_t \equiv b > 0$, then $m_t \to 0$ and $V_t \to b^2/(2a)$ as $t \to \infty$. In engineering notations this means that the output of the low pass filter, driven by the "white" noise, has the power $b^2/(2a)$ in the steady state.

Since this system is linear, the probability distribution of X_t has Gaussian density, if η is a Gaussian random variable

$$p_t(x) = \frac{d}{dx}P(X_t \le x) = \frac{1}{\sqrt{2\pi V_t}} \exp\left\{\frac{-(x-m_t)^2}{2V_t}\right\}.$$

The solution of this equation can be found explicitly:

$$X_t = e^{\int_0^t a_s ds}\left(X_0 + \int_0^t b_s e^{-\int_0^s a_u du} dW_s\right)$$

which is verified by the (vector[7]) Ito formula applied to X_t.

4.1.2. *Vector case.* Consider now a vector diffusion

$$d\xi_t = a(t,\xi_t)dt + b(t,\xi_t)dW_t$$

where $a(t,x) : \mathbb{R}_+ \times \mathbb{R}^d \mapsto \mathbb{R}^d$ and $b(t,x) : \mathbb{R}_+ \times \mathbb{R}^d \mapsto \mathbb{R}^{d\times m}$ and W is a vector of m independent Wiener processes.

Theorem 4.6. *(vector Ito formula) Let $f(t,x)$ be a $\mathbb{R}_+ \times \mathbb{R}^d \mapsto \mathbb{R}$ continuous function with continuous derivatives f'_t, f'_{x_j} and $f''_{x_j x_j}$. Then the process $f(t,\xi_t)$ has the stochastic Ito differential*

$$df(t,\xi_t) = \big[f'_t(t,\xi_t) + \sum_{i=1}^d f'_{x_i}(t,\xi_t)a_i(t,\xi_t) +$$

$$\frac{1}{2}\sum_{i,j=1}^d f''_{x_i x_j}(t,\xi_t)\sum_{k=1}^m b_{ik}(t,\xi_t)b_{jk}(t,\xi_t)\big]dt + \sum_{i=1}^d\sum_{j=1}^m f'_{x_i}(t,\xi_t)b_{ij}(t,\xi_t)dW_t(j). \quad (4.5)$$

[7] see the following section

Remark 4.7. The equation (4.5) can be written compactly as

$$df(t,\xi_t) = \big[f_t'(t,\xi_t) + \nabla^* f(t,\xi_t)a(t,\xi_t) + \frac{1}{2}\Big(\nabla^* b(t,\xi_t)b^*(t,\xi_t)\nabla\Big)f(t,\xi_t)\big]dt + \\ \nabla^* f(t,\xi_t)b(t,\xi_t)dW_t, \quad (4.6)$$

where ∇f is the column gradient vector and the differential operator $\nabla^* bb^* \nabla$ is determined by the formal rules of multiplication.

Example 4.8. Consider the linear system

$$dX_t = aX_t dt + bdW_t, \quad X_0 = \eta$$

where η is square integrable random vector, a and b are $d\times d$ and $d \times m$ matrices and W_t is the vector Wiener process of dimension m. Lets find the equations for $m_t = EX_t$ and $P_t = \mathrm{cov}(X_t)$. Taking the expectation from both sides one immediately obtains the differential equation for $m_t = EX_t$

$$\dot{m}_t = am_t, \quad m_0 = E\eta.$$

The process $D_t = X_t - m_t$ satisfies

$$dD_t = aD_t dt + bdW_t, \quad D_0 = \eta - E\eta.$$

Let $f_{pq}(x) = x_p x_q$, $x \in \mathbb{R}^d$ and apply the vector Ito formula to $\Gamma_t^{pq} = f_{pq}(D_t)$

$$d\Gamma_t^{pq} = D_t^p dD_t^q + D_t^q dD_t^p + \frac{1}{2}\sum_{i,j=1}^d \big[\delta(i=p,j=q) + \delta(i=q,j=p)\big]\sum_{k=1}^m b_{ik}b_{jk}dt =$$

$$D_t^p \sum_{i=1}^d a_{qi} D_t^i dt + D_t^p \sum_{j=1}^m b_{qj} dW_t^j + D_t^q \sum_{i=1}^d a_{pi} D_t^i dt + D_t^q \sum_{j=1}^m b_{pj} dW_t^j +$$

$$\frac{1}{2}\Big[\sum_{k=1}^m b_{pk}b_{qk}dt + \sum_{k=1}^m b_{qk}b_{pk}dt\Big] =$$

$$\sum_{i=1}^d a_{qi}\Gamma_t^{pi} dt + D_t^p \sum_{j=1}^m b_{qj} dW_t^j + \sum_{i=1}^d a_{pi}\Gamma_t^{qi} dt + D_t^q \sum_{j=1}^m b_{pj} dW_t^j + \sum_{k=1}^m b_{pk}b_{qk}dt$$

