

# Analyzing communication schemes using methods from nonlinear filtering

Jochen Bröcker<sup>a)</sup> and Ulrich Parlitz

*Drittes Physikalisches Institut, Universität Göttingen, Bürgerstraße 42-44, D-37073 Göttingen, Germany*

(Received 23 April 2002; accepted 18 June 2002; published 21 February 2003)

The performance of a class of communication systems is investigated in a probabilistic framework. We investigate the bit error probability of the optimal as well as approximately optimal receivers. In general the latter turn out to be unavoidable due to the computational complexity of the former. © 2003 American Institute of Physics. [DOI: 10.1063/1.1499256]

**We investigate a certain class of communication schemes including chaotic systems. Nonlinear filtering theory is employed to obtain a representation of the optimal receiver. Using known results on the filtering process we investigate the bit error probability. It is well known that in general there is no closed form expression of the nonlinear filter. Therefore, in practice approximations are necessary for the nonlinear filter in general and the optimal receiver in particular. We obtain bounds on the approximation error using stability properties of the filter. These bounds also apply to approximations of the optimal receiver.**

## I. INTRODUCTION

Since the invention of telecommunication its technical aspects have been subject to vivid research. Usually the telecommunication engineer's goal is to quantify and to optimally payoff between the demands of low cost, low error and high rate of information transfer. Of course, to obtain non-trivial results certain restrictions on the given setup have to be imposed.

A new era of information theory was heralded by the pioneering works of Shannon and Weaver<sup>1</sup> and Wiener.<sup>2</sup> The book of Shannon and Weaver contains the basic ideas and results on channel (and source) coding. Wiener's work addresses the problem of reconstructing a stationary time series that was received in error due to corruption by noise. Although the aim of both works is to combat a nonreliable transmission channel, the respective setups and assumptions are quite different in detail. While Wiener solves his problem by salient handling of elaborated stochastic tools, Shannon applied elementary methods and a couple of completely new and ingenious ideas.

We will briefly review both concepts in Sec. II. The main reason is that in this paper we will talk about communication, and the reader may have the (completely justified) question, how the presented results are related to Shannon's theory. Probably to his or her disappointment, however, it will turn out that our paper, although concerned with the transmission of messages, is more in spirit of Wiener's work.

As will be explained in Sec. II in Shannon's setup it is assumed that the message is manipulated *before and after* it

is sent, in contrast to Wiener's setup, where a manipulation is possible at the receiver side only. This is an important technical difference between the two theories.

In this paper we will consider a setup that is more related to Wiener's setup. More specifically, we assume that a certain signal  $Y_n$  is transmitted, where  $Y_n$  is a real number and  $n \in \mathbb{N}$  plays the role of time. In this paper, only discrete time processes will be considered. The signal  $Y_n$  is the sum of two parts  $Y_n = X_n + W_n$ , where  $X_n$  in turn is the information carrying quantity and  $W_n$  is an unwanted part or noise.  $X_n$ , however, is not the desired information itself but a physical carrier signal that meets some technical demands imposed on the transmission system. The information is in our case just a stream of bits  $M_n \in \{0,1\}$  which we assume to be independent and identically distributed. The assumption here is that the signal is optimally coded and all redundancy is removed. This information is modulated into  $X_n$ , i.e.,  $X_n$  is (not equal to but) depends on  $M_1 \dots M_n$ . More specific assumptions on this dependence will be imposed in Sec. III. The basic question is: How can we recover the message  $M_n$  from the received time series  $Y_1 \dots Y_n$ ?

Of course this question referred to as the *receiver problem* is very complicated to answer in general and has various theoretical as well as computational aspects. We will try to give a partial answer to this question for a specific setup.

The outline of the paper is as follows: In Sec. II we will give a brief overview over Shannon's and Wiener's work and the main differences. As already mentioned, our paper is more in spirit of Wiener's work. However, this does not mean that it has no significance for results following Shannon's work. This is explained in detail in Sec. II. Since this section is not necessary for an understanding of the rest of the paper it may be skipped at first reading.

In Sec. III we will present the theory of nonlinear filtering. This theory generalizes Wiener's original question to the problem of finding the best estimator (in a mean square sense) of  $X_n$  among *all* possible estimators (and not just the linear ones, as considered by Wiener). Of course, this estimator depends not only on spectral properties of the involved processes but on their entire probability distributions. It turns out that the fundamental quantity emerging from nonlinear filtering theory is the conditional probability of  $X_n$  given  $Y_1 \dots Y_n$ . We will try to give a fairly general presentation of this subject, keeping the mathematical level as elementary as possible.

<sup>a)</sup>Electronic mail: jbroe@physik3.gwdg.de

Section IV explains why for the considered models the results from nonlinear filtering can be used to build receivers. It turns out that the *optimal receiver* evaluates a simple decision criterion that involves the conditional probability calculated in nonlinear filtering. Furthermore, results concerning the asymptotic properties of nonlinear filters are employed to calculate asymptotic bit error rates.

In Sec. III the reader already gets acquainted not only with the benefits but also with the difficulties of nonlinear filtering. A general problem is that the nonlinear filter obeys an infinite dimensional dynamics. Referring to known results we will explain that an explicit expression for the optimal estimator is seldom available in a nonlinear context. Therefore, approximations are essential. This is the subject of Sec. V.

In Sec. VI we will show how error bounds on the approximations can be obtained. The filter can be viewed as a dynamical system on the space of probability distributions which is infinite dimensional but in some cases insensitive with respect to its initial condition. The approximation consists, roughly speaking, of replacing each iteration step of this system by a simpler step. If the filter is stable, then the error obtained in every step will *not* be amplified and therefore a bounded total error remains.

Finally, in Sec. VII we apply these results to the receiver problem. It turns out, that an explicit bound on the maximal achievable bit error rate can be obtained.

**II. SHANNON VS WIENER**

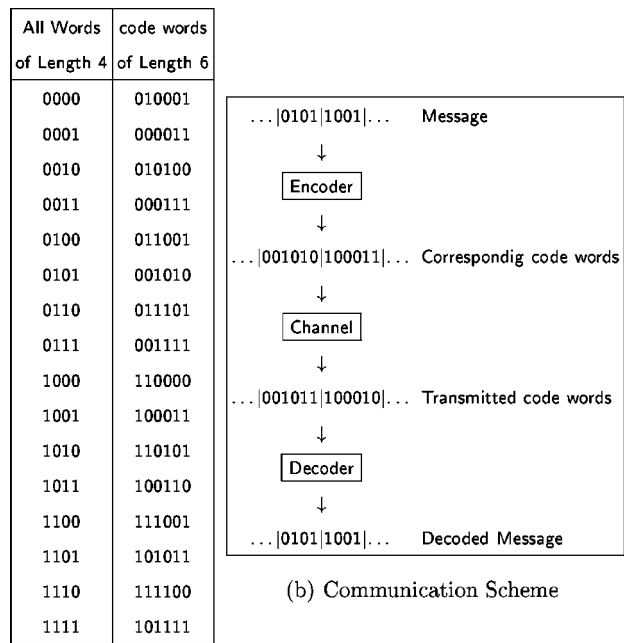
In Shannon’s channel coding theory the problem of information transmission over a not fully reliable channel is considered, i.e., it is assumed that with a certain probability the transmitted message is decoded in error. It is pretty obvious that a certain amount of errors can be corrected at the receiver’s side if a certain amount of the transmitted message is redundant. For example, every bit (assuming the message is represented as a stream of zeros and ones) can be sent twice. The surprising result of Shannon and Weaver<sup>1</sup> was that a finite and fixed amount of redundancy is sufficient to achieve an *arbitrarily small* amount of errors. This will be explained a little more in detail now. The reader already acquainted with this theory may skip this section. It contains no material inevitably necessary for the remainder of the paper.

Suppose we are concerned with binary messages only, i.e., the message is a stream of bits. Suppose the transmitter allows for one bit per second to be sent. At the receiver side we also obtain a bit per second, but with a certain probability this bit is different from the bit that was transmitted. This behavior can be expressed in a  $2 \times 2$ -matrix,

$$p_{ij} := \text{Prob. of receiving } i \text{ when } j \text{ was sent,}$$

where  $i, j = 1, 2$ .

Let us now formalize the procedure of introducing redundancy. The basic idea is to use a *code* as follows (see also Fig. 1). Consider all possible words of, say,  $N$  bits. There are  $2^N$  such words. Let  $R \leq 1$  and specify a subset containing only  $2^{\lfloor RN \rfloor}$  words. Here  $\lfloor \cdot \rfloor$  means the integer part. This sub-



(a) Codetable

(b) Communication Scheme

FIG. 1. Panel (a) shows all words of length 4 (first column). The second column contains possible code words of length 6. Panel (b) shows a communication setup using the code.

set is called a *code of rate R*. The elements of this set are called code words, hence there are  $2^{\lfloor RN \rfloor}$  code words. We can transmit a message using this code by simply dividing the message into blocks of length  $\lfloor RN \rfloor$  (at most  $2^{\lfloor RN \rfloor}$  different blocks can appear) and assigning a code word to each such block. Now the code word can be sent through the channel. Recall that the code word has length  $N$ , but the message block that is assigned to the code word has length only  $\lfloor RN \rfloor$ . So using the code effectively reduces the transmission rate by a factor of  $R$ . In Fig. 1 we used  $N=6$  and  $R=2/3$ , i.e.,  $1/3$  of the bits are redundant.

If a code word is transmitted, at the receiver’s side a word of  $N$  bits obtains. However, some of the  $N$  bits are received in error (in Fig. 1 the last bits of both code words are incorrect). So a received block of  $N$  bits forms a word that is usually *not* a code word (although this may accidentally be the case). Here in general a *decoder* is needed that maps any word of length  $N$  back onto a code word. For example, we may take the code word that has the smallest amount of bits different from the received word (minimum Hamming distance). Finally, inverting the message-code assignment, we get back what is supposed to be the transmitted message.

For a given channel, the performance of this scheme obviously depends on the rate  $R$ , the length  $N$ , the chosen set of code words and the decoder. The outstanding theorem of Shannon states that associated to the channel there is a number  $C$  called the *capacity* with the following property: By taking  $N$  sufficiently large we can find a code of rate  $R$  arbitrarily close to  $C$  and a decoder yielding arbitrarily small transmission error. This is called the direct part of the coding theorem. If however  $R$  is larger than  $C$ , the error is bounded

away from zero. This statement is called the converse part. Actually, Shannon and Weaver<sup>1</sup> proved this result (together with an explicit expression for  $C$ ) in the case of memoryless channels, i.e., the probability that a transmission error occurs at time  $n$  does not depend on what has happened in the past.

