Journal of Mathematical Psychology 94 (2020) 102307

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Tutorial

The Hitchhiker's guide to nonlinear filtering

Anna Kutschireiter [a,b,*], Simone Carlo Surace [a,c], Jean-Pascal Pfister [a,c]

[a] Department of Physiology, University of Bern, Bühlplatz 5, 3012 Bern, Switzerland
[b] Department of Neurobiology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, United States
[c] Institute of Neuroinformatics, UZH / ETH Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 5 June 2019
Received in revised form 21 November 2019
Accepted 25 November 2019
Available online xxxx

Keywords:
Nonlinear filtering theory
Bayesian inference
Change of measure
Particle filters

ABSTRACT

Nonlinear filtering is used in online estimation of a dynamic hidden variable from incoming data and has vast applications in different fields, ranging from engineering, machine learning, economic science and natural sciences. We start our review of the theory on nonlinear filtering from the simplest 'filtering' task we can think of, namely static Bayesian inference. From there we continue our journey through discrete-time models, which are usually encountered in machine learning, and generalize to continuous-time filtering theory. The idea of changing the probability measure connects and elucidates several aspects of the theory, such as the parallels between the discrete- and continuous-time problems and between different observation models. Furthermore, it provides insight into the construction of particle filtering algorithms. This tutorial is targeted at scientists and engineers and should serve as an introduction to the main ideas of nonlinear filtering, and as a segway to more advanced and specialized literature.

Contents

* Corresponding author at: Department of Neurobiology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, United States.
E-mail address: anna_kutschireiter@hms.harvard.edu (A. Kutschireiter).

https://doi.org/10.1016/j.jmp.2019.102307

1. Introduction: A guide to the guide

"The introduction begins like this:
Space, it says, is big. Really big.
You just won't believe how vastly hugely mind-bogglingly big it is."

[Douglas Adams]

Filtering is the problem of estimating a dynamically changing state, which cannot be directly observed, from a stream of noisy incoming data. To give a concrete example, assume that you are a Vogon in charge of a spaceship. Since you had a particularly bad day, you decide to destroy a small asteroid to make yourself feel better. Before you push the red button, you need to know the current position of the asteroid, which corresponds to the hidden state X_t. You have some idea about the physics of movement in space, but there is also a stochastic component in the movement of your target. Overall, the asteroid's movement is described by a stochastic dynamical model. In addition, you cannot directly observe its position (because you like to keep your safe distance), so you have to rely on your own ship's noisy measurements Y_t of the position of the asteroid. Because of these uncertainties, it would not only be useful to have an estimate of the asteroid's current position X_t based on the history of measurements $Y_{0:t} = \{Y_0, Y_1, \dots, Y_t\}$, but also an estimate of the uncertainty of the estimate. Thus generally, the conditional probability density $p(X_t|Y_{0:t})$ is the complete solution to your problem (and the beginning of the problem of how to find this solution).

These sorts of problems are not only relevant for bad-tempered Vogons, but in fact are encountered in a wide variety of applications from different fields. Initial applications of filtering were centered mostly around engineering. After the seminal contributions to *linear* filtering problems by Kalman (1960) and Kalman and Bucy (1961), the theory was largely applied to satellite orbit determination, submarine and aircraft navigation as well as space flight (Jazwinski, 1970). Nowadays, applications of (nonlinear) filtering range from engineering, machine learning (Bishop, 2006), economic science (in particular mathematical finance, some examples are found in Brigo & Hanzon, 1998) and natural sciences such as geoscience (Van Leeuwen, 2010), in particular data assimilation problems for weather forecasting, neuroscience and psychology. As a particular example for its usefulness in neuroscience, the modeling of neuronal spike trains as point processes (Brillinger, 1988; Truccolo, 2004) has led to interesting filtering tasks, such as the problem of decoding a stimulus from the spiking activity of neurons (e.g. Koyama, Eden, Brown, & Kass, 2010; Macke, Buesing, & Sahani, 2011). In psychology, nonlinear filtering techniques are not only used for data analysis, but can also provide qualitative insight into psychological processes such as perception (Körding, Tenenbaum, & Shadmehr, 2007; Wolpert, Ghahramani, & Jordan, 1995) or decision making (Drugowitsch, DeAngelis, Klier, Angelaki, & Pouget, 2014; Glaze, Kable, & Gold, 2015; Piet, Hady, Brody, Hady, & Brody, 2018; Radillo, Veliz-Cuba, Josić, & Kilpatrick, 2017; Veliz-Cuba, Kilpatrick, & Josić, 2016). To tackle these kinds of questions, knowledge about nonlinear filtering is indispensable. Theoretical understanding can further help in connecting and unifying specific applications of filters and is paramount for understanding more advanced topics in filtering (Jazwinski, 1970, Section 1.2).

The aim of this tutorial is to present – in an easily accessible and intuitive way – the theoretical basis for continuous-time nonlinear filtering with diffusion and point-process observations. The tutorial highlights the change of measure as a powerful tool to derive the fundamental equations of nonlinear filtering as well as numerical approximations. In addition, the unification provided by the concept of change of measure provides a solid basis for diving into the huge body of literature on nonlinear filtering. Our tutorial complements the more advanced theoretical literature (e.g. Bain & Crisan, 2009; Bremaud, 1981; Jazwinski, 1970) or more specialized tutorials, e.g. on particle filtering (Arulampalam, Maskell, Gordon, & Clapp, 2002; Doucet & Johansen, 2009; Speekenbrink, 2016), Hidden Markov Models (Rabiner, 1989; Visser, 2011) or variational Bayes for latent linear systems (Ostwald, Kirilina, Starke, & Blankenburg, 2014).

2. A view from space: from Bayes' rule to filtering

"Even the most seasoned star tramp can't help but shiver
at the spectacular drama of a sunrise seen from space,
but a binary sunrise is one of the marvels of the Galaxy".

[Douglas Adams]

Suppose that we observe a random variable Y and want to infer the value of an (unobserved) random variable X. Bayes' rule tells us that the conditional distribution of X given Y, the so-called posterior, can be computed in terms of three ingredients: the prior distribution $p(X)$, the likelihood $p(Y|X)$, and the marginal likelihood $P(Y)$, which acts as a normalizing constant:

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}. \tag{1}$$

This tutorial is concerned with the application of the above idea to a situation where X, Y are continuous-time stochastic processes and we want to perform the inference online as new data from Y comes in. In this section, we want to gradually build up the stage: as some readers might be more familiar with discrete-time filtering due to its high practical relevance and prevalence, we will start our journey from there, picking up important recurring concepts as we make our way to continuous-time filtering.

