Nonlinear Noise Reduction

Jochen Bröcker*, Ulrich Parlitz and Maciej Ogorzalek

December 12, 2001

Abstract

Different methods for removing noise contaminating time series are presented, which all exploit the underlying (deterministic) dynamics. All approaches are embedded in a probabilistic framework for stochastic systems and signals, where the two main tasks, state and orbit estimation are distinguished. Estimation of the true current state (without noise) is based on previously sampled elements of the time series, only, and corresponds to filtering. With orbit estimation the entire measured time series is used to determine a less noisy orbit. In this case not only past values but also future samples are used, which or course improves performance.

Keywords

time series analysis, noise, stochastic systems, nonlinear filters

1 Introduction

Noise is an omnipresent phenomenon. Any physical object or process is subject to changes in its environment, external or internal fluctuations, influences of different nature. Losely speaking in a technical context we consider as noise any disturbance tending to interfere with the normal operation of a device or system. Looking at text-books in different domains one can see slightly different ways of understanding of noise.

The easiest to understand for everyone is the notion of noise in audio. Noise is any unpleasant sound and, more technically, any unwanted sound that is unintentionally added to a desired sound. Ambient sound itself is a series of changes in air pressure transmitted in waves from the sound source to anyone with the sensory apparatus to detect the waves (human beings and other animals with ears, for example). Sound waves are expressed as a series of analog sine waves. The combination and blend of these waves gives sounds their individual characteristics, making them pleasant or unpleasant to listen to. In recording sound, noise is often present on analog tape or low-fidelity digital recordings.

*e-mail: broecker@dpi.physik.uni-goettingen.de.

The standard audio cassette includes a layer of hiss on every recording. When doing digital recording, the conversion of a sound file from 16-bit to 8-bit adds a layer of noise.

Slightly different is the understanding of noise in communications. Any interference or disturbance which affects the signals on a line and may affect the information carried is considered as noise. It can come from a variety of sources, including radio waves, nearby electrical wires, lightning, and bad connections.

In a hard-wired circuit such as any telephone-line-based installation, external noise is picked up from appliances in the vicinity, from electrical transformers, from the atmosphere, and even from outer space. The transmission channel properties may be extremely bad in industrial environments, in areas densly cover with buildings which might reflect waves causing multipath effects etc. In a modem installation noise slows down the data transfer rate, because the system must adjust its speed to match conditions on the line. In a voice telephone conversation, noise rarely sounds like anything other than a faint hissing or rushing. Noise is a more significant problem in wireless systems than in hard-wired systems. At low frequencies atmospheric and electrical noise are much more severe than at a high frequency. Noise generated inside wireless receivers, known as internal noise, is less dependent on frequency. Engineers are more concerned about internal noise at high frequencies than at low frequencies, because the less external noise there is, the more significant the internal noise becomes.

Noise is also present in any measurement system made using any type of sensing device. Unwanted disturbances coming from the environments affect the measuring device itself, change characteristics of the output signals generated by these devices and further affect the signals that are sent and stored. Front-ends of equipement processing data in digital way communicating with the external environment (already mentioned A/D and D/A conversion) and all computer-based operations on signals are sources of significant disturbances which commonly are modeled as noise. A particular case of crucial measurements are those which are used as input of a control system used for example for satellite orbit estimation ([19, 42]), orbit estimation in reentry problems ([42] and references therein), aircraft radar guidance ([19, 42]), positioning of autonomous vehicles, or navigation of autonomous robots ([74] and references therein). In these cases noisy sensor signals may result in wrong state estimations and poor control actions. In contrast to most of the examples mentioned previously in these cases more or less wellknown dynamical systems are involved and this is exactly the situation in which nonlinear noise reduction methods of the kind we are going to present in this review may be applied.

Before we shall discuss methods for noise reduction we would like to mention that there exists a wide range of applications in which noise is beneficial and very useful [35]. Let us just mention here the applications of noise as a broadband signal useful in microwave heating, source detection and location, jamming signals in electronic countermeasures, and even as a carrier signal in communications, test signal in several different types of measurements, (e.g. antenna characteristics, insertion loss, impulse response, linearity and intermodulation etc.) or probe signal into microscopic phenomena (carrier lifetime, uniformity testing in semiconductors, determination of physical parameters etc.) and finally as a conceptual tool (model).

When observing signals noise appears as random variations of one or more characteristics of any entity such as voltage, current, or data, or random signal of (un)known statistical properties of amplitude, distribution, and spectral density.

In most applications engineers are constantly striving to develop better ways to deal with noise. As nonlinear approaches for a long time were very difficult to apply, commonly used were linear filtering techniques. Thus spectral separation became a generally used method to cope with noise. The traditional method has been to minimize the signal bandwidth to the greatest possible extent. The less spectrum space a signal occupies, the less noise is passed through the receiving circuitry. However, reducing the bandwidth limits the maximum speed of the data that can be delivered. Another, more recently developed schemes for minimizing the effects of noise and separating the useful signal from noise are based on advanced (digital) filtering techniques and signal processing.

In this paper we will concentrate on de-contamination of useful signals i.e. separation of noise (stochastic signal) from information-carrying signals (of deterministic type). Specific goals of noise removal vary depending on a particular application. One can concentrate on optimization of the signal energy, improvement of quality of digital communication (reduction of Bit Error Rate), improvement of quality of audio transmission, or recovery of fine details of the measured signal (otherwise hidden in noise as eg. Electro-Cardio-Grammes in bio-medical measurements).

To separate a signal from noise we have to use specific characteristics of both types of signals. One of such characteristics used in classical methods are spectral properties of the signals and their amplitudes. When the information-carrying signal is narrow-band and noise appears to be a wide-band signal (or inversely) it is usually relatively easy to attenuate all unwanted frequencies. Hard problems arise when both the useful signal and noise have similar spectral characteristics and/or comparable amplitudes. Standard filtering techniques are of little use in such cases and more advanced techniques which will be described below could be used that try to separate the signals in some appropriate state space. It should be stressed that here we will mostly use the notion of filtering as introduced in control textbooks (comp. Wiener filters, Kalman filters) i.e. equivalent to state estimation and not as frequency-domain signal shaping!

When dealing with signals generated by some physical systems (measured at some output using some kind of probe or sensor) we encounter basically two extreme classes of problems (and, of course, all cases between them).

- Class 1 in which we know some model of the underlying time evolution of the system.
- Class 2 in which nothing is known about systems dynamics.

One has to take into account that in both cases the measured signal might be corrupted by noise of different proveniences and properties. For the purpose

Goal of calculation	Data used for calculation: past values $Y_1, \ldots, Y_n$ only	Data used for calculation: entire time series $Y_1, \ldots, Y_N$
Current state X_n	1. Find the most likely current state $\hat{X}_n$ exploiting $P(X_n \mid Y_1, \ldots, Y_n)$	2. Find the most likely current state $\hat{X}_n$ exploiting $P(X_n \mid Y_1, \ldots, Y_n)$ (sub-problem of 4.)
Full trajectory $X_1, \ldots X_N$	3. Calculate whole time series using only knowledge of past data (prediction!)	4. Find the most likely orbit exploiting $p(X_1, \ldots X_N \mid Y_1, \ldots Y_N)$

Table 1: Depending on the goals and available date four approaches can be distinguished.

of this paper we will assume that some model of the the underlying dynamical (Markov) process of the system is available. This restriction is in our opinion reasonable because on the one hand one has to understand this case first before investigation the more difficult class 2 and on the other hand Class 1 covers already many interesting areas such as communication or control systems. The case of unknown dynamics will briefly be discussed in Sect. 4.3 (including a list of references for further reading).

We will also assume without loss of generality that all the signals are sampled in time - so what is observed is a discrete time-series (such an assumption is a valid one as in most cases the measured variables are converted by the data acquisition equipment).

Further the system itself or the output (measured) signals can be corrupted by noise, i.e. we have to cope with *dynamical noise* and/or *measurement noise* as will be introduced in more detail in the next section.

Finally we would like to stress that we shall illustrate the different approaches using some well defined (stochastic) nonlinear dynamical models generating chaotic dynamics, because these are typical examples where the frequency bands of the (chaotic) signal and of the noise overlap and conventional methods for noise reduction/separation fail. Furthermore, such chaotic sources are still considered as potential components in some digital communication systems.

Thinking about the possibilities of noise reduction/separation we can distinguish several cases depending on the goal chosen and the information which is used in calculations. The approaches currently under study are summarised in Table 1.

1.1 Motivating example

To give the readers some motivation let us consider a simple example in which we will try to reduce the influence of noise corrupting a wide-band signal. The features of different noise reduction methods will be illustrated using time series

$\{X_n\}$ generated by the chaotic Hénon map

$$X_{n+1}^{(1)} = 1 - a\left[X_n^{(2)}\right]^2 + bX_n^{(2)} \quad (1)$$

$$X_{n+1}^{(2)} = X_n^{(1)} \quad (2)$$

with parameters $a = 1.4$ and $b = 0.3$. The chaotic attractor reconstructed from the clean time series $\{X_n^{(1)}\}$ is shown in Fig. 1a. Figure 1b shows a reconstruction from a time series

$$Y_n = X_n^{(1)} + R_n$$

where gaussian noise $\{R_n\}$ with SNR 13 dB has been added to the purely deterministic data. Since the chaotic dynamics generates already a broadband signal the added noise occupies the same frequency band (inband noise) and can thus not be removed using linear (spectral) methods. Exploiting the underlying determinism, however, noise reduction is still possible as illustrated in Fig. 1c where one of the noise reduction methods to be presented in the following has been applied. [1]

The achievable improvement of the SNR of a noisy time series depends on the noise reduction method and on the SNR of the noise contained in the data. In this paper, the SNR and the SNR improvement will be defined as follows: If X_n, $n = 1 \ldots N$ is a real valued, time discrete signal, its *empirical mean* is defined as

$$\langle X \rangle := \frac{1}{N} \sum_n X_n.$$

The *power* of X is defined to be $P_X := \langle X^2 \rangle - \langle X \rangle^2$. Now if X_n is any signal, Y_n another signal supposed to be X_n corrupted by errors (noise) we define the *signal to noise ratio* in Y_n as

$$\mathrm{SNR}_Y := 10 \log_{10}(P_X / P_{Y-X}).$$

So SNR is the fraction of the power of the signal and the power of the errors on logarithmic scale. If by a noise reduction algorithm we generate from Y_n another signal Y'_n supposed to be a better estimate of X_n, we define the SNR improvement of the respective algorithm as

$$\text{SNR–improvement}_Y := \mathrm{SNR}_{Y'} - \mathrm{SNR}_Y.$$

We have computed this improvement for a noisy time series from the Hénon system using different algorithms that will be presented in detail in the following sections. Figure 2 shows the improvement of the SNR vs. the SNR of the given data set consisting of 1024 samples. As a benchmark the dotted curve shows the performance of a linear optimal Wiener filter that turns out to be competitive for

[1] Readers already aquainted with noise reduction may think that the resulting attractor still looks very noisy compared to other common noise reduction methods. The displayed result however was obtained using a causal method, in contrast to most other methods, which are acausal.

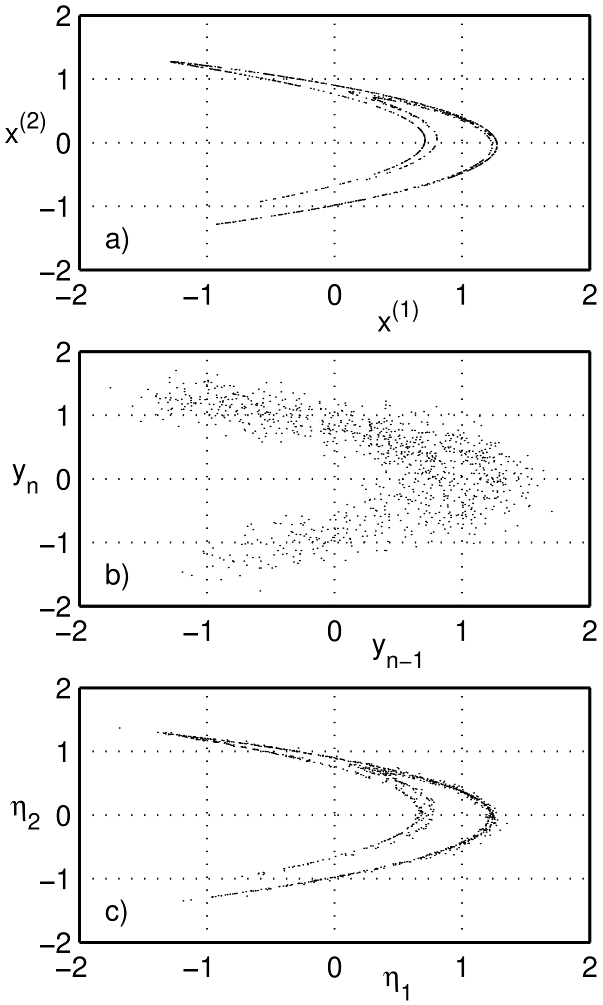

Figure 1: Chaotic attractor of the Hénon map (1) for $a = 1.4$ and $b = 0.3$. (a) Clean data $\{X_n^{(1)}\}$, (b) noisy data (13 dB SNR) $\{Y_n\}$, (c) result of noise reduction using the exponential families (see Section 3.1.2).

negative SNRs, only. Note that for vanishing relative signal power (SNR $\to -\infty$) the dotted curve approaches the straight line given by: SNR improvement = –SNR. This line gives the improvement when using as filtered signal a constant time series given by the empirical mean of the data. Clearly any algorithm should outperform this simple approach.

The solid curve in Fig. 2 shows the SNR improvement of another standard method, the extended Kalman filter (EKF) which gives satisfying results for low noise amplitudes (SNR $>$ 30dB). The dashed-doted curve is obtained with a noise reduction scheme where some underlying probability density functions are approximated by functions from an exponential family. This approach was used in Fig. 1b and yields good SNR improvement for SNR $>$ 10dB, but fails for large noise amplitudes (SNR < 0). The dashed line in Fig. 2 denotes results obtained with Monte Carlo sampling. For large noise amplitudes the performance of this method compares to that of the Wiener filter and for medium noise (0-30 dB) it turned out to be better than the other methods mentioned so far. The sudden decrease of the SNR improvement at about 40dB is due to the finite ensemble used in the Monte Carlo simulation. With a larger ensemble also for higher values of the SNR good improvement is achieved, for smaller ensembles the curve bends already at smaller SNR values. We emphasize that all the methods mentioned so far provide (less noisy) estimates of the state based on information from the past, only. These estimates can be improved considerably when more and more future values of the time series are taken into account. If the entire available data set is used the task changes from state estimation to orbit estimation (comp.Table 1). For comparison the SNR-improvement of an orbit estimation algorithm (called LSS) is shown as a solid curve with plus symbols in Fig. 2. Using the full information from past and future this method outperforms all state estimation schemes. The reason for the saturation at a high level of SNR improvement for large SNR values will be discussed in more detail in Sect. 4.1.