Note that in a practical situation to establish a reliable communication with rates close to the channel capacity it is necessary to manipulate the information carrying quantity *before and after* it is sent. The situation considered by Wiener<sup>2</sup> that will now be described briefly however does not permit a manipulation of the signal before transmission, which from a communication theoretic viewpoint constitutes the main difference between Wiener's and Shannon's setup.

As in Shannon and Weavers work, Wiener considers a stationary stochastic process  $X_n, n \in \mathbb{Z}$  as the quantity carrying the desired information (in contrast to Shannon however, it is not explicitly considered as a message). He assumes that at the receiver the process  $Y_n = X_n + S_n$  obtains, where  $S_n$  is the unwanted part or the noise. Hence effectively he assumes a very specific form of a channel.

Wiener now considers the problem of reconstructing  $X_n$  from  $Y_n$  in a linear manner. More specifically, taking the ansatz,

$$\hat{X}_n := \sum_{k=-\infty}^{\infty} a_k^{(n)} Y_k \quad (1)$$

and the least mean square optimality criterion,

$$E(X_n - \hat{X}_n)^2 = \min!$$

he obtains an equation (Wiener–Hopf equation) for the coefficients,  $a_k^{(n)}$ , where  $E(\cdot)$  denotes the mathematical expectation. It turns out that only auto and cross correlations of  $X_n$  and  $Y_n$  enter the Wiener–Hopf equation. Wiener studies a variety of related problems, mainly differing in how much  $Y$ 's enter the right-hand side of the ansatz (1). The resulting Wiener–Hopf equations are tackled by spectral methods.

Wiener finished his work already in 1942, but due to its significance for war time issues (radar tracking, automatic fire control) it was not classified and published until 1950. Since then, a huge amount of improvements and generalizations to Wiener's theory have been conceived. The most important were probably the Kalman filter,<sup>3</sup> where Gaussian processes admitting a state space description where considered and the extension to nonlinear stochastic differential equations given by Stratonovich and independently by Kushner (see Ref. 4 for an overview).

All the works following Wiener show various attempts to the solution of the general problem: Which was the signal that led to the aquired data? Looking back to Shannon's setup, we see that this is basically the decoding problem: which was the code word that led to the received data? One may even say that *first* Wiener's theory (or its extensions) have to be applied to build good receivers and *then* results in spirit of Shannon's work (i.e., coding) are applied to improve the performance of the receiver. In other words, to apply error correcting codes the receiver problem already has to be solved! Coding theory does not circumvent the receiver problem, it merely addresses the question how the perfor-

mance of a given receiver can be improved using manipulations of the message before and after transmission.

Actually, Shannon's classical result can be established using a quite suboptimal receiver. Nevertheless, the tradeoff between  $N$  and the error depends heavily on the decoder, which is important in practical applications. Furthermore, Shannon's result is valid in full generality only for memoryless channels. For channels having memory, the problem turns out to be quite difficult. In general, a different  $C$  appears in the direct and the converse part of the coding theorem, i.e., it is stated that at rates  $< C_1$ , reliable communication is definitely possible and at rates  $> C_2$  definitely not, but in general  $C_1 < C_2$ . Furthermore, the results usually depend heavily on the employed decoders. In general, to obtain a larger  $C$  in the direct coding theorem, more sophisticated decoders are necessary, probably having a complexity prohibiting their practical implementation.

Thus for application and extension of Shannon's theorem, good decoders are mandatory. By good we mean as reliable as necessary to obtain the direct coding theorem at high rates, but as simple as possible to be implementable in applications. Of course, in this paper we will not solve the problem completely. The basic aim of this paper is to convince the reader that a possible route to good decoders goes via the beforementioned theory of filtering.

### III. NONLINEAR FILTERING

In this section we present the theoretic background of the paper, namely the theory of nonlinear filtering and some auxiliary stochastic tools.

In general message transmission is done employing a (usually electronic) device called the transmitter. The internal state of the transmitter at time instant  $n \in \mathbb{N}$  is assumed to be determined by a variable  $X_n$  in an appropriate space. The state  $X_n$  depends on its predecessors  $X_1 \cdots X_{n-1}$ , the message to be transmitted and some additional random influences. In this paper we will only allow for the simplest possible messages, namely, a sequence  $\{M_n\}$  of independent, identically distributed random variables assuming the values 0 or 1 only. Furthermore, we assume  $M_n$ , the message element at time  $n$ , to be independent of  $X_1 \cdots X_{n-1}$ , whence the message element has influence only on the present and future evolution of the transmitter state.

Based on this general consideration a lot of transmitter models can be considered differing basically in how much past information enters the future evolution of  $X_n$ . The simplest model of interest obtains if we assume that  $X_n$  is, up to random disturbances, determined by  $M_n$  and  $X_{n-1}$ .

Usually a transmitter is employed to generate a signal that is capable of passing through a channel. For example, consider a radio transmitter. The channel here is the atmosphere and the signal transmitted by the channel is the voltage at the antenna, which is a function of the transmitter state. Of course, atmospheric disturbances will take place and lead to a corruption of the transmitted signal. In our model channel noise is taken into account by additive iid random variables. Thus our model of a transmission channel is again a very simple one, namely we assume that the chan-

nel output  $Y_n$  is a function of the transmitter state corrupted by additive noise. As a simple example let us consider the following stochastic process on the unit interval:

$$X_{n+1} = f_{M_{n+1}}(X_n), \tag{2}$$

where

$$\begin{aligned} f_0: [0,1] &\rightarrow [0,1], & x &\rightarrow |2x-1|, \\ f_1: [0,1] &\rightarrow [0,1], & x &\rightarrow 1-|2x-1|, \end{aligned} \tag{3}$$

are the usual and the inverted tentmap. As received signal we take simply  $X_n$  itself. As random noise due to channel disturbances we take random variable  $\{W_n\}$  which are independent, have a centered normal distribution with unit variance and are independent of  $\{X_n\}$ . The signal arriving at the receiver now is assumed to be

$$Y_n = X_n + \sigma W_n,$$

where  $\sigma$  is a given positive constant. The basic question, the *receiver problem* now is: Assume a sample of values  $Y_1, \dots, Y_n$  has been recorded. What is the value of the message  $M_n$ ?

Let us mention also the basic question of nonlinear filtering, which reads slightly different: Assume a sample of values  $Y_1, \dots, Y_n$  has been recorded. What is the value of the system state  $X_n$ ?

At first sight it seems superfluous to calculate the estimator for  $X_n$ , since we are eventually interested in getting estimators for the message  $M_n$  rather than  $X_n$ . It will, however, turn out that all these problems can be encompassed by calculating a fundamental device, namely the conditional probability of  $X_n$  given  $Y_1, \dots, Y_n$ . This in turn is the main aim of filtering theory. How it can be employed to solve the receiver problem will be subject to Sec. IV.

Of course, the answer to the basic question in filtering cannot be given with infinite accuracy due to the unknown noise  $W_n$  (except if  $\sigma=0$ ). What is desired are estimators [i.e., functions  $\hat{X}_n = \hat{X}_n(Y_1, \dots, Y_n)$ ] having a good *average* performance. Wiener's theory provides the best *linear* estimator with respect to an average quadratic error criterion. For nonlinear systems, however, this estimator is outperformed by nonlinear estimators calculated by the theory of nonlinear filters which we now will present.

Let us formalize now our basic model of a transmitter. For a comprehensive presentation of the basic notions of probability theory used in the following we recommend Ref. 5. A few notations are explained in the Appendix. Let  $(\Omega, P, \mathcal{A})$  be a probability space. Let  $E$  be a complete separable metric space and  $\{X_n\}_{n \in \mathbb{N}_0}: \Omega \rightarrow E$  (the transmitter state) as well as  $\{M_n\}_{n \in \mathbb{N}}: \Omega \rightarrow \{0,1\}$  (the message) be random processes. Furthermore we assume that the joint process  $\{M_{n+1}, X_n\}_{n \in \mathbb{N}_0}$  is Markov, the variables  $\{M_n\}$  are all identically distributed and  $M_{n+1}$  is independent of  $\{M_{k+1}, X_k\}_{k=0 \dots n-1}$ . Let  $\mu(A) := P(X_0 \in A)$  and  $p_i := P(M_n = i)$ , where  $i=0$  or  $1$ . Then the initial distribution of the process  $\{M_{n+1}, X_n\}$  is given by  $P(X_0 \in A, M_1 = i) = \mu(A)p_i$  and the transition probability is

$$\begin{aligned} P(X_n \in A, M_{n+1} = i | X_{n-1} = x, M_n = j) \\ = p_i \cdot \varphi_j(A, x), \end{aligned}$$

where we define

$$\varphi_j(A, x) := P(X_n \in A | X_{n-1} = x, M_n = j). \tag{4}$$

Our setup allows us to assume that conditional probabilities as well as conditional expectations are *regular*. This means, for  $A$  held fixed,  $\varphi(A, x)$  is an integrable function in  $x$  and for any  $x$  held fixed,  $\varphi(\cdot, x)$  is a measure (see Ref. 5 for regular conditional probabilities).

It is easy to see that  $\{X_n\}$  alone is a Markov process with transition probability  $\varphi(A, x) := P(X_n \in A | X_{n-1} = x) = \sum_j \varphi_j(A, x)p_j$ . Let us shortly stop here for a short remark about *canonical representations* of a stochastic process. If  $\{X_n\}_{n \geq 0}$  is a Markov process on a probability space  $(\Omega_X, P_X, \mathcal{B}_X)$  in discrete time (i.e.,  $n \in \mathbb{N}_0$ ) assumed to have values in a complete separable metric space (i.e., a polish space)  $E$  equipped with a Borel  $\sigma$ -algebra  $\mathcal{B}_E$ , we can always assume the probability space to be canonical, i.e.,  $\Omega_X = E^\infty$ ,  $\mathcal{B}_X = \mathcal{B}_E^\infty$ . According to Kolmogorov's theorem,  $P_X$  is well defined by specifying the finite dimensional distributions of  $\{X_n\}$ . Since  $\{X_n\}$  is Markov, the finite dimensional distributions are determined by the distribution  $\nu$  of  $X_0$  and the transition kernel,

$$\varphi: \mathcal{B}_E \times E \rightarrow \mathbb{R}_+,$$

$$(A, x) \rightarrow \varphi(A, x) := P(X_{n+1} \in A | X_n = x)$$

by the equation

$$\begin{aligned} P_X^\nu(X_0 \in A_0, \dots, X_k \in A_k) \\ = \int_{A_k \times \dots \times A_0} \varphi(dx_k, x_{k-1}) \cdots \varphi(dx_1, x_0) \nu(dx_0), \end{aligned}$$

where  $A_0, \dots, A_k \in \mathcal{B}_E$ . The dependence on  $\nu$  is denoted by the superscript and in fact we consider not only one measure  $P_X$  on  $\Omega_X$  but a whole family  $P_X^\nu$ . If  $\nu$  assigns probability one to a single point  $z \in E$  we write  $P_X^z$ . Further properties of Markov processes (mainly concerning their ergodic behavior) are summarized in the Appendix.

