# Introduction to Nonlinear Filtering

P. Chigansky

# Contents

# Preface

These lecture notes were prepared for the course, taught by the author at the Faculty of Mathematics and CS of the Weizmann Institute of Science. The course is intended as the first encounter with stochastic calculus with a nice engineering application: estimation of signals from the noisy data. Consequently the rigor and generality of the presented theory is often traded for intuition and motivation, leaving out many interesting and important developments, either recent or classic. Any suggestions, remarks, bug reports etc. are very welcome and can be sent to `pavel.chigansky@weizmann.ac.il`.

*Pavel Chigansky*
*WIS, February 2005*

# Instead of Introduction

### An example

Consider a simple random walk on integers (e.g. randomly moving particle)

$$X_j = X_{j-1} + \varepsilon_j, \quad j \in \mathbb{Z}_+ \tag{1}$$

starting from the origin, where $\varepsilon_j$ is a sequence of independent random signs $\mathrm{P}(\varepsilon_j = \pm 1) = 1/2$, $j \geq 1$. Suppose the position of the particle at time $j$ is to be estimated (guessed or *filtered*) on the basis of the noisy observations

$$Y_i = X_i + \xi_i, \quad i = 1, ..., j \tag{2}$$

where $\xi_j$ is a sequence of independent identically distributed (i.i.d.) random variables (so called discrete time *white noise*) with Gaussian distribution, i.e.

$$P\big(\xi_j \in [a, b]\big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} du, \quad \forall j \geq 1.$$

Formally an estimate is a rule, which assigns a real number[1]to any outcome of the observation vector $Y_{[1,j]} = \{Y_1, ..., Y_j\}$, in other words it is a map $\varphi_j(y) : \mathbb{R}^j \mapsto \mathbb{R}$. How different guesses are compared ? One possible way is to require minimal square error on average, i.e. $\varphi_j$ is considered better than $\psi_j$ if

$$\mathrm{E}\big(X_j - \varphi_j(Y_{[1,j]})\big)^2 \leq \mathrm{E}\big(X_j - \psi_j(Y_{[1,j]})\big)^2, \tag{3}$$

where $\mathrm{E}(\cdot)$ denotes *expectation*, i.e. average with respect to all possible outcomes of the experiment, e.g. for $j = 1$

$$\mathrm{E}\big(X_1 - \varphi_1(Y_1)\big)^2 = \frac{1}{2} \int_{-\infty}^{\infty} \Big(\big(1 - \varphi_1(1+u)\big)^2 + \big(-1 - \varphi_1(-1+u)\big)^2\Big) \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

Note that even if (3) holds,

$$\big(X_j - \varphi_j(Y_{[1,j]})\big)^2 > \big(X_j - \psi_j(Y_{[1,j]})\big)^2$$

may happen in an individual experiment. However this is *not expected*[2]to happen.

Once the criteria (3) is accepted, we would like to find the best (optimal) estimate. Let's start with the simplest guess

$$\widetilde{X}_j := \widetilde{\varphi}_j(Y_{[1,j]}) \equiv Y_j.$$

The corresponding mean square error is

$$\widetilde{P}_j = \mathrm{E}(X_j - Y_j)^2 = \mathrm{E}(X_j - X_j - \xi_j)^2 = \mathrm{E}\xi_j^2 = 1.$$

[1]Though $X_j$ takes only integer values, we allow a guess to take real values, i.e. "soft" decisions are admissible

[2]think of an unfair coin with probability of heads equal to 0.99: it is not expected to give tails, though it may!

This simple estimate does not take into account past observations and hence potentially can be improved by using more data. Let's try

$$\tilde{\tilde{X}}_j = \frac{Y_j + Y_{j-1}}{2}.$$

The corresponding mean square error is

$$\begin{aligned}\tilde{\tilde{P}}_j =& \mathrm{E}\left(X_j - \tilde{\tilde{X}}_j\right)^2 = \mathrm{E}\left(X_j - \frac{Y_j + Y_{j-1}}{2}\right)^2 = \\ & \mathrm{E}\left(X_j - \frac{X_j + X_{j-1} + \xi_{j-1} + \xi_j}{2}\right)^2 = \\ & \mathrm{E}\big((X_j - X_{j-1})/2 - (\xi_{j-1} + \xi_j)/2\big)^2 = \\ & \mathrm{E}\big(\varepsilon_j/2 - (\xi_{j-1} + \xi_j)/2\big)^2 = 1/4 + 1/2 = 0.75\end{aligned}$$

which is an improvement by 25% ! Let's try to increase the "memory" of the estimate:

$$\begin{aligned}\mathrm{E}\left(X_j - \frac{Y_j + Y_{j-1} + Y_{j-2}}{3}\right)^2 &= ... = \\ & \mathrm{E}\left(\frac{2}{3}\varepsilon_j + \frac{1}{3}\varepsilon_{j-1} + \frac{\xi_j + \xi_{j-1} + \xi_{j-2}}{3}\right)^2 = \frac{4}{9} + \frac{1}{9} + \frac{3}{9} \approx 0.89\end{aligned}$$

i.e. the error increased! The reason is that the estimate gives the "old" and the "new" measurements the same weights - it is reasonable to rely more on the latest samples. So what is the optimal way to weigh the data ?

It turns out that the optimal estimate can be generated very efficiently by the difference equation ($j \ge 1$)

$$\widehat{X}_j = \widehat{X}_{j-1} + P_j\big(Y_j - \widehat{X}_{j-1}\big), \quad \widehat{X}_0 = 0 \tag{4}$$

where $P_j$ is a sequence of numbers, generated by

$$P_j = \frac{P_{j-1} + 1}{P_{j-1} + 2}, \quad P_0 = 0. \tag{5}$$

Let's us calculate the mean square error. The sequence $\Delta_j := X_j - \widehat{X}_j$ satisfies

$$\Delta_j = \Delta_{j-1} + \varepsilon_j - P_j\big(\Delta_{j-1} + \varepsilon_j + \xi_j\big) = \big(1 - P_j\big)\Delta_{j-1} + (1 - P_j)\varepsilon_j - P_j\xi_j$$

and thus $\widehat{P}_j = \mathrm{E}\Delta_j^2$ satisfies

$$\widehat{P}_j = \big(1 - P_j\big)^2\widehat{P}_{j-1} + (1 - P_j)^2 + P_j^2, \quad \widehat{P}_0 = 0$$

where the independence of $\varepsilon_j$, $\xi_j$ and $\Delta_{j-1}$ has been used. Note that the sequence $P_j$ satisfies the identity (just expand the right hand side using (5))

$$P_j = \big(1 - P_j\big)^2 P_{j-1} + (1 - P_j)^2 + P_j^2, \quad \widehat{P}_0 = 0.$$

So the difference $\widehat{P}_j - P_j$ obeys the linear time varying equation

$$\big(\widehat{P}_j - P_j\big) = \big(1 - P_j\big)^2\big(\widehat{P}_{j-1} - P_{j-1}\big), \quad t \ge 1$$

and since $\widehat{P}_0 - P_0 = 0$, it follows that $\widehat{P}_j \equiv P_j$ for all $j \ge 0$, or in other words $P_j$ is the mean square error, corresponding to $\widehat{X}_j$ ! Numerically we get

| $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $P_j$ | 0.5 | 0.6 | 0.6154 | 0.6176 | 0.618 |

In particular $P_j$ converges to the limit $P_\infty$, which is the unique positive root of the equation

$$P = \frac{P+1}{P+2} \qquad \Longrightarrow P_\infty = \sqrt{5}/2 - 1/2 \approx 0.618.$$

This is nearly a 40% improvement over the accuracy of $\widetilde{X}_j$! As was mentioned before, no further improvement is possible among linear estimates.

What about nonlinear estimates? Consider the simplest nonlinear estimate of $X_1$ from $Y_1$: guess 1 if $Y_1 \geq 0$ and $-1$ if $Y_1 < 0$, i.e.

$$\check{X}_1 = \mathrm{sign}(Y_1).$$

The corresponding error is

$$\check{P}_1 = \mathrm{E}\big(X_1 - \check{X}_1\big)^2 = \frac{1}{2}\mathrm{E}\big(1 - \mathrm{sign}(1+\xi_1)\big)^2 + \frac{1}{2}\mathrm{E}\big(-1 - \mathrm{sign}(-1+\xi_1)\big)^2 =$$
$$\frac{1}{2}2^2\mathrm{P}(\xi_1 \leq -1) + \frac{1}{2}2^2\mathrm{P}(\xi_1 \geq 1) = 4\mathrm{P}(\xi_1 \geq 1) = 4\frac{1}{\sqrt{2\pi}}\int_1^\infty e^{-u^2/2}du \approx 0.6346$$

which is even worse than the linear estimate $\widehat{X}_1$! Let's try the estimate

$$\bar{X}_1 = \tanh(Y_1),$$

which can be regarded as a "soft" sign. The corresponding mean square error is

$$\bar{P}_1 = \mathrm{E}\big(X_1 - \bar{X}_1\big)^2 =$$
$$\frac{1}{2}\int_\infty^\infty \Big[\big(1 - \tanh(u+1)\big)^2 + \big(1 + \tanh(u-1)\big)^2\Big]\frac{1}{\sqrt{2\pi}}\exp\{-u^2/2\}du \approx 0.4496$$

which is the best estimate up to now (in fact it is the best possible!).

How can we compute the best nonlinear estimate of $X_j$ efficiently (meaning recursively)? Let $\rho_j(i)$, $i \in \mathbb{Z}$, $j \geq 0$ be generated by the nonlinear recursion

$$\rho_j(i) = \exp\{Y_j i - i^2/2\}\big(\rho_{j-1}(i-1) + \rho_{j-1}(i+1)\big), \quad j \geq 1 \tag{6}$$

subject to $\rho_0(0) = 1$ and $\rho_0(i) = 0$, $i \neq 0$. Then the best estimate of $X_j$ from the observations $\{Y_1, ..., Y_j\}$ is given by

$$\bar{X}_j = \frac{\sum_{i=-\infty}^\infty i\rho_j(i)}{\sum_{i=-\infty}^\infty \rho_j(i)}. \tag{7}$$

How good is it ? The exact answer is hard to calculate. E.g. the empirical mean square error $\bar{P}_{100}$ is around 0.54 (note that it should be less than 0.618 and greater than 0.4496).

How the same problem could be formulated in continuous time, i.e. when the time parameter (denoted in this case by $t$) can be any nonnegative real number ? The signal defined in (1) is a Markov[3]chain with integer values, starting from zero and making equiprobable transitions to the nearest neighbors. Intuitively the

[3]Recall that a sequence called Markov if the conditional distribution of $X_j$, given the "history" $\{X_0, ..., X_{j-1}\}$, depends only on the last entry $X_{j-1}$ and not on the whole path. Verify this property for the sequence defined by (1).

analogous Markov chain in continuous time should satisfy

$$\mathrm{P}\big(X_{t+\varepsilon}=i|X_s, 0\le s\le t\big)=\begin{cases}1-2\varepsilon, & i=X_t\\ \varepsilon & i=X_t\pm 1\\ 0 & \text{otherwise}\end{cases} \tag{8}$$

for sufficiently small $\varepsilon>0$. In other words, the process is not expected to jump on short time intervals and eventually jumps to one of the nearest neighbors. It turns out that (8) uniquely defines a stochastic process. For example it can be modelled by a pair of independent Poisson processes. Let $(\tau_n)_{n\in\mathbb{Z}_+}$ be an i.i.d sequence of positive random variables with standard exponential distribution

$$\mathrm{P}\big(\tau_n\le t\big)=\begin{cases}1-e^{-t}, & t\ge 0\\ 0, & t<0\end{cases} \tag{9}$$

Then a standard Poisson process is defined as [4]

$$\Pi_t=\max\{n:\sum_{\ell=1}^n \tau_\ell\le t\},$$

Clearly $\Pi_t$ starts at zero ($\Pi_0=0$) and increases, jumping to the next integer at random times separated by $\tau_\ell$'s. Let $\Pi_t^-$ and $\Pi_t^+$ be a pair of independent Poisson process. Then the process

$$X_t=\Pi_t^+-\Pi_t^-,\quad t\ge 0$$

satisfies (8). Remarkably the exponential distribution is the only one which can lead to a Markov process.

To define an analogue of $Y_t$, the concept of "white noise" is to be introduced in continuous time. The origin of the term "white noise" stems from the fact that the spectral density of an i.i.d. sequence $\xi$ is flat, i.e.

$$S_\xi(\lambda):=\sum_{j=-\infty}^{\infty}\mathrm{E}\xi_0\xi_j e^{-i\lambda j}=\sum_{j=-\infty}^{\infty}\delta(j)e^{-i\lambda j}=1\quad\forall\lambda\in(-\pi,\pi].$$

So any random sequence with flat spectral density is called (discrete time) white noise and its variance is recovered by integration over the spectral density

$$E\xi_t^2=\frac{1}{2\pi}\int_{-\pi}^{\pi}1d\lambda=1.$$

The same definition leads to a paradox in continuous time: suppose that a stochastic process have flat spectral density, then it should have infinite variance[5]

$$E\xi_t^2=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\lambda=\infty.$$

This paradox is resolved if the observation process is defined as

$$Y_t=\int_0^t X_s ds+W_t, \tag{10}$$

[4] with convention $\sum_{\ell=1}^0=0$.

[5] recall that the spectral density for continuous time processes is supported on the whole real line, rather than being condensed to $(-\pi,\pi]$ as in the case of sequences.

where $W=(W_t)_{t\ge 0}$ is the Wiener process or mathematical Brownian motion. The Wiener process is characterized by the following properties: $W_0=0$, the trajectories of $W_t$ are continuous functions and it has independent increments with

$$\mathrm{E}\big(W_t|W_u, u\le s\big)=W_s, \quad \mathrm{E}\big((W_t-W_s)^2|W_u,u\le s\big)=t-s.$$

Why is the model (10) compatible with the "white noise" notion? Introduce the process

$$\nu_t^\Delta=\frac{W_t-W_{t-\Delta}}{\Delta}, \quad \Delta>0$$

Then $\mathrm{E}\nu_t^\Delta=0$ and[6]

$$\mathrm{E}\nu_t^\Delta\nu_s^\Delta=\frac{1}{\Delta^2}\mathrm{E}\big(W_t-W_{t-\Delta}\big)\big(W_s-W_{s-\Delta}\big)=\frac{1}{\Delta^2}\begin{cases}\Delta-|t-s|, & |t-s|\le\Delta\\ 0, & |t-s|\ge\Delta\end{cases}.$$

So the process $\nu_t^\Delta$ is stationary with the correlation function

$$R_\nu^\Delta(\tau)=\frac{1}{\Delta^2}\begin{cases}\Delta-|\tau|, & |\tau|\le\Delta\\ 0, & |\tau|\ge\Delta\end{cases}.$$

For small $\Delta>0$, $R_\nu^\Delta(\tau)$ approximates the Dirac $\delta(\tau)$ in the sense that for any continuous and compactly supported test function $\varphi(\tau)$

$$\int_{-\infty}^{\infty}\varphi(\tau)R_\nu^\Delta(\tau)d\tau\xrightarrow{\Delta\to 0}\varphi(0)$$

and if the limit process $\nu:=\lim_{\Delta\to 0}\nu_t^\Delta$ existed, it would have flat spectral density as required. Then the observation process (10) would contain the same information as

$$\dot{Y}_t=X_t+\nu_t,$$

with $\nu_t$ being the derived white noise. Of course, this is only an intuition and $\nu_t$ does not exists as a limit in any reasonable sense (e.g. its variance at any point $t$ grows to infinity with $\Delta\to 0$, which is the other side of the "flat spectrum" paradox). It turns out that the axiomatic definition of the Wiener process leads to very unusual properties of its trajectories. For example, almost all trajectories of $W_t$, though continuous, are not differentiable at any point.

After a proper formulation of the problem is found, what would be the analogs of the filtering equations (4)-(5) and (6)-(7)? Intuitively, instead of the difference equations in discrete time, we should obtain differential equations in continuous time, e.g.

$$\dot{\widehat{X}}_t=P_t\big(\dot{Y}_t-\widehat{X}_t\big), \quad \widehat{X}_0=0.$$

However the right hand side of this equation involves derivative of $Y_t$ and hence also of $W_t$, which is impossible in view of aforementioned irregularity of the latter. Then instead of differential equations we may write (and implement!) the corresponding integral equation

$$\widehat{X}_t=\int_0^t P_s dY_s-\int_0^t P_s\widehat{X}_s ds,$$

[6]Note that $\mathrm{E}W_tW_s=\min(t,s):=t\wedge s$ for all $t,s\ge 0$.

where the first integral may be interpreted as Stieltjes integral with respect to $Y_t$ or alternatively defined (in the spirit of integration by parts formula) as

$$\int_0^t P_s dY_s := Y_t P_t - \int_0^t Y_s \dot{P}_s ds.$$

Such a definition is correct, since the integrand function is deterministic and differentiable ($Y_t$ turns to be Riemann integrable as well). Of course, we should define precisely what is the solution of such equation and under what assumptions it exists and is unique. The optimal linear filtering equations then can be derived:

$$\begin{aligned} \widehat{X}_t &= \int_0^t P_s\big(dY_s - \widehat{X}_s ds\big) \\ \dot{P}_t &= 2 - P_t^2, \quad P_0 = 0. \end{aligned} \tag{11}$$

Now what about the nonlinear filter? The equations should realize a nonlinear map of the data and thus their right hand side would require integration of some stochastic process with respect to $Y_t$. This is where the classical integration theory completely fails! The reason is again irregularity of the Wiener process - it has unbounded variation! Thus the construction similar to Stieltjes integral would not lead to a well defined limit in general. The foundations of the integration theory with respect to the Wiener process were laid by K.Itô in 40's. The main idea is to use Stieltjes like construction for a specific class of integrands (non-anticipating processes). In terms of Itô integral the nonlinear filtering formulae are[7]

$$\rho_t(i) = \delta(i) + \int_0^t \big(\rho_s(i+1) + \rho_s(i-1) - 2\rho_s(i)\big)ds + \int_0^t i\rho_s(i)dY_s \tag{12}$$

and

$$\bar{X}_t = \frac{\sum_{m=-\infty}^{\infty} m\rho_t(m)}{\sum_{\ell=-\infty}^{\infty} \rho_t(\ell)}.$$

This example is the particular case of the *filtering problem*, which is the main subject of these lectures:

> *Given a pair of random process $(X_t, Y_t)_{t\ge 0}$ with known statistical description, find a recursive realization for the optimal in the mean square sense estimate of the signal $X_t$ on the basis of the observed trajectory $\{Y_s, s \le t\}$ for each $t \ge 0$.*

### The brief history of the problem

The estimation problem of signals from the noisy observations dates back to Gauss (the beginning of XIX century), who studied the motion of planets on the basis of celestial observations by means of his least squares method. In the modern probabilistic framework the filtering type problems were addresses independently by N.Wiener (documented in the monograph [**26**]) and A.Kolmogorov ([**20**]). Both treated linear estimation of stationary processes via the spectral representation. Wiener's work seems to be partially motivated by the radar tracking problems and gunfire control. This part of the filtering theory won't be covered in this course and the reader is referred to the classical text [**28**] for further exploration.

---

[7]From now on $\delta(i)$ denotes the Kronecker symbol, i.e. $\delta(i) = \begin{cases} 1 & i = 0 \\ 0 & i \neq 0 \end{cases}$

The Wiener-Kolmogorov theory in many cases had serious practical limitation - all the processes involved are assumed to be stationary. R.Kalman and R.Bucy (1960-61) [**13**], [**14**] addressed the same problem from a different perspective: using state space representation they relaxed the stationarity requirement and obtained closed form recursive formulae realizing the best estimator. The celebrated Kalman-Bucy filter today plays a central role in various engineering applications (communications, signal processing, automatic control, etc.) Besides being of significant practical importance, the Kalman-Bucy approach stimulated much research in the theory of stochastic processes and their applications in control and estimation. The state space approach allowed nonlinear extensions of the filtering problem. The milestone contributions in this field are due to H.Kushner [**29**], R. Stratonovich [**37**] and Fujisaki, Kallianpur and Kunita [**10**] (the dynamic equations for conditional probability distribution), Kallianpur and Striebel [**17**] (Bayes formula for white noise observations), M. Zakai [**41**] (reference measure approach to nonlinear filtering).

There are several excellent books and monographs on the subject including R.Lipster and A.Shiryaev [**21**] (the main reference for the course), G.Kallianpur [**15**], S.Mitter [**23**], G. Kallianpur and R.L. Karandikar [**16**] (a different look at the problem), R.E. Elliott, L. Aggoun and J.B. Moore [**8**]. Classic introductory level texts are B.Anderson and J. Moore [**1**] and A. Jazwinski [**12**].

CHAPTER 1

# Probability preliminaries

*Probability theory is simply a branch of measure theory, with its own special emphasis and field of application* (J.Doob).

This chapter gives a summary of the probabilistic notions used in the course, which are assumed to be familiar (the book [**34**] is the main reference hereafter).

## 1. Probability spaces

The basic object of probability theory is the *probability space* $(\Omega, \mathcal{F}, \mathrm{P})$, where $\Omega$ is a collection of elementary events $\omega \in \Omega$ (points), $\mathcal{F}$ is an appropriate family of all considered events (or sets) and P is the probability measure on $\mathcal{F}$. While $\Omega$ can be quite arbitrary, $\mathcal{F}$ and P are required to satisfy certain properties to provide sufficient applicability of the derived theory. The mainstream of the probability research relies on the axioms, introduced by A.Kolmogorov in 30's (documented in [**19**]). $\mathcal{F}$ is required to be a $\sigma$-algebra of events, i.e. to be closed under countable intersections and compliment operations[1]

$$\begin{aligned}
&\Omega \in \mathcal{F} \\
&A \in \mathcal{F} \implies \Omega / A \in \mathcal{F} \\
&A_n \in \mathcal{F} \implies \cap_{n=1}^{\infty} A_n \in \mathcal{F}
\end{aligned}$$

P is a $\sigma$-additive nonnegative measure on $\mathcal{F}$ normalized to one, in other words P is a set function $\mathcal{F} \mapsto [0,1]$, satisfying

$$\begin{aligned}
&\mathrm{P}\left(\biguplus_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathrm{P}(A_n), \quad A_n \in \mathcal{F} && \sigma\textit{-additivity} \\
&\mathrm{P}(\Omega) = 1 && \textit{normalization.}
\end{aligned}$$

Here are some examples of probability spaces:

**1.1. A finite probability space.** For example

$$\begin{aligned}
&\Omega := \{1, 2, 3\} \\
&\mathcal{F} := \{\emptyset, 1, 2, 3, 1 \cup 2, 1 \cup 3, 2 \cup 3, \Omega\} \\
&\mathrm{P}(A) = \sum_{\omega_\ell \in A} 1/3, \quad \forall A \in \mathcal{F}
\end{aligned}$$

Note that the $\sigma$-algebra $\mathcal{F}$ coincides with the (finite) algebra, generated by the points of $\Omega$ and P is defined on $\mathcal{F}$ by specifying its values for each $\omega \in \Omega$, i.e. $\mathrm{P}(1) = \mathrm{P}(2) = \mathrm{P}(3) = 1/3$.

[1] These imply that $\mathcal{F}$ is also closed under countable unions as well, i.e. $A_n \in \mathcal{F} \implies \cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

EXAMPLE 1.1. Tossing a coin $n$ times. The elementary event $\omega$ is a string of $n$ zero-one bits, i.e. the sampling space $\Omega$ consists of $2^n$ points. $\mathcal{F}$ consists of all subsets of $\Omega$ (how many are there?). The probability measure is defined (on $\mathcal{F}$) by setting $\mathrm{P}(\omega) = 2^{-n}$, for all $\omega \in \Omega$. What is the probability of the event $A =$ *"the first bit of a string is one"*?

$$\mathrm{P}(A) = \mathrm{P}\big(\omega : \omega(1) = 1\big) = \sum_{\ell:\omega_\ell(1)=1} 2^{-n} = 1/2 \quad \text{(by symmetry)}.$$

■

**1.2. The Lebesgue probability space** $([0,1], \mathcal{B}, \lambda)$**.** Here $\mathcal{B}$ denotes the *Borel* $\sigma$-algebra on $[0,1]$, i.e. the minimal $\sigma$-algebra containing all open sets from $[0,1]$. It can be generated by the algebra of all intervals. The probability measure $\lambda$ is uniquely defined (by Caratheodory extension theorem) on $\mathcal{B}$ by its restriction e.g. to the algebra of semi-open intervals

$$\lambda((a,b]) = b - a, \quad b \ge a.$$

Similarly a probability space is defined on $\mathbb{R}$ (or $\mathbb{R}^d$). The probability measure in this case can be defined by any nondecreasing right continuous (why?) nonnegative function $F : \mathbb{R} \mapsto [0,1]$, satisfying $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$:

$$\mathrm{P}\big((a,b]\big) = F(b) - F(a).$$

What is the analogous construction in $\mathbb{R}^d$ ?

EXAMPLE 1.2. An infinite series of coin tosses. The elementary event is an infinite binary sequence or equivalently [2] a point in $[0,1]$, i.e. $\Omega = [0,1]$. For the event $A$ from the previous example:

$$\lambda(A) = \lambda\big(\omega : \omega(1) = 1\big) = \lambda\big(\omega \ge 1/2\big) = 1/2.$$

■

**1.3. The space of infinite sequences.** $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \mathrm{P})$. The Borel $\sigma$-algebra $\mathcal{B}(\mathbb{R}^\infty)$ can be generated by the cylindrical sets of the form

$$A = \{x \in \mathbb{R}^\infty : x_{i_1} \in (a_1, b_1], ..., x_{i_n} \in (a_n, b_n]\}, \quad b_i \ge a_i$$

The probability $\mathrm{P}$ is uniquely defined on $\mathcal{B}(\mathbb{R}^\infty)$ by a *consistent* family of probability measures $\mathrm{P}^n$ on $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$, $n \ge 1$ (Kolmogorov theorem), i.e. if $\mathrm{P}^n$ satisfies

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

EXAMPLE 1.3. Let $p(x,y)$ be a measurable[3] $\mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}_+$ nonnegative function, such that

$$\int_{\mathbb{R}} p(x,y)dy = 1, \quad \text{a.s.}\forall x$$

and let $\nu(x)$ be a probability density (i.e. $\nu(x) \ge 0$ and $\int_{\mathbb{R}} \nu(x)dx = 1$). Define a family of probability measures on $\mathcal{B}(\mathbb{R}^{n+1})$ by the formula:

$$\mathrm{P}^{n+1}\big(A_0 \times ... \times A_n\big) = \int_{A_0} ... \int_{A_n} \nu(x_1)p(x_1,x_2)...p(x_{n-1},x_n)dx_1...dx_n.$$

[2] some sequences represent the same numbers (e.g. 0.10000... and 0.011111...), but there are countably many of them, which can be neglected while calculating the probabilities.

[3] measurability with respect to the Borel field is mean by default

This family is consistent:

$$\mathrm{P}^{n+1}\big(A_0 \times ... \times \mathbb{R}\big) = \int_{A_0} ... \int_{\mathbb{R}} \nu(x_1)p(x_1,x_2)...p(x_{n-1},x_n)dx_1...dx_n =$$

$$\int_{A_0} ... \int_{\mathbb{R}} \nu(x_1)p(x_1,x_2)...p(x_{n-2},x_{n-1})dx_1...dx_{n-1} := \mathrm{P}^n\big(A_1 \times ... \times A_{n-1}\big),$$

and hence there is a unique probability measure P (on $\mathcal{B}(\mathbb{R}^\infty)$), such that

$$\mathrm{P}(A) = \mathrm{P}^n(A_n) \quad \forall A_n \in \mathcal{B}(\mathbb{R}^n), \quad n = 1, 2, ...$$

The constructed measure is called Markov. ■

## 2. Random variables and random processes

A random variable is a measurable function on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ to a metric space (say $\mathbb{R}$ hereon), i.e a map $X(\omega) : \Omega \mapsto \mathbb{R}$, such that

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

Due to measurability requirement $X$ (the argument $\omega$ is traditionally omitted) induces a measure on $\mathcal{B}(\mathbb{R})$:

$$\mathrm{P}_X(B) := \mathrm{P}\big(\omega : X(\omega) \in B\big), \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

The function $F_X : \mathbb{R} \mapsto [0,1]$

$$F_X(x) = \mathrm{P}_X\big((-\infty, x]\big) = P(X \le x), \quad x \in \mathbb{R}$$

is called the *distribution* function of $X$. Note that by definition $F_X(x)$ is a right-continuous function.

A stochastic (random) process is a collection of random variables $X_n(\omega)$ on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, parameterized by time $n \in \mathbb{Z}_+$. Equivalently, a stochastic process can be regarded as a probability measure (or probability distribution) on the space of real valued sequences. The finite dimensional distributions $F_X^n : \mathbb{R}^n \mapsto [0,1]$ of $X$ are defined as

$$F_X^n(x_1, ..., x_n) = \mathrm{P}\big(X_1 \le x_1, ..., X_n \le x_n\big), \quad n \ge 1$$

The existence of a random process with given finite dimensional distributions is guaranteed by the Kolmogorov theorem if and only if the family of probability measures on $\mathbb{R}^n$, corresponding to $F_X^n$, is consistent. Then one may realize $X$ as a coordinate process on an appropriate probability space, in which case the process is called canonical.

## 3. Expectation and its properties

The expectation of a real random variable $X \ge 0$, defined on $(\Omega, \mathcal{F}, \mathrm{P})$, is the Lebesgue integral of $X$ with respect to the measure P, i.e. the limit (either finite of infinite)

$$\mathrm{E}X = \int_\Omega X(\omega)\mathrm{P}(d\omega) := \lim_{n\to\infty} \mathrm{E}X_n,$$

where $X_n$ is an approximation of $X$ by simple ("piecewise constant") functions, e.g.

$$X_n(\omega) = \sum_{\ell=1}^{n2^n} \frac{\ell-1}{2^n}\mathbf{1}\left\{\frac{\ell-1}{2^n} \le X(\omega) < \frac{\ell}{2^n}\right\} + n\mathbf{1}\big(X(\omega) \ge n\big) \tag{1.1}$$

for which

$$\mathrm{E}X_n := \sum_{\ell=1}^{n2^n} \frac{\ell-1}{2^n}\mathrm{P}\left\{\frac{\ell-1}{2^n} \le X(\omega) < \frac{\ell}{2^n}\right\} + n\mathrm{P}\big(X(\omega) \ge n\big)$$

is defined. Such limit always exists and is independent of the specific choice of the approximating sequence. For a general random variable, taking values with both signs, the expectation is defined[4]

$$\mathrm{E}X = \mathrm{E}(0 \wedge X) - \mathrm{E}(0 \vee X) := \mathrm{E}X^+ + \mathrm{E}X^-$$

if at least one of the terms is finite. If $EX$ exists and is finite $X$ is said to be Lebesgue integrable with respect to P. Note that expectation can be also realized on the induced probability space, e.g.

$$\mathrm{E}X = \int_\Omega X(\omega)\mathrm{P}(d\omega) = \int_{\mathbb{R}} x\mathrm{P}_X(dx) = \int_{-\infty}^{\infty} x dF_X(x).$$

(the latter stands for the Lebesgue-Stieltjes integral).

Example 1.4. Consider a random variable $X(\omega) = \omega^2$ on the Lebesgue probability space. Then

$$\mathrm{E}X = \int_{[0,1]} \omega^2 \lambda(d\omega) = 1/3$$

Another way to calculate $\mathrm{E}X$ is to find its distribution function:

$$F_X(x) = \mathrm{P}(X(\omega) \le x) = \mathrm{P}(\omega^2 \le x) = \mathrm{P}(\omega \le \sqrt{x}) = \begin{cases} 0 & x < 0 \\ \sqrt{x} & 0 \le x < 1 \\ 1 & 1 \le x \end{cases}$$

and then to calculate the integral

$$\mathrm{E}X = \int_{-\infty}^{\infty} x dF_X(x) = \int_{[0,1]} x d(\sqrt{x}) = 1 - \int_{[0,1]} \sqrt{x} dx = 1/3.$$

■

The expectation have the following basic properties:

(A) if $\mathrm{E}X$ is well defined, then $\mathrm{E}cX = c\mathrm{E}X$ for any $c \in \mathbb{R}$
(B) if $X \le Y$ P-a.s., then $\mathrm{E}X \le \mathrm{E}Y$
(C) if $\mathrm{E}X$ is well defined, then $\mathrm{E}X \le \mathrm{E}|X|$
(D) if $\mathrm{E}X$ is well defined, then $\mathrm{E}X\mathbf{1}_A$ is well defined for all $A \in \mathcal{F}$. If $\mathrm{E}X$ is finite, so is $\mathrm{E}X\mathbf{1}_A$
(E) if $\mathrm{E}|X| < \infty$ and $\mathrm{E}|Y| < \infty$, then $\mathrm{E}(X + Y) = \mathrm{E}X + \mathrm{E}Y$
(F) if $X = 0$ P-a.s., then $\mathrm{E}X = 0$
(G) if $X = Y$ P-a.s. and $\mathrm{E}|X| < \infty$, $\mathrm{E}|Y| < \infty$, then $\mathrm{E}X = \mathrm{E}Y$
(H) if $X \ge 0$ and $\mathrm{E}X = 0$, then $X = 0$ P-a.s.

The random variables $\{X_1, ..., X_n\}$ are *independent* if for any subset of indices $\{i_1, ..., i_m\} \subseteq \{1, ..., n\}$ and Borel sets $A_1, ..., A_m$,

$$\mathrm{P}\big(X_{i_1} \in A_1, ..., X_{i_m} \in A_m\big) = \mathrm{P}\big(X_{i_1} \in A_1\big)...\mathrm{P}\big(X_{i_m} \in A_m\big).$$

For example $X$ and $Y$ are independent if

$$P(X \in A, Y \in B) = P(X \in A)P(X \in B)$$

[4] $a \wedge b = a \min b$ and $a \vee b = a \max b$

for any Borel sets $A$ and $B$. Note that pairwise independence is not enough in general for independence of e.g. three random variable. Also note that independence is the joint property of random variables and the measure P. Being dependent under P, the same random variables may be independent under another measure $\widetilde{\mathrm{P}}$ (defined on the same probability space).

The characteristic function of $X$ is the Fourier transform of its distribution, i.e.

$$\varphi_X(\lambda) := \mathrm{E}\exp\big(i\lambda X\big), \quad \lambda \in \mathbb{R}.$$

The independence can be alternatively formulated via distribution or characteristic functions (How?).

## 4. Convergence of random variables

A sequence of random variables $X_n$ converges to a random variable $X$

(1) P-*almost surely*, if $\mathrm{P}\big(\lim_{n\to\infty} X_n = X\big) = 1$.
(2) *in probability* P if $\lim_{n\to\infty} \mathrm{P}\big(|X_n - X| \ge \varepsilon\big) = 0, \quad \forall \varepsilon > 0$.
(3) *in* $\mathbb{L}^p(\Omega, \mathcal{F}, \mathrm{P})$, $p \ge 1$ if $\lim_{n\to\infty} \mathrm{E}\big|X_n - X\big|^p = 0$ and $\mathrm{E}|X|^p < \infty$.
(4) *weakly* or *in law*, if for any bounded and continuous function $f$

$$\lim_{n\to\infty} \mathrm{E}f(X_n) = \mathrm{E}f(X).$$

Other types of convergence are possible, but these are used mostly. Note that the convergence in law is actually not a convergence of the random variables, but rather of their distributions: for example, an i.i.d. random sequence converges in law and does not converge in any other aforementioned sense.

The following implications can be easily verified

$$\left.\begin{array}{l} \xrightarrow{\mathrm{P}-a.s.} \\ \xrightarrow{\mathbb{L}^p} \end{array}\right\} \implies \xrightarrow{\mathrm{P}} \implies \xrightarrow{w}$$

while the other are wrong in general.

Example 1.5. Let $X_n$ be an sequence of independent random variables with

$$\mathrm{P}(X_n = 1) = 1/n, \quad \mathrm{P}(X_n = 0) = 1 - 1/n.$$

Then $X_n$ converges in probability: for $0 < \varepsilon < 1$

$$\mathrm{P}\big(X_n \ge \varepsilon\big) = \mathrm{P}\big(X_n = 1\big) = 1 - 1/n \to 0.$$

However it doesn't converge P-a.s. Let $A_n = \{X_n = 1\}$ and let

$$A_{i.o} = \bigcap_{n\ge 0} \bigcup_{m\ge n} A_m$$

i.e. the event of $X_n$ being equal to 1 infinitely often. Let us show that $P(A_{i.o.}) = 1$ or alternatively[5] $P(A_{i.o.}^c) = 0$:

$$P(A_{i.o.}^c) = \mathrm{P}\big(\bigcup_{n\ge 0} \bigcap_{m\ge n} A_m^c\big) \le \sum_n \mathrm{P}\big(\bigcap_{m\ge n} A_m^c\big).$$

[5] the superscript $c$ stands for compliment, i.e. $A^c = \Omega \backslash A$.

For any fixed $n$ and $\ell \geq 1$, due to independence

$$\mathrm{P}\Big(\bigcap_{m=n}^{n+\ell} A_m^c\Big) = \prod_{m=n}^{n+\ell} \mathrm{P}\big(A_m^c\big) = \prod_{m=n}^{n+\ell} (1-1/m) = \exp\left\{\sum_{m=n}^{n+\ell} \log(1-1/m)\right\} \leq$$

$$\exp\left\{-\sum_{m=n}^{n+\ell} 1/m\right\} \xrightarrow{\ell\to\infty} 0,$$

so, by continuity of P (which is implied by $\sigma$-additivity!),

$$\mathrm{P}\Big(\bigcap_{m=n}^{\infty} A_m^c\Big) = 0$$

for any $n$ and thus $\mathrm{P}(A_{i.o}) = 1$, meaning that $X_n$ does not converge to zero P-a.s. Is the independence crucial? Yes ! For example take dependent (why?) random variables on the Lebesgue space, $X_n = \mathbf{1}(\omega \leq 1/n)$. Then the set $\{\omega : X_n(\omega) \not\to 0\}$ is just the singleton $\{0\}$, whose probability is zero and so $\mathrm{P}(X_n \to 0) = 1!$ ■

This example is the particular case of the Borel-Cantelli lemmas:

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

and

$$\left.\begin{array}{l} \sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty \\ A_n \text{ are independent} \end{array}\right\} \implies \mathrm{P}(A_{i.o.}) = 1.$$

## 5. Conditional expectation

The *conditional expectation* of a random variable $X \geq 0$ with respect to a $\sigma$-algebra $\mathcal{G}$ (under measure P) is a *random variable*, denoted by $\mathrm{E}(X|\mathcal{G})(\omega)$, which satisfies the properties:

(1) $\mathrm{E}(X|\mathcal{G})(\omega)$ is $\mathcal{G}$-measurable
(2) $\mathrm{E}\big(X - \mathrm{E}(X|\mathcal{G})\big)\mathbf{1}_A = 0$ for all $A \in \mathcal{G}$.

The conditional expectation is characterized by these properties up to almost sure equivalence.

EXAMPLE 1.6. Suppose $\mathcal{G}$ is generated by a finite partition $G$ of $\Omega$, i.e.

$$G = \{G_1, ..., G_n\}, \quad G_i \cap G_j = \emptyset, \quad \biguplus_{j=1}^{n} G_j = \Omega.$$

Then (why?)

$$\mathrm{E}(X|\mathcal{G}) = \sum_{\ell=1}^{n} \frac{\mathrm{E}X\mathbf{1}_{G_\ell}(\omega)}{\mathrm{P}(G_\ell)} \mathbf{1}_{G_\ell}(\omega),$$

where $0/0 = 0$ is understood. ■

For a general random variable $X$, $\mathrm{E}(X|\mathcal{G}) = \mathrm{E}(X^+|\mathcal{G}) + \mathrm{E}(X^-|\mathcal{G})$ if no uncertainty of the type "$\infty - \infty$" arises.

The inverse images $\{\omega : Y \in B\}$, $B \in \mathcal{B}(\mathbb{R})$ of a random variable $Y$ form a $\sigma$-algebra $\mathcal{G}^Y \subseteq \mathcal{F}$. The conditional expectation $\mathrm{E}(X|\mathcal{G}^Y)$ is usually denoted by $\mathrm{E}(X|Y)$ and there always exists[6] a Borel function $\psi$, such that $\mathrm{E}(X|Y) = \psi(Y)$.

[6]if the space is not too wild, e.g. Polish spaces are OK

The conditional expectation enjoys the same properties as the expectation and in addition

$(A')$ if $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then $\mathrm{E}\Big(\mathrm{E}\big(X|\mathcal{G}_2\big)\big|\mathcal{G}_1\Big) = \mathrm{E}(X|\mathcal{G}_1)$ P-a.s.

$(B')$ if $\mathrm{E}|X|^2 < \infty$, then for any Borel function $g$

$$\mathrm{E}\big(X - \mathrm{E}(X|Y)\big)^2 \le \mathrm{E}\big(X - g(Y)\big)^2. \tag{1.2}$$

The latter property can be interpreted as optimality in the mean square sense of the conditional expectation among all estimates of $X$ given the realization of $Y$ (cf. (7) from the previous chapter). The main tool in calculation of the conditional expectation is the Bayes formula.

EXAMPLE 1.7. Let $(X, Y)$ be a pair of random variables and suppose that their distribution has density (with respect to the Lebesgue measure on the plane), i.e.

$$\mathrm{P}\big(X \le x, Y \le y\big) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv.$$

Suppose that $\mathrm{E}X^2 < \infty$, then (why?)

$$\mathrm{E}(X|Y)(\omega) = \frac{\int_{\mathbb{R}} x f\big(x, Y(\omega)\big)dx}{\int_{\mathbb{R}} f\big(u, Y(\omega)\big)du}.$$

■

Later we will prove and use a more abstract version of this formula.

## 6. Gaussian random variables

A random variable $X$ is Gaussian with mean $\mathrm{E}X = m$ and variance $\mathrm{E}(X - \mathrm{E}X)^2 = \sigma^2 > 0$ if

$$F_X(x) := \mathrm{P}(X \le x) = \int_{(-\infty,x]} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(u-m)^2}{2\sigma^2}\right\} du.$$

The corresponding characteristic function is

$$\varphi_X(\lambda) = \mathrm{E}e^{i\lambda X} = \exp\left\{im\lambda - \frac{1}{2}\sigma^2\lambda^2\right\}.$$

If the latter is taken as definition (since there is a one to one correspondence between $F_X$ and $\varphi_X$), then the degenerate case $\sigma = 0$ is included as well, i.e. a constant random variable can be considered as Gaussian.

Analogously a random vector $X$ with values in $\mathbb{R}^d$ is Gaussian with mean $\mathrm{E}X = m \in \mathbb{R}^d$ and the covariance matrix $C = \mathrm{E}(X - \mathrm{E}X)(X - \mathrm{E}X)^* \ge 0$ (semi positive definite matrix!), if

$$\varphi_X(\lambda) = \mathrm{E}\exp\{iX^*\lambda\} = \exp\left\{im^*\lambda - \frac{1}{2}\lambda^* C\lambda\right\}.$$

Finally a random process is Gaussian if its finite dimensional distributions are Gaussian. Gaussian processes have a special place in probability theory and in particular in filtering as we will see soon.

## Exercises

(**1**) Let $A_n$ $n \geq 1$ be a sequence of events and define the events $A_{i.o} = \bigcap_{n\geq 1}\bigcup_{m\geq n} A_n$ and $A_e = \bigcup_{n\geq 1}\bigcap_{m\geq n} A_n$.
- (a) Explain the terms "i.o." (*infinitely often*) and "e" (*eventually*) in the notations.
- (b) Is $A_{i.o} = A_e$ if $A_n$ is a monotonous sequence, i.e. $A_n \subseteq A_{n+1}$ or $A_n \supseteq A_{n+1}$ for all $n \geq 1$?
- (c) Explain the notation $A_{i.o} = \overline{\lim}_{n\to\infty} A_n$ and $A_e = \underline{\lim}_{n\to\infty} A_n$.
- (d) Show that $A_e \subseteq A_{i.o.}$

(**2**) Prove the Borel-Cantelli lemmas.

(**3**) Using the Borel-Cantelli lemmas, show that
- (a) a sequence $X_n$ converging in probability has a subsequence converging almost surely
- (b) a sequence $X_n$, converging exponentially[7] in $\mathbb{L}^2$, converges P-a.s.
- (c) if $X_n$ is an i.i.d. sequence with $\mathrm{E}|X_1| < \infty$, then $X_n/n$ converges to zero P-a.s.
- (d) if $X_n$ is an i.i.d. sequence with $\mathrm{E}|X_1| = \infty$, then $\overline{\lim}_{n\to\infty} |X_n|/n = \infty$ P-a.s.
- (e) show that if $X_n$ is a standard Gaussian i.i.d. sequence, then
$$\overline{\lim_{n\to\infty}} |X_n|/\sqrt{2\ln n} = 1, \quad \mathrm{P}-a.s.$$

(**4**) Give counterexamples to the following false implications:
- (a) convergence in probability implies $\mathbb{L}^2$ convergence
- (b) P-a.s. convergence implies $\mathbb{L}^2$ convergence
- (c) $\mathbb{L}^2$ convergence implies P-a.s. convergence

(**5**) Let $X$ be a r.v. with uniform distribution on $[0,1]$ and $\eta$ be a r.v. given by:
$$\eta = \begin{cases} X & X \leq 0.5 \\ 0.5 & X > 0.5 \end{cases}$$
Find $\mathrm{E}(X|\eta)$.

(**6**) Let $\xi_1, \xi_2, ...$ be an i.i.d. sequence. Show that:
$$\mathrm{E}(\xi_1|S_n, S_{n+1}, ...) = \frac{S_n}{n}$$
where $S_n = \xi_1 + ... + \xi_n$.

(**7**)
- (a) Consider an event $A$ that does not depend on itself, i.e. $A$ and $A$ are independent. Show that:
$$\mathrm{P}\{A\} = 1 \quad \text{or} \quad \mathrm{P}\{A\} = 0$$
- (b) Let $A$ be an event so that $\mathrm{P}\{A\} = 1$ or $\mathrm{P}\{A\} = 0$. Show that $A$ and any other event $B$ are independent.
- (c) Show that a r.v. $\xi(\omega)$ doesn't depend on itself if and only if $\xi(\omega) \equiv$ const.

(**8**) Consider the Lebesgue probability space and define a sequence of random variables[8]
$$X_n(\omega) = \lfloor 2^n\omega \rfloor \mod 2.$$

---
[7] i.e. $\mathrm{E}|X_n - X|^2 \leq C\rho^n$ for all $n \geq 1$ with $C \geq 0$ and $\rho \in [0,1)$

[8] $\lfloor x \rfloor$ is the integer part of $x$

Show that $X_n$ is an i.i.d. sequence.

**(9)** Let $Y$ be a nonnegative random variable with probability density:

$$f(y) = \frac{1}{\sqrt{2\pi}} \frac{e^{-y/2}}{\sqrt{y}}, \quad y \geq 0$$

Define the conditional density of $X$ given fixed $Y$:

$$f(x;Y) = \frac{\sqrt{Y}}{\sqrt{2\pi}} e^{-Yx^2/2},$$

i.e. for any bounded function $f$

$$\mathrm{E}\big(f(X)|Y\big) = \int_{\mathbb{R}} f(x) f(x;Y) dx.$$

Does the formula $\mathrm{E}\big(\mathrm{E}(X|Y)\big) = \mathrm{E}X$ hold ? If not, explain why.

**(10)** Give an example of three dependent random variables, any two of which are independent.

**(11)** Let $X$ and $Z$ be a pair of independent r.v. and $\mathrm{E}|X| < \infty$. Then $\mathrm{E}(X|Z) = EX$ with probability one. Does the formula

$$\mathrm{E}(X|Z,Y) = \mathrm{E}(X|Y)$$

holds for an arbitrary $Y$ ?

**(12)** Let $X_1$ and $X_2$ be two random variables such that, $\mathrm{E}X_1 = 0$ and $\mathrm{E}X_2 = 0$. Suppose we can find a linear combination $Y = X_1 + \alpha X_2$, which is independent of $X_2$. Show that $\mathrm{E}(X_1|X_2) = -\alpha X_2$.

**(13)** Show that the coordinate (canonical) process on the space from Example 1.3 is Markov, i.e.

$$\mathrm{E}\big(f(X_n)|X_0, ..., X_{n-1}\big) = \mathrm{E}(f(X_n)|X_{n-1}), \quad \mathrm{P}-a.s. \tag{1.3}$$

for any bounded Borel $f$.

**(14)** Let $(X_n)_{n\geq 0}$ be a sequence of random variables and let

$$\mathcal{G}_{\leq n} := \sigma\{X_0, ..., X_n\} \quad \text{and} \quad \mathcal{G}_{>n} := \sigma\{X_n, X_{n+1}, ...\}.$$

Show that the Markov property (1.3) is equivalent to the property

$$\mathrm{E}\big(\pi\phi|X_n\big) = \mathrm{E}\big(\phi|X_n\big)\mathrm{E}\big(\pi|X_n\big)$$

for all bounded random variables $\pi$ and $\phi$, $\mathcal{G}_{\leq n}$ and $\mathcal{G}_{>n}$ measurable respectively. In other words, the Markov property is equivalently stated as "the future and the past are conditionally independent, given the present".

**(15)** Let $X$ and $Y$ be i.i.d. random variables with finite variance and twice differentiable probability density. Show that if $X+Y$ and $X-Y$ are independent, then $X$ and $Y$ are Gaussian.

**(16)** Let $X_1$, $X_2$ and $X_3$ be independent standard Gaussian random variables. Show that

$$\frac{X_1 + X_2 X_3}{\sqrt{1+X_3^2}}$$

is a standard Gaussian random variable as well.

**(17)** Let $\{X_1, X_2, X_3, X_4\}$ be a Gaussian vector with zero mean. Show that

$$\mathrm{E}X_1X_2X_3X_4 = \mathrm{E}X_1X_2\mathrm{E}X_3X_4 + \mathrm{E}X_1X_3\mathrm{E}X_2X_4 + \mathrm{E}X_1X_4\mathrm{E}X_2X_3.$$

Recall that the moments, if exist, can be recovered from the derivatives of the characteristic function at $\lambda = 0$.