The structure of the paper is as follows. We shall start with a brief review of relevant notions from probability theory in Sec. 2. Then state estimation and orbit estimation are addressed in Secs. 3 and 4, respectively. In Sec. 5 the task of symbol extraction from a noisy time series is discussed.

For the purpose of this paper we will consider as noise any unwanted disturbance of stochastic nature which degrades the quality of signals and data. Noise occurs in digital and analog systems, and can affect files and communications of all types, including text, programs, images, audio, and measurements. Throughout this paper we will not deal with distortion which often corrupts the signal quality (although it should be stressed that in some situations distortion is modelled by an "equivalent noise source" as eg. truncation errors introduced in data converters). As opposed to such concept of noise we will assume that all the usefful signals are deterministic.

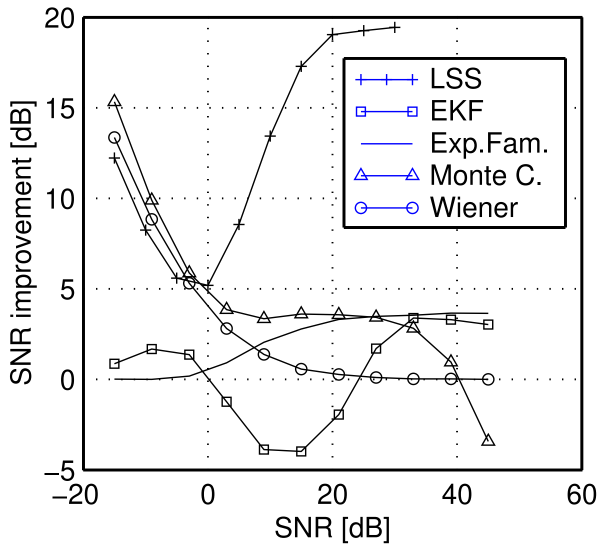

Figure 2: Improvement of the SNR vs. SNR for different noise reduction methods applied to noisy data from a chaotic Hénon map.

2 Probabilistic Framework

The dynamical systems we deal with in this paper are Markov processes in discrete time, which are often also referred to as Markov chains. A Markov process is in some sense a generalisation of a deterministic dynamical system. For a deterministic system, the future evolution is completely determined by the actual state of the system. For a Markov process, the actual state defines the future evolution of the system *up to unpredictable random influences*. In other words, the actual state of the system completely determines the *probability distribution* of the future states.

For example, consider dynamical systems of the following form:

$$X_{n+1} = F(X_n) + R_{n+1}, \tag{3}$$

where $F(\cdot)$ is a map on $\mathbb{R}^d$ and R_n is a series of independent identically distributed (iid) random vectors referred to as *dynamical noise*. Additionally R_n is independent of $\{X_1, \ldots, X_{n-1}\}$ and therefore an unpredictable random influence. Further examples of Markov processes are given in Subsection 2.2 after basic concepts of probability theory have been introduced.

Now assume that the state X_n of the Markov process itself is unaccessible, i.e. we cannot measure it directly. Instead we rely on measurements giving only incomplete state information. For example in an electric circuit having some internal degrees of freedom the only accessible quantity may be the voltage at a certain resistor. Of course, this quantity does not provide complete information about the full state of the circuit. In general, we assume the accessible quantity to be a function of the system state *plus some stochastic corruptions*. This means we collect measurements of the form

$$Y_n = G(X_n) + S_n, \tag{4}$$

where G is a function and S_n is an iid random variable independent of the whole process $\{X_n\}_{n=1\ldots\infty}$ with probability density function g. This random variable S_n is called *measurement noise* and accounts for random disturbances the measurement is corrupted with. A Markov process available only through measurements of the form (4) is known as *Hidden Markov Process* (HMP). HMP's serve as models in a lot of different applications, see [28] Let $\mathcal{Y}^n := \{Y_1, \ldots, Y_n\}$ be the available data set up to time n which will be referred to as a *time series* from now on. Everything we want to know about (3) has to be extracted from the time series $\mathcal{Y}^n$. Of course, the amount of information in the time series depends on n, and in general the larger n is, the more information about the underlying system we will be able to retrieve. In this paper we will adress the following problems:

orbit estimation: Suppose a time series $\mathcal{Y}^N$ has been recorded. Now try to recover the parameter vector α and the orbit $\{X_1, \ldots, X_N\}$ that is "most likely" to be the orbit that produced the time series.

state estimation: Suppose at every time instant n we are obliged to estimate the actual state X_n. This has to be done of course by means of the collected measurements $\mathcal{Y}^n$ up to time n, i.e. in a *causal* manner.

Both concepts will be defined and discussed in more detail in section 2.3. In most of the mathematics and control literature the problem of state estimation is called *filtering*, a term we will use synonymously to state estimation. In the engineering community filtering is often used synonymously to spectral manipulation. The reason is that for linear systems, the filtering problem (in our sense) can be solved by a certain spectral manipulation (Wiener filter). The Wiener filter is in fact the optimal *linear* filter. For nonlinear systems however this filter is not the optimal of all possible filters.

2.1 Basic notions of probability theory

The problems of state and orbit estimation will be tackled by means of probability theory. Although this paper is intended to be as self contained as possible, we assume the reader to be familiar with the very basic concepts of this theory. Before proceeding, we now review some facts needed here. An informal but very comprehensive introduction to probability and statistics is [52], a more mathematical treatment is [7].

For any event M its probability will be denoted by $\mathsf{P}(M)$. Suppose X is a random variable taking values in $\mathbb{R}^d$. Then for a set $A \subset \mathbb{R}^d$, by $\mathsf{P}(X \in A)$ we denote the probability of the event $\{X \text{ takes values in } A\}$.

Under some conditions, a random variable X has a probability density function (pdf) $p_X(x)$ that provides the probability $\mathsf{P}(X \in A)$ for any set $A \subset \mathbb{R}^d$ by

$$\mathsf{P}(X \in A) = \int_A p_X(x) \mathrm{d}x.$$

Likely we will write p_Y or p_Z for random variables Y or Z, for example.[2]

For two random variables X, Y and any set A, the *conditional probability* $\mathsf{P}(X \in A|Y)$ is very important. For fixed A, it is a function of Y and for fixed values of Y, it features a probability measure. Therefore, we can as well consider its density denoted by $p_{X|Y}(x; y)$.

The conditional probability (or its density) is an important concept in estimation theory. We shall explain the reason briefly. Assume we have access to a certain quantity Y, for example the outcome of a physical experiment. However, we want to know something about a certain quantity X we *do not* have access to. The quantity X is yet not completely determined by Y (this would then be an analytical problem), but the deviations are irregular disturbances having well known statistical features. Hence, Y determines X up to random disturbances, and this relation is fully described by the conditional probability density $p_{X|Y}$. Estimation theory is therefore concerned with calculating $p_{X|Y}$ for given problems and obtaining "reasonable estimates" from it.

A function $\hat{x}(y)$ that maps the measured quantity Y on possible values of X is called an *estimator* for X. The random variable $\hat{X} = \hat{x}(Y)$ is called an *estimate* of X. We will use the following three estimators in this paper:

1. The *conditional expectation*

$$\hat{x}(y) := E(X|Y)(y) := \int x \, p_{X|Y}(x; y) \, \mathrm{d}x \tag{5}$$

2. The *maximum likelihood estimator*

$$\hat{x}(y) := \lambda_{Y|X}(y) := \mathrm{argmax}_x \, p_{Y|X}(y; x). \tag{6}$$

3. The *maximum aposteriory estimator*

$$\hat{x}(y) := \lambda_{X|Y}(y) := \mathrm{argmax}_x \, p_{X|Y}(x; y). \tag{7}$$

The quantity

$$b := E(\hat{x}(Y)) - E(X)$$

is called the *bias* of an estimator $\hat{x}(y)$. If the bias vanishes, the estimator is called *unbiased*. Obviously, by subtracting the bias we always obtain an

[2] Often in the literature, the pdf of a random variable X is simply denoted as $p(x)$ and the x serves both as the argument and to indicate the corresponding random variable. Then however $p(x)$ and $p(y)$, may be two entirely different functions and not just the same function taken at different points. This notation becomes very inconvenient when $p(x)$ appears under an integral sign and substitutions of variables have to be employed. Then, sometimes, notations like $p(X = x)$ are used which we, however, consider very awkward.

unbiased estimator. The quantity $E(\hat{x}(Y)-X)^2$ is called the *mean square error* (mse) of the estimator. It can be shown that the conditional expectation has the least mean square error among all unbiased estimators.

Very often $p_{X|Y}$ is an everywhere positive function, i.e. it may be written as $p_{X|Y}(x;y) = e^{-J(x,y)}$. If J considered as a function of x is comparably large except for a certain minimum $\hat{x}(y)$, we may write

$$E(X|Y)(y) = \int x\, e^{-J(x,y)} \mathrm{d}x \cong \hat{x}(y).$$

So by minimizing the *cost function* $J(x,y)$ we may obtain an approximation for $E(X|Y)$. However, $\hat{x}(y)$ is the maximum aposteriory estimator for X. So if J has the property to be large except for a certain stationary point, we can expect the maximum aposteriory estimator and the conditional expectation to be almost equal.

The main facts about conditional pdf's needed in this paper are immediate consequences of the Bayes rule: Suppose X, Y, Z are random variables. If $P(X \in A, Y \in B|Z)$ has a density $p_{X,Y|Z}$, then also $P(X \in A|Y,Z)$ as well as $P(Y \in B|Z)$ have densities (say $p_{X|Y,Z}$ and $p_{Y|Z}$ resp.) and the following formula holds

$$p_{X,Y|Z}(x,y;z) = p_{X|Y,Z}(x;y,z) \cdot p_{Y|Z}(y;z) \tag{8}$$

Interchanging X and Y in the above formula yields

$$p_{Y,X|Z}(y,x;z) = p_{Y|X,Z}(y;x,z) \cdot p_{X|Z}(x;z). \tag{9}$$

Because the left hand sides of this two formulas are equal, we see

$$p_{X|Y,Z}(x;y,z) \cdot p_{Y|Z}(y;z) = p_{Y|X,Z}(y;x,z) \cdot p_{X|Z}(x;z). \tag{10}$$

Integrating equation (8) over y yields

$$p_{X|Z}(x;z) = \int p_{X|Y,Z}(x;y,z) \cdot p_{Y|Z}(y;z)\mathrm{d}y \tag{11}$$

The Bayes rule (10) can be used to establish a connection between maximum likelihood estimator and maximum aposteriory estimator: Choosing Z a constant in equation (10), we get

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

where c is a constant not depending on X. Now assume that p_X is as well a constant, which basically means a lack of *prior information* about X, we have

$$p_{X|Y} = c \cdot p_{Y|X},$$

whence maximum aposteriory estimator and maximum likelihood estimator are the same. So if no information about X is available exept for what is provided by the measurements, the concepts of mle and mae are equivalent. If however prior information is available, it is quite logical that it should be exploited, whence in this case the maximum aposteriory estimator should be preferred over the maximum likelihood estimator.

2.2 Markov Processes

We now look a little closer at Markov processes. For a general introduction see e.g. [7]. We already mentioned that for Markov processes the actual state defines completely the probability distribution of the future states. In other words, for predicting the future, all information about the past is redundant if the present state of the process is known. More formally, let $\{X_n\}_{n\in\mathbb{N}}$ be a Markov process. Fix $n \in \mathbb{N}$ and let Z be a random variable depending only on $X_{n+1}, X_{n+2}, \ldots$,i.e. on future values and Y be a random variable depending only on $X_{n-1}, X_{n-2}, \ldots$,i.e. on past values. Then

$$p_{Z|X_n,Y} = p_{Z|X_n}.$$

This is the *Markov property.* An important consequence is the following:

$$\begin{aligned} p_{X_n,\ldots,X_1} &= p_{X_n|X_{n-1}\ldots,X_1} \cdot p_{X_{n-1}|X_{n-2}\ldots,X_1} \cdot \cdots \cdot p_{X_1|X_0} \cdot p_{X_0} \\ &= p_{X_n|X_{n-1}} \cdot p_{X_{n-1}|X_{n-2}} \cdot \cdots \cdot p_{X_1|X_0} \cdot p_{X_0} \end{aligned}$$

where the first equality is an iterative application of Bayes rule and the second equality is the Markov property. So all statistical properties of a Markov process are completely defined by the initial pdf p_{X_0} and the *transition densities*

$$\varphi_n(x,z) := p_{X_n|X_{n-1}}(x,z).$$

A Markov process is called *homogenous* if its transition densities do not depend on n. This case is assumed throughout this paper and we will adopt the notation $\varphi(x,z)$ for the transition density from now on. In the following we will consider some examples of Markov processes and give their corresponding transition densities.

Example (Deterministic dynamical system):

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a map. Then

$$X_{n+1} = F(X_n)$$

defines a Markov process with transition function

$$\varphi(x;z) = \delta\left(x - F(z)\right).$$

The only stochastic influence in this model results from a random choice of initial condition X_0. Considering a deterministic system as a Markov process requires in fact a little bit more general notions from probability theory. First the transition function is actually a δ–function, which of course cannot be starting point of a rigorous theory. Furthermore, it is not clear (as tacitly assumed) that all X_n posses a probability density function.

Example (Additive dynamical noise):

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a map and R_n be a sequence of iid random variables having a gaussian pdf with covariance matrix R. Then

$$X_{n+1} = F(X_n) + R_{n+1} \tag{12}$$

defines a Markov process with transition function

$$\varphi(x;z) = \frac{1}{\sqrt{(2\pi)^d \det R}} \exp\left[-0.5\,(x - F(z))\,R^{-1}\,(x - F(z))\right]$$

Example (Binary message transmitter):

Let $F : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}^d$ be a map and M_n be a sequence of iid random variables taking the values 0 or 1 with probability p_0 or p_1, respectively. Then

$$X_{n+1} = F(X_n, M_{n+1}) \tag{13}$$

defines a Markov process with transition function

$$\varphi(x;z) = p_0\delta\,(x - F(z,0)) + p_1\delta\,(x - F(z,1))\,.$$

Ordinary and stochastic differential equations may also be considered in the framework of Markov processes. It should be noted that a process of the form

$$X_{n+1} = F(X_n, R_{n+1})$$

is *not* the most general Markov process. Discretely sampled nonlinear stochastic differential equations for example cannot be written in this form.