Now we turn to the channel. Let  $\{W_n\}_{n \geq 1}$  be a process of i.i.d. random variables having values in  $\mathbb{R}$ . We assume that the  $W_n$  have a probability density function  $g$  with respect to Lebesgue measure  $\lambda$ . We assume the  $\{W_n\}$  to be of zero mean and unit standard deviation. By using Kolmogorov's theorem again we can assume the corresponding probability space to be canonical, i.e.,  $\Omega_W = \mathbb{R}^\infty$ ,  $\mathcal{B}_W = \mathcal{B}^\infty$ , where  $\mathcal{B}$  is the Borel algebra on  $\mathbb{R}$ . The probability measure is defined by the finite dimensional distributions,

$$P_W(W_1 \in A_1, \dots, W_k \in A_k) = \prod_{i=1}^k \int_{A_i} g(x) dx.$$

Furthermore, let  $\{W_n\}$  be independent of  $\{X_n\}$ . It is well known that the corresponding probability space covering both  $\{X\}$  and  $\{W\}$  can be chosen as  $\Omega := \Omega_X \times \Omega_W$ ,  $P^\nu := P_X^\nu \times P_W$ ,  $\mathcal{B} := \mathcal{B}_X \otimes \mathcal{B}_W$ . The filtration of the process  $(X_n, W_n)$  is denoted by  $\mathcal{F}_n$ . Expectation with respect to  $P^\nu$  or  $P^x$  will be denoted by  $E_\nu$  or  $E_x$ , respectively.

Finally we introduce a third process called *measurement process*. Let  $h: E \rightarrow \mathbb{R}$  be a measurable function. Now define the process  $\{Y_n\}_{n \geq 1}$  by

$$Y_n = h(X_n) + \sigma W_n.$$

The  $\sigma$ -algebra  $\sigma(Y_1, \dots, Y_k)$  is denoted by  $\mathcal{G}_n$ . Note that  $W_n$  and also  $Y_n$  are defined for  $n \geq 1$ , while  $X_n$  is defined for  $n \geq 0$ .

The aim of *filtering* now is to estimate the ‘‘hidden’’ process  $\{X_n\}$  from the measurements  $\{Y_n\}$  in a causal manner, i.e., the estimator  $\hat{X}_n$  of  $X_n$  shall depend only on  $Y_1, \dots, Y_n$ , that is, it shall be  $\mathcal{G}_n$ -measurable. It can be shown (see, e.g., Ref. 6) that for any such estimator,

$$E[(X_n - \hat{X}_n)^2] \geq E[(X_n - E(X_n | \mathcal{G}_n))^2],$$

while if the equality holds,  $\hat{X}_n = E(X_n | \mathcal{G}_n)$  almost sure. This is a general property of the conditional expectation and also holds for estimators  $\hat{f}_n$  of  $f(X_n)$  for any measurable function for which  $E|f(X_n)| < \infty$ , e.g., if  $f$  is bounded and continuous.

To calculate conditional expectations we consider the *filtering process*  $\pi_n^\nu$  defined as

$$\pi_n^\nu(A) := P^\nu(X_n \in A | \mathcal{G}_n),$$

where the conditioning on  $\mathcal{G}_n$  can be viewed just as a shorthand notation for  $Y_1 \cdots Y_n$ . Define also

$$\pi_n^\nu(f) := E_\nu(f(X_n) | \mathcal{G}_n) = \int f(x) \pi_n^\nu(dx)$$

for a given bounded continuous function  $f: E \rightarrow \mathbb{R}$ . The aim of filtering is to give convenient formulas for  $\pi_n^\nu$  as an explicit function of  $Y_1, \dots, Y_n$ .

It follows from the Kallianpur–Striebel formula (see Ref. 6) that

$$\pi_n^\nu(f) = c \cdot \int_E f(z) \cdot g\left(\frac{Y_n - h(z)}{\sigma}\right) \int_E \varphi(dz, x) \pi_{n-1}^\nu(dx), \quad (5)$$

where  $c$  is the normalization constant,

$$c = \int_E g\left(\frac{Y_n - h(z)}{\sigma}\right) \int_E \varphi(dz, x) \pi_{n-1}^\nu(dx).$$

Let  $\mathcal{M}_E$  be the space of all finite positive measures on  $E$ . Elements of  $\mathcal{M}_E$  will be denoted by small Greek letters in the following. Furthermore, denote by  $\mathcal{P}_E$  the subset of probability measures. Define the operator,

$$S: \mathbb{R} \times \mathcal{M}_E \rightarrow \mathcal{P}_E,$$

$$S(y, \nu)(A) := c \cdot \int_A g\left(\frac{y - h(z)}{\sigma}\right) \int_E \varphi(dz, x) \nu(dx), \quad (6)$$

where  $c$  is again the normalizing constant. So  $S(y, \cdot)$  maps finite positive measures to probability measures. With this definitions we have the iterative formula,

$$\pi_{n+1}^\nu = S(Y_n + 1, \pi_n^\nu).$$

Furthermore,  $\pi_0^\nu = \nu$ . The process  $\pi_n^\nu$ , called the filtering process is a random process on  $\mathcal{P}_E$  and turns out to be a Markov process. Introducing the weak topology on  $\mathcal{P}_E$ , the transition kernel

$$\Pi(\Lambda, \mu) := P^\nu(\pi_n^\nu \in \Lambda | \pi_{n-1}^\nu = \mu)$$

turns out to be Feller, i.e., for any function  $F: \mathcal{P}_E \rightarrow \mathbb{R}$  bounded and continuous in the weak topology, also  $\Pi F(\nu) := \int F(\mu) \Pi(d\mu, \nu)$  is bounded and continuous in the weak topology. To compute average quantities in the filtering problem like average filtering errors or approximation errors, ergodic properties of the filtering process are required. The main results needed in this paper are due to Stettner<sup>7</sup> and Kunita<sup>8</sup> to which we refer the interested reader. This section is finished with the presentation of three examples serving as standard models throughout this paper.

*Example 1: (CSK scheme)* Let  $\{M_n\}$  again be a binary message and  $f_0, f_1$  be two continuous mappings of a closed interval  $I$  (which might be the whole real line) to itself. Let  $X_0$  be a random variable and define the process,

$$X_{n+1} = f_{M_{n+1}}(X_n).$$

The measurement process is taken as

$$Y_n = X_n + \sigma W_n,$$

where  $W_n$  is Gaussian (noise). The transition probability of  $X_n$  is

$$\varphi(A, x) = p_1 \delta_{f_1(x)}(A) + p_0 \delta_{f_0(x)}(A),$$

where the *delta measure*  $\delta_z(A)$  is 1 if  $z \in A$  and 0 else. Often  $I$  is chosen as the unit interval and  $f_0, f_1$  are piecewise expanding Markov maps. In this case this setup is called *chaotic shift keying* (CSK) scheme.

If the distribution of  $X_n$  has a density  $h$  with respect to Lebesgue measure, then so has the distribution of  $X_{n+1}$ , and the Markov transition kernel translates into an operator on  $L_1$ , called the *Frobenius–Perron–operator* (FPO). The FPO of a CSK-scheme is given by

$$\mathcal{L}h(x) = p_1 \sum_{y \in f_1^{-1}(x)} \frac{h(y)}{|f_1'(y)|} + p_0 \sum_{y \in f_0^{-1}(x)} \frac{h(y)}{|f_0'(y)|}.$$

If the distribution  $\nu$  of  $X_0$  has an  $L^1$  density  $\pi_0(x)$  with respect to Lebesgue measure, then also the filtering process  $\pi_n^\nu$  has a representation in terms of densities (denoted by  $\pi_n^\nu(x)$ ) given by

$$\pi_n^\nu(x) = c \cdot g\left(\frac{Y_n - x}{\sigma}\right) \mathcal{L} \pi_{n-1}^\nu(x),$$

where again  $c$  is normalization and  $g$  the density of  $W_n$ .

Piecewise expanding Markov maps are thoroughly investigated in Ref. 9. It is shown that there exists an invariant measure  $\nu$  on the unit interval which has a density  $h$  with respect to Lebesgue measure that is of bounded variation. Furthermore, if  $f$  is aperiodic, this measure is exact (in particular, ergodic and the only one having a density with respect to Lebesgue measure). The density  $h$  is everywhere positive and for any continuous function  $f$ ,

$$\mathcal{L}^n f(z) \rightarrow \int f dx \cdot h(z)$$

uniformly in  $z$ . This analysis depends entirely on the FPO, and it turns out that much of it carries over to our setup.

Especially there is an invariant measure  $\nu$  on the unit interval which has a density  $h$  with respect to Lebesgue measure that is of bounded variation. The corresponding measure  $P^\nu$  is therefore stationary and the finite dimensional distributions have all densities. Furthermore it can be shown that under a modified aperiodicity assumption, any function  $g \in L_1(\nu)$  on the interval which is invariant under  $\varphi$  is  $\nu$ -almost sure equal to a constant. It follows then from Lemma 16 of the Appendix that  $P^\nu$  is even ergodic.

The relevance of CSK-schemes as models for a real time electronic transmitting device may of course be doubted. They are however subject to vivid research on a more abstract level. They are used to generate signals having desired statistical properties (see, e.g., Ref. 10).