2.1. Changes of measure

Before we talk about dynamic models, let us briefly highlight a concept in Bayesian inference that will be very important in the sequel: that of changing the probability measure. A probability

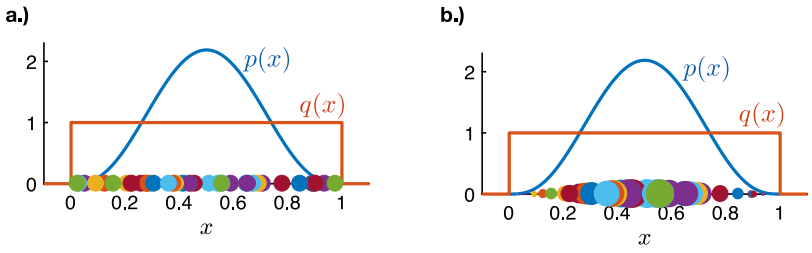

Fig. 1. Consider the problem of empirically approximating the beta distribution $p(x) = \text{Beta}(x; 4, 4)$ (blue) with samples from the uniform distribution between 0 and 1, $q(x) = \mathcal{U}(x; 0, 1)$ (red). (a) The density of those samples does not represent the Beta distribution, but (b) a combination of the density of samples together with their respective importance weights according to Eq. (8). Here, the size of a dot represents the weight of the respective sample. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

measure is a function that assigns numbers ('probabilities') to events. If we have two such measures $\mathbb{P}$ and $\mathbb{Q}$, then $\mathbb{P}$ is called *absolutely continuous* w.r.t. $\mathbb{Q}$ if every nullset of $\mathbb{Q}$ is a nullset of $\mathbb{P}$. Moreover, $\mathbb{P}$ and $\mathbb{Q}$ are called *equivalent* if they have the same nullsets. In other words, if A denotes an event, and $\mathbb{P}(A)$ denotes its probability, then equivalence means that $\mathbb{Q}(A) = 0$ if and only if $\mathbb{P}(A) = 0$.

But why would we want to change the measure in the first place? Changing the measure allows us to compute expectations of a measurable function $\phi(x)$ with respect to a measure $\mathbb{Q}$, which were originally expressed with respect to another measure $\mathbb{P}$. To see this, consider the two measures $\mathbb{P}$ and $\mathbb{Q}$ for some real-valued random variable X, and write them in terms of their densities p, q (with respect to the Lebesgue measure).[1] We then have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[\phi(X)] &= \int dx\, \phi(x)p(x) \\ &= \int dx\, \frac{p(x)}{q(x)}\phi(x)q(x) = \mathbb{E}_{\mathbb{Q}}[L(X)\phi(X)]\,, \end{aligned} \tag{2}$$

where we introduced the likelihood ratio $L(x) := \frac{p(x)}{q(x)}$ and $\mathbb{E}_{\mathbb{Q}}$ denotes expectation under the distribution q. Thus, changing the measure proves to be very useful whenever expectations under $\mathbb{Q}$ are easier to compute than under $\mathbb{P}$.

A fundamental problem in filtering is that of computing a conditional expectation (i.e. an expected value under the posterior distribution) of this sort:

$$\mathbb{E}_{\mathbb{P}}[\phi(X)|Y] = \int dx\, \phi(x)p(x|Y) \tag{3}$$

for some function ϕ, and we want to use Eq. (1) to compute $p(X|Y)$. We therefore have to compute the two integrals here

$$\mathbb{E}_{\mathbb{P}}[\phi(X)|Y] = \int dx\, \phi(x)\frac{p(Y|x)p(x)}{p(Y)} = \frac{\int dx\, \phi(x)p(Y|x)p(x)}{\int dx\, p(Y|x)p(x)}, \tag{4}$$

but the structure of the model (interactions between X and Y) might make it very hard to compute the integrals, either analytically or numerically. Thus, we again change the measure to a reference measure $\mathbb{Q}$ with joint density $q(x, y)$, and rewrite Eq. (4):

$$\mathbb{E}_{\mathbb{P}}[\phi(X)|Y] = \frac{\int dx\, \phi(x)\frac{p(x,Y)}{q(x,Y)}q(x,Y)}{\int dx\, \frac{p(x,Y)}{q(x,Y)}q(x,Y)} = \frac{\mathbb{E}_{\mathbb{Q}}[L(X,Y)\phi(X)|Y]}{\mathbb{E}_{\mathbb{Q}}[L(X,Y)|Y]}, \tag{5}$$

where now the likelihood ratio $L(x, y) = \frac{p(x,y)}{q(x,y)}$ is a function of both x and y.

The hope is that we can pick a reference measure $\mathbb{Q}$ such that both $L(x, y)$ and $q(x, y)$ are simple enough to make Eq. (5) more tractable than Eq. (3). For instance, some simplification might be achieved by switching from a model $p(x, y)$ of $\mathbb{P}$ in which X and Y are coupled, i.e. statistically dependent, to a model $p(x)q(y)$ of $\mathbb{Q}$ where they are independent (while preserving the distribution of X), i.e. under model $\mathbb{Q}$ we find $q(x, y) = p(x)q(y)$. A potential added advantage of changing the measure is when the distribution $q(y)$ is computationally simple. Then, the likelihood ratio $L(x, y)$ reads

$$L(x,y) = \frac{p(x,y)}{q(x,y)} = \frac{p(y|x)p(x)}{p(x)q(y)} = \frac{p(y|x)}{q(y)}, \tag{6}$$

and conditional expectations under $\mathbb{Q}$ can simply be taken with respect to the prior probability $p(x)$.

Please take a moment to appreciate the value of this idea: the change of measure has allowed us to replace the expectation with respect to the posterior $p(x|y)$ of $\mathbb{P}$ (which might be hard to get) with an expectation with respect to the prior $p(x)$ of $\mathbb{Q}$ (which might be easy to compute). This 'trick' will become the central theme of this manuscript.

2.1.1. Importance sampling

A numerical example where a change of measure is directly used is **importance sampling**. Here, the goal is to approximate expectations with respect to a distribution $p(x)$ (under $\mathbb{P}$) by using M empirical samples $X^{(i)} \sim p(x)$, such that

$$\mathbb{E}_{\mathbb{P}}[\phi(X)] \approx \frac{1}{M}\sum_{i=1}^{M}\phi(X^{(i)}). \tag{7}$$

However, there might be situations where we cannot draw samples from $p(x)$, but only from another distribution $q(x)$ (under $\mathbb{Q}$). Thus, we first perform a change of measure to $\mathbb{Q}$, and then use the samples $X^{(i)} \sim q(x)$ to approximate the expectation:

$$\mathbb{E}_{\mathbb{P}}[\phi(X)] = \mathbb{E}_{\mathbb{Q}}[L(X)\,\phi(X)] \approx \frac{1}{M}\sum_{i=1}^{M} L(X^{(i)})\,\phi(X^{(i)}). \tag{8}$$

In this context, the likelihood ratio $L(X^{(i)}) = \frac{p(X^{(i)})}{q(X^{(i)})} =: w^{(i)}$ is referred to as (unnormalized) *importance weight*. Hence, the target distribution $p(x)$ is jointly represented by the *density* of empirical samples (or 'particles'), i.e. how many samples can be found in a specific interval in the state space, and by their respective importance weights (see simple example in Fig. 1).