2.3 State and orbit estimation

The introduced concepts from probability theory will now be applied to orbit and state estimation. Recall that the latter problem is to estimate the state X_n by the measurements $\{Y_1, \ldots, Y_n\}$. To calculate the conditional expectation, the maximum likelihood estimator or the maximum aposteriory estimator, we have to calculate the conditional probability $P(X_n \in A|Y_1, \ldots Y_n)$ or its density

$$\pi_n(x) := p_{X_n|Y_1,\ldots Y_n}(x; y_1, \ldots, y_n), \tag{14}$$

Note that the notation $\pi_n(x)$ suppresses the dependency on the Y's. In case of orbit correction, we have to calculate the conditional probability $P(X_1 \in A_1, \ldots, X_N \in A_N|Y_1, \ldots Y_n)$ or its density

$$\Pi_N(x_1, \ldots, x_N) := p_{X_1,\ldots,X_N|Y_1,\ldots Y_N}(x_1, \ldots, x_N; y_1, \ldots, y_N). \tag{15}$$

Again, the dependency on the Y_n is suppressed. We will first discuss state estimation.

2.3.1 State estimation

Our aim is to give an iterative formula for π_n. Therefore we write

$$\begin{aligned}
&\pi_n(x) \\
&\quad = \quad p_{X_n|Y_1,\ldots,Y_n}(x; y_1, \ldots, y_n) \\
&\quad = \quad c \cdot p_{Y_n|X_n,Y_1,\ldots,Y_{n-1}}(y_n; x, y_1, \ldots, y_{n-1}) \\
&\qquad\quad \times p_{X_n|Y_1,\ldots,Y_{n-1}}(x; y_1, \ldots, y_{n-1}).
\end{aligned} \tag{16}$$

Here we applied equation (10) with $X := X_n$, $Y := (Y_1, \dots, Y_{n-1})$, $Z := Y_n$ and $c := 1/p_{Y|Z}(y; z)$, which does not depend on x.

To identify the first term on the right hand side of equation (16), recall that $Y_n = G(X_n) + R_n$, so if X_n is given, Y_n is distributed as R_n but with a mean shifted by $G(X_n)$. Now, since R_n has pdf g and is independent of $Y_1, \dots, Y_{n-1}$, we obtain

$$p_{Y_n|X_n,Y_1,\dots,Y_{n-1}}(y_n; x, y_1, \dots, y_{n-1}) = g(G(x) - y_n).$$

To identify the second term in (16) we simply apply equation (11) with $X := X_n$, $Y := X_{n-1}$ and $Z := (Y_1, \dots, Y_{n-1})$. Since $\{X_n\}$ is a Markov Process , Z is a random quamtity depending only on the past and therefore is redundant. Hence we can replace $p_{X_n|X_{n-1},Y_1,\dots,Y_{n-1}}$ by $p_{X_n|X_{n-1}}$ which yields

$$\begin{aligned} & p_{X_n|Y_1,\dots,Y_{n-1}}(x; y_1, \dots, y_{n-1}) \\ = & \int p_{X_n|X_{n-1},Y_1,\dots,Y_{n-1}}(x; z) \\ & \times p_{X_{n-1}|Y_1,\dots,Y_{n-1}}(z; y_1, \dots, y_{n-1})\mathrm{d}z \\ = & \int p_{X_n|X_{n-1}}(x; z) \cdot p_{X_{n-1}|Y_1,\dots,Y_{n-1}}(z; y_1, \dots, y_{n-1})\mathrm{d}z \end{aligned}$$

The last term however is $\pi_{n-1}(z)$. So substituting in equation (16) we finally have

$$\pi_n(x) = c \cdot g(G(x) - y_n) \cdot \int p_{X_n|X_{n-1}}(x; z) \cdot \pi_{n-1}(z)\mathrm{d}z. \tag{17}$$

This is the desired formula for computing the evolution of the pdf $\pi_n(x)$ using information from past values $\{y_1, \dots, y_n\}$ of the measured time series. The constant c may be calculated from the condition $\int \pi_n(x)\mathrm{d}x = 1$.

Calculating π_n by means of equation (17) for a given time series $\{Y_1 \dots Y_n\}$ consists basically of the following steps:

Initial condition

$$\pi_0(x) := p_{X_0}(x).$$

Here $p_{X_0}(x)$ represents our prior knowledge about the initial condition of the underlying process.

Prediction step

$$\pi_n^+(x) := \int \varphi(x; z)\pi_n(z)\mathrm{d}z$$

This describes the evolution of the pdf due to the underlying Markov process (or dynamical system).

Update step

$$\pi_{n+1} := \frac{g(Y_{n+1} - G(x)) \cdot \pi_n^+(x)}{\int[\text{numerator}]\mathrm{d}x}$$

This describes the “update” of the pdf using information provided by the latest sample Y_n given by multiplication with the pdf g of the noise centered at Y_n.

Figure 3 shows an illustration of these two steps for noisy data

$$Y_n = X_n + S_n$$

generated by the tent map:

$$X_{n+1} = 1 - 2|X_n - 0.5| \tag{18}$$

The noise is gaussian with pdf g. In Fig. 3a the pdf $\pi_n(x)$ is shown that is transformed according Eq. (17) to $\pi_{n+1}(x)$ given in Fig. 3c. Figure 3b shows intermediate results. The dotted line denotes the result of the integration in (17) (the "prediction step") and the solid line is the gaussian distribution g centred at the measured sample Y_n. These two pdf are multiplied in Eq. (17) to combined the information from time evolution of the previous state and the new measurement.

Knowing the conditional pdf π_n almost any estimation problem can be solved. For any (Borel) function f, the estimator $E(f(X_n)|Y_1, \ldots, Y_n)$ can be calculated by

$$E(f(X_n)|Y_1, \ldots, Y_n) = \int f(x)\pi_n(x)\mathrm{d}x.$$

Furthermore, the maximum likelyhood estimate for X_n can be obtained by

$$\hat{X}_n = \mathrm{argmax}_x \pi_n(x).$$

The question arises whether π_n can actually be calculated from equation (17) in closed form. Unfortunately this is possible in only very few cases (including the linear case). Hence for practically all nonlinear cases approximations are essential. This will be discussed in more detail in section 3.

2.3.2 Orbit estimation

Very similar considerations as in the case of state estimation lead to a representation of $\Pi_n(x_0, \ldots, \ldots x_N; y_1, \ldots, y_N) := p_{X_0,\ldots,X_N|Y_1,\ldots Y_N}(x_0, \ldots, \ldots x_N; y_1, \ldots, y_N)$ required for orbit estimation. The result will be presented here without derivation

$$\begin{aligned} &\Pi_N(x_1, \ldots, x_N; y_1, \ldots, y_N) \\ &= \prod_{i=1}^{N} g(G(x_i) - y_i) \cdot \varphi(x_i, x_{i-1}) \cdot p_{X_0}(x_0) \end{aligned} \tag{19}$$

Recall that $p_{X_0}$is the pdf of X_0. In the special case of example (12) we get, inserting the expressions for φ and g,

$$\begin{aligned} &\Pi_N(x_1, \ldots, x_N; y_1, \ldots, y_N) \\ &= c \cdot \exp -\frac{1}{2} \left\{ \sum_{i=1}^{N} (G(x_i) - y_i) S^{-1} (G(x_i) - y_i) + \right. \\ &\left. \sum_{i=1}^{N} (x_i - F(x_{i-1})) R^{-1} (x_i - F(x_{i-1})) \right\} \cdot p_{X_0}(x_0). \end{aligned}$$

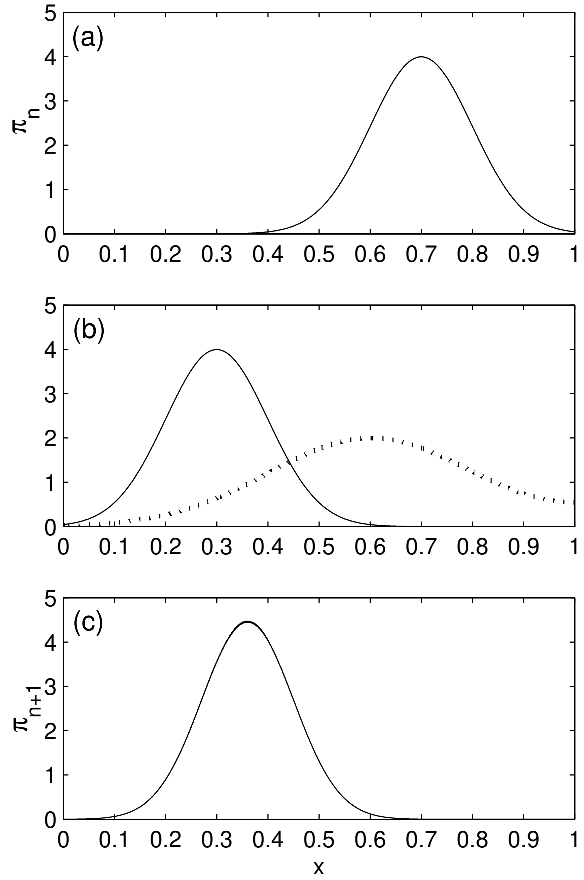

Figure 3: Illustration of Eq.(17) using the tent map (18). (a) pdf $\pi_n(x)$, (b) result of the integration in Eq.(17) (dotted curve) and pdf g of the measurement centered at the Y_n (solid curve), (c) pdf π_{n+1} as product of both pdfs shown in (b).

If $p_{X_0}(x_0) = c \cdot \exp -\frac{1}{2}\left((x-\mu)\Lambda^{-1}(x-\mu)\right)$ with given Λ and μ, we have the representation

$$\begin{aligned} &\Pi_N(x_1,\ldots,x_N;y_1,\ldots,y_N) \\ &= \quad c \cdot e^{-\frac{1}{2}J(x_1,\ldots,x_N;y_1,\ldots,y_N)} \end{aligned}$$

with the cost functional

$$\begin{aligned} &J(x_0,\ldots,x_N;y_1,\ldots,y_N) \\ &= \quad (x-\mu)\Lambda^{-1}(x-\mu) + \sum_{i=1}^{N}(G(x_i)-y_i)S^{-1}(G(x_i)-y_i) + \\ &\quad \sum_{i=1}^{N}(x_i - F(x_{i-1}))R^{-1}(x_i - F(x_{i-1})). \end{aligned} \tag{20}$$

Estimators may now be obtained by either conditional expectations or by stationary points of the cost functional (with respect to the arguments $x_1,\ldots,x_N$), which yield the maximum aposteriory estimator.

In many cases X_1 is not gaussian but equidistributed. This means that no information about X_1 is available a priori. In this case, p_{X_0} contributes no information to the problem and we get the maximum likelihood estimator by minimizing the cost functional

$$\begin{aligned} &J(x_1,\ldots,x_N;y_1,\ldots,y_N) \\ &= \quad \sum_{i=1}^{N}(G(x_i)-y_i)S^{-1}(G(x_i)-y_i) + \\ &\quad \sum_{i=1}^{N-1}(x_{i+1} - F(x_i))R^{-1}(x_{i+1} - F(x_i)). \end{aligned}$$

Another problem of interest occurs for vanishing dynamical noise. In this case, $\varphi(x,z) = \delta(x - F(z))$. Obviously, J is no longer defined. For a maximum likelyhood estimate however one may seek a minimizer of the cost function

$$J(x_1,\ldots,x_N;y_1,\ldots,y_N) = \sum_{i=1}^{N}(G(x_i)-y_i)S^{-1}(G(x_i)-y_i)$$

with respect to the *constraint*

$$x_{i+1} - F(x_i) = 0 \qquad \text{for all } i = 1,\ldots,N-1.$$

This minimisation problem may formally be obtained by letting $R \to 0$ in equation (20).

3 State Estimation (Filtering)

The objective of this section is to provide (approximative) solutions for equation (17) describing the temporal evolution of the pdf $\pi_n(x)$. This is far from being straigth forward, since it features a functional equation. Approximation techniques will turn out to be essential. We again assume that $\{X_n\}_{n\geq 0}$ is a Markov process with transition density function $\varphi(x,y)$. We will also encounter Markov processes having transition densities that can be expressed only by delta functions, but we will use this slightly informal yet very convenient notation. In all cases, observations are taken according to the equation

$$Y_n = G(X_n) + S_n$$

where $G : \mathbb{R}^d \to \mathbb{R}^l$ is a mapping and S_n is a series of iid random variables having the common pdf g.

In general, for Markov processes there exists no closed form solution of equation (17). One important exception is the linear case, where the

Markov process is given by the equation

$$X_{n+1} = FX_n + R_{n+1}$$

where R_n has a gaussian distribution with covariance matrix R, and F is a $d\times d$–matrix. Furthermore assume X_0 has a gaussian distribution with covariance matrix Γ_0. Let the measurement process be given by the equation

$$Y_n = GX_n + S_n$$

where S_n has a gaussian distribution with covariance matrix S, and G is a $d \times l$–matrix. Then

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

where Γ_n and μ_n are given by

$$\begin{aligned} \Gamma_{n+1}^{-1} &= \left(F\Gamma_n F^t + R\right)^{-1} + G^t S^{-1} G \\ \mu_{n+1} &= F\mu_n + \Gamma_{n+1} G^t S^{-1}\left(Y_{n+1} + GF\mu_n\right) \end{aligned}$$

This is a direct consequence of equation (17).

This result provides not only a solution to equation (17) but shows that in this case the infinite dimensional problem can be reduced to a finite dimensional problem, namely the equations for μ_n and Γ_n characterizing the *Kalman Filter*. In the general nonlinear case, no such theorem is available. A number of recent papers (see [30, 63] and references therein) was concerned with the question of existence of finite dimensional filters for nonlinear systems. Basically, it turned out that obtaining a finite dimensional filter for a given state space model is a very unusual event. Therefore, approximations are required for most of the cases of interest.

The classical approach to nonlinear problems is the *Extended Kalman Filter* (EKF) already encountered in the first section. The idea is to linearize the nonlinear equations along the estimated trajectory and then apply the Kalman filter. The EKF has seen successful application in many technical problems. For an overview, see [19]. The EKF has, however, a couple of serious drawbacks. First, it tends to fail if bad initial conditions are given. Furthermore, it will yield bad results if the problem is not well described by its linearisation, especially for chaotic maps.

In this paper we will consider basically two different kinds of approximations. The first one considers $\pi_n(x)$ as a sequence in a function space. The problem is the infinite dimensionality of this space, which is usually exploited by the sequence π_n. The idea of *parametric approximation* is to consider a finite dimensional submanifold in this space and to project π_n onto this manifold. The coordinates of this manifold then serve as parameters yielding a finite dimensional representation. For continuous time problems, this program has been carried out by several authors, see [9, 8]. The second approach is an extension of the *Monte Carlo method.* Basically any Monte Carlo method employs a large ensemble of particles. This ensemble provides an empirical distribution. The idea is to prepare the ensemble so that if the number of particles goes to infinity the limit distribution yields the required distribution. We will now discuss the parametric and the Monte Carlo approximation in detail.