(**18**) Let $f(x)$ be a probability density function of a Gaussian variable, i.e:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-a)^2/(2\sigma^2)}$$

Define a function:

$$g_n(x_1, ..., x_n) = \Big[\prod_{j=1}^{n} f(x_j)\Big]\Big[1 + \prod_{k=1}^{n}(x_k - a)f(x_k)\Big], \quad (x_1, ..., x_n) \in \mathbb{R}^n$$

(a) Show that $g_n(x_1, ..., x_n)$ is a valid probability density function of some random vector $X = (X_1, ..., X_n)$.
(b) Show that any subvector of $X$ is Gaussian, while $X$ is not Gaussian.

(**19**) Let $f(x, y, \rho)$ be a two dimensional Gaussian probability density, so that the marginal densities have zero means and unit variances and the correlation coefficient is $\rho = \int_{\mathbb{R}}\int_{\mathbb{R}} xyf(x, y, \rho) = \rho$. Form a new density:

$$g(x, y) = c_1 f(x, y, \rho_1) + c_2 f(x, y, \rho_2)$$

with $c_1 > 0, c_2 > 0, c_1 + c_2 = 1$.
(a) Show that $g(x, y)$ is a valid probability density of some vector $\{X, Y\}$.
(b) Show that each of the r.v. $X$ and $Y$ is Gaussian.
(c) Show that $c_1$, $c_2$ and $\rho_1$, $\rho_2$ can be chosen so that $\mathrm{E}XY = 0$. Are $X$ and $Y$ independent ?

CHAPTER 2

# Linear filtering in discrete time

Consider a pair of random square integrable random variables $(X, Y)$ on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Suppose that the following (*second order*) probabilistic description of the pair is available,

$$\begin{aligned}
&\mathrm{E}X, \quad \mathrm{E}Y \\
&\operatorname{cov}(X) := \mathrm{E}(X-\mathrm{E}X)^2, \ \operatorname{cov}(Y) := \mathrm{E}(Y-\mathrm{E}Y)^2, \\
&\operatorname{cov}(X,Y) := \mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)
\end{aligned}$$

and it is required to find a pair of constants $a_0'$ and $a_1'$, such that

$$\mathrm{E}\big(X - a_0' - a_1'Y\big)^2 \le \mathrm{E}\big(X - a_0 - a_1 Y\big)^2, \quad \forall a_0, a_1 \in \mathbb{R}.$$

The corresponding estimate $\widehat{X} = a_0' + a_1'Y$ is then the optimal linear estimate of $X$, given the observation (realization) of $Y$. Clearly

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

where $\operatorname{cov}(Y) > 0$ was assumed. The minimizers are

$$a_1' = \frac{\operatorname{cov}(X,Y)}{\operatorname{cov}(Y)}, \quad a_0' = \mathrm{E}X - \frac{\operatorname{cov}(X,Y)}{\operatorname{cov}(Y)}\mathrm{E}Y.$$

If $\operatorname{cov}(Y) = 0$ (or in other words $Y = \mathrm{E}Y$, P-a.s.), then the same arguments lead to

$$a_1' = 0, \quad a_0' = \mathrm{E}X.$$

So among all linear functionals of $\{1, Y\}$ (or *affine* functionals of $Y$), there is the unique optimal one [1], given by

$$\widehat{X} := \mathrm{E}X + \operatorname{cov}(X,Y)\operatorname{cov}^{\oplus}(Y)\big(Y - \mathrm{E}Y\big) \tag{2.1}$$

with the corresponding minimal mean square error

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

where for any $x \in \mathbb{R}$

$$x^{\oplus} = \begin{cases} x^{-1}, & x \ne 0 \\ 0, & x = 0 \end{cases}$$

[1] Note that the pair of optimal coefficients $(a_0', a_1')$ is unique, though the random variable $a_0' + a_1'Y(\omega)$ can be modified on a P-null set, without altering the mean square error. So the uniqueness of the estimate is understood as uniqueness among the equivalence classes of random variables (all equal with probability one within each class)

Note that the optimal estimate satisfies the *orthogonality* property

$$\begin{aligned} &\mathrm{E}\big(X-\widehat{X}\big)1=0\\ &\mathrm{E}\big(X-\widehat{X}\big)Y=0 \end{aligned}$$

that is the residual estimation error is orthogonal to any linear functional of the observations. It is of course not a coincidence, since (2.1) is nothing but the orthogonal projection of $X$ on the linear space spanned by the random variables 1 and $Y$. These simple formulae are the basis for the optimal linear filtering equations of Kalman-Bucy and Bucy ([**13**], [**14**]), which is the subject of this chapter.

## 1. The Hilbert space $\mathbb{L}^2$, orthogonal projection and linear estimation

Let $\mathbb{L}^2(\Omega,\mathcal{F},\mathrm{P})$ (or simply $\mathbb{L}^2$) denote the space of all square integrable random variables [2]. Equipped with the scalar product

$$\langle X,Y\rangle := \mathrm{E}XY, \quad X,Y\in\mathbb{L}^2$$

and the induced norm $\|X\|:=\sqrt{\langle X,X\rangle}$, $\mathbb{L}^2$ is a Hilbert space (i.e. infinite dimensional Euclidian space). Let $\mathcal{L}$ be a closed linear subspace of $\mathbb{L}^2$ (either finite or infinite dimensional at this point). Then

Theorem 2.1. *For any $X\in\mathbb{L}^2$, there exists a unique*[3] *random variable $\widehat{X}\in\mathcal{L}$, called the* orthogonal projection *and denoted by $\widehat{\mathrm{E}}(X|\mathcal{L})$, such that*

$$\mathrm{E}\big(X-\widehat{X}\big)=\inf_{\widetilde{X}\in\mathcal{L}}\mathrm{E}\big(X-\widetilde{X}\big)^2 \tag{2.2}$$

*and*

$$\mathrm{E}\big(X-\widehat{X}\big)Z=0 \tag{2.3}$$

*for any $Z\in\mathcal{L}$.*

Proof. Let $d^2:=\inf_{\widetilde{X}\in\mathcal{L}}\mathrm{E}\big(X-\widetilde{X}\big)^2$ and let $\widetilde{X}_j$ be the sequence in $\mathcal{L}$, such that $d_j^2:=\mathrm{E}(X-\widetilde{X}_j)^2\to d^2$. Then $\widetilde{X}_j$ is a Cauchy sequence in $\mathbb{L}^2$

$$\begin{aligned}\mathrm{E}\big(\widetilde{X}_j-\widetilde{X}_i\big)^2 =&2\mathrm{E}\big(X-\widetilde{X}_i\big)^2+2\mathrm{E}\big(X-\widetilde{X}_j\big)^2-4\mathrm{E}\left(X-\frac{\widetilde{X}_i+\widetilde{X}_j}{2}\right)^2\le\\ &2\mathrm{E}\big(X-\widetilde{X}_i\big)^2+2\mathrm{E}\big(X-\widetilde{X}_j\big)^2-4d^2\xrightarrow{i,j\to\infty}0,\end{aligned}$$

where the inequality holds since $\widetilde{X}_i+\widetilde{X}_j\in\mathcal{L}$. The space $\mathbb{L}^2$ is complete and so $\widetilde{X}_j$ converges to a random variable $\widetilde{X}_\infty$ in $\mathbb{L}^2$ and since $\mathcal{L}$ is closed, $\widetilde{X}_\infty\in\mathcal{L}$. Then

$$\|X-X_\infty\|=\sqrt{\mathrm{E}\big(X-X_\infty\big)^2}\le\sqrt{\mathrm{E}(X-\widetilde{X}_j)^2}+\sqrt{\mathrm{E}\big(\widetilde{X}_j-X_\infty\big)^2}\xrightarrow{j\to\infty}d$$

and so $X_\infty$ is a version of $\widehat{X}$. To verify (2.3), fix a $t\in\mathbb{R}$: then for any $Z\in\mathcal{L}$

$$\mathrm{E}\big(X-\widehat{X}\big)^2\le\mathrm{E}\big(X-\widehat{X}-tZ\big)^2 \quad\Longrightarrow\quad 2t\mathrm{E}(X-\widehat{X})Z\le t^2\mathrm{E}Z^2$$

The latter cannot hold for arbitrary small $t$ unless $\mathrm{E}(X-\widehat{X})Z=0$. Finally $\widehat{X}$ is unique: suppose that $\widehat{X}'\in\mathcal{L}$ satisfies (2.2) as well, then

$$\mathrm{E}(X-\widehat{X}')^2=\mathrm{E}(X-\widehat{X}+\widehat{X}-\widehat{X}')^2=\mathrm{E}(X-\widehat{X})^2+\mathrm{E}(\widehat{X}-\widehat{X}')^2$$

[2] more precisely of the equivalence classes with respect the relation $\mathrm{P}(X=Y)=1$

[3] actually a unique equivalence class

which implies $\mathrm{E}(\widehat{X}-\widehat{X}')^2=0$ or $\widehat{X}=\widehat{X}'$, P-a.s. □

The orthogonal projection satisfies the following main properties:

(a) $\mathrm{E}\widehat{\mathrm{E}}(X|\mathcal{L})=\mathrm{E}X$
(b) $\widehat{\mathrm{E}}(X|\mathcal{L})=X$ if $X\in\mathcal{L}$ and $\widehat{\mathrm{E}}(X|\mathcal{L})=0$ if $X\perp\mathcal{L}$
(c) linearity: for $X_1,X_2\in\mathbb{L}^2$ and $c_1,c_2\in\mathbb{R}$,

$$\widehat{\mathrm{E}}(c_1X_1+c_2X_2|\mathcal{L})=c_1\widehat{\mathrm{E}}(X_1|\mathcal{L})+c_2\widehat{\mathrm{E}}(X_2|\mathcal{L})$$

(d) for two linear subspaces $\mathcal{L}_1\subseteq\mathcal{L}_2$,

$$\widehat{\mathrm{E}}(X|\mathcal{L}_1)=\widehat{\mathrm{E}}\big(\widehat{\mathrm{E}}(X|\mathcal{L}_2)\big|\mathcal{L}_1\big)$$

Proof. (a)-(c) are obvious from the definition. (d) holds, if

$$\mathrm{E}\Big(X-\widehat{\mathrm{E}}\big(\widehat{\mathrm{E}}(X|\mathcal{L}_2)\big|\mathcal{L}_1\big)\Big)Z=0$$

for all $Z\in\mathcal{L}_1$, which is valid since

$$\begin{aligned}\mathrm{E}\Big(X-\widehat{\mathrm{E}}\big(\widehat{\mathrm{E}}(X|\mathcal{L}_2)\big|\mathcal{L}_1\big)\Big)Z=&\\ \mathrm{E}\Big(X-\widehat{\mathrm{E}}(X|\mathcal{L}_2)\Big)Z+\Big(\widehat{\mathrm{E}}(X|\mathcal{L}_2)-\widehat{\mathrm{E}}\big(\widehat{\mathrm{E}}(X|\mathcal{L}_2)\big|\mathcal{L}_1\big)\Big)Z=0 \quad (2.4)\end{aligned}$$

where the first term vanishes since $\mathcal{L}_1\subseteq\mathcal{L}_2$. □

Theorem 2.1 suggests that the optimal in the mean square sense estimate of a random variable $X\in\mathbb{L}^2$ from the observation (realization) of the collection of random variables $Y_j\in\mathbb{L}^2$, $j\in\mathcal{J}\subseteq\mathbb{Z}_+$ is given by the orthogonal projection of $X$ onto $\mathcal{L}^Y_{\mathcal{J}}:=\overline{\mathrm{span}}\{Y_j,j\in\mathcal{J}\}$. While for finite $\mathcal{J}$ the explicit expression for $\widehat{\mathrm{E}}(X|\mathcal{L}^Y_{\mathcal{J}})$ is straightforward and is given in Proposition 2.2 below, calculation of $\widehat{\mathrm{E}}(X|\mathcal{L}^Y_{\mathcal{J}})$ in the infinite case is more involved. In this chapter the finite case is treated (still we'll need generality of Theorem 2.1 in continuous time case).

Proposition 2.2. *Let $X$ and $Y$ be random vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ with square integrable entries. Denote*[4] *by $\widehat{\mathrm{E}}(X|\mathcal{L}^Y)$ the orthogonal projection*[5] *of $X$ onto the linear subspace, spanned by the entries of $Y$ and 1. Then*[6]

$$\widehat{\mathrm{E}}(X|\mathcal{L}^Y)=\mathrm{E}X+\mathrm{cov}(X,Y)\,\mathrm{cov}(Y)^{\oplus}\big(Y-\mathrm{E}Y\big) \tag{2.5}$$

*and*

$$\mathrm{E}\big(X-\widehat{\mathrm{E}}(X|\mathcal{L}^Y)\big)\big(X-\widehat{\mathrm{E}}(X|\mathcal{L}^Y)\big)^*=\mathrm{cov}(X)-\mathrm{cov}(X,Y)\,\mathrm{cov}(Y)^{\oplus}\,\mathrm{cov}(Y,X), \tag{2.6}$$

*where $Q^{\oplus}$ stands for the generalized inverse of $Q$ (see* (2.8) *below).*

Proof. Let $A$ and $a$ be a matrix and a vector, such that $\widehat{\mathrm{E}}(X|\mathcal{L}^Y)=a+AY$. Then by Theorem 2.1 (applied componentwise!)

$$0=\mathrm{E}\big(X-a-AY\big)$$

[4] sometimes the notation $\widehat{\mathrm{E}}(X|Y)=\widehat{\mathrm{E}}(X|\mathcal{L}^Y)$ is used.

[5] Naturally the orthogonal projection of a random vector (on some linear subspace) is a vector of the orthogonal projections of its entries.

[6] the constant random variable 1 is always added to the observations, meaning that the expectations $\mathrm{E}X$ and $\mathrm{E}Y$ are known (available for the estimation procedure)

and

$$\begin{aligned} 0 =&\mathrm{E}\big(X-a-AY\big)\big(Y-\mathrm{E}Y\big)^* = \\ &\mathrm{E}\big(X-\mathrm{E}X-A(Y-\mathrm{E}Y)-a+\mathrm{E}X-A\mathrm{E}Y\big)\big(Y-\mathrm{E}Y\big)^* = \\ &\mathrm{cov}(X,Y)-A\,\mathrm{cov}(Y) \end{aligned} \tag{2.7}$$

If $\mathrm{cov}(Y)>0$, then (2.5) follows with $\mathrm{cov}(Y)^{\oplus}=\mathrm{cov}(Y)^{-1}$. If only $\mathrm{cov}(Y)\ge 0$, there exists a unitary matrix $U$ (i.e. $UU^*=I$) and a diagonal matrix $D\ge 0$, so that $\mathrm{cov}(Y)=UDU^*$. Define[7]

$$\mathrm{cov}(Y)^{\oplus} := UD^{\oplus}U^* \tag{2.8}$$

where $D^{\oplus}$ is a diagonal matrix with the entries

$$D^{\oplus}_{ii} = \begin{cases} 1/D_{ii}, & D_{ii}>0 \\ 0, & D_{ii}=0 \end{cases}. \tag{2.9}$$

Then

$$\begin{aligned} &\mathrm{cov}(X,Y)-\mathrm{cov}(X,Y)\,\mathrm{cov}(Y)^{\oplus}\,\mathrm{cov}(Y) = \\ &\qquad \mathrm{cov}(X,Y)U\big(I-D^{\oplus}D\big)U^* = \sum_{\ell:D_{\ell\ell}=0}\mathrm{cov}(X,Y)u_\ell u_\ell^* \end{aligned} \tag{2.10}$$

by the definition of $D^{\oplus}$, where $u_\ell$ is the $\ell$-th column of $U$. Clearly

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

and so

$$\mathrm{cov}(X,Y)u_\ell = \mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)^*u_\ell = 0,$$

i.e. (2.7) holds. The equation (2.6) is verified directly by substitution of (2.5) and using the obvious properties of the generalized inverse.

□

Remark 2.3. Note that if instead of (2.9), $D^{\oplus}$ were defined as

$$D^{\oplus}_{ii} = \begin{cases} 1/D_{ii}, & D_{ii}>0 \\ c, & D_{ii}=0 \end{cases}$$

with $c\ne 0$, the same estimate would be obtained.

## 2. Recursive orthogonal projection

Consider a pair of random processes $(X,Y)=(X_j,Y_j)_{j\in\mathbb{Z}_+}$ with entries in $\mathbb{L}^2$ and let $\mathcal{L}^Y_j=\overline{\mathrm{span}}\{1,Y_0,...,Y_j\}$. Calculation of the optimal estimate $\widehat{\mathrm{E}}\big(X_j|\mathcal{L}^Y_j\big)$ by the formulae of Proposition 2.2 would require inverting matrices of sizes, growing linearly with $j$. The following lemma is the key to a much more efficient calculation algorithm of the orthogonal projection. Introduce the notations

$$\widehat{X}_j := \widehat{\mathrm{E}}\big(X_j|\mathcal{L}^Y_j\big), \quad \widehat{X}_{j|j-1} := \widehat{\mathrm{E}}(X_j|\mathcal{L}^Y_{j-1}), \quad \widehat{Y}_{j|j-1} := \widehat{\mathrm{E}}(Y_j|\mathcal{L}^Y_{j-1})$$

$$P^X_j := \mathrm{E}\big(X_j-\widehat{X}_j\big)\big(X_j-\widehat{X}_j\big)^*, \quad P^X_{j|j-1} := \mathrm{E}\big(X_j-\widehat{X}_{j|j-1}\big)\big(X_j-\widehat{X}_{j|j-1}\big)^*$$

$$P^{XY}_{j|j-1} := \mathrm{E}\big(X_j-\widehat{X}_{j|j-1}\big)\big(Y_j-\widehat{Y}_{j|j-1}\big)^*, \quad P^Y_{j|j-1} := \mathrm{E}\big(Y_j-\widehat{Y}_{j|j-1}\big)\big(Y_j-\widehat{Y}_{j|j-1}\big)^*$$

Then

[7] this is the generalized inverse of Moore and Penrose, in the special case of nonnegative definite matrix. Note that it coincides (as should be) with the ordinary inverse if the latter exists.

PROPOSITION 2.4. *For $j \ge 1$*

$$\widehat{X}_j = \widehat{X}_{j|j-1} + P^{XY}_{j|j-1}\big[P^{Y}_{j|j-1}\big]^{\oplus}\big(Y_j - \widehat{Y}_{j|j-1}\big) \tag{2.11}$$

*and*

$$P^X_j = P^X_{j|j-1} - P^{XY}_{j|j-1}\big[P^{Y}_{j|j-1}\big]^{\oplus}P^{XY*}_{j|j-1}. \tag{2.12}$$

PROOF. To verify (2.11), check that

$$\eta := X_j - \widehat{X}_{j|j-1} + P^{XY}_{j|j-1}\big[P^{Y}_{j|j-1}\big]^{\oplus}\big(Y_j - \widehat{Y}_{j|j-1}\big)$$

is orthogonal to $\mathcal{L}^Y_j$. Note that $\eta$ is orthogonal to $\mathcal{L}^Y_{j-1}$ and so it suffices to show that $\eta \perp Y_j$ or equivalently $\eta \perp (Y_j - \widehat{Y}_{j|j-1})$:

$$\begin{aligned}\mathrm{E}\eta(Y_j - \widehat{Y}_{j|j-1}) = P^{XY}_{j|j-1} - P^{XY}_{j|j-1}\big[P^{Y}_{j|j-1}\big]^{\oplus}P^{Y}_{j|j-1} = \\ P^{XY}_{j|j-1}\Big(I - \big[P^{Y}_{j|j-1}\big]^{\oplus}P^{Y}_{j|j-1}\Big) = 0\end{aligned}$$

where the last equality is verified as in (2.10). The equation (2.12) is obtained similarly to (2.6). □

## 3. The Kalman-Bucy filter in discrete time

Consider a pair of processes $(X, Y) = (X_j, Y_j)_{j\ge 0}$, generated by the linear recursive equations ($j \ge 1$)

$$X_j = a_0(j) + a_1(j)X_{j-1} + a_2(j)Y_{j-1} + b_1(j)\varepsilon_j + b_2(j)\xi_j \tag{2.13}$$

$$Y_j = A_0(j) + A_1(j)X_{j-1} + A_2(j)Y_{j-1} + B_1(j)\varepsilon_j + B_2(j)\xi_j \tag{2.14}$$

where

* $X_j$ and $Y_j$ have values in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively
* $\varepsilon = (\varepsilon_j)_{j\ge 1}$ and $\xi = (\xi_j)_{j\ge 1}$ are orthogonal (discrete time) white noises with values in $\mathbb{R}^\ell$ and $\mathbb{R}^k$, i.e.

  $$\mathrm{E}\varepsilon_j = 0, \quad \mathrm{E}\varepsilon_j\varepsilon_i^* = \begin{cases} I, & i = j \\ 0, & i \ne j \end{cases} \in \mathbb{R}^{\ell\times\ell}$$

  $$\mathrm{E}\xi_j = 0, \quad \mathrm{E}\xi_j\xi_i^* = \begin{cases} I, & i = j \\ 0, & i \ne j \end{cases} \in \mathbb{R}^{k\times k}$$

  and

  $$\mathrm{E}\varepsilon_j\xi_i^* = 0 \quad \forall i, j \ge 0.$$
* the coefficients $a_0(j), a_1(j)$, etc. are deterministic (known) sequences of matrices of appropriate dimensions [8]. From here on we will omit the time dependence from the notation for brevity.
* the equations are solved subject to random vectors $X_0$ and $Y_0$, uncorrelated with the noises $\varepsilon$ and $\xi$, whose means and covariances are known.

[8]Note the customary abuse of notations, now time parameter is written in the parenthesis instead of subscript

Denote the optimal linear estimate of $X_j$, given $\mathcal{L}_j^Y = \overline{\mathrm{span}}\{1, Y_1, ..., Y_j\}$, by

$$\widehat{X}_j = \widehat{\mathrm{E}}(X_j|\mathcal{L}_j^Y)$$

and the corresponding error covariance matrix by

$$P_j = \mathrm{E}\big(X_j - \widehat{X}_j\big)\big(X_j - \widehat{X}_j\big)^*$$

THEOREM 2.5. *The estimate $\widehat{X}_j$ and the error covariance $P_j$ satisfy the equations*

$$\widehat{X}_j = a_0 + a_1\widehat{X}_{j-1} + a_2 Y_{j-1} + \big(a_1 P_{j-1}A_1^* + b\circ B\big)\cdot \\ \big(A_1 P_{j-1} A_1^* + B\circ B\big)^{\oplus}\big(Y_j - A_0 - A_1\widehat{X}_{j-1} - A_2 Y_{j-1}\big) \quad (2.15)$$

*and*

$$P_j = a_1 P_{j-1} a_1^* + b\circ b - \big(a_1 P_{j-1} A_1^* + b\circ B\big)\cdot \\ \big(A_1 P_{j-1} A_1^* + B\circ B\big)^{\oplus}\big(a_1 P_{j-1} A_1^* + b\circ B\big)^* \quad (2.16)$$

*where*

$$b\circ b = b_1 b_1^* + b_2 b_2^*, \quad b\circ B = b_1 B_1^* + b_2 B_2^*, \quad B\circ B = B_1 B_1^* + B_2 B_2^*$$

(2.15) *and* (2.16) *are solved subject to*

$$\widehat{X}_0 = \mathrm{E}X_0 + \mathrm{cov}(X_0, Y_0)\,\mathrm{cov}(Y_0)^{\oplus}(Y_0 - \mathrm{E}Y_0)$$
$$P_0 = \mathrm{cov}(X_0) - \mathrm{cov}(X_0, Y_0)\,\mathrm{cov}(Y_0)^{\oplus}\,\mathrm{cov}(X_0, Y_0)^*.$$

PROOF. Apply the formulae of Proposition 2.4 and the properties of orthogonal projections. For example

$$\widehat{X}_{j|j-1} = \widehat{\mathrm{E}}\Big(a_0 + a_1 X_{j-1} + a_2 Y_{j-1} + b_1\varepsilon_j + b_2\xi_j|\mathcal{L}_{j-1}^Y\Big) \stackrel{\dagger}{=} \\ a_0 + a_1\widehat{\mathrm{E}}\big(X_{j-1}|\mathcal{L}_{j-1}^Y\big) + a_2 Y_{j-1} = a_0 + a_1\widehat{X}_{j-1} + a_2 Y_{j-1},$$

where the equality † holds since $\varepsilon_j$ and $\xi_j$ are orthogonal to $\mathcal{L}_{j-1}^Y$. □

EXAMPLE 2.6. Consider an autoregressive scalar signal, generated by

$$X_j = aX_{j-1} + \varepsilon_j, \quad X_0 = 0$$

where $a$ is a constant and $\varepsilon$ is a white noise sequence. Suppose it is observed via a noisy linear sensor, so that the observations are given by

$$Y_j = X_{j-1} + \xi_j$$

where $\xi_j$ is another white noise, orthogonal to $\varepsilon$. Applying the equations from Theorem 2.5, one gets

$$\widehat{X}_j = a\widehat{X}_{j-1} + \frac{aP_{j-1}}{P_{j-1}+1}\big(Y_j - \widehat{X}_{j-1}\big), \quad \widehat{X}_0 = 0$$

where

$$P_j = a^2 P_{j-1} + 1 - \frac{a^2 P_{j-1}^2}{P_{j-1}+1}, \quad P_0 = 0. \quad (2.17)$$

■

Many more interesting examples are given as exercises in the last section of this chapter.

**3.1. Properties of the Kalman-Bucy filter.**

**1.** The equation for $P_j$ is called difference (discrete time) Riccati equation (analogously to differential Riccati equation arising in continuous time). Note that it does not depend on the observations and so can be solved off-line (before the filter is applied to the data). Even if all the coefficients of the system (2.13) and (2.14) are constant matrices, the optimal linear filter has in general time-varying coefficients.

**2.** Existence, uniqueness and strict positiveness of the limit $P_\infty := \lim_{j\to\infty} P_j$ is a non-trivial question, the answer to which is known under certain conditions on the coefficients. If the limit exists and is unique, then one may use the stationary version of the filter, where all the coefficients are calculated with $P_{j-1}$ replaced by $P_\infty$. In this case, the error matrix of this "suboptimal" filter converges to $P_\infty$ as well, i.e. such stationary filter is asymptotically optimal as $j \to \infty$. Note that the infinite sequence $(X, Y)$ generated by (2.13) and (2.14) may not have an $\mathbb{L}^2$ limit (e.g. if $|a| \ge 1$ in Example 2.6), so the infinite horizon problem actually is beyond the scope of Theorem 2.1. When $(X, Y)$ is in $\mathbb{L}^2$, then the filter may be used e.g. to realize the orthogonal projection[9] $\widehat{\mathrm{E}}(X_0|\mathcal{L}^Y_{(-\infty,0]})$. This would coincide with the estimates, obtained via Kolmogorov-Wiener theory for stationary processes (see [**28**] for further exploration).

**3.** The propagation of $\widehat{X}_j$ and $P_j$ is sometimes regarded in two-stages: *prediction*

$$\widehat{X}_{j|j-1} = a_0 + a_1\widehat{X}_{j-1} + a_2 Y_{j-1}, \quad \widehat{Y}_{j|j-1} = A_0 + A_1\widehat{X}_{j-1} + A_2 Y_{j-1}$$

and *update*

$$\widehat{X}_j = \widehat{X}_{j|j-1} + K_j(Y_j - \widehat{Y}_{j|j-1})$$

where $K_j$ is the *Kalman gain* matrix from (2.15). Similar interpretation is possible for $P_j$.

**4.** The sequence

$$\bar{\varepsilon}_j = Y_j - A_0 - A_1\widehat{X}_{j-1} - A_2 Y_{j-1} \tag{2.18}$$

turns to be an orthogonal sequence and is called the *innovations*: it is the residual "information" borne by $Y_j$ after its prediction on the basis of the past information is subtracted.

## Exercises

(**1**) Prove that $\mathbb{L}^2$ is complete, i.e. any Cauchy sequence converges to a random variable in $\mathbb{L}^2$. **Hint:** show first that from any Cauchy sequence in $\mathbb{L}^2$ a P-a.s. convergent subsequence can be extracted (Exercise (**3**a) on page 22)

(**2**) Complete the proof of Proposition 2.2 (verify (2.6))

(**3**) Complete the proof of Proposition 2.4.

(**4**) Show that the innovation sequence $\bar{\varepsilon}_j$ from (2.18) is orthogonal. Find its covariance sequence $\mathrm{E}\bar{\varepsilon}_j\bar{\varepsilon}_j^*$.

(**5**) Show that the limit $\lim_{j\to\infty} P_j$ in (2.17) exists[10] and is positive. Find the explicit expression for $P_\infty$. Does it exist when the equation (2.17) is started from any nonnegative $P_0$ ?

[9] here $\mathcal{L}^Y_{(-\infty,0]} = \overline{\mathrm{span}}\{..., Y_1, Y_0\}$

[10] Note that the filtering error $P_j$ is finite even if the signal is "unstable" ($|a| \ge 1$), i.e. all its trajectories diverge to $\infty$ as $j \to \infty$.

(**6**) Derive the Kalman-Bucy filter equations for the model, similar to Example 2.6, but with non-delayed observations

$$\begin{aligned} X_j &= aX_{j-1} + \varepsilon_j \\ Y_j &= X_j + \xi_j \end{aligned}$$

(**7**) Derive the equations (4) and (5) on the page 8.

(**8**) Consider the continuous-time AM[11] radio signal $X_t = A(s_t+1)\cos(ft+\varphi)$, $t \in \mathbb{R}_+$ with the carrier frequency $f$, amplitude $A$ and phase $\varphi$. The time function $s_t$ is the information message to be transmitted to the receiver, which recovers it by means of synchronous detection algorithm: it generates a cosine wave of frequency $f'$, phase $\varphi'$ and amplitude $A'$, and forms the base-band signal as follows

$$\widehat{s}_t = \big[A'\cos(f't+\varphi')X_t\big]_{\mathrm{LPF}}\,, \tag{2.19}$$

where $[\cdot]_{\mathrm{LPF}}$ is the (ideal) low pass filter operator, defined by

$$[q_t + r_t\cos(c_1t+c_2)]_{\mathrm{LPF}} = q_t, \quad \forall c_1, c_2 \in \mathbb{R}, \quad c_1 \neq 0$$

for any time functions $q_t$ and $r_t$.

(a) Show that to get $\widehat{s}_t = s_t$ for all $t \geq 0$, the receiver has to know $f$, $A$ and $\varphi$ (and choose $f'$, $\varphi'$ and $A'$ appropriately).

(b) Suppose the receiver knows $f$ (set $f = 1$), but not $A$ and $\varphi$. The following strategy is agreed between the transmitter and the receiver: $s_t \equiv 0$ for all $0 \leq t \leq T$ (the training period), i.e. the transmitter chooses some $A$ and $\varphi$ and sends $X_t = A\cos(t+\varphi)$ to the channel till time $T$. The digital receiver is used for processing the transmission, i.e. the received wave is sampled at times $t_j = \Delta j$, $j \in \mathbb{Z}_+$ with some fixed $\Delta > 0$, so that the following observations are available for processing:

$$Y_{j+1} = A\cos(\Delta j+\varphi) + \sigma\xi_{j+1}, \quad j = 0, 1, ... \tag{2.20}$$

where $\xi$ is a white noise sequence of intensity $\sigma > 0$. Define

$$\zeta_t = \begin{pmatrix} X_t \\ \dot{X}_t \end{pmatrix}$$

and let $Z_j := \zeta_{\Delta j}$, $j \in \mathbb{Z}_+$. Find the recursive equations for $Z_j$, i.e. the matrix $\theta(\Delta)$ (depending on $\Delta$) such that

$$Z_{j+1} = \theta(\Delta)Z_j. \tag{2.21}$$

(c) Using (2.21) and (2.20) and assuming that $A$ and $\varphi$ are random variables with uniform distributions on $[a_1, a_2]$, $0 < a_1 < a_2$ and $[0, 2\pi]$ respectively, derive the Kalman-Bucy filter equations for the estimate $\widehat{Z}_j = \widehat{\mathrm{E}}(Z_j|\mathcal{L}^Y_j)$ and the corresponding error covariance $P_j$.

(d) Find the relation between the estimates $\widehat{Z}_j$, $j = 0, 1, ...$ and the signal estimate[12]

$$\widehat{X}^\Delta_t := \widehat{\mathrm{E}}\big(X_t|\mathcal{L}^Y_{\lfloor t/\Delta \rfloor]}\big)$$

for all $t \in \mathbb{R}_+$

[11]AM - amplitude modulation

[12]recall that $\lfloor x \rfloor$ is the integer part of $x$

(e) Solve the Riccati difference equation from (c) explicitly[13]

(f) Is exact asymptotical synchronization possible, i.e.

$$\lim_{T\to\infty} \mathrm{E}\big(X_T - \widehat{X}_T^{\Delta}\big)^2 = 0 \tag{2.22}$$

for any $\Delta > 0$ ? For those $\Delta$ (2.22) holds, find the decay rate of the synchronization error, i.e. find the sequence $r_j > 0$ and positive number $c$, such that

$$\lim_{j\to\infty} \mathrm{E}\big(X_{\Delta j} - \widehat{X}_{\Delta j}^{\Delta}\big)^2 / r_j = c.$$

(g) Relying on the asymptotic result from (e) and assuming $\Delta = 1$, what should be $T$ to attain synchronization error of 0.001 ?

(h) Simulate numerically the results of this problem (using e.g. MATLAB)

(**9**) (taken from R.Kalman [**13**]) A number of particles leaves the origin at time $j = 0$ with random velocities; after $j = 0$, 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 $x_1(j)$ be the position and $x_2(j)$ the velocity of the particle; $x_3(j)$ is the noise. The problem is then represented by the model:

$$\begin{aligned} x_1(j+1) &= x_1(j) + x_2(j) \\ x_2(j+1) &= x_2(j) \\ x_3(j+1) &= \varphi x_3(j) + u(j) \\ y(j) &= x_1(j) + x_3(j) \end{aligned} \tag{2.23}$$

and the additional conditions

* $Ex_1^2(0) = Ex_2(0) = 0$, $Ex_2^2(0) = a^2 > 0$
* $Eu(j) = 0$, $Eu^2(j) = b^2$

(a) Derive Kalman-Bucy filter equations for the signal

$$X_j = \begin{pmatrix} x_1(j) \\ x_2(j) \\ x_3(j) \end{pmatrix}$$

(b) Derive Kalman-Bucy filter equations for the signal

$$X_j = \begin{pmatrix} x_2(j) \\ x_3(j) \end{pmatrix}$$

using the obvious relation $x_1(j) = jx_2(j) = jx_2(0)$.

(c) Solve the Riccati equation from (b) explicitly[14]

---

[13]**Hint:** you may need the very useful Matrix Inversion Lemma (verify it): for any matrices $A$,$B$,$C$ and $D$ (such that the required inverses exist), the following implication holds

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

[14]**Hint:** use the fact that the error covariance matrix is two dimensional and symmetric, i.e. there are only three parameters to find. Let the tedious calculations not scare you - the reward is coming!

(d) Show that for $\varphi \neq 1$ (both $|\varphi| < 1$ and $|\varphi| > 1$!), the mean square errors of the velocity and position estimates converge to 0 and $b^2$ respectively. Find the convergence rate for the velocity error.

(e) Show that for $\varphi = 1$, the mean square error for of the position diverges[15]!

(f) Define the new observation sequence

$$\delta y(j+1) = y(j+1) - \varphi y(j), \quad j \geq 0$$

and $\delta y(0) = y(0)$. Then (why?)

$$\overline{\text{span}}\{\delta y(j), 0 \leq j \leq n\} = \overline{\text{span}}\{y(j), 0 \leq j \leq n\}.$$

Derive the Kalman-Bucy filter for the signal $X_j := x_2(j)$ and observations $\delta y_j$. Verify your answer in (e).

(**10**) Consider the linear system of algebraic equations $Ax = b$, where $A$ is an $m \times n$ matrix and $b$ is an $n \times 1$ column vector. The *generalized* solution of these equations is a vector $x'$, which solves the following minimization problem (the usual Euclidian norm is used here)

$$x' := \begin{cases} \operatorname{argmin}_{x \in \Gamma} \|x\|^2 & \Gamma \neq \emptyset \\ \operatorname{argmin}_{x \in \mathbb{R}} \|Ax - b\|^2 & \Gamma = \emptyset \end{cases}$$

where $\Gamma = \{x \in \mathbb{R} : \|Ax - b\| = 0\}$. If $A$ is square and invertible then $x = A^{-1}b$. If the equations $Ax = b$ are satisfied by more than one vector, then the vector with the least norm is chosen. If $Ax = b$ has no solutions, then the vector which minimizes the norm $\|Ax - b\|$ is chosen. This defines $x'$ uniquely, moreover

$$x' := A^{\oplus} b = (A^* A)^{\oplus} A^* b$$

where $A^{\oplus}$ is the Moore-Penrose generalized inverse (recall that $(A^*A)^{\oplus}$ has been defined in (2.8)).

(a) Applying the Kalman-Bucy filter equations, show that $x'$ can be found by the following algorithm:

$$\widehat{x}_j = \widehat{x}_{j-1} + (b_j - \widehat{x}_{j-1}) \begin{cases} \frac{P_{j-1} a^{j*}}{a^j P_{j-1} a^{j*}}, & a^j P_{j-1} a^{j*} > 0 \\ 0 & a^j P_{j-1} a^{j*} = 0 \end{cases}$$

and

$$P_{j-1} = P_{j-1} + \begin{cases} \frac{P_{j-1} a^{j*} a^j Pj-1}{a^j P_{j-1} a^{j*}}, & a^j P_{j-1} a^{j*} > 0 \\ 0 & a^j P_{j-1} a^{j*} = 0 \end{cases},$$

where $a^j$ is the $j$-th row of the matrix $A$ and $b_j$ are the entries of $b$. To calculate $x$, these equations are to be started from $P_0 = I$ and $\widehat{x}_0 = 0$ and run for $j = 1, ..., m$. The solution is given by $x' = \widehat{x}_m$.

(b) Show that for each $j \leq m$,

$$a^j P_{j-1} a^{j*} = \min_{c_1, \dots, c_{j-1}} \left\| a^j - \sum_{\ell=1}^{j-1} c_j a^\ell \right\|^2$$

[15]Note that for $|\varphi| \geq 1$ the noise is "unstable" in the sense that its trajectories escape to $\pm\infty$. When $|\varphi| > 0$ this happens exponentially fast (in appropriate sense) and when $\varphi = 1$, the divergence is "linear". Surprisingly (for the author at least) the position estimate is "worse" in the latter case!

so that $a^j P_{j-1} a^{j*} = 0$ indicates that a row, linearly dependent on the previous ones, is encountered. So counting the number of times zero was used to propagate the above equations, the rank of $A$ is found as a byproduct.

**(11)** Let $X = (X_j)_{j\in\mathbb{Z}_+}$ be a Markov chain with values in a finite set of numbers $\mathbb{S} = \{a_1, ..., a_d\}$, the matrix $\Lambda$ of transition probabilities $\lambda_{ij}$ and initial distribution $\nu$[16], i.e.

$$\mathrm{P}(X_j = a_\ell | X_{j-1} = a_m) = \lambda_{\ell m}, \quad \mathrm{P}(X_0 = a_\ell) = \nu_\ell, \quad 1 \le \ell, m \le d.$$

(a) Let $p_n$ be the vector with entries $p_j(i) = \mathrm{P}(X_j = a_i)$, $j \ge 0$. Show that $p_j$ satisfies

$$p_j = \Lambda^* p_{j-1}, \quad \text{s.t. } p_0 = \nu \quad j \ge 0.$$

(b) Let $I_j$ be the vectors with entries $I_j(i) = \mathbf{1}(X_j = a_i)$, $j \ge 0$. Show that there exists a sequence of orthogonal random vectors $\varepsilon_j$, such that

$$I_j = \Lambda^* I_{j-1} + \varepsilon_j, \quad j \ge 0$$

Find its mean and covariance matrix.

(c) Suppose that the Markov chain is observed via noisy samples

$$Y_j = h(X_j) + \sigma\xi_j, j \ge 1$$

where $\xi$ is a white noise (with square integrable entries) and $\sigma > 0$ is its intensity. Let $h$ be the column vector with entries $h(a_i)$. Verify that

$$Y_j = h^* I_j + \sigma\xi_j.$$

(d) Derive the Kalman-Bucy filter for $\widehat{I}_j = \widehat{\mathrm{E}}(I_j|\mathcal{L}^Y_j)$.

(e) What would be the estimate of $\widehat{\mathrm{E}}\big(g(X_j)|\mathcal{L}^Y_j\big)$ for any $g : \mathbb{S} \mapsto \mathbb{R}$ in terms of $\widehat{I}_j$ ? In particular, $\widehat{X}_j = \widehat{\mathrm{E}}(X_j|\mathcal{L}^Y_j)$?

**(12)** Consider the ARMA(p,q) signal[17] $X = (X_j)_{j\ge 0}$, generated by the recursion

$$X_j = -\sum_{k=1}^p a_k X_{j-k} + \sum_{\ell=0}^q a_\ell \varepsilon_{j-\ell}, \quad j \ge p$$

subject to say $X_0 = X_1 = ... = X_p = 0$. Suppose that

$$Y_j = X_{j-1} + \xi_j, \quad j \ge 1.$$

Suggest a recursive estimation algorithm for $X_j$, given $\mathcal{L}^Y_j$, based on the Kalman-Bucy filter equations.

---

[16]Such a chain is a particular case of the Markov processes as in Example 1.3 on page 16 and can be constructed in the following way: let $X_0$ be a random variable with values in $\mathbb{S}$ and $\mathrm{P}(X_0 = a_\ell) = \nu_\ell$, $0 \le \ell \le d$ and

$$X_j = \sum_{i=1}^d \eta^i_j \mathbf{1}_{\{X_{j-1}=a_i\}}, \quad j \ge 0$$

where $\eta^i_j$ is a table of independent random variables with the distribution

$$\mathrm{P}(\eta^i_j = a_\ell) = \lambda_{i\ell}, \quad j \ge 0, \quad 1 \le i, \ell \le d$$

[17]ARMA(p,q) stands for "auto regressive of order $p$ and moving average of order $q$". This model is very poplar in voice recognition (LPC coefficients), compression, etc.

CHAPTER 3

# Nonlinear filtering in discrete time

Let $X$ and $Z$ be a pair of independent real random variables on $(\Omega, \mathcal{F}, \mathrm{P})$ and suppose that $\mathrm{E}X^2 < \infty$. Assume for simplicity that both have probability densities $f_X(u)$ and $f_Z(u)$, i.e.

$$\mathrm{P}(X \le u) = \int_{-\infty}^{u} f_X(x)dx, \quad \mathrm{P}(Z \le u) = \int_{-\infty}^{u} f_Z(x)dx.$$

Suppose it is required to estimate $X$, given the observed realization of the sum $Y = X + Z$ or, in other words, to find a function[1] $\bar{g} : \mathbb{R} \mapsto \mathbb{R}$, so that

$$\mathrm{E}\big(X - \bar{g}(Y)\big)^2 \le \mathrm{E}\big(X - g(Y)\big)^2 \tag{3.1}$$

for any other function $g : \mathbb{R} \mapsto \mathbb{R}$. Note that such a function should be square integrable as well, since (3.1) with $g = 0$ and $\bar{g}^2(Y) \le 2X^2 + 2\big(X - \bar{g}(Y)\big)^2$ imply

$$\mathrm{E}\bar{g}^2(Y)^2 \le 4\mathrm{E}X^2 < \infty.$$

Moreover, if $\bar{g}$ satisfies

$$\mathrm{E}\big(X - \bar{g}(Y)\big)g(Y) = 0 \tag{3.2}$$

for *any* $g : \mathbb{R} \mapsto \mathbb{R}$, such that $\mathrm{E}g^2(Y) < \infty$, then (3.1) would be satisfied too. Indeed, if $\mathrm{E}\big(X - g(Y)\big)^2 = \infty$, the claim is trivial and if $\mathrm{E}\big(X - g(Y)\big)^2 < \infty$, then $\mathrm{E}g^2(Y) \le 2\mathrm{E}X^2 + 2\mathrm{E}(g(Y) - X)^2 < \infty$ and

$$\mathrm{E}\big(X - g(Y)\big)^2 = \mathrm{E}\big(X - \bar{g}(Y) + \bar{g}(Y) - g(Y)\big)^2 = \\ \mathrm{E}\big(X - \bar{g}(Y)\big)^2 + \mathrm{E}\big(\bar{g}(Y) - g(Y)\big)^2 \ge \mathrm{E}\big(X - \bar{g}(Y)\big)^2$$

Moreover, the latter suggests that if another function satisfies (3.1), then it should be equal to $\bar{g}$ on any set $A$, such that $\mathrm{P}(Y \in A) > 0$. Does such a function exist ? Yes - we give an explicit construction using (3.2)

$$\mathrm{E}\big(X - \bar{g}(Y)\big)g(Y) = \int_{\mathbb{R}}\int_{\mathbb{R}} \big(x - \bar{g}(x+z)\big)g(x+z)f_X(x)f_Z(z)dxdz = \\ \int_{\mathbb{R}}\int_{\mathbb{R}} \big(x - \bar{g}(u)\big)g(u)f_X(x)f_Z(u-x)dxdu = \\ \int_{\mathbb{R}} g(u)\left(\int_{\mathbb{R}} \big(x - \bar{g}(u)\big)f_X(x)f_Z(u-x)dx\right)du$$

The latter would vanish if

$$\int_{\mathbb{R}} \big(x - \bar{g}(u)\big)f_X(x)f_Z(u-x)dx = 0$$

[1] $g$ should be a Borel function (measurable with respect to Borel $\sigma$-algebra on $\mathbb{R}$) so that all the expectations are well defined

is satisfied for all $u$, which leads to

$$\bar{g}(u) = \frac{\int_{\mathbb{R}} x f_X(x) f_Z(u-x)dx}{\int_{\mathbb{R}} f_X(x) f_Z(u-x)dx}.$$

So the best estimate of $X$ given $Y$ is the random variable

$$\mathrm{E}(X|Y)(\omega) = \frac{\int_{\mathbb{R}} x f_X(x) f_Z(Y(\omega)-x)dx}{\int_{\mathbb{R}} f_X(x) f_Z(Y(\omega)-x)dx}, \tag{3.3}$$

which is nothing but the familiar Bayes formula for the conditional expectation of $X$ given $Y$.

## 1. The conditional expectation: a closer look

**1.1. The definition and the basic properties.** Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, carrying a random variable $X \geq 0$ with values in $\mathbb{R}$ and let $\mathcal{G}$ be a sub-$\sigma$-algebra of $\mathcal{F}$.

DEFINITION 3.1. The conditional expectation[2] of $X \geq 0$ with respect to $\mathcal{G}$ is a real random variable, denoted by $\mathrm{E}(X|\mathcal{G})(\omega)$, which is $\mathcal{G}$-measurable, i.e.

$$\big\{\omega : \mathrm{E}(X|\mathcal{G})(\omega) \in A\big\} \in \mathcal{G}, \quad \forall A \in \mathcal{B}(\mathbb{R})$$

and satisfies

$$\mathrm{E}\Big(X - \mathrm{E}(X|\mathcal{G})(\omega)\Big)\mathbf{1}_A(\omega) = 0, \quad \forall A \in \mathcal{G}.$$

Why is this definition correct, i.e. is there indeed such a random variable and is it unique? The positive answer is provided by the Radon-Nikodym theorem from analysis

THEOREM 3.2. *Let $(\mathbb{X}, \mathscr{X})$ be a measurable space*[3]*, $\mu$ be a $\sigma$-finite* [4] *measure and $\nu$ is a signed measure*[5]*, absolutely continuous* [6] *with respect to $\mu$. Then there is exists an $\mathscr{X}$-measurable function $f = f(x)$, taking values in $\mathbb{R} \cup \{\pm\infty\}$, such that*

$$\nu(A) = \int_A f(x)\mu(dx), \quad A \in \mathscr{X}.$$

*$f$ is called the Radon-Nikodym derivative (or density) of $\nu$ with respect to $\mu$ and is denoted by $\frac{d\nu}{d\mu}$. It is unique up to $\mu$-null sets*[7]*.*

Now consider the measurable space $(\Omega, \mathcal{G})$ and define a nonnegative set function on[8] $\mathcal{G}$

$$\mathrm{Q}(A) = \int_A X \mathrm{P}(d\omega) = \mathrm{E}X\mathbf{1}_A, \quad A \in \mathcal{G}. \tag{3.4}$$

[2]Note that the conditional probability is a special case of the conditional expectation: $\mathrm{P}(B|\mathcal{G}) = \mathrm{E}(I_B|\mathcal{G})$

[3]i.e. a collection of points $\mathbb{X}$ with a $\sigma$-algebra of sets $\mathscr{X}$

[4]i.e. $\mu(\mathbb{X}) = \infty$ is allowed, only if there is a countable partition $D_j \in \mathscr{X}$, $\biguplus_j D_j = \mathbb{X}$, so that $\mu(D_j) < \infty$ for any $j$. For example, the Lebesgue measure on $\mathcal{B}(\mathbb{R})$ is not a finite measure (the "length" of the whole line is $\infty$). It is $\sigma$-finite, since $\mathbb{R}$ can be partitioned into e.g. intervals of unit Lebesgue measure.

[5]i.e. which can be represented as $\nu = \nu_1 - \nu_2$, with at least one of $\nu_i$ is finite

[6]A measure $\mu$ is absolutely continuous with respect to $\nu$ (denoted $\mu \ll \nu$), if for any $A \in \mathscr{X}$ $\nu(A) = 0 \implies \mu(A) = 0$. The measures $\mu$ and $\nu$ are said to be equivalent $\mu \sim \nu$, if $\mu \ll \nu$ and $\nu \ll \mu$.