Similarly, we can use importance sampling to approximate a posterior expectation $\mathbb{E}_{\mathbb{P}}[\phi(X)|Y]$ with samples from the prior distribution $p(x)$. For this, consider changing to a measure $\mathbb{Q}$ with

[1] In this section, all 'densities' are with respect to the Lebesgue measure.

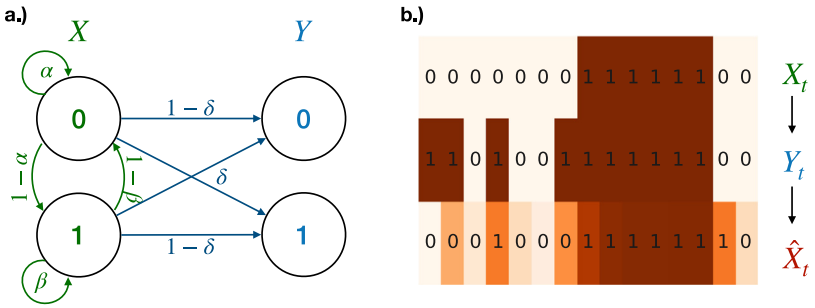

Fig. 2. (a) A two-state HMM with binary observation channel. α and β denote the probability to stay in state 0 and 1, respectively. The probability of making an error in the observation channel is given by δ. (b) Sample state trajectory, sample observation and filtered density $\hat{p}_{t_n}$ (color intensity codes for the probability to be in state 1), as well as estimated state trajectory $\hat{X}_{t_n}$ (where $\hat{X}_{t_n} = 1$ if $\hat{p}^{(2)}_{t_n} > 1/2$ and $\hat{X}_{t_n} = 0$ otherwise). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

density $q(x, y) = p(x)q(y)$, such that with the likelihood ratio in Eq. (6) we find

$$\mathbb{E}_{\mathbb{P}}[\phi(X)|Y] = \frac{\mathbb{E}_{\mathbb{Q}}[L(X,Y)\phi(X)|Y]}{\mathbb{E}_{\mathbb{Q}}[L(X,Y)|Y]} \approx \frac{1}{Z}\sum_{i=1}^{M} p(Y|X^{(i)})\phi(X^{(i)})$$
$$= \frac{1}{Z}\sum_{i=1}^{M} w^{(i)}\phi(X^{(i)}), \quad X^{(i)} \sim p(x) \tag{9}$$

where the unnormalized importance weights are given by $w^{(i)} = p(Y|X^{(i)})$, and the normalization Z is given by

$$Z = \sum_{i=1}^{M} p(Y|X^{(i)}) = \sum_{i=1}^{M} w^{(i)}. \tag{10}$$

Thus, in order to approximate a posterior with empirical samples from the prior, each sample $X^{(i)} \sim p(x)$ has to be weighted according to how likely it is that this particular sample has generated the observed value of the random variable Y by evaluating the likelihood $p(Y|X^{(i)})$ for this sample. Generalizing this to a dynamical inference setting will give rise to the bootstrap particle filter, and we will show in Section 6 how changing the measure in empirical sampling for a dynamical system results in dynamical equations for the particles and weights.

2.2. Filtering in discrete time - an introductory example

The inference problems in the previous section were of purely static nature.[2] However, this being the Hitchhiker's guide to nonlinear *filtering*, let us now start to consider dynamical models for filtering.

Filtering means computing the conditional distribution of the hidden state X_t at time t using the observations up to that time $Y_{0:t} = \{Y_0, \dots, Y_t\}$. There are two important ingredients to this problem: first, the **signal model** describes the dynamics of the hidden state X_t. In order to perform the inference recursively, the usual minimum assumption for the **hidden, or latent, state process** X_t with state space S is that it is a first-order Markov process, which, roughly speaking, means that the (probability of the) current state just depends on the last state, rather than on the whole history. In discrete time,[3] we can write more formally

$$p(X_{t_n}|X_{t_{0:n-1}}) = p(X_{t_n}|X_{t_{n-1}}). \tag{11}$$

Thus, the dynamics of the whole process is captured by the **transition probability** $p(X_{t_n}|X_{t_{n-1}})$, which is assumed to be known.

Second, the **observation model** describes the (stochastic) generation of the observation process Y_{t_n}, and is captured by the **emission probability** $p(Y_{t_n}|X_{t_n})$, which is also assumed to be known. Together, the transition and emission probability form a so-called state space model (SSM).[4] With these ingredients, the filtering problem in discrete time reduces to a simple application of Bayes' rule (Eq. (1)) at each time step, which may be written recursively:

$$p(X_{t_n}|Y_{0:t_n}) = \frac{p(Y_{t_n}|X_{t_n})p(X_{t_n}|Y_{0:t_{n-1}})}{p(Y_{t_n}|Y_{0:t_{n-1}})} \tag{12}$$

$$= \frac{p(Y_{t_n}|X_{t_n})\int_S dx_{t_{n-1}}\, p(X_{t_n}|x_{t_{n-1}})p(x_{t_{n-1}}|Y_{0:t_{n-1}})}{\int_S dx_{t_n}\, p(Y_{t_n}|x_{t_n})\int_S dx_{t_{n-1}}\, p(x_{t_n}|x_{t_{n-1}})p(x_{t_{n-1}}|Y_{0:t_{n-1}})}. \tag{13}$$

The simplest dynamic model for filtering is a Hidden Markov Model (HMM). To see that you do not need rocket science for hitchhiking and applying Eq. (13), let us consider an HMM with two hidden states and two observed states, i.e. X_{t_n} and Y_{t_n} can take values of 0 or 1 for each time t_n. The transition probabilities for X_{t_n} are given by

$$p(X_{t_n} = 0|X_{t_{n-1}} = 0) = \alpha, \quad p(X_{t_n} = 1|X_{t_{n-1}} = 1) = \beta. \tag{14}$$