Taking the expectation of the latter equation one gets

$$dP_t^{pq} = \sum_{i=1}^d a_{qi} P_t^{pi} dt + \sum_{i=1}^d a_{pi} P_t^{qi} dt + \sum_{k=1}^m b_{pk}b_{qk}dt,$$

for $P_t^{pq} = E\Gamma_t^{pq}$ or in the matrix form

$$\dot{P}_t = aP_t^* + P_t a^* + bb^* = aP_t + P_t a^* + bb^*, \quad P_0 = \mathrm{cov}(\eta) \qquad (4.7)$$

where $P_t = ED_t D_t^*$. The *Lyapunov* equation (4.7) is linear and it can be further analyzed to verify whether X_t has a non degenerate Gaussian density or whether this density stabilizes as $t \to \infty$, etc.

The vector Ito formula can be conveniently remembered as follows. Let X_t and Y_t be a pair of Ito processes with differentials

$$dX_t = a_1(X_t,Y_t)dt + b_{11}(X_t,Y_t)dW_t + b_{12}(X_t,Y_t)dW_t'$$
$$dX_t = a_2(X_t,Y_t)dt + b_{21}(X_t,Y_t)dW_t + b_{22}(X_t,Y_t)dW_t',$$

where W' and W are independent Wiener processes. Let $f(t, x, y)$ be a real function of three arguments, sufficiently differentiable so that Ito formula is applicable.

Use the regular calculus rules and consequent Taylor expansion up to order two to write formally

$$df(t, X_t, Y_t) = f_t dt + f_x dX_t + f_y dY_t + \frac{1}{2} f_{xx}(dX_t)^2 + f_{xy}(dX_t dY_t) + \frac{1}{2} f_{yy}(dY_t)^2,$$

where e.g.

$$f_{xx} = \frac{\partial^2}{\partial^2 x} f(t, x, y)\Big|_{x := X_t, y := Y_t}, \quad \text{etc.}$$

To proceed use the following "multiplication" table

$\cdot$	dt	dW_t	dW'_t
dt	0	0	0
dW_t	0	dt	0
dW'_t	0	0	dt

For example

$$\begin{aligned}(dX_t)^2 =&(a_1 dt + b_{11} dW_t + b_{12} dW'_t)^2 = a_1^2 (dt)^2 + b_{11}^2 (dW_t)^2 + b_{12}^2 (dW'_t)^2 + \\ &2a_1 b_{11}(dW_t dt) + 2a_1 b_{12}(dW'_t dt) + 2b_{11} b_{12}(dW_t dW'_t) = b_{11}^2 dt + b_{12}^2 dt.\end{aligned}$$

Similar calculations lead to

$$\begin{aligned}df(t, X_t, Y_t) = f_t + f_x dX_t + f_y dY_t + \frac{1}{2} f_{xx}(b_{11}^2 dt + b_{12}^2) dt + \\ f_{xy}(b_{11} b_{21} + b_{12} b_{22}) dt + \frac{1}{2} f_{yy}(b_{21}^2 + b_{22}^2) dt\end{aligned}$$

Verify that the proposed procedure leads to the correct answer (suggested by formal Ito formula).

Example 4.9. Consider the two dimensional Ito system

$$\begin{aligned}du'_t &= -\sin(\xi_t) dW_t + \cos(\xi_t) dV_t \\ du''_t &= \cos(\xi_t) dW_t + \sin(\xi_t) dV_t\end{aligned}$$

subject to $u'_0 = u''_0 = 0$, where W and V are independent Wiener processes and ξ_t is some $\mathcal{F}_t^{V,W}$-adapted process. Let's show that u' and u'' are independent Gaussian random variables for any fixed $t > 0$, i.e.

$$E\Big(\exp\left\{ i\lambda u'_t + i\mu u''_t \right\}\Big) = e^{-1/2\lambda^2 t} e^{-1/2\mu^2 t} \tag{4.8}$$