3.1 Parametric approximation of π_n

Consider a set $\mathcal{P}$ of positive integrable functions on $\mathbb{R}^d$ which are normalized, i.e. $\int p \mathrm{d}x = 1$ for all $p \in \mathcal{P}$. A *parametrisation* of $\mathcal{P}$ is a mapping

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

where Θ is a subset of a finite dimensional vector space. The parametrisation 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 will be called a *parametrized set of pdfs.* Parametric sets of pdfs are well known in statistics, especially in parametric estimation theory, see [2, 52].

The basic idea of parametric approximation is to chose a parametrized family $(\mathcal{P}, p)$ and replace π_n by a sequence in $\mathcal{P}$ which by p can be pulled back to a sequence θ_n in Θ. We will consider two different classes of $\mathcal{P}$'s, namely *exponential* and *linear* families. They will be introduced after the general scheme has been explained.

As a metric between two pdfs p_1, p_2 we introduce the *Kullback–Leibler* distance

$$d(p_1, p_2) := \int -\log \frac{p_1}{p_2} p_2 \mathrm{d}x.$$

The Kullback–Leibler distance is neither symmetric nor fulfills the triangle inequality. Nevertheless, $d(p_1, p_2)$ is always positive (Kullback–Leibler inequality) and vanishes if and only if $p_1 = p_2$. The Kullback–Leibler distance is an interesting object occuring in many different fields of statistics. For its improtance in limit theorems see [4].

There are many other useful metrics for probability densities that in principle could be employed for our purposes. We found yet the Kullback–Leibler distance the most convenient one from a computational point of view, it is however by no means obvious whether the resulting estimators have any conceptual advantage over estimators that one obtains using other metrics.

Approximating π_n by parametrized families of pdfs can be done in many ways, and we decided to do it as follows. We want to approximate π_n by a simpler sequence of pdfs $\tilde{\pi}_n$ of a parametrized family of pdfs. We have basically two schemes differing only in when the approximation takes place: after the prediction and update steps are performed or between prediction and update step (see discussion after equation (17)).

Scheme I: Here the approximation step takes place at the end after both the prediction and update steps are performed. Let $p(\cdot, \theta)$ be the parametrisation of a parametrized family and let $d(\cdot, \cdot\cdot)$ denote the Kullback–Leibler distance.

Initial condition

$$\tilde{\pi}_0(x) := p_{X_0}(x)$$

Prediction and update step By applying the equation (17), calculate the intermediate result

$$\bar{\pi}_{n+1}(x) := \frac{g(Y_{n+1} - G(x)) \cdot \int \varphi(x; z)\tilde{\pi}_n(z)\mathrm{d}z}{\int [\text{numerator}]\mathrm{d}x}$$

Approximation step Calculate

$$\theta_{n+1} := \arg\min_{\theta} d(p(\cdot, \theta), \bar{\pi}_{n+1})$$

and then set $\tilde{\pi}_{n+1} = p(\cdot, \theta_{n+1})$.

Scheme II: Here the approximation step takes place between the prediction and the update step. Let $p(\cdot, \theta)$ be the parametrisation of a parametrized family and let $d(\cdot, \cdot\cdot)$ denote the Kullback–Leibler distance.

Initial condition

$$\tilde{\pi}_0(x) := p_{x_0}(x)$$

Prediction step

$$\tilde{\pi}_n^+(x) := \int \varphi(x; z)\tilde{\pi}_n(z)\mathrm{d}z$$

Approximation step Calculate

$$\theta_n^+ := \arg\min_{\theta} d(p(\cdot, \theta), \tilde{\pi}_n^+)$$

Update step

$$\tilde{\pi}_{n+1}(x) := \frac{g(Y_{n+1} - G(x)) \cdot p(x, \theta_n^+)}{\int [\text{numerator}] \mathrm{d}x}$$

The main difference between the schemes is that for scheme I, $\tilde{\pi}_n$ is always a member of the parametrized family. This is not the case for scheme II. There, however, the update step is performed exactly, which may be an advantage. We now turn to the explanation of linear and exponential families.

3.1.1 Linear Families

The use of these families is motivated by the quite obvious idea of approximating the pdf π_n by a piecewise constant function, where the partition on which the function is constant is kept fixed. The piecewise constant function can also be considered as a convex combination of fixed characteristic functions. To generalize this idea, let $p_i(x), i = 1 \dots k$ be a set of normalized positive functions, i.e. $\int p_i(x)\,\mathrm{d}x = 1$ for all $i = 1 \dots k$. Then the resulting *linear family* $\mathcal{P}$ consists of all convex combinations of the form

$$p(x, \theta) = \sum_{i=1}^{k} \theta_i p_i(x)$$

where the θ's form the convex simplex

$$\Theta = \{\theta \in \mathbb{R}^k; 0 \leq \theta_i \leq 1, \sum_{i=1}^{k} \theta_i = 1\}.$$

For any pdf q the Kullback–Leibler distance $d(p(\cdot, \theta), q)$ reads as

$$d(p(\cdot, \theta), q) = \int \log(q)\, q\, \mathrm{d}x - \int \log\left(\sum_{i=1}^{k} \theta_i p_i(x)\right) q\, \mathrm{d}x.$$

Since $-\log(x)$ is a strictly convex function, again minimizing the Kullback–Leibler distance is a convex problem having a unique solution.

As a special class of linear models we consider the case of nonoverlapping support of the p_i's. This is, for example the case when a pdf is represented by a piecewise constant function (histogram) on a grid. Using the method of Lagrange multipliers it turns out that minimizing the Kullback–Leibler distance is equivalent to maximizing the function

$$J(\theta, \mu) := \int \log\left(\sum_{i=1}^{k} \theta_i p_i(x)\right) q\, \mathrm{d}x - \mu\left(\sum_{i=1}^{k} \theta_i - 1\right)$$

for θ and the Lagrange multiplier μ. Taking the derivative with respect to θ_j and setting it equal to zero yields

$$\int \frac{p_j}{\sum_{i=1}^{k} \theta_i p_i(x)} \cdot q\, \mathrm{d}x = \mu.$$

The integral however only has to be extended over the support of p_j, where all other $p_i, i \neq j$ vanish which yields

$$\int_{\text{support}(p_j)} q \, \mathrm{d}x = \mu\theta_j.$$

Setting

$$\mu = \sum_j \int_{\text{support}(p_j)} q \, \mathrm{d}x$$

we get finally

$$\frac{\int_{\text{support}(p_j)} q \, \mathrm{d}x}{\sum_j [\text{numerator}]} = \theta_j$$

fulfilling also the requirement $\sum_j \theta_j = 1$.

Unfortunately, approximations of pdfs using piecewise constant functions on a grid are feasible in low dimensions, only. The simulations shown in Fig. 3 for the one-dimensional tent map have been performed in this way. Already in two dimensions, however, the chaotic stretch and fold mechanism renders the pdf very complicated. In particular for low noise it becomes a wildly fluctuating function that is localized along the underlying chaotic attractor (compa. fig. 1). In this case adapted grids are necessary to resolve the pdf properly.

3.1.2 Exponential Families

Exponential families are very important in statistical estimation theory, since many important random quantities obey an exponential distribution law. A thorough discussion of this subject is given in [5]. Their application to filtering of discrete time series was discussed in [10, 11]. We already mentioned the succesful application to the continuous time case in [9]. Let $\lambda(x)$ be a positive function on $\mathbb{R}^d$ and $c_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1 \ldots k$ is a set of functions. Let Θ be the set of all $\theta \in \mathbb{R}^k$ for which

$$\psi(\theta) := \log \int \exp(\sum_i \theta_i c_i(x)) \, \lambda(x) \mathrm{d}x$$

is finite. A straight forward application of Hölders inequality yields that ψ is a convex function and Θ a convex set. Define the parametrisation

$$p : \Theta \to \mathcal{P}, \theta \to \exp(\sum_i \theta_i c_i(x) - \psi(\theta)) \cdot \lambda(x)$$

Then $\mathcal{P}$ is an exponential family that contains in particular the well-known gaussian pdf. If we require the functions $c_i(x)$ to be affinely independent, i.e. the function $\theta_0 + \sum \theta_i c_i(x)$ vanishes if and only if all θ are equal to zero, then the parametrisation turns out to be faithful. The function ψ called the *potential* renders the pdf normalized: $\int p(x, \theta) \, \mathrm{d}x = 1$. Taking the derivative with respect to θ_i on both sides one obtains

$$\eta_i := \int c_i(x) \, p(x, \theta) \, \mathrm{d}x = \frac{\partial \psi}{\partial \theta_i}(\theta).$$

The η_i are called the c_i–*moments* or *expectation parameters*, in contrast to the θ_i, which are called *canonical parameters*. The c_i's are called *canonical statistics*. One easily obtains the following identity:

$$g_{ij} := \frac{\partial \eta_i}{\partial \theta_j} = \int \frac{\partial \log p}{\partial \theta_i} \frac{\partial \log p}{\partial \theta_j} \, p \, \mathrm{d}x = \frac{\partial^2 \psi}{\partial \theta_i \partial \theta_j}.$$

Since the c_i's are affinely independent, the functions $\partial \log p / \partial \theta_i$ are linear independent and therefore g_{ij} is a nonsingular positiv definite matrix, called the FISHER *metric*. Hence, the function ψ is *strictly* convex. Furthermore, it is easy to see that the expectation parameters η and the canonical parameters θ are connected by a Legendre transform of the function ψ, i.e.

$$\theta(\eta) = \arg\max_{\theta} \left[\theta\eta - \psi(\theta) \right].$$

The Legendre transform of a strictly convex function however uniquely connects θ and η, hence the expectation parameters η_i are *globally* diffeomorphic functions of the θ_i. Therefore the expectation parameters form another coordinate system for $\mathcal{P}$, which is of great use in the following.

Now we consider the Kullback–Leibler distance for exponential families. More specific, let $p_1 = p(x, \theta)$ belong to an exponential family and p_2 be an arbitrary measure. This yields

$$d(p_1, p_2) = \int \log \frac{p_2}{\lambda} p_2 \, \mathrm{d}x - \left(\sum \theta_i \int c_i(x) p_2(x) \mathrm{d}x - \psi(\theta) \right)$$

Minimizing this expression w.r.t. θ is again related to a Legendre transform of $\psi$and therefore a convex optimisation problem. Unfortunately, for many convenient exponential families there is neither a closed form expression for ψ nor its Legendre transform. Therefore numerical schemes have to be employed to compute them. We will briefly discuss some possible approaches in the appendix.

A few remarks concerning the approximation by parametrized densities seem to be in order:

1. We have seen that in the approximation scheme II $\tilde{\pi}_n$ is not a member of the parametrized family. One may ask if for some special cases one can achieve that even in scheme II, $\tilde{\pi}_n$ stays in the parametrized family. This is possible if $g(y - G(x))$ as a function of x is of exponential type. One may then chose an exponential family containing also g, and since the multiplication of two exponential densities again yields an exponential density the update step in scheme II will keep $\tilde{\pi}_{n+1}$ a member of the exponential family.

   An important example is the following. Suppose g is a gaussian density. Then
   $$g(y - G(x)) = c \cdot \exp\left(\frac{1}{s} \left[yG(x) - 0.5G^2(x) \right]\right).$$
   So if $G(x)$ and $G^2(x)$ affinely depend on the canonical statistics, then the update step will keep $\tilde{\pi}_{n+1}$ a member of the exponential family.

2. It turns out that the presented framework unifies a lot of known different approaches. E.g. approximation by linear families (in our language) was already proposed in the 1970's (see [61]). Furthermore, using exponential families amounts to compute a few moments of the actual distribution and discard the higher order ones (see the Appendix). This is in fact the main idea behind the *assumed density principle* (see [42, 76]). This approach has however been carried out only for gaussian densities. Based on the assumed gaussian density filter, a further simplification has been proposed by Julier and Ullmann (the *unscented filter*, see [43]). The main idea here is to replace the exact calculation of the moments by an approximation that is applicable also in cases where the discrete time dynamical system equations are not given in mathematically closed form (e.g. if a continuous time system is investigated and a numerical integration scheme is employed).

3. The simplest approximation of pdfs within the framework of exponential families are gaussian functions and the noise reduction of the Henon time series shown in Fig. 1b as well as the dash–dotted SNR improvement curve in Fig. 2 were computed in this way.

3.2 Monte Carlo Methods

Classically Monte Carlo methods where conceived to evaluate certain integrals that can be understood as the expectation value of a random quantity. Suppose for example f is a function and p a probability density and we want to calculate

$$\int f(x) \cdot p(x) \,\mathrm{d}x$$

The idea of Monte Carlo simulation simply is to generate a large number of other independent random variables $X_1, X_2, \ldots, X_M$ (called the *ensemble*) featuring the same statistical properties, that is in this case having the distribution $p(x)$. Then (according to the law of large numbers) the expectation is approximately given by the empirical mean over the ensemble, which in our case means

$$\int f(x) \cdot p(x) \,\mathrm{d}x \cong \frac{1}{M} \sum_{k=1}^{M} f(X_k).$$

The problem is of course how to generate the ensemble, which may be quite difficult for complicated p.

These ideas can be modified for the purpose of state estimation in two ways. The idea of weighted particles is, roughly speaking, to work with a large number of independent Markov Processes featuring the same statistical properties as the original signal process $\{X_n\}$. The ensemble does not provide an approximation of π_n, but allows for approximative calculation of any integral of the form

$$\int f(x) \cdot \pi_n(x) \mathrm{d}x$$

by a *weighted* average over the ensemble of Markov processes. The weights depend on the observations Y_n.

The SNR improvement result in Fig. 2 denoted "Monte C." was computed with the weighted particle approach. The decrease of the SNR improvement around 40 dB is due to the (finite) ensemble size of $M = 1000$. Using more "particles" one may achieve better SNR improvement also for higher SNR of the time series (with higher computational costs, of course).

An alternative is the method of evolutionary particles which in contrast to the weighted particles directly provides an ensemble approximating π_n. This method consists of two steps resembling the prediction and the update step. The ensemble points are not independent like in the weighted particle method, whence the method is often called interacting particle method.