Thus, α is the probability of staying in state 0, whereas β is the probability of staying in state 1, and leaving those states has to have probability $1-\alpha$ and $1-\beta$, respectively, where we assume that $\alpha, \beta \in (0, 1)$ such that each state is visited. This can be represented by a matrix.[5]

$$P^\top = \begin{pmatrix} \alpha & 1-\beta \\ 1-\alpha & \beta \end{pmatrix}, \tag{15}$$

which recursively determines the distribution of the hidden Markov chain at each time: if $p_{t_{n-1}} = (p^{(1)}_{t_{n-1}}, p^{(2)}_{t_{n-1}})^\top$ is a two-dimensional vector, denoting probability of state occupancy at time t_{n-1}, i.e. $p^{(1)}_{t_{n-1}} = P(X_{t_{n-1}} = 0)$ and $p^{(2)}_{t_{n-1}} = P(X_{t_{n-1}} = 1)$, the corresponding vector at time t_n is given by

$$p_{t_n} = P^\top p_{t_{n-1}}. \tag{16}$$

In our example, the emission probabilities of Y are given by a binary symmetric channel (random bit flip) with error probability $0 < \delta < 1$

$$p(Y_{t_n} = 1|X_{t_n} = 0) = \delta, \quad p(Y_{t_n} = 0|X_{t_n} = 1) = \delta. \tag{17}$$

The structure of this model is depicted in Fig. 2a.

[2] If you ask yourself why we needed 4 pages to get to this point, please bear with us: the concept of changing the measure is very straightforward in a static setting, and might help to grasp the (seemingly) more complicated applications in a dynamic setting later on.

[3] In discrete time $t_n = n\Delta t$.

[4] Somewhat oddly, the name 'state space model' usually refers to a model with continuous state space, i.e. $X_t \in \mathbb{R}^n$, which is distinct from models with finite state space such as the Hidden Markov Model below. In this tutorial, the state space can both be discrete or continuous, and if necessary, will be further clarified in the text.

[5] Here $P^\top$ is used to denote the transpose of the matrix P. It will become clear later why we define the transition matrix as $P^\top$.

For filtering, we can directly apply Eq. (13), and since the state space is discrete, the integral reduces to a sum over the possible states 0 and 1. Thus, Eq. (13) may be expressed as

$$p(X_{t_n}|Y_{t_0:t_n}) =: \hat{p}_{t_n} = \frac{\mathrm{diag}(e(Y_{t_n}))P^\top \hat{p}_{t_{n-1}}}{e(Y_{t_n})^\top P^\top \hat{p}_{t_{n-1}}}, \tag{18}$$

where $\mathrm{diag}(v)$ is a diagonal matrix with the vector v along the diagonal and e is a vector encoding the emission likelihood:

$$e(Y_{t_n}) = \begin{pmatrix} p(Y_{t_n}|X_{t_n}=0) \\ p(Y_{t_n}|X_{t_n}=1) \end{pmatrix}. \tag{19}$$

Fig. 2b shows a sample trajectory of the hidden state X_{t_n}, the corresponding observations Y_{t_n} as well as the filtered probabilities p_{t_n} and the estimated state $\hat{X}_{t_n}$. Even though what is presented here is a very simple setting (discrete time and finite number of states), it illustrates nicely that the filter takes into account both the dynamics of the hidden states as well as the reliability of the observations.

2.3. Continuous (state) space

Remarkably, in the previous example the filtering problem could be solved in closed form because it was formulated in discrete time for a discrete state space. We will now continue our journey towards more complex filtering problems involving a continuous state space. For this, we have to go back to Eq. (13), which is actually the filtering recursion for any state space. While being straightforward to write down — is it possible to solve it in closed form? Depending on the specific transition and emission densities, the integrals in Eq. (13) might not admit a closed-form solution. In fact, this is almost always the case! Except…

2.3.1. The Kalman filter

…if the transition and emission probabilities are Gaussians and linear, i.e.

$$p(X_{t_n}|X_{t_{n-1}}) = \mathcal{N}\left(X_{t_n}; AX_{t_{n-1}}, \Sigma_x\right), \tag{20}$$

$$p(Y_{t_n}|X_{t_n}) = \mathcal{N}\left(X_{t_n}; BX_{t_n}, \Sigma_y\right), \tag{21}$$

where we consider $X_{t_n} \in \mathbb{R}^k$ and $Y_{t_n} \in \mathbb{R}^l$ to be vector-valued random processes. Further, $A \in \mathbb{R}^{k\times k}$ and $B \in \mathbb{R}^{l\times k}$ are the transition and emission matrices, respectively, and $\Sigma_x \in \mathbb{R}^{n\times n}$ and $\Sigma_y \in \mathbb{R}^{l\times l}$ are state and observation noise covariances, respectively.

Let us assume that at time t_{n-1} the posterior is given by a Gaussian

$$p(X_{t_{n-1}}|Y_{0:t_{n-1}}) = \mathcal{N}(X_{t_{n-1}}; \mu_{t_{n-1}}, \Sigma_{t_{n-1}}). \tag{22}$$

We can immediately plug Eqs. (20) and (21) together with this assumption into Eq. (13). After a bit of tedious but straightforward algebra (see Bishop, 2006, Section 13.3.1), we find that the posterior is also a Gaussian $\mathcal{N}(X_t; \mu_{t_n}, \Sigma_{t_n})$. The famous Kalman filter equations give us update rules for its mean and variance:

$$\mu_{t_n} = A\mu_{t_{n-1}} + K_t(Y_{t_n} - BA\mu_{t_{n-1}}), \tag{23}$$

$$\Sigma_{t_n} = (\mathbb{I} - K_t B)\tilde{\Sigma}_{t_{n-1}}, \tag{24}$$

where $\tilde{\Sigma}_{t_{n-1}} = A\Sigma_{t_{n-1}}A^\top + \Sigma_x$ is the covariance matrix of $p(X_t|Y_{0:t_{n-1}})$, obtained after performing the marginalization over the state transition. The so-called Kalman gain K_t is given by

$$K_t = \tilde{\Sigma}_{t_{n-1}}B^\top(B\tilde{\Sigma}_{t_{n-1}}B^\top + \Sigma_y)^{-1}. \tag{25}$$

The immediate implication of this result is that for this particular model, given that the initial distribution is a Gaussian, the posterior stays Gaussian at all times.

2.3.2. Particle filtering in discrete time

In those cases where transition and emission probabilities are not Gaussian, we cannot expect Eq. (13) to take an analytically accessible form. In other words: as time goes by (in terms of time steps n), we will have to keep track of an ever-growing amount of integrals, which is clearly not desirable. Alternatively, we can try to approach this task numerically, by considering empirical samples and propagating these samples through time to keep track of this posterior. This idea is the very basis of particle filters (PF).