*Example 2: (Mixing process)* Let  $W'_n$  be a process of iid random variables on  $\mathbb{R}^d$  having a continuous and strictly positive pdf  $d(x)$  with respect to Lebesgue measure. Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}^d$  be a continuous and bounded function. Then the process

$$X_{n+1} = f(X_n) + W'_n$$

is a Markov process satisfying the conditions of theorem (13). The transition kernel is given by

$$\varphi(A, x) = \int_A d(z - f(x)) dz$$

and the FPO by

$$\mathcal{L}h(x) = \int_{\mathbb{R}^d} d(x - f(z))h(z) dz.$$

This setup can also be extended to a message transmission scheme by letting  $\{M_n\}$  be the usual message process and taking two functions  $f_0, f_1: \mathbb{R}^d \rightarrow \mathbb{R}^d$ , both bounded and continuous.  $\{X_n\}$  is now defined by

$$X_{n+1} = f_{M_{n+1}}(X_n) + W'_n.$$

Again  $\{X_n\}$  is a Markov process satisfying the conditions of theorem (13). The transition kernel is given by

$$\begin{aligned} \varphi(A, x) &= p_0 \varphi_0(A, x) + p_1 \varphi_1(A, x) \\ &= \int_A p_0 \cdot d(z - f_0(x)) + p_1 \cdot d(z - f_1(x)) dz \end{aligned}$$

and the FPO by

$$\mathcal{L}h(x) = \int_{\mathbb{R}^d} p_0 \cdot d(x - f_0(z)) + p_1 \cdot d(x - f_1(z))h(z) dz.$$

Again, if the distribution  $\nu$  of  $X_0$  has an  $L^1$  density  $\pi_0(x)$  with respect to the Lebesgue measure, then also the filtering process  $\pi_n^\nu$  has a representation in terms of densities (denoted by  $\pi_n^\nu(x)$ ) given by

$$\pi_n^\nu(x) = c \cdot g\left(\frac{Y_n - x}{\sigma}\right) \mathcal{L}\pi_{n-1}^\nu(x),$$

where again  $c$  is normalization.

*Example 3: (Linear Gaussian process):* The first system class for which the filtering process was calculated explicitly

was of course the linear Gaussian case. This example contains no message transmission and is presented here as a standard example of filtering,

$$X_{n+1} = F_n X_n + a_n + W'_n,$$

where  $W'_n$  has a Gaussian distribution with covariance matrices  $\{R_n\}$ ,  $\{F_n\}$  is a sequence of  $d \times d$ -matrices, and  $\{a_n\}$  a sequence of  $d$ -dimensional vectors. Furthermore assume  $X_0$  has a Gaussian distribution with covariance matrix  $\Gamma_0$ . Let the measurement process be given by the equation,

$$Y_n = G_n X_n + b_n + W_n,$$

where  $W_n$  has a Gaussian distribution with covariance matrices  $S_n$ ,  $\{G_n\}$  is a sequence of  $d \times l$ -matrices and  $\{b_n\}$  a sequence of  $l$ -dimensional vectors. Then,

$$\pi_n(x) = \frac{1}{\sqrt{(2\pi)^d \det \Gamma_n}} \exp[-0.5(x - \mu_n) \Gamma_n^{-1} (x - \mu_n)],$$

where  $\Gamma_n$  and  $\mu_n$  are given by

$$\Gamma_{n+1}^{-1} = (F_n \Gamma_n F_n^t + R_n)^{-1} + G_n^t S_n^{-1} G_n,$$

$$\mu_{n+1} = F_n \mu_n + a_n$$

$$+ \Gamma_{n+1} G_n^t S_n^{-1} (Y_{n+1} - G_n (F_n \mu_n + a_n) - b_n).$$

These equations are due to Kalman<sup>3</sup> and a direct consequence of Eq. (5).

The Kalman filter is an example where the filtering process admits a parametrization. This is,  $\pi_n(x) = \pi(x, \theta_n)$  and  $\theta_n$  is given iteratively by a finite dimensional dynamical system of the form  $\theta_n = F(Y_n, \theta_{n-1})$ . We will discuss in Sec. V that this is in some sense a very unusual situation.

#### IV. MESSAGE TRANSMISSION

The receiver is any device that produces a reasonable estimate  $\hat{M}_n$  for the actual message  $M_n$  based on the time series  $Y_1, \dots, Y_n$ . We will show that this problem can be solved if the conditional probability  $\rho_n(m) := P(M_n = m | \mathcal{G}_n)$  is known.

We now give an expression for  $\rho_n(m)$  in terms of the filtering process. This establishes the beforementioned condition between the receiver problem and the theory of nonlinear filtering.

*Lemma 4:* Let [slightly different from the examples of Sec. III and formula (4)],

$$\varphi_1(A, x) := p_1 P(X_n \in A | X_{n-1} = x, M_n = 1),$$

$$\varphi_0(A, x) := p_0 P(X_n \in A | X_{n-1} = x, M_n = 0).$$

Then we have (with  $\varphi$  defined as before)

$$\varphi = \varphi_1 + \varphi_0.$$

Then

$$\rho_n^\nu(m) = \int \frac{d\varphi_m \pi_{n-1}^\nu}{d\varphi \pi_{n-1}^\nu} \pi_n^\nu(dx).$$

*Proof:* This follows easily using change of measure like in the Kallianpur–Striebel formula. An informal derivation is

given in Ref. 11. For the meaning of the Radon–Nykodim derivative  $d\cdot/d\cdot$  see Ref. 5 and the Appendix.  $\square$

The performance of a binary communication channel is usually measured by the bit error rate (BER) which is defined as

$$\text{BER} = \frac{1}{N} \sum_{k=1}^N |M_k - \hat{M}_k|,$$

where  $M_k$  is the transmitted message and  $\hat{M}_k$  is the received message. It should be kept in mind that in general, the bit error rate is a random quantity and depends on  $N$ . It is an interesting question whether the bit error rate converges to a (possibly random) limit or not.

In any case (ergodic or stationary or nothing) we will call

$$P^\nu(M_k \neq \hat{M}_k)$$

the bit error probability (denoted by  $\text{BEP}_k^\mu$ ) where  $\hat{M}_k$  is used as an estimator for  $M_k$  and  $\nu$  is the distribution of  $X_0$ . We now define the receiver  $\hat{M}_k$  we will use throughout the rest of this paper.

*Definition 5:* We set  $\hat{M}_k = 1$  if  $\rho_k^\nu(1) > \rho_k^\nu(0)$  and  $\hat{M}_k = 0$  else. Since in fact  $\hat{M}_k$  depends on  $\nu$  we will write  $\hat{M}_k^\nu$  in the following.

Obviously,  $\hat{M}_k^\nu$  is a function of  $Y_1 \cdots Y_k$ . Furthermore, this estimator turns out to have a certain minimum property. If  $\bar{M}_k$  is an estimator depending on  $Y_1 \cdots Y_k$  and assuming the values 0 or 1 only, it can be shown that

$$P^\nu(M_k = \bar{M}_k) = E_\nu(\rho_k^\nu(\bar{M}_k)),$$

whence we have that for any such estimator,

$$P(M_k = \bar{M}_k) \leq P(M_k = \hat{M}_k^\nu).$$

Hence the estimator  $\hat{M}_k^\nu$  yields the least bit error probability and, in this sense, provides an optimal estimator. Our first theorem concerning ergodicity of the bit error rate can be obtained using ergodic theory of nonlinear filtering.

**Theorem 6:** *Suppose  $\mu$  is a  $\varphi$ -invariant measure. Then the bit error rate*

$$\text{BER}_N = \frac{1}{N} \sum_{k=1}^N |M_k - \hat{M}_k^\mu|$$

*converges almost surely to a (possibly random) limit. If  $\mu$  is even a unique  $\varphi$ -invariant measure satisfying condition (A2), then the limit is almost sure equal to a constant.*

*Proof:* We only sketch the main ideas. A full outline will be given elsewhere. If  $\mu$  is  $\varphi$ -invariant it follows from theorem 1 in Ref. 7 that the distribution of  $\pi_n^\mu$  converges to an invariant measure of  $\Pi$ , the transition semigroup of the filter. Calling this invariant measure  $\Phi$  it turns out that the joint random variable  $(M_{n+1}, \pi_n^\mu)$  has asymptotic distribution  $p_i \cdot \Phi$ . Furthermore, the process  $\{M_{n+1}, \pi_n^\mu\}$  is Markov having invariant measure  $\Phi \cdot p_i$ . So the compound process is asymptotically stationary. Since  $|M_k - \hat{M}_k|$  can be expressed as a function of  $M_k$ ,  $\pi_k^\mu$ , and  $\pi_{k-1}^\mu$  it turns out to be stationary as well. The first assertion now follows from

Birkhoff’s theorem. The second assertion follows if the filtering process turns out to be ergodic. Under condition (A2), the invariant measure  $\Phi$  of  $\Pi$  having barycenter  $\mu$  is unique (see Ref. 7, Theorem 2). However, any other  $\Pi$ -invariant measure *must* have a barycenter which is  $\varphi$ -invariant. Since there are no such measures except for  $\mu$  it turns out that  $\Phi$  is the unique invariant measure of the filtering process. By Lemma 15, (3) the filtering process is ergodic.  $\square$

Concerning the asymptotic properties of the bit error probability we have the following theorem:

**Theorem 7:** *If  $\mu$  is an  $\varphi$ -invariant measure satisfying condition (A2), then the  $\text{BEP}_k^\mu$  is convergent and decreasing in  $k$ . Call the limit  $\text{BEP}^\mu$ . If furthermore  $\nu$  satisfies the assumption  $\nu \varphi^k \rightarrow \mu$ , then  $\text{BEP}_k^\nu \rightarrow \text{BEP}^\mu$ .*

*Proof:* This follows from the fact that the bit error probability  $\text{BEP}_k^\nu$  can be written as

$$\begin{aligned} \text{BEP}_k^\nu = \frac{1}{2} E_\nu \left[ 1 - \int \left| \int \frac{1}{\sigma} g \left( \frac{y-h(x)}{\sigma} \right) \right. \right. \\ \left. \left. \times (\varphi_1 - \varphi_0) \pi_{n-1}^\nu(dx) \right| dy \right], \end{aligned}$$

which is an expectation over a concave function of  $\pi_{n-1}^\nu$ . The theorem now follows from the results in Ref. 7.  $\square$

We remark that the transmitter model introduced in example 2 actually satisfies the conditions of Theorem 13 (see Appendix), hence there is a unique invariant measure satisfying condition (A2). Thus, both theorems apply.

Theorems 6 and 7 may be of restricted practical use since a quite restricted receiver model is assumed. However, the main purpose was to show that theoretical methods of nonlinear filtering translate into the framework of message transmission.