3.2.1 Weighted Particle Method

The method of weighted particles was proposed and investigated theoretically by [25]. The idea is to generate M independent Markov processes $\{X_n^{(k)}\}_{n\leq 0, k=1\ldots M}$, where n is, as before, the time and k denotes the k'th member of the ensemble. All copies have the same statistics as the original signal process $\{X_n\}$, i.e. the same initial distribution and the same transition pdf. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary function. As already mentioned before, the method provides an approximation to the quantity

$$E(f(X_n)|Y_1, \ldots, Y_n) = \int f(x) \cdot \pi_n(x)\,\mathrm{d}x$$

by a weighted average over the ensemble points $\{X_n^{(k)}\}_{k=1\ldots M}$ at fixed time n, i.e.

$$\int f(x) \cdot \pi_n(x)\,\mathrm{d}x \cong \sum_{k=1}^{M} w_n^{(k)} \cdot f(X_n^{(k)}).$$

It only remains to give an expression for $w_n^{(k)}$. Define the quantities

$$g_j^{(k)} := g(Y_j - G(X_j^{(k)})) \qquad \text{for all } j \leq 0, k = 1, \ldots, M.$$

Theoretically one can prove that

$$w_n^{(k)} = c \cdot \prod_{j=1}^{n} g_j^{(k)},$$

where c is a constant chosen to yield

$$\sum_k w_n^{(k)} = 1.$$

A profound analysis of the problem however shows that this method tends to diverge, and one should rather implement a *limited memory version* of the filter, where the memory depends on the ensemble size M. This is done as follows: Let q_M be a certain positive integer depending on the ensemble size M. Then define the weights to be

$$w_n^{(k)} = c \cdot \prod_{j=n-q_M}^{n} g_j^{(k)},$$

where we define $g_j^{(k)} := 1$ if j is negative or zero. Practically this method can be implemented as follows: Choose M and $q_M =$ integer closest to $2\sqrt{\log(M)}$.

Initial condition: Let $X_0^{(1)}, \ldots, X_0^{(M)}$ be independent samples of the pdf p_{X_0}. For each $k = 1, \ldots, M$ allocate a *weight vector* $g^{(k)} := [g_1^{(k)}, \ldots, g_{q_M}^{(k)}]$ and set all entries equal to one.

From n to $n+1$: Assume the ensemble $X_n^{(1)}, \ldots, X_n^{(M)}$ and the weight vectors $g^{(1)}, \ldots, g^{(M)}$ for time instant n are given. For each $k = 1, \ldots, M$ let

1. $X_{n+1}^{(k)}$ be a sample point of the pdf $\varphi(\cdot, X_n^{(k)})$,
2. $\bar{g}_j^{(k)} = g_{j+1}^{(k)}$ for $j = 1, \ldots, q_M - 1$
3. $\bar{g}_{q_M}^{(k)} = g(Y_{n+1} - G(X_{n+1}^{(k)}))$.

Then $X_{n+1}^{(1)}, \ldots, X_{n+1}^{(M)}$ is the new ensemble and $\bar{g}^{(1)}, \ldots, \bar{g}^{(M)}$ the new weight vectors at time $n+1$, which are renamed $g^{(1)}, \ldots, g^{(M)}$ for the next time step.

For any function f, the conditional expectation then is approximately given by

$$E(f(X_n)|Y_1 \ldots Y_n) \cong \frac{\sum_{k=1}^{M} \prod_{j=1}^{q_M} g_j^{(k)} \cdot f(X_n^{(k)})}{\sum_{k=1}^{M} \prod_{j=1}^{q_M} g_j^{(k)}}$$

3.2.2 Evolutionary Particle Method

The evolutionary particle method uses an ensemble of particles generated by a sampling procedure resembling the prediction/update mechanism of equation (17) which renders the ensemble a particle approximation of π_n itself. Let again $X_n^{(1)}, \ldots, X_n^{(M)}$ denote the ensemble at time n. Fix another positive integer M_C called the *number of children*. The dynamics of the particle ensemble is given by the following evolutionary process:

Initialisation: Let $X_0^{(1)}, \ldots, X_0^{(M)}$ be independent samples of the pdf p_{X_0}.

Prediction step: For each $k = 1, \ldots, M$ fixed produce an ensemble $X_{n+1}^{(k,1)}, \ldots, X_{n+1}^{(k,M_C)}$ of M_C *children* of $X_n^{(k)}$ by sampling M_C times independently from the pdf $\varphi(\cdot, X_n^{(k)})$.

Update step: To each $X_{n+1}^{(k,l)}$ assign the probability

$$p_{k,l} := \frac{g(Y_{n+1} - G(X_{n+1}^{(k,l)}))}{\sum_{k,l}[\text{numerator}]}$$

and select M of the children $X_{n+1}^{(k,l)}$ each with probability $p_{k,l}$. This gives the new ensemble $X_{n+1}^{(1)}, \ldots, X_{n+1}^{(M)}$.

For any function f, the conditional expectation then is approximately given by

$$E(f(X_n)|Y_1 \ldots Y_n) \cong \frac{1}{M} \sum_{k=1}^{M} f(X_n^{(k)}).$$

As far as we know, the Evolutionary Particle Method remains to be investigated numerically as well as theoretically. Hopefully, also large deviation results like in [25] can be obtained.

4 Orbit Estimation

Most of the papers considering noise reduction that emerged from the nonlinear dynamics community adress the case of orbit estimation rather than state estimation which is more an issue in the control engineering community.

4.1 An Optimisation Method for Deterministic Systems

We consider the case of a deterministic nonlinear system with observation noise, only:

$$\begin{aligned} X_{n+1} &= F(X_n) \\ Y_n &= G(X_n) + S_n. \end{aligned} \tag{21}$$

The aim is to find estimates $\{\hat{X}_1, \ldots, \hat{X}_N\}$ of the original states $\{X_1, \ldots, X_N\}$ from the collected measurements $\{Y_1, \ldots, Y_N\}$. In section 2 we derived a constrained minimisation principle

$$\text{minimize (for the } x_n\text{'s)} \qquad \sum_{n=1}^{N} (G(x_n) - Y_n)^2$$

subject to the constraints

$$x_{n+1} - F(x_n) = 0 \text{ for } n = 1, \ldots, N-1.$$

The minimizer is the required estimate $\{\hat{X}_1, \ldots, \hat{X}_N\}$. In [36] it was suggested to include the constraints into the cost function with an appropriate weight factor λ yielding an unconstrained minimisation principle

$$\sum_{n=1}^{N} (G(x_n) - Y_n)^2 + \lambda \sum_{n=1}^{N-1} \|x_{n+1} - F(x_n)\|^2$$

We have seen that the minimizing argument of the cost function is the maximum likelyhood estimate of the orbit in the presence of dynamic noise. In this case, the weight factor turns out to be the ratio of the dynamic and the measurement noise. If no dynamic noise is present, one may as well apply the principle iteratively letting λ go to infinity. However, finding a minimizer of such a highly complex cost function is not easy. At least, we found simple methods like Powells method or gradient descent (see [57]) to fail.

The first step is to transform the system into a simpler form using the so called *observability map* Ψ, which is defined as

$$\Psi : \mathbb{R}^d \to \mathbb{R}^d, x \to [G(x), G \circ F(x), \ldots, G \circ F^{d-1}(x)].$$

Since we want to use the observability map as a transformation we assume it to be a global diffeomorphism. Furthermore, we assume G to be a scalar valued function. A dynamical system featuring a diffeomorphic observability map is called *observable*. Defining the state variable

$$Z_n := \Psi(X_n)$$

it is easy to see that the system has the form

$$\begin{aligned} Z_{n+1}^{(1)} &= Z_n^{(2)} \\ Z_{n+1}^{(2)} &= Z_n^{(3)} \\ &\vdots \\ Z_{n+1}^{(d)} &= f(Z_n) \\ \\ Y_n &= Z_n + S_n, \end{aligned} \tag{22}$$

Here f is given by

$$f(z) := G \circ F^d \circ \Psi^{-1}(z),$$

and by $Z_n^{(k)}$ we denote the k'th element in the vector Z_n, i.e.

$$Z_n^{(1)} := G(X_n), Z_n^{(2)} := G \circ F(X_n), \ldots$$

Note that $Z_n = [Z_n^{(1)}, \ldots, Z_{n+d-1}^{(1)}]$. Therefore in (22) all variables can be replaced by $Z_n^{(1)}$ and its temporal predecessors yielding the structure

$$\begin{aligned} Z_n^{(1)} &= f([Z_{n-d}^{(1)}, \ldots, Z_{n-1}^{(1)}]) \\ Y_n &= Z_n^{(1)} + S_n, \end{aligned} \tag{23}$$

We henceforth assume that the system has the structure (23). The minimisation principle now reads as

$$\begin{array}{ll} \text{minimize (for the } z_n\text{'s)} & \displaystyle\sum_{n=1}^{N} (z_n - y_n)^2 \\ \text{subject to the constraints} & z_n - f(z_{n-d}, \ldots, z_{n-1}) = 0 \\ & \text{for } n = d+1, \ldots, N \end{array}$$

where $\{y_1, ..., y_N\}$ is the measured time series. In the following we shall use the abbreviation

$$z_1^N = \{z_1, \ldots, z_N\}$$

and analogous definitions for y_1^N, z_1^N, etc. Furthermore we define the functions

$$\begin{aligned} J(y_1^N, z_1^N) \quad &:= \quad \sum_{n=1}^{N} (z_n - y_n)^2, & (24) \\ \Phi(z_1^N) \quad &:= \quad \{z_n - f(z_{n-d}, \ldots, z_{n-1})\}_{n=d+1,\ldots,N}. & (25) \end{aligned}$$

Using these abbreviations we have to minimize $J(y_1^N, z_1^N)$ as a function of z_1^N subject to the nonlinear $(N-d)$–dimensional constraint $\Phi(z_1^N) = 0$. Solving this problem is very difficult in general. Note that if the initial condition z_1 is considered as the independent variable, all other z_n's are completely determined. Therefore, the choice of the initial condition is the only degree of freedom in the problem. But for chaotic dynamics the quadratic error $J(y_1^N, z_1^N)$ will depend very sensitive on z_1, resulting in a rather complicated error landscape with many local extrema. So any minimisation of $J(y_1^N, z_1^N)$ with respect to z_1^N will be very difficult. Therefore, in the following, we shall *not* try to minimise $J(y_1^N, z_1^N)$ directly but use an alternative strategy to find a less noisy deterministic orbit z_1^N close to the measured orbit y_1^N.

Suppose we could obtain an exact orbit z_1^N by iteratively adding small corrections to y_1^N. Then z_1^N is a shadowing orbit [3] for y_1^N. The original orbit, on the other hand, is also a shadowing orbit for y_1^N. For hyperbolic systems one can conclude from this that z_1 lies on the stable manifold of x_1 if N tends to infinity [6] and therefore $\lim_{n\to\infty}(x_n - z_n) = 0$, i.e. both orbits converge to the same asymptotic evolution.

In the following we shall briefly describe the *LSS (Least–Squares–Shadowing)* method that has been proved to converge locally [12]. The main steps of this algorithm are the following. Suppose there is an estimate z_1^N. Now linearize the nonlinear function (25) at z_1^N. The remaining problem is a quadratic minimisation problem with linear constraints. This problem is solved to obtain a new estimate $\bar{z}_1^N$. Then restart the process using this new estimate.

Suppose we have an estimate z_1^N for x_1^N (with $z_1^N = y_1^N$ in the first step). Our aim is to improve this estimate obtaining $\bar{z}_1^N$. Therefore we linearize F at every $z_n \in z_1^N$ and abbreviate the resulting Jacobian matrices by

$$M_n = \mathrm{D}f(z_{n-d+1}, \ldots, z_n).$$

Defining $\theta_n := \bar{z}_n - z_n$ $(n = 1, ..., N)$ and $\theta_1^N := \{\theta_1, ..., \theta_N\}$ the minimization problem for the functions J and Φ then reads as:

$$\begin{aligned} \text{minimize} \qquad J(z_1^N, \bar{z}_1^N) \quad &= \quad \sum_{n=1}^{N} \theta_n^2 & (26) \\ \text{with constrains} \qquad 0 = \Phi(\bar{z}_1^N) \quad &\cong \quad \Phi(z_1^N) + \mathrm{D}\Phi(z_1^N) \cdot \theta_1^N & (27) \end{aligned}$$

[3] z_1^N is called an ε–*shadowing orbit* for y_1^N, if z_1^N is an orbit and $\|z_n - y_n\| < \varepsilon$ for all $n = 1, \ldots, N$.

where

$$\mathrm{D}\Phi(z_1^N) = \begin{pmatrix} -M_d & I & 0 & \dots & 0 \\ & \ddots & \ddots & & \\ 0 & \dots & 0 & -M_N & I \end{pmatrix}.$$

Note that in (27) there are N unknown scalars θ_n but only $N-d$ equations. We therefore want to select from the (at least d-dimensional) solution space the vector θ_1^N which minimizes condition (26). The constrained minimisation problem (26)-(27) can in principle be solved by computing a pseudo inverse using a singular value decomposition (SVD) [57]. However, for an orbit of N points in d dimensions the SVD of a $N-d \times N$ matrix is needed. So even only for a few hundred points, the amount of storage and computational time needed is far too large. Therefore we will employ in the following an alternative way for computing the pseudo inverse which is then used to solve the set of linear equations obtained from Eq. (27). Using the pseudo inverse $\mathrm{D}\Phi(z_1^N)^\#$ the solution can be written as:

$$\theta_1^N = -\mathrm{D}\Phi(z_1^N)^\# \cdot \Phi(z_1^N). \tag{28}$$

Since the $N-d \times N$ matrix $\mathrm{D}\Phi(z_1^N)$ is surjective, its pseudo inverse can be expressed as

$$\mathrm{D}\Phi(z_1^N)^\# = \mathrm{D}\Phi(z_1^N)^{\mathrm{tr}} \cdot \left[\mathrm{D}\Phi(z_1^N) \cdot \mathrm{D}\Phi(z_1^N)^{\mathrm{tr}}\right]^{-1}. \tag{29}$$

The matrix inversion on the r.h.s. of Eq. (29) can be performed by a multistep iteration scheme [67]. More precisely, to compute (28) we solve equation

$$\Phi(z_1^N) = \left[\mathrm{D}\Phi(z_1^N) \cdot \mathrm{D}\Phi(z_1^N)^{\mathrm{tr}}\right] \cdot \eta_1^{N-d} \tag{30}$$

for η_1^{N-d} and then set

$$\theta_1^N = -\mathrm{D}\Phi(z_1^N)^{\mathrm{tr}} \cdot \eta_1^{N-d}. \tag{31}$$