The only remaining problem is that direct sampling from the true posterior is usually not possible. In Section 2.1.1 we have motivated importance sampling for a static setting from a change of measure perspective, and we will now use the same reasoning to motivate *sequential* importance sampling. In other words: we will replace samples from the true posterior (under $\mathbb{P}$) by weighted samples from a proposal density under $\mathbb{Q}$. Importantly, a 'sample' i here refers to a single realization of the whole path $X_{0:t_n} = \{X_0, \dots, X_{t_n}\}$, and the measure change needs to be done with respect to the whole sequence of state and observations.

Let us first note that the posterior expectation can be understood as an expectation with respect to the whole sequence

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\phi(X_{t_n})|Y_{0:t_n}\right] &= \int_S dx_{t_n}\, \phi(x_{t_n})p(x_{t_n}|Y_{0:t_n}) \\ &= \int_S dx_{0:t_n}\, \phi(x_{t_n})p(x_{0:t_n}|Y_{0:t_n}),\end{aligned} \tag{26}$$

where in the last step we simply used that $\int_S dx_{0:t_{n-1}} p(x_{0:t_{n-1}}|Y_{0:t_n}) = 1$. Now, we perform the measure change according to Eq. (5):

$$\mathbb{E}_{\mathbb{P}}\left[\phi(X_{t_n})|Y_{0:t_n}\right] = \frac{\mathbb{E}_{\mathbb{Q}}[L(X_{0:t_n}, Y_{0:t_n})\phi(X_{t_n})|Y_{0:t_n}]}{\mathbb{E}_{\mathbb{Q}}[L(X_{0:t_n}, Y_{0:t_n})|Y_{0:t_n}]}, \tag{27}$$

with

$$L(x_{0:t_n}, y_{0:t_n}) = \frac{p(x_{0:t_n}, y_{0:t_n})}{q(x_{0:t_n}, y_{0:t_n})} = \frac{p(x_{0:t_n}|y_{0:t_n})p(y_{0:t_n})}{q(x_{0:t_n}|y_{0:t_n})q(y_{0:t_n})}, \tag{28}$$

where p and q denote densities of $\mathbb{P}$ and $\mathbb{Q}$, respectively.

Let us now choose the measure $\mathbb{Q}$ such that the conditional density $q(x_{0:t_n}|y_{0:t_n})$ factorizes, i.e.

$$\begin{aligned}q(x_{0:t_n}|y_{0:t_n}) &= \prod_{j=0}^{n} \pi(x_{t_j}|x_{0:t_{j-1}}, y_{0:t_j}) \\ &= \pi(x_{t_n}|x_{0:t_{n-1}}, y_{0:t_n})q(x_{0:t_{n-1|}}|y_{0:t_{n-1}}).\end{aligned} \tag{29}$$

Further, we can rewrite the conditional density $p(x_{0:t_n}|y_{0:t_n})$ using the structure of the SSM

$$\begin{aligned}&p(x_{0:t_n}|y_{0:t_n}) \\ &= \frac{p(y_{t_n}|x_{0:t_n}, y_{0:t_{n-1}})p(x_{0:t_n}|y_{0:t_{n-1}})}{p(y_{t_n}|y_{0:t_{n-1}})} \\ &= \frac{p(y_{t_n}|x_{0:t_n}, y_{0:t_{n-1}})p(x_{t_n}|x_{0:t_{n-1}}, y_{0:t_{n-1}})}{p(y_{t_n}|y_{0:t_{n-1}})}p(x_{0:t_{n-1}}|y_{0:t_{n-1}}) \\ &= \frac{p(y_{t_n}|x_{t_n})p(x_{t_n}|x_{t_{n-1}})}{p(y_{t_n}|y_{0:t_{n-1}})}p(x_{0:t_{n-1}}|y_{0:t_{n-1}}).\end{aligned} \tag{30}$$

Thus, using that all factors independent of the state variable x can be taken out of the expectations in Eq. (28) and cancel subsequently, we find

$$\begin{aligned}L(x_{0:t_n}, y_{0:t_n}) &\propto \frac{p(y_{t_n}|x_{t_n})p(x_{t_n}|x_{t_{n-1}})}{\pi(x_{t_n}|x_{0:t_{n-1}}, y_{0:t_n})}\frac{p(x_{0:t_{n-1}}|y_{0:t_{n-1}})}{q(x_{0:t_{n-1|}}|y_{0:t_{n-1}})} \\ &\propto \frac{p(y_{t_n}|x_{t_n})p(x_{t_n}|x_{t_{n-1}})}{\pi(x_{t_n}|x_{0:t_{n-1}}, y_{0:t_n})}L(x_{0:t_{n-1}}, y_{0:t_{n-1}}).\end{aligned} \tag{31}$$

In analogy to Section 2.1.1, we now take M i.i.d. samples from the proposal density, i.e. we sample $X^{(i)}_{0:t_n} \sim q(X_{0:t_n}|Y_{0:t_n})$,

and weigh them according to the value of the likelihood ratio evaluated at the particle positions (cf. Eq. (9)). Since the proposal in Eq. (29) was chosen to factorize, both the sampling process as well as the evaluation of the unnormalized importance weights $w_{t_n}^{(i)}$ (according to Eq. (31)) can be done recursively. More specifically, the problem of sampling (and weighing) the whole sequences $X_{0:t_n}^{(i)}$ is replaced by sampling just a single transition $X_{t_n}^{(i)}$ for each of the M particles at each time step n and updating the associated particle weights.

$$X_{t_n}^{(i)} \sim \pi(X_{t_n}|X_{0:t_{n-1}}^{(i)}, Y_{0:t_n}), \tag{32}$$

$$w_{t_n}^{(i)} = L(X_{0:t_n}^{(i)}, Y_{0:t_n}) = w_{t_{n-1}}^{(i)} \frac{p(Y_{t_n}|X_{t_n}^{(i)})\, p(X_{t_n}^{(i)}|X_{t_{n-1}}^{(i)})}{\pi(X_{t_n}^{(i)}|X_{0:t_{n-1}}^{(i)}, Y_{0:t_n})}, \tag{33}$$

such that the posterior expectation is approximated by

$$\mathbb{E}_{\mathbb{P}}\left[\phi(X_{t_n})|Y_{0:t_n}\right] = \frac{1}{Z_{t_n}} \sum_{i=1}^{M} w_{t_n}^{(i)} \phi(X_{t_n}^{(i)}), \tag{34}$$

with $Z_{t_n} = \sum_{i=1}^{P} w_{t_n}^{(i)}$.