[7]i.e. if there is another function $h$, such that $\nu(A) = \int_A h(x)\mu(dx)$ then $\mu(h \neq f) = 0$

[8]Note that the integral here is well defined for $A \in \mathcal{F}$ as well, but we restrict it to $A \in \mathcal{G}$ only

This set function is a nonnegative $\sigma$-finite measure: take for example the partition $D_j = \{j \le X < j+1\}$, $j = 0, 1, ...$, then $\mathrm{Q}(D_j) = \mathrm{E}X\mathbf{1}_{\{X\in[j,j+1)\}} < \infty$ even if $\mathrm{E}X = \infty$. To verify $\mathrm{Q} \ll \mathrm{P}$, let $A$ be such that $\mathrm{P}(A) = 0$ and let $X_j$ be a sequence of simple random variables, such that $X_j \nearrow X$ (for example as in (1.1) on page 17), i.e.

$$X_j = \sum_k x_k^j \mathbf{1}_{B_k^j}, \quad B_k^j \in \mathcal{F}, \quad x_k^j \in \mathbb{R}$$

Since

$$\mathrm{E}X_j\mathbf{1}_A = \sum_k x_k^j \mathrm{P}(B_k^j \cap A) = 0,$$

by monotone convergence (see Theorem A.1 in the Appendix ) $\mathrm{Q}(A) = \mathrm{E}X\mathbf{1}_A = \lim_j \mathrm{E}X_j\mathbf{1}_A = 0$. Now by Radon-Nikodym theorem there exists the unique up to P-null sets random variable $\xi$, measurable with respect to $\mathcal{G}$ (unlike $X$ itself!), such that

$$\mathrm{Q}(A) = \int_A \xi \mathrm{P}(A), \quad \forall A \in \mathcal{G}.$$

This $\xi$ is said to be a version of the conditional expectation $\mathrm{E}(X|\mathcal{G})$ to emphasize its uniqueness only up to P-null sets:

$$\mathrm{E}(X|\mathcal{G}) = \frac{d\mathrm{Q}}{d\mathrm{P}}(\omega).$$

For a general random variable $X$, taking both positive and negative values, define $\mathrm{E}(X|\mathcal{G}) = \mathrm{E}(X^+|\mathcal{G}) - \mathrm{E}(X^-|\mathcal{G})$, if no $\infty - \infty$ confusion occurs with positive probability. Note that $\infty - \infty$ is allowed on the P-null sets, in which case an arbitrary value can be assigned. For this reason, the conditional expectation $\mathrm{E}(X|\mathcal{G})$ may be well defined even, when $\mathrm{E}X$ is not. For example, let $\mathcal{F}^X$ be the $\sigma$-algebra generated by the pre-images $\{X \in A\}$, $A \in \mathcal{B}(\mathbb{R})$. Suppose that $\mathrm{E}X^+ = \infty$ and $\mathrm{E}X^- = \infty$, so that $\mathrm{E}X$ is not defined. Since $\{X^+ = \infty \cap X^- = \infty\}$ is a null set, the conditional expectation is well defined and equals

$$\mathrm{E}(X|\mathcal{F}^X) = \mathrm{E}(X^+|\mathcal{F}^X) - \mathrm{E}(X^-|\mathcal{F}^X) = X^+ - X^- = X.$$

EXAMPLE 3.3. Let $\mathcal{G}$ be the (finite) $\sigma$-algebra generated by the finite partition $D_j \in \mathcal{F}$, $j = 1, ..., n$, $\uplus D_j = \Omega$, $\mathrm{P}(D_j) > 0$. Any $\mathcal{G}$-measurable random variable (with real values) $\xi$ is necessarily constant on each set $D_j$: suppose it takes two distinct values on e.g. $D_1$, say $x' < x''$, then $\{\omega : X(\omega) \le x'\} \cap D_1$ and $\{\omega : X(\omega) \ge x''\} \cap D_1$ are disjoint subsets of $D_1$ and hence not in any other $D_i$, $i \ne j$. Thus both events clearly cannot be in $\mathcal{G}$. So for any random variable $X$,

$$\mathrm{E}(X|\mathcal{G}) = \sum_{j=1}^n a_j \mathbf{1}_{D_j}(\omega).$$

The constants $a_j$ are found from

$$\mathrm{E}\Big(X - \sum_{j=1}^n a_k \mathbf{1}_{D_j}\Big)\mathbf{1}_{D_i} = 0, \quad i = 1, ..., n,$$

which leads to

$$\mathrm{E}(X|\mathcal{G}) = \sum_{j=1}^n \frac{\mathrm{E}X\mathbf{1}_{D_j}}{\mathrm{P}(D_j)} \mathbf{1}_{D_j}(\omega).$$

■

The conditioning with respect to $\sigma$-algebras generated by the pre-images of random variables (or more complex random objects), i.e. by the sets of the form

$$\mathcal{F}^Y = \sigma\{\omega : Y \in A\}, \quad A \in \mathcal{B}(\mathbb{R})$$

are of special interest. Given a pair of random variables $(X, Y)$, $\mathrm{E}(X|Y)$ is sometimes[9] written shortly for $\mathrm{E}(X|\mathcal{F}^Y)$. It can be shown, that for any $\mathcal{F}^Y$-measurable random variable $Z(\omega)$, there exists a Borel function $\varphi$, such that $Z = \varphi(Y(\omega))$. In particular, there always can be found a Borel function $g$, so that $\mathrm{E}(X|Y) = g(Y)$. This function is sometimes denoted by $\mathrm{E}(X|Y = y)$.

The main properties of the conditional expectations are[10]

(A) if $C$ is a constant and $X = C$, then $\mathrm{E}(X|\mathcal{G}) = C$
(B) if $X \le Y$, then $\mathrm{E}(X|\mathcal{G}) \le \mathrm{E}(Y|\mathcal{G})$
(C) $|\mathrm{E}(X|\mathcal{G})| \le \mathrm{E}(|X||\mathcal{G})$
(D) if $a, b \in \mathbb{R}$, and $a\mathrm{E}X + b\mathrm{E}Y$ is well defined, then
$$\mathrm{E}(aX + bY|\mathcal{G}) = a\mathrm{E}(X|\mathcal{G}) + b\mathrm{E}(Y|\mathcal{G})$$
(E) if $X$ is $\mathcal{G}$-measurable, then $\mathrm{E}(X|\mathcal{G}) = X$
(F) if $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then $\mathrm{E}\big(\mathrm{E}(X|\mathcal{G}_2)\big|\mathcal{G}_1\big) = \mathrm{E}(X|\mathcal{G}_1)$
(G) if $X$ and $Y$ are independent and $f(x, y)$ is such that $\mathrm{E}|f(X, Y)| < \infty$, then
$$\mathrm{E}\big(f(X, Y)|Y\big) = \int_\Omega f\big(X(\omega'), Y(\omega)\big)\mathrm{P}(d\omega')$$
In particular, if $X$ is independent of $\mathcal{G}$ and $\mathrm{E}X$ is well defined, then $\mathrm{E}(X|\mathcal{G}) = \mathrm{E}X$.
(H) if $Y$ is $\mathcal{G}$-measurable and $\mathrm{E}|Y| < \infty$ and $\mathrm{E}|YX| < \infty$, then
$$\mathrm{E}\big(XY|\mathcal{G}\big) = Y\mathrm{E}(X|\mathcal{G})$$
(I) let $(X, Y)$ be a pair of random variables and $\mathrm{E}|X|^2 < \infty$, then
$$\mathrm{E}\big(X - \mathrm{E}(X|Y)\big)^2 = \inf_\varphi \mathrm{E}\big(X - \varphi(Y)\big)^2 \tag{3.5}$$
where all the Borel functions $\varphi$ are taken.

Let $A_j$ be a sequence of disjoint events, then

$$P\big(\uplus A_j|\mathcal{G}\big) = \sum_j P(A_j|\mathcal{G}). \tag{3.6}$$

So one is tempted to think that for any fixed $\omega$, $\mathrm{P}(A|\mathcal{G})(\omega)$ is a measure on $\mathcal{F}$. This is wrong in general, since (3.6) holds only up to P-null sets. Denote by $N_i$ the set of points at which (3.6) fails for the specific sequence $A_j^{(i)}$, $j = 1, 2, ....$ And let $N$ be the set of all null sets of the latter form. Since in general there can be uncountably many sequences of events, $N$ may have positive probability ! So in general, the function

$$F_X(x; \omega) = \mathrm{P}\big(X \le x|\mathcal{G}\big)(\omega)$$

may not be a proper distribution function for $\omega$ from a set of positive probability.

It turns out that for any random variable $X$ with values in a complete separable metric space $\mathbb{X}$, there exists so called *regular* conditional measure of $X$, given $\mathcal{G}$, i.e. a function $\mathrm{P}_X(B; \omega)$, which is a probability measure on $\mathcal{B}(\mathbb{X})$ for each fixed

[9] throughout these notations are freely switched

[10] as usual any relations, involving comparison of random variables are understood P-a.s.

$\omega \in \Omega$ and is a version of $\mathrm{P}(X \in B|\mathcal{G})(\omega)$. Obviously regular conditional expectation plays the central role in statistical problems, where typically it is required to find an explicit formula (function), which can be applied to the realizations of the observed random variables. For example regular conditional expectation was explicitly constructed in (3.3).

**1.2. The Bayes formula: an abstract formulation.** The Bayes formula (3.3) involves explicit distribution functions of the random variables involved in the estimation problem. On the other hand, the abstract definition of the conditional expectation of the previous section, allows to consider the setups, where the conditioning $\sigma$-algebra is not necessarily generated by random variables, whose distribution have explicit formulae: think for example of $\mathrm{E}(X|\mathcal{F}^Y_t)$, when $\mathcal{F}^Y_t = \sigma\{Y_s, 0 \le s \le t\}$ with $Y_t$, being a continuous time process.

THEOREM 3.4. *(the Bayes formula) Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space, carrying a real random variable* $X$ *and let* $\mathcal{G}$ *be a sub-*$\sigma$*-algebra of* $\mathcal{F}$. *Assume that there exists a regular conditional probability measure* [11] $\mathrm{P}(d\omega|X = x)$ *on* $\mathcal{G}$ *and it has Radon-Nikodym density* $\rho(\omega; x)$ *with respect to a* $\sigma$*-finite measure* $\lambda$ *(on* $\mathcal{G}$*):*

$$\mathrm{P}(B|X = x) = \int_B \rho(\omega; x)\lambda(d\omega).$$

*Then for any* $\varphi : \mathbb{R} \mapsto \mathbb{R}$, *such that* $\mathrm{E}|\varphi(X)| < \infty$,

$$\mathrm{E}\big(\varphi(X)|\mathcal{G}\big) = \frac{\int_{\mathbb{R}} \varphi(u)\rho(\omega; u)\mathrm{P}_X(du)}{\int_{\mathbb{R}} \rho(\omega; u)\mathrm{P}_X(du)}, \tag{3.7}$$

*where* $\mathrm{P}_X$ *is the probability measure induced by* $X$ *(on* $\mathcal{B}(\mathbb{R})$*).*

PROOF. Recall that

$$\mathrm{E}\big(\varphi(X)|\mathcal{G}\big)(\omega) = \frac{d\mathrm{Q}}{d\mathrm{P}}(\omega) \tag{3.8}$$

where Q is a signed measure, defined by

$$\mathrm{Q}(B) = \int_B \varphi(X(\omega))\mathrm{P}(d\omega), \quad B \in \mathcal{G}.$$

Let $\mathcal{F}^X = \sigma\{X\}$. Then for any $B \in \mathcal{G}$

$$\mathrm{P}(B) = \mathrm{E}\mathrm{E}(\mathbf{1}_B|\mathcal{F}^X) = \int_\Omega \mathrm{P}(B|\mathcal{F}^X)(\omega)\mathrm{P}(d\omega) \stackrel{\dagger}{=} \int_{\mathbb{R}} \mathrm{P}(B|X = u)\mathrm{P}_X(du) = \\ \int_{\mathbb{R}}\int_B \rho(\omega; u)\lambda(d\omega)\mathrm{P}_X(du) \stackrel{\ddagger}{=} \int_B \left(\int_{\mathbb{R}} \rho(\omega; u)\mathrm{P}_X(du)\right)\lambda(d\omega) \tag{3.9}$$

where the equality † is changing variables under the Lebesgue integral and ‡ follows from the Fubini theorem (see Theorem A.5 Appendix for quick reference). Also for any $B \in \mathcal{G}$

$$\mathrm{Q}(B) := \mathrm{E}\varphi(X)\mathbf{1}_B = \mathrm{E}\varphi(X)\mathrm{E}\big(\mathbf{1}_B|\mathcal{F}^X\big)(\omega) = \int_{\mathbb{R}} \varphi(u)\mathrm{P}(B|X = u)d\mathrm{P}_X(du) = \\ \int_{\mathbb{R}} \varphi(u)\int_B \rho(\omega; u)\lambda(d\omega)\mathrm{P}_X(du) = \int_B \left(\int_{\mathbb{R}} \varphi(u)\rho(\omega; u)\mathrm{P}_X(du)\right)\lambda(d\omega). \tag{3.10}$$

[11]i.e. a measurable function $P(B; x)$, which is a probability measure on $\mathcal{F}$ for any fixed $x \in \mathbb{R}$ and $P(B; X(\omega))$ coincides with $P(B|\mathcal{F}^X)(\omega)$ up to P-null sets.

Note that $\mathrm{Q} \ll \mathrm{P}$ and by (3.9) $\mathrm{P} \ll \lambda$ (on $\mathcal{G}$!) and thus also $\mathrm{Q} \ll \lambda$. So for any $B \in \mathcal{G}$

$$\mathrm{Q}(B) = \int_B \frac{d\mathrm{Q}}{d\mathrm{P}}(\omega)\mathrm{P}(d\omega) = \int_B \frac{d\mathrm{Q}}{d\mathrm{P}}(\omega)\frac{d\mathrm{P}}{d\lambda}(\omega)\lambda(d\omega)$$

while on the other hand

$$\mathrm{Q}(B) = \int_B \frac{d\mathrm{Q}}{d\lambda}(\omega)d\lambda, \quad \forall B \in \mathcal{G}.$$

By arbitrariness of $B$, it follows that

$$\frac{d\mathrm{Q}}{d\lambda}(\omega) = \frac{d\mathrm{Q}}{d\mathrm{P}}(\omega)\frac{d\mathrm{P}}{d\lambda}(\omega), \quad \lambda - a.s.$$

Now since

$$\mathrm{P}\left\{\omega : \frac{d\mathrm{P}}{d\lambda}(\omega) = 0\right\} = \int_\Omega \mathbf{1}\left(\frac{d\mathrm{P}}{d\lambda}(\omega) = 0\right)\mathrm{P}(d\omega) =$$
$$\int_\Omega \mathbf{1}\left(\frac{d\mathrm{P}}{d\lambda}(\omega) = 0\right)\frac{d\mathrm{P}}{d\lambda}(\omega)\lambda(d\omega) = 0$$

it follows

$$\frac{d\mathrm{Q}}{d\mathrm{P}}(\omega) = \frac{d\mathrm{Q}/d\lambda(\omega)}{d\mathrm{P}/d\lambda(\omega)}, \quad \mathrm{P} - a.s.$$

The latter and (3.8), (3.9), (3.10) imply (3.7). $\square$

Corollary 3.5. *Suppose that $\mathcal{G}$ is generated by a random variable $Y$ and there is a $\sigma$-finite measure $\nu$ on $\mathcal{B}(\mathbb{R})$ and a measurable function (density) $r(u;x) \geq 0$ so that*

$$\mathrm{P}(Y \in A | X = x) = \int_A r(u;x)\nu(du).$$

*Then for $|\varphi(X)| < \infty$,*

$$\mathrm{E}\big(\varphi(X)|\mathcal{G}\big) = \frac{\int_{\mathbb{R}} \varphi(u) r\big(Y(\omega), u\big)\mathrm{P}^X(du)}{\int_{\mathbb{R}} r\big(Y(\omega), u\big)\mathrm{P}^X(du)}. \tag{3.11}$$

Proof. By the Fubini theorem (see Appendix)

$$\mathrm{P}(Y \in A) = \mathrm{E}\mathrm{P}(Y \in A|X) = \mathrm{E}\int_A r(u; X(\omega))\nu(du) = \int_A \mathrm{E}r(u; X(\omega))\nu(du).$$

Denote $\bar{r}(u) := \mathrm{E}r(u; X(\omega))$ and define

$$\rho(\omega; x) = \begin{cases} \frac{r\left(Y(\omega), x\right)}{\bar{r}\left(Y(\omega)\right)}, & \bar{r}\big(Y(\omega)\big) > 0 \\ 0, & \bar{r}\big(Y(\omega)\big) = 0 \end{cases}$$

Any $\mathcal{G}$-measurable set is by definition a preimage of some $A$ under $Y(\omega)$, i.e. for any $B \in \mathcal{G}$, there is $A \in \mathcal{B}(\mathbb{R})$ such that $B = \{\omega : Y(\omega) \in A\}$. Then

$$\int_B \rho(\omega; x)\mathrm{P}(d\omega) = \int_A \frac{r\big(u, x\big)}{\bar{r}\big(u\big)}\bar{r}\big(u\big)\nu(du) =$$
$$\int_A r(u;x)\nu(du) = \mathrm{P}(Y \in A|X = x) = \mathrm{P}(B|X = x).$$

Now (3.11) follows from (3.7) with the specific $\rho(\omega; x)$ and $\lambda(d\omega) := \mathrm{P}(d\omega)$, where the denominators cancel. $\square$

REMARK 3.6. Let $(\check{\Omega}, \check{\mathcal{F}}, \check{\mathrm{P}})$ be a copy of the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, then (3.11) reads

$$\mathrm{E}\big(\varphi(X)|\mathcal{G}\big) = \frac{\check{\mathrm{E}}\varphi(X(\check{\omega}))r\big(Y(\omega), X(\check{\omega})\big)}{\check{\mathrm{E}}r\big(Y(\omega), X(\check{\omega})\big)}, \tag{3.12}$$

where $\check{\mathrm{E}}$ denotes expectation on $(\check{\Omega}, \check{\mathcal{F}}, \check{\mathrm{P}})$ (and $X(\check{\omega})$ is a copy of $X$, defined on this auxiliary probability space).

REMARK 3.7. The formula (3.11) (and its notation (3.12)) holds when $X$ and $Y$ are random vectors.

REMARK 3.8. Often the following notation is used

$$\mathrm{P}\big(X \in du|Y = y\big) = \frac{r\big(y, u\big)\mathrm{P}^X(du)}{\int_{\mathbb{R}} r\big(y, u\big)\mathrm{P}^X(du)}$$

for the regular conditional distribution of $X$ given $\mathcal{F}^Y$. Note that it is absolutely continuous with respect to the measure induced by $X$.

## 2. The nonlinear filter via the Bayes formula

Let $(X_j, Y_j)_{j\geq 0}$ be a pair of random sequences with the following structure:

* $X_j$ is a Markov process with the transition kernel[12] $\Lambda(x, du)$ and initial distribution $p(du)$, that is

$$\mathrm{P}(X_j \in B|\mathcal{F}^X_{j-1} \vee \mathcal{F}^Y_{j-1}) = \int_B \Lambda(X_{j-1}, du), \quad \mathrm{P} - a.s.$$

  where[13] $\mathcal{F}^X_{j-1} = \sigma\{X_0, ..., X_{j-1}\}$

$$\mathrm{P}(X_0 \in B) = \int_B p(du), \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

* $Y_j$ is a random sequence, such that for all[14] $j \geq 0$

$$\mathrm{P}(Y_j \in B|\mathcal{F}^X_j \vee \mathcal{F}^Y_{j-1}) = \int_B \Gamma(X_j, du), \quad \mathrm{P} - a.s \tag{3.13}$$

  with a Markov kernel $\Gamma(x, du)$, which has density $\gamma(x, u)$ with respect to some $\sigma$-finite measure $\nu(du)$ on $\mathcal{B}(\mathbb{R})$.
* $f : \mathbb{R} \mapsto \mathbb{R}$ be a measurable function, such that $\mathrm{E}|f(X_j)| < \infty$ for each $j \geq 0$.

THEOREM 3.9. *Let $\pi_j(dx)$ be the solution of the recursive equation*

$$\pi_j(dx) = \frac{\int_{\mathbb{R}} \gamma\big(u, Y_j(\omega)\big)\Lambda(u, dx)\pi_{j-1}(du)}{\int_{\mathbb{R}}\int_{\mathbb{R}} \gamma\big(u, Y_j(\omega)\big)\Lambda(u, dx)\pi_{j-1}(du)}, \quad j \geq 0 \tag{3.14}$$

*subject to*

$$\pi_0(dx) = \frac{\int_{\mathbb{R}} \gamma\big(u, Y_0(\omega)\big)p(du)}{\int_{\mathbb{R}}\int_{\mathbb{R}} \gamma\big(u, Y_0(\omega)\big)p(du)}. \tag{3.15}$$

---

[12]a function $\Lambda : \mathbb{R} \times \mathcal{B}(\mathbb{R}) \mapsto [0, 1]$ is called a Markov (transition) kernel, if $\Lambda(x, B)$ is a Borel measurable function for each $B \in \mathcal{B}(\mathbb{R})$ and is a probability measure on $\mathcal{B}(\mathbb{R})$ for each fixed $x \in \mathbb{R}$.

[13]a family $\mathcal{F}_j$ of increasing $\sigma$-algebras is called filtration

[14]by convention $\mathcal{F}^Y_{-1} = \{\emptyset, \Omega\}$

*Then*

$$\mathrm{E}\big(f(X_j)|\mathcal{F}_j^Y\big) = \int_{\mathbb{R}} f(x)\pi_j(dx), \quad \mathrm{P}-a.s. \tag{3.16}$$

PROOF. Note that by the above assumptions the pair process $(X_j, Y_j)$ is Markov with the transition kernel $\Lambda(x, du)\gamma(u, v)\nu(dv)$:

$$\mathrm{P}\Big(X_j \in A, Y_j \in B\big|\mathcal{F}_{j-1}^X \vee \mathcal{F}_{j-1}^Y\Big) = \int_A\int_B \gamma(u,v)\nu(dv)\Lambda(X_{j-1}, du),$$

and hence the regular conditional measure for the vector $\{Y_0, ..., Y_j\}$, given $\mathcal{F}_j^X = \sigma\{X_0, ..., X_j\}$ is

$$\mathrm{P}\big(Y_0 \in A_0, ..., Y_j \in A_j|\mathcal{F}_j^X\big) = \\ \int_{A_0}...\int_{A_j}\gamma(X_0, u_0)\cdots\gamma(X_j, u_j)\nu(du_0)\cdots\nu(du_j). \tag{3.17}$$

Then by Remark 3.7

$$\mathrm{E}\big(\varphi(X_j)|\mathcal{F}_j^Y\big) = \frac{\check{\mathrm{E}}\varphi(X_j(\check{\omega}))\prod_{i=0}^j\gamma\big(X_i(\check{\omega}), Y_i\big)}{\check{\mathrm{E}}\prod_{i=0}^j\gamma\big(X_i(\check{\omega}), Y_i\big)} \tag{3.18}$$

Introduce the notation

$$L_j(X(\check{\omega}), Y) = \prod_{i=0}^j\gamma\big(X_i(\check{\omega}), Y_i\big) \tag{3.19}$$

and note that

$$\check{\mathrm{E}}\big(\varphi(X_j(\check{\omega}))L_j(X(\check{\omega}), Y)|\mathcal{F}_{j-1}^X\big) = \\ L_{j-1}(X(\check{\omega}), Y)\check{\mathrm{E}}\big(\varphi(X_j(\check{\omega}))\gamma(X_j(\check{\omega}), Y_j)\big|\mathcal{F}_{j-1}^X\big) = \\ L_{j-1}(X(\check{\omega}), Y)\int_{\mathbb{R}}\varphi(u)\gamma(u, Y_j)\Lambda(X_{j-1}(\check{\omega}), du)$$

Then

$$\mathrm{E}\big(\varphi(X_j)|\mathcal{F}_j^Y\big) = \frac{\check{\mathrm{E}}\varphi(X_j(\check{\omega}))L_j(X(\check{\omega}), Y)}{\check{\mathrm{E}}L_j(X(\check{\omega}), Y)} = \\ \frac{\check{\mathrm{E}}L_{j-1}(X(\check{\omega}), Y)\int_{\mathbb{R}}\varphi(u)\gamma(u, Y_j)\Lambda(X_{j-1}(\check{\omega}), du)}{\check{\mathrm{E}}L_{j-1}(X(\check{\omega}), Y)\int_{\mathbb{R}}\gamma(u, Y_j)\Lambda(X_{j-1}(\check{\omega}), du)} = \\ \frac{\check{\mathrm{E}}L_{j-1}(X(\check{\omega}), Y)\int_{\mathbb{R}}\varphi(u)\gamma(u, Y_j)\Lambda(X_{j-1}(\check{\omega}), du)/\check{\mathrm{E}}L_{j-1}(X(\check{\omega}), Y)}{\check{\mathrm{E}}L_{j-1}(X(\check{\omega}), Y)\int_{\mathbb{R}}\gamma(u, Y_j)\Lambda(X_{j-1}(\check{\omega}), du)/\check{\mathrm{E}}L_{j-1}(X(\check{\omega}), Y)} = \\ \frac{\mathrm{E}\Big(\int_{\mathbb{R}}\varphi(u)\gamma(u, Y_j)\Lambda(X_{j-1}, du)\big|\mathcal{F}_{j-1}^Y\Big)}{\mathrm{E}\Big(\int_{\mathbb{R}}\gamma(u, Y_j)\Lambda(X_{j-1}, du)\big|\mathcal{F}_{j-1}^Y\Big)}$$

Now let $\pi_j(dx)$ be the regular conditional distribution of $X_j$, given $\mathcal{F}_j^Y$. Then the latter reads (again the Fubini theorem is used)

$$\int_{\mathbb{R}}\varphi(x)\pi_j(dx) = \mathrm{E}\big(\varphi(X_j)|\mathcal{F}_j^Y\big) = \int_{\mathbb{R}}\varphi(u)\frac{\int_{\mathbb{R}}\gamma(u, Y_j)\Lambda(x, du)\pi_{j-1}(dx)}{\int_{\mathbb{R}}\int_{\mathbb{R}}\gamma(u, Y_j)\Lambda(x, du)\pi_{j-1}(dx)}$$

and by arbitrariness of $\varphi$ (3.14) follows. The equation (3.15) is obtained similarly. $\square$

REMARK 3.10. The proof may seem unnecessarily complicated at the first glance: in fact, a simpler and probably more intuitive derivation is possible (see Exercise **10**). This (and an additional derivation in the next section) is given for two reasons: (1) to exercise the properties and notations, related to conditional expectations and (2) to demonstrate the technique, which will be very useful when working in continuous time case.

## 3. The nonlinear filter by the reference measure approach

Before proceeding to discuss the properties of (3.14), we give another proof of it, using so called *reference* measure approach. This powerful and elegant method requires stronger assumptions on $(X, Y)$, but gives an additional insight into the structure of (3.14) and turns to be very efficient in the continuous time setup. It is based on the following simple fact

LEMMA 3.11. *Let $(\Omega, \mathcal{F})$ be a probability space and let $\mathrm{P}$ and $\widetilde{\mathrm{P}}$ be equivalent probability measures on $\mathcal{F}$, i.e. $\mathrm{P} \sim \widetilde{\mathrm{P}}$. Denote by $\mathrm{E}(\cdot|\mathcal{G})$ and $\widetilde{\mathrm{E}}(\cdot|\mathcal{G})$ the conditional expectations with respect to $\mathcal{G} \subseteq \mathcal{F}$ under $\mathrm{P}$ and $\widetilde{\mathrm{P}}$. Then for any $X$, $\mathrm{E}|X| < \infty$*

$$\mathrm{E}\big(X|\mathcal{G}\big) = \frac{\widetilde{\mathrm{E}}\big(X\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\big)}{\widetilde{\mathrm{E}}\big(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\big)}. \tag{3.20}$$

PROOF. Note first that the right hand side of (3.20) is well defined (on the sets of full P-probability[15]) , since

$$\mathrm{P}\left(\widetilde{\mathrm{E}}\Big(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\Big) = 0\right) = \widetilde{\mathrm{E}}\mathbf{1}\left(\widetilde{\mathrm{E}}\Big(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\Big) = 0\right)\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega) = \\ \widetilde{\mathrm{E}}\mathbf{1}\left(\widetilde{\mathrm{E}}\Big(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\Big) = 0\right)\widetilde{\mathrm{E}}\left(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)\Big|\mathcal{G}\right) = 0.$$

Clearly the right hand side of (3.20) is $\mathcal{G}$-measurable and for any $A \in \mathcal{G}$

$$\mathrm{E}\left(X - \frac{\widetilde{\mathrm{E}}\big(X\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\big)}{\widetilde{\mathrm{E}}\big(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\big)}\right)\mathbf{1}_A(\omega) = \widetilde{\mathrm{E}}\left(X - \frac{\widetilde{\mathrm{E}}\big(X\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\big)}{\widetilde{\mathrm{E}}\big(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\big)}\right)\mathbf{1}_A(\omega)\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega) = \\ = \widetilde{\mathrm{E}}X\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)\mathbf{1}_A - \widetilde{\mathrm{E}}\frac{\widetilde{\mathrm{E}}\big(X\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\big)}{\widetilde{\mathrm{E}}\big(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\big)}\mathbf{1}_A\widetilde{\mathrm{E}}\left(\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega)|\mathcal{G}\right) = 0,$$

which verifies the claim. □

This lemma suggests the following way of calculating the conditional probabilities: find a *reference* measure $\widetilde{\mathrm{P}}$, equivalent to P, under which calculation of the conditional expectation would be easier (typically, $\widetilde{\mathrm{P}}$ is chosen so that $X$ is independent of $\mathcal{G}$) and use (3.20).

Assume the following structure for the observation process [16] (all the other assumptions remain the same)

---

[15] and thus also $\widetilde{\mathrm{P}}$-probability

[16] greater generality is possible with the reference measure approach, but is sacrificed here for the sake of clarity

* $Y_j = h(X_j) + \xi_j$, where $h$ is a measurable function $\mathbb{R} \mapsto \mathbb{R}$ and $\xi = (\xi_j)_{j\ge 0}$ is an i.i.d. sequence, independent of $X$, such that $\xi_1$ has a positive density $q(u) > 0$ with respect to the Lebesgue measure:

$$\mathrm{P}(\xi_1 \le u) = \int_{-\infty}^{u} q(s)ds.$$

Let's verify the claim of Theorem 3.9 under this assumption. For a fixed $j$, let $\mathcal{F}_j = \mathcal{F}_j^X \vee \mathcal{F}_j^Y$ (or equivalently $\mathcal{F}_j = \mathcal{F}_j^X \vee \mathcal{F}_j^\xi$). Introduce the (positive) random process

$$\Phi_j(X,Y) := \prod_{i=0}^{j} \frac{q(Y_i)}{q\big(Y_i - h(X_i)\big)}. \tag{3.21}$$

and define the probability measure $\widetilde{\mathrm{P}}$ (on $\mathcal{F}_j$) by means of the Radon-Nikodym derivative

$$\frac{d\widetilde{\mathrm{P}}}{d\mathrm{P}}(\omega) = \Phi_j\big(X(\omega), Y(\omega)\big),$$

with respect to the restriction of P on $\mathcal{F}_j$. $\widetilde{\mathrm{P}}$ is indeed a probability measure, since $\Phi_j$ is positive and

$$\begin{aligned}\widetilde{\mathrm{P}}(\Omega) =& \mathrm{E}\Phi_j(X,Y) = \mathrm{E}\prod_{i=0}^{j} \frac{q(Y_i)}{q\big(Y_i - h(X_i)\big)} = \mathrm{E}\prod_{i=0}^{j} \frac{q\big(h(X_i)+\xi_i\big)}{q(\xi_i)} = \\ &\mathrm{E}\int_{\mathbb{R}}\dots\int_{\mathbb{R}}\prod_{i=0}^{j}\frac{q\big(h(X_i)+u_i\big)}{q(u_i)}\prod_{\ell=0}^{j} q(u_\ell)du_0\dots du_j = \\ &\mathrm{E}\int_{\mathbb{R}}\dots\int_{\mathbb{R}}\prod_{i=0}^{j} q\big(h(X_i)+u_i\big)du_0\dots du_j = \\ &\mathrm{E}\prod_{i=0}^{j}\int_{\mathbb{R}} q\big(h(X_i)+u_i\big)du_i = 1\end{aligned}$$

Under measure $\widetilde{\mathrm{P}}$, the random processes $(X,Y)$ "look" absolutely different:

(i) the distribution of the process[17] $Y$ under $\widetilde{\mathrm{P}}$, coincides with the distribution of $\xi$ under P
(ii) the distribution of the process $X$ is the same under both measures P and $\widetilde{\mathrm{P}}$
(iii) the processes $X$ and $Y$ are independent under $\widetilde{\mathrm{P}}$

[17] of course the restriction of $Y$ to $[0,j]$ is meant here

Let $\psi(x_0, ..., x_j)$ and $\phi(x_0, ..., x_j)$ be measurable bounded $\mathbb{R}^{j+1} \mapsto \mathbb{R}$ functions. Then

$$\widetilde{\mathrm{E}}\psi(X_0, ..., X_j)\phi(Y_0, ..., Y_j) = \mathrm{E}\psi(X_0, ..., X_j)\phi(Y_0, ..., Y_j)\Phi_j(X, Y) =$$
$$\mathrm{E}\psi(X_0, ..., X_j)\phi(Y_0, ..., Y_j)\prod_{i=0}^{j}\frac{q(Y_i)}{q\big(Y_i - h(X_i)\big)} =$$
$$\mathrm{E}\psi(X_0, ..., X_j)\phi\big(h(X_0) + \xi_0, ..., h(X_j) + \xi_j\big)\prod_{i=0}^{j}\frac{q\big(h(X_i) + \xi_i\big)}{q\big(\xi_i\big)} =$$
$$\mathrm{E}\psi(X_0, ..., X_j)\int_{\mathbb{R}} ... \int_{\mathbb{R}} \phi\big(h(X_0) + u_0, ..., h(X_j) + u_j\big)\cdot$$
$$\prod_{i=0}^{j}\frac{q\big(h(X_i) + u_i\big)}{q\big(u_i\big)}\prod_{\ell=0}^{j} q(u_\ell)du_0...du_j =$$

$$\mathrm{E}\psi(X_0, ..., X_j)\int_{\mathbb{R}} ... \int_{\mathbb{R}} \phi\big(h(X_0) + u_0, ..., h(X_j) + u_j\big)\prod_{i=0}^{j} q\big(h(X_i) + u_i\big)du_0...du_j =$$
$$\mathrm{E}\psi(X_0, ..., X_j)\int_{\mathbb{R}} ... \int_{\mathbb{R}} \phi\big(u_0, ..., u_j\big)\prod_{i=0}^{j} q\big(u_i\big)du_0...du_j =$$
$$\mathrm{E}\psi(X_0, ..., X_j)\mathrm{E}\phi\big(\xi_0, ..., \xi_j\big).$$

Now the claim (i) holds by arbitrariness of $\phi$ with $\psi \equiv 1$. Similarly the (ii) holds by arbitrariness of $\psi$ with $\phi \equiv 1$. Finally, if (i) and (ii) hold then,

$$\widetilde{\mathrm{E}}\psi(X_0, ..., X_j)\phi(Y_0, ..., Y_j) = \mathrm{E}\psi(X_0, ..., X_j)\phi\big(\xi_0, ..., \xi_j\big) =$$
$$\widetilde{\mathrm{E}}\psi(X_0, ..., X_j)\widetilde{\mathrm{E}}\phi\big(Y_0, ..., Y_j\big),$$

which is nothing but (iii) by arbitrariness of $\phi$ and $\psi$.

Now by Lemma 3.11 for any bounded function $g$,

$$\mathrm{E}\big(g(X_j)|\mathcal{F}_j^Y\big) = \frac{\widetilde{\mathrm{E}}\big(g(X_j)\Phi_j^{-1}(X, Y)|\mathcal{F}_j^Y\big)}{\widetilde{\mathrm{E}}\big(\Phi_j^{-1}(X, Y)|\mathcal{F}_j^Y\big)} = \frac{\check{\mathrm{E}}g(X_j(\breve{\omega}))\Phi_j^{-1}(X(\breve{\omega}), Y(\omega))}{\check{\mathrm{E}}\Phi_j^{-1}(X(\breve{\omega}), Y(\omega))} \tag{3.22}$$

where $\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega) = \Phi_j^{-1}(X, Y)$. The latter equality is due to independence of $X$ and $Y$ under $\widetilde{\mathrm{P}}$ (the notations of Remark 3.6 are used here).

Now for arbitrary (measurable and bounded) function $g$

$$\check{\mathrm{E}}g(X_j(\breve{\omega}))\Phi_j^{-1}(X(\breve{\omega}), Y(\omega)) = \check{\mathrm{E}}g(X_j(\breve{\omega}))\Big(\Phi_j^{-1}(X(\breve{\omega}), Y(\omega))|X_j\Big) =$$
$$\int_{\mathbb{R}} g(u)\check{\mathrm{E}}\Big(\Phi_j^{-1}(X(\breve{\omega}), Y(\omega))|X_j = u\Big)P^{X_j}(du) := \int_{\mathbb{R}} g(u)\rho_j(du)$$

On the other hand

$$
\begin{aligned}
&\check{\mathrm{E}} g(X_j(\check\omega))\Phi_j^{-1}(X(\check\omega),Y(\omega)) = \\
&\check{\mathrm{E}}\Phi_{j-1}^{-1}(X(\check\omega),Y)\check{\mathrm{E}}\left( g(X_j(\check\omega))\frac{q\big(Y_j-h(X_j(\check\omega))\big)}{q(Y_j)}\Big|\mathcal{F}_{j-1}^X\right) = \\
&\check{\mathrm{E}}\Phi_{j-1}^{-1}(X(\check\omega),Y)\int_{\mathbb{R}} g(u)\frac{q\big(Y_j-h(u)\big)}{q(Y_j)}\Lambda(X_{j-1}(\check\omega),du) = \\
&\int_{\mathbb{R}} g(u)\int_{\mathbb{R}}\frac{q\big(Y_j-h(u)\big)}{q(Y_j)}\Lambda(s,du)\rho_{j-1}(ds).
\end{aligned}
$$

By arbitrariness of $g$, the recursion

$$
\rho_j(du) = \int_{\mathbb{R}}\frac{q\big(Y_j-h(u)\big)}{q(Y_j)}\Lambda(s,du)d\rho_{j-1}(s). \tag{3.23}
$$

is obtained. Finally by (3.22)

$$
\mathrm{E}\big(g(X_j)|\mathcal{F}_j^Y\big) = \frac{\int_{\mathbb{R}} g(u)\rho_j(du)}{\int_{\mathbb{R}}\rho_j(du)}
$$

and hence the conditional distribution $\pi_j(du)$ from Theorem 3.9 can be calculated by normalizing

$$
\pi_j(du) = \frac{\rho_j(du)}{\int_{\mathbb{R}}\rho_j(ds)}. \tag{3.24}
$$

Besides verifying (3.14), the latter suggests that $\pi_j(du)$ can be calculated by solving linear (!) equation (3.23), whose solution $\rho_j(du)$ (which is called the *unnormalized* conditional distribution) is to be normalized at the final time $j$. In fact this remarkable property can be guessed directly from (3.14) (under more general assumptions on $Y$).

## 4. The curse of dimensionality and finite dimensional filters

The equation (3.14) (or its unnormalized counterpart (3.23)) are not very practical solutions to the estimation problem: at each step they require at least two integrations! Clearly the following property would be very desirable

Definition 3.12. The filter is called *finite* dimensional with respect to a function $f$, if the right hand side of (3.16) can be parameterized by a finite number of sufficient statistics, i.e. solutions of real valued difference equations, driven by $Y$.

The evolution of $\pi_j$ can be infinite-dimensional, while the integral of $\pi_j$ versus specific function $f$ may admit a finite dimensional filter (see Exercise **21**). Unfortunately there is no easy way to determine whether the nonlinear filter at hand is finite dimensional. Moreover sometimes it can be proved to be infinite dimensional. In fact few finite dimensional filters are known, the most important of which are described in the following sections.

**4.1. The Hidden Markov Models (HMM).** Suppose that $X_j$ is a Markov chain with a finite state space $\mathbb{S} = \{a_1, ..., a_d\}$. Then its Markov kernel is identified[18] with the matrix $\Lambda$ of transition probabilities $\lambda_{\ell m} = \mathrm{P}(X_j = a_m | X_{j-1} = a_\ell)$. Let $p_0$ be the initial distribution of $X$, i.e. $p_0(\ell) = \mathrm{P}(X_0 = a_\ell)$. Suppose that the observation sequence $Y = (Y_j)_{j\ge 1}$ satisfies

$$\mathrm{P}(Y_j \in A | \mathcal{F}_j^X \vee \mathcal{F}_{j-1}^Y) = \int_A \nu_\ell(du), \quad \ell = 1, ..., d.$$

Note that each $\nu_\ell(du)$ is absolutely continuous with respect to the measure $\nu(du) = \sum_{m=1}^d \nu_m(du)$ and so no generality is lost if $\nu_\ell(du) = f_\ell(u)\nu(du)$ is assumed for some fixed $\sigma$-finite measure on $\mathcal{B}(\mathbb{R})$ and densities $f_\ell(u)$. This statistical model is extremely popular in various areas of engineering (see [**7**] for a recent survey).

Clearly the conditional distribution $\pi_j(dx)$ is absolutely continuous with respect to the point measure with atoms at $a_1, ..., a_d$ and so can be identified with the density $\pi_j$, which is just a vector of conditional probabilities $\mathrm{P}(X_j = a_\ell | \mathcal{F}_j^Y)$, $\ell = 1, ..., d$. Then by the formulae (3.14),

$$\pi_j = \frac{D(Y_j)\Lambda^* \pi_{j-1}}{\big|D(Y_j)\Lambda^* \pi_{j-1}\big|}, \tag{3.25}$$

subject to $\pi_0 = p_0$, where $|x| = \sum_{\ell=1}^d |x_\ell|$ ($\ell^1$ norm) of a vector $x \in \mathbb{R}^d$ and $D(y)$ is a scalar matrix with $f_\ell(y)$, $y \in \mathbb{R}$, $\ell = 1, ..., d$ on the diagonal. Alternatively the unnormalized equation can be solved

$$\rho_j = D(Y_j)\Lambda^* \rho_{j-1}, \quad j \ge 1$$

subject to $\rho_0 = p_0$ and then $\pi_j$ is recovered by normalizing $\pi_j = \rho_j / |\rho_j|$. Finite dimensional filters are known for several filtering problems, related to HMM - see Exercise **21**.

**4.2. The linear Gaussian case: Kalman-Bucy filter revisited.** The Kalman-Bucy filter from Chapter 2 has a very special place among the nonlinear filters due to the properties of Gaussian random vectors. Recall that

DEFINITION 3.13. A random vector $X$, with values in $\mathbb{R}^d$, is Gaussian if

$$\mathrm{E}\exp\Big\{i\lambda^* X\Big\} = \exp\Big\{i\lambda^* m - \frac{1}{2}\lambda^* K \lambda\Big\}, \quad \forall \lambda \in \mathbb{R}^d$$

for a vector $m$ and a nonnegative definite matrix $K$.

REMARK 3.14. It is easy to check that $m = \mathrm{E}X$ and $K = \mathrm{cov}(X)$.

It turns out that if characteristic function of a random vector is exponential of a quadratic form, this vector is necessarily Gaussian. Gaussian vectors (processes) play a special role in probability theory. The following properties make them special in the filtering theory in particular:

LEMMA 3.15. *Assume that the vectors $X$ and $Y$ (with values in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively) form a Gaussian vector $(X, Y)$ in $\mathbb{R}^{m+n}$. Then*

(1) *Any random variable from the linear subspace, spanned by the entries of $(X, Y)$ is Gaussian. In particular $Z = b + AX$ with a vector $b$ and a matrix $A$, is a Gaussian vector with* $\mathrm{E}Z = b + A\mathrm{E}X$ *and* $\mathrm{cov}(Z) = A\,\mathrm{cov}(X)A^*$.

---

[18] In this case the Markov kernel is absolutely continuous to the point measure $\sum_{i=1}^d \delta_{a_i}(du)$ and the matrix $\Lambda$ is formally the density w.r.t this measure.

(2) *If $X$ and $Y$ are orthogonal, they are independent (the opposite direction is obvious)*
(3) *The regular conditional distribution of $X$, given $Y$ is Gaussian* P*-a.s., moreover*[19] $\mathrm{E}(X|Y)=\widehat{\mathrm{E}}(X|Y)$ *and*

$$\operatorname{cov}(X|Y):=\mathrm{E}\Big(\big(X-\mathrm{E}(X|Y)\big)\big(X-\mathrm{E}(X|Y)\big)^*|Y\Big)=\\ \operatorname{cov}(X)-\operatorname{cov}(X,Y)\operatorname{cov}^{\oplus}(Y)\operatorname{cov}(Y,X). \quad (3.26)$$

REMARK 3.16. Note that in the Gaussian case the conditional covariance does not depend on the condition !

PROOF. For fixed $b$ and $A$

$$\mathrm{E}\exp\Big\{i\lambda^*(b+AX)\Big\}=\exp\big\{i\lambda^*(b+A\mathrm{E}X)\big\}\mathrm{E}\exp\Big\{i(\lambda^*A)(X-\mathrm{E}X)\Big\}=\\ \exp\Big\{i\lambda^*(b+A\mathrm{E}X)\Big\}\exp\Big\{-\frac{1}{2}\lambda^*\big(A\operatorname{cov}(X)A^*\big)\lambda\Big\},$$

and the claim (1) holds, since the latter is a characteristic function of a Gaussian vector.

Let $\lambda_x$ and $\lambda_y$ be vectors from $\mathbb{R}^m$ and $\mathbb{R}^n$ (so that $\lambda=(\lambda_x,\lambda_y)\in\mathbb{R}^{m+n}$), then due to orthogonality $\operatorname{cov}(X,Y)=0$ and

$$\mathrm{E}\exp\Big\{i\lambda^*(X,Y)\Big\}=\exp\Big\{i\lambda_x^*\mathrm{E}X-\frac{1}{2}\lambda_x^*\operatorname{cov}(X)\lambda_x\Big\}\exp\Big\{i\lambda_y^*\mathrm{E}Y-\frac{1}{2}\lambda_y^*\operatorname{cov}(Y)\lambda_y\Big\},$$

which verifies the second claim.

Recall that $X-\widehat{\mathrm{E}}(X|Y)$ is orthogonal to $Y$, and thus by (2), they are also independent. Then

$$\mathrm{E}\Big(\exp\Big\{i\lambda_x^*\big(X-\widehat{\mathrm{E}}(X|Y)\big)\Big\}|Y\Big)=\mathrm{E}\exp\Big\{i\lambda_x^*\big(X-\widehat{\mathrm{E}}(X|Y)\big)\Big\}$$

and on the other hand

$$\mathrm{E}\Big(\exp\Big\{i\lambda_x^*\big(X-\widehat{\mathrm{E}}(X|Y)\big)\Big\}|Y\Big)=\exp\Big\{-i\lambda_x^*\widehat{\mathrm{E}}(X|Y)\Big\}\mathrm{E}\Big(\exp\Big\{i\lambda_x^*X\Big\}|Y\Big)$$

and so

$$\mathrm{E}\Big(\exp\Big\{i\lambda_x^*X\Big\}|Y\Big)=\exp\Big\{i\lambda_x^*\widehat{\mathrm{E}}(X|Y)\Big\}\mathrm{E}\exp\Big\{i\lambda_x^*\big(X-\widehat{\mathrm{E}}(X|Y)\big)\Big\}.$$

Since $X-\widehat{\mathrm{E}}(X|Y)$ is in the linear span of $(X,Y)$, the latter term equals

$$\exp\Big\{i\lambda_x^*\mathrm{E}(X-\widehat{\mathrm{E}}(X|Y))-\frac{1}{2}\lambda_x^*\operatorname{cov}\big(X-\widehat{\mathrm{E}}(X|Y)\big)\lambda_x\Big\},$$

and the third claim follows, since $\mathrm{E}(X-\widehat{\mathrm{E}}(X|Y))=0$ and $\operatorname{cov}\big(X-\widehat{\mathrm{E}}(X|Y)\big)$ equals (3.26). □

Consider now the Kalman-Bucy linear model (2.13) and (2.14) (on page 29), where the sequences $\xi$ and $\varepsilon$ are Gaussian, as well as the initial condition $(X_0,Y_0)$. Then the processes $(X,Y)$ are Gaussian (i.e. any finite dimensional distribution is Gaussian) and by Lemma 3.15, the conditional distribution of $X_j$ given $\mathcal{F}_j^Y$ is Gaussian too. Moreover its parameters - the mean and the covariance are governed by the Kalman-Bucy filter equations from Theorem 2.5.

[19] in other notations $\mathrm{E}(X|\mathcal{F}^Y)=\widehat{\mathrm{E}}(X|\mathcal{L}^Y)$

REMARK 3.17. The recursions of Theorem 2.5 can be obtained via the nonlinear filtering equation (3.14), using certain properties of the Gaussian densities. Note however that guessing the Gaussian solution to (3.14) would not be easy !

In particular for any measurable $f$, such that $\mathrm{E}|f(X_j)| < \infty$ (the scalar case is considered for simplicity)

$$\mathrm{E}\big(f(X_j)|\mathcal{F}_j^Y\big) = \int_{\mathbb{R}} f(u)\frac{1}{\sqrt{2\pi P_j}}\exp\left\{-\frac{(u-\widehat{X}_j)^2}{2P_j}\right\}du,$$

where $P_j$ nd $\widehat{X}_j$ are generated by the Kalman-Bucy equations. In Exercise **24** an important generalization of the Kalman-Bucy filter is considered. More models, for which finite dimensional filter exists are known, but their practical applicability is usually limited.