## V. APPROXIMATIONS OF THE FILTERING PROCESS

It was already stated informally in Sec. III that the filtering process in general has a very high complexity rendering it unfeasible for direct applications. We will make this a little more precise in this section by stating some well known results about (non)existence of finite dimensional filters due to Sawitzki<sup>12</sup> we will explain later in this section.

For this suitable approximations of the filtering process turn out to be essential. This is the main subject of this section. A lot of methods have been conceived. For an overview and further references, see Ref. 11.

The idea of *parametric approximation* is to consider a set  $\mathcal{P}$  of strictly positive integrable functions on  $E$  which are normalized, i.e.,  $\int p d\lambda = 1$  for all  $p \in \mathcal{P}$  and a fixed (not necessarily finite) carrier measure  $\lambda$  on  $E$ . A *parametrization* of  $\mathcal{P}$  is a mapping,

$$p: \Theta \rightarrow \mathcal{P}; \theta \mapsto p(\cdot, \theta),$$

where  $\Theta$  is a subset of a finite dimensional vector space. The parametrization is called *faithful* if  $\theta_1 \neq \theta_2$  necessarily yields  $p(\cdot, \theta_1) \neq p(\cdot, \theta_2)$ . Such a set  $\mathcal{P}$  together with a faithful  $p$  and a carrier measure  $\lambda$  can be considered as a parametrized family of probability density functions (pdf’s) with respect to  $\lambda$  as well as a parametrized set in  $\mathcal{P}_E$ .

The basic idea of parametric approximation is to choose a parametrized family  $(\mathcal{P}, p, \lambda)$  and replace  $\pi_n$  by a sequence  $\tilde{\pi}_n$  in  $\mathcal{P}$  which by  $p$  can be pulled back to a sequence  $\theta_n$  in  $\Theta$ . More formally, this means that there is a map,

$$F: \mathbb{R} \times \Theta \rightarrow \Theta, \tag{7}$$

defining the stochastic parameter process

$$\theta_{n+1} = F(y_{n+1}, \theta_n) \tag{8}$$

so that  $\tilde{\pi}_n = p(\cdot, \theta_n)$ . Actually, the filtering process is called finite-dimensional if such a representation can be found that holds *exactly* rather than just approximately. The result of Savitzki states that this is the case if and only if  $P(X_n \in A | Y_n)$  and  $P(X_n \in A | Y_1 \cdots Y_{n-1})$  are of exponential form in  $X_n$ . Whether a given model admits a finite dimensional filter is relatively easy to decide. However, it is not easy to use Savitzki's results to create state space models admitting finite dimensional filters. In Ref. 13 a Laplace transform approach is used. However, starting from a linear observation equation Runggaldier and Spizzichino arrive at a *linear* state space model as well.

For the approximation, a certain fitness criterion between the true and the approximated filtering process is required. In Ref. 11 we employed the Kullback–Leibler distance which proved to be suitable from a computational point of view. We will now describe the general approximation scheme. Details as well as numerical simulations may be found in Ref. 11.

Let  $\mu, \nu \in \mathcal{P}_E$ ,  $\nu \ll \mu$ . Then the Kullback–Leibler distance,

$$\text{KL}(\mu, \nu) := \int_E \log \left( \frac{d\nu}{d\mu} \right) d\nu = \int_E \frac{d\nu}{d\mu} \log \left( \frac{d\nu}{d\mu} \right) d\mu$$

is positive (but maybe  $\infty$ ), vanishes if and only if  $\nu = \mu$  and is a convex function of both  $\mu$  and  $\nu$ . We will now define the approximative filtering process in the course of the following:

*Definition 8:* Suppose a parametrized set of pdf's  $(\mathcal{P}, p, \lambda)$  is given as well as a filtering process  $\pi_n$ . Suppose the following requirements are met:

- (1) We assume the initial measure  $\pi_0 := P(X_0 \in \cdot)$  to be given and fixed throughout this chapter.
- (2)  $\pi_n \ll \lambda$ . The densities will be denoted by  $\pi_n(x)$ . Since  $p(\cdot, \theta) \sim \lambda$ , we also have  $\pi_n \ll p(\cdot, \theta)$  for all  $\theta$ .
- (3) For any  $\nu \ll \lambda$ ,  $\text{KL}(p(\cdot, \theta), \nu)$  is a convex function on the convex parameter space  $\Theta \subset \mathbb{R}^d$ .
- (4) For any  $\nu \ll \lambda$ , there is a  $\theta(\nu) \in \Theta$  with the property

$$\text{KL}(p(\cdot, \theta(\nu)), \nu) \leq \text{KL}(p(\cdot, \theta), \nu) \quad \forall \theta \in \Theta$$

and equality implies  $\theta = \theta(\nu)$ . So  $\theta(\nu)$  is the unique minimizer of  $\text{KL}(p(\cdot, \theta), \nu)$ .

Then we can define the *approximative filtering process*  $\{\theta_n\}_{n \geq 0}$  on  $\Theta$  by

$$\begin{aligned} \theta_0 &:= \theta(\pi_0), \\ \theta_n &:= \theta(S(Y_n, p(\cdot, \theta_{n-1}))). \end{aligned}$$

Obviously  $\theta_n$  is a function of  $\theta_{n-1}$  and  $Y_n$ . We will also call

$$\tilde{\pi}_n := p(\cdot, \theta_n)$$

the approximative filtering process.

## VI. GENERAL ERROR BOUND FOR THE APPROXIMATIVE FILTERING PROCESS

In Sec. V we proposed a scheme to approximate the filtering process on the level of probability distributions. Suppose a stochastic process  $\tilde{\pi}_n$  on  $\mathcal{P}_E$  intended to be an approximation of the correct filtering process  $\pi_n$  is given. The question is whether  $\tilde{\pi}_n$  is a good approximation of  $\pi_n$  or not. A possible way to characterize “good approximations” is to calculate the KL-distance between  $\tilde{\pi}_n$  and  $\pi_n$ . However, a more natural criterion is the accuracy up to which expectations like  $\int f(x) \pi_n(dx)$  are reproduced using  $\tilde{\pi}_n$  instead of  $\pi_n$ . If  $\tilde{\pi}_n \ll \pi_n$  and  $f \in C_B(E)$ , then

$$\begin{aligned} & \left| \int f(x) \pi_n(dx) - \int f(x) \tilde{\pi}_n(dx) \right| \\ & \leq \int |f(x)| \left| \frac{d\tilde{\pi}_n}{d\pi_n} - 1 \right| \pi_n(dx) \\ & \leq \max_x |f(x)| \int \left| \frac{d\tilde{\pi}_n}{d\pi_n} - 1 \right| \pi_n(dx). \end{aligned}$$

The quantity  $\text{TV}(\mu, \nu) := \int |(d\mu/d\nu) - 1| \nu(dx)$  is called the total variation distance. If  $\mu, \nu$  have densities with respect to Lebesgue measure it can be written in the form,

$$\text{TV}(\mu, \nu) := \int |\mu - \nu| dx.$$

It turns out that TV is symmetric, convex in both arguments, vanishes iff  $\mu = \nu$  and satisfies the triangle inequality (in contrast to the KL-distance).

*Lemma 9:* (1) If  $\varphi$  is a Markov kernel, then

$$\text{TV}(\varphi\mu, \varphi\nu) \leq \text{TV}(\mu, \nu).$$

(2) Between KL and TV the following relations hold:

$$\text{TV}(\mu, \nu) \leq 2 \sqrt{1 - \exp(-\text{KL}(\mu, \nu))},$$

$$\text{TV}(\mu, \nu) \leq 2 \sqrt{\text{KL}(\mu, \nu)}.$$

The first inequality is called *Bretagnole–Huber inequality*, the second *Furstemberg inequality*. In both inequalities, the right hand-side is a concave function of KL.

*Proof:* For the first assertion, see Ref. 14. For the Bretagnole–Huber inequality, see Ref. 15. The second inequality is an easy consequence of the first.  $\square$

So far there is not the least bit of evidence that the algorithms defined in the definition will actually work. The process  $\tilde{\pi}_n$  is controlled only by  $\theta_n$  which in turn depends on  $\theta_{n-1}$  and  $Y_n$ , that is, has only restricted input from outside. In the course of the approximation, no reference to  $\pi_n$  is made. This may lead to an unbounded amplification of errors. However, results on the stability property of the nonlin-

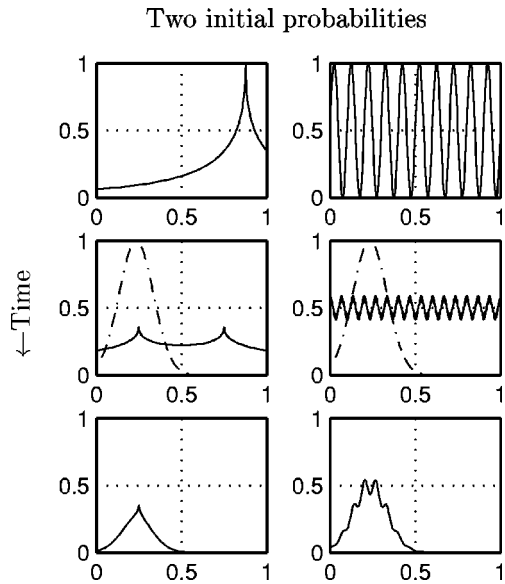


FIG. 2. The time evolution of two probability distributions (more specifically their densities) under the operator  $S(0.25, \cdot)$  [see Eq. (6)] is shown in the left resp. right column. The first row shows the initial densities. The second row shows the probabilities after applying the transition kernel (solid line). The dashed-dotted line is  $g((0.25-x)/\sigma)$ . The third row shows the final product after normalization.

ear filter<sup>16-18</sup> give hope that the filter may be insensitive to errors in the initial condition which leads to a damping of the errors in the course of the approximation. We will first consider again the tentmap example [Eq. (3)] already encountered in Sec. III exemplifying our assertion and then formalize the statements.

Figure 2 shows the time evolution of two probability distributions (more specifically their densities) under the operator  $S(0.25, \cdot)$  [see Eq. (6)] in the left resp. right column. The first row shows the initial densities which have been chosen *ad libitum*. The solid line curves in the second row show the densities after applying the transition kernel. The dashed-dotted line is  $g((0.25-x)/\sigma)$ . The third row shows the final product after normalization. It is apparent that in effect the densities are much more similar than at the beginning. This observation is of general nature. We will however not prove the stability for this model (see end of this section) but for the models given in Example 2.