A simple (but not necessarily efficient) choice is to use the transition probability $p(X_{t_n}|X_{t_{n-1}})$ as the proposal function in Eq. (32). Then, computation of the unnormalized weights simplifies to

$$w_{t_n}^{(i)} = w_{t_{n-1}}^{(i)} p(Y_{t_n}|X_{t_n}^{(i)}). \tag{35}$$

This scheme is the basis of the famous **Bootstrap PF** (BPF, Gordon, Salmond, & Smith, 1993).[6] Doucet, Godsill, and Andrieu (2000) state that the BPF is "inefficient in simulations as the state space is explored without any knowledge of the observations". To account for this, alternative proposal densities can be crafted in discrete time, which may take into account the observations in the particle transitions (e.g. the 'optimal proposal' in Doucet et al., 2000).

3. Knowing where your towel is: setting the stage for continuous-time models

> *"A towel is about the most massively useful thing an interstellar hitchhiker can have.*
> *Partly it has great practical value. More importantly,*
> *a towel has immense psychological value".*
>
> [Douglas Adams]

So far, we have made our journey from Bayes' theorem to discrete-time filtering, first for discrete state spaces and then made the transition towards continuous state space models. The next logical step would be the transition to continuous time models. In the following three sections, we will see that the mindset is very similar to the approaches taken before, just in their respective continuous-time limit, i.e. $dt = t_n - t_{n-1} \to 0$. In particular, we will use the change of measure approach to derive the filtering equations, i.e. dynamical equations for the posterior expectations $\mathbb{E}[\phi(X_t)|Y_{0:t}]$ or, equivalently, the posterior density $p(X_t|Y_{0:t})$.

But let us take a step back here and first explain the model assumptions under which we will present continuous-time filtering theory. For the purpose of this tutorial, we have seen that a generative model consists of two parts:

1. A *signal* model or *hidden process* model that describes the dynamics of some system whose states we want to estimate. In continuous-time, we will consider two very general classes of signal model, namely continuous-time Markov chains (countable or finite state space) and jump–diffusion processes (continuous state space).
2. An *observation model* that describes how the system generates the information that we can observe and utilize in order to estimate the state. We will elaborate the filtering theory for two types of observation noise, namely continuous-time Gaussian noise and Poisson noise.

3.1. Signal models

As in Section 2, we will restrict ourself to the treatment of Markovian processes for the signal, i.e. $p(X_t|X_{0:t-dt}) = p(X_t|X_{t-dt})$. Our goal in this subsection will be to obtain dynamical equations that fully describe the signal process.

3.1.1. Markov chain

An important example is when X_t is a continuous-time time-homogeneous Markov chain with a finite number of states, i.e. $S = \{1, \dots, m\}$. In this case we may represent the function $\phi : \{1, \dots, m\} \to \mathbb{R}$ as a vector $\phi = (\phi(1), \dots, \phi(m))^\top$ and we have a transition probability matrix $P(t)^\top$. The entry $P_{ji}(t)$ gives the probability of going from state j to state i within a time interval of length t, so it is a time-dependent generalization of Eq. (15). This allows us to compute the distribution at time t, $p(t)$ from the initial distribution $p(0)$ as $p(t) = P(t)^\top p(0)$. We therefore have two equivalent ways of computing the expectation of ϕ:

$$\begin{aligned}\mathbb{E}[\phi(X_t)] &= p(t)^\top \phi \\ &= p(0)^\top P(t) \phi = p(0)^T \phi(t).\end{aligned} \tag{36}$$

In the first one, the observable is fixed while the distribution changes as a function of time, while in the second, the distribution is fixed to the initial distribution, and the observable evolves in time, i.e. $\phi(t) = P(t)\phi$.

By differentiating with respect to time, we obtain differential equations for the distribution $p(t)$ and the observable $\phi(t)$,

$$\dot{\phi}(t) = \dot{P}(t)\phi, \tag{37}$$

$$\dot{p}(t) = \dot{P}(t)^\top p(0). \tag{38}$$

The Markov property ensures that $P(t+s) = P(t)P(s) = P(s)P(t)$. Further, since $P(0) = \mathbb{I}$ is the unit matrix,[7] the time derivative of the matrix $P(t)$ can be simplified to

$$\begin{aligned}\dot{P}(t) &= \lim_{s\to 0} \frac{P(t+s) - P(t)}{s} = P(t) \lim_{s\to 0} \frac{P(s) - \mathbb{I}}{s} \\ &= P(t)\dot{P}(0).\end{aligned} \tag{39}$$

We denote $A = \dot{P}(0)$ and then get

$$\dot{\phi}(t) = A\phi(t), \tag{40}$$

$$\dot{p}(t) = A^\top p(t). \tag{41}$$

Equivalently, we find for the time derivative of the expectation

$$\frac{d}{dt}\mathbb{E}[\phi(X_t)] = p(0)^\top \dot{P}(t)\phi = p(t)^\top A\phi = \mathbb{E}[A\phi]. \tag{42}$$

So conceptually, the whole temporal evolution of the stochastic process X_t is encapsulated in the matrix A, the so-called **generator** matrix. In other words, the generator matrix, together with the initial distribution, is all we need to completely characterize the Markov chain.

[6] Although technically, the BPF requires a resampling step at every iteration step.

[7] Because $p(t=0) = P(0)^\top p(0)$ is only fulfilled if $P(0) = \mathbb{I}$.