Apply the Ito formula to the function $\varphi_t = \exp\{i\lambda u'_t + i\mu u''_t\}$:

$$\begin{aligned}&d\varphi_t = \varphi_t(i\lambda du'_t + i\mu du''_t) - 1/2\lambda^2 \varphi_t[\sin^2(\xi_t) + \cos^2(\xi_t)]dt \\ &- 1/2\mu^2 \varphi_t[\cos^2(\xi_t) + \sin^2(\xi_t)]dt - \varphi_t \lambda\mu[-\sin(\xi_t)\cos(\xi_t) + \cos(\xi_t)\sin(\xi_t)]dt = \\ &\varphi_t(i\lambda du'_t + i\mu du''_t) - 1/2[\lambda^2 + \mu^2]\varphi_t dt\end{aligned}$$

The characteristic function $\psi_t = E\varphi_t$ then satisfies

$$\dot{\psi}_t = -1/2[\lambda^2 + \mu^2]\psi_t$$

subject to $\psi_0 = 1$ and thus is given by $\psi_t = e^{-1/2t\lambda^2 - 1/2t\mu^2}$. In fact, similarly it can be verified that

$$E\Big(\exp\big\{i\lambda(u'_t - u'_s) + i\mu(u''_t - u''_s)\big\}\big|\mathcal{F}^W_s \vee \mathcal{F}^V_s\Big) = e^{-1/2\lambda^2(t-s)}e^{-1/2\mu^2(t-s)}$$

which implies that u'_t and u''_t are independent Wiener processes, since both have continuous trajectories.

Example 4.10. What SDE does the process (W_t is the Wiener process)

$$\xi_t = \frac{e^{W_t}}{1 + \int_0^t e^{W_s}ds}$$

satisfy ? Let $X_t = e^{W_t}$ and $Y_t = \int_0^t e^{W_s}ds$, then

$$dX_t = e^{W_t}dW_t + \frac{1}{2}e^{W_t}dt = \frac{1}{2}X_t dt + X_t dW_t$$

and

$$dY_t = e^{W_t}dt = X_t dt$$

Denote $\xi_t = f(X_t, Y_t)$, where $f = x/(1+y)$. Clearly

$$f_x = \frac{1}{1+y}, \quad f_y = \frac{-x}{(1+y)^2}.$$

By vector Ito formula

$$d\xi_t = \frac{1}{1+Y_t}dX_t - \frac{X_t}{(1+Y_t)^2}dY_t = \frac{1}{1+Y_t}\big(\frac{1}{2}X_t dt + X_t dW_t\big) - \frac{X_t}{(1+Y_t)^2}X_t dt =$$
$$\frac{1}{2}\xi_t dt + \xi_t dW_t - \xi_t^2 dt = \xi_t(1/2 - \xi_t)dt + \xi_t dW_t.$$

5. Applications

5.1. PDE for the marginal density of diffusions. Consider the scalar SDE

$$dX_t = a(X_t)dt + b(X_t)dW_t \tag{5.1}$$

where a and b are twice continuously differentiable functions and W is a Wiener process. Suppose that this equation is solved subject to $X_0 = \eta$, which has smooth probability density $p_0(x)$. Below we give a heuristic derivation for the probability density

$$p_t(x) = \frac{\partial}{\partial x}P(X_t \le x).$$

Let's assume that this density exists and is twice continuously differentiable as well. Fix a bounded compactly supported function f: by Ito formula

$$df(X_t) = f'(X_t)a(X_t)dt + f'(X_t)b(X_t)dW_t + \frac{1}{2}f''(X_t)b^2(X_t)dt.$$

Thus

$$Ef(X_t) = Ef(X_0) + \int_0^t E\big\{f'(X_s)a(X_s) + \frac{1}{2}f''(X_s)b^2(X_s)\big\}ds =$$
$$\int_{\mathbb{R}} f(u)p_0(u)du + \int_0^t \int_{\mathbb{R}} \big\{f'(u)a(u) + \frac{1}{2}f''(u)b^2(u)\big\}p_s(u)duds.$$

Integration by parts gives

$$\int_{\mathbb{R}} \big\{f'(u)a(u) + \frac{1}{2}f''(u)b^2(u)\big\}p_s(u)du =$$
$$-\int_{\mathbb{R}} f(u)\frac{\partial}{\partial u}\{a(u)p_s(u)\}du - \int_{\mathbb{R}} \frac{1}{2}f'(u)\frac{\partial}{\partial u}\{b^2(u)p_s(u)\}du =$$
$$-\int_{\mathbb{R}} f(u)\frac{\partial}{\partial u}\{a(u)p_s(u)\}du + \int_{\mathbb{R}} \frac{1}{2}f(u)\frac{\partial^2}{\partial u^2}\{b^2(u)p_s(u)\}du$$