## Exercises

(**1**) Verify the properties of the conditional expectations on page 40
(**2**) Prove that pre-images of Borel sets of $\mathbb{R}$ under a measurable function (random variable) is a $\sigma$-algebra
(**3**) Prove (3.6) (use monotone convergence theorem - see Appendix).
(**4**) Obtain the formula (3.3) by means of (3.11).
(**5**) Verify the claim of Remark 3.7.
(**6**) Explore the definition of the Markov process on page 43: argue the existence, etc. How such process can be generated, given say a source of i.i.d. random variables with uniform distribution ?
(**7**) Is $Y$, defined in (3.13) a Markov process? Is the pair $(X_j, Y_j)$ a (two dimensional) Markov process?
(**8**) Show that $\mathrm{P}\Big(\check{\mathrm{E}}L_j(X(\check{\omega}), Y) = 0\Big) = 0$ ($L_j(X,Y)$ is defined in (3.19)).
(**9**) Complete the proof of Theorem 3.9 (i.e. verify (3.15)).
(**10**) Derive (3.14) and (3.15), using the orthogonality property of the conditional expectation (similarly to derivation of (3.3)).
(**11**) Show that (3.23) and (3.24) imply (3.14).
(**12**) Derive the nonlinear filtering equations when $Y$ is defined with "delay":

$$\mathrm{P}(Y_j \in B|\mathcal{F}_{j-1}^X, \mathcal{F}_{j-1}^Y) = \int_B \gamma(X_{j-1}; du), \quad \mathrm{P}-a.s$$

(**13**) Discuss the changes, which have to be introduced into (3.14), when $X$ and $Y$ take values in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively (the multivariate case)
(**14**) Discuss the changes, which have to be introduced into (3.14), when the Markov kernels $\Lambda$ and $\gamma$ are allowed to depend on $j$ (time dependent case) and $\mathcal{F}_{j-1}^Y$ (dependence on the past observations).
(**15**) Show that if the transition matrix $\Lambda$ of the finite state chain $X$ is $q$-primitive, i.e. the matrix $\Lambda^q$ has all positive entries for some integer $q \geq 1$, then the limits $\lim_{j\to\infty} \mathrm{P}\big(X_j = a_\ell\big) = \mu_\ell$ exist, are positive for all $a_\ell \in \mathbb{S}$ and independent of the initial distribution (such chain is called ergodic).

(**16**) Find the filtering recursion for the signal/observation model

$$\begin{aligned} X_j &= g(X_{j-1}) + \varepsilon_j, \quad j \ge 1 \\ Y_j &= f(X_j) + \xi_j \end{aligned}$$

subject to a random initial condition $X_0$ (and $Y_0 \equiv 0$), independent of $\varepsilon$ and $\xi$. Assume that $g : \mathbb{R} \mapsto \mathbb{R}$ and $f : \mathbb{R} \mapsto \mathbb{R}$ are measurable functions, such that $\mathrm{E}|g(X_{j-1})| < \infty$ and $\mathrm{E}|f(X_j)| < \infty$ for any $j \ge 0$. The sequences $\varepsilon = (\varepsilon_j)_{j\ge1}$ and $\xi = (\xi_j)_{j\ge1}$ are independent and i.i.d., such that $\varepsilon_1$ and $\xi_1$ have densities $p(u)$ and $q(u)$ with respect to the Lebesgue measure on $\mathcal{B}(\mathbb{R})$.

(**17**) Let $X$ be a Markov chain as in Section 4.1 and $Y_j = h(X_j) + \xi_j$, $j \ge 1$, where $\xi = (\xi_j)_{j\ge0}$ is an i.i.d. sequence. Assume that $\xi_1$ has probability density $f(u)$ (with respect to the Lebesgue measure). Write down the equations (3.25) in componentwise notation. Simulate the filter with MATLAB.

(**18**) Show that the filtering process $\pi_j$ from the previous problem is Markov.

(**19**) Under the setting of Section 4.1, denote by $\mathscr{Y}_j$ the family of $\mathcal{F}^Y_j$ - measurable random variables with values in $\mathbb{S}$ (detectors which guess the current symbol of $X_j$, given the observation of $\{Y_1, ..., Y_j\}$). For a random variable $\eta_j \in \mathscr{Y}_j$, let $P_d$ denote the detection error:

$$P_d = \mathrm{P}\Big(\eta_j \ne X_j\Big).$$

Show that the optimal detector, minimizing the detection error in the class $\mathscr{Y}_j$ is given by

$$\widehat{\eta}_j = \mathrm{argmax}_{a_\ell \in \mathbb{S}}\, \pi_j(\ell).$$

Find (an implicit) expression for the minimal detection error.

(**20**) A random switch $\theta_j \in \{0, 1\}$, $j \ge 0$ is a discrete-time two-state Markov chain with transition matrix:

$$\Lambda = \begin{bmatrix} \lambda_1 & 1-\lambda_1 \\ 1-\lambda_2 & \lambda_2 \end{bmatrix}.$$

Assume that $\theta_0 = 1$.

A counter $\xi_j$, counts arrivals (of e.g. particles) from two independent sources with different intensities $\alpha$ and $\beta$. The counter is connected according to the state of the switch $\theta_j$ to one source or another, so that:

$$\xi_j = \xi_{j-1} + \mathbf{1}(\theta_j = 1)\varepsilon^\alpha_j + \mathbf{1}(\theta_j = 0)\varepsilon^\beta_j, \quad j = 1, 2, ...$$

subject $\xi_0 = 0$. Here $\beta$ and $\alpha$ are constants from the interval $(0, 1)$ and $\varepsilon^\gamma_j \in \{0, 1\}$ stands for an i.i.d. sequence with $\mathrm{P}\{\varepsilon^\gamma_j = 1\} = \gamma$ $(0 < \gamma < 1)$.

(a) Find the optimal estimate of the switch state, given the counter data up to the current moment, i.e. derive the recursion for $\pi_j = \mathrm{E}(\theta_j|\mathcal{F}^\xi_j)$.

(b) Study the behavior of the filter in the limit cases:
   (i) $\alpha = 1$ and $\beta = 0$ (simultaneously).
   (ii) $\lambda_1 = 1$ and $\lambda_2 = 0$ (and vice versa).
   (iii) $\lambda_1 = \lambda_2 = 1$

(**21**) Let $\theta_j$ be the number of times, a finite state Markov chain $X$ visited ("occupied") the state $a_1$ (or any other fixed state) up to time $j$. Find the recursion for calculation of the optimal estimate of the occupation time $\mathrm{E}(\theta_j|\mathcal{F}_j^Y)$, where $Y$ is defined as in Section 4.1.

(a) Let $I_j$ be the vector of indicators $\mathbf{1}_{\{X_j=a_i\}}$, $i=1,...,d$ and define $Z_j := \theta_j I_j$. Find the expression for $\bar{Z}_{j|j-1} := \mathrm{E}(Z_j|\mathcal{F}_{j-1}^Y)$ in terms of $\bar{Z}_{j-1} = \mathrm{E}(Z_{j-1}|\mathcal{F}_{j-1}^Y)$ and $\pi_{j|j-1} = \Lambda^* \pi_j$.

(b) Find the expression of $\bar{Z}_j$ in terms of $\bar{Z}_{j|j-1}$ and thus "close" the recursion for $\bar{Z}_j$.

(c) How $\mathrm{E}(\theta_j|\mathcal{F}_j^Y)$ is recovered from $\bar{Z}_j$?

**(22)** Let $\tau_j$ be the number of transitions from state $a_1$ to state $a_2$ (or any other fixed pair of states), a finite state Markov chain $X$ made on the time interval $[1,j]$. Find the finite dimensional filter for $\mathrm{E}(\tau_j|\mathcal{F}_j^Y)$. **Hint:** use the approach suggested in the previous problem.

**(23)** Check the claim of Remark 3.14.

**(24)** Consider the signal/observation model $(X_j, Y_j)_{j\ge 0}$:

$$\begin{aligned} X_j &= a_0(Y_0^{j-1}) + a_1(Y_0^{j-1})X_{j-1} + b\varepsilon_j, \quad j=1,2,... \\ Y_j &= A_0(Y_0^{j-1}) + A_1(Y_0^{j-1})X_{j-1} + B\xi_j \end{aligned}$$

where $b$ and $B$ are constants and $A_i(Y_0^{j-1})$ and $a_i(Y_0^{j-1})$, $i=0,1$ are some functionals of the vector $\{Y_0, Y_1, ..., Y_{j-1}\}$. $\varepsilon = (\varepsilon_j)_{j\ge 1}$ and $\xi = (\xi_j)_{j\ge 1}$ are independent i.i.d. standard Gaussian random sequences. The initial condition $(X_0, Y_0)$ is a standard Gaussian vector with unit covariance matrix, independent of $\varepsilon$ and $\xi$.

(a) Is the pair of processes $(X_j, Y_j)_{j\ge 0}$ necessarily Gaussian ? Give a proof or a counterexample.

(b) Find the recursion for $\widehat{X}_j = \mathrm{E}(X_j|\mathcal{F}_j^Y)$ and $P_j = \mathrm{E}\Big((X_j - \widehat{X}_j)^2|\mathcal{F}_j^Y\Big)$. Is the obtained filter linear w.r.t. observations ? Does the error $P_j$ depend on the observations ?

**Hint:** prove first that $X_j$ is Gaussian, conditioned on $\mathcal{F}_j^Y$.

Remark 3.18. The filtering recursion in this case is sometimes referred as *conditionally Gaussian filter*. It plays an important role in control theory, where the coefficients usually depend on the past observations.

(c) Verify that in the case of $a_i(Y_0^{j-1}) \equiv a_i$ and $A_i(Y_0^{j-1}) \equiv A_i$, $i=0,1$ ($a_i$ and $A_i$ constants) your solution coincides with the Kalman-Bucy filter.

**(25)** Consider the recursion

$$X_j = aX_{j-1} + \varepsilon_j, \quad j \ge 1$$

subject to a standard Gaussian random variable $X_0$ and where $\varepsilon$ is a Gaussian i.i.d. sequence, independent of $X_0$. Assuming that the parameter $a$ is a Gaussian random variable independent of $\varepsilon$ and $X_0$, derive a recursion for $\mathrm{E}(a|\mathcal{F}_j^X)$ and for the square error

$$P_j = \mathrm{E}\Big(\big(a - \mathrm{E}(a|\mathcal{F}_j^X)\big)^2 \big|\mathcal{F}_j^X\Big).$$

Is the recursion for $\mathrm{E}(a|\mathcal{F}_j^X)$ linear ? Does $P_j$ converge ? If yes, to which limit and in which sense? **Hint:** use the results of the previous exercise.

**(26)** Consider a signal/observation pair $(\theta, \xi_j)_{j\ge 1}$, where $\theta$ is a random variable distributed uniformly on $[0,1]$ and $(\xi_j)$ is a sequence generated by:

$$\xi_j = \theta U_j$$

where $(U_j)_{j\ge 1}$ is a sequence of i.i.d. random variables with uniform distribution on $[0,1]$. $\theta$ and $U$ are independent.

(a) Derive the Kalman-Bucy filter for $\widehat{\theta}_j = \widehat{\mathrm{E}}(\theta|\xi_1^j)$.

(b) Find the corresponding mean square error $P_j = \mathrm{E}(\theta - \widehat{\theta}_j)^2$. Show that it converges to zero as $j \to \infty$ and determine the rate of convergence [20]

(c) Consider the recursive filtering estimate $(\widetilde{\theta}_j)_{j\ge 0}$

$$\widetilde{\theta}_j = \max(\widetilde{\theta}_{j-1}, \xi_j), \quad \widetilde{\theta}_0 = 0$$

Find the corresponding mean square error, $Q_j = \mathrm{E}(\theta - \widetilde{\theta}_j)^2$.

(d) Show that $Q_j$ converges to zero and find the rate of convergence. Does this filter give better accuracy, compared to Kalman-Bucy filter, uniformly in $j$ ? Asymptotically in $j \to \infty$?

(e) Verify whether $\widetilde{\theta}_j$ is the optimal in the mean square sense filtering estimate. If not, find the optimal estimate $\bar{\theta}_j = \mathrm{E}(\theta|\mathcal{F}_j^\xi)$.

[20] i.e. find a sequence $r_j$, such that $\lim_{j\to\infty} r_j P_j$ exists and positive

CHAPTER 4

# The white noise in continuous time

A close look at the derivation of nonlinear filtering recursions reveals that one of the crucial assumptions is independence of the observation noise on the past. The model (3.13) is in fact a generalization of the following "additive white noise" observation scenario

$$Y_j = h(X_j) + \xi_j, \quad j \geq 0 \tag{4.1}$$

where $h$ is a measurable function and $\xi$ is an i.i.d. sequence. As was mentioned in the introduction, the term "white noise" stems from the fact that power spectral density of the sequence $\xi$ (when $\mathrm{E}\xi_1^2 < \infty$), defined as the Fourier transform of the correlation sequence $R(n) = \mathrm{E}\xi_0\xi_n$, is constant. In the continuous time case similar definition would be meaningless both for mathematical and physical reasons: the sample pathes of such process would be extremely irregular (e.g. not even continuous in any point) and its variance is infinite. It turns out that overcoming this difficulty is not an easy mathematical task. It is accomplished in several steps

**i.** Introduce a continuous time process with independent increments. The motivation is that a formal derivative of such process is a "white noise" (recall the discussion on page 10). It turns out that such a process can be constructed (the Wiener process), but it is not differentiable in any reasonable sense. At this point the hope for real "white noise" is abandoned and instead of considering problems involving differentials (e.g. differential equations, etc.), their integral analogues are considered.

**ii.** This naturally leads to considering integration with respect to the Wiener process. It turns out however that the Wiener process has irregular trajectories, so that all the classical integration approaches (e.g. Stieltjes, Lebesgue, etc.) fail. However integration can be carried out if the family of integrands is chosen in a special way. Specifically we will use the stochastic integral introduced by K.Itô

**iii.** After introducing the integral, one is led to establish the rules to manipulate the new object: e.g. the change of integration variable, chain rule, etc. Surprisingly (or not!) the Itô integral have properties, dramatically different from the classical integration. The particularly useful tool in, what is called by now, *the stochastic calculus*, is the Itô formula.

**iv.** Once there is a new calculus, the ultimate goal is accomplished: the stochastic differential equations are introduced. The term "differential" is in fact misleading, though customary: actually the *integral* equations involving usual Riemann integrals and Itô integrals are considered. It turns out that besides *strong* solutions (roughly speaking analogous to the usual solutions of ODE), one may consider *weak* solutions, which have no analogue in classical ODE's. We will be concerned mainly with the first kind of solutions, though weak solutions play an important role in filtering in particular.

REMARK 4.1. The introductory scope of these lectures doesn't include many important concepts and details from the vast theory of random processes in continuous time. The reader may and should consult basic books in this area for deeper understanding. The author's choice was and still is: the classic J.Doob's book [**5**] and the modern [**39**] for general concepts of stochastic processes in continuous time, the book [**18**] is a good starting point for further study of the Brownian motion and stochastic calculus, the first volume of [**21**] is a confined but very accessible coverage of stochastic Itô calculus and its applications (collected in the second volume).

## 1. The Wiener process

The main building block of the white noise theory is the Wiener process (or mathematical Brownian motion), which is defined (on some probability space $(\Omega, \mathcal{F}, \mathrm{P})$) as a stochastic process $W = (W_t(\omega))_{t\in\mathbb{R}_+}$, satisfying the properties

(1) $W_0(\omega) = 0$, $\mathrm{P} - a.s.$
(2) the trajectories of $W$ are continuous functions
(3) the increments of $W$ are independent Gaussian random variables with zero mean and $\mathrm{E}(W_t - W_s)^2 = t - s$, $t \ge s$.

**1.1. Construction.** The existence of such process is not at all clear. There are many constructions of $W$ (see e.g. [**18**]) of which we choose the one due to P.Levy (Section 2.3 in [**18**])

THEOREM 4.2. *The Wiener process $W = (W_t)_{t\in[0,1]}$ exists.*

PROOF. Let $I(n)$ denote the odd integers from $\{0, 1, ..., 2^n\}$. Define the Haar functions as $H_1^0(t) = 1$, $t \in [0,1]$ and $n \ge 1$, $k \in I(n)$

$$H_k^n(t) = \begin{cases} 2^{-(n-1)/2}, & \frac{k-1}{2^n} \le t < \frac{k}{2^n} \\ -2^{-(n-1)/2}, & \frac{k}{2^n} \le t < \frac{k+1}{2^n} \\ 0 & \text{otherwise} \end{cases}.$$

The Schauder functions are

$$S_k^n = \int_0^t H_k^n(s)ds,$$

which do not overlap for different $k$, when $n$ is fixed, and have a "tent" like shape.

Let $\xi_j^n$, $j \in I(n)$, $n = 1, ...$ be an array of i.i.d. standard Gaussian random variables. Introduce the sequence of random processes, $n \ge 0$

$$W_t^n = \sum_{m=0}^n \sum_{k\in I(m)} \xi_k^m S_k^m(t), \quad t \in [0,1], \tag{4.2}$$

Note that $W_t^n$ has continuous trajectories for all $n$. If the sequence $W_t^n$ converges P-a.s. uniformly in $t \in [0,1]$, then the limit process has continuous trajectories as required in axiom 2.

Let's verify the convergence of the series

$$\sum_{m=1}^n \sum_{j\in I(m)} \left|\xi_j^m\right| S_j^m(t) \le \sum_{m=1}^n \max_{j\in I(m)} \left|\xi_j^m\right| \sum_{j\in I(m)} S_j^m(t) \le \sum_{m=1}^n 2^{-(m-1)} \max_{j\le 2^m} \left|\xi_j^m\right| \tag{4.3}$$

(recall that $S^m_j(t)$ do not overlap for a fixed $m$ and different $j$). Since

$$\mathrm{P}\big(|\xi^m_j| \ge x\big) = \frac{2}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du \le \sqrt{\frac{2}{\pi}} \int_x^\infty \frac{u}{x} e^{-u^2/2} du = \sqrt{\frac{2}{\pi}} \frac{e^{-x^2/2}}{x}$$

for $m \ge 1$

$$\mathrm{P}\big(\max_{j \le 2^m} \big|\xi^m_j\big| \ge m\big) = \mathrm{P}\Big(\bigcup_{j \le 2^m} \{|\xi^m_j| > m\}\Big) \le 2^m \mathrm{P}\big(|\xi^m_j| \ge m\big) \le \sqrt{\frac{2}{\pi}} \frac{2^m e^{-m^2/2}}{m}.$$

Since $\sum_{m=1}^\infty 2^m e^{-m^2/2} m^{-1} < \infty$, by Borel-Cantelli Lemma

$$\mathrm{P}\big(\max_{j \le 2^m} \big|\xi^m_j\big| \ge m,\ i.o.\big) = 0.$$

In other words, there is a set $\Omega'$ of full P-measure and a random integer $n(\omega)$, such that $\max_{j \le 2^m} \big|\xi^m_j\big| \le m$ for all $m \ge n(\omega)$ for all $\omega \in \Omega'$. Then the series in (4.3) converge on $\Omega'$ since

$$\sum_{m=n(\omega)}^n 2^{-m} \max_{j \le 2^m} \big|\xi^m_j\big| \le \sum_{m=n(\omega)}^n 2^{-m} m < \infty.$$

So the processes $W^n_t$ converge P-a.s. uniformly in $t$ to a continuous process $W_t$. It is left to verify the axiom 3. The Haar basis forms a complete orthonormal system in the Hilbert space $\mathbb{L}^2[0,1]$ with the scalar product $\langle g, f\rangle = \int_{[0,1]} f(s)g(s)ds$ and so by Parseval equality

$$\langle g, f\rangle = \sum_{n=0}^\infty \sum_{k \in I(n)} \langle g, H^n_k\rangle \langle f, H^n_k\rangle.$$

For $g_u = \mathbf{1}(u \le t)$ and $f(u) = \mathbf{1}(u \le s)$, the latter implies

$$s \wedge t = \sum_{n=0}^\infty \sum_{k \in I(n)} S^n_k(t) S^n_k(s).$$

Now let $\lambda_j$, $j = 1, ..., n$ be real numbers and fix $n$ distinct times $t_1 < ... < t_n$. Then (with $\lambda_{n+1} = 0$)

$$
\begin{aligned}
&\mathrm{E}\exp\Big(-i\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)W^{\ell}_{t_j}\Big) = \\
&\qquad \mathrm{E}\exp\Big(-i\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)\sum_{m=0}^{\ell}\sum_{k\in I(m)}\xi_k^m S_k^m(t_j)\Big) = \\
&\qquad \mathrm{E}\exp\Big(-\sum_{m=0}^{\ell}\sum_{k\in I(m)}\xi_k^m\sum_{j=1}^{n} i(\lambda_{j+1}-\lambda_j)S_k^m(t_j)\Big) = \\
&\qquad \prod_{m=0}^{\ell}\prod_{k\in I(m)}\mathrm{E}\exp\Big(-\xi_k^m\sum_{j=1}^{n} i(\lambda_{j+1}-\lambda_j)S_k^m(t_j)\Big) = \\
&\qquad \prod_{m=0}^{\ell}\prod_{k\in I(m)}\exp\Big(-\frac{1}{2}\Big\{\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)S_k^m(t_j)\Big\}^2\Big) = \\
&\qquad \exp\Big(-\frac{1}{2}\sum_{m=0}^{\ell}\sum_{k\in I(m)}\Big\{\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)S_k^m(t_j)\Big\}^2\Big) = \\
&\qquad \exp\Big(-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\lambda_{j+1}-\lambda_j)(\lambda_{i+1}-\lambda_i)\sum_{m=0}^{\ell}\sum_{k\in I(m)}S_k^m(t_j)S_k^m(t_i)\Big) \xrightarrow{\ell\to\infty} \\
&\qquad \exp\Big(-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\lambda_{j+1}-\lambda_j)(\lambda_{i+1}-\lambda_i)(t_j\wedge t_i)\Big)
\end{aligned}
$$

Then

$$
\begin{aligned}
&\mathrm{E}\exp\Big(i\sum_{j=1}^{n}\lambda_j\big(W_{t_j}-W_{t_{j-1}}\big)\Big) = \mathrm{E}\exp\Big(-i\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)W_{t_j}\Big) = \\
&\exp\Big(-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\lambda_{j+1}-\lambda_j)(\lambda_{i+1}-\lambda_i)(t_j\wedge t_i)\Big) = \\
&\exp\Big(-\sum_{j=1}^{n-1}\sum_{i=j+1}^{n}(\lambda_{j+1}-\lambda_j)(\lambda_{i+1}-\lambda_i)(t_j\wedge t_i)-\frac{1}{2}\sum_{j}^{n}(\lambda_{j+1}-\lambda_j)^2 t_j\Big) = \\
&\exp\Big(-\sum_{j=1}^{n-1}(\lambda_{j+1}-\lambda_j)t_j\sum_{i=j+1}^{n}(\lambda_{i+1}-\lambda_i)-\frac{1}{2}\sum_{j}^{n}(\lambda_{j+1}-\lambda_j)^2 t_j\Big) = \\
&\exp\Big(\sum_{j=1}^{n-1}(\lambda_{j+1}-\lambda_j)t_j\lambda_{j+1}-\frac{1}{2}\sum_{j}^{n}(\lambda_{j+1}-\lambda_j)^2 t_j\Big) = \\
&\exp\Big(\sum_{j=1}^{n-1}t_j\Big\{(\lambda_{j+1}-\lambda_j)\lambda_{j+1}-\frac{1}{2}(\lambda_{j+1}-\lambda_j)^2\Big\}-\frac{1}{2}\lambda_n^2 t_n\Big) =
\end{aligned}
$$

$$\exp\Big(\frac{1}{2}\sum_{j=1}^{n-1} t_j\{\lambda_{j+1}^2-\lambda_j^2\}-\frac{1}{2}\lambda_n^2 t_n\Big)=\exp\Big(-\frac{1}{2}\sum_{j=1}^{n}\lambda_j^2(t_j-t_{j-1})\Big)=$$
$$\prod_{j=1}^{n}\exp\Big(-\frac{1}{2}\lambda_j^2(t_j-t_{j-1})\Big),$$

which verifies axiom 2. □

REMARK 4.3. The Wiener process on $[0,\infty)$ can be constructed by patching the Wiener processes on the intervals $[j,j+1]$, $j=0,1,....$

REMARK 4.4. Though Gaussian distribution of the i.i.d. random variables in this proof plays crucial role, the Gaussian property of the limit $W$ is "universal": it turns out that any continuous time process with independent increments (a martingale!), continuous trajectories and variance $t$ is the Wiener process. Roughly speaking, this suggests that the "white noise", which originates from a random process with these properties is necessarily Gaussian! More exactly

THEOREM 4.5. *(P. Levy) Let $B_t$ be a continuous process with* $\mathrm{E}B_t\equiv 0$, *$t\geq 0$ and*

$$\mathrm{E}\big(B_t^2-t|\mathcal{F}_s^B\big)=B_s^2-s,\quad t\geq s\geq 0.$$

*Then $B_t$ is a Wiener process.*

REMARK 4.6. Sometimes it is convenient to relate the Wiener process to some filtration $\mathcal{F}_t$, by extending the definition in the following way: $W_t$ is the Wiener process with respect to a filtration $\mathcal{F}_t$, if $W$ has continuous pathes, starts from zero and for any $t\geq s\geq 0$, $W_t-W_s$ is a Gaussian random variable, independent of $\mathcal{F}_s$, with zero mean and variance $(t-s)$. The previous definition reduces to the case $\mathcal{F}_t\equiv\mathcal{F}_t^W=\{W_s, s\leq t\}$.

**1.2. Nondifferentiability of the pathes.** The properties of the trajectories of $W$ are really amazing and up to now do not cease to attract attention of mathematicians. We will verify a few of them, which are crucial to understanding the origins of stochastic calculus.

For a function $f:[0,1]\mapsto\mathbb{R}$, denote by $D^{\pm}$ the upper left and right Deni derivatives at $t$:

$$D^{\pm}f(t)=\overline{\lim_{h\to 0\pm}}\frac{f(t+h)-f(t)}{h}$$

and by $D_{\pm}(t)$ the lower left and right Deni derivatives at $t$:

$$D_{\pm}f(t)=\underline{\lim_{h\to 0\pm}}\frac{f(t+h)-f(t)}{h}.$$

The function is differentiable at $t$ from the right if $D^+f(t)$ and $D_+f(t)$ are finite and coincide. Similarly left differentiability is defined by means of $D^-f(t)$ and $D_-f(t)$. If all the Deni derivatives are equal, $f$ is differentiable at $t$. Differentiability at $t=0$ and $t=1$ is defined as right and left differentiability respectively.

THEOREM 4.7. *(Paley, Wiener and Zygmund, 1933) The Wiener process has nowhere differentiable trajectories, more precisely*

$$\mathrm{P}\Big(\omega: \textit{for each } t<1, \textit{either } D^+W_t=\infty \textit{ or } D_+W_t=-\infty\Big)=1.$$

PROOF. For fixed $j,k\geq 0$, define the sets

$$A_{jk}=\bigcup_{t\in[0,1]}\bigcap_{h\in[0,1/k]}\Big\{\omega:\big|W_{t+h}-W_t\big|\leq jh\Big\}.$$

Clearly

$$\Big\{\omega:-\infty<D_+W_t\leq D^+W_t<\infty\Big\}\subseteq\bigcup_{j\geq 1}\bigcup_{k\geq 1}A_{jk}$$

and so to verify the claim, it would be enough to show that $\mathrm{P}\big(A_{jk}\big)=0$ for any $j,k$. Fix a trajectory in the set $A_{jk}$. For this trajectory there exists a number $t\in[0,1]$, such that $\big|W_{t+h}-W_t\big|\leq jh$ for any $0\leq h\leq 1/k$. Fix an integer $n\geq 4k$ and let $1\leq i\leq n$ be such that $(i-1)/n\leq t\leq i/n$. Then we have

$$\begin{aligned}
&\big|W_{(i+1)/n}-W_{i/n}\big|\leq\big|W_{(i+1)/n}-W_t\big|+\big|W_t-W_{i/n}\big|\leq\frac{2j}{n}+\frac{j}{n}=\frac{3j}{n}\\
&\big|W_{(i+2)/n}-W_{(i+1)/n}\big|\leq\big|W_{(i+2)/n}-W_t\big|+\big|W_t-W_{(i+1)/n}\big|\leq\frac{3j}{n}+\frac{2j}{n}=\frac{5j}{n}\\
&\big|W_{(i+3)/n}-W_{(i+2)/n}\big|\leq\big|W_{(i+3)/n}-W_t\big|+\big|W_t-W_{(i+2)/n}\big|\leq\frac{4j}{n}+\frac{3j}{n}=\frac{7j}{n}.
\end{aligned}$$

Then $A_{jk}\subseteq\bigcup_{i=1}^n C_i^{(n)}$ with

$$C_i^{(n)}=\bigcap_{r=1}^3\Big\{\big|W_{(i+r)/n}-W_{(i+r-1)/n}\big|\leq\frac{(2r+1)j}{n}\Big\}.$$

hold for any $n\geq 4k$ or in other words

$$A_{jk}\subseteq\bigcap_{n\geq 4k}\bigcup_{i=1}^n C_i^{(n)}:=C.$$

Note that since $W_{(i+r)/n}-W_{(i+r-1)/n}$ are independent and Gaussian with zero mean and variance $1/\sqrt{n}$,

$$\mathrm{P}\big(C_i^{(n)}\big)\leq\frac{3\cdot 5\cdot 7j^3}{n^{3/2}},$$

where the inequality $\mathrm{P}(|\xi|\leq\varepsilon)\leq\varepsilon$ for a standard Gaussian r.v. $\xi$, have been used. Then $\mathrm{P}(A_{jk})\leq P(C)\leq\inf_{n\geq 4k}\mathrm{P}(\cup_{i=1}^n C_i^{(n)})=0$, where the latter holds since

$$\mathrm{P}\big(\cup_{i=1}^n C_i^{(n)}\big)\leq\sum_{i=1}^n\mathrm{P}(C_i^{(n)})=\frac{105j^3}{n^{1/2}}\xrightarrow{n\to\infty}0.$$

$\square$

Recall that the $p$-variation of the function $f:[0,1]\mapsto\mathbb{R}$ on the partition $\Pi^n=\{t_i\}$, $0=t_0<...<t_{n+1}=1$ is

$$\bigvee_{\Pi^n}^p f(t):=\sum_{t_{i+1}\leq t}\big|f_{t_{i+1}}-f_{t_i}\big|^p,\quad t\in[0,1].$$

The function $f$ is said to be of finite $p$-variation on $[0,1]$ if the limit is finite

$$\bigvee^p f(t):=\sup_{\Pi^n,n\in\mathbb{Z}}\sum_{t_{i+1}\leq t}^n\big|f_{t_{i+1}}-f_{t_i}\big|^p,\quad t\in[0,1].$$

THEOREM 4.8. *The quadratic variation of the Wiener process trajectories equals $t$ in the sense, that*

$$\bigvee^2 W(t) = \lim_{|\Pi^n|\to 0} \bigvee^2_{\Pi^n} W(t) = t,$$

*where*[1] *the limit in* $\mathbb{L}^2$ *is understood*[2].

PROOF. Use the Gaussian properties of the Wiener process

$$\mathrm{E}\Big(\sum_{t_{i+1}\le t}\big(W_{t_{i+1}}-W_{t_i}\big)^2 - t\Big)^2 = \mathrm{E}\Big(\sum_{t_{i+1}\le t}\big(W_{t_{i+1}}-W_{t_i}\big)^2 - (t_{i+1}-t_i)\Big)^2 =$$

$$\sum_{t_{i+1}\le t}\mathrm{E}\Big(\big(W_{t_{i+1}}-W_{t_i}\big)^2 - (t_{i+1}-t_i)\Big)^2 = \sum_{t_{i+1}\le t} 2(t_{i+1}-t_i)^2 \le 2\big|\Pi^n\big|t \xrightarrow{n\to\infty} 0.$$

□

THEOREM 4.9. *The Wiener process has trajectories with infinite variation, in particular*

$$\mathrm{P}\Big(\lim_{n\to 0}\sum_{0\le i\le n}\big|W_{i/n}-W_{(i-1)/n}\big| = \infty\Big) = 1.$$

PROOF. The random variables $\big(W_{i/n}-W_{(i-1)/n}\big)\sqrt{n}$ form an i.i.d. standard Gaussian sequence, so that by the law of large numbers

$$\mathrm{P}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \big|W_{i/n}-W_{(i-1)/n}\big|\sqrt{n} = \mathrm{E}|W_1|\right) = 1.$$

Since $\mathrm{E}|W_1| > 0$, this implies in particular

$$\mathrm{P}\left(\sum_{i=1}^n \big|W_{i/n}-W_{(i-1)/n}\big| \ge n^{1/2-\varepsilon}, \text{ eventually}\right) = 1.$$

for any $\varepsilon > 0$. □

## 2. The Itô Stochastic Integral

Recall the following fact from the classical analysis Vol.3, Ch. 15, §4-5 in [**9**].

THEOREM 4.10. *(Stieltjes integral) Let*[3] $f : [0,1] \mapsto \mathbb{R}$ *be a uniformly continuous function and* $g_t : [0,1] \mapsto \mathbb{R}$ *be a function of finite variation. Let* $0 = t_0 < t_1 < ... < t_n = 1$ *be a sequence of partitions and denote* $\delta^n = \max_j |t_j - t_{j-1}|$. *Then the limit*

$$\int_0^1 f_s dg_s := \lim_{\delta^n\to 0}\sum_{j=1}^n f(t^*_{j-1})\big(g_{t_j}-g_{t_{j-1}}\big) \tag{4.4}$$

*exists and is unique for any choice of points* $t^*_{j-1} \in [t_{j-1}, t_j]$, $j = 1, ..., n$. *It is called the Stieltjes integral of* $f_t$ *with respect to* $g_t$.

[1] $|\Pi^n| = \max_{0\le i\le n+1}|t_{i+1}-t_i|$ is the size of the partition.

[2] Stronger convergence is possible if the partitions sizes are allowed to decrease fast enough.

[3] For the sake of notation simplicity, the dependence of the partition $\{t_j\}$ on $n$ is always assumed implicitly.

PROOF. Assume first that $g$ does not decrease. Define the Darboux sums

$$s^n = \sum_{j=1} m_{j-1}\big(g_{t_j} - g_{t_{j-1}}\big), \quad S^n = \sum_{j=1} M_{j-1}\big(g_{t_j} - g_{t_{j-1}}\big)$$

where $m_{j-1} = \min_{s\in[t_{j-1},t_j]} f_s$ and $M_{j-1} = \max_{s\in[t_{j-1},t_j]} f_s$. It is easy to see that $S^n$ ($s^n$) does not increase (decrease) with $n$ and moreover $S^n \geq s^m$ for any $m, n \geq 1$. Then the limit in (4.4) exists and is unique if $I^* := \inf_n S^n = \sup_n s^n =: I_*$. The latter holds if

$$\lim_{\delta^n\to\infty} \sum_{j=1}^n (M_{j-1} - m_{j-1})\big(g_{t_j} - g_{t_{j-1}}\big) = 0.$$

If $f$ is uniformly continuous, then for any $\varepsilon > 0$, one may choose $\delta^n > 0$ such that $M_j - m_j \leq \varepsilon/(g_1 - g_0)$ uniformly in $j$. Then

$$\sum_{j=1}^n (M_{j-1} - m_{j-1})\big(g_{t_j} - g_{t_{j-1}}\big) \leq \varepsilon,$$

and the claim of the Theorem holds for nondecreasing $g$. The general case follows from the fact that $g$ with finite variation can be decomposed into sum of a nonincreasing and nondecreasing functions. □

The Wiener process has infinite variation and hence it is not clear how Stieltjes integral with respect to its trajectories can be constructed. This is clarified in the following example:

EXAMPLE 4.11. Suppose we would like to define the integral $\int_0^t W_s dW_s$ as the limit $n \to \infty$ of the sums

$$\sum_{i=0}^{[tn]} W_{s_i^*}\big(W_{s_{i+1}} - W_{s_i}\big), \quad t \in [0, 1]$$

where $s_i = i/n$ and $s_i^*$ is a point from interval $[s_{i-1}, s_i]$ for each $i$. Consider the two choices: $s_i^* = s_i$ and $s_i^* = (s_{i+1} + s_i)/2$, which lead to

$$I_t^n = \sum_{i=0}^{[tn]} W_{s_i}\big(W_{s_{i+1}} - W_{s_i}\big)$$

and

$$J_t^n = \sum_{i=0}^{[tn]} W_{(s_i+s_{i+1})/2}\big(W_{s_{i+1}} - W_{s_i}\big)$$

respectively. Clearly $\mathrm{E}I_t^n = 0$ for all $t$ and $n \geq 1$. On the other hand

$$\mathrm{E}J_t^n = \sum_{i=0}^{[tn]} \mathrm{E}W_{(s_i+s_{i+1})/2}\big(W_{s_{i+1}} - W_{s_i}\big) =$$

$$\sum_{i=0}^{[tn]} \big((s_i + s_{i+1})/2 - s_i\big) = \frac{1}{2}[tn]/n \xrightarrow{n\to\infty} t/2.$$

It is not hard to see that the limits in probability $I_t = \lim_{n\to\infty} I_t^n$ and $J_t = \lim_{n\to\infty} J_t^n$ exist and satisfy $\mathrm{E}I_t = 0$ and $\mathrm{E}J_t = t/2$ for all $t \in [0, 1]$. So one does not obtain the same limit for different partitions as promised in Theorem 4.10. This is a manifestation of the trajectories irregularity of $W$: if their variation were finite the same limit would be obtained! Let us note that both examples are in fact the prototypes of the stochastic integrals in the sense of Itô and Stratonovich respectively. See Exercise **7** for further exploration. ∎

**2.1. Construction.** The Itô integral will be defined in this course[4] under the following setup. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a complete[5] probability space with the increasing family of sub-$\sigma$-algebras (filtration) $\mathcal{F}_t \subseteq \mathcal{F}$. Sometimes $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge 0}, \mathrm{P})$ is referred as *filtered* probability space or *stochastic basis.*

DEFINITION 4.12. A random process $X$ is said to be adapted to filtration $\mathcal{F}_t$ if for each fixed $t \ge 0$, the random variable $X_t$ is $\mathcal{F}_t$-measurable.

From here on all the random processes are assumed to be adapted to $\mathcal{F}_t$, if not stated otherwise. For example the Wiener process $W_t$ is trivially adapted to its natural filtration $\mathcal{F}^W_t = \sigma\{W_s, s \le t\}$, but is also assumed to be adapted to $\mathcal{F}_t$. This allows to define the integral more generally and is of no limitation, since $\mathcal{F}_t$ can be usually defined to be the least filtration, to which all the processes are adapted. For example it allows to define integrals like $\int_0^t V_t dW_t$ where $W$ and $V$ are independent Wiener processes: $V$ is not adapted to $\mathcal{F}^W_t$, but both $W$ and $V$ are adapted to $\mathcal{F}_t := \mathcal{F}^V_t \vee \mathcal{F}^W_t$.

Construction of the Itô integral is based on two main ideas: (1) to restrict the choice of the sampling points of the integrand in the prelimit sums to the beginning of the sub-intervals of the partition and (2) to consider integrands for which this restriction leads to the unique limit.

DEFINITION 4.13. The process $X_t(\omega)$ is said to belong to the family $\mathcal{H}^2_{[0,T]}$ if

(1) the mapping $(t, \omega) \mapsto X_t(\omega)$ is measurable with respect to $\mathcal{B}([0, T]) \times \mathcal{F}$ (as a function of both arguments)
(2) $X_t(\omega)$ is $\mathcal{F}_t$ adapted
(3) $\mathrm{E} \int_0^T X^2_s(\omega) ds < \infty$

REMARK 4.14. The stochastic integral can be constructed for a more general class of integrands, satisfying only

$$\mathrm{P}\left(\int_0^T X_t^2 dt < \infty\right) = 1,$$

instead of (3). In what follows the stochastic integral will be used with the integrands satisfying the stronger condition, if not specified otherwise. It turns out that the properties of the stochastic integral may crucially depend on the integrand type - this point is demonstrated in Example 4.25 below.

Generally stochastic integration can be defined with respect to processes, more general than the Wiener process: the *martingales*. For further exploration see the introductory text [**4**] and [**22**] for a more advanced treatment.

DEFINITION 4.15. The process $X_t$ is $\mathcal{H}^2_{[0,T]}$-*simple* (or just simple) if it belongs to $\mathcal{H}^2_{[0,T]}$ and has the form $X^n_t = \sum_{j=1}^n \xi_{j-1} \mathbf{1}_{[t_{j-1},t_j)}$ for some fixed partition $0 = t_0 \le t_1 \le ... \le t_n = T$ and random variables $\xi_j$.

[4]The text [**25**] is followed here.

[5]standard technical requirement which is usually imposed on probability spaces: it means that $\mathcal{F}$ contains all the sets $A$, such that $\underline{A} \subseteq A \subseteq \overline{A}$ for some measurable sets $\overline{A}$ and $\underline{A}$ (on which P is defined) with $\mathrm{P}(\underline{A}) = \mathrm{P}(\overline{A})$. Then $\mathrm{P}(A) = 0$ is set.

Assume that $\mathcal{F}_t^W \subseteq \mathcal{F}_t$ and define the Itô integral for a simple process $X_t^n$ as

$$I(X^n) := \int_0^T X_t^n dW_t := \sum_{j=1}^n \xi_{j-1}\big(W_{t_j} - W_{t_{j-1}}\big).$$

Then[6]

$$\begin{aligned}
\mathrm{E}I^2(X^n) =& \mathrm{E}\Big(\sum_{j=1}^n \xi_{j-1}\big(W_{t_j} - W_{t_{j-1}}\big)\Big)^2 = \\
& \sum_{j=1}^n \mathrm{E}\xi_{j-1}^2\big(W_{t_j} - W_{t_{j-1}}\big)^2 + \\
& 2\sum_{i=1}^{n-1}\sum_{j<i} \mathrm{E}\xi_{i-1}\xi_{j-1}\big(W_{t_j} - W_{t_{j-1}}\big)\big(W_{t_i} - W_{t_{i-1}}\big) = \\
& \sum_{j=1}^n \mathrm{E}\xi_{j-1}^2 \mathrm{E}\Big(\big(W_{t_j} - W_{t_{j-1}}\big)^2 \big| \mathcal{F}_{t_{j-1}}\Big) + \\
& 2\sum_{i=1}^{n-1}\sum_{j<i} \mathrm{E}\xi_{i-1}\xi_{j-1}\big(W_{t_j} - W_{t_{j-1}}\big)\Big(\mathrm{E}\big(W_{t_i} - W_{t_{i-1}}\big) \big| \mathcal{F}_{t_{i-1}}\Big) = \\
& \sum_{j=1}^n \mathrm{E}\xi_{j-1}^2(t_j - t_{j-1}) = \int_0^T \mathrm{E}\big(X_t^n\big)^2 dt.
\end{aligned} \tag{4.5}$$

The latter property is called the *Itô isometry* and is the main feature in the construction of the stochastic integral.

LEMMA 4.16. *Let $\{t_j\}$ be a sequence of partitions on $[0,T]$, such that $\delta^n = \max_j |t_j - t_{j-1}| \to 0$ , as $n \to 0$. Then*

**1.** *for any continuous*[7] *and bounded $\mathcal{H}^2_{[0,T]}$ process $X_t^{bc}$, there is a sequence of simple $\mathcal{H}^2_{[0,T]}$ processes $X_t^\ell$, $\ell \ge 0$, such that*

$$\lim_{\ell\to\infty} \int_0^T \mathrm{E}\big(X_t^{bc} - X_t^\ell\big)^2 dt = 0 \tag{4.6}$$

**2.** *for any bounded $\mathcal{H}^2_{[0,T]}$ process $X_t^b$ there is a sequence of continuous $\mathcal{H}^2_{[0,T]}$ processes $X_t^{c,m}$, $m \ge 1$, such that*

$$\lim_{m\to\infty} \int_0^T \mathrm{E}\big(X_t^b - X_t^{c,m}\big)^2 dt = 0 \tag{4.7}$$

**3.** *for any $\mathcal{H}^2_{[0,T]}$ process $X_t$ there is a sequence of bounded $\mathcal{H}^2_{[0,T]}$ processes $X_t^{b,n}$, $n \ge 1$ such that*

$$\lim_{n\to\infty} \int_0^T \mathrm{E}\big(X_t - X_t^{b,n}\big)^2 dt = 0. \tag{4.8}$$

---

[6] It can be shown that the filtration $\mathcal{F}_t^W$ is continuous, i.e. $\mathcal{F}_{t+}^W := \bigcap_{\varepsilon>0}\mathcal{F}_{t+\varepsilon}^W$ and $\mathcal{F}_{t-}^W := \bigvee_{\varepsilon>0}\mathcal{F}_{t-\varepsilon}^W$ coincide. It is customary to assume that $\mathcal{F}_t$ is continuous (or at least right continuous) as well. This and the definition of $X_t^n$ implies that $\xi_{j-1}$ is $\mathcal{F}_{t_{j-1}}$-measurable.

[7] i.e. a process with continuous trajectories

PROOF. **1.** Let $X^\ell_t = \sum_{t_j \le t} X^{bc}_{t_{j-1}} \mathbf{1}_{[t_{j-1},t_j)}$. Clearly $X^\ell_t$ is a simple bounded $\mathcal{H}^2_{[0,T]}$ process, which converges to $X^{bc}_t$ uniformly in $t$ due to its continuity. Then (4.6) follows by dominated convergence.

**2.** Let $\psi^m_t$, $m \ge 1$ be a sequence of continuous functions supported on $(-n^{-1}, 0)$ and satisfying $\int_{\mathbb{R}} \psi^n_s ds = 1$. Define

$$X^{c,m}_t = \int_0^t X^b_s \psi^m_{s-t} ds.$$

Clearly $X^{c,m}_t$ are continuous $\mathcal{H}^2_{[0,T]}$ processes (since $\psi^m_t$ was chosen in a "casual" way) and

$$\lim_{m\to\infty} \int_0^T \mathrm{E}\big(X^b_t - X^{c,m}_t\big)^2 dt = 0, \quad \mathrm{P}-a.s.$$

since convolution with $\psi^m_s$ approximates the identity operator for bounded functions. Again (4.7) follows by dominated convergence.

**3.** Fix an integer $n \ge 1$ and define

$$X^{b,n}_t = \begin{cases} X_t & |X_t| \le n \\ \mathrm{sign}(X_t) n & |X_t| > n \end{cases}.$$

Clearly $|X^{b,n}_t| \le |X_t|$ and so

$$\int_0^T \mathrm{E}\big(X^{b,n}_t - X_t\big)^2 dt \le 2 \int_0^T \mathrm{E} X^2_t dt < \infty$$

and hence (4.8) follows by dominated convergence. □

THEOREM 4.17. *(Itô stochastic integral) For any $X_t \in \mathcal{H}^2_{[0,T]}$, the $\mathbb{L}^2$-limit*

$$\int_0^T X_s dW_s := \lim_{\delta^n \to \infty} \int_0^T X^n_s dW_s$$

*exists and is independent of the specific sequence $X^n$ of simple processes, approximating $X$ in the sense*

$$\int_0^T \mathrm{E}(X_s - X^n_s)^2 ds \xrightarrow{n\to\infty} 0.$$

PROOF. By Lemma 4.16 for any $\mathcal{H}^2_{[0,T]}$-process $X_t$ there is a sequence of simple processes $X^n_t$ for which $I(X^n)$ is well defined. Note that for any $n, m$, $X^n_t - X^m_t$ is a simple $\mathcal{H}^2_{[0,T]}$-process. Then sequence $I(X^n)$, $n \ge 1$ satisfies the Cauchy property

$$\mathrm{E}\big(I(X^n) - I(X^m)\big)^2 = \mathrm{E}\Big(\int_0^T (X^n_t - X^m_t) dW_t\Big)^2 = \int_0^T \mathrm{E}\big(X^n_t - X^m_t\big)^2 dt \xrightarrow{n,m\to\infty} 0,$$

where the latter holds since $X^n$ is a convergent sequence in[8] $\mathbb{L}^2$. The existence of the limit $I(X) = \lim_{n\to\infty} I(X^n)$ follows since any Cauchy sequence converges in $\mathbb{L}^2$.

The uniqueness is obtained by standard arguments. Let $X^{(1)}_n$ and $X^{(2)}_n$ be two approximating sequences and let $X^n$ denote the sequence obtained by taking $X^{(1)}_n$ for odd $n$ and taking $X^{(2)}_n$ for even $n$. Suppose that different limits $I_1(X)$ and $I_2(X)$ are obtained when using $X^{(1)}_n$ and $X^{(1)}_n$. Then the approximating sequence

[8] $\mathbb{L}^2\big(\Omega \times [0,T], \mathcal{F} \times \mathcal{B}, \mathrm{P} \times \lambda\big)$ is meant here

$X^n$ will not converge to any limit. This however contradicts the existence of a limit for $X^n$. □

REMARK 4.18. Calculation of the Itô integral is possible by applying the construction used in its definition - see Exercise **8**. Another way is to apply the Itô formula to be given below.

**2.2. Properties.** Let $X$ and $Y$ be $\mathcal{H}^2_{[0,T]}$-processes, then (all "random" equalities hold P-a.s.)

(i) $\int_0^T X_t dW_t = \int_0^S X_t dW_t + \int_S^T X_t dW_t$, $S \le T$
(ii) $\int_0^T (aX_t + bY_t) dW_t = a\int_0^T X_t dW_t + b\int_0^T Y_t dW_t$, for constants $a$ and $b$
(iii) $\mathrm{E}\int_0^T X_t dW_t = 0$
(iv) $\mathrm{E}\Big(\int_0^T X_t dW_t \int_0^S Y_t dW_t) = \int_0^{S\wedge T} \mathrm{E}X_t Y_t dt$. In particular

$$\mathrm{E}\Big(\int_0^T X_t dW_t\Big)^2 = \int_0^T \mathrm{E}X_t^2 dt.$$

(v) $\int_0^t X_s dW_s$ is $\mathcal{F}_t$-adapted
(vi) $\int_0^t X_s dW_s$, $t \in [0,T]$ admits a continuous version[9], i.e. there exists a random process $I_t(X)$, $t \in [0,T]$ with continuous trajectories, so that

$$\mathrm{P}\Big(\int_0^t X_s dW_s = I_t(X)\Big) = 1, \quad \forall t \in [0,T].$$

PROOF. The properties (i)-(v) are inherited from the simple functions approximation. Let's verify, say (i): take a sequence $X^n \to X$, in the sense

$$\int_0^T \mathrm{E}(X_t^n - X_t)^2 dt \to 0.$$

Then

$$\int_0^T X_t^n dW_t = \int_0^S X_t^n dW_t + \int_S^T X_t^n dW_t$$

and so

$$\begin{aligned}
&\mathrm{E}\Big(\int_0^T X_t dW_t - \int_0^S X_t dW_t - \int_S^T X_t dW_t\Big)^2 \le \\
&4\mathrm{E}\Big(\int_0^T X_t dW_t - \int_0^T X_t^n dW_t\Big)^2 + 4\mathrm{E}\Big(\int_0^S X_t dW_t - \int_0^S X_t^n dW_t\Big)^2 + \\
&4\mathrm{E}\Big(\int_S^T X_t dW_t - \int_S^T X_t^n dW_t\Big)^2 \xrightarrow{n\to\infty} 0.
\end{aligned}$$

---

[9]Several types of equalities between continuous time random process are usually considered. The processes $X$ and $Y$ are said to be indistinguishable if

$$\mathrm{P}\Big(\exists t \in [0,T] : X_t \ne Y_t\Big) = \mathrm{P}\Big(\sup_{t\le T} |X_t - Y_t| > 0\Big) = 0.$$

This is the strongest kind of equality, which is sometimes hard to establish. $X$ is said to be a version of $Y$ if for any $t \in [0,T]$

$$\mathrm{P}(X_t \ne Y_t) = 0 \tag{4.9}$$

Clearly indistinguishable processes are versions of each other. Note that if $X$ and $Y$ satisfy (4.9), then their finite dimensional distributions coincide.