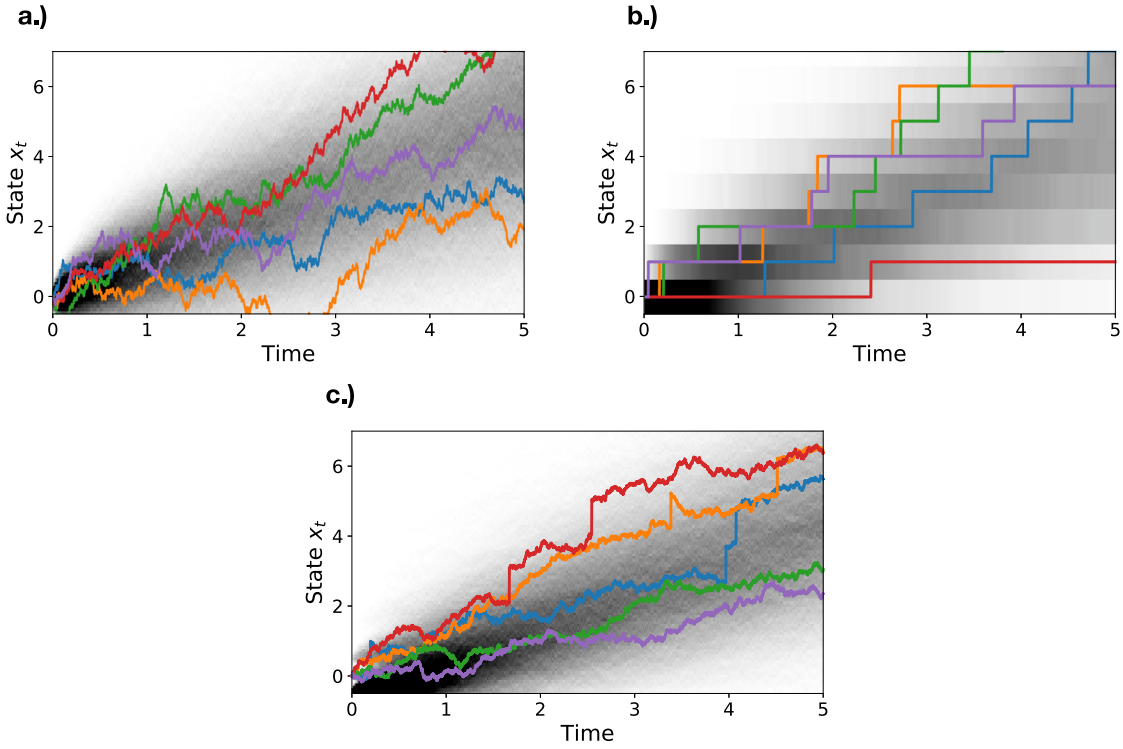

Fig. 3. Example trajectories from Eq. (43). Shading denotes density of 10'000 simulated trajectories. (a) Drift–diffusion process ($f(x) = 1, G(x) = 1, J(x) = 0$). (b) Jump process ($f(x) = 0, G(x) = 0, J(x) = 1$) with rate $\lambda(x) = 1$. c.) Jump–diffusion process ($f(x) = 1/2, G(x) = 1/2, J(x) = 1$) with rate $\lambda(x) = 1/2$.

3.1.2. Jump–diffusion process

Intuitively, in order to make the transition to a continuous state space, we have to exchange "sums by integrals and matrices by linear operators". We will now see that this holds for the hidden state dynamics by characterizing a continuous-time stochastic process with continuous state space S similarly to Eqs. (41) and (42).

An important signal model, which is a generalization of the classical diffusion model in continuous time, is a hidden state X_t that is a **jump–diffusion process**, i.e. it evolves according to a stochastic differential equation (SDE) in $S = \mathbb{R}^n$,

$$dX_t = f(X_t, t)\,dt + G(X_t, t)\,dW_t + J(X_t, t)dN_t. \tag{43}$$

Here, $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, $G : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{n\times m}$, and $J : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{n\times k}$ are called the drift, diffusion, and jump coefficients of X_t, respectively. The process noise is modeled by two types of noise sources: $W_t \in \mathbb{R}^m$ is a vector Brownian motion that models white Gaussian noise in continuous time, and we may consider $dW_t \sim \mathcal{N}(0, \mathbb{I}^{m\times m}dt)$. N_t is a k-dimensional point process with k-dimensional rate (or intensity) vector $\lambda(X_t)$, i.e. $dN_t^i \sim \text{Poisson}(\lambda_i(X_t)dt)$. Note that dN_t^i takes only values 0 or 1, because in the limit $dt \to 0$, the Poisson distribution becomes a Bernoulli distribution. In Fig. 3, we show example trajectories from Eq. (43), one being a drift–diffusion (where the jump term vanishes), one being a pure jump process, and the last one being a jump–diffusion process.

Dealing with this type of SDE model is considerably more technical than the Markov chains above. Therefore, we will outline the theory of diffusion processes for readers who are new to them. Unless stated otherwise, derivations presented here roughly follow Bremaud (1981) and Gardiner (2009).

We can choose to describe the process in terms of transition probability densities $p(x, t|x', s)$, which give the probability density at a point $x \in \mathbb{R}^n$ at time t conditioned on starting at a point $x' \in \mathbb{R}^n$ at time $s < t$. This transition density can be combined with the initial density $p_0(x')$ by integrating in order to compute an expectation:

$$\begin{aligned}\mathbb{E}[\phi(X_t)] &= \iint \phi(x)p(x, t|x', 0)p_0(x')\,dx\,dx' \\ &= \int \phi(x)p(x, t)\,dx,\end{aligned} \tag{44}$$

in complete analogy with the vector–matrix–vector product for the Markov chains in Eq. (36). Taking this analogy further, differentiating with respect to time gives rise to two different (equivalent) ways of writing the time evolution of the expected value:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}[\phi(X_t)] &= \int \phi(x)\partial_t p(x, t)\,dx \\ &= \int \phi(x)\mathcal{A}^\dagger p(x, t)\,dx \\ &= \int \mathcal{A}\phi(x)p(x, t)\,dx,\end{aligned} \tag{45}$$

where in analogy to Eq. (42) we have introduced the adjoint operator $\mathcal{A}^\dagger$ that describes the time evolution of the probability density. Thus, in analogy to Eq. (37) we can set out to find the appropriate form of the **infinitesimal generator** $\mathcal{A}$, which generalizes the generator matrix A of the Markov chain, and then, by integration by parts, we may derive the corresponding adjoint operator $\mathcal{A}^\dagger$.

Itô lemma for jump diffusions. The form of the generator $\mathcal{A}$ can be obtained by changing the variables in Eq. (43) from the random variable X_t to the random variable $\phi_t := \phi(X_t)$. The following calculation will be performed for a scalar process X_t.[8] Consider

[8] Generalization to a multivariate state process X_t is straightforward.

an infinite Taylor expansion of its increment $d\phi_t$ around $dX_t = 0$ up to $\mathcal{O}(dt)$:

$$\begin{aligned} d\phi_t &= \phi(X_t + dX_t) - \phi(X_t) \\ &= \sum_{n=1}^{\infty} \frac{1}{n!} \phi_t^{(n)} (dX_t)^n, \end{aligned} \tag{46}$$

with $\phi_t^{(n)} := \left(\partial_x^n \phi(x)\right)|_{x=X_t}$.