For our analysis we need yet another metric for measures called the Hilbert metric. Call two measures  $\mu, \nu \in \mathcal{M}_E$  comparable if there are two positive constants  $c_1, c_2$  so that

$$c_1 \leq \frac{\mu(A)}{\nu(A)} \leq c_2 \quad \forall A.$$

This is actually equivalent to  $\mu \sim \nu$  and

$$c_1 \leq \frac{d\mu}{d\nu} \leq c_2.$$

The Hilbert distance is defined as

$$H(\mu, \nu) := \inf(\log(c_2/c_1)),$$

where the infimum is taken over all such  $c_1, c_2$ . We have the following properties of  $H$ :

- (1)  $H$  is symmetric;
- (2)  $H$  vanishes iff  $\mu = c \cdot \nu$  for a positive  $c$ ;
- (3)  $H$  fulfills the triangle inequality;
- (4)  $H(a\mu, b\nu) = H(\mu, \nu)$  for positive constants  $a, b$ ;
- (5) If  $f$  is positive and in  $L_1$ , define the measures  $d\bar{\mu} = f d\mu, d\bar{\nu} = f d\nu$ . Then  $H(\bar{\mu}, \bar{\nu}) = H(\mu, \nu)$ .

This immediately yields for the filtering process

$$H(S(y, \mu), S(y, \nu)) = H(\varphi\mu, \varphi\nu)$$

since the normalization and the multiplication with  $g(\dots)$  cancels out (see Ref. 16). Furthermore, the Hilbert distance has outstanding properties in connection with positive operators. We restrict our attention to Markov transition kernels. If  $\mu, \nu$  are comparable, then so are  $\varphi\mu, \varphi\nu$  and

$$H(\varphi\mu, \varphi\nu) \leq H(\mu, \nu).$$

Furthermore,

$$\sup_{0 < H(\mu, \nu) < \infty} \frac{H(\varphi\mu, \varphi\nu)}{H(\mu, \nu)} \leq \tanh\left(\frac{\Delta}{4}\right),$$

where

$$\Delta := \sup_{\mu, \nu \in \mathcal{P}_E} H(\varphi\mu, \varphi\nu)$$

is the projective diameter of  $\varphi$  (see Refs. 16 and 19). For a Markov kernel satisfying the conditions of Theorem 13 we have  $\Delta \leq 2 \log(c_1/c_2)$ , whence

$$H(\varphi\mu, \varphi\nu) \leq \tau H(\mu, \nu),$$

where  $\tau = \tanh(\log(c_1/c_2)/2) < 1$ .

This analysis shows that such Markov kernels have a negative Lyapunov exponent with respect to the Hilbert metric. According to the properties of the Hilbert metric, this behavior immediately carries over to the filter and will be exploited in our error analysis to follow soon. The technique of using  $H$  in connection with filtering was (to our knowledge) first used in Ref. 16.

A connection to the total variation distance is given in the following lemma:

*Lemma 10: In general,*

$$\text{TV}(\mu, \nu) \leq \frac{2}{\log 3} H(\mu, \nu),$$

where the right-hand side is maybe infinite. Furthermore, if  $\varphi$  satisfies the conditions of Theorem 13, then

$$H(\varphi\mu, \varphi\nu) \leq 2 \log\left(1 + \frac{c_2}{c_1} \text{TV}(\mu, \nu)\right).$$

*Proof:* The first inequality is due to Atar and Zeitouni.<sup>16</sup> The second inequality is due to Kushner and Budhiraja.<sup>18</sup>  $\square$

Now we are ready to embark for the first estimate on the error of our approximative filtering process. Let

- (1)  $\{X_n\}$  be a Markov process satisfying the conditions of Theorem 13;
- (2)  $\pi_n$  be the true filtering process;

- (3)  $\tilde{\pi}_n$  be a process obtained by the approximation scheme 8;
- (4)  $S_k^n(\mu) := S(Y_n, S(Y_{n-1}, S(\dots S(Y_k, \mu) \dots)))$  be the  $k+1$  fold iterate of  $S$  with arguments  $\mu$  and  $Y_k \dots Y_n$ , where, if  $k > n$  we set  $S_k^n(\mu) = \mu$ . We also write  $S_n(\mu) := S(Y_n, \mu)$ .

Then a direct application of the triangle inequality yields

$$\begin{aligned} \text{TV}(\pi_n, \tilde{\pi}_n) &\leq \text{TV}(S_n(\tilde{\pi}_{n-1}), \tilde{\pi}_n) \\ &+ \sum_{k=1}^{n-1} \text{TV}(S_k^n(\tilde{\pi}_{k-1}), S_{k+1}^n(\tilde{\pi}_k)) \\ &+ \text{TV}(S_1^n(\pi_0), S_1^n(\tilde{\pi}_0)). \end{aligned} \tag{9}$$

The first term is by the Bretagnole–Huber inequality bounded by

---


$$\text{TV}(S_n(\tilde{\pi}_{n-1}), \tilde{\pi}_n) \leq 2 \sqrt{1 - \exp(-\text{KL}(S_n(\tilde{\pi}_{n-1}), \tilde{\pi}_n))}.$$

The other terms concern comparable probability measures, so we apply the Atar–Zeitouni inequality and the Kushner–Budhiraja inequality to get

$$\begin{aligned} \text{TV}(S_k^n(\tilde{\pi}_{k-1}), S_{k+1}^n(\tilde{\pi}_k)) &\leq \frac{2}{\log 3} \text{H}(S_k^n(\tilde{\pi}_{k-1}), S_{k+1}^n(\tilde{\pi}_k)) \leq \frac{2}{\log 3} \tau^{n-(k+1)} \text{H}(S_k^{k+1}(\tilde{\pi}_{k-1}), S_{k+1}(\tilde{\pi}_k)) \\ &\leq \frac{2}{\log 3} \tau^{n-(k+1)} \text{H}(\varphi S_k(\tilde{\pi}_{k-1}), \varphi \tilde{\pi}_k) \leq \frac{4}{\log 3} \tau^{n-(k+1)} \log \left( 1 + \frac{c_2}{c_1} \text{TV}(\varphi S_k(\tilde{\pi}_{k-1}), \varphi \tilde{\pi}_k) \right). \end{aligned}$$

Since  $\log(1 + (c_2/c_1)x) \leq (c_2/c_1)x$  we get finally using again the Bretagnole–Huber inequality,

$$\text{TV}(S_k^n(\tilde{\pi}_{k-1}), S_{k+1}^n(\tilde{\pi}_k)) \leq 8 \frac{c_2}{c_1 \log 3} \tau^{n-(k+1)} \sqrt{1 - \exp(-\text{KL}(S_k(\tilde{\pi}_{k-1}), \tilde{\pi}_k))}.$$

In a similar manner, the last term can be treated to give

$$\text{TV}(S_1^n(\pi_0), S_1^n(\tilde{\pi}_0)) \leq 8 \frac{c_2}{c_1 \log 3} \tau^{n-1} \sqrt{1 - \exp(-\text{KL}(\pi_0, \tilde{\pi}_0))}.$$

Since  $2 \leq 8(c_2/c_1 \log 3) =: C$  we get finally

$$\text{TV}(\pi_n, \tilde{\pi}_n) \leq \frac{C}{\tau} \sum_{k=0}^n \tau^{n-k} R_k, \tag{10}$$

where  $R_k$  is the approximation residual given by

$$R_k := \sqrt{1 - \exp(-\text{KL}(S_k(\tilde{\pi}_{k-1}), \tilde{\pi}_k))}$$

if  $k > 0$  and by

$$R_0 := \sqrt{1 - \exp(-\text{KL}(\pi_0, \tilde{\pi}_0))}$$

if  $k = 0$ .

Recall that  $\text{KL}(S_k(\tilde{\pi}_{k-1}), \tilde{\pi}_k)$  (resp.  $\text{KL}(\pi_0, \tilde{\pi}_0)$ ) is exactly the quantity we minimize in the approximation scheme. Indeed,  $\tilde{\pi}_k$  is chosen to minimize  $\text{KL}(S_k(\tilde{\pi}_{k-1}), q)$ , where  $q$  varies over the parametric family of distributions. Since  $\sqrt{1 - \exp(-x)}$  is increasing, the algorithm “optimizes” the bound (10).

Furthermore, remark that the calculations circumvent Hilbert distances of the form  $\text{H}(S_k(\tilde{\pi}_{k-1}), \tilde{\pi}_k)$ , whence we do not need to assume that  $S_k(\tilde{\pi}_{k-1})$  and  $\tilde{\pi}_k$  are comparable. This would be a quite inconvenient restriction of the parametrized families.

---

The whole analysis was done under the assumption that  $\{X_n\}$  satisfies the conditions of Theorem 13. If this is not the case, the Kushner–Budhiraja inequality is not valid and furthermore the Lyapunov exponent with respect to the Hilbert metric may be 1, so the filter is not so easily proved to be stable. In this case a more involved analysis of the TV distance is required. Basically the same considerations apply if we know that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \sup \frac{\text{TV}(S_1^n(\mu), S_1^n(\nu))}{\text{TV}(\mu, \nu)} < 1, \tag{11}$$

where the supremum should *not* be taken over all  $\mu, \nu \in \mathcal{P}_E$  but only over the restricted set

$$\{S(y_n, S(y_{n-1}, \dots, S(y_1, \mu) \dots))\}; \mu \in \mathcal{P}, y_i \in \mathbb{R}\},$$

where  $\mathcal{P}$  is the parametrized set of distributions chosen for the approximations. This is basically the set of distributions that may appear in the course of the approximation. The

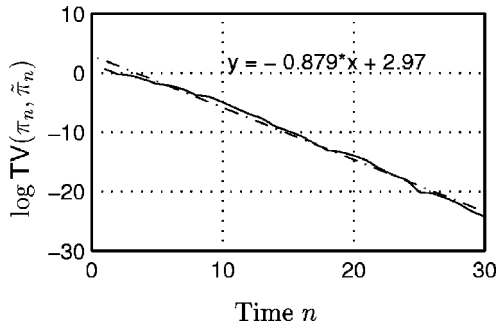


FIG. 3. A logarithmic plot of the TV of two filtering processes for the tentmap example initialized differently. The stability of the filter is apparent in this case.

existence of the limit in (11) may be shown using Kingman’s subadditive ergodic theorem.<sup>20</sup> We have not carried out the analysis, but we would like to remark that the nonlinear filter for CSK-schemes does *not* satisfy the conditions of Theorem 13 and in fact is *not* insensitive to its initial condition *in general*. Suppose that in the CSK setup,  $f_0$  and  $f_1$  have two distinct periodic orbits in common, that is  $\bar{x} = \{x_1 \cdots x_p\}$  and  $\bar{z} = \{z_1 \cdots z_q\}$  are periodic orbits for both  $f_0$  and  $f_1$ . Then the filtering process initialized with a measure supported on  $\bar{x}$  will always have support on  $\bar{x}$ . The filtering process correctly reproduces the fact that  $X_n$  cannot escape from  $\bar{x}$ . The same holds for the periodic orbit  $\bar{z}$ . So the two filtering processes initialized with a measure supported on  $\bar{x}$  and on  $\bar{z}$ , respectively, will never become similar in any sense if  $n$  goes to infinity.