The property (vi) stems from continuity of $W$. It's proof relies on the fact that $\int_0^t X_t^n dW_t$ is continuous for a fixed $n \geq 1$ and that this sequence converges uniformly in $t$, making the limit a continuous function of $t$ as well (the proof uses Doob's inequality for martingales). □

REMARK 4.19. If the assumption

$$\int_0^T \mathrm{E}X_t^2 dt < \infty$$

is replaced by

$$\mathrm{P}\left(\int_0^T X_t^2 dt < \infty\right) = 1,$$

the integral is still well defined (as mentioned before in Remark 4.14), however the properties (iii) and (iv) may fail to hold (!) - see Example 4.25 below.

## 3. The Itô formula

Consider the scalar random process

$$X_t = X_0 + \int_0^t a_s(\omega)ds + \int_0^t b_s(\omega)dW_s, \quad t \leq T, \tag{4.10}$$

where $a_t$ and $b_t$ are $\mathcal{H}^2_{[0,T]}$ processes and $W = (W_t)_{t\leq T}$ is the Wiener process, defined on a stochastic basis $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathrm{P})$. A random process is an *Itô process*, if it satisfies (4.10), which is usually written in a "differential" form

$$dX_t = a_t(\omega)dt + b_t(\omega)dW_t. \tag{4.11}$$

Note that this Itô differential is nothing more than a brief notation in the spirit of classical calculus.

Let $f(t,x)$ be a $\mathbb{R}_+ \times \mathbb{R} \mapsto \mathbb{R}$ function with one and two continuous derivatives in $t$ and $x$ respectively. It turns out that the process $\xi_t := f(t, X_t)$ admits unique integral representation, similar to (4.10), or in other words, it is also an Itô process.

THEOREM 4.20. *(the Itô formula) Assume $f$ and its partial derivatives with respect to $t$ and $x$ variables $f'_t$, $f'_x$ and $f''_x$ are bounded and continuous, then the process $\xi_t = f(t, X_t)$ admits the Itô differential*

$$d\xi_t = f'_t(t,X_t)dt + f'_x(t,X_t)a_t dt + \frac{1}{2}f''_x(t,X_t)b_t^2 dt + f'_x(t,X_t)b_t dW_t, \tag{4.12}$$

*subject to $\xi_0 = f(0, X_0)$.*

REMARK 4.21. Consider the similar setting in the classical nonrandom case: let $V_t$ be a function of bounded variation and $dX_t = a_t dt + b_t dV_t$, where the latter is the Stieltjes differential. Then the differential for $\xi_t = f(t, X_t)$ is obtained by the well known chain rule

$$d\xi_t = f'_t(t,X_t)dt + f'_x(t,X_t)dX_t = f'_t(t,X_t)dt + f'_x(t,X_t)a_t dt + f'_x(t,X_t)b_t dV_t.$$

The major difference between the classic differentiation and (4.12) is the extra term $\frac{1}{2}f''_x(t,X_t)b_t^2 dt$, which is again the manifestation of trajectories irregularity of $W$.

This non-classic chain rule is called Itô formula and is the central tool of stochastic calculus with respect to Wiener process.

REMARK 4.22. The requirements for $f$ and its derivatives to be bounded can be relaxed even if working under the condition, mentioned in Remark 4.14. Moreover the second derivative in $x$ can be discontinuous at a countable number of points. One should be careful to make further relaxations: for example if the first derivative has a discontinuity point, the local time process arises - see Example 4.26.

REMARK 4.23. The Itô formula remains valid under condition mentioned in Remark 4.14 (recall that the stochastic integral itself may have different properties depending on the integrability conditions of the integrand - see Remark 4.19).

PROOF. (Sketch) Let $a_t^n(\omega)$ and $b_t^n(\omega)$ be simple $\mathcal{H}^2_{[0,T]}$ processes, approximating $a_t$ and $b_t$:

$$\int_0^T \mathrm{E}\big|a_t^n - a_t\big|dt \xrightarrow{n\to\infty} 0$$

$$\int_0^T \mathrm{E}\big(b_t^n - b_t\big)^2 dt \xrightarrow{n\to\infty} 0,$$

Let $X_t^n = X_0 + \int_0^t a_s^n ds + \int_0^t b_s^n dW_s$ and suppose that (4.12) holds for $\xi^n := f(t, X_t^n)$. Then (4.12) holds for $\xi_t$ by continuity and boundedness of $f$ and its derivatives:

$$\begin{aligned}
&\mathrm{E}\Big|f(t,X_t) - f(0,X_0) - \\
&\qquad \int_0^t \big(f_t'(s,X_s) + f_x'(s,X_s)a_s ds + \frac{1}{2}f_x''(s,X_s)b_s^2\big)ds - \int_0^t f_x'(s,X_s)b_s dW_s\Big| \le \\
&\mathrm{E}\Big|f(t,X_t) - f(t,X_t^n)\Big| + \int_0^T \mathrm{E}\Big|f_t'(s,X_s) - f_t'(s,X_s^n)\Big|ds + \\
&\int_0^T \mathrm{E}\Big|f_x'(s,X_s) - f_x'(s,X_s^n)\Big|a_s ds + \int_0^T \frac{1}{2}\mathrm{E}\Big|f_x''(s,X_s) - f_x''(s,X_s^n)\Big|b_s^2 ds + \\
&\Big(\int_0^T \mathrm{E}\big(f_x'(s,X_s) - f_x'(s,X_s^n)\big)^2 b_s^2 ds\Big)^{1/2} \xrightarrow{n\to 0} 0.
\end{aligned}$$

So it is enough to verify (4.12), when $a_t$ and $b_t$ are simple. Due to additivity of the stochastic integral, it even suffices to consider constant $a(\omega)$ and $b(\omega)$ (such that the Itô integral is well defined), in which case $X_t = at + bW_t$. Since $f(t, at + bW_t)$ is now a function of $t$ and $W_t$, the formula (4.12) holds, if

$$u(t,W_t) = u(0,0) + \int_0^t u_t'(s,W_s)ds + \int_0^t u_x'(s,W_s)dW_s + \frac{1}{2}\int_0^t u_x''(s,W_s)ds \quad (4.13)$$

for a bounded $u(t,x)$ with two bounded continuous derivatives. Using the Taylor expansion for $u(t,x)$, the telescopic sum is obtained (with $\Delta t_i := t_i - t_{i-1}$ and $\Delta W_i = W_{t_i} - W_{t_{i-1}}$)

$$\begin{aligned}
u(t,W_t) =& u(0,0) + \sum_{i=1}^n u_t'(t_{i-1},W_{t_{i-1}})\Delta t_i + \sum_{i=1}^n u_x'(t_{i-1},W_{t_{i-1}})\Delta W_i + \\
&\frac{1}{2}\sum_{i=1}^n u_x''(t_{i-1},W_{t_{i-1}})(\Delta W_i)^2 + R^n
\end{aligned}$$

where $R^n$ is the residual term, consisting of sums over $(\Delta t_i)^2$, $\Delta t_i \Delta W_i$ and $(\Delta W_i)^3$ with coefficients obtained by the Mean Value Theorem. Clearly the first three terms on the right hand side of the latter converge to the corresponding terms in (4.13). By the same arguments, used in the proof of Theorem 4.8

$$\mathrm{E}\Big(\sum_{i=1}^n u''_x(t_{i-1}, W_{t_{i-1}})(\Delta W_i)^2 - \sum_{i=1}^n u''_x(t_{i-1}, W_{t_{i-1}})\Delta t_i\Big)^2 =$$

$$\sum_{i=1}^n \mathrm{E}\big(u''_x(t_{i-1}, W_{t_{i-1}})\big)^2\big((\Delta W_i)^2 - \Delta t_i\big)^2 \le$$

$$2T \sup_{t,x\in[0,T]\times\mathbb{R}} |u''_x(t,x)|^2 \max_i \Delta t_i \xrightarrow{n\to\infty} 0$$

Similarly the residual term $R^n$ is shown to vanish as $n \to \infty$. □

EXAMPLE 4.24. Apply the Itô formula to $W_t^2$:

$$d(W_t)^2 = 2W_t dW_t + dt$$

or in other words

$$W_t^2 = 2\int_0^t W_s dW_s + t.$$

■

EXAMPLE 4.25. (Example 8 Ch. 6.2 in [**21**]) Let $\beta_t$ be a random process, adapted to $\mathcal{F}_t$ and satisfying

$$\mathrm{P}\left(\int_0^1 \beta_t^2 dt < \infty\right) = 1. \tag{4.14}$$

Then the process

$$\varphi_t = \exp\left(\int_0^t \beta_s dW_s - \frac{1}{2}\int_0^t \beta_s^2 ds\right)$$

is well defined and by the Itô formula, satisfies the integral identity (which is also an example of stochastic differential equation (SDE) to be introduced in Section 5)

$$\varphi_t = 1 + \int_0^t \varphi_s \beta_s dW_s, \quad t \in [0,1].$$

If $\int_0^1 \mathrm{E}\beta_s^2 ds < \infty$, then the stochastic integral has zero mean and thus $\mathrm{E}\varphi_1 = 1$. If however only (4.14) holds, then $\mathrm{E}\varphi_1 < 1$ is possible, meaning that the stochastic integral is no longer a martingale. Consider a specific $\beta_t$

$$\beta_t = -\frac{2W_t}{(1-t)^2}\mathbf{1}_{\{t\le\tau\}},$$

where $\tau = \inf\{t \le 1 : W_t^2 = 1-t\}$, i.e. the first time $W_t^2$ hits the line $1-t$. The event $\{\tau \le t\}$ is $\mathcal{F}_t^W$ measurable (and a fortiori $\mathcal{F}_t$ measurable), since it can be resolved on the basis of trajectory of $W$ up to time $t$ and hence $\beta_t$ is $\mathcal{F}_t$-adapted. Note that $\mathrm{P}(\tau < 1) = 1$, since

$$\mathrm{P}(\tau = 1) \le \mathrm{P}(W_1 = 0) = 0,$$

and so

$$\int_0^1 \beta_t^2 dt = \int_0^1 \frac{4W_t^2}{(1-t)^4}\mathbf{1}_{\{t\le\tau\}} dt = \int_0^\tau \frac{4W_t^2}{(1-t)^4} dt < \infty, \quad \mathrm{P}-a.s.$$

By the Itô formula

$$d\left(\frac{W_t^2}{(1-t)^2}\right) = \frac{2W_t^2}{(1-t)^3}dt + \frac{2W_t}{(1-t)^2}dW_t + \frac{1}{(1-t)^2}dt,$$

which implies

$$\int_0^1 \beta_s dW_s - \frac{1}{2}\int_0^1 \beta_s^2 ds = -\int_0^\tau \frac{2W_t}{(1-t)^2}dW_s - \int_0^\tau \frac{2W_t^2}{(1-t)^4}dt =$$
$$-\frac{W_\tau^2}{(1-\tau)^2} + \int_0^\tau \frac{2W_t^2}{(1-t)^3}dt + \int_0^\tau \frac{1}{(1-t)^2}dt - \int_0^\tau \frac{2W_t^2}{(1-t)^4}dt =$$
$$-\frac{1}{(1-\tau)^2} + \int_0^\tau 2W_t^2\left(\frac{1}{(1-t)^3} - \frac{1}{(1-t)^4}\right)dt + \int_0^\tau \frac{1}{(1-t)^2}dt \le$$
$$-\frac{1}{1-\tau} + \int_0^\tau \frac{1}{(1-t)^2}dt = -1.$$

Then $\mathrm{E}\varphi_t \le 1/e < 1$, i.e. the stochastic integral $\int_0^t \varphi_s\beta_s dW_s$ has nonzero mean! ■

Example 4.26. *(The Tanaka formula and the local time)* Let $\varepsilon > 0$ and

$$f_\varepsilon(x) = |x|\mathbf{1}_{\{|x|\ge\varepsilon\}} + \frac{1}{2}(\varepsilon + \frac{x^2}{\varepsilon})\mathbf{1}_{\{|x|<\varepsilon\}}.$$

Since $f_\varepsilon(x)$ is twice differentiable with the second derivative discontinuous at two points $x = \pm\varepsilon$, the Itô formula still applies and gives

$$f_\varepsilon(W_t) = \int_0^t f'_\varepsilon(W_s)dW_s + \frac{1}{2}\int_0^t f''_\varepsilon(W_s)ds =$$
$$\int_0^t \mathrm{sign}(W_s)\mathbf{1}_{\{|W_s|\ge\varepsilon\}}dW_s + \int_0^t \varepsilon^{-1}W_s\mathbf{1}_{\{|W_s|<\varepsilon\}}dW_s +$$
$$\frac{1}{2\varepsilon}\int_0^t \mathbf{1}_{\{|W_s|\le\varepsilon\}}ds$$

Note that

$$\mathrm{E}\left(\int_0^t \varepsilon^{-1}W_s\mathbf{1}_{\{|W_s|<\varepsilon\}}dW_s\right)^2 = \int_0^t \varepsilon^{-2}\mathrm{E}W_s^2\mathbf{1}_{\{|W_s|<\varepsilon\}}ds \le$$
$$\int_0^t \varepsilon^{-2}\varepsilon^2\mathrm{E}\mathbf{1}_{\{|W_s|<\varepsilon\}}ds = \int_0^t \mathrm{P}(|W_s|<\varepsilon)ds \xrightarrow{\varepsilon\to 0} 0.$$

Hence the *local time* process corresponding to $W_t$

$$L_t = \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\int_0^t \mathbf{1}_{\{|W_s|\le\varepsilon\}}ds \tag{4.15}$$

exists at least as $\mathbb{L}^2$ limit. In fact it exists in a stronger sense and moreover the Tanaka formula holds

$$|W_t| = \int_0^t \mathrm{sign}(W_t)dW_t + L_t, \tag{4.16}$$

as the preceding limit procedure hints ($f_\varepsilon(x) \to |x|$ for all $x$). By definition $L_t$ is the rate at which the amount of time spent by the Wiener process in the vicinity of zero decays as it shrinks. This is another manifestation of pathes irregularity of the Wiener process: e.g. the limit (4.15) would vanish if $W_t$ had a countable number of zeros on $[0, T]$. ■

More examples are collected in the Exercises section. The following Theorem gives the multivariate version of the Itô formula

THEOREM 4.27. *Let $X_t$ have the Itô differential*

$$dX_t = a_t dt + b_t dW_t, \quad t \in [0,T],$$

*where $a_t$ and $b_t$ are $n \times 1$ vector and $n \times m$ matrix of $\mathcal{H}^2_{[0,T]}$-random processes and $W_t$ is a vector of $m$ independent Wiener processes. Assume $f : \mathbb{R}_+ \times \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable in $t$ variable and twice continuously differentiable in the $x$ variables. Then*

$$\begin{aligned} df(t,X_t) = \frac{\partial}{\partial t} f(t,X_t)dt + \sum_{i=1}^d \frac{\partial}{\partial x_i} f(t,X_t) dX_t + \\ \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} f(t,X_t) \sum_{k=1}^n b_t(i,k) b_t(j,k) dt. \end{aligned} \tag{4.17}$$

REMARK 4.28. Denote by $\nabla$ the (row vector) gradient operator with respect to $x$ and let $\nabla b_t b_t^* \nabla^*$ be the second order differential operator, obtained by formal multiplication of partial derivatives. Denote by $\dot{f}(t,x)$ the partial derivative w.r.t. time variable $t$. Then (4.17) can be compactly written as

$$df(t,X_t) = \dot{f}(t,X_t)dt + \nabla f(t,X_t) dX_t + \frac{1}{2}(\nabla b_t b_t^* \nabla^*) f(t,X_t) dt.$$

The vector Itô formula can be conveniently encoded into the mnemonic multiplication rules between differentials, summarized in Table 4.28, used with formal Taylor expansion of $f$ as demonstrated in the following example.

| $\times$ | $1$ | $dt$ | $dW_t(1)$ | $dW_t(2)$ |
|---|---|---|---|---|
| $dt$ | $dt$ | $0$ | $0$ | $0$ |
| $dW_t(1)$ | $dW_t(1)$ | $0$ | $dt$ | $0$ |
| $dW_t(2)$ | $dW_t(2)$ | $0$ | $0$ | $dt$ |

TABLE 1. The formal Itô differential multiplication rules

EXAMPLE 4.29. Consider the two dimensional system

$$\begin{aligned} dX_t &= a_1 X_t dt + b_{11} dW_t + b_{12} dV_t \\ dY_t &= a_2 Y_t dt + b_{21} dW_t + b_{22} dV_t. \end{aligned}$$

and let $r_t = f(X_t, Y_t)$. Then formally

$$\begin{aligned} dr_t = & df(X_t,Y_t) = f_x(X_t,Y_t)dX_t + f_y(X_t,Y_t)dY_t + \\ & \frac{1}{2} f_{xx}(X_t,Y_t)(dX_t)^2 + f_{xy}(X_t,Y_t)dX_t dY_t + \frac{1}{2} f_{yy}(X_t,Y_t)(dY_t)^2. \end{aligned}$$

and using the rules from the table.

$$(dX_t)^2 = \big(a_1 X_t dt + b_{11} dW_t + b_{12} dV_t\big)^2 = b_{11}^2 dt + b_{12}^2 dt.$$

Proceeding similarly for the rest of terms, one gets

$$dr_t = f_x(X_t,Y_t)dX_t + f_y(X_t,Y_t)dY_t + \frac{1}{2}f_{xx}(X_t,Y_t)(b_{11}^2+b_{12}^2)dt+$$
$$f_{xy}(X_t,Y_t)(b_{11}b_{21}+b_{12}b_{22})dt + \frac{1}{2}f_{yy}(X_t,Y_t)(b_{21}^2+b_{22}^2)dt$$

Verify the answer by applying (4.17) directly. ■

## 4. The Girsanov theorem

The following theorem, proved by I.Girsanov, plays the crucial role in stochastic analysis and in filtering particularly

THEOREM 4.30. *Let $\beta_t$ be an $\mathcal{F}_t$-adapted process, defined on $(\Omega,\mathcal{F},\mathcal{F}_t,\mathrm{P})$ and satisfying*

$$\mathrm{P}\left(\int_0^T \beta_t^2 dt < \infty\right) = 1$$

*and let*

$$\varphi_t = \exp\left(\int_0^t \beta_s dW_s - \frac{1}{2}\int_0^t \beta_s^2 ds\right).$$

*Assume that $\mathrm{E}\varphi_T = 1$ and define the probability measure $\widetilde{\mathrm{P}}$ by*

$$\frac{d\widetilde{\mathrm{P}}}{d\mathrm{P}}(\omega) = \varphi_T(\omega).$$

*Then*

$$V_t = W_t - \int_0^t \beta_s ds, \quad t\in[0,T]$$

*is the Wiener process with respect to $\mathcal{F}_t$ under probability $\widetilde{\mathrm{P}}$.*

PROOF. (Sketch) Clearly $V_t$ has continuous pathes and starts at zero. Thus it is left to verify

$$\widetilde{\mathrm{E}}\big(\exp\{i\lambda(V_t-V_s)\}|\mathcal{F}_s\big) = \exp\big\{-0.5\lambda^2(t-s)\big\}, \quad t\ge s. \tag{4.18}$$

It turns out that the assumption $\mathrm{E}\varphi_T = 1$ implies $\mathrm{P}(\inf_{t\le T}\varphi_t = 0) = 0$ and hence also $\widetilde{\mathrm{P}}(\inf_{t\le T}\varphi_t = 0) = 0$. Then $\mathrm{P}\sim\widetilde{\mathrm{P}}$ and

$$\frac{d\mathrm{P}}{d\widetilde{\mathrm{P}}}(\omega) = \varphi_T^{-1}(\omega).$$

By Lemma 3.11

$$\widetilde{\mathrm{E}}\big(\exp\{i\lambda(V_t-V_s)\}|\mathcal{F}_s\big) = \frac{\mathrm{E}\big(\exp\{i\lambda(V_t-V_s)\}\varphi_T|\mathcal{F}_s\big)}{\mathrm{E}\big(\varphi_T|\mathcal{F}_s\big)} =$$
$$\exp\{-i\lambda V_s\}\frac{\mathrm{E}\big(\exp\{i\lambda V_t\}\varphi_T|\mathcal{F}_s\big)}{\mathrm{E}\big(\varphi_T|\mathcal{F}_s\big)}$$

Moreover under the assumption $\mathrm{E}\varphi_T = 1$, the process $\varphi_t$ is a martingale, i.e. it is $\mathcal{F}_t$-adapted and $\mathrm{E}(\varphi_t|\mathcal{F}_s) = \varphi_s$. Indeed by the Itô formula $\varphi_t$ satisfies

$$\varphi_t = \varphi_s + \int_s^t \varphi_r\beta_t dW_r \quad \Longrightarrow \quad \mathrm{E}(\varphi_t|\mathcal{F}_s) = \varphi_s,$$

where the (nontrivial!) fact $\mathrm{E}\Big(\int_s^t \varphi_r\beta_t dW_r|\mathcal{F}_s\Big) = 0$ has been used. Then

$$\widetilde{\mathrm{E}}\big(\exp\{i\lambda(V_t-V_s)\}|\mathcal{F}_s\big) = \frac{\mathrm{E}\big(\exp\{i\lambda V_t\}\varphi_t|\mathcal{F}_s\big)}{\exp\{i\lambda V_s\}\varphi_s}. \tag{4.19}$$

By the Itô formula the process $\zeta_t := \exp\{i\lambda V_t\}\varphi_t$ satisfies

$$d\zeta_t = i\lambda\zeta_t dV_t - \frac{1}{2}\lambda^2\zeta_t dt + \exp\{i\lambda V_t\}d\varphi_t + i\lambda\exp\{i\lambda V_t\}\varphi_t\beta_t dt =$$
$$i\lambda\zeta_t dW_t - i\lambda\zeta_t\beta_t dt - \frac{1}{2}\lambda^2\zeta_t dt + \zeta_t\beta_t dW_t + i\lambda\zeta_t\beta_t dt$$

which implies

$$\zeta_t = \zeta_s - \int_s^t \frac{1}{2}\lambda^2\zeta_u du + \int_s^t \zeta_u(i\lambda+\beta_u)dW_u$$

and in turn

$$\mathrm{E}(\zeta_t|\mathcal{F}_s) = \zeta_s - \frac{1}{2}\lambda^2\int_s^t \mathrm{E}(\zeta_u|\mathcal{F}_s)du,$$

where once again the martingale property of the stochastic integral has been used. This linear equation is explicitly solved for $\zeta_t$

$$\zeta_t = \zeta_s\exp\Big(-\frac{1}{2}\lambda^2(t-s)\Big)$$

and the claim (4.18) holds by (4.19). □

Remark 4.31. As we have seen in the Example 4.25, the verification of $\mathrm{E}\varphi_T = 1$ is not a trivial task. It holds if the process $\beta_t$ satisfies Novikov condition (e.g. Theorem 6.1 in [**21**])

$$\mathrm{E}\exp\left(\frac{1}{2}\int_0^T \beta_t^2 dt\right) < \infty. \tag{4.20}$$

Remark 4.32. The Girsanov theorem basically states that if $W$ is shifted by a sufficiently smooth function, then the obtained process induces a measure, absolutely continuous with respect to the Wiener measure. Obviously this wouldn't be possible if the shift is done by a function, say, with a jump - the obtained process won't have continuous trajectories. Let's try to shift $W$ by a continuous function: an independent Wiener process $W'$. In this case $V = W - W'$ is again a Wiener process with quadratic variation $2t$. Since quadratic variation is measurable with respect to natural filtration, the induced measure cannot be equivalent to the standard Wiener measure, corresponding to quadratic variation $t$. This indicates that certain degree of trajectories smoothness is required.

## 5. Stochastic Differential Equations

Let $(\Omega,\mathcal{F},\mathcal{F}_t,\mathrm{P})$ be a stochastic basis, carrying a Wiener process $W$. Let $a(t,x)$ and $b(t,x)$ be a pair of functionals on the space of continuous functions $C_{[0,T]}$, which are non-anticipating in the sense

$$x_1(s)\equiv x_2(s),\quad s\le t \quad\Longrightarrow\quad \begin{matrix} a(t,x_1)\equiv a(t,x_2) \\ b(t,x_1)\equiv b(t,x_2)\end{matrix} \qquad \forall t\in[0,T].$$

Equivalently this property can be formulated as measurability of $a(t,x)$ with respect to the Borel $\sigma$-algebra $\mathcal{B}_t$, generated by the open sets of $C_{[0,t]}$.

DEFINITION 4.33. A continuous random process $X$ is a unique *strong* solution of the stochastic differential equation (SDE)

$$dX_t = a(t,X)dt + b(t,X)dW_t \tag{4.21}$$

subject to a random $\mathcal{F}_0$-measurable initial condition $X_0 = \eta$, if

(1) $X$ is $\mathcal{F}_t$-adapted
(2) $X$ satisfies[10]

$$\mathrm{P}\left(\int_0^T |a(t,X)|dt < \infty\right) = 1, \quad \mathrm{P}\left(\int_0^T b^2(t,X)dt < \infty\right) = 1$$

(3) for each $t \in [0,T]$

$$X_t = \eta + \int_0^t a(s,X)ds + \int_0^T b(s,X)dW_s, \quad \mathrm{P}-a.s.$$

(4) (uniqueness) any two processes, satisfying (1)-(3) are indistinguishable.

The simplest conditions to guarantee the existence and uniqueness of the strong solutions are e.g.

THEOREM 4.34. *Assume that $a(t,x)$ and $b(t,x)$ satisfy the functional Lipschitz condition*

$$|a(t,x) - a(t,y)|^2 + |b(t,x) - b(t,y)|^2 \le L_1 \int_0^t |x_s - y_s|^2 dK_s + L_2|x_t - y_t|^2 \tag{4.22}$$

*and the linear growth condition*

$$a^2(t,x) + b^2(t,x) \le L_1 \int_0^t (1 + x_s^2)dK_s + L_2(1 + x_t^2) \tag{4.23}$$

*where $L_1$,$L_2$ are constants, $K_s$ is a nondecreasing right continuous function* [11]*, such that $0 \le K_s \le T$. Then the equation* (4.21) *has a unique strong solution.*

PROOF. (only the main idea - see Theorem 4.6 in [**21**] for details) The proof is in the spirit of classical differential equations by the Picard iterations method. Let $X_t^{(0)} \equiv X_0$ and define $X^{(n)}$ recursively

$$X_t^{(n)} = X_0 + \int_0^t a\big(s, X^{(n-1)}\big)ds + \int_0^t b\big(s, X^{(n-1)}\big)dW_s.$$

Now one shows, using the properties of Itô integral, that $\sup_{t\le T}|X_t^{(n)} - X_t^{(n-1)}|$ converges to zero as $n \to \infty$ P-a.s. and define the process

$$X_t := X_t^{(0)} + \sum_{n=0}^{\infty}\big(X_t^{(n+1)} - X_t^{(n)}\big).$$

Then it is verified that $X_t$ satisfies all the four properties in Definition 4.33. □

[10]Note that the strong solution actually employs the definition of the stochastic integral under weaker condition than $\mathcal{H}^2_{[0,T]}$, usually considered in these notes

[11]e.g. $K_s = s$

COROLLARY 4.35. *Let $a(t, x)$ and $b(t, x)$ be functions on $\mathbb{R}_+ \times \mathbb{R}$ satisfying the Lipschitz condition*

$$|a(t, x) - a(t, y)|^2 + |b(t, x) - b(t, y)|^2 \le L|x - y|^2, \quad x, y \in \mathbb{R}$$

*and the linear growth condition*

$$a^2(t, x) + b^2(t, x) \le L(1 + x^2).$$

*Then the SDE*

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t, \quad X_0 = \eta$$

*has a unique strong solution.*

REMARK 4.36. Analogous definition and proofs apply in the multivariate case, with appropriate adjustments in the conditions to be satisfied by the coefficients $a$ and $b$.

REMARK 4.37. Sometimes the existence and uniqueness can be verified under significantly weaker conditions: for example (first shown in [**43**]) the scalar equation with $b(t, x) \equiv 1$, has a unique strong solution if $a(t, x)$ is a bounded function on $\mathbb{R}_+ \times \mathbb{R}$ (without Lipschitz condition). This is a remarkable fact, since it is well known that classic ordinary differential equation may not have a unique solution if the drift $a(t, x)$ is not Lipschitz (e.g. $\dot{X} = 3/2\sqrt[3]{X}$, $X_0 = 0$ has two distinct solutions $X_t \equiv 0$ and $X_t = t^{3/2}$). Loosely speaking the equation is regularized if a small amount of white noise is plugged in! Even more remarkably, the strong solution ceases to exist in general if $a(t, x)$, being still bounded, is allowed to depend on the past of $x$ - a celebrated counterexample was given by B.Tsirelson in [**38**].

EXAMPLE 4.38. As in the world of ODE's, the explicit solutions to SDEs are rarely available. The Itô formula and a good guess are usually the main tools. For example the strong solution of the equation

$$dX_t = aX_t dt + bX_t dW_t, \quad X_0 = 1,$$

is

$$X_t = \exp\left(at - b^2/2t + bW_t\right).$$

Indeed,

$$dX_t = X_t d(at - b^2/2t + bW_t) + \frac{1}{2}b^2 X_t dt = aX_t dt + bX_t dW_t.$$

Sometimes it is easier to calculate various statistical parameters of the process, directly via the corresponding SDE. Let e.g. $m_t = \mathrm{E}X_t$ and $P_t = \mathrm{E}X_t^2$. Then

$$\mathrm{E}X_t = \mathrm{E}X_0 + a\int_0^t \mathrm{E}X_s ds, \quad \Longrightarrow \quad m_t = \mathrm{E}X_0 e^{at}.$$

Apply Itô formula to $X_t^2$ to get

$$X_t^2 = X_0^2 + \int_0^t 2X_s dX_s + \int_0^t b^2 X_s^2 ds = X_0^2 + \int_0^t (2a + b^2)X_s^2 ds + \int_0^t 2X_s b dW_s$$

and so

$$P_t = \mathrm{E}X_0^2 + \int_0^t (2a + b^2)\mathrm{E}X_s^2 ds \quad \Longrightarrow \quad P_t = \mathrm{E}X_0^2 \exp\{(2a + b^2)t\}.$$

■

Along with the strong solutions, *weak* solutions of (4.21) are defined.

DEFINITION 4.39. The equation (4.21) has a weak solution if there exists a probability basis $(\Omega', \mathcal{F}', \mathcal{F}'_t, \mathrm{P}')$, carrying a Wiener process $W$ and a continuous $\mathcal{F}'_t$-adapted process $X$, such that (4.21) is satisfied and $\mathrm{P}'(X_0 \le x) = \mathrm{P}(\eta \le x)$. If all weak solutions induce the same probability distribution, the equation (4.21) is said to have a unique weak solution.

REMARK 4.40. Note that in the case of strong solutions the random process $X$ is defined on the original probability space and thus $X$ is by definition adapted to $\mathcal{F}_t = \mathcal{F}^W_t \vee \sigma\{\eta\}$, i.e. the driving Wiener process $W$ "generates" $X$:

$$\mathcal{F}^X_t \subseteq \mathcal{F}^W_t \vee \sigma\{\eta\}.$$

In particular any strong solution is trivially also a weak solution with the choice $(\Omega', \mathcal{F}', \mathcal{F}'_t, \mathrm{P}') = (\Omega, \mathcal{F}, \mathcal{F}_t, \mathrm{P})$. In the case of weak solutions, one is allowed to choose a probability space and to construct on it a process $X$ to satisfy the relation (4.21). Typically (as we'll see shortly) the opposite inclusion holds for weak solutions

$$\mathcal{F}^X_t \supseteq \mathcal{F}^W_t \vee \sigma\{\eta\}$$

on the new probability space.

THEOREM 4.41. *Let $b(t,x) \equiv 1$ and $a(t,x)$ satisfy*

$$\mu^W\left(x \in C_{[0,T]} : \int_0^T a^2(t,x)dt < \infty\right) = 1,$$

*and*

$$\int_{C_{[0,T]}} \exp\left\{\int_0^T a(t,x)dW_t(x) - \frac{1}{2}\int_0^T a^2(t,x)dt\right\}\mu^W(dx) = 1$$

*where $\mu^W$ is the Wiener measure on $C_{[0,T]}$ and $W_t(x)$ is the coordinate process on the measure space $(C_{[0,T]}, \mathcal{B}, \mu^W)$, i.e. $W_t(x) := x_t$, $x \in C_{[0,T]}$, $t \in [0,T]$. Then* (4.21) *subject to $X_0 = 0$ has a weak solution.*

PROOF. Define

$$\varphi_T(x) = \exp\left(\int_0^T a(t,x)dW_t(x) - \frac{1}{2}\int_0^T a^2(t,x)dt\right)$$

and introduce a new measure $\mu$ on $(C_{[0,T]}, \mathcal{B})$ by

$$\frac{d\mu}{d\mu^W}(x) = \varphi_T(x).$$

Then by Girsanov theorem the process

$$W'_t := W_t - \int_0^t a\big(s, W\big)ds$$

is a Wiener process on $(C_{[0,T]}, \mathcal{B}, \mu)$ and hence $W$ is the weak solution of

$$dW_t = a(t, W_t)dt + dW'_t$$

on this probability space. □

REMARK 4.42. As the notion of "weak" suggests, (4.21) may have a weak solution, without having a strong one. The classical example is the Tanaka equation (see e.g. Chapter 5.3 in [**25**])

$$dX_t = \text{sign}(X_t)dW_t, \quad X_0 = 0.$$

To show that $X_t$ is not measurable with respect to $\mathcal{F}_t^W$ (and thus the equation does not have a strong solution) use the Tanaka formula (see Example 4.26).

Since the stochastic integral $\int_0^t \text{sign}(X_s)dW_s$ is a martingale [12] and its quadratic variation is $\int_0^t 1ds = t$, it is a Wiener process itself (by the Levy Theorem 4.5) and so by Tanaka formula (applied to $|X_t|$)

$$W_t = \int_0^t \text{sign}(X_t)dX_t = |X_t| - L_t,$$

where $L_t$ is the local time of (the Wiener process) $X_t$. Since the local time is measurable with respect to $\mathcal{F}_t^{|X|} = \sigma\{X_s, s \le t\}$, $W_t$ is measurable with respect to $\mathcal{F}_t^{|X|}$, which is strictly less than $\mathcal{F}_t^X$, hence

$$\mathcal{F}_t^W \subseteq \mathcal{F}_t^{|X|} \subset \mathcal{F}_t^X,$$

and $X_t$ cannot be a strong solution.

A weak solution is easily constructed by taking a Wiener process $W_t$ on some probability space and letting $dX_t = \text{sign}(W_t)dW_t$. Then $\text{sign}(W_t)dX_t = dW_t$, which is nothing but Tanaka equation with respect to the Wiener process $W_t$ on the new probability space. Note that on the original probability space $dX_t = \text{sign}(W_t)dW_t$ does not satisfy $dX_t = \text{sign}(X_t)dW_t$!

Another example of an SDE without strong solution (with nonzero drift with memory!) is the already mentioned Tsirelson equation (see e.g. Example in Section 4.4.8 in [**21**]).

**5.1. A connection to PDEs.** The theory and applications of SDEs with respect to Wiener process are vast (see e.g. [**36**], [**33**]), especially in the case of *diffusions*, i.e. when $a(t,x)$ (called the drift coefficient) and $b(t,x)$ (called diffusion matrix) are pointwise functions of $x$. In particular there is a close relation between various statistical properties of diffusions and PDEs.

As an example[13] consider the scalar diffusion

$$dX_t = a(X_t)dt + b(X_t)dW_t, \quad t \ge 0 \tag{4.24}$$

subject to a random variable $X_0$ with distribution $F(x)$, having density $q(x)$ with respect to the Lebesgue measure. Assume that the coefficients are such that the unique strong solution exists.

Define the second order differential (*forward* Kolmogorov-Focker-Planck) operator

$$(\mathcal{L}^* f)(x) = -\frac{\partial}{\partial x}\Big(a(x)f(x)\Big) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\Big(b^2(x)f(x)\Big). \tag{4.25}$$

and consider the Cauchy problem

$$\frac{\partial}{\partial t}p_t(x) = (\mathcal{L}^* p_t)(x) \tag{4.26}$$

$$p_0(x) = q(x). \tag{4.27}$$

---

[12] its integrand is bounded and thus satisfies the Novikov condition trivially

[13] to be revisited in the context of filtering below

Suppose that the unique solution $p_t(x)$ exists, such that for each $t \geq 0$ the function $p_t(x)$ decays sufficiently fast as $|x| \to \infty$. The conditions for this are well known from the theory of PDEs and can be found in textbooks.

Then $p_t(x)$ is the distribution density (with respect to the Lebesgue measure) of $X_t$ for a fixed $t$. Take a twice continuously differentiable function $f$. Then by the Itô formula, for any fixed $t \geq 0$

$$f(X_t) = f(X_0) + \int_0^t f'(X_s)a(X_s)ds + \int_0^t f'(X_s)b(X_s)dW_s + \frac{1}{2}\int_0^t f''(X_s)b^2(X_s)ds$$

and so

$$\mathrm{E}f(X_t) = \mathrm{E}f(X_0) + \int_0^t \mathrm{E}\Big(f'(X_s)a(X_s) + \frac{1}{2}f''(X_s)b^2(X_s)\Big)ds.$$

Let $F_t^X(dx)$ be the probability distribution of $X_t$, then the latter equation reads

$$\int_{\mathbb{R}} f(x)F_t^X(dx) = \int_{\mathbb{R}} f(x)q(x)dx + \\ \int_0^t \int_{\mathbb{R}} \Big(f'(x)a(x) + \frac{1}{2}f''(x)b^2(x)\Big)F_s^X(dx)ds. \quad (4.28)$$

Let's verify that $F_t^X(dx) = p_t(x)dx$ is a solution:

$$\int_{\mathbb{R}} \Big(f'(x)a(x) + \frac{1}{2}f''(x)b^2(x)\Big)F_s^X(dx) = \int_{\mathbb{R}} \Big(f'(x)a(x) + \frac{1}{2}f''(x)b^2(x)\Big)p_s(x)dx =$$
$$-\int_{\mathbb{R}} f(x)\frac{\partial}{\partial x}\Big(a(x)p_s(x)\Big)dx + \frac{1}{2}\int_{\mathbb{R}} f(x)\frac{\partial^2}{\partial x^2}\Big(b^2(x)p_s(x)\Big)dx = \int_{\mathbb{R}} f(x)(\mathcal{L}^*p_t)(x)dx$$

where the tail decay properties of $p_t(x)$ are to be used to ensure proper integration by parts. The right hand side of (4.28) becomes

$$\int_{\mathbb{R}} f(x)q(x)dx + \int_{\mathbb{R}} f(x)\int_0^t (\mathcal{L}^*p_t)(x)dx = \int_{\mathbb{R}} f(x)q(x)dx + \int_{\mathbb{R}} f(x)\int_0^t \frac{\partial}{\partial t}p_t(x)dx$$
$$= \int_{\mathbb{R}} f(x)q(x)dx + \int_{\mathbb{R}} f(x)\Big(p_t(x) - p_0(x)\Big)dx = \int_{\mathbb{R}} f(x)p_t(x)dx$$

and (4.28) holds. Of course these naive arguments leave many unanswered questions: e.g. it is not clear whether (4.28) defines the distribution of $X_t$ uniquely, etc. But nevertheless they give the correct intuition and the correct answer.

It can be shown that under certain conditions on the coefficients (e.g. $a(x)x \leq -x^2$ and $b^2(x) \geq C > 0$), the nonnegative solution $p(x)$ of the ODE

$$(\mathcal{L}^*p)(x) = 0$$

exists and is unique and

$$\lim_{t\to\infty} \int_{\mathbb{R}} |p_t(x) - p(x)|dx = 0.$$

In other words, the unique stationary distribution of $X_t$ exists and has density $p(x)$. In the scalar case it may be even found explicitly

$$p(x) = \frac{C}{b^2(x)} \exp\left\{\int_0^x \frac{2a(u)}{b^2(u)}du\right\}, \quad (4.29)$$

where $C$ is the normalization constant.

## 6. Martingale representation theorem

Martingales have been mentioned before on several occasions:

DEFINITION 4.43. The process $X_t$ is an $\mathcal{F}_t$-martingale[14] if $X_t$ is $\mathcal{F}_t$-adapted and $\mathrm{E}(X_t|\mathcal{F}_s) = X_s$ for any $t \geq s \geq 0$.

The Wiener process and the stochastic integral (under appropriate conditions imposed on the integrand) are examples of martingales. It turns out that any martingale with respect to the filtration $\mathcal{F}_t^W$ generated by a Wiener process $W_t$ is necessarily a stochastic integral with respect to $W_t$. We chose the simplified approach of [**25**] to hint how this deep result emerges. The more complete treatment of the subject can be found in Chapter 5 of [**21**].

THEOREM 4.44. *(The Itô representation theorem) Let $\xi$ be a square integrable $\mathcal{F}_T^W$ measurable random variable, i.e. $\xi \in \mathbb{L}^2(\Omega, \mathcal{F}_T^W, \mathrm{P})$. Then there is an $\mathcal{H}^2_{[0,T]}$ process $f(t,\omega)$, such that*

$$\xi = \mathrm{E}\xi + \int_0^T f(s,\omega)dW_s, \quad \mathrm{P}-a.s. \tag{4.30}$$

REMARK 4.45. When $(\xi, W)$ form a Gaussian process, deterministic $f(t,\omega) \equiv f(t)$ in (4.30) always exists - see Example 4.47.

PROOF. The idea is to show[15] that the linear closed subspace $\mathscr{E}$ of random variables of the form[16]

$$\eta_T := \exp\left\{\int_0^T h_s dW_s - \frac{1}{2}\int_0^T h_s^2 ds\right\}, \quad \forall h : [0,T] \mapsto \mathbb{R}, \int_0^T h_s^2 ds < \infty \tag{4.31}$$

is dense in $\mathbb{L}^2(\Omega, \mathcal{F}_T^W, \mathrm{P})$ (all square integrable functionals of the Wiener process on $[0,T]$). By the Itô formula

$$\eta_T = 1 + \int_0^T h_s \eta_s dW_s,$$

and thus $\eta_T$ admits the representation (4.30) (with $f(t,\omega) = h_t\eta_t$). Due to linearity of the stochastic integral the linear combinations of random variables from $\mathscr{E}$ are also of the form (4.31). If the subspace $\mathscr{E}$ is dense in $\mathbb{L}^2(\Omega, \mathcal{F}_T^W, \mathrm{P})$, any $\mathcal{F}_T^W$-measurable random variable $\xi$ can be approximated by a convergent sequence $\xi^n \in \mathscr{E}$:

$$\xi^n = \mathrm{E}\xi^n + \int_0^T f^n(s,\omega)dW_s.$$

Then by the Itô isometry,

$$\mathrm{E}\big(\xi^n - \xi^m\big)^2 = \big(\mathrm{E}\xi^n - \mathrm{E}\xi^m\big)^2 + \int_0^T \mathrm{E}\big(f^n(s,\omega) - f^m(s,\omega)\big)^2 ds$$

---

[14] Sometimes the pair $(X_t, \mathcal{F}_t)$ us referred as martingale

[15] the proof is taken from §4.3 [**25**] (the same proof is used in Ch. V, §3[**27**]). Different proof is given in §5.2 [**21**].

[16] the functions $h$ are deterministic

and since $\xi^n$ converges in $\mathbb{L}^2(\Omega, \mathcal{F}_T^W, \mathrm{P})$, $f^n(t,\omega)$ is a Cauchy sequence and hence is also convergent, i.e. the limit $f(t,\omega)$ exists in the sense

$$\int_0^T \mathrm{E}\big(f^n(s,\omega) - f(s,\omega)\big)^2 ds \xrightarrow{n\to\infty} 0.$$

Since $f^n$ are adapted, $f$ is adapted as well and again by the Itô isometry

$$\xi^n = \mathrm{E}\xi^n + \int_0^T f^n(s,\omega)dW_s \xrightarrow[\mathbb{L}^2]{n\to\infty} \mathrm{E}\xi + \int_0^T f(s,\omega)dW_s.$$

and hence $\xi$ admits (4.30).

Suppose that $f$ is non-unique, i.e. there are $f_1$ and $f_2$, so that

$$\xi = \mathrm{E}\xi + \int_0^T f_1(s,\omega)dW_s = \mathrm{E}\xi + \int_0^T f_2(s,\omega)dW_s.$$

This implies $\int_0^T \mathrm{E}\big(f_1(s,\omega) - f_2(s,\omega)\big)^2 ds = 0$, i.e. $f_1 = f_2$, $ds \times \mathrm{P}$-a.s.

So the main issue is to verify that $\mathscr{E}$ is dense in $\mathbb{L}^2(\Omega, \mathcal{F}_T^W, \mathrm{P})$, or equivalently to check that if $\zeta \in \mathbb{L}^2(\Omega, \mathcal{F}_T^W, \mathrm{P})$ satisfies

$$\mathrm{E}\eta\zeta = 0, \quad \forall \eta \in \mathscr{E}, \tag{4.32}$$

then $\zeta \equiv 0$, P-a.s. If (4.32) holds, then in particular

$$\mathrm{E}\exp\left\{\sum_{i=1}^n \lambda_i\big(W_{t_{i+1}} - W_{t_i}\big) - \frac{1}{2}\sum_{i=1}^n \lambda_i^2(t_{i+1} - t_i)\right\}\zeta = 0$$

for any finite number of $0 = t_1 < ... < t^n = T$ and any constants $\lambda_i$, $i = 1, ..., n$, which is equivalent to

$$\mathrm{E}\exp\left\{\sum_{i=1}^n \alpha_i W_{t_i}\right\}\zeta = 0,$$

for any real numbers $\alpha_i$. It is easy to verify that the function

$$G(\alpha) = \mathrm{E}\exp\left\{\sum_{i=1}^n \alpha_i W_{t_i}\right\}\zeta, \quad \alpha \in \mathbb{R}^n$$

is real analytic (i.e. has derivatives of any order at any $\alpha \in \mathbb{R}^n$). Then the complex function

$$G(z) = \mathrm{E}\exp\left\{\sum_{i=1}^n z_i W_{t_i}\right\}\zeta, \quad z \in \mathbb{C}^n$$

is analytic as well (i.e. satisfies the Cauchy-Riemann condition or equivalently has a complex derivative at any point of $\mathbb{C}^n$). The analytic function, which vanishes on the real line (or on the real lines in this case), vanishes everywhere on the complex plain and thus in particular vanishes on the complex axes

$$G(i\alpha) = \mathrm{E}\exp\left\{\sum_{i=1}^n i\alpha_i W_{t_i}\right\}\zeta, \quad \alpha \in \mathbb{R}^n.$$

Now for an arbitrary real analytic function $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ with compact support

$$\mathrm{E}\varphi(W_{t1}, ..., W_{t_n})\zeta = \mathrm{E}(2\pi)^{-n/2}\int_{\mathbb{R}^n} \hat{\varphi}(u)\exp\big\{iu_1 W_{t_1} + ... + iu_n W_{t_n}\big\}\zeta =$$

$$(2\pi)^{-n/2}\int_{\mathbb{R}^n} \hat{\varphi}(u)\mathrm{E}\exp\big\{iu_1 W_{t_1} + ... + iu_n W_{t_n}\big\}\zeta = 0.$$

The claim holds, since smooth compactly supported functions approximate Borel functions in $\mathbb{L}^2$. □

REMARK 4.46. The integrand in (4.30) is an adapted random process. It turns out that functionals of the Wiener process can be expanded into *multiple* integrals with respect to $W$ with *non-random* kernels - this is so called Wiener chaos expansion.