For approximately solving Eq. (30) the Jacobi iteration scheme [67] is used with zero initial condition [12]. In our simulations we observed that a single step of the Jacobi algorithm is sufficient to achieve convergence of LSS. Since the iteration starts at the origin only the inversion of the diagonal blocks of

$$\left[\mathrm{D}\Phi(z_1^N) \cdot \mathrm{D}\Phi(z_1^N)^{\mathrm{tr}}\right] =$$

$$\begin{pmatrix} I + M_d M_d^{\mathrm{tr}} & -M_{d+1}^{\mathrm{tr}} & 0 & \dots & 0 \\ M_2 & I + M_{d+1} M_{d+1}^{\mathrm{tr}} & -M_{d+2}^{\mathrm{tr}} & \dots & \vdots \\ 0 & \vdots & \vdots & \vdots & 0 \\ \vdots & \dots & M_{N-1} & I + M_{N-1} M_{N-1}^{\mathrm{tr}} & -M_N^{\mathrm{tr}} \\ 0 & \dots & 0 & M_N & I + M_N M_N^{\mathrm{tr}} \end{pmatrix}$$

is required. Denoting by $\Delta(z_1^N)$ the matrix consisting of these diagonal blocks

$$\Delta(z_1^N) = \begin{pmatrix} I + M_d {M_d}^{\mathrm{tr}} & & 0 \\ & \ddots & \\ 0 & & I + M_N {M_N}^{\mathrm{tr}} \end{pmatrix} \tag{32}$$

we can write for (31) in the first step of the Jacobi iteration:

$$\theta_1^N \cong -\mathrm{D}\Phi(z_1^N)^{\mathrm{tr}} \cdot \Delta(z_1^N)^{-1} \cdot \Phi(z_1^N) \tag{33}$$

where the inverse $\Delta(z_1^N)^{-1}$ is given by the inverse block matrices $\left[I + M_n {M_n}^{\mathrm{tr}}\right]^{-1}$. In this way we obtain some explicite equations for θ_1^N. We re–insert the definition of $M_n = \mathrm{D}f(Z_n)$ and of $Z_n := [z_n, \ldots, z_{n+d-1}]$

$$\begin{array}{rcl} \theta_1 & = & -\,\mathrm{D}f^{\mathrm{tr}}(Z_1)\left[I + \mathrm{D}f(Z_1)\,\mathrm{D}f^{\mathrm{tr}}(Z_1)\right]^{-1}(f(Z_1) - Z_2)\,, \\ \theta_n & = & \left[I + \mathrm{D}f(Z_{n-1})\,\mathrm{D}f^{\mathrm{tr}}(Z_{n-1})\right]^{-1}(f(Z_{n-1}) - Z_n) \\ & & -\,\mathrm{D}f^{\mathrm{tr}}(Z_n)\left[I + \mathrm{D}f(Z_n)\,\mathrm{D}f^{\mathrm{tr}}(Z_n)\right]^{-1}(f(Z_n) - Z_{n+1})\,, \\ \theta_N & = & \left[I + \mathrm{D}f(Z_{N-1})\,\mathrm{D}f^{\mathrm{tr}}(Z_{N-1})\right]^{-1}(f(Z_{N-1}) - Z_N)\,. \end{array} \tag{34}$$

With θ_1^N the estimates $z_1, ..., z_N$ of the desired states can be updated

$$\bar{z}_n = z_n + \delta \cdot \theta_n \tag{35}$$

where δ is a damping constant that can be used to enforce convergence of the LSS iteration scheme.

We finish this section with three remarks:

1. We first point out that the block matrices $I + \mathrm{D}f\mathrm{D}f^{\mathrm{tr}}$ are always invertible since $\mathrm{D}f\mathrm{D}f^{\mathrm{tr}}$ a positive *semi*–definite matrix and therefore cannot have the eigenvalue -1.

2. It can be shown that if the system is not observable there are components that effectively do not enter the output and therefore cannot be recovered using output data. Furthermore, if the system is observable the problem of calculating Ψ^{-1} remains. Since time is not a constraint in this approach it can be done using a numerical scheme. How this can be incorporated into LSS requires further investigation.

3. Note, however, that LSS in general will *not* find the optimal solution, i.e. the global minimum of (24). LSS solves a consecutive series of linear minimisation problems that in general does not converge to the global minimum of the nonlinear problem. This remark, by the way, applies also to the methods considered in [29]. What LSS in fact does is to reduce the dynamic error $z_{n+1} - F(z_n)$ in an orbit by a series of small steps. Although LSS does not find the optimal solution, it will converge to an *exact* orbit at least locally [12].

4.2 Other Approaches

To solve either the constrained or the unconstrained minimisation principle (in the absence or presence of noise resp.) a lot of different approaches have been suggested. We shall briefly mention some of the original works as well as some articles providing a survey of methods.

Unconstrained minimisation was considered in [47, 48, 46] more as an ad hoc scheme rather than a probability theory based approach. To solve the minimisation problem, mostly gradient descend was employed [20]. The interpretation of the unconstrained minimisation as a maximum likelihood principle was, to our knowledge, pointed out first by [22]. Hammel [36] discusses noise reduction as a shadowing problem.

The constrained minimisation also was investigated by many authors. An approach using the method of Lagrange multipliers was suggested by [29]. In their original paper, the manifold decomposition technique was suggested to solve the resulting nonlinear Lagrange equations, which is of theoretical interest but seems to be quite hard to implement. In [68] it was suggested to solve the Lagrangian equations by Newtons method. Again, however, the techniques does not seem to work offhand. At least we were not able to reproduce their results.

If the state space of the system has only finitely many points (or can suitably be approximated by such a system), noise reduction corresponds to finding an optimal path in a decision tree. This tree grows of course exponentially with the number of orbit points, but presumably there are branches having small probability compared to the other branches and therefore can be discarded. This method was carried out in [23, 50].

So far only the maximum likelihood approach, which effectively results in an optimisation problem, was adressed. One may ask whether other approaches like calculating the conditional expectation yield convenient noise reduction methods. The problem here turns out to be the enormous complexity of the conditional pdf Π_N. To calculate the expectation by direct quadrature is merely impossible in practically all cases. In [22], however, an interesting Monte Carlo approach to this problem was suggested. The difficulty is of course to sample from the very complex conditional pdf. This is a general problem often encountered also in other fields where Monte Carlo methods are applied, for example thermodynamics, and therefore sophisticated methods have been conceived to overcome this problem, namely the Markov–Chain–Monte–Carlo or Metropolis–Hastings method. For a general introduction into these methods see e.g. [38]. To our knowledge, noise reduction using this techniques has been investigated only by [22], and we think that this approach merits further research.

4.3 Approaches for Unknown Dynamics

So far we assumed the measurement function G and the dynamical system F (or equivalently the transition pdf) to be known. Of course, the presented approaches cannot be applied directly if the underlying equations are not known. Therefore, in this section we shall briefly discuss different approaches for orbit estimation with limited pre–knowledge.

All methods assuming no or little knowledge of the signal process need a certain idea of what kind the signal process should be. Basically it is implicitely or explicitely assumed that the underlying signal belongs to a certain class $\mathcal{S}$. Then the method finds the member of $\mathcal{S}$ that is "at most in accordance" with the given data with respect to a certain measure of accordance.

If a model for the data is available (derived from first principles, for example) and only some parameters are unknown then the filtering task can be incorporated into the framework of this paper. The idea is to specify the model up to an unknown parameter vector α and to treat this α as a further state vector to be reconstructed. So suppose the dynamical system has the following form:

$$X_{n+1} = F(X_n, \alpha) + R_{n+1}.$$

Now introducing a further state equation of the form

$$\alpha_{n+1} = \alpha_n,$$

the unknown parameter can be considered as a part of the unknown state and is recovered by the usual filtering process.

If no dynamical equations are known for the data, some modeling has to be done in combination with or previous to noise reduction. This modeling is usually done assuming the dynamics to be esssentially deterministic. For deterministic dynamical systems state space reconstruction from scalar time series is possible using delay coordinates [55, 72, 58, 59]. The resulting states $Z_n = (y_n, y_{n-1}, ...)$ consist of sequences of measurement values and provide a faithful representation of the underlying dynamics if the dimension d is large enough.

Having successfully embedded the underlying dynamics in a reconstructed state space one may approximate the induced flow using *globally* or *locally* defined models. Global modelling using a superposition of radial basis functions has, for example, be used in Ref. [40] for subsequent noise reduction. Most noise reduction methods based on delay reconstruction, however, are based on *local* approximations of the flow in reconstruction space [64, 60, 27]. or of some (sub–)manifold containing the reconstructed states [17, 16], because local approximations are very efficient and flexible (and suffer not from a possible poor choice of basis functions). Well written reviews comparing different implementations of the basic idea may be found in Refs. [46, 45, 31]. Such methods have been applied succesfully for denoising speech signals [32, 33] or EEG data [34] and in studies on chaos based communication schemes [62]. However, they have to applied very carefully to avoid artefacts and misinterpretations [53]. When applied iteratively (what is typically done) most of the algorithms first improve the SNR but start to destroy/scramble the data completely when iterations are continued.

Finally we want to mention some recent approaches for noise reduction exploiting the existence of (unstable) periodic orbits in chaotic systems [75] and the availability of some simultaneously measured reference data set [71].

4.4 Linear Prefiltering

With all noise reduction methods introduced above one may have the idea to apply a linear filter first (for example, to remove some high-frequency noise) and then some of the methods to filter out the signal of interest.

The influence of FIR and IIR filters on attractor reconstruction was investigated by several authors [51, 21, 56, 18, 59, 58, 15]

It turned out that almost all FIR-filter do not destroy the faithful reconstruction of the dynamics via delay reconstruction of the filtered signal. However, with chaotic signals strong filtering may lead to additional folding of the reconstructed attractor due to the underlying stretch and fold mechanism. This makes any subsequent modeling and nonlinear noise reduction more difficult and one should check whether it is for the given data and task perhaps better to abandon linear FIR filtering at all. With IIR filters the situation is even more involved because these filters may even destroy important features of the embedding when applied prior to delay reconstruction. If the noise comes from a deterministic (chaotic) source it may be canceled using some nonlinear inverses of linear filters which are constructed using state space reconstruction [14]. Discrete valued signals can be separated from chaotic time series using some feedback stabilization method [13].

5 Message transmission

Many important applications of signal processing concern the problems occuring in message transmission. A schematic communication device is plotted in Fig.

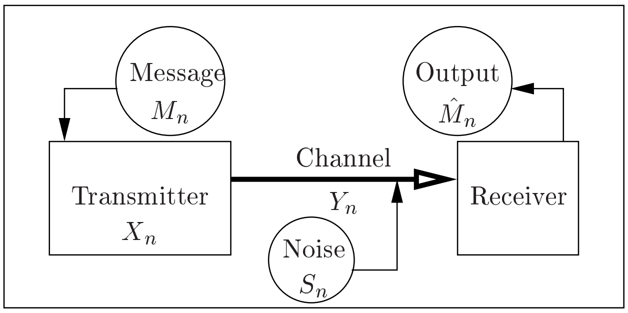

Figure 4: Schematic description of an electronic communication device. The message is put into the transmitter which produces a signal capable of passing the channel over long distance. In the channel, the signal is corrupted by noise. The receiver produces an estimate of the message.

4. We first rederive the mathematical description given in example (13). The *message* M_n is assumed to be a sequence of iid random variables assuming the values 0 or 1 with probability 0.5 each. Furthermore, the state of the

transmitter at time instant n is described by a d–dimensional vector X_n, and X_{n+1} is assumed to be fully described as a function of X_n and the incoming message M_{n+1}, i.e.

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

The state Y_n of the channel (as seen by the receicer) is assumed to be a function of the state of the transmitter plus corrupting noise, i.e.

$$Y_n = G(X_n) + S_n$$

as usual. One may assume even more complex models for the channel including feedback or even channel echoes causing a delay dynamics. Usual feedback can be incorporated into the transmitter dynamics, delay systems however are not considered here.

The *receiver* is any device that produces a reasonable estimate $\hat{M}_n$ for the actual message M_n using only what has been transmitted through the channel up to time n. This is, $\hat{M}_n$ shall be an estimator for M_n based on the time series $Y_1, \ldots, Y_n$. We have seen that this problem can be solved if the conditional probability

$$\mu_n(m) := P(M_n = m | Y_1 \ldots Y_n)$$

is known. This function tells us the probability of M_n being equal to m. Although M_n is zero or one each with probability 0.5, we of course expect $\mu_n(m)$ much more close to either 1 or 0 depending on wether the measurements $Y_1 \ldots Y_n$ are in accordance with the event $M_n = m$ or not, respectively.

The performance of a binary communication channel is usually measured by the Bit Error Rate (BER) which is defined as

$$\mathrm{BER} = \langle |M - \hat{M}| \rangle = \frac{1}{N} \sum_{k=1}^{N} |M_k - \hat{M}_k|$$

where M is the transmitted and $\hat{M}$ is the received message. If the process $\{M_k, \hat{M}_k\}$ is ergodic, we have that for $N \to \infty$

$$\mathrm{BER} = E|M_k - \hat{M}_k|$$

It is therefore reasonable to seek an estimator that minimizes $E|M_k - \hat{M}_k|$. It turns out that the mae does the job. To see this, note that for a constant m beeing either 0 or 1,

$$\begin{aligned} & E(|M_n - m| \mid Y_1 \ldots Y_n) \\ &= P(M_n \neq m \mid Y_1 \ldots Y_n) \\ &= 1 - P(M_n = m \mid Y_1 \ldots Y_n) \\ &= 1 - \mu_n(m) \end{aligned}$$

If we assume the estimator $\hat{M}_n$ to be $\{Y_1 \ldots Y_n\}$–measurable, we can replace m by $\hat{M}_n$ in the preceeding equations, which yields

$$E(|M_n - \hat{M}_n| \mid Y_1 \ldots Y_n) = 1 - \mu_n(\hat{M})$$

Taking $\hat{M}_n$ to be the minimizer of $\mu_n(m)$ we hence minimize $E(|M_n - m| \mid Y_1 \dots Y_n)$ over all such estimators. But by an elementary property of the conditional expectation,

$$E(|M_n - \hat{M}_n|) = E(E(|M_n - \hat{M}_n| \mid Y_1 \dots Y_n)),$$

whence we see that we also minimize the expected bit error which is (in the ergodic case) equal to the BER.