So

$$\int_{\mathbb{R}} f(u)p_t(u)du = \int_{\mathbb{R}} f(u)p_0(u)du +$$
$$\int_0^t \int_{\mathbb{R}} f(u)\Big[-\frac{\partial}{\partial u}\{a(u)p_s(u)\} + \frac{1}{2}\frac{\partial^2}{\partial u^2}\{b^2(u)p_s(u)\}\Big]duds$$

which by arbitrariness of f implies

$$p_t(x) = p_0(x) + \int_0^t \Big[-\frac{\partial}{\partial u}\{a(u)p_s(u)\} + \frac{1}{2}\frac{\partial^2}{\partial u^2}\{b^2(u)p_s(u)\}\Big]ds$$

or in differential form

$$\frac{\partial}{\partial t}p_t(x) = -\frac{\partial}{\partial x}\{a(x)p_t(x)\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{b^2(x)p_t(x)\}.$$

The latter is known as the Fokker-Planck or the forward Kolmogorov equation.

Exercise 5.1. *Consider the linear equation from Example* (4.5)*, subject to Gaussian random variable $X_0 = \eta$. The FPK PDE reads*

$$\frac{\partial}{\partial t}p_t(x) = -\frac{\partial}{\partial x}\{axp_t(x)\} + \frac{b^2}{2}\frac{\partial^2}{\partial x^2}\{p_t(x)\} =$$
$$-ap_t(x) - ax\frac{\partial}{\partial x}p_t(x) + \frac{b^2}{2}\frac{\partial^2}{\partial x^2}\{p_t(x)\}$$

Verify that the solution is the Gaussian density with mean m_t and variance V_t, satisfying the equations derived in Example (4.5)*.*

5.2. Filtering of linear diffusions. Consider the signal/observation pair of processes $(X_t, Y_t)_{t\ge 0}$, generated by the linear system

$$dX_t = aX_tdt + bdW_t$$
$$dY_t = AX_tdt + BdV_t$$

subject to $Y_0 = X_0 = 0$, where a, A and $b, B > 0$ are known parameters and W and V are independent Wiener processes. Suppose that X_t is estimated by the linear filter of the form

$$d\widehat{X}_t = a\widehat{X}_tdt + \gamma\big(dY_t - A\widehat{X}_tdt\big), \tag{5.2}$$

subject to $\widehat{X}_0 = 0$. Choose γ so that the filter is stable in the sense $Q_\infty(\gamma) = \lim_{t\to\infty} E\big(X_t - \widehat{X}_t\big)^2$ exists and finite and the steady state filtering error $Q_\infty(\gamma)$ is minimal.

Define $\Delta_t = X_t - \widehat{X}_t$, then

$$d\Delta_t = a\Delta_tdt + bdW_t - \gamma\big(A\Delta_tdt + BdV_t\big) = \big(a - \gamma A\big)\Delta_tdt + bdW_t - \gamma BdV_t$$

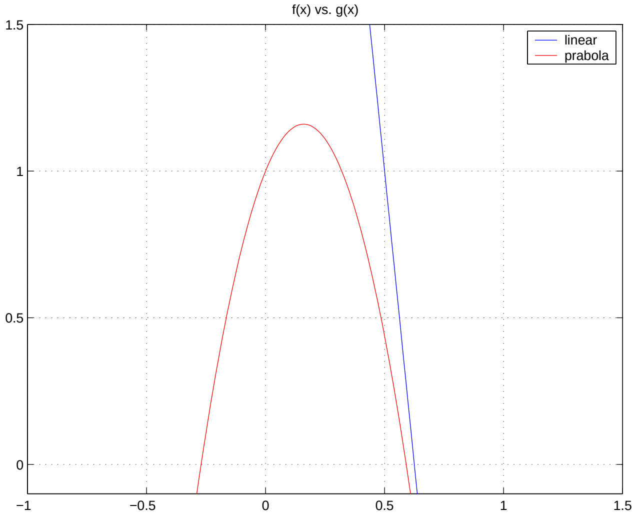

FIGURE 1. $f(x)$ versus $g(x)$

and

$$d(\Delta_t^2) = 2\Delta_t d\Delta_t + (b^2 + \gamma^2 B^2)dt = \\ 2\big(a - \gamma A\big)\Delta_t^2 dt + 2\Delta_t(bdW_t - \gamma BdV_t) + (b^2 + \gamma^2 B^2)dt.$$

Thus $Q_t = E\Delta_t^2$ satisfies

$$\dot{Q}_t = 2\big(a - \gamma A\big)Q_t + b^2 + \gamma^2 B^2.$$

The latter is a linear equation and it is stable (i.e. $Q_\infty = \lim_{t\to\infty} Q_t < \infty$ exists) if $a - \gamma A < 0$. In this case Q_∞ solves the linear equation

$$2\big(a - \gamma A\big)Q_\infty + b^2 + \gamma^2 B^2 = 0.$$

Now γ is to be chosen to minimize $Q_\infty(\gamma)$ under constraint $a - \gamma A < 0$.