EXAMPLE 4.47. The random variable $\xi = \int_0^T W_s ds$ is $\mathcal{F}_T^W$-measurable with

$$\xi = \int_0^T (T-t)dW_t.$$

■

THEOREM 4.48. *(The martingale representation theorem) Let $X_t$ be an square integrable*[17] *$\mathcal{F}_t^W$-martingale. Then there is a unique $\mathcal{H}^2_{[0,T]}$ process $g(s,\omega)$, adapted to $\mathcal{F}_t^W$, such that*

$$X_t = \mathrm{E}X_0 + \int_0^t g(s,\omega)dW_s, \quad t \in [0,T], \quad \mathrm{P}-a.s.$$

PROOF. By Theorem 4.44, for each fixed $t \in [0,T]$, there is a unique $\mathcal{F}_t^W$-measurable process $f^{(t)}(s,\omega)$, such that ($\mathrm{E}\xi_t = \mathrm{E}\xi_0$)

$$\xi_t = \mathrm{E}\xi_0 + \int_0^t f^{(t)}(s,\omega)dW_s,$$

and we shall verify that $f^{(t)}(s,\omega)$ can be chosen independently of $t$. Let $T \ge t_2 \ge t_2 \ge 0$, then

$$\mathrm{E}\big(\xi_{t_2}|\mathcal{F}_{t_1}^W\big) = \mathrm{E}\xi_0 + \mathrm{E}\Big(\int_0^{t_2} f^{(t_2)}(s,\omega)dW_s\Big|\mathcal{F}_{t_1}^W\Big) = \mathrm{E}\xi_0 + \int_0^{t_1} f^{(t_2)}(s,\omega)dW_s.$$

On the other hand

$$\mathrm{E}\big(\xi_{t_2}|\mathcal{F}_{t_1}^W\big) = \xi_{t_1} = \mathrm{E}\xi_0 + \int_0^{t_1} f^{(t_1)}(s,\omega)dW_s$$

and hence by Itô isometry, $f^{(t_2)}(s,\omega)$ and $f^{(t_1)}(s,\omega)$ coincide on $[0,t_1]$, namely

$$\int_0^{t_1} \mathrm{E}\big(f^{(t_2)}(s,\omega) - f^{(t_1)}(s,\omega)\big)^2 ds = 0.$$

Then one can choose

$$f(s,\omega) = f^{(T)}(s,\omega),$$

so that

$$\xi_t = \mathrm{E}\xi_0 + \int_0^t f^{(T)}(s,\omega)dW_s = \mathrm{E}\xi_0 + \int_0^t f^{(t)}(s,\omega)dW_s.$$

□

[17] $\sup_{t\in[0,T]} \mathrm{E}X_t^2 < \infty$

EXAMPLE 4.49. Let $\xi = W_1^4$ and consider the martingale $X_t = \mathrm{E}\big(W_1^4|\mathcal{F}_t^W\big)$, $t \le 1$. By the Markov property of $W$, $X_t = \mathrm{E}(W_1^4|W_t)$. Since $(W_1, W_t)$ is a Gaussian pair, the conditional distribution of $W_1$ given $W_t$ is Gaussian as well with the mean $W_t$ and variance $1-t$. Hence

$$\begin{aligned}\mathrm{E}(W_1^4|W_t) =& \mathrm{E}\big((W_1 - W_t + W_t)^4|W_t\big) = \\ & \mathrm{E}\big((W_1-W_t)^4|W_t\big) + 4\mathrm{E}\big((W_1-W_t)^3 W_t|W_t\big) + \\ & 6\mathrm{E}\big((W_1-W_t)^2 W_t^2|W_t\big) + 4\mathrm{E}\big((W_1-W_t)W_t^3|W_t\big) + W_t^4 = \\ & 3(1-t)^2 + 6(1-t)W_t^2 + W_t^4.\end{aligned}$$

Applying the Itô formula one gets

$$\begin{aligned}dX_t = -6(1-t)dt - 6W_t^2 dt + 12(1-t)dW_t + 6(1-t)dt \\ + 4W_t^3 dW_t + 6W_t^2 dt = 12(1-t)dW_t + 4W_t^3 dW_t.\end{aligned}$$

and hence

$$\xi = X_1 = X_0 + \int_0^1 \big(12(1-t) + 4W_t^3\big)dW_t = 3 + \int_0^1 \big(12(1-t) + 4W_t^3\big)dW_t.$$

■

EXAMPLE 4.50. This representation is not always easy to find explicitly. Here is one amazing formula: the random variable $S_1 = \sup_{s\in[0,1]} W_s$ satisfies

$$S_1 = \mathrm{E}S_1 + 2\int_0^1 \left(1 - \Phi\Big(\frac{S_t - B_t}{\sqrt{1-t}}\Big)\right) dW_t$$

where $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-r^2/2} dr$. ■

The following theorem will be extensively used in the derivation of nonlinear filtering equations.

THEOREM 4.51. *Let $Y = (Y_t)_{t\in[0,T]}$ be the strong solution*[18] *of the SDE*

$$dY_t = a_t(Y)dt + dW_t,$$

*where $a_t(\cdot)$ is a non-anticipating functional on $C_{[0,T]}$, satisfying*

$$\int_0^T \mathrm{E}a_t^2(Y)dt < \infty, \quad \text{and} \quad \int_0^T \mathrm{E}a_t^2(W)dt < \infty$$

*Then any square integrable $\mathcal{F}_t^Y$-martingale $Z_t$ has a continuous version satisfying*

$$Z_t = Z_0 + \int_0^t g(s,\omega)dW_s$$

*with an $\mathcal{H}^2_{[0,T]}$ process $g(s,\omega)$, adapted to $\mathcal{F}_t^Y$.*

PROOF. Due to the assumptions on $a_t(\cdot)$, the process

$$\begin{aligned}\varphi_T(\omega) = \exp\left\{-\int_0^t a_s(Y)dW_s - \frac{1}{2}\int_0^t a_s^2(Y)ds\right\} = \\ \exp\left\{-\int_0^t a_s(Y)dY_s + \frac{1}{2}\int_0^t a_s^2(Y)ds\right\}\end{aligned}$$

[18] in other words $a$ is such that the strong solution exists

is an $\mathcal{F}_t^Y$-martingale under P and thus the Radon-Nikodym density

$$\frac{d\widetilde{\mathrm{P}}}{d\mathrm{P}}(\omega) = \varphi_T(\omega),$$

defines probability $\widetilde{\mathrm{P}}$. Moreover by Girsanov theorem, $Y_t$ is a Wiener process under $\widetilde{\mathrm{P}}$. The process $z_t := Z_t/\varphi_t$ is an $\mathcal{F}_t^Y$-martingale under $\widetilde{\mathrm{P}}$:

$$\widetilde{\mathrm{E}}|z_t| = \mathrm{E}|z_t|\varphi_T = \mathrm{E}|z_t|\mathrm{E}\big(\varphi_T|\mathcal{F}_t^Y\big) = \mathrm{E}|z_t|\varphi_t = \mathrm{E}|Z_t| < \infty$$

and by Lemma 3.11

$$\widetilde{\mathrm{E}}(z_t|\mathcal{F}_s^Y) = \widetilde{\mathrm{E}}\Big(\frac{Z_t}{\varphi_t}|\mathcal{F}_s^Y\Big) = \frac{\mathrm{E}\Big(\frac{Z_t}{\varphi_t}\varphi_T|\mathcal{F}_s^Y\Big)}{\mathrm{E}(\varphi_T|\mathcal{F}_s^Y)} = \frac{\mathrm{E}(Z_t|\mathcal{F}_s^Y)}{\varphi_s} = z_s.$$

Then by Theorem 4.48, $z_t$ admits the representation ($Y$ is a Wiener process under $\widetilde{P}$)

$$z_t = z_0 + \int_0^t f(s,\omega)dY_s = z_0 + \int_0^t f(s,\omega)a_s(Y)ds + \int_0^t f(s,\omega)dW_s$$

with an $\mathcal{F}_t^Y$-adapted process $f$. Applying the Itô formula to $Z_t = z_t\varphi_t$ one gets (recall that $d\varphi_t = -a_t(Y)\varphi_t dW_t$)

$$\begin{aligned} dZ_t = z_t d\varphi_t + \varphi_t dz_t - a_t\varphi_t f(t,\omega)dt = -z_t a_t\varphi_t dW_t + \varphi_t f(t,\omega)a_t dt + \\ \varphi_t f(t,\omega)dW_t - a_t\varphi_t f(t,\omega)dt = \big(\varphi_t f(t,\omega) - z_t a_t\big)dW_t, \end{aligned}$$

and thus the required representation holds with $g(s,\omega) := \varphi_t f(t,\omega) - z_t a_t(Y)$. $\square$

## Exercises

(**1**) Prove that the limit of a sequence of uniformly convergent continuous functions $f^n : [0,1] \mapsto \mathbb{R}$ is continuous.

(**2**) Plot a typical path of $W_t^n$, defined in (4.2) for $n = 1, 2, 3$

(**3**) Prove

$$\mathrm{P}\big(D^+W_t = \infty \text{ and } D_+W_t = -\infty\big) = 1, \quad \forall t \in [0,T]$$

(**4**) Verify that for a standard Gaussian r.v. $\xi$, $\mathrm{P}(|\xi| \le \varepsilon) \le \varepsilon$ for any $\varepsilon > 0$.

(**5**) Prove the law of large numbers

$$\mathrm{P}\big(\lim_{t\to\infty} W_t/t = 0\big) = 1.$$

(**6**) Let $W_t$, $t \in [0,1]$ be the Wiener process (with respect to its natural filtration $\mathcal{F}_t^W$). Verify that each of the following processes is a Wiener process with respect to appropriate filtration.

(a) *Scaling invariance*: for any constant $c > 0$

$$W_t^c := \frac{1}{\sqrt{c}}W_{ct}, \quad t \le 1$$

(b) *Time inversion:*

$$Y_t = \begin{cases} tW_{1/t}, & t \in (0,1] \\ 0, & t = 0. \end{cases}$$

(c) *Time reversal:*

$$Z = W_1 - W_{1-t}, \quad t \le 1.$$

(d) *Symmetry:*
$$V_t = -W_t, \quad t \le 1.$$

**(7)** Let $f : \mathbb{R} \mapsto [-K, K]$ for some constant $0 < K < \infty$ be a twice continuously differentiable function with bounded derivatives. For a fixed number $q \in [0,1]$, define
$$I_t^{q,n} = \sum_{i=1}^{[nt]} f(W_{s_i^q})\big(W_{s_i} - W_{s_{i-1}}\big)$$
where $s_i = i/n$, $i \le n$ and $s_i^q = q s_{i-1} + (1-q)s_i$.

(a) Show that the $\mathbb{L}$ limit $I_t^q = \lim_{n\to\infty} I_t^{q,n}$ exists (in particular for $q = 1$, the Itô integral $I_t := I_t^1$ is obtained). Calculate the expectation of $I_t^q$.

(b) Verify the Wong-Zakai correction formula
$$I_t^q = I_t + (1-q)\int_0^t f'(W_s)ds.$$

**(8)** Prove directly from the definition of Itô integral with respect to the Brownian motion $B$ that

(a) $\int_0^t s dB_s = tB_t - \int_0^t B_s ds$

(b) $\int_0^t B_s^2 dB_s = \frac{1}{3}B_t^3 - \int_0^t B_s ds$

**(9)** Use the Itô formula to verify the integration by parts rule. Let $f_t : \mathbb{R}_+ \mapsto \mathbb{R}$ be a deterministic differentiable function, then
$$\int_0^t f_s dW_s = W_t f_t - \int_0^t W_s \dot{f}_t dt.$$
Use the multivariate Itô formula to derive the analogue of integration by parts rule, when $f_t$ is another Itô process with respect to the same Wiener process: $df_t = a_t dt + b_t dW_t$.

**(10)** Let $a_t$ and $b_t$ be a pair of deterministic functions. Find the differential of the process
$$X_t = \exp\left\{\int_0^t a_s ds\right\}\left\{x + \int_0^t \exp\Big(-\int_0^s a_u du\Big) b_s dW_s\right\},$$
where $x \in \mathbb{R}$. Show that the mean $m_t = \mathrm{E}X_t$, variance $V_t = \mathrm{E}(X_t - m_t)^2$ and covariance $K(t,s) = \mathrm{E}(X_t - m_s)(X_s - m_s)$ functions satisfy the equations
$$\begin{aligned} \dot{m}_t &= a_t m_t, \quad m_0 = x \\ \dot{V}_t &= 2a_t V_t + b_t^2, \quad V_0 = 0 \\ K(t,s) &= \exp\left\{\int_s^t a_u ds\right\} V_s, \quad t \ge s \end{aligned}$$

**(11)** Use the multivariate Itô formula to show that the process
$$R_t = \sqrt{(W_t^1)^2 + \ldots + (W_t^d)^2}, \quad t \ge 0$$
where $W_t^i$ are independent Wiener processes, satisfies
$$dR_t = \sum_{i=1}^d \frac{W_t^i dW_t^i}{R_t} + \frac{d-1}{2R_t}dt.$$

This is so called $d$-dimensional Bessel process. For the case $d = 2$, show that

$$R_3 \le \mathrm{E}(R_4|W_3, V_3) \le \sqrt{2 + R_3^2}.$$

**Hint:** the upper bound can be obtained by Jensen inequality.

(**12**) Let $\beta_k(t) = \mathrm{E}W_t^k$, $k = 0, 1, 2, ....$ Use the Itô formula to derive the recursion

$$\beta_k(t) = \frac{1}{2}k(k-1)\int_0^t \beta_{k-2}(s)ds, \quad k \ge 2.$$

Deduce that $\mathrm{E}W_t^4 = 3t^2$ and find $\mathrm{E}W_t^6$.

(**13**) Explain the origins of mnemonic rules in Remark 4.28 by sketching the proof of multivariate Itô formula

(**14**) Obtain the answer in Example 4.29 by applying the Itô formula directly (avoiding the use of table).

(**15**) Verify the existence and uniqueness of the strong solution of the following equations (check the conditions of Theorem 4.34). Check whether the given processes solve the corresponding equations as claimed.

(a) $X_t = e^{B_t}$ solves

$$dX_t = 0.5X_t dt + X_t dB_t, \quad X_0 = 1$$

(b) $X_t = B_t/(t+1)$ solves

$$dX_t = -\frac{1}{1+t}X_t dt + \frac{1}{1+t}dB_t, \quad X_0 = 0$$

(c) $X_t = \sin(W_t)$ solves

$$dX_t = -\frac{1}{2}X_t dt + \sqrt{1 - X_t^2}dB_t, \quad B_0 \in (-\pi/2, \pi/2)$$

(d) $X_1(t) = X_1(0)+t+B_1$ and $X_2(t) = X_2(0)+X_1(0)B_2(t)+\int_0^t sdB_2(s)+\int_0^t B_1(s)dB_2(s)$ solve

$$\begin{aligned} dX_1 &= dt + dB_1 \\ dX_2 &= X_1 dB_2 \end{aligned}$$

(e) $X_t = e^{-t}X_0 + e^{-t}B_t$ solves

$$dX_t = -X_t dt + e^{-t}dB_t.$$

(f) $Y_t = \exp(aB_t - 0.5a^2t)\big[Y_0 + r\int_0^t \exp(-aB_s + 0.5a^2s)ds\big]$ solves

$$dY = rdt + aY dB_t.$$

(g) The processes $X_1(t) = X_1(0)\cosh(t) + X_2(0)\sinh(t) + \int_0^t a\cosh(t-s)dB_1 + \int_0^t b\sinh(t-s)dB_2$ and $X_2(t) = X_1(0)\sinh(t) + X_2(0)\cosh(t) + \int_0^t a\sinh(t-s)dB_1 + \int_0^t b\cosh(t-s)dB_2$ solve

$$\begin{aligned} dX_1 &= X_2 dt + adB_1 \\ dX_2 &= X_1 dt + bdB_2, \end{aligned}$$

which can be seen as stochastically excited vibrating string equations.

(h) The process $X_t = \big(X_1(t), X_2(t)\big) = \big(\cosh(B_t), \sinh(B_t)\big)$ solve

$$dX_t = \frac{1}{2}X_t dt + X_t dB_t.$$

(**16**) Let $X$ and $Y$ be the strong solution of

$$\begin{aligned} dX_t &= -0.5X_t dt - Y_t dB_t \\ dY_t &= -0.5Y_t dt + X_t dB_t. \end{aligned}$$

subject to $X_0 = x$ and $Y_0 = y$ with $B_t$ being a Wiener process (Brownian motion).

(a) Show that $X_t^2 + Y_t^2 \equiv x^2 + y^2$ for all $t \geq 0$, i.e. the vector $(X_t, Y_t)$ revolves on a circle.

(b) Find the SDE, satisfied by $\theta_t = \arctan(X_t/Y_t)$.

(**17**) Consider the multivariate linear SDE

$$dX_t = AX_t dt + BdW_t, \quad X_0 = \eta,$$

where $A$ and $B$ are $n \times n$ and $n \times m$ matrices, $W$ is the vector of $m$ independent Wiener process (usually referred as vector Wiener process) and $\eta$ is a random variable independent of $W$ and $\mathrm{E}\|\eta\|^2 < \infty$.

(a) Find the explicit strong solution of the vector linear equation

(b) Find the explicit expressions for $M_t = \mathrm{E}X_t$ and $Q_t = \mathrm{cov}(X_t) = \mathrm{E}(X_t - m_t)(X_t - m_t)^*$ (**Hint**: find first the ODE's for $m_t$ and $Q_t$)

(c) Find the explicit expression for the correlation matrix $K_{t,s} = \mathrm{E}(X_t - m_t)(X_s - m_s)^*$ in terms of $Q_t$

(d) Give simple sufficient conditions on $A$,$B$ and $\eta$ so that the process $X_t$ is stationary, i.e. $m_t \equiv m$ and $Q_t \equiv Q$ for certain (what?) $m$ and $Q$.

(e) The linear one dimensional diffusion $X_t$ is called Ornstein-Uhlenbeck process. Specify your answers in the previous questions in this case.

(**18**) Consider the equation of a harmonic oscillator, driven by the "white noise" $N_t$

$$\ddot{X}_t + (1 + \varepsilon N_t)X = 0, \quad X_0 = 1, \dot{X}_0 = 1$$

where $\varepsilon > 0$ is a parameter.

(a) Write this equation as a two dimensional linear Itô SDE with respect to the Wiener process

(b) Find the mean, variance and covariance functions of the oscillator position

(c) Verify that the position satisfies the stochastic Volterra equation

$$X_t = X_0 + \dot{X}_0 t + \int_0^t (r - t)X_r dr + \int_0^t \varepsilon(r - t)X_r dW_r$$

(**19**) Write down the KFP PDE, corresponding to the linear SDE

$$dX_t = -aX_t dt + bdW_t, \quad X_0 \sim \eta$$

where $\eta$ is a standard Gaussian random variable, $b > 0$ and $a > 0$ are constants. Find the stationary density $p(x)$ and calculate the stationary mean and the variance. Compare to Exercise (**17**).

(**20**) Find explicit Itô representation for the following functionals of $W$ on $[0, T]$: $W_T$, $W_T^2$, $W_T^3$, $e^{W_T}$, $\sin W_T$. **Hint:** use the Itô formula.

CHAPTER 5

# Linear filtering in continuous time

The continuous time linear filtering problem is addressed in this chapter, using the white noise formalism, developed in the preceding one. In continuous time setting the filtering formulae are derived by solving the Wiener-Hopf equation, rather than using the general recursive formulae for orthogonal projection as in the discrete time.

## 1. The Kalman-Bucy filter: scalar case

Consider the following system of linear SDEs:

$$dX_t = a_t X_t dt + b_t dW_t \tag{5.1}$$
$$dY_t = A_t X_t dt + B_t dV_t \tag{5.2}$$

where $W$ and $V$ are independent Wiener processes and the (scalar) coefficients are deterministic functions of $t$, such that the system has a unique strong solution. These equations are solved subject to random variables $X_0$ and $Y_0$ with the bounded covariance matrix, assumed independent of $(W, V)$. Hereafter $B_t^2 \geq C > 0$ for some constant $C$.

In what follows $\mathcal{L}_t^Y$ denotes the closed linear subspace generated by the random variables $Y_s, s \leq t$ and $\widehat{\mathrm{E}}(\cdot|\mathcal{L}_t^Y)$ is the orthogonal projection[1] on $\mathcal{L}_t^Y$. As discussed in Chapter 2, $\widehat{X}_t := \widehat{\mathrm{E}}(X_t|\mathcal{L}_t^Y)$ is the best linear estimate of $X_t$, given the observations $\{Y_s, s \leq t\}$.

Theorem 5.1. *(Kalman-Bucy filter) The optimal linear estimate $\widehat{X}_t$ and the corresponding mean square error $P_t = \mathrm{E}(X_t - \widehat{X}_t)^2$ satisfy the equations*

$$\begin{aligned} \widehat{X}_t &= a_t \widehat{X}_t dt + \frac{P_t A_t}{B_t^2}\big(dY_t - A_t \widehat{X}_t dt\big) \\ \dot{P}_t &= 2a_t P_t + b_t^2 - \frac{A_t^2 P_t^2}{B_t^2} \end{aligned} \tag{5.3}$$

*subject to*

$$\begin{aligned} \widehat{X}_0 &= \mathrm{E}X_0 + \mathrm{cov}(X_0, Y_0)\,\mathrm{cov}^{\oplus}(Y_0)\big(Y_0 - \mathrm{E}Y_0\big) \\ P_0 &= \mathrm{cov}(X_0) - \mathrm{cov}^2(X_0, Y_0)\,\mathrm{cov}^{\oplus}(Y_0). \end{aligned} \tag{5.4}$$

Proof. The proof is done in several steps:

Step 1 *(getting rid of $\widehat{X}_0$)*

[1] as usual a constant is added to any linear subspace

It would be easier to treat the case $\widehat{X}_0 \equiv 0$ and we claim that it is enough to prove the theorem under this assumption: introduce

$$X'_t = X_t - \widehat{X}_0 \exp\left(\int_0^t a_s ds\right), \quad Y'_t = Y_t - \int_0^t A_s \widehat{X}_0 \exp\left(\int_0^s a_u du\right).$$

The process $(X'_t, Y'_t)$ satisfies

$$\begin{aligned} dX'_t &= a_t X'_t dt + b_t dW_t \\ dY'_t &= A_t X'_t dt + B_t dV_t, \end{aligned}$$

subject to $X'_0 = X_0 - \widehat{X}_0$ and $Y'_0 = Y_0$. Clearly $\mathcal{L}^Y_t = \mathcal{L}^{Y'}_t$ and hence

$$\widehat{X}_t = \widehat{\mathrm{E}}\big(X_t | \mathcal{L}^Y_t\big) = \widehat{\mathrm{E}}\big(X_t | \mathcal{L}^{Y'}_t\big) = \widehat{\mathrm{E}}\big(X'_t | \mathcal{L}^{Y'}_t\big) + \widehat{X}_0 \exp\left(\int_0^t a_s ds\right).$$

Note that $\mathrm{E}\big(X'_0 | Y'_0\big) = 0$ and suppose that $\widehat{X}'_t = \widehat{\mathrm{E}}(X'_t | \mathcal{L}^{Y'}_t)$ and $P'_t = \mathrm{E}(X'_t - \widehat{X}'_t)^2$ satisfy (5.3), subject to $\widehat{X}'_0 = 0$ and $P'_0 = \mathrm{E}(X'_0 - \widehat{X}'_0)^2$. Then

$$\begin{aligned} d\widehat{X}_t =& d\widehat{X}'_t + a_t \widehat{X}_0 \exp\left(\int_0^t a_s ds\right) dt = \\ & a_t \widehat{X}_t dt + \frac{P_t A_t}{B_t^2}\Big(dY_t - A_t \widehat{X}_0 \exp\Big\{\int_0^t a_s ds\Big\} - A_t \widehat{X}'_t dt\Big) = \\ & a_t \widehat{X}_t dt + \frac{P_t A_t}{B_t^2}\big(dY_t - A_t \widehat{X}_t dt\big), \end{aligned}$$

which means that $\widehat{X}_t$ satisfies (5.3) equation as well, subject to $\widehat{X} = \widehat{\mathrm{E}}(X_0|Y_0)$, given by the first equation of (5.4). Moreover

$$\begin{aligned} P_t = \mathrm{E}\Big(X_t - \widehat{X}_t\Big)^2 = \mathrm{E}\Big(X'_t + \widehat{X}_0 \exp\Big\{\int_0^t a_s ds\Big\} - \\ \widehat{X}'_t - \widehat{X}_0 \exp\Big\{\int_0^t a_s ds\Big\}\Big)^2 = \mathrm{E}(X'_t - \widehat{X}'_t)^2 = P'_t, \end{aligned}$$

i.e. $P_t$ satisfies the equation from (5.3).

<u>Step 2</u> *(the general form of the estimate)*

From here on $\widehat{\mathrm{E}}(X_0|Y_0) = 0$ is assumed P-a.s. Let $0 = t_1 < ... < t_n = T$ be a partition of $[0, T]$ and denote by $\mathcal{L}^Y_t(n)$ the subspace, spanned by $\{Y_{t_1}, ..., Y_{t_n}\}$. This subspace coincides with the one spanned by the increments $\{Y_{t_1}, Y_{t_2} - Y_{t_1}, ..., Y_{t_n} - Y_{t_{n-1}}\}$ and so

$$\widehat{\mathrm{E}}\big(X_t | \mathcal{L}^Y_t(n)\big) = \widehat{\mathrm{E}}(X_t|Y_0) + \sum_{j=1}^{n-1} g_j\big(Y_{t_{j+1}} - Y_{t_j}\big) = \widehat{\mathrm{E}}(X_t|Y_0) + \int_0^t G^n(t,s) dY_s,$$

where $g_j$ are real numbers and $G(t,s) = \sum_{j \le n} g_j \mathbf{1}_{\{s \in [t_j, t_{j+1})\}}$. Since $\mathcal{L}^Y_t$ is a closed subspace,

$$\lim_{n \to \infty} \widehat{\mathrm{E}}\big(X_t | \mathcal{L}^Y_t(n)\big) = \widehat{\mathrm{E}}\big(X_t | \mathcal{L}^Y_t\big),$$

and hence

$$\mathrm{E}\left(\int_0^t G^n(t,s) dY_t - \int_0^t G^m(t,s) dY_t\right)^2 \xrightarrow{n,m \to \infty} 0.$$

Since $X$ and $V$ are independent, the latter implies

$$\left(\int_0^t \big(G^n(t,s)-G^m(t,s)\big)^2 A_s X_s ds\right)^2 + \int_0^t \big(G^n(t,s)-G^m(t,s)\big)^2 B_s^2 ds \xrightarrow{n,m\to\infty} 0$$

Then due to the assumption $B_s^2 \ge C > 0$, $G^n(t,s)$ is a Cauchy sequence and hence converges to a limit $G(t,s)$, so that

$$\widehat{\mathrm{E}}(X_t|\mathcal{L}_t^Y) = \widehat{\mathrm{E}}(X_t|Y_0) + \int_0^t G(t,s)dY_s.$$

<u>Step 3</u> *(using orthogonality)*

Recall that $\widehat{\mathrm{E}}(X_0|Y_0) = 0$, P-a.s. is assumed, so that $\mathrm{E}X_t = 0$ and $\widehat{\mathrm{E}}(X_t|Y_0) = 0$. The function $G(t,s)$ satisfies the Wiener-Hopf equation

$$K(t,u)A_u = \int_0^t G(t,s)A_s K(s,u)A_u ds + G(t,u)B_u^2, \quad t \ge u \ge 0, \tag{5.5}$$

where $K(t,s) = \mathrm{cov}(X_t,X_s)$. Indeed, by orthogonality property of the orthogonal projection, for any fixed $t\in[0,T]$ and any measurable and bounded deterministic function $\lambda$

$$\mathrm{E}\left(X_t - \widehat{\mathrm{E}}(X_t|\mathcal{L}_t^Y)\right)\int_0^t \lambda_s dY_s = \mathrm{E}\left(X_t - \int_0^t G(t,s)dY_s\right)\int_0^t \lambda_u dY_u = 0.$$

Then (5.5) holds, since

$$\mathrm{E}X_t \int_0^t \lambda_u dY_u = \int_0^t \lambda_u A_u K(t,u)du$$

and

$$\mathrm{E}\int_0^t G(t,s)dY_s \int_0^t \lambda_u dY_u = \int_0^t\int_0^t G(t,s)A_s K(s,u)A_u\lambda_u duds + \int_0^t \lambda_u G(t,u)B_u^2 du$$

for arbitrary $\lambda$. Under the assumption $B_t^2 \ge C > 0$, the Wiener-Hopf equation has a unique solution: suppose it doesn't, i.e. both $G_1(t,s)$ and $G_2(t,s)$ satisfy (5.5) and let $\Delta(t,s) = G_1(t,s) - G_2(t,s)$. Then $\Delta(t,s)$ satisfies

$$\int_0^t \Delta(t,s)A_s K(s,u)A_u ds + \Delta(t,u)B_u^2 = 0, \quad t \ge u \ge 0.$$

Multiply this equation by $\Delta(t,u)$ and integrate with respect to $u$:

$$\int_0^t\int_0^t \Delta(t,u)A_u K(s,u)\Delta(t,s)A_s dsdu + \int_0^t \Delta^2(t,u)B_u^2 = 0.$$

The first term is nonnegative, since the covariance function $K(s,u)$ is nonnegative definite, and thus for $t\in[0,T]$

$$\int_0^t \Delta^2(t,u)B_u^2 = 0 \quad \Longrightarrow \quad \Delta^2(t,u) = 0, \quad du-a.s.$$

<u>Step 4</u> (*solving the Wiener-Hopf equation)*

The uniqueness allows us to look for differentiable $G(t,s)$, since once found it should be *the* solution. Differentiating (5.5) with respect to $t$ one obtains

$$\frac{\partial}{\partial t}K(t,u)A_u = G(t,t)A_tK(t,u)A_u + \int_0^t \frac{\partial}{\partial t}G(t,s)A_sK(s,u)A_u ds + \frac{\partial}{\partial t}G(t,u)B_u^2$$

Recall that (Exercise **10** of the previous chapter)

$$\frac{\partial}{\partial t}K(t,u) = a_tK(t,u), \quad K(u,u) = \mathrm{E}X_u^2$$

and hence the latter equation reads

$$K(t,u)A_u\Big(a_t - G(t,t)A_t\Big) - \int_0^t \frac{\partial}{\partial t}G(t,s)A_sK(s,u)A_u ds - \frac{\partial}{\partial t}G(t,u)B_u^2 = 0.$$

Now using the expression for $K(t,u)A_u$ from (5.5), one gets

$$\left(\int_0^t G(t,s)A_sK(s,u)A_u ds + G(t,u)B_u^2\right)\Big(a_t - G(t,t)A_t\Big) - \int_0^t \frac{\partial}{\partial t}G(t,s)A_sK(s,u)A_u ds - \frac{\partial}{\partial t}G(t,u)B_u^2 = 0.$$

or

$$\int_0^t \Big\{G(t,s)\Big(a_t - G(t,t)A_t\Big) - \frac{\partial}{\partial t}G(t,s)\Big\}A_sK(s,u)A_u ds + \Big\{G(t,u)\Big(a_t - G(t,t)A_t\Big) - \frac{\partial}{\partial t}G(t,u)\Big\}B_u^2 = 0$$

Multiply the latter equality by

$$\Psi(t,u) := G(t,u)\Big(a_t - G(t,t)A_t\Big) - \frac{\partial}{\partial t}G(t,u)$$

and integrate:

$$\int_0^t\int_0^t \Psi(t,s)A_sK(s,u)\Psi(t,u)A_u dsdu + \int_0^t \Psi(t,u)^2B_u^2 du = 0,$$

which gives the differential equation for $G(t,s)$:

$$\frac{\partial}{\partial t}G(t,s) = G(t,s)\Big(a_t - G(t,t)A_t\Big). \tag{5.6}$$

With $u=t$ in (5.5), one gets

$$0 = K(t,t)A_t - A_t\int_0^t G(t,s)A_sK(s,t)ds - G(t,t)B_t^2,$$

which implies

$$\begin{aligned}
0 &= A_t \mathrm{E} X_t \Big( X_t - \int_0^t G(t,s) A_s X_s ds \Big) - G(t,t) B_t^2 = \\
& A_t \mathrm{E} X_t \Big( X_t - \int_0^t G(t,s) dY_s \Big) - G(t,t) B_t^2 \stackrel{\dagger}{=} \\
& A_t \mathrm{E} \Big( X_t - \int_0^t G(t,s) dY_s \Big)^2 - G(t,t) B_t^2 = A_t P_t - G(t,t) B_t^2,
\end{aligned}$$

where the equality † is due to the orthogonality property and $P_t = (X_t - \widehat{X}_t)^2$. Hence the ODE (5.6) reads

$$\frac{\partial}{\partial t} G(t,s) = G(t,s) \Big( a_t - \frac{A_t^2 P_t}{B_t^2} \Big). \tag{5.7}$$

Being a linear equation, the latter admits the representation $G(t,s) = \Phi(s,t) G(s,s)$, where $\Phi(s,t)$ is the Cauchy[2] (or fundamental) solution corresponding to (5.7). Then

$$\widehat{X}_t = \int_0^t G(t,s) dY_s = \int_0^t \Phi(s,t) G(s,s) Y_s = \Phi(0,t) \int_0^t \Phi^{-1}(0,s) G(s,s) dY_s$$

and applying the Itô formula one gets the first equation in (5.3)

$$\begin{aligned}
d\widehat{X}_t = & \int_0^t \Phi^{-1}(0,s) G(s,s) dY_s \frac{\partial}{\partial t} \Phi(0,t) dt + \Phi(0,t) \Phi^{-1}(0,t) G(t,t) dY_t = \\
& \int_0^t \Phi^{-1}(0,s) G(s,s) dY_s \Big( a_t - \frac{A_t^2 P_t}{B_t^2} \Big) \Phi(0,t) dt + G(t,t) dY_t = \\
& a_t \widehat{X}_t dt + \frac{A_t P_t}{B_t^2} \big( dY_t - A_t \widehat{X}_t \big).
\end{aligned}$$

The process $D_t = X_t - \widehat{X}_t$ satisfies

$$\begin{aligned}
dD_t = a_t D_t dt + b_t dW_t - \frac{A_t P_t}{B_t^2} \big( A_t X_t dt + B_t dV_t - A_t \widehat{X}_t \big) = \\
\Big( a_t - \frac{A_t^2 P_t}{B_t^2} \Big) D_t dt + b_t dW_t - \frac{A_t P_t}{B_t} dV_t.
\end{aligned}$$

Applying the Itô formula to $D_t^2$ one gets

$$\begin{aligned}
dD_t^2 = 2 D_t dD_t + b_t^2 dt + \left( \frac{A_t P_t}{B_t} \right)^2 dt = 2 \Big( a_t - \frac{A_t^2 P_t}{B_t^2} \Big) D_t^2 dt + \\
b_t^2 dt + \left( \frac{A_t P_t}{B_t} \right)^2 dt + 2 D_t \left( b_t dW_t - \frac{A_t P_t}{B_t} dV_t \right)
\end{aligned}$$

and taking the expectation

$$dP_t = 2 \Big( a_t - \frac{A_t^2 P_t}{B_t^2} \Big) P_t dt + b_t^2 dt + \left( \frac{A_t P_t}{B_t} \right)^2 dt = 2 a_t dt + b_t^2 dt - \frac{A_t^2 P_t^2}{B_t^2} dt,$$

subject to $P_0 = \mathrm{E}(X_0 - \widehat{X}_0)^2$ (recall the construction of Step 1). □

[2]Since solution of linear equation depends linearly on the initial condition, it can be written as a time dependent linear operator (just multiplication by $\Phi(s,t)$ in this case), acting on the initial condition. The Cauchy operator satisfies $\Phi(0,s)\Phi(s,t) = \Phi(0,t)$ and is invertible.

The Kalman-Bucy filter is a linear SDE with time varying coefficients, which depend on $P_t$, being the solution of the Riccati equation (5.3). The innovation process

$$\bar{W}_t = \int_0^t \frac{dY_s - A_s\widehat{X}_s ds}{B_s}$$

has uncorrelated increments and in the case of Gaussian $(X_0, Y_0)$ is a Wiener process (!), with respect to the filtration $\mathcal{F}_t^Y$ (this is worked out in details in the next chapter, dealing with nonlinear filtering).

EXAMPLE 5.2. Consider the system (5.1)-(5.2) with constant coefficients: $a_t \equiv a$, etc. and subject to a random square integrable $X_0$ and $Y_0 = 0$. The Kalman-Bucy filter in this case is

$$\begin{aligned} \widehat{X}_t &= a\widehat{X}_t dt + \frac{P_t A}{B^2}\big(dY_t - A\widehat{X}_t dt\big) \\ \dot{P}_t &= 2aP_t + b^2 - \frac{A^2 P_t^2}{B^2} \end{aligned} \tag{5.8}$$

subject to $\widehat{X}_0 = \mathrm{E}X_0$ and $P_0 = \mathrm{E}(X_0 - \mathrm{E}X_0)^2$.

Consider the quadratic equation

$$2aP + b^2 - A^2P^2/B^2 = 0. \tag{5.9}$$

If $A \neq 0$ and $b \neq 0$ are assumed, then it has two solutions

$$P_\pm = \frac{B^2}{A^2}\Big(a \pm \sqrt{a^2 + \frac{A^2b^2}{B^2}}\Big),$$

with $P_- < 0$ and $P_+ > 0$. Consider the suboptimal filter

$$\widetilde{X}_t = a\widetilde{X}_t dt + \frac{AP_+}{B^2}\big(Y_t - A\widetilde{X}_t dt\big), \quad \widetilde{X}_0 = 0.$$

The error process $\delta_t = X_t - \widetilde{X}_t$, satisfies

$$d\delta_t = \Big(a - \frac{A^2P_+}{B^2}\Big)\delta_t dt + bdW_t + \frac{AP_+}{B}dV_t, \quad \delta_0 = X_0.$$

Since

$$a - \frac{A^2P_+}{B^2} = a - \Big(a + \sqrt{a^2 + \frac{A^2b^2}{B^2}}\Big) = -\sqrt{a^2 + \frac{A^2b^2}{B^2}} < 0, \tag{5.10}$$

the mean square error of this filter is bounded: $\sup_{t\ge 0}\mathrm{E}\delta_t^2 < \infty$ and thus by optimality of $\widehat{X}_t$

$$\sup_{t\ge 0} P_t \le \mathrm{E}\delta_t^2 < \infty.$$

The function $R_t := P_t - P_+$, satisfies

$$\dot{R}_t = 2aR_t - \frac{A^2}{B^2}\big(P_t^2 - P_+^2\big) = 2aR_t - \frac{A^2}{B^2}R_t\big(P_t + P_+\big)$$

and hence

$$\begin{aligned} |R_t| = |R_0|\exp\Big\{2at - \frac{A^2}{B^2}\int_0^t \big(P_s + P_+\big)ds\Big\} &\le |R_0|\exp\Big\{2at - \frac{A^2}{B^2}P_+ t\Big\} \\ &= |R_0|\exp\Big\{at - \sqrt{a^2 + \frac{A^2b^2}{B^2}}t\Big\} \xrightarrow{t\to\infty} 0, \end{aligned}$$

due to (5.10). In other words, if $A \neq 0$ and $b \neq 0$, the solution of the Riccati equation stabilizes and the limit mean square error $P_\infty = \lim_{t\to\infty} P_\infty$ equals the unique positive solution of the algebraic Riccati equation (5.9). If $A = 0$ and $b \neq 0$, then $P_t = \mathrm{E}(X_t - \mathrm{E}X_t)^2$ and the limit $P_\infty$ exists and is finite if $a < 0$, otherwise $P_t$ grows to infinity. Finally if $b = 0$ and $A \neq 0$, then $P_\infty = 0$, either if $a < 0$ (since $X_t \to 0$ in $\mathbb{L}^2$) or if $a > 0$ (since then $a/Ae^{-at}Y_t \to X_0$ in $\mathbb{L}^2$) or if $a = 0$ (since $A^{-1}Y_t/t \to X_0$ in $\mathbb{L}^2$).

Unlike in the discrete time case, the scalar Riccati equation in (5.8) has an explicit solution:

$$P_t = \frac{\alpha_- - K\alpha_2 \exp\Big(\frac{(\alpha_+-\alpha_-)A^2 t}{B^2}\Big)}{1 - K\exp\Big(\frac{(\alpha_+-\alpha_-)A^2 t}{B^2}\Big)}, \tag{5.11}$$

where

$$\alpha_\pm = A^{-2}\big(aB^2 \pm B\sqrt{a^2B^2 + A^2b^2}\big), \quad K = \frac{P_0 - \alpha_-}{P_0 - \alpha_+}.$$

■

## 2. The Kalman-Bucy filter: the general case

In this section we give the general formulation of linear filtering problem and the corresponding Kalman-Bucy equations. The proof uses the very same arguments as in the scalar case and is left as an exercise. Let $X = (X_t)_{t\in[0,T]}$ and $Y = (Y_t)_{t\in[0,T]}$ be the process with values in $\mathbb{R}^m$ and $\mathbb{R}^n$, generated by the system of linear SDEs

$$dX_t = \big(a_0(t) + a_1(t)X_t + a_2(t)Y_t\big)dt + b_1(t)dW_t + b_2(t)dV_t \tag{5.12}$$

$$dY_t = \big(A_0(t) + A_1(t)X_t + A_2(t)Y_t\big)dt + B_1(t)dW_t + B_2(t)dW_t, \tag{5.13}$$

with respect to independent vector Wiener processes $W$ and $V$ and subject to a square integrable random vector $(X_0, Y_0)$ independent of $(W, V)$. The coefficients are deterministic matrix functions of appropriate dimensions, such that the unique strong solution of the system exists[3] and $(B\circ B)(t) := B_1B_1^* + B_2B_2^*$ is uniformly nonsingular matrix.

THEOREM 5.3. *The the orthogonal projection $\widehat{X}_t = \widehat{\mathrm{E}}(X_t|\mathcal{L}_t^Y)$ and the corresponding error covariance matrix $P_t = \mathrm{E}\big(X_t - \widehat{X}_t\big)\big(X_t - \widehat{X}_t\big)^*$ satisfy the Kalman-Bucy equations*[4]

$$\begin{aligned} d\widehat{X}_t =& \big(a_0 + a_1\widehat{X}_t dt + a_2\widehat{Y}_t\big)dt + \big(b\circ B + P_tA_1^*\big)\big(B\circ B\big)^{-1}\cdot \\ & \qquad\qquad \big(dY_t - (A_0 - A_1\widehat{X}_t - A_2Y_t)dt\big) \end{aligned} \tag{5.14}$$

$$\dot{P}_t = a_1P_t + P_ta_1^* + b\circ b - \big(b\circ B + P_tA_1^*\big)\big(B\circ B\big)^{-1}\big(b\circ B + P_tA_1^*\big)^* \tag{5.15}$$

*subject to*

$$\widehat{X}_0 = \mathrm{E}X_0 - \mathrm{cov}(X_0,Y_0)\,\mathrm{cov}^{\oplus}(Y_0)(Y_0 - \mathrm{E}Y_0),$$

$$P_0 = \mathrm{cov}(X_0) - \mathrm{cov}(X_0,Y_0)\,\mathrm{cov}^{\oplus}(Y_0)\,\mathrm{cov}(Y_0,X_0)$$

*and where*

$$b\circ B = b_1B_1^* + b_2B_2^*, \quad b\circ b = b_1b_1^* + b_2b_2^*.$$

---

[3]for example if the drift coefficients are integrable and the diffusion coefficients are square integrable functions of $t$ with respect to the Lebesgue measure.

[4]the time dependence of the coefficients is omitted for brevity

## 3. Linear filtering beyond linear diffusions

The Kalman-Bucy filtering formulae are applicable in somewhat more general setting than (5.1)-(5.2) (or (5.12)-(5.13)).

DEFINITION 5.4. $w_t$ is a Wiener process in wide sense, if $w_0 = 0$, $\mathrm{E}w_t = 0$ and $\mathrm{E}w_t w_s = s \wedge t$, $t, s \geq 0$.

EXAMPLE 5.5. The stochastic integral $w_t = \int_0^t X_s/\sqrt{\mathrm{E}X_s^2} dW_s$ with a positive process $X_t \geq C > 0$ is a Wiener process in the wide sense:

$$\mathrm{E}w_t w_s = \int_{t \wedge s} \mathrm{E}\left(\frac{X_u}{\sqrt{\mathrm{E}X_u^2}}\right)^2 du = t \wedge s.$$

■

Since $w_t$ has uncorrelated increments, one may define the stochastic integral

$$I_t(f) = \int_0^t f_s dw_s := \lim_{n\to\infty} \sum_{i=1}^n f_{t_{i-1}}\big(w_{t_i} - w_{t_{i-1}}\big),$$

where $f$ is an $\mathbb{L}^2_{[0,T]}$ deterministic function and $0 = t_0 < ... < t_n = T$, such that $\max_i |t_i - t_{i-1}| \to 0$ as $n \to \infty$ (by construction similar to the Itô integral).

Since the linear SDE

$$dX_t = a_t X_t dt + b_t dW_t,$$

has an explicit solution

$$X_t = \exp\left\{\int_0^t a_u du\right\}\left(X_0 + \int_0^t \exp\left\{-\int_0^s a_u du\right\} b_s dW_s\right),$$

analogously one may define the process

$$X_t = \exp\left\{\int_0^t a_u du\right\}\left(X_0 + \int_0^t \exp\left\{-\int_0^s a_u du\right\} b_s dw_s\right)$$

to be the solution of

$$dX_t = a_t X_t dt + b_t dw_t.$$

With these definitions it is almost obvious that the Kalman-Bucy filtering equations generate the optimal linear estimates, if the Wiener processes are replaced by the Wiener processes in the wide sense. Let's demonstrate the application of this generalization in the following example:

EXAMPLE 5.6. Consider the SDE system

$$\begin{aligned} dX_t &= -X_t dt + dW_t \\ dY_t &= X_t^3 dt + dV_t \end{aligned} \tag{5.16}$$

subject to random $X_0$ with zero mean and $\mathrm{E}X_0^2 = 1/2$, $Y_0 = 0$. By the Itô formula

$$dX_t^3 = 3X_t^2 dX_t + 3X_t dt = -3X_t^3 dt + 3X_t dt + 3X_t^2 dW_t.$$

Define $Z_t = X_t^3$ and

$$w_t = \sqrt{2}\int_0^t X_s^2 dW_s - \frac{W_t}{\sqrt{2}}.$$

Then $w_t$ is the Wiener process in the wide sense ($t \geq s$):

$$\mathrm{E}w_t w_s = \mathrm{E}\left(\sqrt{2}\int_0^s X_u^2 dW_u - \frac{W_s}{\sqrt{2}}\right)^2 =$$
$$2\int_0^s \mathrm{E}X_u^4 du + \frac{s}{2} - 2\int_0^s \mathrm{E}X_u^2 du = 2\frac{3}{4}s + \frac{s}{2} - s = s,$$

where the Gaussian property of $X_t$ have been used ($\mathrm{E}X_t^2 = 1/2$, $\mathrm{E}X_t^4 = 3(\mathrm{E}X_t^2)^2 = 3/4$, etc.). Analogously

$$\mathrm{E}w_t W_t = \mathrm{E}\left(\sqrt{2}\int_0^t X_u^2 dW_u - \frac{W_t}{\sqrt{2}}\right) W_t = \sqrt{2}t\mathrm{E}X_t^2 - \frac{t}{\sqrt{2}} = 0.$$

So $(w_t, W_t, V_t)$ is a three-dimensional Wiener process in wide sense. Consider now the linear system

$$\begin{aligned} dX_t &= -X_t dt + dW_t \\ dZ_t &= -3Z_t dt + 3X_t dt + \frac{3}{\sqrt{2}} dw_t + \frac{3}{2} dW_t \\ dY_t &= Z_t dt + dV_t, \end{aligned} \tag{5.17}$$

subject to $(X_0, Z_0) = (X_0, X_0^3)$ (i.e. $\mathrm{E}Z_0 = 0$, $\mathrm{E}Z_0^2 = \mathrm{E}X_0^6 = 15/8$, etc.). The estimate $\mathrm{E}(X_t|\mathcal{L}_t^Y)$ can be obtained by means of the Kalman-Bucy equations for (5.17).

■

## Exercises

(**1**) Verify that if $X_0$ and $Y_0$ are such that $\widehat{\mathrm{E}}(X_0|Y_0) = 0$, P-a.s. in the model (5.1)-(5.2), then $\mathrm{E}X_t = 0$ and $\widehat{\mathrm{E}}(X_t|Y_0) = 0$, P-a.s.

(**2**) Show that the innovation process

$$\bar{W}_t = B^{-1}\int_0^t (dY_s - A\widehat{X}_s ds)$$

satisfies the following properties ($t \geq s \geq 0$)

(a) $\widehat{\mathrm{E}}\big(\bar{W}_t|\mathcal{L}_s^Y\big) = \bar{W}_s$

(b) $\mathrm{E}\big(\bar{W}_t - \bar{W}_s\big)^2 = t - s$

(c) Derive the Kalman-Bucy equations, assuming that $\bar{W}$ is a Wiener process (in the wide sense) and that $\widehat{\mathrm{E}}(X_t|\mathcal{L}_t^Y) = \int_0^t \Gamma(t,s)d\bar{W}_s$ for some $\Gamma(t,s)$.

(**3**) Let $Y_t = \int_0^t W_s ds + V_t$, where $W$ and $V$ are independent Wiener processes.

(a) Find the optimal linear filter for $\widehat{W_t} = \widehat{\mathrm{E}}(W_t|\mathcal{L}_t^Y)$

(b) Find the explicit form for the optimal kernel $G(t,s)$, such that

$$\widehat{W_t} = \int_0^t G(t,s)dY_s.$$

**Hint:** use the explicit solution (5.11).

(c) Derive the equation for linear estimate $\widehat{V_t} = \widehat{\mathrm{E}}(V_t|\mathcal{L}_t^Y)$.

**Hint:** use the two dimensional formulae of Theorem (5.3)).

(**4**) Derive the equations (11), claimed in the Introduction (page 12).

(**5**) Prove that the equations (5.3) have the unique strong solution.

(**6**) Reformulate and solve the problem (**8**) (page 32) in continuous time
(**7**) Reformulate and solve the problem (**9**) (page 33) in continuous time

CHAPTER 6

# Nonlinear filtering in continuous time

In this chapter the two main approaches to nonlinear filtering problem in continuous time are presented. The first one relies on the representation of the conditional expectation as a stochastic integral with respect to the *innovation* Wiener process. The second one uses the abstract version of the Bayes formula, involving the Girsanov change of measure to define a *reference* probability, under which the dependence between the signal and the observations is cancelled and thus the calculations are carried out in a particularly simple way. This approach gives an additional insight into the structure of FKK equation: it turns out that its solution is a normalized version of the measure valued stochastic process, generated by a linear Zakai equation.

As in the discrete time case, both approaches lead to measure valued equations which at best characterize the conditional law of the signal given the observation $\sigma$-algebra. Remarkably for certain particular systems the filtering process turns to be finite dimensional, i.e. can be parameterized by a finite number of computable parameters. For example, Kalman-Bucy filtering equations turn to be the finite dimensional parametrization in the linear Gaussian case.

## 1. The innovation approach

The typical filtering problem in continuous time is to find a recursive realization for the conditional expectation of the signal Markov process at the current time, given the past of its noisy trajectory. Let's consider the following general framework of this problem: let $(X,Y) = (X_t, Y_t)_{t\in[0,T]}$ be supported on a stochastic basis $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathrm{P})$ and satisfy the following assumptions:

(a) $X$ admits the decomposition

$$X_t = X_0 + \int_0^t H_s ds + M_t, \tag{6.1}$$

where $(M_t, \mathcal{F}_t)$ is a martingale[1] and $H_t$ is an $\mathcal{H}^2_{[0,T]}$-process.