μ_n can easily be calculated if π_n is available. We will show this now. First recall that M_n is a discrete variable, so probabilities and pdfs are the same. Furthermore, the notation $\mu_n(m)$ supresses the fact that it is a function of $Y_1 \dots Y_n$ as well. We have according to Bayes rule

$$\begin{aligned}\mu_n(m) = c \quad \cdot \quad & p_{Y_n \mid M_n, Y_1 \dots Y_{n-1}}(y_n; m, y_1 \dots y_{n-1}) \\ \cdot \quad & p_{M_n \mid Y_1 \dots Y_{n-1}}(m; y_1 \dots y_{n-1})\end{aligned} \tag{36}$$

The second pdf is equal to p_{M_n}, since M_n and $Y_1 \dots Y_{n-1}$ are independent. Furthermore $p_{M_n} = 0.5$. The first term turns out to be

$$\begin{aligned}& p_{Y_n \mid M_n, Y_1 \dots Y_{n-1}}(y_n; m, y_1 \dots y_{n-1}) \\ &= c \cdot \textstyle\int g(y_n - G \circ F(x, m)) \pi_{n-1}(x) \mathrm{d}x\end{aligned}$$

All constants are normalisation constants. So defining the function

$$\bar{\mu}_{n-1}(y, m) := \int g(y - G \circ F(x, m)) \pi_{n-1}(x) \mathrm{d}x$$

we have that

$$\mu_n(m) = \frac{\bar{\mu}_{n-1}(Y_n, m)}{\bar{\mu}_{n-1}(Y_n, 0) + \bar{\mu}_{n-1}(Y_n, 1)}.$$

To tackle a message transmission problem numerically one has to calculate the functions $\bar{\mu}_n$ replacing π_n by any of the approximations $\tilde{\pi}_n$ introduced in the preceeding sections. Figure 5 shows the Bit Error Rate that is achievable with this approach for a chaotic signal contaminated by additive gaussian noise.

The model was as follows: We used

$$\begin{aligned} F(x, m) \quad &:= \quad \exp\left[-\left(\frac{x - 0.5 - A \cdot m}{\omega}\right)^2\right] \\ A \quad &= \quad 0.25 \\ \omega \quad &= \quad 0.3 \end{aligned}$$

For G we used the identity map, i.e.

$$Y_n = X_n + S_n.$$

As an exponential family we used truncated gaussian densities, i.e.

$$p(x; \theta_1, \theta_2) := \exp\left[\theta_1 x + \theta_2 x^2 - \psi(\theta)\right], \qquad x \in [0, 1].$$

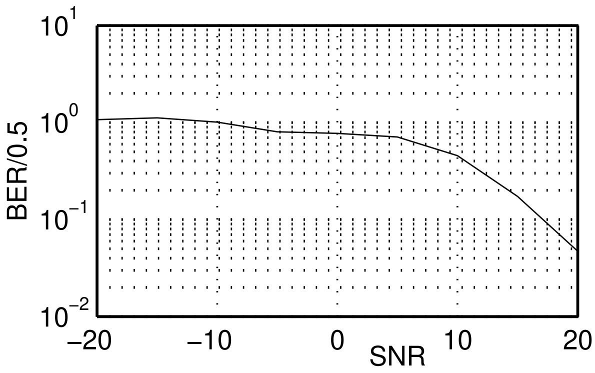

Figure 5: Bit error rate vs. SNR.

In principle the potential could be expressed using the errorfunction, which is provided by custom mathematical software packages. However, for large arguments, the errorfunction is close to one. Since ψ is basically the logarithm of the difference of two error functions, spurious cancellation may take place, leading to wrong results. We therefore employed a direct power series expansion explained in the appendix.

6 Conclusion

This paper adresses the problem of estimating the state of dynamical systems from uncertain information. We restricted ourselves to dynamical systems that are Markov processes, for this case is probably the most studied so far and also proves applicable to recently proposed communication schemes. We pointed out that the term noise reduction commonly denotes a couple of different tasks. Precise definitions of these tasks are formulated using a probabilistic framework. In the main part of the paper we focussed mainly on two of them, namely state and orbit estimation.

We presented a fundamental approach to these problems using a probabilistic formalism providing a unifying framework for conventional and new algorithms. Furthermore, a couple of common ad hoc procedures could be identified as certain approximative solutions. The presented selection is by no means exhaustive. Our goal was to convince the reader that new effective tools are available for many practically interesting cases, outperforming standard methods (e.g. Extended Kalman Filter) in many numerical testing cases.

We have mentioned also a variety of other problems related to nonlinear noise reduction and showed how minor modifications of the theory yield useful approaches for various cases, e.g. parameter estimation (identification of dynamical systems), message transmission problems and even solutions for unknown dynamics.

In this paper, the performance of the proposed algorithms was illustrated by means of (obviously not exhaustive) numerical simulations. A more rigorous analysis of the performance of filtering algorithms however requires some much more elaborated mathematical tools. The space available in this paper as well as its intention as a collection of applicable recipes did not allow for a detailed representation of available results on performance analysis.

Furthermore, supposing a filtering algorithm yields unsatisfactory results, we do not know whether this is due to an insufficient algorithm or a principal intrinsic indeterminism of the problem. Since the problem we are dealing with is stochastic, no algorithm will ever perform without any error. So an important goal for future research is to obtain an estimate of the maximal achievable accuracy for the given dynamical system. Considering the few and specialized results in this field we have the impression that sufficiently general and convenient results are quite difficult to obtain.

7 Acknowledgments

* The authors achnowledge support by the NATO Linkage Collaboration Grant PST.CLG 974981. For Maciej Ogorzalek this paper has been prepared during his stay at the National Microelectronics Center in Sevilla thanks to the "Ano sabatico" Award received from the Ministry of Education and Culture of Spain. Jochen Bröcker gratefully acknowledges support by the DFG, Graduiertenkolleg "Strömungsinstabilitäten und Turbulenz".

Appendix

The Kullback–Leibler distance for exponential families

Let $p(x, \theta)$ be the parametrisation of an exponential family with canonical statistics $c_i(x)$, $i = 1 \dots k$ and carrier measure λ. Let q be an arbitrary pdf and define the quantities

$$\mu_i := \int c_i(x) \cdot q(x) \cdot \mathrm{d}x, \qquad i = 1 \dots k.$$

Then for the Kullback–Leibler distance $d(p, q)$ we have

$$d(p, q) = \int \log(\frac{q}{\lambda}) q \mathrm{d}x - \left(\sum_{i=1}^{k} \theta_i \mu_i - \psi(\theta) \right).$$

So the minimisation of the Kullback–Leibler distance is equivalent to

$$\psi^*(\mu) := \max_\theta [\sum_{i=1}^{k} \theta_i \mu_i - \psi(\theta)] \tag{37}$$

which is a Legendre transform of ψ. This appendix will be concerned with numerical methods to compute ψ as well as the Legendre transform, or more exactly the minimizing argument θ^*.

Suppose first that convenient expressions for ψ as well as its first and second derivatives, i.e. the expectation parameters η and their jacobian $\frac{\partial\eta}{\partial\theta}$ respectively, are available. To solve the maximisation problem (37), consider a scheme of the form

$$\theta^{(n+1)} = \theta^{(n)} + \delta_n \cdot \Delta_n$$

where Δ_n is the Newton–direction

$$\Delta_n := \frac{\partial\eta}{\partial\theta}^{-1}[\mu - \eta(\theta^{(n)})]$$

and δ_n is a damping factor taken from the one dimensional maximisation problem

$$\delta_n := \arg\max_{\delta}\left[(\theta^{(n)} + \delta\Delta_n)\cdot\mu - \psi(\theta^{(n)} + \delta\Delta_n)\right].$$

So δ is chosen to maximize the problem not globally but along the Newton–direction. This scheme can be proved to converge globally. Locally the convergence is even quadratic, i.e. the number of valid digits doubles at every iteration step. There are further modifications simplifying both the computation of δ as well as the Newton–direction, which may be both very costly.

Remark that the Newton algorithm in principle *cannot* fail. However, the problem is that all quantities needed for the Newton scheme are given only approximately, and we observed that computing ψ and its derivatives far from the usual range with necessary accuracy is not easy. Therefore, if the Newton scheme does not converge, the reason is often failure of the routines computing the quantities entering Newtons scheme. We will now turn to the problem of computing them for a certain class of families called *convenient*.

Design of exponential families

The exponential family we want to use for a state estimation problem is not given in general but has to be designed. Usually it is required that certain functions $c_i(x), i = 1 \ldots k$ are among the canonical statistics. Is it possible to extend them to canonical statistics of an exponential family? Suppose the following growth condition

$$\|c(x)\| \leq K(1 + |x|^s)$$

is fulfilled for a certain $s \geq 0$ and $K > 0$. Then setting

$$\xi(x) := \sum_{i=1}^{d} |x_i|^r$$

with $r > s$ we have that

$$p(x,\theta) = \exp(\sum_{i=1}^{k} \theta_i c_i(x) - \theta_{k+1}\xi(x) - \psi(\theta))$$

is an integrable exponential family with parameter space

$$\Theta := \{(\theta_1 \dots \theta_{k+1}) \in \mathbb{R}^{k+1}, \theta_{k+1} > 0\}$$

and $\lambda(x) = 1$. We can also set θ_{k+1} to a constant positive value a and define an exponential family

$$p(x,\theta) = \lambda(x) \exp(\sum_{i=1}^{k} \theta_i c_i(x) - \psi(\theta))$$

with $\lambda(x) = \exp(-a\xi(x))$. We will call both exponential families *convenient* with respect to the $c_i(x)$. Of course, our growth condition and function ξ is only one possible choice to get an exponential family with a prescribed set of canonical statistics. However, our ξ has the advantage to factorize.

The potential function for convenient exponential families

We now proceed to derive a power series expansion for the function $\varphi := \exp(\psi)$ for convenient exponential families. The notation φ will be adopted only in this appendix and is not to be confused with the transition pdf. The coefficients in the power series expansion of φ are closely related to higher moments of $p(x,\theta)$. For calculations involving higher order moments, the concept of multiindices is useful. A k–dimensional multiindex $\alpha = (\alpha_1, \dots, \alpha_k)$ is an element of $\mathbb{N}^k$. For a multiindex α use the following notations

$$\begin{aligned} \alpha! &:= \alpha_1! \cdots \alpha_k! \\ |\alpha| &:= \alpha_1 + \ldots + \alpha_k \\ x^\alpha &:= x_1^{\alpha_1} \cdots x_k^{\alpha_k} \\ \frac{\partial^{|\alpha|}}{\partial x^\alpha} &:= \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots x_k^{\alpha_k}} \end{aligned}$$

Now we let

$$\eta_\alpha(\theta) = \int c_\alpha(x)\, p(x,\theta)\, \mathrm{d}x = \int c_{\alpha_1}(x) \cdots c_{\alpha_k}(x)\, p(x,\theta)\, \mathrm{d}x.$$

Defining the point $\bar{\theta} := [0 \dots 0, \theta_{k+1}]$ we have the following straight forward identity

$$\begin{aligned} \varphi(\theta) &= \int \exp\left(\sum_{i=1}^{k} \theta_i c_i(x) - \theta_{k+1}\xi(x)\right) \mathrm{d}x \\ &= \int \left(\sum_{\alpha \in \mathbb{N}^k} \frac{1}{\alpha!} (\theta_1 \dots \theta_k)^\alpha c_\alpha(x)\right) \exp(-\theta_{k+1}\xi(x))\, \mathrm{d}x \\ &= \sum_{\alpha \in \mathbb{N}^k} \frac{1}{\alpha!} \eta_\alpha(\bar{\theta}) (\theta_1 \dots \theta_k)^\alpha \end{aligned} \tag{38}$$

where $\eta_0 := \varphi$. This reduces the problem of computing ψ at an arbitrary point θ to the computation of ψ as well as higher order moments at a certain fixed point $\bar{\theta} = [0 \ldots 0, \theta_{k+1}]$. For these quantities we have

$$\begin{aligned}
\eta_\alpha(\bar{\theta}) &= \int c_\alpha(x) \cdot \exp(-\theta_{k+1}\xi(x) - \psi(\bar{\theta})) \cdot \mathrm{d}x \\
\varphi(\bar{\theta}) &= \int \exp(-\theta_{k+1}\xi(x)) \cdot \mathrm{d}x.
\end{aligned}$$

Now ξ is a homogenous function of degree r. Substituting

$$x = \theta_{k+1}^{-1/r} z, \qquad \mathrm{d}x = \theta_{k+1}^{-d/r} \mathrm{d}z$$

we get

$$\varphi(\bar{\theta}) = \theta_{k+1}^{-d/r} \int \exp(-\xi(x))\, \mathrm{d}x$$

The integral is a constant which may be computed offline. So far we used only the homogenity of ξ. Now for our convenient exponential families we have

$$\int \exp(-\xi(x))\, \mathrm{d}x = \left(\int \exp(-|t|^r)\, \mathrm{d}t\right)^d = \left(\frac{2\Gamma(\frac{1}{r})}{r}\right)^d$$

which finally yields

$$\psi(\bar{\theta}) = d \cdot \left(\log(\frac{2\Gamma(\frac{1}{r})}{r}) - \frac{\log(\theta_{k+1})}{r}\right).$$

To calculate the $\eta_\alpha(\bar{\theta})$ it can be very helpful to exploit symmetries and invariance properties of the c_i's. For example, the c_i's may be functions that factorize into functions depending on one coordinate only. Then all c_α factorize as well, and finally by choice of ξ, all η_α factorize into integrals over only one coordinate. If for example the c_i's are monomials, these integrals can be expressed in closed form using again the Γ–function.

The problem of representing φ is now solved. In principle we could get the moments $\eta_\alpha(\theta)$ for general θ from computing higher formal derivatives of the power series for φ. i.e. use the general equation

$$\eta_\alpha(\theta) = \frac{1}{\varphi(\theta)} \frac{\partial^{|\alpha|}}{\partial \theta_\alpha} \varphi(\theta)$$

and apply it to Eq.(38). The approximation by taking a truncated power series expansion for φ and compute formal derivatives is however of less order than the approximation for φ itself. There are, however, other methods exploiting the fact that the moments cannot be completely independent in general.