However, looking back to Fig. 2 we see apparently a stability property. The main point here seems to be that the densities we start within our numerical example are smooth. Then it is pretty obvious how the stability emerges: The Frobenius–Perron operator stretches the function and thus reduces all slopes (in this example, by a factor of 2). Then multiplying with  $g(\cdots)$  (dashed–dotted line) effectively cuts out a small part of the function leading to a further regularization. So for CSK-schemes, the filter seems to be stable for a restricted class of initial conditions. We conjecture that this is the case for initial distributions having a density of bounded variation with respect to the Lebesgue measure. This would suffice for most applications.

In Fig. 3 we show a logarithmic plot of the TV of two filtering processes for the tentmap example [Eq. (3) in Sec. III] initialized with two different densities. The Lyapunov exponent of the filter seems to be  $\cong -0.88$  and the stability of the filter is apparent in this case.

**VII. A BOUND ON THE BIT ERROR RATE**

We have seen how to build the optimal causal receiver using the nonlinear filtering process. Since the nonlinear filtering process cannot be calculated in general, we suggested several approximation schemes. In Sec. VI we gave a bound on the error between the true and the approximative filtering process. In this section we show the implications of this result on the bit error rate obtained by receivers based on approximative filtering processes rather than the true one. Consider the function

$$f_n(y) := \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) \pi_{n-1}^v(dx) \right|.$$

From Sec. IV we know that

$$\text{BEP}_n^v = \frac{1}{2} E_v \left[ 1 - \int f_n(y) dy \right].$$

We now have using the triangle inequality,

$$\begin{aligned} & \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) \pi_{n-1}^v(dx) \right| \\ &= \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) \right. \\ & \quad \times (\varphi_1 - \varphi_0) (\pi_{n-1}^v - \tilde{\pi}_{n-1}^v + \tilde{\pi}_{n-1}^v)(dx) \left. \right| \\ &\leq \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) \tilde{\pi}_{n-1}^v(dx) \right| \\ & \quad + \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) (\pi_{n-1}^v - \tilde{\pi}_{n-1}^v)(dx) \right|. \end{aligned} \tag{12}$$

The second term can be bounded using the triangle inequality,

$$\begin{aligned} & \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) (\pi_{n-1}^v - \tilde{\pi}_{n-1}^v)(dx) \right| \\ &\leq \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) |\varphi(\pi_{n-1}^v - \tilde{\pi}_{n-1}^v)|(dx). \end{aligned}$$

The integrand is an integrable function of  $x$  and  $y$  so we can replace the second term in (12), integrate over  $y$ , and reverse the order of integration in the second term to get

$$\begin{aligned} & \int \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) \pi_{n-1}^v(dx) \right| dy \\ &\leq \int \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) \tilde{\pi}_{n-1}^v(dx) \right| dy \\ & \quad + \text{TV}(\varphi \pi_{n-1}^v, \varphi \tilde{\pi}_{n-1}^v) \\ &\leq \int \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) \tilde{\pi}_{n-1}^v(dx) \right| dy \\ & \quad + \text{TV}(\pi_{n-1}^v, \tilde{\pi}_{n-1}^v), \end{aligned}$$

since  $\text{TV}(\varphi \cdot, \varphi \cdot) \leq \text{TV}(\cdot, \cdot)$  (Lemma 9). In exactly the same manner (exchanging the role of  $\pi$  and  $\tilde{\pi}$ ) one obtains

$$\begin{aligned} & \int \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) \pi_{n-1}^v(dx) \right| dy \\ &\geq \int \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) (\varphi_1 - \varphi_0) \tilde{\pi}_{n-1}^v(dx) \right| dy \\ & \quad - \text{TV}(\pi_{n-1}^v, \tilde{\pi}_{n-1}^v). \end{aligned}$$

If we define the quantity,

$$\begin{aligned} \text{B}\tilde{\text{E}}\text{P}_k^\nu := & \frac{1}{2} E_\nu \left[ 1 - \int \left| \int \frac{1}{\sigma} g\left(\frac{y-h(x)}{\sigma}\right) \right. \right. \\ & \left. \left. \times (\varphi_1 - \varphi_0) \tilde{\pi}_{n-1}^\nu(dx) \right| dy \right], \end{aligned}$$

which is the same as  $\text{BEP}_k^\nu$  but with  $\pi$  replaced by  $\tilde{\pi}$  we can write our estimate as

$$|\text{BEP}_k^\nu - \text{B}\tilde{\text{E}}\text{P}_k^\nu| \leq \frac{1}{2} E_\nu \text{TV}(\pi_{n-1}^\nu, \tilde{\pi}_{n-1}^\nu).$$

We assume now the validity of the estimate (10). Then we get

$$|\text{BEP}_k^\nu - \text{B}\tilde{\text{E}}\text{P}_k^\nu| \leq \frac{C}{2} \sum_{k=0}^n \tau^{n-k} E_\nu R_k.$$

So far this estimate is of restricted practical use since both the approximated bit error probability  $\text{B}\tilde{\text{E}}\text{P}$  as well as the right-hand side of the above estimate involve the *true* expectation  $E_\nu$ .

Under the additional assumption that the compound process  $\{Y_n, \theta_n\}$  is ergodic, then we can compute  $E_\nu R_n$  and  $\text{B}\tilde{\text{E}}\text{P}_n$  in an offline experiment, since  $R_n$  is a function of  $\theta_{n-1}$  and  $Y_n$ , and furthermore  $\text{B}\tilde{\text{E}}\text{P}_n$  is the expectation over a function depending on  $\theta_{n-1}$  only. Then both  $\text{B}\tilde{\text{E}}\text{P}_n$  and  $E_\nu R_n$  are asymptotically equal to a constant depending on the system and the approximation algorithm and can be computed numerically by an empirical mean over a long realization.

## VIII. CONCLUSION

In this paper we investigated a certain class of communication schemes. Basically we consider a transmitter whose internal state is a Markov process depending on a message which is in turn a binary sequence of independent and identically distributed random variables. At the receiver a time series  $\{Y_n\}$  is recorded, where  $Y_n$  is a function of the transmitter state plus additive noise. We formulate the receiver problem as the question: What is the value of the actual message bit  $M_n$  given the time series  $Y_1, \dots, Y_n$ . We show that the optimal receiver  $\hat{M}_n$  (giving the least probability of errors) can be obtained using results from nonlinear filtering. Furthermore, straightforward application of known results on ergodic properties of the nonlinear filter leads to results on asymptotic properties of  $\hat{M}_n$ .

A quite well known problem of the optimal nonlinear filter is the unbounded growth of its complexity. This problem appears to be present also in our communication setup. This underlines the necessity of approximations of the nonlinear filter and in turn of the optimal receiver  $\hat{M}_n$ . Assuming a quite general approximation scheme we derive error bounds on the approximative nonlinear filter as well as on the optimal receiver. It turns out that the validity of these calculations essentially depends on a stability property of the nonlinear filter or, roughly speaking, on its largest Lyapunov exponent.

The following questions left open in the paper merit further investigation. First, a lot of interesting and frequently

investigated communication schemes (either novel or classical) do not fall into the classes investigated here. These are (among others)

- (1) Shift keying schemes where two independent dynamical systems run in parallel. The output of the first or the second system is transmitted (for a certain amount of time) depending on whether the bit to be transmitted was zero or one, respectively. Sometimes two chaotic systems are used (see, for example, Ref. 21), whence this setup, although different from Example 1, is called chaotic shift keying as well.
- (2) Setups where the transmitted signal is again the output of a dynamical system, but now the state space is divided into two regions representing the bit zero or one, respectively. By nonlinear control methods this system is steered into one of the regions after the other depending on the bits to be sent. Of course, at the receiver side the problem is to locate the position of the system in state space to recover the bits. Of particular interest are systems where the control input vanishes for a possibly restricted set of messages. This indeed can be the case for chaotic systems, where the remaining set of messages is still sufficiently large for communication.<sup>22</sup> At the receiver side the system then can be considered as autonomous.
- (3) System with more users involved.
- (4) Communication systems based on *synchronization*. In Ref. 23 a setup is considered where the message (actually a continuous valued message is permitted) appears in the state space equation *and* in the transmitted signal. In the case of no measurement noise a synchronization based receiver reveals the message with asymptotically vanishing error. Synchronization in the presence of chaos, however is known to be quite sensitive to noise in the transmission line, so the receiver fails for even small amounts of measurement noise. In other words, the receiver for the noise free case appears to be a too simple an approximation of the optimal receiver and has to be replaced by more sophisticated devices.

A further field of investigation may concern CSK-schemes. It turned out that CSK-schemes often do not fulfill the assumptions required to obtain the results of the paper. However, this does not mean that these results cannot be extended to CSK-schemes, maybe in weaker form. A first step obviously is to give conditions on the ergodicity of CSK-schemes, which apparently in many investigations published so far was tacitly assumed. We already mentioned that a modified aperiodicity assumption seem to make the analysis of Ref. 9 applicable also in this case. Furthermore, proving the negativity of Lyapunov exponents for the nonlinear filtering process associated with CSK-schemes is quite important. We already mentioned that a possible route goes via densities of bounded variation.

## ACKNOWLEDGMENTS

We thank the members of the nonlinear dynamics group of the "Drittes Physikalisches Institut" for stimulating dis-

cussions and support. J.B. acknowledges support from the DFG Graduiertenkolleg ‘‘Str6mungsinstabilitäten und Turbulenz.’’