[1]As mentioned before, the definition of the stochastic integral can be extended to martingales, more general than Wiener process. In this introductory course we don't really need this generality. In fact $M_t$ will be either a stochastic integral with respect to Wiener process or a Poisson like jump processes

(b) $Y$ is the Itô process, satisfying[2]

$$Y_t = \int_0^t A_s ds + BW_t, \tag{6.2}$$

where $A$ is an $\mathcal{H}^2_{[0,T]}$ process, $B > 0$ is a fixed constant and $W$ is a Wiener process, independent of $X$.

The following generic notation will be used throughout: $\pi_t(\xi) = \mathrm{E}\big(\xi_t|\mathcal{F}^Y_t\big)$ for a process $\xi = (\xi_t)_{t\in[0,T]}$, where $\mathcal{F}^Y_t$ is the natural filtration of $Y$.

**1.1. The innovation Wiener process.** The innovation process $\bar{W}$ was already encountered in the Kalman-Bucy filtering setting.

THEOREM 6.1. *The process $Y$, satisfying* (b)*, admits the representation*

$$Y_t = Y_0 + \int_0^t \pi_s(A)ds + B\bar{W}_t, \tag{6.3}$$

*where*

$$\bar{W}_t = B^{-1}(Y_t - \int_0^t \pi_s(A)ds). \tag{6.4}$$

*is a Wiener process with respect to $\mathcal{F}^Y_t$.*

PROOF. Clearly $\bar{W}$ has continuous trajectories, starting at zero. For brevity let $B = 1$, then

$$\bar{W}_t = W_t + \int_0^t \big(A_s - \pi_s(A)\big)ds.$$

Show that

$$\mathrm{E}\Big(e^{i\lambda(\bar{W}_t-\bar{W}_s)}|\mathcal{F}^Y_t\Big) = e^{-\frac{1}{2}\lambda^2(t-s)}. \tag{6.5}$$

Applying the Itô formula to $\eta_t = \exp\big\{i\lambda\bar{W}_t\big\}$ one gets

$$d\eta_t = i\lambda\eta_t d\eta_t - \frac{1}{2}\lambda^2\eta_t dt = i\lambda\eta_t dW_t + i\lambda\eta_t\big(A_t - \pi_t(A)\big)dt - \frac{1}{2}\lambda^2\eta_t dt$$

and hence

$$e^{i\lambda\bar{W}_t} = e^{i\lambda\bar{W}_s} + i\lambda\int_s^t e^{i\lambda\bar{W}_u}dW_u + \\ i\lambda\int_s^t e^{i\lambda\bar{W}_u}\big(A_u - \pi_u(A)\big)du - \frac{1}{2}\lambda^2\int_s^t e^{i\lambda\bar{W}_u}dt$$

Since $W$ is a Wiener process with respect to the filtration $\mathcal{F}^W_t \vee \mathcal{F}^Y_t$,

$$\mathrm{E}\left(\int_s^t e^{i\lambda\bar{W}_u}dW_u\Big|\mathcal{F}^Y_s\right) = 0.$$

[2]With an additional effort, the diffusion coefficient $B$ can be allowed to depend on $Y$ and time $t$. The essential requirement is then $B^2_t(Y) \geq C > 0$, which prevents the filtering problem from being singular. Also note that if $B$ is allowed to depend on the signal $X$, the filtering problem becomes ill-posed. For example, if $B(x) = x$, $x \in \mathbb{R}$, then $X^2_t$ can be recovered from the quadratic variation of $Y$ and thus $X^2_t$ is $\mathcal{F}^Y_t$-measurable, i.e. known up to its sign. These situations are customary taboo in filtering

Note that for $u \geq s$

$$\mathrm{E}\Big(e^{i\lambda \bar W_u}\pi_u(A)\big|\mathcal{F}^Y_s\Big) = \mathrm{E}\Big(e^{i\lambda \bar W_u}\mathrm{E}(A_u|\mathcal{F}^Y_u)\big|\mathcal{F}^Y_s\Big) =$$
$$\mathrm{E}\Big(\mathrm{E}(A_u e^{i\lambda \bar W_u}|\mathcal{F}^Y_u)\big|\mathcal{F}^Y_s\Big) = \mathrm{E}\big(A_u e^{i\lambda \bar W_u}|\mathcal{F}^Y_s\big)$$

and thus

$$\mathrm{E}\left(\int_s^t e^{i\lambda \bar W_u}\big(A_u - \pi_u(A)\big)du\Big|\mathcal{F}^Y_s\right) = 0.$$

Then $\eta_t = \mathrm{E}(e^{i\lambda \bar W_t}|\mathcal{F}^Y_s)$ satisfies

$$\eta_t = \eta_s - \frac{1}{2}\lambda^2\int_s^t \eta_u du,$$

which verifies (6.5). □

REMARK 6.2. Note that $\bar W$ need not be (and in general is not) a Wiener process with respect to other filtrations, e.g. $\mathcal{F}^W$.

REMARK 6.3. Note that the equation (6.6) is driven not by the observation process $Y$ itself, but rather by a Wiener process, generated by $Y$. Loosely speaking, this Wiener process is a minimal representation of the information carried by $Y$, sufficient for estimation of $X$, which is the origin of the term "*innovation*". Clearly $\mathcal{F}^{\bar W}_t \subseteq \mathcal{F}^Y_t$, since $\bar W_t$ is a measurable functional of $Y$ on $[0,t]$ or in other words, the information carried by $\bar W$ is less than information carried by $Y$. Naturally the question arises: does $\bar W_t$ encodes all the information, i.e. $\mathcal{F}^Y_t \subseteq \mathcal{F}^{\bar W}_t$ ? The answer to this question is affirmative if the SDE (6.3) has a strong solution. However, in view of the Tsirelson's counterexample, mentioned in Remark 4.37, the latter is not at all clear. Some positive results in this direction can be found in Section 12.2 in [**21**].

REMARK 6.4. Recall the statement of the Girsanov theorem: given a Wiener process $(W_t,\mathcal{F}_t)$ on a fixed probability basis $(\Omega,\mathcal{F},\mathcal{F}_t,\mathrm{P})$, there is a probability $\widetilde{\mathrm{P}}$ on $(\Omega,\mathcal{F})$, equivalent to P and such that the process, obtained by shifting $W$ by a random process with sufficiently smooth trajectories (absolutely continuous with respect to the Lebesgue measure), is again a Wiener process with respect to $\mathcal{F}_t$ under $\widetilde{\mathrm{P}}$. On the other hand, the innovations (6.4)

$$\bar W_t = W_t + \int_0^t \big(A_s - \pi_s(A)\big)ds$$

exhibit a different phenomenon: $W$ shifted by a special function becomes a Wiener process under the original measure P but with respect to another filtration $\mathcal{F}^Y_t$!

**1.2. Fujisaki-Kallianpur-Kunita equation.** Using the innovation form of $Y$ and the martingale representation theorem an equation for the measure valued filtering process $\pi_t(\cdot)$ is derived below.

THEOREM 6.5. *Assume* (a) *and* (b)*, then* $\pi_t(X)$ *satisfies satisfies the Fujisaki-Kallianpur-Kunita (FKK) equation: for any* $t\in[0,T]$ P-*a.s.*

$$\pi_t(X) = \pi_0(X) + \int_0^t \pi_s(H)ds + \int_0^t \Big(\pi_s(AX) - \pi_s(A)\pi_s(X)\Big)B^{-1}d\bar W_t, \qquad (6.6)$$

*where* $(\bar W_t,\mathcal{F}^Y_t)$ *is the innovation Wiener process defined in* (6.4).

Remark 6.6. FKK equation (6.6) is a measure valued equation: its (strong) solution, say $\pi_t(dx)$, can be defined as a stochastic process taking values in the space of probability measures on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$, adapted to $\mathcal{F}_t^Y$ and satisfying (6.6) with probability one. For example, if the process $\pi_t(dx)$ has a density, (6.6) can be used to derive an equation for the conditional density process (Kushner-Stratonovich equation (6.13)). The existence and uniqueness of the strong solution is not an easy issue.

Proof. The filtering process admits the following decomposition

$$\pi_t(X) = \pi_0(X) + \int_0^t \pi_s(H)ds + \bar{M}_t, \quad t \in [0, T], \tag{6.7}$$

where

$$\bar{M}_t := \mathrm{E}(X_0|\mathcal{F}_t^Y) - \pi_0(X) + \mathrm{E}\left(\int_0^t H_s ds|\mathcal{F}_t^Y\right) - \int_0^t \pi_s(H)ds + \mathrm{E}(M_t|\mathcal{F}_t^Y).$$

is a square integrable $\mathcal{F}_t^Y$-martingale. The square integrability of each component follows from the assumptions on $X$ and the martingale property is verified as follows: the first term is a martingale, since $(t \geq s \geq 0)$

$$\mathrm{E}\Big(\mathrm{E}(X_0|\mathcal{F}_t^Y) - \pi_0(X)|\mathcal{F}_s^Y\Big) = \mathrm{E}(X_0|\mathcal{F}_s^Y) - \pi_0(X).$$

The second one satisfies

$$\begin{aligned}
&\mathrm{E}\left(\mathrm{E}\Big(\int_0^t H_u du|\mathcal{F}_t^Y\Big) - \int_0^t \pi_u(H)du\Big|\mathcal{F}_s^Y\right) = \\
&\int_0^t \mathrm{E}\big(H_u|\mathcal{F}_s^Y\big)du - \int_0^t \mathrm{E}\big(\pi_u(H)|\mathcal{F}_s^Y\big)du = \\
&\mathrm{E}\left(\int_0^s H_u du|\mathcal{F}_s^Y\right) - \int_0^s \pi_u(H)du + \int_s^t \mathrm{E}\big(H_u|\mathcal{F}_s^Y\big)du - \int_s^t \mathrm{E}\big(\pi_u(H)\big|\mathcal{F}_s^Y\big)du = \\
&\mathrm{E}\left(\int_0^s H_u du|\mathcal{F}_s^Y\right) - \int_0^s \pi_u(H)du
\end{aligned}$$

and thus is also a martingale. Finally the third term inherits martingale properties from $M_t$:

$$\mathrm{E}\big(\mathrm{E}(M_t|\mathcal{F}_t^Y)|\mathcal{F}_s^Y\big) = \mathrm{E}(M_t|\mathcal{F}_s^Y) = \mathrm{E}\big(\mathrm{E}(M_t|\mathcal{F}_s)\big|\mathcal{F}_s^Y\big) = \mathrm{E}(M_s|\mathcal{F}_s^Y).$$

Since $Y_t$ is an Itô process, generated by (6.3), where $\bar{W}_t$ is a Wiener process, by Theorem 4.51, being a square integrable $\mathcal{F}_t^Y$-martingale, $\bar{M}_t$ has the representation

$$\bar{M}_t = \int_0^t g_s(Y)d\bar{W}_s,$$

with $g_s$ being $\mathcal{F}_t^Y$-adapted process. To verify (6.6) one should show that

$$g_s(Y) = \big(\pi_s(AX) - \pi_s(A)\pi_s(X)\big)/B, \quad ds \times \mathrm{P} - a.s., \tag{6.8}$$

which is equivalent to

$$\int_0^t \mathrm{E}\lambda_s(Y)\Big(g_s(Y) - \big(\pi_s(AX) - \pi_s(A)\pi_s(X)\big)/B\Big)ds = 0, \tag{6.9}$$

for any bounded $\mathcal{F}_t^Y$-adapted[3] $\lambda_s(Y)$.

Let $z_t = \int_0^t \lambda_s(Y)d\bar{W}_s$ and $\xi_t = \int_0^t g_s(Y)d\bar{W}_s$, then

$$\int_0^t \mathrm{E}\lambda_s(Y)g_s(Y)ds = \mathrm{E}z_t\xi_t. \tag{6.10}$$

On the other hand,

$$\mathrm{E}z_t\xi_t = \mathrm{E}z_t\left(\pi_t(X) - \pi_0(X) - \int_0^t \pi_s(H)ds\right) = \mathrm{E}\left(z_t X_t - \int_0^t z_s H_s ds\right),$$

since $\mathrm{E}z_t\pi_0(X) = \mathrm{E}\pi_0(X)\mathrm{E}(z_t|\mathcal{F}_0^Y) = 0$, $\mathrm{E}z_t\pi_t(X) = \mathrm{E}z_t\mathrm{E}\big(X_t|\mathcal{F}_t^Y\big) = \mathrm{E}z_t X_t$ and

$$\mathrm{E}z_t\int_0^t \pi_s(H)ds = \mathrm{E}\int_0^t \mathrm{E}(z_t|\mathcal{F}_s^Y)\pi_s(H)ds = \\ \int_0^t z_s\pi_s(H)ds = \int_0^t \mathrm{E}(z_s H_s|\mathcal{F}_s^Y)ds = \mathrm{E}\int_0^t z_s H_s ds.$$

Using the definition of $\bar{W}$

$$z_t = \int_0^t \lambda_s dW_s + \int_0^t \lambda_s \frac{A_s - \pi_s(A)}{B}ds.$$

Then

$$\mathrm{E}z_t\xi_t = \mathrm{E}\left(X_t\int_0^t \lambda_s dW_s - \int_0^t \Big(\int_0^s \lambda_u dW_u\Big)H_s ds\right) + \\ \mathrm{E}\left(X_t\int_0^t \lambda_s\frac{A_s - \pi_s(A)}{B}ds - \int_0^t \Big(\int_0^s \lambda_u\frac{A_u - \pi_u(A)}{B}du\Big)H_s ds\right) \tag{6.11}$$

We claim that the first expectation vanishes: indeed

$$\mathrm{E}X_0\int_0^t \lambda_s(Y)dW_s = \mathrm{E}X_0\mathrm{E}\Big(\int_0^t \lambda_s(Y)dW_s|\mathcal{F}_0\Big) = 0$$

and

$$\mathrm{E}\int_0^t \Big(\int_0^s \lambda_u dW_u\Big)H_s ds = \mathrm{E}\int_0^t \mathrm{E}\Big(\int_0^t \lambda_u dW_u\Big|\mathcal{F}_s\Big)H_s ds = \\ \mathrm{E}\int_0^t \mathrm{E}\Big(H_s\int_0^t \lambda_u dW_u\Big|\mathcal{F}_s\Big)ds = \mathrm{E}\int_0^t \lambda_u dW_u\int_0^t H_s ds$$

and hence

$$\mathrm{E}\left(X_t\int_0^t \lambda_s dW_s - \int_0^t \Big(\int_0^s \lambda_u dW_u\Big)H_s ds\right) = \\ \mathrm{E}\int_0^t \lambda_s dW_s\left(X_t - X_0 - \int_0^t H_s ds\right) = \mathrm{E}\int_0^t \lambda_s dW_s M_t = 0,$$

where the latter equality holds[4] since the martingale $M$ is independent of $W$.

---

[3]if $\alpha$ is $\mathcal{F}_t^Y$-adapted and satisfies $\int_0^t \mathrm{E}\beta_s\alpha_s ds = 0$ for any bounded $\mathcal{F}_t^Y$-adapted $\beta$, then with particular $\beta_t = \mathrm{sign}(\alpha_t)$ one gets $\int_0^t \mathrm{E}|\alpha_s|ds = 0$ and so $\alpha_s = 0$ $ds \times$ P-a.s. on $[0,t]$.

[4]verify this claim when $M_t$ is another Wiener process, independent of $W$. By the way, $M$ and $W$ can be assumed to be correlated and then the correlation will enter the filtering formula (6.6) at this point.

Consider the first term in the second expectation in the right hand side of (6.11):

$$
\begin{aligned}
&\mathrm{E}X_t\int_0^t\lambda_s\frac{A_s-\pi_s(A)}{B}ds=\\
&\mathrm{E}\int_0^t\lambda_s\frac{X_s\big(A_s-\pi_s(A)\big)}{B}ds+\mathrm{E}\int_0^t\lambda_s(X_t-X_s)\frac{A_s-\pi_s(A)}{B}ds=\\
&\mathrm{E}\int_0^t\lambda_s\frac{\pi_s(XA)-\pi_s(X)\pi_s(A)\big)}{B}ds+\mathrm{E}\int_0^t\lambda_s(M_t-M_s)\frac{A_s-\pi_s(A)}{B}ds+\\
&\mathrm{E}\int_0^t\lambda_s\int_s^t H_u du\frac{A_s-\pi_s(A)}{B}ds=\\
&\mathrm{E}\int_0^t\lambda_s\frac{\pi_s(XA)-\pi_s(X)\pi_s(A)\big)}{B}ds+\mathrm{E}\int_0^t H_s\left(\int_0^s\lambda_u\frac{A_u-\pi_u(A)}{B}du\right)ds
\end{aligned}
$$

Assembling all parts together we obtain

$$
\mathrm{E}z_t\xi_t=\int_0^t\mathrm{E}\lambda_s\frac{\pi_s(XA)-\pi_s(X)\pi_s(A)\big)}{B}ds
$$

which along with (6.10) implies (6.8). □

**1.3. Kushner-Stratonovich equation for conditional density.** The FKK equation (6.6) takes a somewhat more concrete form in the case when $(X_t, Y_t)$ are diffusion processes, namely the (strong) solution of SDE[5]

$$
\begin{aligned}
dX_t &= a(X_t)dt + b(X_t)dV_t, &\qquad X_0 = \xi,\\
dY_t &= A(X_t)dt + BW_t, & Y_0 = 0
\end{aligned}
\tag{6.12}
$$

where $\xi$ is a random variable with probability density $p_0(x)$, independent of the Wiener processes $V$ and $W$.

THEOREM 6.7. *Assume that there is an $\mathcal{F}_t^Y$-adapted random field[6] $q_t(x)$, satisfying the Kushner-Stratonovich stochastic partial integral-differential equation*

$$
q_t(x)=p_0(x)+\int_0^t\big(\mathcal{L}^*q_s\big)(x)ds+B^{-1}\int_0^t q_s(x)\big(A(x)-\pi_s(A)\big)d\bar{W}_s \tag{6.13}
$$

*where $\mathcal{L}^*$ is defined in* (4.25) *and*

$$
\pi_t(A)=\int_{\mathbb{R}}A(x)q_t(x)dx.
$$

*Then $q_t(x)$ is a version of the conditional density of $X_t$ given $\mathcal{F}_t^Y$, i.e. for any bounded function $\varphi$*

$$
\mathrm{E}\big(\varphi(X_t)|\mathcal{F}_t^Y\big)=\int_{\mathbb{R}}\varphi(x)q_t(x)dx.
$$

---

[5]Hereon $Y_0 = 0$ is usually set for brevity

[6]by random field we mean a random process, parameterized by time variable $t$ and space variable $x$. All the usual properties (e.g. adaptedness) are assumed to be satisfied uniformly in $x$. In our case sufficient smoothness (e.g. twice differentiability) in $x$ is required.

PROOF. Verify that $q_t(x)$ is a solution of (6.6) and thus is a version of the required conditional expectation. For any twice continuously differentiable function $f$,

$$f(X_t) = f(X_0) + \int_0^t (\mathcal{L}f)(X_s)ds + \int_0^t f'(X_s)b(X_s)dV_s, \quad t \in [0,T],$$

where $\mathcal{L}$ is the backward Kolmogorov operator

$$\big(\mathcal{L}f\big)(x) = a(x)\frac{\partial}{\partial x}f(x) + \frac{b^2(x)}{2}\frac{\partial^2}{\partial x^2}f(x). \tag{6.14}$$

Then the random measure $\pi_t(dx) = q_s(x)dx$ satisfies FKK equation (6.6) for $f(X_t)$ with arbitrary $f$:

$$\pi_s\big((\mathcal{L}f)(X)\big) = \int_{\mathbb{R}} \Big(a(x)\frac{\partial}{\partial x}f(x) + \frac{b^2(x)}{2}\frac{\partial^2}{\partial x^2}f(x)\Big)q_s(x)dx =$$
$$\int_{\mathbb{R}} \Big(-\frac{\partial}{\partial x}a(x)q_s(x) + \frac{1}{2}\frac{\partial^2}{\partial x^2}b^2(x)q_s(x)\Big)f(x)dx = \int_{\mathbb{R}} \big(\mathcal{L}^* q_s\big)(x)f(x)dx \tag{6.15}$$

and

$$\pi_s(fA) - \pi_s(f)\pi_s(A) = \int_{\mathbb{R}} f(x)A(x)q_s(x)dx - \pi_s(A)\int_{\mathbb{R}} f(x)q_s(x)dx =$$
$$\int_{\mathbb{R}} f(x)q_s(x)\Big(A(x) - \pi_s(A)\Big)dx.$$

Then the right hand side of (6.6) reads

$$\pi_0(f) + \int_0^t \pi_s\big(\mathcal{L}f\big)ds + B^{-1}\int_0^t \big(\pi_s(fA) - \pi_s(f)\pi_s(A)\big)d\bar{W}_s =$$
$$\int_{\mathbb{R}} f(x)\left(p_0(x) + \int_0^t \big(\mathcal{L}^* q_s\big)(x)ds + B^{-1}\int_0^t q_s(x)\Big(A(x) - \pi_s(A)\Big)d\bar{W}_s\right)dx =$$
$$\int_{\mathbb{R}} f(x)q_t(x)dx,$$

where (6.13) has been used. □

REMARK 6.8. Due to complicated structure of (6.13), the assumption of the Theorem 6.7 are not easy to verify.

## 2. Reference measure approach

The nonlinear filtering equation can be derived by the Girsanov change of measure. For the clarity of presentation, we chose a specific form of $A_s$ in (6.2):

$$dY_t = \int_0^t g(s, X_s)ds + BW_t, \tag{6.16}$$

where $g$ is a measurable $\mathbb{R}_+ \times \mathbb{R} \mapsto \mathbb{R}$ function.

### 2.1. Kallianpur-Striebel formula.

THEOREM 6.9. *(Kallianpur-Striebel formula) Assume that $g(s, X_s)$ is an $\mathcal{H}^2_{[0,T]}$ process and $Y$ satisfies* (6.16). *Let $(\check{\Omega}, \check{\mathcal{F}}, \check{\mathrm{P}})$ be an auxiliary copy of $(\Omega, \mathcal{F}, \mathrm{P})$, then for any bounded and measurable function $f : \mathbb{R} \mapsto \mathbb{R}$*

$$\mathrm{E}\big(f(X_t)|\mathcal{F}^Y_t\big)(\omega) = \frac{\check{\mathrm{E}} f\big(X_t(\check{\omega})\big)\psi_t\big(X(\check{\omega}), Y(\omega)\big)}{\check{\mathrm{E}}\psi_t\big(X(\check{\omega}), Y(\omega)\big)}, \quad \mathrm{P} - a.s. \tag{6.17}$$

*where*

$$\psi_t(X, Y) = \exp\left\{\frac{1}{B^2}\int_0^t g(s, X_s)dY_s - \frac{1}{2B^2}\int_0^t g^2(s, X_s)ds\right\}. \tag{6.18}$$

REMARK 6.10. The integral $J(\check{\omega}, \omega) := \int_0^t g\big(X_s(\check{\omega})\big)dY_s(\omega)$ is a well defined random variable on the product space $\big(\check{\Omega}\times\Omega, \check{\mathcal{F}}\times\mathcal{F}, \check{\mathrm{P}}\times\mathrm{P}\big)$. In fact the integration over $\check{\omega}$ could have been done on the original probability space by means of an independent copy of $X$.

REMARK 6.11. The function $f$ need not to be bounded, but should rather satisfy appropriate integrability conditions.

REMARK 6.12. The expression in (6.18) is sometimes referred as the likelihood ratio, being the Radon-Nikodym density of the law of $Y$ under the hypothesis that $Y$ either has a drift or not.

PROOF. Consider $B = 1$ for brevity ($B \neq 1$ is treated completely analogously). Denote by $\mu^W$ the Wiener measure on $C_{[0,T]}$, i.e. the probability measure induced by $W$. Let

$$z_t(X, W) = \exp\left(-\int_0^t g(s, X_s)dW_s - \frac{1}{2}\int_0^t g^2(s, X_s)ds\right), \quad t \in [0, T].$$

Under the assumption on $g$, $z_t$ is a martingale and so

$$\frac{d\widetilde{\mathrm{P}}}{d\mathrm{P}}(\omega) = z_T\big(X(\omega), Y(\omega)\big), \tag{6.19}$$

defines the probability measure $\widetilde{P}$.

Let $Y^x$ be given by[7]

$$Y^x_t = \int_0^t g(s, x_s)ds + W_t, \quad t \in [0, T], \quad x \in D_{[0,T]}.$$

Then by Girsanov theorem (recall that $\mathrm{P} \sim \widetilde{\mathrm{P}}$ and $Y^x$ is a Wiener process under $\widetilde{P}$)

$$\mathrm{E}\big(z_T(x, W)\Psi(Y^x)\big) = \int_{C_{[0,T]}} \Psi(y)\mu^W(dy), \quad \mu^X - a.s,$$

where $\mu^X$ is the probability measure induced by $X$. Now by independence of $X$ and $W$ under P, for any bounded and measurable functionals $\Phi$ and $\Psi$

$$\widetilde{\mathrm{E}}\Psi(Y)\Phi(X) = \mathrm{E}z_T(X, W)\Psi(Y)\Phi(X) =$$

$$\int_{D_{[0,T]}} \mathrm{E}z_T(x, W)\Psi(Y^x)\Phi(x)\mu^X(dx) = \int_{C_{[0,T]}} \Psi(y)\mu^W(dy)\int_{D_{[0,T]}} \Phi(x)\mathrm{Q}^X(dx)$$

This implies that under $\widetilde{\mathrm{P}}$, $Y$ is a Wiener process (take $\Phi \equiv 1$ and arbitrary $\Psi$), $X$ has the same distribution as under P (take $\Psi \equiv 1$ and arbitrary $\Phi$) and $Y$ and $X$ are independent.

Since $z_t(X,W)$ is $\mathcal{F}_t$-martingale and

$$z_t(X,W) = \exp\left(-\int_0^t g(s,X_s)dY_s + \frac{1}{2}\int_0^t g^2(s,X_s)ds\right) = \psi_t^{-1}(X,Y),$$

by Lemma 3.11

$$\mathrm{E}\big(f(X_t)|\mathcal{F}_t^Y\big) = \frac{\widetilde{\mathrm{E}}\big(f(X_t)z_T^{-1}(X,W)|\mathcal{F}_t^Y\big)}{\widetilde{\mathrm{E}}\big(z_T^{-1}(X,W)|\mathcal{F}_t^Y\big)} = \frac{\widetilde{\mathrm{E}}\big(f(X_t)z_t^{-1}(X,W)|\mathcal{F}_t^Y\big)}{\widetilde{\mathrm{E}}\big(z_t^{-1}(X,W)|\mathcal{F}_t^Y\big)} =$$
$$\frac{\widetilde{\mathrm{E}}\big(f(X_t)\psi_t(X,Y)|\mathcal{F}_t^Y\big)}{\widetilde{\mathrm{E}}\big(\psi_t(X,Y)|\mathcal{F}_t^Y\big)} = \frac{\check{\mathrm{E}}f\big(X_t(\check{\omega})\big)\psi_t\big(X(\check{\omega}),Y(\omega)\big)}{\check{\mathrm{E}}\psi_t\big(X(\check{\omega}),Y(\omega)\big)},$$

where the latter holds by independence of $X$ and $Y$ under $\widetilde{\mathrm{P}}$. □

REMARK 6.13. The drift term in (6.16) can be allowed to depend on $Y$: let

$$Y_t = \int_0^t g(s,X_s,Y)ds + BW_t,$$

where $g$ is a non-anticipating measurable $\mathbb{R}_+ \times \mathbb{R} \times C_{[0,t]} \mapsto \mathbb{R}$ functional, such that the SDE has the unique strong solution. Let $\psi_t(X,Y)$ be defined by (6.18) with $g(s,X_s)$ replaced by $g(s,X_s,Y)$. Then for any measurable and bounded $f:\mathbb{R}\mapsto\mathbb{R}$

$$\mathrm{E}\Big(f(X_t)\big|\mathcal{F}_t^Y\Big) = \frac{\widetilde{\mathrm{E}}\big(f(X_t)\psi_t(X,Y)|\mathcal{F}_t^Y\big)}{\widetilde{\mathrm{E}}\big(\psi_t(X,Y)|\mathcal{F}_t^Y\big)}, \tag{6.20}$$

where $\widetilde{\mathrm{E}}$ is the expectation with respect to probability $\widetilde{\mathrm{P}}$ (defined similarly to (6.19)), under which $X$ and $Y$ are independent, $X$ is distributed as under P and $Y$ is a Wiener process.

REMARK 6.14. The Kallianpur-Striebel formula can be reformulated as

$$\mathrm{E}\big(f(X_t)|\mathcal{F}_t^Y\big)(\omega) = \frac{\int_{C_{[0,T]}} f(x_t)\psi_t\big(x,Y(\omega)\big)\mu^X(dx)}{\int_{C_{[0,T]}} \psi_t\big(x,Y(\omega)\big)\mu^X(dx)}, \tag{6.21}$$

where $\mu^X$ is the probability measure (distribution) induced by $X$ on $D_{[0,T]}$ under either P or P$'$.

EXAMPLE 6.15. Consider the Bayesian estimation problem of a random variable $\theta$ ("constant unknown signal") from the observations

$$Y_t = \int_0^t g(s,\theta)ds + W_t.$$

[7] $X$ is assumed to have right continuous pathes with finite left limits. Such functions are usually referred as *cadlag* (French abbreviature) or corlol (English one). In other words, the trajectories are allowed to have countable number of finite jumps. This space, denoted by $D_{[0,T]}$ is not complete under the usual supremum metric. The so called Skorohod metric turns it into a complete separable space

Then by Kallianpur-Stribel formula

$$\mathrm{E}(\theta|\mathcal{F}_t^Y) = \frac{\check{\mathrm{E}}\theta(\check{\omega})\exp\left\{\int_0^t g\big(s,\theta(\check{\omega})\big)dY_s - \frac{1}{2}\int_0^t g^2\big(s,\theta(\check{\omega})\big)ds\right\}}{\exp\left\{\int_0^t g\big(s,\theta(\check{\omega})\big)dY_s - \frac{1}{2}\int_0^t g^2\big(s,\theta(\check{\omega})\big)ds\right\}} =$$

$$\frac{\int_{\mathbb{R}} x\exp\left\{\int_0^t g(s,x)dY_s - \frac{1}{2}\int_0^t g^2(s,x)ds\right\}dF_\theta(x)}{\int_{\mathbb{R}} \exp\left\{\int_0^t g(s,x)dY_s - \frac{1}{2}\int_0^t g^2(s,x)ds\right\}dF_\theta(x)},$$

where $F_\theta(x)$ is the distribution function of $\theta$. In particular, if $g(s,x)\equiv g(x)$

$$\mathrm{E}(\theta|\mathcal{F}_t^Y) = \frac{\int_{\mathbb{R}} x\exp\{g(x)Y_t - \frac{1}{2}g^2(x)t\}dF_\theta(x)}{\int_{\mathbb{R}} \exp\{g(x)Y_t - \frac{1}{2}g^2(x)t\}dF_\theta(x)}.$$

■

**2.2. The Zakai equation.** Note that the Kallianpur-Striebel formula does not impose much structure on $X$. If the signal satisfies (6.1), an SDE can be derived for the *unnormalized* conditional law of $X_t$ given $\mathcal{F}_t^Y$. Below we use the generic notation $\sigma_t(\xi) = \widetilde{\mathrm{E}}(\xi_t\psi_t|\mathcal{F}_t^Y)$, where $\xi$ is an $\mathcal{F}_t$ adapted random process.

THEOREM 6.16. *Assume that in addition to the assumptions of Theorem 6.9, $X$ obeys the representation* (6.1)*, then*

$$d\sigma_t(X) = \sigma_t(H)dt + B^{-2}\sigma_t(Xg)dY_t, \quad t\in[0,T], \tag{6.22}$$

*subject to $\sigma_0(X) = \mathrm{E}X_0$ and*

$$\pi_t(f) = \frac{\sigma_t(f)}{\sigma_t(1)}$$

*for any bounded and measurable $f$.*

REMARK 6.17. Similarly to (6.6), the Zakai equation (6.22) is a measure valued stochastic equation - see Remark 6.6.

PROOF. The process $\psi_t$ satisfies SDE (again $B=1$ is set for brevity)

$$d\psi_t = \psi_t g(t,X_t)dY_t, \quad \psi_0 = 1. \tag{6.23}$$

Then by the Itô formula[8]

$$X_t\psi_t = X_0 + \int_0^t \psi_s dX_s + \int_0^t X_s d\psi_s =$$

$$X_0 + \int_0^t \psi_s H_s dt + \int_0^t \psi_s dM_s + \int_0^t X_s g(s,X_s)\psi_s dY_s.$$

The equation (6.22) is obtained by taking the conditional expectation given $\mathcal{F}_t^Y$, under $\widetilde{\mathrm{P}}$. First note that

$$\widetilde{\mathrm{E}}\left(\int_0^t \psi_s H_s ds\Big|\mathcal{F}_t^Y\right) = \int_0^t \widetilde{\mathrm{E}}\big(\psi_s H_s\big|\mathcal{F}_t^Y\big)ds = \int_0^t \widetilde{\mathrm{E}}\big(\psi_s H_s\big|\mathcal{F}_s^Y\big)ds,$$

[8] Here we use the extension of the Itô formula for general martingales (not necessarily Wiener processes or their stochastic integrals). In the case when it is applied to $f(x,y)=xy$ and independent martingales, it reduces to the usual differentiation rule for product. Verify this in the case of a pair of independent Wiener processes.

where the latter equality holds since $(\psi_s, H_s)$ is $\mathcal{F}^X_s \vee \mathcal{F}^Y_s$-measurable and thus independent of $\mathcal{F}^Y_{[s,T]} = \sigma\{Y_u - Y_s, s \le u \le T\}$ under $\widetilde{\mathrm{P}}$. For the same reason

$$\widetilde{\mathrm{E}}\left(\int_0^t \psi_s dM_s | \mathcal{F}^Y_t\right) = 0, \tag{6.24}$$

and

$$\widetilde{\mathrm{E}}\left(\int_0^t X_s g(s, X_s)\psi_s dY_s | \mathcal{F}^Y_t\right) = \int_0^t \widetilde{\mathrm{E}}\big(X_s g(s, X_s)\psi_s | \mathcal{F}^Y_s\big) dY_s. \tag{6.25}$$

The vulgar proof of these facts can be done by verifying them for simple processes and then extending to the general case by an approximation argument (refer Corollaries 1 and 2 of Theorem 5.13 in [**21**] for a more solid reasoning). □

The FKK equation (6.6) can be recovered from (6.22)

COROLLARY 6.18. *Under the setup of Theorem 6.16, the conditional expectation* $\pi_t(X) = \mathrm{E}\big(X_t | \mathcal{F}^Y_t\big)$ *satisfies*

$$\pi_t(X) = \pi_0(X) + \int_0^t \pi_s(H) ds + \int_0^t \Big(\pi_s(gX) - \pi_s(g)\pi_s(X)\Big) B^{-1} d\bar{W}_s, \tag{6.26}$$

*where*

$$\bar{W}_t = B^{-1}\big(Y_t - \int_0^t \pi_s(g) ds\big).$$

PROOF. By Kallianpur-Striebel formula $\pi_t(X) = \sigma_t(X)/\sigma_t(1)$. By (6.22) the process $\sigma_t(1)$ satisfies

$$d\sigma_t(1) = B^{-2}\sigma_t(g) dY_t, \quad \sigma_0(1) = 1.$$

and by the Itô formula

$$d\pi_t = d\left(\frac{\sigma_t(X)}{\sigma_t(1)}\right) = \frac{d\sigma_t(X)}{\sigma_t(1)} - \frac{\sigma_t(X)}{\sigma_t^2(1)} d\sigma_t(1) + \frac{\sigma_t(X)\sigma_t^2(g)}{B^2\sigma_t^3(1)} dt - \frac{\sigma_t(g)\sigma_t(Xg)}{B^2\sigma_t^2(1)} dt =$$
$$\frac{\sigma_t(H_t)}{\sigma_t(1)} dt + \frac{\sigma_t(Xg)}{B^2\sigma_t(1)} dY_t - \frac{\sigma_t(X)\sigma_t(g)}{B^2\sigma_t^2(1)} dY_t + \frac{\sigma_t(X)\sigma_t^2(g)}{B^2\sigma_t^3(1)} dt - \frac{\sigma_t(g)\sigma_t(Xg)}{B^2\sigma_t^2(1)} dt =$$
$$\pi_t(H) dt + \frac{\pi_t(Xg)}{B^2} dY_t - \frac{\pi_t(X)\pi_t(g)}{B^2} dY_t + \frac{\pi_t(X)\pi_t^2(g)}{B^2} dt - \frac{\pi_t(g)\pi_t(Xg)}{B^2} dt =$$
$$\pi_t(H) dt + B^{-2}\Big(\pi_t(Xg) - \pi_t(X)\pi_t(g)\Big)\Big(dY_t - \pi_t(g) dt\Big)$$

which verifies (6.26). □

**2.3. Stochastic PDE for the unnormalized conditional density.** Similarly to the Kushner-Stratonovich PDE (6.13) for the conditional density in the case of diffusions, the corresponding PDE for the unnormalized conditional density can be derived using (6.22). Consider the diffusion signal, given by the SDE

$$dX_t = a(t, X_t) dt + b(t, X_t) dV_t, \quad X_0 \sim \eta \tag{6.27}$$

where $V$ is a Wiener process, independent of $W$, the coefficients guarantee existence and uniqueness of the strong solution and $\eta$ is a random variable with density $p_0(x)$, with $\int_{\mathbb{R}} x^2 p_0(x) dx < \infty$.

THEOREM 6.19. *Assume that there is an $\mathcal{F}_t^Y$-adapted nonnegative random field $\rho_t(x)$, satisfying*[9] *the Zakai PDE*

$$d\rho_t(x) = \big(\mathcal{L}^*\rho_t\big)(x)dt + B^{-2}g(s,x)\rho_t(x)dY_s, \quad \rho_0(x) = p_0(x). \tag{6.28}$$

*Then $\rho_t(x)$ is a version of the unnormalized conditional density of $X_t$ given $\mathcal{F}_t^Y$, so that for any measurable $f$, such that $\mathrm{E}f^2(X_t) < \infty$,*

$$\mathrm{E}\big(f(X_t)|\mathcal{F}_t^Y\big) = \frac{\int_\mathbb{R} f(x)\rho_t(x)dx}{\int_\mathbb{R}\rho_t(x)dx}, \quad \mathrm{P}-a.s. \tag{6.29}$$

PROOF. Let $f$ be a twice continuously differentiable function (again $B=1$ is treated). Then by the Itô formula

$$f(X_t) = f(X_0) + \int_0^t (\mathcal{L}f)(X_s)ds + \frac{1}{2}\int_0^t f''(X_s)b^2(X_s)dV_s,$$

where $\mathcal{L}$ is defined in (6.14). Applying (6.22) to $f(X_t)$ one obtains

$$\sigma_t(f) = \sigma_0(f) + \int_0^t \sigma_s(\mathcal{L}f)ds + \int_0^t \sigma_s(fg)dY_s.$$

Let's verify that the (random) measure corresponding to the density $\rho_t(x)$, is a solution of the latter equation:

$$\int_0^t \sigma_s(\mathcal{L}f)ds + \int_0^t \sigma_s(fg)dY_s =$$
$$\int_0^t\int_\mathbb{R}\Big(a(x)f'(x) + \frac{b^2(x)}{2}f''(x)\Big)\rho_s(x)dxds + \int_0^t\int_\mathbb{R} f(x)g(s,x)\rho_s(x)dxdY_s =$$
$$\int_\mathbb{R} f(x)\left(\int_0^t (\mathcal{L}^*\rho_s)(x)ds + \int_0^t g(s,x)\rho_s(x)dY_s\right)dx =$$
$$\int_\mathbb{R} f(x)\big(\rho_t(x) - \rho_0(x)\big)dx = \sigma_t(f) - \sigma_0(f).$$

□

REMARK 6.20. The solution existence and uniqueness for (6.28) is the issue far beyond the scope of these lecture notes. The density $\rho_t(x)$ even at the first glance is not an easy mathematical object to treat: being twice differentiable in $x$, it is very nonsmooth in time $t$, as should be a diffusion. Still (6.28) is much easier to deal with compared to (6.13).

**2.4. The robust filtering formulae.** The stochastic PDE (6.28) involves stochastic integral, which is defined on the continuous functions only in the support of the Wiener measure. It turns out, that it may be rewritten as a PDE without stochastic integral, but rather with random coefficients, depending on $Y$ continuously and thus well defined for all continuous functions. Let for simplicity $g(s,x) \equiv g(x)$ and define

$$\widetilde{\rho}_t(x) = R_t(x)\rho_t(x), \tag{6.30}$$

where

$$R_t(x) = \exp\left\{-\frac{1}{B^2}Y_t g(x) + \frac{1}{2B^2}g^2(x)t\right\}.$$

[9]The natural question arises at this point: what is the (strong) solution of stochastic PDE ? Clearly besides the obvious property of adaptedness to $\mathcal{F}_t$, a solution should satisfy some integrability properties in $x$ variable, etc. This issue is beyond the scope of these notes.

Then by the Itô formula

$$d\widetilde{\rho}_t(x) = -\frac{g(x)\widetilde{\rho}_t}{B^2}dY_t + \frac{g^2(x)\widetilde{\rho}_t}{2B^2}dt + \frac{g^2(x)\widetilde{\rho}_t}{2B^2}dt + \\ R_t(x)d\rho_t(x) - \frac{g^2(x)\widetilde{\rho}_t}{B^2}dt = R_t(x)\big(\mathcal{L}^*\rho_t\big)(x)dt,$$

which leads to

$$\begin{aligned} d\widetilde{\rho}_t(x) &= R_t(x)\Big(\mathcal{L}^* R_t^{-1}(x)\widetilde{\rho}_t\Big)(x)dt, \quad \widetilde{\rho}_0(x) = p_0(x) \\ \rho_t(x) &= R_t^{-1}(x)\widetilde{\rho}_t(x). \end{aligned} \tag{6.31}$$

The PDE (6.31) is sometimes referred as *robust* filtering equation, corresponding to the *gauge* transformation (6.30).

## 3. Finite dimensional filters

The nonlinear filtering equations (6.6) and (6.22), as well as the corresponding PDE versions (6.13) and (6.28), are in general infinite dimensional, meaning that their solutions may not belong to a family of stochastic fields, parameterizable by a finite number of sufficient statistics. The importance of the latter is obvious in applications. This section covers some special settings when a finite dimensional filter exists. There is no constructive way to derive or even to verify the existence of the finite dimensional filters in general. However there is a beautiful connection between this issue and Lie algebras generated by the coefficients of the signal/observation equations - see the survey [**31**]. Some negative results about the existence of the finite dimensional realization of the filtering equation with cubic observation nonlinearity are available [**24**], [**11**].

**3.1. The Kalman-Bucy filter revisited.** The Kalman-Bucy filtering formulae can be obtained from the general nonlinear filtering equations.

Theorem 6.21. *The solution of* (5.12) *and* (5.13)*, subject to a Gaussian vector* $(X_0, Y_0)$ *is a Gaussian process. In particular the conditional distribution of* $X_t$*, given* $\mathcal{F}^Y_t$ *is Gaussian with mean* $\widehat{X}_t$ *and covariance* $P_t$*, generated by* (5.14) *and* (5.15) *respectively.*

Proof. Let's verify the claim for the simple scalar example (of course the general vector case is obtained similarly with more tedious calculations). Consider the two dimensional system of linear SDEs

$$\begin{aligned} dX_t &= aX_t dt + bdW_t \\ dY_t &= AX_t dt + BdV_t \end{aligned} \tag{6.32}$$

subject to $Y_0 = 0$ and a Gaussian random variable $X_0$, where $W$ and $V$ are independent Wiener processes, independent of $X_0$, and all the coefficients are scalars. The process $(X, Y)$ form a Gaussian system and hence the conditional law of $X_t$, given $\mathcal{F}^Y_t$ is Gaussian as well, so that we are left with the problem of finding the equations for the conditional mean and variance.

Applying the equation (6.6) to $X_t$ one gets the familiar equation for $\widehat{X}_t := \pi_t(X)$

$$\widehat{X}_t = \mathrm{E}X_0 + \int_0^t a\widehat{X}_s ds + \int_0^t A\big(\pi_s(X^2) - \pi_s^2(X)\big)B^{-2}\big(dY_s - A\widehat{X}_t ds\big) = \mathrm{E}X_0 + \int_0^t a\widehat{X}_s ds + \int_0^t \frac{AP_s}{B^2}\big(dY_s - A\widehat{X}_t ds\big), \quad (6.33)$$

where

$$P_t = \pi_t(X^2) - \pi_t^2(X) = \mathrm{E}(X_t^2|\mathcal{F}_t^Y) - \big(\mathrm{E}(X_t|\mathcal{F}_t^Y)\big)^2 = \mathrm{E}\Big(\big(X_t - \mathrm{E}(X_t|\mathcal{F}_t^Y)\big)^2|\mathcal{F}_t^Y\Big).$$

By the Itô formula

$$X_t^2 = X_0^2 + \int_0^t 2aX_s^2 ds + \int_0^t b^2 ds + \int_0^t 2X_s b dW_s,$$

and thus (6.6) gives

$$\pi_t(X^2) = \pi_0(X_0^2) + \int_0^t \big(2a\pi_s(X^2) + b^2\big)ds + \int_0^t A\Big(\pi_s(X^3) - \pi_s(X)\pi_s(X^2)\Big)B^{-2}\big(dY_s - A\widehat{X}_s ds\big) \quad (6.34)$$

Note that $\pi_t(X^2) = \widehat{X}_t^2 + P_t$ and moreover since the conditional law of $X_t$ is Gaussian $\mathrm{E}\big((X_t - \widehat{X}_t)^p|\mathcal{F}_t^Y\big) = 0$ for any odd $p$ and so

$$\pi_t(X^3) = \mathrm{E}\big(X_t^3|\mathcal{F}_t^Y\big) = \mathrm{E}\big((X_t - \widehat{X}_t + \widehat{X}_t)^3|\mathcal{F}_t^Y\big) = 3\mathrm{E}\big((X_t - \widehat{X}_t)^2|\mathcal{F}_t^Y\big)\widehat{X}_t + \widehat{X}_t^3 = 3P_t\widehat{X}_t + \widehat{X}_t^3.$$

Then (6.34) gives

$$\widehat{X}_t^2 + P_t = \widehat{X}_0^2 + P_0 + \int_0^t \big(2a\widehat{X}_s^2 + 2aP_s + b^2\big)ds + \int_0^t 2AP_s\widehat{X}_s B^{-2}\big(dY_s - A\widehat{X}_s ds\big).$$

Recall that $\bar{W}_t = \big(dY_s - A\widehat{X}_s ds\big)/B$ is a Wiener process and thus by (6.33),

$$d\widehat{X}_t^2 = \widehat{X}_0^2 + \int_0^t 2a\widehat{X}_s^2 ds + \int_0^t \frac{A^2P_s^2}{B^2}ds + 2\widehat{X}_s\frac{AP_s}{B}d\bar{W}_s.$$

The latter two equations imply

$$\dot{P}_t = 2aP_t + b^2 - \frac{A^2P_t^2}{B^2}, \quad P_0 = \mathrm{E}(X_0 - \mathrm{E}X_0)^2,$$

which is the familiar Riccati equation for the filtering error. □

Remark 6.22. In particular in the linear Gaussian case the conditional density equation (6.13) is solved by

$$p_t(x) = \frac{1}{\sqrt{2\pi P_t}}\exp\left\{\frac{-(x - \widehat{X}_t)^2}{2P_t}\right\}.$$

**3.2. Conditionally Gaussian filter.** In the previous section the key reason for the FKK to be finite (two) dimensional was the Gaussian property of the pair $(X, Y)$. In fact the very same arguments would be applicable, if only the conditional distribution of $X_t$ given $\mathcal{F}_t^Y$ is Gaussian. This leads to the following generalization of the Kalman-Bucy filter due to R.Liptser and A.Shiryaev (see Chapters 11, 12 in [**21**])

THEOREM 6.23. *(Conditionally Gaussian filter) Consider the SDE system*

$$dX_t = \big(a_0(t,Y) + a_1(t,Y)X_t\big)dt + b(t,Y_t)dW_t \tag{6.35}$$

$$dY_t = \big(A_0(t,Y) + A_1(t,Y)X_t\big)dt + BdV_t \tag{6.36}$$

*subject to $Y_0 = 0$ and Gaussian random variable $X_0$, where $B$ is a positive constant and the rest of the coefficients are non-anticipating functionals of $Y$, satisfying the conditions under which the unique strong solution $(X,Y) = (X_t, Y_t)_{t\in[0,T]}$ exists and $\mathrm{E}X_t^2 < \infty$ $t \in [0,T]$. Then the conditional distribution of $X_t$ given $\mathcal{F}_t^Y$ is Gaussian with the mean $\widehat{X}_t$ and variance $P_t$, given by*

$$\begin{aligned} d\widehat{X}_t = \Big(a_0(t,Y) + a_1(t,Y)\widehat{X}_t\Big)dt + \\ \frac{A_1(t,Y)P_t}{B^2}\Big(dY_t - A_0(t,Y)dt - A_1(t,Y)\widehat{X}_t dt\Big) \\ \dot{P}_t = 2a_1(t,Y)dt + b^2(t,Y)dt - \frac{A_1^2(t,Y)P_t^2}{B^2}, \end{aligned} \tag{6.37}$$

*subject to $\widehat{X}_0 = \mathrm{E}X_0$ and $P_0 = \mathrm{E}(X_0 - \widehat{X}_0)^2$.*

REMARK 6.24. Note that in general the processes $(X,Y)$ do not form a Gaussian system anymore. The only essential constrain on the structure of (6.35) and (6.36) is linear dependence on $X_t$. Despite of similarity, the difference between the Kalman-Bucy filter (5.3) and the equations (6.37) is significant: the latter are no longer linear and the conditional filtering error is no longer deterministic ! This nonlinear generalization plays an important role in various problems of control and optimization (see e.g. the "Applications" volume of [**21**]). The multidimensional version of the filter is derived similarly.