We will investigate $\varphi = \exp(\psi)$ in the following for the special case that the canonical statistics are monomials and λ is the Lebesgue measure. To denote these monomials, let $A = \{\alpha^{(1)} \ldots \alpha^{(k)}\}$ be a set of d–dimensional multiindices, i.e. each $\alpha^{(i)}$ denotes a whole d–dimensional multiindex and $\alpha_j^{(i)}$ denotes the j'th entry so that the canonical statistics can be written as $c_i(x) = x^{\alpha^{(i)}}$.

$$\varphi(\theta) = \int \exp(\sum_{i=1}^{k} \theta_i x^{\alpha^{(i)}}) \mathrm{d}x,$$

It is readily seen that φ satisfies a set of partial differential equations in this case. The partial differential equations will be valid in the interior of the parameter space, so we can assume it to be open. Now substitute $z_j := \tau_j x_j$ for every $j = 1, \ldots, d$ and τ_j close to one. Then we have

$$\varphi(\theta) = \int \tau \exp(\sum_j \theta_j \tau^{\alpha^{(i)}} z^{\alpha^{(i)}}) \mathrm{d}z = \tau\varphi(\tau^{\alpha^{(1)}}\theta_1, \ldots, \tau^{\alpha^{(k)}}\theta_k),$$

where $\tau := \tau_1 \cdots \tau_d$. This property will be referred to as the *invariance property* of φ. Now take the gradient of both sides with respect to $(\tau_1, \ldots, \tau_d)$ and set them all equal to 1, which yields

$$0 = \varphi(\theta) + \sum_{i=1}^{k} \alpha_j^{(i)} \theta_i \frac{\partial\varphi}{\partial\theta_i}, \qquad j = 1, \ldots, d. \tag{39}$$

This is the desired set of equations. These equations can be extremely helpful for computing higher moments. Since $\eta_i = \varphi(\theta)^{-1} \frac{\partial\varphi}{\partial\theta_i}$ we get from the preceding equation

$$0 = 1 + \sum_{i=1}^{k} \alpha_j^{(i)} \theta_i \eta_i, \qquad j = 1, \ldots, d.$$

by dividing through $\varphi(\theta)$. This equation may be analyzed by linear tools for the solution space. To obtain similar relations for higher order moments, take the derivative $\frac{\partial^{|\beta|}}{\partial\theta_\beta}$ of eq. (39) and aqain divide by $\varphi(\theta)$. There are again further symmetries in this problem: Let β, γ be two k–dimensional multiindices. Now if

$$\beta_1\alpha^{(1)} + \ldots + \beta_k\alpha^{(k)} = \gamma_1\alpha^{(1)} + \ldots + \gamma_k\alpha^{(k)},$$

then $\eta_\beta = \eta_\gamma$. The problem of course becomes more and more involved and may require computational algebraic methods. It seems that symmetry plays an essential role here, and maybe employing group theory one can get more general results.

References

[1] H.D.I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer Verlag 1996

[2] Sun-Ichi Amari. *Differential Geometric Methods in Statistics*, volume 28 of *Lecture Notes in Statistics*. Springer–Verlag, Berlin, 1985.

[3] M. Aoki *Optimisation of Stochastic Systems*. vol. of *Mathematics in Science and Engineering*, Academic Press, 1972

[4] R. R. Bahadur. *Some Limit Theorems in Statistics*. SIAM, 1971.

[5] Ole E. Barndorff-Nielsen. *Information and exponential families*. John Wiley & Sons, New York, 1978.

[6] R. Bowen. *Equilibrium States and the Ergodic Theory of Anosov diffeomorphisms*. Lecture notes in mathematics 470, Springer–Verlag, 1975.

[7] Leo Breiman. *Probability*. Addison-Wesley-Publishing, 1973.

[8] Damiano Brigo. *Filtering by projecting on the manifold of exponential densities*. PhD thesis, Vrije Universiteit Amsterdam, 1996.

[9] Damiano Brigo, Bernard Hanzon, and Francois LeGland. A differential geometric approach to nonlinear filtering: the projection filter. Technical report, Institut de Recherche en Informatique et Systèmes Aléatoires, 1995.

[10] J. Bröcker and U. Parlitz, "Filtering using exponential probability densities," *Proc. Int. Symposium on Nonlinear Theory and Appl. NOLTA'99*, Hawaii, 1999, pp.573-576.

[11] J. Bröcker and U. Parlitz, "Noise reduction and filtering of chaotic time series," *Proc. Int. Symposium on Nonlinear Theory and Appl. NOLTA2000*, Dresden 2000, pp.381-384.

[12] J. Bröcker and U. Parlitz, "Efficient noncausal noise reduction for deterministic time series,", *Chaos*, vol.11, No.2, 2001, pp.319-326.

[13] D.S. Broomhead, J.P.Huke and R. Jones, "Signals in Chaos: A method for the cancellation of deterministic noise from discrete signals", *Physica D*, vol.D80, 1995, pp.413-432.

[14] D.S. Broomhead, J.P. Huke, M.A.S. Potts, "Cancelling deterministic noise by constructing nonlinear inverses to linear filters", *Physica D*, vol. 89, 1996, pp. 439-458.

[15] D.S. Broomhead, J. Huke and M. Muldoon, "Linear Filters and Nonlinear Signals", *J. Royal Statistical Society B 54*, No. 2, pp. 373-382, (1992).

[16] R. Cawley and G.-H. Hsu, "SNR performance of a noise reduction algorithm applied to coarsely sampled chaotic data," *Phys. Lett. A*, vol.166, 1992, pp.188-196.

[17] R. Cawley and G.-H. Hsu, "Local–geometric–projection method for noise reduction in chaotic maps and flows", *Phys. Rev. A*, vol. 46, no. 6, 1992, pp. 3057-3082

[18] A. Chennaoui, K. Pawelzik, W. Liebert, H.G. Schuster, G. Pfister, “Attractor reconstruction from filtered chaotic time series”, *Phys. Rev. A*. A vol. 41, no. 8, 1990, pp. 4151-4159

[19] C. K. Chui and G. Chen. *Kalman Filtering*, vol.17 of *Springer Series in Information Sciences.* Springer-Verlag, 1987.

[20] M. Davies, ”Noise reduction by gradient decent,” *Int. J. of Bif. and Chaos*, vol.3, 1992, pp.113-118.

[21] M.E. Davies, “Reconstructing attractors from filtered time series”, *Physica D*, vol. 101, 19997, pp. 195-206.

[22] M.E. Davies, ”Nonlinear noise reduction through Monte Carlo sampling”, *Chaos*, vol.8, No.4, 1998, pp.775-781.

[23] H. Dedieu and A. Kisel, “Communications with chaotic time series: probabilistic methods for noise reduction,” *Int. J. Circ. Theor. Appl.* vol. 27, 1999, pp. 577-587.

[24] H. Dedieu and M. Ogorzałek, ”Overview of Noise Reduction Algorithms for Systems with Known Dynamics,” *Proc. International Symposium on Nonlinear Theory and its Applications NOLTA*, Crans-Montana, Switzerland, 1999, pp. 1297-1300.

[25] Pierre Del Moral. Nonlinear filtering: monte-carlo particle resolution. Technical Report 02, Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 31062 Toulouse, 1996.

[26] Pierre Del Moral. A uniform convergence theorem for the numerical solving of the nonlinear filtering problemng. Technical Report 14, Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 31062 Toulouse, 1996.

[27] N. Enge, Th. Buzug and G. Pfister, ”Noise reduction on chaotic attractors,” *Phys. Lett. A*, vol.175, 1993, pp.178-186.

[28] J. R. Elliott, L. Aggoun and J. B. Moore, *Hidden Markov Models*, vol.29 of Appl. of Math., Springer Verlag 1995.

[29] J. D. Farmer and J. J. Sidorovich, ”Optimal shadowing and noise reduction,” *Physica D*, vol.47, 1991, pp. 373-392.

[30] M. Ferrante, ”On the existence of finite dimensional filters in discrete time,” *Stochastics and Stochastics Report*, vol.40, pp. 169-179.

[31] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath, and T. Schreiber, “On noise reduction methods for chaotic data,” *Chaos*, vol.3, 1993, pp.127-141.

[32] R. Hegger, H. Kantz, and L. Matassini,“Denoising Human Speech Signals Using Chaoslike Features”, *Phys. Rev. Lett.*, vol. 84, no. 14, 2000, pp. 3197-32000.

[33] R. Hegger, H. Kantz, and L. Matassini, “Noise Reduction for Human Speech Signals by Local Projections in Embedding Space”, *IEEE Trans. Circ. Syst. II*, accepted for publication.

[34] T. Schreiber and D.T. Kaplan, “Nonlinear noise reduction for electrocardiograms”, *Chaos* vol. 6, no. 1, 1996, pp. 87-92.

[35] M-S. Gupta, "Applications of electrical noise", *Proc. IEEE*, vol.63, no.7, pp.996-1010, 1975.

[36] S. M. Hammel, "A noise reduction method for chaotic systems," *Phys. Lett. A*, vol.148, No.8-9, 1990, pp.421-428.

[37] S. M. Hammel, *Noise Reduction for Chaotic Systems*, Naval Surface Warfare Center, Silver Spring, MD 1989.

[38] J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods*, Methuen and Co., London, 1964.

[39] S. Haykin, X.-B. Li, "Detection of Signals in Chaos," *Proc. IEEE*, Vol.83, No.1, 1995, pp.95-122.

[40] J. Holzfuss and J. Kadtke, “Global Nonlinear Noise Reduction Using Radial Basis Functions,” Int. J. of Bif. and Chaos 3(3), 589-596 (1993).

[41] Z. Jako, G. Kolumban and H. Dedieu, "On some recent developments of noise cleaning algorithms for chaotic signals," *IEEE Trans. Circuits and Systems*, Part I, vol.47, No.9, 2000, pp.1403-1406.

[42] A. H. Jazwinsky. *Stochastic Processes and Filtering Theory*, vol.64 of *Mathematics in Science and Engineering*, Academic Press, 1970

[43] S. Julier and J. Uhlmann *A General Method for Approximating Nonlinear Transformations of Probability Densities*, Tech. Rep. Department of engineering Science, University of Oxford, Academic Press, 1970

[44] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis.* Cambridge UP, Cambridge 1997

[45] H. Kantz, T. Schreiber, I. Hoffmann, T. Buzug, G. Pfister, L.G. Flepp, J. Simonet, R. Badii, E. Brun, "Nonlinear noise reduction: A case study on experimental data," *Phys. Rev. E*, vol.48, 1993, pp.1529-1538.

[46] E. J. Kostelich and T. Schreiber, "Noise reduction in chaotic time-series: A survey of common methods," *Phys. Rev. E*, vol.48, No.3, 1993, pp.1752-1763.

[47] E. J. Kostelich and J. A. Yorke, "Noise reduction in dynamical systems," *Physical Review A*, vol.38, 1988, pp.1649-1652.

[48] E.J. Kostelich and J.A. Yorke, “Noise reduction: Finding simplest dynamical system consistent with the data”, *Physica D*, vol.D41, 1990, pp.183-196.

[49] C. Lee, D. B. Williams, “Generalized Iterative Methods for Enhancing Contaminated Chaotic Systems”, *IEEE Trans. on Circuits and Systems-I*, Vol. 44, No. 6, June 1997.

[50] P.F. Marteau and H.D.I. Abarbanel, ”Noise reduction in chaotic time series using scaled probabilistic methods,” *J. Nonlin. Science*, Vol.1, 1991, pp.313-343.

[51] F. Mitschke, ”Acausal filters for chaotic signals,”, *Phys. Review A*, vol.41, No.2, 1990, pp.1169-1171.

[52] Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to the Theory of Statistics.* McGraw-Hill Series in Probability and Statistics. McGraw-Hill, 1974.

[53] A.I. Mees and K. Judd, “Dangers of geometric filtering”, *Physica* D vol. 68, 1993, pp. 427-436.

[54] E. Ott, T. Sauer, J.A. Yorke, *Coping with chaos. Analysis of Chaotic Data and the Exploitation of Chaotic Systems*, Wiley Series in Nonlinear Science, 1994.

[55] N.H. Packard et al., “Geometry from a Time Series”, *Phys. Rev. Lett.* vol. 45, no. 9, 1980, pp. 712-717

[56] P.Paoli, A.Politi, G. Broggi, M. Ravani, R. Badii, “Phase Transitions in Filtered Chaotic Signals”, *Phys. Rev. Lett.* vol.62, no. 21, 1989, pp. 2429-2432.

[57] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery. *Numerical Recipes in C.* Cambridge UP, Cambridge 1996.

[58] T. Sauer, J.A. Yorke, M. Casdagli, “Embedology”, *J. Stat. Phys.* vol. 65, nos. 3/4, 1991, pp. 579-616.

[59] T. Sauer, “How many delay coordinates do you need?” *Int. J. Bif. Chaos* vol. 3, no. 3, 1993, 737-744.

[60] T. Sauer, ”On the Development of Practical Nonlinear Filters,” *Information Sciences*, vol.7, pp.253-270, 1974.

[61] H. W. Sorenson, ”A noise reduction method for signals from nonlinear systems,” *Physica D*, vol.58, pp.193-201, 1992 .

[62] E. Rosa, S. Hayes and C. Grebogi, ”Noise filtering in communication with chaos”, *Phys. Review Letters*, vol.78, No.7, 1997, pp.1247-1250.

[63] W. J. Runggaldier and F. Spizzichino, "Sufficient conditions for finite dimensionality of filters in discrete time: a Laplace transform based approach" *Bernoulli* (to appear).

[64] T. Schreiber, "An extremely simple nonlinear noise reduction method," *Phys. Rev. E*, vol.47, 1993, pp.2401.

[65] T. Schreiber and H. Kantz, "Noise in chaotic data: Diagnosis and treatment," *Chaos*, vol.5, No.1, 1995, pp.133-142.

[66] T. Schreiber and D.T. Kaplan, "Nonlinear noise reduction for electrocardiograms," *Chaos*, vol.6, No.1, 1996, pp.87-92.

[67] J. Stoer and R. Burlisch. *Numerische Mathematik*, volume I, II. Springer–Verlag, 1990.

[68] T. Schimming, H. Dedieu, M. Hasler, M. Ogorzalek. "Noise Filtering in Chaos–Based Communication," in *Chaotic Electronics in Telecommunications.* M. P. Kennedy, R. Rovatti, G. Setti (Eds.) CRC Press LLC, Boca Raton, Fl. 2000.

[69] J. Schweizer, *Application of Chaos to Communication*, Ph.D. Thesis EPFL No. 1953, Lausanne, Switzerland, 1999.

[70] N. Sharma and E. Ott, "Synchronization-based noise reduction method for communication with chaotic systems," *Phys. Rev. E* 58(6), 8005-8008 (1998).

[71] K. Sternickel, A. Effern, K. Lehnertz, T. Schreiber and P. David, "Nonlinear noise reduction using reference data", *Phys. Review E*, vol.63, 2001.

[72] F. Takens, "Detecting strange attractors in turbulence", *Lecture Notes in Math.* vol, 898 Springer Verlag, 1981

[73] N. Tanaka, H. Okamoto, and M. Naito, "Estimating the amplitude of measurement noise present in chaotic time series," *Chaos* vol.9, no. 2, 1999, pp. 436-444.

[74] S. Thrun, D. Fox, W. Burgard and F. Dellaert, "Robust Monte Carlo Localisation for Mobile Robots," to appear in *Artificial Intelligence* Summmer 2001.

[75] D.M. Walker, S.P. Allie, A. Mees, "Exploiting the periodic structure of chaotic systems for noise reduction of nonlinear signals", *Phys. Lett. A* vol. 242, 1998, pp. 63-73.

[76] D.M. Walker, *Reconstruction and Noise Reduction of Nonlinear Dynamics using Nonlinear Filters*, Ph–th., University of Western Australia, 1998.