Consider the function $f(x) = 2(a - \gamma A)x + b^2 + \gamma^2 B^2$ and note that (examine this claim geometrically)

$$f(x) = 2ax + b^2 - 2\gamma Ax + \gamma^2 B^2 \geq 2ax + b^2 - \frac{A^2 x^2}{B^2} := g(x)$$

It can be verified directly that the equation $g(x)$ always has a positive root P (even when $a > 0$, i.e. when the diffusion X escapes to infinity as $t \to \infty$). Since $f(x) \geq g(x)$ and the parabola $g(x)$ is concave it follows $Q_\infty \geq P$ (again refer the geometry Figure 1 for intuition). The lower bound P is attained, i.e. $Q_\infty = P$, when both equations

$$2(a - \gamma A)Q_\infty + b^2 + \gamma^2 B^2 = 0$$

$$2aQ_\infty + b^2 - \frac{A^2 Q_\infty^2}{B^2} = 0$$

are simultaneously satisfied for some γ. The appropriate solution is

$$\gamma^\circ = \frac{AQ_\infty}{B^2}.$$

In this case $a-\gamma^\circ A = -(b^2+[\gamma^\circ]^2B^2)/(2Q_\infty) < 0$ as required (recall that $Q_\infty > 0$). The minimal steady state error for this gain is found from

$$aQ_\infty + b^2 - \frac{A^2Q_\infty^2}{B^2} = 0.$$

The equations derived above are the special case of the Kalman-Bucy filter:

Theorem 5.2. *Let (X, Y) satisfy the equations*

$$\begin{aligned} dX_t &= a_tX_tdt + b_tdW_t \\ dY_t &= A_tX_tdt + B_tdV_t \end{aligned}$$

subject to the Gaussian vector (X_0, Y_0), where a_t, b_t, A_t and $B_t \ge C > 0$ are square integrable functions and W and V are independent Wiener processes, independent of (X_0, Y_0). The conditional mean $\widehat{X}_t = E(X_t|Y_0^t)$ and covariance $Q_t = E\big((X_t - \widehat{X}_t)^2|Y_0^t\big)$ satisfy

$$d\widehat{X}_t = a_t\widehat{X}_tdt + \frac{A_tQ_t}{B_t^2}\big(dY_t - A_t\widehat{X}_tdt\big) \tag{5.3}$$

$$\dot{Q}_t = 2a_tQ_t + b_t^2 - \frac{A_t^2Q_t^2}{B_t^2} \tag{5.4}$$

subject to $\widehat{X}_0 = E(X_0|Y_0)$ and $Q_0 = E(X_0 - \widehat{X}_0)^2$.

General References on Probability Theory

[1] J.L. Doob. *Stochastic Processes.* Wiley, 1953.

[2] R. Durrett. *Stochastic calculus : a practical introduction.* CRC press, 1996.

[3] W. Feller. *An introduction to probability theory and its applications.* Wiley, 1971.

[4] I.I. Gikhman and A.V. Skorokhod. *Introduction to the Theory of Random Processes.* Dover Publications, 1996.

[5] B. Oksendal. *Stochastic differential equatons: an introduction with applicatons.* Springer, 2007.

[6] Yu.A. Rozanov. *Stationary random processes.* Holden-Day, Inc., 1967.

[7] Z. Schuss. *Theory and applications of stochastic differential equations.* Wiley, 1980.

[8] A. Shiryaev. *Probability.* Springer, 1995.

References on Nonlinear Filtering

[9] B.D.O. Anderson and J.B. Moore. *Optimal Filtering.* Dover Publications, 2005.

[10] R. J. Elliott, L. Aggoun, and J.B. Moore. *Hidden Markov Models: Estimation and Control.* Springer-Verlag, 1995.

[11] A.H. Jazwinski. *Stochastic processes and filtering theory.* Dover Publications, 2007.

[12] R. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D. J. Basic Engrg.*, 82:35–45, 1960.

[13] R. Liptser and A. Shiryaev. *Statistics of random processes: General theory (I) and applications (II).* Springer, 2000.

[14] E. Wong and B. Hajek. *Stochastic processes in engineering systems.* Springer, 1985.