PROOF. Only the conditional Gaussian property of $(X,Y)$ is to be verified

$$\mathrm{E}\Big(e^{i\lambda X_t}|\mathcal{F}_t^Y\Big) = \exp\left\{i\lambda m_t(Y) - \frac{1}{2}\lambda^2 V_t(Y)\right\}, \quad \lambda \in \mathbb{R} \tag{6.38}$$

where $m_t(Y)$ and $V_t(Y)$ are some non-anticipating functionals of $Y$. Once (6.38) is established the very same arguments of the preceding section lead to the equations (6.37), i.e. $m_t(Y) \equiv \widehat{X}_t$ and $V_t(Y) \equiv P_t$.

The equation (6.35) has a closed form solution

$$X_t = \gamma(t,Y)\left(X_0 + \int_0^t \gamma^{-1}(s,Y)b(s,Y)dW_s\right) := \Phi_t(X_0, W, Y). \tag{6.39}$$

where $\gamma(t,Y) = \exp\Big\{\int_0^t \big(a_0(s,Y) + a_1(s,Y)\big)ds\Big\}$.

The (6.20) version of Kallianpur-Striebel formula implies

$$\mathrm{E}\big(e^{i\lambda X_t}|\mathcal{F}_t^Y\big) = \frac{\widetilde{\mathrm{E}}\big(e^{i\lambda X_t}\psi_t(X,Y)|\mathcal{F}_t^Y\big)}{\widetilde{\mathrm{E}}\big(\psi_t(X,Y)|\mathcal{F}_t^Y\big)}, \tag{6.40}$$

where

$$\psi_t(X,Y) = \exp\Big\{\int_0^t \big(A_0(s,Y)+A_1(s,Y)X_s\big)dY_s - \frac{1}{2}\int_0^t \big(A_0(s,Y)+A_1(s,Y)X_s\big)^2 ds\Big\}.$$

Insert the expression (6.39) into the right hand side of (6.40). Since $Y$ and $(W, X_0)$ are independent under $\widetilde{\mathrm{E}}$ (which follows from the independence of $Y$ and $X$), the expectation $\widetilde{\mathrm{E}}$ averages over $(X_0, W)$, keeping $Y$ fixed. This results in the quadratic form of the type (6.38), due to Gaussian property of the system $(X_0, W)$, which enter the exponent linearly. In fact its precise expression is identical to the one that would have been obtained in the usual Kalman-Bucy setting. □

REMARK 6.25. Another (much more harder!) way to verify the claim of the Theorem 6.23 is to check that Gaussian density with the mean and variance driven by (6.37) is the unique solution of FKK equation (or Kushner-Stratonovich equation).

**3.3. Linear systems with non-Gaussian initial condition.** If the initial condition $X_0$ is non-Gaussian, the conditional law of $X_t$ given $\mathcal{F}_t^Y$ is no longer Gaussian and thus the Kalman-Bucy equations do not necessarily generate the conditional mean and variance. It turns out that a finite dimensional filter exists and even can be derived in a number of ways, of which we choose the elegant approach due to A.Makowski [**30**].

THEOREM 6.26. *Consider the processes $(X, Y)$ generated by the linear system (with $B = 1$) (6.32), started from a random variable $X_0$ with distribution $F(x)$, $\int_{\mathbb{R}} x^2 dF(x) < \infty$. Then for any measurable $f$, such that $\mathrm{E}f^2(X_t) < \infty$, $t \in [0, T]$*

$$\mathrm{E}\big(f(X_t)|\mathcal{F}_t^Y\big) = \frac{\int_{\mathbb{R}^2}\int_{\mathbb{R}} f(x_1 + e^{at}u)\psi_t(u, x_2)dF(u)\Gamma_t(x_1, x_2)dx_1dx_2}{\int_{\mathbb{R}}\int_{\mathbb{R}} \psi_t(u, x_2)dF(u)\gamma_t(x_2)dx_2} \tag{6.41}$$

*where*

$$\psi_t(u, x) = \exp\left\{ux - \frac{u^2}{2}\frac{A^2}{2a}(e^{2at}-1)\right\},$$

*$\Gamma_t(x, y)$ is the two dimensional Gaussian density with the mean and covariance satisfying the equations*

$$\begin{aligned} d\widehat{X}_t &= a\widehat{X}_t dt + AP_t^2\big(dY_t - A\widehat{X}_t\big), \quad \widehat{X}_0 = 0 \\ d\widehat{\xi}_t &= A\big(e^{at} + Q_t\big)\big(dY_t - A\widehat{X}_t\big), \quad \widehat{\xi}_0 = 0 \end{aligned} \tag{6.42}$$

*and*

$$\begin{aligned} \dot{P}_t &= 2aP_t + b^2 - A^2P_t^2, \quad P_0 = 0 \\ \dot{Q}_t &= aQ_t - P_tA^2\big(Q_t + e^{at}\big), \quad Q_0 = 0 \\ \dot{R}_t &= A^2e^{2at} - A^2\big(Q_t + e^{at}\big)^2, \quad R_0 = 0, \end{aligned} \tag{6.43}$$

*and $\gamma_t(x)$ is its marginal with the mean $\widehat{\xi}_t$ and variance $R_t$.*

PROOF. Let $X^\circ$ be the solution of $\dot{X}_t^\circ = aX_t^\circ$, subject to $X_0^\circ = X_0$, i.e.

$$X_t^\circ = e^{at}X_0, \quad t \in [0, T],$$

and $X'_t$ be the solution of

$$dX'_t = aX'_t dt + b dW_t, \quad X'_0 = 0.$$

Then $X_t = X^\circ_t + X'_t$, $t \in [0,T]$ and

$$Y_t = \int_0^t AX'_s ds + \int_0^t AX^\circ_s ds + V_t. \tag{6.44}$$

Define

$$\varphi_t = \exp\left\{ -\int_0^t AX^\circ_s dV_s - \frac{1}{2}\int_0^t \big(AX^\circ_s\big)^2 ds \right\}$$

Since $\mathrm{E}X_0^2 < \infty$ is assumed, $\varphi_t$ is a martingale and by Girsanov theorem the Radon-Nikodym derivative

$$\frac{d\widetilde{\mathrm{P}}}{d\mathrm{P}}(\omega) = \varphi_T(\omega)$$

defines the probability measure $\widetilde{\mathrm{P}}$, under which

$$V'_t := \int_0^t AX^\circ_s ds + V_t$$

is a Wiener process, independent of $X^\circ$ (or equivalently of $X_0$) and $X'$ (which is verified as in the proof of Kallianpur-Striebel formula of Theorem 6.9), whose distributions are preserved. Moreover

$$\mathrm{E}\big(f(X_t)|\mathcal{F}^Y_t\big) = \frac{\widetilde{\mathrm{E}}\big(f(X'_t + e^{at}X_0)\psi_t(X_0,\xi)|\mathcal{F}^Y_t\big)}{\widetilde{\mathrm{E}}\big(\psi_t(X_0,\xi)|\mathcal{F}^Y_t\big)} \tag{6.45}$$

where

$$\begin{aligned}
\psi_t(X^\circ,\xi) := \varphi_t^{-1} = &\exp\left\{\int_0^t AX^\circ_s dV'_s - \frac{1}{2}\int_0^t \big(AX^\circ_s\big)^2 ds\right\} = \\
&\exp\left\{X_0\int_0^t Ae^{as} dV'_s - \frac{X_0^2}{2}\int_0^t \big(Ae^{as}\big)^2 ds\right\} = \\
&\exp\left\{X_0\int_0^t d\xi_s - \frac{X_0^2}{2}\int_0^t \big(Ae^{as}\big)^2 ds\right\},
\end{aligned}$$

where $d\xi_t = Ae^{at}dV'_t$ was defined. Note that under $\widetilde{\mathrm{P}}$, $(X',\xi,Y)$ form a Gaussian system (independent of $X_0$) and thus the conditional distribution of $(X'_t,\xi_t)$ given $\mathcal{F}^Y_t$ is Gaussian, whose parameters can be found by the Kalman-Bucy filter for the linear model

$$\begin{aligned}
dX'_t &= aX'_t dt + b dW_t, \quad X'_0 = 0 \\
d\xi_t &= Ae^{at}dV'_t, \quad \xi_0 = 0 \\
dY_t &= AX'_t dt + dV'_t, \quad Y_0 = 0.
\end{aligned}$$

Applying the equations (5.14) and (5.15), one gets (6.42) and (6.43) and the formula (6.41) follows from (6.45).

□

### 3.4. Markov chains with finite state space.

3.4.1. *The Poisson process.* Similarly to the role played by the Wiener process $W$ in the theory of diffusion, the Poisson process $\Pi$ is the main building block of purely discontinuous martingales, counting processes, etc.

DEFINITION 6.27. A Markov process $\Pi$ with piecewise constant (right continuous) trajectories with unit positive jumps, $\Pi_0 = 0$, P-a.s. and stationary independent increments, such that[10]

$$\mathrm{P}\big(\Pi_t - \Pi_s = k|\mathcal{F}^\Pi_s\big) = \frac{\big(\lambda(t-s)\big)^k e^{-\lambda(t-s)}}{k!}, \quad k \in \mathbb{Z}_+, \tag{6.46}$$

is called Poisson process with intensity[11] $\lambda \geq 0$.

The existence of $\Pi$ is a relatively easy matter: let $(\tau_n)_{n\geq 1}$ be an i.i.d sequence of exponential random variables

$$\mathrm{P}\big(\tau_1 \geq t\big) = e^{-\lambda t}, \quad t \geq 0,$$

and let[12]

$$\Pi_t = \max_{n\geq 0}\left\{n : \sum_{i=1}^n \tau_i \leq t\right\}, \quad t \geq 0. \tag{6.47}$$

THEOREM 6.28. *$\Pi$ defined in (6.47) is a Poisson process.*

PROOF. Clearly $\Pi_0 = 0$ and the trajectories of (6.47) are piecewise constant as required. Introduce $\sigma_k = \sum_{i=1}^k \tau_i$. Then

$$\mathrm{P}(\Pi_t = k|\mathcal{F}^\Pi_s) = \sum_{\ell=0}^k \mathrm{P}(\Pi_t = k|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell)\mathbf{1}_{\{\Pi_s = \ell\}}$$

and thus

$$\mathrm{P}(\Pi_t = k|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell) = \frac{(\lambda(t-s))^{(k-\ell)}e^{-\lambda(t-s)}}{(k-\ell)!}$$

is to be verified:

$$\begin{aligned}
&\mathrm{P}(\Pi_t = k|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell) = \mathrm{P}(\sigma_k \leq t < \sigma_{k+1}|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell) = \\
&\mathrm{E}\Big(\mathrm{P}(\sigma_k \leq t < \sigma_{k+1}|\tau_1, ..., \tau_{\ell+1})\Big|\tau_1, ..., \tau_\ell, \tau_{\ell+1} > s - \sigma_\ell\Big) = \\
&\mathrm{E}\Big(\mathrm{P}(\tau_{\ell+2} + ... + \tau_k \leq t - \sigma_\ell - \tau_{\ell+1} < \tau_{\ell+2} + ... + \tau_{k+1}|\sigma_\ell, \tau_{\ell+1})\Big|\sigma_\ell, \tau_{\ell+1} > s - \sigma_\ell\Big) \\
&= P\big(\tau_{\ell+2} + ... + \tau_k \leq t - \sigma_\ell - \tau_{\ell+1} < \tau_{\ell+2} + ... + \tau_{k+1}\big|\sigma_\ell, \tau_{\ell+1} > s - \sigma_\ell\big) = \\
&e^{\lambda(s-\sigma_\ell)}\int_{s-\sigma_\ell}^\infty P\big(\tau_{\ell+2} + ... + \tau_k \leq t - \sigma_\ell - u < \tau_{\ell+2} + ... + \tau_{k+1}\big|\sigma_\ell\big)\lambda e^{-\lambda u}du = \\
&= \int_0^\infty P\big(\tau_{\ell+2} + ... + \tau_k \leq t - s - u' < \tau_{\ell+2} + ... + \tau_{k+1}\big)\lambda e^{-\lambda u'}du' = \\
&= P\big(\tau_{\ell+1} + \tau_{\ell+2} + ... + \tau_k \leq t - s < \tau_{\ell+1} + \tau_{\ell+2} + ... + \tau_{k+1}\big) = \\
&= P\big(\tau_1 + ... + \tau_{k-\ell} \leq t - s < \tau_1 + ... + \tau_{k-\ell+1}\big) = P\big(\Pi_{t-s} = k - \ell\big).
\end{aligned}$$

---

[10] extra care should be taking, when manipulating the filtrations of point processes. This delicate matter is left out (as many others) - see the last chapter in [**21**] for a discussion

[11] in (6.46) $0^0 = 1$ is understood and so $\lambda = 0$ is allowed

[12] $\sum_{i=1}^0 ... \equiv 0$ is understood

Now (6.46) holds, if

$$P(\Pi_t = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \quad k \ge 0. \tag{6.48}$$

Note that

$$P(\Pi_t = k) = P(\sigma_k \le t < \sigma_k + \tau_{k+1}) = EI(\sigma_k \le t)I(\tau_{k+1} > t - \sigma_k) = \\ EI(\sigma_k \le t)e^{-\lambda(t-\sigma_k)} = \int_0^t e^{-\lambda(t-s)} dP(\sigma_k \le s). \tag{6.49}$$

and

$$P(\sigma_k \le s) = P(\tau_k \le s - \sigma_{k-1}) = EP(\tau_k \le s - \sigma_{k-1}|\sigma_{k-1}) = \\ EI(s - \sigma_{k-1} \ge 0)(1 - e^{-\lambda(s-\sigma_{k-1})}) = \int_0^s (1 - e^{-\lambda(s-u)}) dP(\sigma_{k-1} \le u) \tag{6.50}$$

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 (6.50) satisfies

$$\frac{dP(\sigma_k \le s)}{ds} = \lambda \int_0^s e^{-\lambda(s-u)} \frac{dP(\sigma_{k-1} \le u)}{du} du$$

and thus[13]

$$\frac{dP(\sigma_k \le s)}{ds} = \lambda \frac{(\lambda s)^{k-1} e^{-\lambda s}}{(k-1)!}.$$

Now the equation (6.48) follows from (6.49). □

A simple consequence of the definition is that $\Pi_t - \lambda t$ is a martingale. Remarkably the converse is true (compare the Levy theorem (Theorem 4.5) for the Wiener process)

THEOREM 6.29. *(S. Watanabe) A process $N_t$ with piecewise constant (right continuous) trajectories with positive unit jumps is a Poisson process with intensity $\lambda$, if $N_t - \lambda t$ is a martingale.*

Since the pathes of $\Pi_t$ are of bounded variation, the stochastic integral with respect to $\Pi$ is understood in Stieltjes sense: for any bounded[14] random process $X$

$$\int_0^t X_{s-} dN_s = \sum_{s \le t} X_{s-} \Delta N_s = \sum_{s \le t} X_{s-} \big(N_s - N_{s-}\big), \tag{6.51}$$

where $X_{s-}$ denotes the left limit of $X$ at point $s$. If $X$ is an $\mathcal{F}_t^N$-adapted process, then $\int_0^t X_{s-}(dN_s - \lambda ds)$ is a martingale[15].

[13] This is known as Erlang distribution

[14] we won't need integrands more complicated than bounded ones

[15] This is again an oversimplification, as many things in these notes

3.4.2. *Markov chains in continuous time.* The Markov chains with finite number of states is the simplest example of Markov processes in continuous time [16]. Among many possible constructions we choose the following: let $\mathbb{S} = \{a_1, ..., a_d\}$ be a finite set of (distinct) real numbers and $N_t$ be $d \times d$ matrix, whose off diagonal entries are independent Poisson processes with intensities $\lambda_{ij} \geq 0$. The diagonal entries are chosen in a special way: $N_t(i,j) = -\sum_{j\neq i} N_t(i,j)$. Now define the vector process $I_t$ by

$$I_t = I_0 + \int_0^t dN^*_s I_{s-}, \tag{6.52}$$

where $I_0$ is a random vector, equal to one of the vectors of the standard Euclidian basis[17] $\{e_1, ..., e_d\}$ with probabilities $p_i \geq 0$. It is easy[18] to see that only one component of $I_t$ equals unity and all others are zeros at any time $t \geq 0$, i.e. $I_t$ takes the values in $\{e_1, ..., e_d\}$ as well. Finally define

$$X_t = \sum_{i=1}^d a_i I_t(i), \quad t \geq 0.$$

THEOREM 6.30. *The process $X$ is a Markov chain with initial distribution*[19] $p_0$ *and transition intensities matrix $\Lambda$ with off-diagonal entries $\lambda_{ij}$ and*

$$\lambda_{ii} := -\sum_{j\neq i} \lambda_{ij}, \quad i = 1, ..., d,$$

*meaning that*

$$p_{s,t}(j) := \mathrm{P}(X_t = a_j | \mathcal{F}^X_s) = \sum_{i=1}^d p_{s,t}(i,j) \mathbf{1}_{\{X_s = a_i\}}, \quad t \geq s \geq 0, \tag{6.53}$$

*where the matrix $p_{s,t}$ solves the forward Kolmogorov equation*[20]

$$\frac{\partial}{\partial t} p_{s,t} = \Lambda^* p_{s,t}, \quad p_{s,s} = E_{d\times d}.$$

PROOF. Since $I_t$ takes values in $\{e_1, ..., e_d\}$, by definition $\mathcal{F}^X_t = \mathcal{F}^I_t$ and thus $\mathrm{P}(X_t = a_i | \mathcal{F}^X_s) = \mathrm{P}(I_t = e_i | \mathcal{F}^I_s) = q_{s,t}(i)$, $i = 1, ..., d.$, where $q_{s,t} := \mathrm{E}(I_t | \mathcal{F}^I_s)$. The latter satisfies

$$q_{s,t} = I_s + \mathrm{E}\left( \int_s^t dN^*_u I_{u-} \Big| \mathcal{F}^I_s \right) =$$

$$I_s + \mathrm{E}\left( \int_s^t (dN^*_u - \Lambda^* du) I_{u-} + \int_s^t \Lambda^* I_{u-} du \Big| \mathcal{F}^I_s \right) = I_s + \int_s^t \Lambda^* q_{s,u} du, \tag{6.54}$$

where[21] the martingale property of the stochastic integral has been used. Reading (6.54) componentwise gives (6.53) and verifies the claim of the theorem. □

---

[16]for the general theory of Markov processes, the reader is referred to the classic text [**6**] - but don't expect easy reading!

[17]i.e. $i$-th entry of $e_i$ is one and the rest are zeros

[18]note that the probability of an event, that any two of a finite number of Poisson processes have a jump simultaneously is zero - this follows directly from the construction of the Poisson process, since exponential distribution does not have atoms.

[19]distributions on $\mathbb{S}$ are identified with vectors of the simplex $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x_i \geq 0\}$ in an obvious way

[20]$E_{d\times d}$ is $d$-dimensional identity matrix

[21]Note that $\int_0^t \Lambda^* I_{s-} ds = \int_0^t \Lambda^* I_s ds$ since the integrator is continuous!

In particular the equation (6.53) implies that the a priori distribution of $X_t$, i.e. the vector of probabilities $p_i = \mathrm{P}(X_t = a_i)$ satisfies the equation

$$\dot{p}_t = \Lambda^* p_t, \quad \text{subject to } p_0, \tag{6.55}$$

whose explicit solution is given by means of the matrix exponential $p_t = e^{\Lambda^* t} p_0$.

3.4.3. *The Shiryaev-Wonham filter.* Consider the filtering problem of a finite state Markov chain $X$ (with known parameters) to be estimated from the trajectory of the observation process $Y$, given by

$$Y_t = \int_0^t g(X_s)ds + BW_t, \quad t \in [0, T]$$

where $g$ is an $\mathbb{S} \mapsto \mathbb{R}$ function, $B > 0$ is a constant and $W$ is a Wiener process, independent of $X$. The sufficient statistics in this problem is the vector[22] $\pi_t$ of conditional probabilities $\pi_t(i) = \mathrm{P}(X_t = a_i | \mathcal{F}_t^Y)$, $i = 1, ..., d$, since

$$\mathrm{E}\big(f(X_t)|\mathcal{F}_t^Y\big) = \mathrm{E}\Big(\sum_{i=1}^d f(a_i)\mathbf{1}_{\{X_t = a_i\}}|\mathcal{F}_t^Y\Big) = \sum_{i=1}^d f(a_i)\pi_t(i).$$

The following theorem gives the complete solution to the filtering problem

THEOREM 6.31. *(Shiryaev [**35**], Wonham [**40**]) The vector $\pi_t$ satisfies the Itô SDE*

$$d\pi_t = \Lambda^* \pi_t dt + \big(\operatorname{diag}(\pi_t) - \pi_t \pi_t^*\big) g \big(dY_t - g^* \pi_t dt\big)/B^2, \quad \pi_0 = p_0, \tag{6.56}$$

*where $g$ stands for $d$-dimensional vector with entries $g(a_i)$, $i = 1, ..., d$. Moreover*[23] *$\pi_t = \rho_t / |\rho_t|$, where*

$$d\rho_t = \Lambda^* \rho_t dt + \operatorname{diag}(g)\rho_t dY_t / B^2, \quad \rho_0 = p_0. \tag{6.57}$$

PROOF. The equation 6.56 follows from the FKK equation (6.6), applied to the process $I_t$, introduced in (6.52). In particular the $i$-th component of $I_t$ satisfies

$$I_t(i) = I_0(i) + \int_0^t \sum_{j=1}^d \lambda_{ji} I_s(j) ds + \int_0^t \sum_{j=1}^d I_{s-}(j)\big(dN_s(ji) - \lambda_{ji} ds\big) :=$$
$$I_0(i) + \int_0^t \sum_{j=1}^d \lambda_{ji} I_s(j) ds + M_t(i),$$

where $M(i)$ is a square integrable martingale. Then (6.6) implies

$$\begin{aligned}\pi_t(i) =& \pi_0(i) + \int_0^t \lambda_{ji}\pi_s(j)ds + \\ & \big(\mathrm{E}(I_s(i) g^* I_s | \mathcal{F}_s^Y) - \pi_s(i)\mathrm{E}(g^* I_s | \mathcal{F}_s^Y)\big)\big(dY_s - \mathrm{E}(g^* I_s|\mathcal{F}_s^Y) ds\big)/B^2 = \\ & \pi_0(i) + \int_0^t \lambda_{ji}\pi_s(j)ds + \big(g_i \pi_s(i) - \pi_s(i)\pi_s^* g\big)\big(dY_s - g^*\pi_s ds\big)/B^2\end{aligned}$$

which is nothing but (6.56) in the componentwise notation. Similarly (6.57) follows from (6.22). ☐

[22] a slight abuse of notation is allowed here - recall that $\pi_t(\cdot)$ stands for the conditional expectation operator in the FKK equation (6.6)

[23] $|x|$ denotes the $\ell^2$ norm: $|x| = \sum_i |x_i|$.

EXAMPLE 6.32. The two dimensional version of (6.56) was derived in [**35**] and shown to play an important role in the problems of quickest change detection. Let $X$ be a symmetric Markov chain with the switching intensity $\lambda > 0$ and with values in $\{0, 1\}$ (often referred as telegraphic signal) and set $\pi_t = \mathrm{P}(X_t = 1|\mathcal{F}_t^Y)$. Suppose that the observations

$$Y_t = \int_0^t X_s ds + W_t$$

are available. Then

$$d\pi_t = \lambda(1 - 2\pi_t)dt + \pi_t(1 - \pi_t)\big(dY_t - \pi_t dt\big), \quad \pi_0 = \mathrm{P}(X_0 = 1).$$

■

3.4.4. *Filtering number of transitions and occupation times.* Clearly the key to the existence of finite dimensional filter for finite state Markov chains is the fact that powers of the indicators process $I_t$ reduce to a linear function of $I_t$! This can be exploited further to get finite dimensional filters for various functionals of $X$: the occupation time of the state $a_i$

$$O_t(i) = \int_0^t \mathbf{1}_{\{X_s = a_i\}} ds = \int_0^t I_s(i) ds, \tag{6.58}$$

the number of transitions from $a_i$ to $a_j$

$$T_t(i, j) = \int_0^t \mathbf{1}_{\{X_{s-} = a_i\}} d\mathbf{1}_{\{X_s = a_j\}} = \int_0^t I_{s-}(i) dI_s(j) \tag{6.59}$$

and the stochastic integrals like

$$J = \int_0^t I_s dY_s. \tag{6.60}$$

Being of interest on their own, the filtering formulae for these quantities can be used to estimate the intensities matrix $\Lambda$ and other parameters in the problem by means of so called EM (Expectation/Minimization) algorithm.[24] We derive the filter for $O_t$ (omitting the index $i$, since the derivation is the same for all $i$'s), leaving the rest as exercises. These problems seem to be initially addressed in [**42**], the derivation below is taken from [**8**].

THEOREM 6.33. *The filtering estimate* $\bar{O}_t = \mathrm{E}\big(O_t|\mathcal{F}_t^Y\big) = |\bar{Z}_t|$, *with* $\bar{Z}_t$ *being the solution of*

$$d\bar{Z}_t = \Lambda^* \bar{Z}_t dt + e_i e_i^* \pi_t dt + \big(\operatorname{diag}(\bar{Z}_t) - \bar{Z}_t \pi_t^*\big) g\big(dY_t - g^* \pi_t dt\big)/B^2, \quad \bar{Z}_0 = 0. \tag{6.61}$$

PROOF. The trick is to introduce an auxiliary process $Z_t = O_t I_t$ with values in $\mathbb{R}^d$. Once the conditional expectation $\bar{Z}_t = \mathrm{E}(Z_t|\mathcal{F}_t^Y)$ is found, the estimate of $O_t$ is recovered by

$$\bar{O}_t = \mathrm{E}\Big(O_t \sum_{i=1}^d I_t(i)|\mathcal{F}_t^Y\Big) = \sum_{i=1}^d \mathrm{E}\big(O_t I_t(i)|\mathcal{F}_t^Y\big) = \sum_{i=1}^d \bar{Z}_t(i) = \big|\bar{Z}_t\big|$$

By the Itô formula[25]

$$dZ_t = d(O_t I_t) = O_t dI_t + I_t dO_t = O_t dN_t^* I_{t-} + I_t I_t(i) dt = dN_t^* Z_{t-} + e_i e_i^* I_t dt$$

---

[24] an iterative procedure for finding maximum of certain likelihood functionals.

[25] in this case it is simply integration by parts: no continuous time martingales or mutual jumps are involved: note that $O_t$ has absolutely continuous trajectories

and hence

$$Z_t = \int_0^t \big(\Lambda^* Z_s ds + e_i e_i^* I_s\big) ds + \int_0^t (dN_s^* - \Lambda^* ds) Z_{s-} := \int_0^t \big(\Lambda^* Z_s ds + e_i e_i^* I_s\big) ds + M_t'$$

where $M_t'$ is a square integrable martingale (check it). Apply (6.6) to the component $Z_t(\ell)$

$$\bar{Z}_t(\ell) = \int_0^t \Big(\sum_{j=1}^d \lambda_{j\ell} \bar{Z}_s(\ell) + \delta_{i\ell} \pi_s(i)\Big) ds \\ + \int_0^t \Big(\mathrm{E}(Z_s(\ell) g^* I_s | \mathcal{F}_s^Y) - \bar{Z}_s(\ell) g^* \pi_s\Big) B^{-2} \big(dY_s - g^* \pi_s ds\big)$$

Since $Z_s(\ell) g^* I_s = g^* O_s I_s(\ell) I_s = g^* e_\ell Z_s(\ell) = g_\ell Z_s(\ell)$, the equation (6.61) is obtained. □

**3.5. Beneŝ filter.** Unlike the preceding finite dimensional filters, Beneŝ filter ([**2**]) is mostly of "academic" interest: it is an example of a filtering problem for nonlinear diffusions admitting finite dimensional realization. This filter does not seem to have an analogue in discrete time.

THEOREM 6.34. *Consider the two dimensional system of SDEs*

$$\begin{aligned} dX_t &= h(X_t)dt + dW_t \\ dY_t &= X_t + dV_t \end{aligned} \tag{6.62}$$

*subject to* $Y_0 = 0$ *and* $X_0 = 0$*, where* $W$ *and* $V$ *are independent Wiener processes. Assume that* $h(x)$ *satisfies the ODE*

$$h' + h = ax^2 + bx + c, \quad a \ge 0, b, c \in \mathbb{R}$$

*and is such that* (6.62) *has a unique strong solution. Then the unnormalized conditional distribution of* $X_t$ *given* $\mathcal{F}_t^Y$ *has density*

$$\rho_t(x) = \exp\Big\{H(x) + xY_t + \frac{1}{2}\sqrt{1+a}x^2 - \frac{1}{2}(c+k)t\Big\} \\ \int_{\mathbb{R}^2} e^{x_2+x_3} \Gamma(x; m_t, V_t) dx_2 dx_3 \tag{6.63}$$

*where* $\Gamma(x; m_t, V_t)$ *is three dimensional Gaussian density with the mean* $m_t$ *and covariance matrix* $V_t$*, corresponding to the Gaussian system*

$$\begin{aligned} d\xi_t &= -\sqrt{1+a}\xi_t dt + dW_t, & \xi_0 &= 0 \\ d\eta_t &= -Y_t dW_t, & \eta_0 &= 0 \\ d\theta_t &= \big(Y_t\sqrt{1+a} - b/2\big)\xi_t dt, & \theta_0 &= 0. \end{aligned} \tag{6.64}$$

REMARK 6.35. For example $h(x) = \tanh(x)$ satisfies the Beneŝ nonlinearity with $a = b = 0$ and $c = 1$, and the Kalman-Bucy case $h(x) = x$ corresponds to $b = c = 1$, $a = 0$.

PROOF. By the Kallianpur-Stribel formula, for any measurable and bounded function $f$

$$\mathrm{E}\big(f(X_t)|\mathcal{F}_t^Y\big)(\omega) = \frac{\int_{C_{[0,T]}} f(x_t)\psi_t\big(x, Y(\omega)\big)\mu^X(dx)}{\int_{C_{[0,T]}} \psi_t\big(x, Y(\omega)\big)\mu^X(dx)},$$

with

$$\psi_t(x,Y) = \exp\left\{\int_0^t x_s dY_s - \frac{1}{2}\int_0^t x_s^2 ds\right\}, \quad \mu^X - a.s.$$

and where $\mu^X$ denotes the probability measure induced by $X$.

The integration with respect to $\mu^X$ can be replaced with integration by the Wiener measure $\mu^W$: indeed by the Girsanov theorem $\mu^X \sim \mu^W$ (checking that $h(X_t)$ satisfies e.g. the Novikov condition (4.20)) and

$$\frac{d\mu^X}{d\mu^W}(x) = \exp\left\{\int_0^t h(x_s)dx_s - \frac{1}{2}\int_0^t h^2(x_s)ds\right\}, \quad \mu^X - a.s.$$

Hence

$$\int_{C_{[0,T]}} f(x_t)\psi_t\big(x,Y(\omega)\big)\mu^X(dx) = \int_{C_{[0,T]}} f(x_t)\psi_t\big(x,Y(\omega)\big)\frac{d\mu^X}{d\mu^W}(x)\mu^W(dx) =$$
$$\int_{C_{[0,T]}} f(x_t)\exp\left\{\int_0^t x_s dY_s - \frac{1}{2}\int_0^t x_s^2 ds + \int_0^t h(x_s)dx_s - \right.$$
$$\left.\frac{1}{2}\int_0^t h^2(x_s)ds\right\}\mu^W(dx)$$

Let $H(x) := \int_0^x h(u)du$, then by the Itô formula

$$H(W_t) = \int_0^t h(W_s)dW_s + \frac{1}{2}\int_0^t h'(W_s)ds$$

and since $h' + h^2 = ax^2 + bx + c$, we have

$$\int_{C_{[0,T]}} f(x_t)\psi_t\big(x,Y(\omega)\big)\mu^X(dx) =$$
$$\int_{C_{[0,T]}} f(x_t)\exp\left\{\int_0^t x_s dY_s - \frac{1}{2}\int_0^t x_s^2 ds + \right.$$
$$\left. H(x_t) - \frac{1}{2}\int_0^t h'(x_s)ds - \frac{1}{2}\int_0^t h^2(x_s)ds\right\}\mu^W(dx) =$$
$$\int_{C_{[0,T]}} f(x_t)e^{H(x_t)}\exp\left\{\int_0^t x_s dY_s - \frac{1}{2}(1+a)\int_0^t x_s^2 ds - \right.$$
$$\left.\frac{1}{2}\int_0^t \big(bx_s + c\big)ds\right\}\mu^W(dx)$$

Now we apply the Girsanov theorem once again: introduce the Ornstein-Uhlnebeck process

$$d\xi_t = -\sqrt{1+a}\,\xi_t dt + dW_t, \quad \xi_0 = 0$$

The induced measure $\mu^\xi$ is equivalent to $\mu^W$ and

$$\frac{d\mu^\xi}{d\mu^W}(x) = \exp\left\{-\int_0^t \sqrt{1+a}\,x_s dx_s - \frac{1}{2}\int_0^t (1+a)x_s^2 ds\right\}, \quad \mu^\xi - a.s.$$

Hence

$$\int_{C_{[0,T]}} f(x_t)\psi_t\big(x, Y(\omega)\big)\mu^X(dx) =$$

$$\int_{C_{[0,T]}} f(x_t)e^{H(x_t)} \exp\left\{ \int_0^t x_s dY_s - \frac{1}{2}(1+a)\int_0^t x_s^2 ds - \right.$$

$$\left. \frac{1}{2}\int_0^t \big(bx_s + c\big)ds \right\} \frac{d\mu^W}{d\mu^\xi}(x)\mu^\xi(dx) =$$

$$\int_{C_{[0,T]}} f(x_t)e^{H(x_t)} \exp\left\{ \int_0^t x_s dY_s - \frac{1}{2}\int_0^t \big(bx_s + c\big)ds + \right.$$

$$\left. \sqrt{1+a}\int_0^t x_s dx_s \right\}\mu^\xi(dx) =$$

$$\int_{C_{[0,T]}} f(x_t)e^{H(x_t)} \exp\left\{ x_t Y_t - \int_0^t Y_s dx_s - \frac{1}{2}\int_0^t \big(bx_s + c\big)ds \right.$$

$$\left. + \sqrt{1+a}\frac{1}{2}(x_t^2 - t) \right\}\mu^\xi(dx),$$

where the latter equality is obtained by the Itô formula (applicable under $\mu^\xi$).

Let $(\xi, \eta, \theta)$ be the solution of the linear system (6.64), then

$$\int_{C_{[0,T]}} f(x_t)\psi_t\big(x, Y(\omega)\big)\mu^X(dx) = \int_{\mathbb{R}^3} f(x_1)\cdot$$

$$\exp\left\{ H(x_1) + x_1 Y_t + \frac{1}{2}\sqrt{1+a}x_1^2 - \frac{1}{2}(c+k)t + x_2 + x_3 \right\} \Gamma(x; m_t, V_t)dx,$$

and (6.63) follows by arbitrariness of $f$. □

## Exercises

**(1)** Let the signal process $X_t = \mathbf{1}_{\{\tau \le t\}}$, where $\tau$ is a nonnegative random variable with probability distribution $G(dx)$. Suppose that the trajectory of

$$Y_t = \int_0^t X_s ds + W_t$$

is observed, where $W$ is a Wiener process, independent of $\tau$.

(a) Is $X_t$ a Markov process for general $G$? Give a counterexample if your answer is negative. Give an example for which $X_t$ is Markov.

(b) Apply the Kallianpur-Striebel formula to obtain a formula for $\mathrm{P}(\tau \le t | \mathcal{F}_t^Y)$.

**(2)** Show that

$$\sigma_t(1) = \exp\left( \int_0^t \pi_s(g)dY_s - \frac{1}{2}\int_0^t \big(\pi_s(g)\big)^2 ds \right).$$

**(3)** (a) Verify the claim of Remark 6.22 directly

(b) Find the solution of the Zakai equation (6.28) in the linear Gaussian case

(**4**) Consider the linear diffusion

$$dX_t = aX_t + dW_t, \quad X_0 = 0,$$

where $W$ is a Wiener process and $a$ is an unknown random parameter, to be estimated from $\mathcal{F}_t^X$. Below $a$ and $W$ are assumed independent.

(a) Assume that $a$ takes a finite number of values $\{\alpha_1, ..., \alpha_d\}$ with positive probabilities $\{p_1, ..., p_d\}$. Find the recursive formulae ($d$ dimensional system of SDEs) for $\pi_t(i) = \mathrm{P}(a = \alpha_i | \mathcal{F}_t^X)$.

(b) Find the explicit solutions to the SDEs in (a).

(c) Does $\pi_t(i)$ converges to $\mathbf{1}_{\{a=\alpha_i\}}$, $i = 1, ..., d$ ? If yes, in what sense ?

(d) Assume that $\mathrm{E}a^2 < \infty$ and find an explicit expression for the orthogonal projection $\widehat{\mathrm{E}}(a|\mathcal{L}_t^X)$ and the corresponding mean square error.

(e) Assume that $a$ is a standard Gaussian random variable. Is the process $X$ Gaussian ? Is the pair $(a, X)$ Gaussian ? Is $X$ conditionally Gaussian, given $a$ ?

(f) Is the optimal nonlinear filter in this case finite dimensional ? If yes, find the recursive equations for the sufficient statistics.

(g) Does the mean square error $P_t = \mathrm{E}\big(a - \mathrm{E}(a|\mathcal{F}_t^X)\big)^2$ converges to zero as $t \to \infty$?

(**5**) Verify that $\mathcal{F}_t^Y \subseteq \mathcal{F}_t^{\bar{W}}$ for the linear Gaussian setting (6.32)

(**6**) Derive the robust version of the Wonham filter (see (6.31) for reference). Elaborate the telegraphic (two dimensional) signal case.

(**7**) Calculate the mean, covariance and one dimensional characteristic function for the Poisson process.

(**8**) Verify the last equality (or equivalently the martingale property of the stochastic integral in this specific case) in (6.54).

(**9**) Let $X_t$ be a finite state Markov chain with values in $\mathbb{S} = \{a_1, ..., a_d\}$, transition intensities matrix $\Lambda$ and initial distribution $p_0$. Let $I_t$ be the $d$-dimensional vector of indicators $\mathbf{1}_{\{X_t=a_i\}}$.

(a) Show that the vector process $M_t = I_t - I_0 - \int_0^t \Lambda^* I_s ds$ is a $\mathcal{F}_t^X$-martingale.

(b) Find its variance $\mathrm{E}M_t M_t^*$

(**10**) For the process $I_t$, defined in the previous exercise, derive the filtering equations for the optimal linear estimate $\widehat{I}_t = \widehat{\mathrm{E}}(I_t|\mathcal{L}_t^Y)$ and the corresponding error covariance, where $Y_t = \int_0^t h(X_s)ds + W_t$.

**Hint:** use the results of Section 3 from the previous chapter

(**11**) Consider a finite automaton with $d$ states. A timer is associated with each state, which is reset upon entering and initiates state transition after a random period of time elapses. The next state is chosen at random, independently of all the timers with probabilities depending on the current state. Let $X_t$ be the state of the automaton at time $t$. Calibrate this model (i.e. choose the timers parameters and transition probabilities, so that $X_t$ is a Markov chain with given intensities matrix $\Lambda$).

(**12**) (a) Derive finite dimensional filtering equations for $T_t(i, j)$ in (6.59) and $J$ in (6.60)

(b) Derive the Zakai type equations for $O_t(i)$, $T_t(i, j)$ and $J$

(c) Elaborate the structure of the optimal filters for telegraphic signal case.

APPENDIX A

# Auxiliary facts

## 1. The main convergence theorems

THEOREM A.1. *(Monotone convergence) Let $Y, X, X_1, \ldots$ be random variables, then*

(a) *If $X_j \geq Y$ for each $j \geq 1$, $\mathrm{E}Y > -\infty$ and $X_j \nearrow X$, then*

$$\mathrm{E}X_j \nearrow \mathrm{E}X.$$

(b) *If $X_j \leq Y$ for each $j \geq 1$, $\mathrm{E}Y < \infty$ and $X_j \searrow X$, then*

$$\mathrm{E}X_j \searrow \mathrm{E}X.$$

COROLLARY A.2. *Let $X_j$ be a sequence of nonnegative random variables, then*

$$\mathrm{E}\sum_{j=1}^{\infty} X_j = \sum_{j=1}^{\infty} \mathrm{E}X_j$$

THEOREM A.3. *(Fatou Lemma) Let $Y, X_1, X_2, \ldots$ be random variables, then*

(a) *If $X_j \geq Y$ for all $j \geq 1$ and $\mathrm{E}Y > -\infty$, then*

$$\mathrm{E}\varliminf_{j\to\infty} X_j \leq \varliminf_{j\to\infty} \mathrm{E}X_j.$$

(b) *If $X_j \leq Y$ for all $j \geq 1$ and $\mathrm{E}Y < \infty$, then*

$$\varlimsup_{j\to\infty} \mathrm{E}X_j \leq \mathrm{E}\varlimsup_{j\to\infty} X_j.$$

(c) *If $|X_j| \leq Y$ for all $j \geq 1$ and $\mathrm{E}Y < \infty$, then*

$$\mathrm{E}\varliminf_{j\to\infty} X_j \leq \varliminf_{j\to\infty} \mathrm{E}X_j \leq \varlimsup_{j\to\infty} \mathrm{E}X_j \leq \mathrm{E}\varlimsup_{j\to\infty} X_j$$

THEOREM A.4. *(Lebesgue dominated convergence) Let $Y, X_1, X_2, \ldots$ be random variables, such that $|X_j| \leq Y$, $\mathrm{E}Y < \infty$ and $X_j \xrightarrow[j\to\infty]{\mathrm{P}-a.s.} X$. Then $\mathrm{E}|X| < \infty$ and*

$$\lim_{j\to\infty} \mathrm{E}X_j = \mathrm{E}X$$

*and*

$$\lim_{j\to\infty} \mathrm{E}|X_j - X| = 0.$$

## 2. Changing the order of integration

Consider the (product) measure space $(\Omega, \mathcal{F}, \mu)$ with $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, i.e. $\mathcal{F}$ is the $\sigma$-algebra of sets $A_1 \times A_2$, $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$, and $\mu = \mu_1 \times \mu_2$, i.e.

$$\mu_1 \times \mu_2(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2), \quad A_1 \in \mathcal{F}_1, \quad A_2 \in \mathcal{F}_2.$$

THEOREM A.5. *(Fubini theorem) Let $X(\omega_1, \omega_2)$ be $\mathcal{F}_1 \times \mathcal{F}_2$-measurable function, integrable with respect to measure $\mu_1 \times \mu_2$, i.e.*

$$\int_{\Omega_1 \times \Omega_2} |X(\omega_1, \omega_2)| d(\mu_1 \times \mu_2) < \infty.$$

*Then the integrals $\int_{\Omega_1} X(\omega_1, \omega_2)\mu(d\omega_1)$ and $\int_{\Omega_2} X(\omega_1, \omega_2)\mu(d\omega_2)$ are well defined for all $\omega_1$ and $\omega_2$ and are measurable functions with respect to $\mathcal{F}_2$ and $\mathcal{F}_1$ respectively:*

$$\mu_2 \left\{ \omega_2 : \int_{\Omega_1} |X(\omega_1, \omega_2)| \mu_1(d\omega_1) = \infty \right\} = 0$$
$$\mu_1 \left\{ \omega_1 : \int_{\Omega_2} |X(\omega_1, \omega_2)| \mu_2(d\omega_2) = \infty \right\} = 0.$$

*Moreover*

$$\int_{\Omega_1 \times \Omega_2} X(\omega_1, \omega_2) d(\mu_1 \times \mu_2) = \int_{\Omega_1} \left[ \int_{\Omega_2} X(\omega_1, \omega_2) \mu_2(d\omega_2) \right] \mu_1(d\omega_1) =$$
$$\int_{\Omega_2} \left[ \int_{\Omega_1} X(\omega_2, \omega_1) \mu_1(d\omega_1) \right] \mu_2(d\omega_2).$$

# Bibliography


[1] B.D.O. Anderson and J.B. Moore, Optimal Filtering, Prentice-Hall, October 1978 available on the author's page:
`http://www.syseng.anu.edu.au/ftp/Publications/by_author/John_Moore/index.html`

[2] V.E. Beneŝ, Exact finite-dimensional filters for certain diffusions with nonlinear drift. Stochastics 5 (1981), no. 1-2, 65–92.

[3] P. Bremaud, Point processes and queues. Martingale dynamics, Springer Series in Statistics, Springer-Verlag, New York-Berlin, 1981

[4] K.L.Chung, R.J.Williams, Introduction to stochastic integration. Progress in Probability and Statistics, 4. Birkhauser Boston, Inc., Boston, MA, 1983.

[5] J.L. Doob, Stochastic processes. John Wiley & Sons, Inc., New York; Chapman & Hall, Limited, London, 1953

[6] E.B.Dynkin, Markov processes. Vols. I, II. , Academic Press Inc., Publishers, New York; Springer-Verlag, Berlin-Gottingen-Heidelberg 1965

[7] Y. Ephraim, N.Merhav, Hidden Markov processes. Special issue on Shannon theory: perspective, trends, and applications. IEEE Trans. Inform. Theory 48 (2002), no. 6, 1518–1569

[8] R.E. Elliott, L. Aggoun and J.B. Moore, Hidden Markov Models: Estimation and Control, Springer-Verlag, 1995

[9] Г.М. Фихтенголц, Курс дифференциального и интегрального исчесления, Наука, Москва.

[10] M. Fujisaki, G. Kallianpur, H.Kunita, Stochastic differential equations for the non linear filtering problem. Osaka J. Math. 9 (1972), 19–40

[11] M.Hazewinkel, S. Marcus, H. Sussmann, Nonexistence of finite-dimensional filters for conditional statistics of the cubic sensor problem. Systems Control Lett. 3 (1983), no. 6, 331–340

[12] A.H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press , 1970

[13] R.E. Kalman, A New Approach to Linear Filtering and Prediction Problems, Trans. ASME Ser. D. J. Basic Engrg. 82 1960 35–45.
(available at `http://www.cs.unc.edu/~Ewelch/kalman/kalmanPaper.html`)

[14] R.E. Kalman, R.S. Bucy,New results in linear filtering and prediction theory, Trans. ASME Ser. D. J. Basic Engrg. 83 1961 95–108.

[15] G. Kallianpur, Stochastic Filtering Theory (Applications of Mathematics Vol 13), Springer-Verlag, 1980

[16] G. Kallianpur, R.L. Karandikar, White Noise Theory of Prediction, Filtering and Smoothing (Stochastics Monographs), Gordon & Breach Science Pub, 1988

[17] G. Kallianpur, C.Striebel, Estimation of stochastic systems: Arbitrary system process with additive white noise observation errors. Ann. Math. Statist. 39 1968 785–801

[18] I. Karatzas, S. Shreve, Brownian motion and stochastic calculus. Graduate Texts in Mathematics, 113. Springer-Verlag, New York, 1988

[19] A.N. Kolmogorov, Foundations of the Theory of Probability. Chelsea Publishing Company, New York, N. Y., 1950

[20] A.N. Kolmogorov, Interpolation and extrapolation of stationary sequences, Izv. Akad. Nauk SSSR, Ser. Mat. 5, 3-14, (1941)

[21] R.Liptser, A.Shirayev, Statistics of random processes, General Thoery and Applications, 2nd ed., Applications of Mathematics (New York), 6. Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2001

[22] R.Sh. Liptser, A.N.Shiryayev, Theory of martingales, Mathematics and its Applications (Soviet Series), 49. Kluwer Academic Publishers Group, Dordrecht, 1989

[23] S.Mitter, Nonlinear Filtering and Stochastic Control (Lecture notes in mathematics) Springer-Verlag , 1983

[24] D. Ocone, Probability densities for conditional statistics in the cubic sensor problem, Math. Control Signals Systems 1 (1988), no. 2, 183–202.

[25] B.Oksendal, Stochastic Differential Equations: an introduction with applications, 5th ed., Springer, 1998

[26] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series. With Engineering Applications. The Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass; John Wiley & Sons, Inc., New York, N. Y.; Chapman & Hall, Ltd., London, 1949. ix+163 pp.

[27] D.Revuz, M.Yor, Continuous martingales and Brownian motion. Third edition. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 293. Springer-Verlag, Berlin, 1999

[28] Yu.A.Rozanov, Stationary Random processes, Holden-Day, 1967

[29] H.J.Kushner On the differential equations satisfied by conditional probablitity densities of Markov processes, with applications. J. Soc. Indust. Appl. Math. Ser. A Control 2 1964 106–119.

[30] A. Makowski, Filtering formulae for partially observed linear systems with non-Gaussian initial conditions, Stochastics 16 (1986), no. 1-2, 1–24

[31] S. Marcus, Algebraic and geometric methods in nonlinear filtering. SIAM J. Control Optim. 22 (1984), no. 6, 817–844.

[32] J.R. Norris, Markov chains, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

[33] L.C.G. Rogers, D. Williams, Diffusions, Markov processes, and martingales. Vol. 1. Foundations and Vol 2. Itô calculus Reprint of the second (1994) edition. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000

[34] A.N. Shiryaev, Probability, 2nd ed., Graduate Texts in Mathematics, 95. Springer-Verlag, New York, 1996

[35] A.N. Shiryaev, Optimal methods in quickest detection problems, Teor. Verojatnost. i Primenen. 8 1963, pp. 26–51

[36] Z. Schuss, Theory and applications of stochastic differential equations. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 1980

[37] R.Stratonovich, "Conditional Markov Processes," Theoretical Probability and Its Applications 5 (1960): 156–178

[38] B.Tsirelson, An example of a stochastic differential equation that has no strong solution, Teor. Verojatnost. i Primenen. 20 (1975), no. 2, 427–430

[39] A.D. Wentzell, A course in the theory of stochastic processes, McGraw-Hill International Book Co., New York, 1981

[40] M. Wonham, Some applications of stochastic differential equations to optimal nonlinear filtering. J. Soc. Indust. Appl. Math. Ser. A Control 2 1965 347–369 (1965).

[41] M. Zakai, On the optimal filtering of diffusion processes. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 11 1969 230–243.

[42] O.Zeitouni, A.Dembo, Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes, IEEE Trans. Inform. Theory 34 (1988), no. 4, 890–893

[43] A.K. Zvonkin, A transformation of the phase space of a diffusion process that will remove the drift, Mat. Sb. (N.S.), 93 (135), 1974, 129-149