**APPENDIX: ERGODIC THEORY OF MARKOV PROCESSES**

We recall some results about ergodic properties of Markov processes. We will keep the same notation as in the paper, namely, let

- (1)  $E$  a polish (complete separable metric) space;
- (2)  $\mathcal{B}_E$  the Borel field;
- (3)  $\mathcal{P}_E$  the space of probability measures on  $E$ ;
- (4)  $C_b(E)$  the spaces of continuous bounded functions on  $E$ ;
- (5)  $\mathcal{B}(\mathcal{P}_E)$  the Borel field of  $\mathcal{P}_E$  endowed with the weak topology.

We write as usual

$$\int f(x)P(dx)$$

for the Lebesgue integral of  $f$  over  $P$ . A measure  $Q$  is *absolutely continuous* with respect to  $P$  (write  $Q \ll P$ ) if  $P(A) = 0$  always implies  $Q(A) = 0$ . In this case there is a function denoted by  $dQ/dP$  (Radon–Nykodym derivative, see Ref. 5) with the property,

$$Q(A) = \int_A \frac{dQ}{dP}(x)P(dx).$$

If both  $Q \ll P$  and  $P \ll Q$ , they are called *equivalent* and we write  $P \sim Q$ .

*Definition 11:* A random process  $\{X_n\}$  is *stationary* if, for any  $k$  and sets  $A_j \in \mathcal{B}_E$  the probability  $P(X_{n+1} \in A_1, \dots, X_{n+k} \in A_k)$  does not depend on  $n$ , i.e., is invariant with respect to time shifts.

*Lemma 12:* A Markov process is stationary iff the probability measure  $\nu(A) := P(X_0 \in A)$  has the property

$$\nu(A) = \int \varphi(A, x)\nu(dx).$$

Such a measure is called *invariant*.

*Proof:* See Ref. 5. □

The question arises whether for a given transition kernel  $\varphi(A, x)$  there is an invariant measure  $\nu$  so that the canonical process on  $(E^\infty, P^\nu, \mathcal{B}_E^\infty)$  is stationary. A fruitful idea is to consider iterates of the kernel: Define  $\varphi^{(1)}(A, x) := \varphi(A, x)$  and iteratively

$$\varphi^{(n)}(A, x) := \int \varphi(A, z) \cdot \varphi^{(n-1)}(dz, x).$$

The following theorem gives conditions under which the sequence  $\varphi^{(n)}(A, x)$  generated by a Markov transition kernel converges to an invariant measure.

**Theorem 13:** Suppose there is a finite measure  $\mu$  and two positive constants  $c_1, c_2$  with the property,

$$c_1\mu \ll \varphi(\cdot, x) \ll c_2\mu \quad \text{for all } x \in E, \tag{A1}$$

then there is an invariant probability measure  $s$  absolutely continuous with respect to  $\mu$ . Furthermore, there are constants  $K \geq 0$  and  $0 < \delta < 1$  independent of  $x$  with

$$\sup_{A \in \mathcal{B}_E} |\varphi^{(n)}(A, x) - s(A)| \leq K\delta^n.$$

*Proof:* The theorem is a slight modification of results presented in Ref. 24, Chap. V (Sec. 5). □

A trivial verification shows that if  $\varphi^{(n)}(\cdot, x)$  satisfies the conditions of Theorem 13, then we also have the property,

$$\lim_{n \rightarrow \infty} \int |f\varphi^{(n)}(x) - s(f)|s(dx) = 0 \quad \forall f \in C_b(E). \tag{A2}$$

Condition (A2) (which is weaker than the result of Theorem 13) will prove to be essential for ergodic properties of the filtering process.

Starting with  $s$  as the initial distribution, the resulting probability on the probability space  $(E^\infty, \mathcal{B}_E^\infty)$  is denoted by  $P^s$ , as for every probability measure  $\nu$  on  $E$  the resulting probability on  $(E^\infty, \mathcal{B}_E^\infty)$  is denoted by  $P^\nu$ .

Stationary processes may or may not be *ergodic*. We recall the basic concepts of ergodic theory. Let  $\{X_n\}_{n \in \mathbb{N}}$  be a stationary process. An event  $A$  is *invariant* if there is a fixed  $B \in \mathcal{B}_\infty$  so that for any  $k$ ,  $A$  can be represented as

$$A := \{\omega \in \Omega; (X_k, X_{k+1}, \dots) \in B\}.$$

The invariant events form a  $\sigma$ -algebra denoted by  $\mathcal{I}$ . This is the basis for the following famous result:

**Theorem 14 (Birkhoff’s ergodic theorem):** Let  $X_n$  be a stationary process,  $E|X_1| < \infty$ . Then the following limit holds a.s. and in  $L_1$ :

$$\frac{1}{n} \sum_{k=1}^n X_k \rightarrow E(X_1 | \mathcal{I}).$$

*Proof:* See Ref. 5. □

If  $X_n$  are iid random variables, all invariant events have probability zero or one (Kolmogorov’s zero-one law). Obviously, then  $E(X_1 | \mathcal{I}) = E(X_1)$ , and Birkhoff’s ergodic theorem translates into the strong law of large numbers. To generalize this, call a process *ergodic*, if all invariant events have probability zero or one. Obviously, a process is ergodic iff all random variables measurable with respect to  $\mathcal{I}$  are a.s. constant. Hence, if  $E|X_1| < \infty$  and the process is ergodic,  $E(X_1 | \mathcal{I}) = E(X_1)$  and Birkhoff’s theorem gives

$$\frac{1}{n} \sum_{k=1}^n X_k \rightarrow E(X_1)$$

both a.s. and in  $L_1$ .

Obviously, conditions for ergodicity are quite essential:

*Lemma 15:* (1) Let  $f: E^\infty \rightarrow \mathbb{R}$  be measurable. Then the process

$$Y_n := f(X_n, X_{n+1}, \dots)$$

is stationary (ergodic) if  $X_n$  is stationary (ergodic).

(2) A stationary process  $X_n$  is ergodic iff all random variables measurable with respect to  $\mathcal{I}$  are a.s. constant.

(3) If a process  $X_n$  admits a unique stationary measure it must be ergodic.

Back to Markov processes we have the following more special criteria:

*Lemma 16:* (1) If  $E$  is compact, there are always invariant measures for  $\varphi$ :

(2) If an invariant measure  $\nu$  for  $\varphi$  is unique, then  $P^\nu$  must be ergodic.

(3) Let  $\nu$  be an invariant measure for  $\varphi$ . Then if any  $f \in L_1(E, \nu)$  with the property

$$f \circ \varphi^{(n)}(x) = f(x)$$

is  $\nu$ -almost sure constant, then  $P^\nu$  must be ergodic.

<sup>1</sup>C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Chicago, 1949).

<sup>2</sup>N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series* (MIT Press, New York, 1950).

<sup>3</sup>R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.* **82**, 35–45 (1960).

<sup>4</sup>A. Jazwinsky, *Stochastic Processes and Filtering Theory*, Vol. 64, in *Mathematics in Science and Engineering* (Academic, New York, 1970).

<sup>5</sup>L. Breiman, *Probability* (Addison-Wesley, Reading, 1973).

<sup>6</sup>G. Kallianpur, *Stochastic Filtering Theory*, No. 13, in *Applications of Mathematics* (Springer Verlag, Berlin, 1980).

<sup>7</sup>L. Stettner, "On invariant measures of filtering processes," in *Stochastic Differential Systems, Proceedings of the 4th Bad Honnef Conference, 1989*, edited by K. Helmes, N. Christopeit, and M. Kohlmann, *Lectures in Control and Information Sciences*, pp. 279–292.

<sup>8</sup>H. Kunita, "Asymptotic behavior of the nonlinear filtering errors of Markov processes," *J. Multivariate Anal.* **1**, 365–393 (1971).

<sup>9</sup>M. Pollicott and M. Yuri, *Dynamical Systems and Ergodic Theory*, Vol. 40

in *Mathematical Society Student Texts* (Cambridge University Press, Cambridge, 1998).

<sup>10</sup>M. P. Kennedy, R. Rovatti, and G. Setti, *Chaotic Electronics in Telecommunications* (CRC Press, Boca Raton, 2000).

<sup>11</sup>J. Bröcker, U. Parlitz, and M. Ocorzalek, "Nonlinear noise reduction," in *Applications of Nonlinear Dynamics to Electronic and Information Engineering*, edited by G. Setti, *Proc. IEEE* **90** (5) (2002).

<sup>12</sup>G. Sawitzki, "Finite-dimensional filters in discrete time," *Stochastics* **5**, 107–114 (1981).

<sup>13</sup>W. J. Runggaldier and F. Spizzichino, "Sufficient conditions for finite dimensionality of filters in discrete time: A Laplace transform-based approach," *Bernoulli* **7**, 211–221 (2001).

<sup>14</sup>F. LeGland, "Stability and approximation of nonlinear filters: An information theoretic approach," Technical Report No. IRISA/INRIA, 1999.

<sup>15</sup>V. N. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

<sup>16</sup>R. Atar and O. Zeitouni, "Exponential stability for nonlinear filters," *Ann. Inst. H. Poincaré Prob. Statist.* **36**, 691–725 (1997).

<sup>17</sup>A. Budhiraja and D. Ocone, "Exponential stability of discrete time filters for bounded observation noise," *Syst. Control Lett.* **30**, 185–193 (1997).

<sup>18</sup>A. Budhiraja and H. Kushner, "Robustness of nonlinear filters over the infinite time interval," *SIAM J. Control Optim.* **36**, 1618–1637 (1998).

<sup>19</sup>G. Birkhoff, *Lattice Theory*, 3rd ed., Vol. 25, in *AMS Colloquium Publications* (American Mathematical Society, Providence, RI, 1967).

<sup>20</sup>U. Krengel, *Ergodic Theorems* (de Gruyter, Paris, 1985).

<sup>21</sup>U. Parlitz, L. Kocarev, L. Chua, K. Halle, and A. Shang, "Transmission of digital signals by chaotic synchronization," *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **4**, 973–977 (1992).

<sup>22</sup>S. Hayes, C. Grebogi, E. Ott, and A. Mark, "Experimental control of chaos for communication," *Phys. Rev. Lett.* **73**, 1781–1784 (1994).

<sup>23</sup>L. Kocarev and U. Parlitz, "General approach for chaotic synchronization with applications to communication," *Phys. Rev. Lett.* **74**, 5028–5031 (1995).

<sup>24</sup>J. Doob, *Stochastic Processes* (Wiley, New York, 1953).