In a deterministic differential, Taylor-expanding up to first order would suffice since $dt^n = 0\ \forall n > 1$. In Eq. (43), the additional stochastic terms add additional orders of dt. Particularly, since the variance of the Brownian motion process grows linearly in time, we have $dW_t^2 = dt$, and thus the diffusion term has to be expanded up to second order. For the jump term, all order up to infinity have to be considered: indeed, N_t is not a continuous process, and elicits jumps of always size 1 irrespectively of the infinitesimally small time interval dt. Therefore, any power of this jump will have the same magnitude, i.e. $dN_t^n = dN_t$, $\forall n$. Thus, we find for a scalar process

$$\begin{aligned} d\phi_t &= \left[f(X_t, t)\phi_t' + \frac{1}{2} G^2(X_t, t)\phi_t'' \right] dt + G(X_t, t)\phi_t'\, dW_t \\ &\quad + \sum_{n=1}^{\infty} \frac{1}{n!} J^n(X_t, t)\phi_t^{(n)}\, dN_t \\ &= \left[f(X_t, t)\phi_t' + \frac{1}{2} G^2(X_t, t)\phi_t'' \right] dt \\ &\quad + \left[\phi\left(X_t + J(X_t)\right) - \phi(X_t)\right] \lambda(X_t) dt \\ &\quad + G(X_t, t)\phi_t'\, dW_t + \left[\phi\left(X_t + J(X_t)\right) - \phi(X_t)\right] \\ &\quad \times \left(dN_t - \lambda(X_t) dt\right) \\ &=: \mathcal{A}\phi_t\, dt + dM_t^{\phi}. \end{aligned} \tag{47}$$

This formula is called **Itô's lemma**. In the last step, we have defined

$$\begin{aligned} \mathcal{A}\phi_t &= f(X_t, t)\phi_t' + \frac{1}{2} G^2(X_t, t)\phi_t'' \\ &\quad + \lambda(X_t)\left(\phi\left(X_t + J(X_t)\right) - \phi(X_t)\right), \end{aligned} \tag{48}$$

$$\begin{aligned} dM_t^{\phi} &= G(X_t, t)\phi_t'\, dW_t + \left[\phi\left(X_t + J(X_t)\right) - \phi(X_t)\right] \\ &\quad \times \left(dN_t - \lambda(X_t) dt\right), \end{aligned} \tag{49}$$

where $\mathcal{A}$ is the infinitesimal generator of the stochastic process X_t. The stochastic process M_t^{ϕ} is a so-called martingale[9] and the contribution from its increment vanishes upon taking expectations, i.e. $\mathbb{E}[dM_t^{\phi}] = 0$. Thus, taking expectations on both sides of Eq. (47) we find indeed

$$\frac{d}{dt}\mathbb{E}[\phi(X_t)] = \mathbb{E}[\mathcal{A}\phi(X_t)], \tag{50}$$

which is the continuous state space analogue to Eq. (42).

The multivariate version is completely analogous:

$$d\phi(X_t) = \mathcal{A}\phi(X_t) dt + dM_t^{\phi},$$

where now the infinitesimal generator of the stochastic process is given by

$$\begin{aligned} \mathcal{A}\phi(x) &= \sum_{i=1}^{n} f_i(x, t)\partial_{x_i}\phi(x) + \frac{1}{2}\sum_{i,j=1}^{n} \left(GG^{\top}(x, t)\right)_{ij} \partial_{x_i}\partial_{x_j}\phi(x) \\ &\quad + \sum_{i=1}^{k} \lambda_i(x)\Big[\phi\big(x + J_i(x, t)\big) - \phi(x)\Big], \end{aligned} \tag{51}$$

and the martingale part reads

$$\begin{aligned} dM_t^{\phi} &= \sum_{i=1}^{n}\sum_{j=1}^{m} G_{ij}(X_t, t)(\partial_{x_i}\phi(x)|_{x=X_t}) dW_s^j \\ &\quad + \sum_{i=1}^{k} \left[\phi\left(X_t + J_i(X_t)\right) - \phi(X_t)\right] \left(dN_t^i - \lambda_i(X_t) ds\right). \end{aligned} \tag{52}$$

The generator $\mathcal{A}$, just like the generator matrix A of the Markov chain, together with the initial distribution, completely characterizes the Markov process and allows us to describe its time evolution on an abstract level. Or in other words: even though the particular form of $\mathcal{A}$ might be different for each of these models presented here, the structure of the mathematics remains the same, and can therefore be generalized to arbitrary $\mathcal{A}$ when the need arises.

The evolution of the probability density. With the explicit form in Eq. (51) of the generator $\mathcal{A}$, we can go back to Eq. (45) and perform the integration by parts to find the adjoint operator $\mathcal{A}^{\dagger}$, which will take the role of $A^{\top}$ in the Markov chain case.

Plugging Eq. (51) into Eq. (45), we obtain

$$\begin{aligned} &\int_{\mathbb{R}^n} \mathcal{A}\phi(x) p(x, t) dx \\ &= \sum_{i=1}^{n} \int_{\mathbb{R}^n} f_i(x, t)(\partial_{x_i}\phi(x)) p(x, t) dx \\ &\quad + \frac{1}{2}\sum_{i,j=1}^{n} \int_{\mathbb{R}^n} \left(GG^{\top}(x, t)\right)_{ij} (\partial_{x_i}\partial_{x_j}\phi(x)) p(x, t) dx \\ &\quad + \sum_{i=1}^{k} \int_{\mathbb{R}^n} \lambda_i(x)\Big[\phi\big(x + J_i(x, t)\big) - \phi(x)\Big] p(x, t) dx. \end{aligned} \tag{53}$$

The first two integrals can be dealt with by ordinary integration by parts,[10] i.e.

$$\int_{\mathbb{R}^n} f_i(x, t)(\partial_{x_i}\phi(x)) p(x, t) dx = -\int_{\mathbb{R}^n} \phi(x)\Big[\partial_{x_i} f_i(x, t) p(x, t)\Big] dx, \tag{54}$$

and

$$\begin{aligned} &\int_{\mathbb{R}^n} \left(GG^{\top}(x, t)\right)_{ij} (\partial_{x_i}\partial_{x_j}\phi(x)) p(x, t) dx \\ &= \int_{\mathbb{R}^n} \phi(x)\partial_{x_i}\partial_{x_j}\Big[\left(GG^{\top}(x, t)\right)_{ij} p(x, t)\Big] dx. \end{aligned} \tag{55}$$

For the third integral in Eq. (53), we perform a change of variables $x \mapsto x + J_i(x, t)$ (where we assume the integral boundaries are not affected by the substitution), thereby obtaining

$$\begin{aligned} &\int_{\mathbb{R}^n} \lambda_i(x)\Big[\phi\big(x + J_i(x, t)\big) - \phi(x)\Big] p(x, t) dx \\ &= \int_{\mathbb{R}^n} \phi(x) \Big[\lambda_i(x - J_i(x, t)) p(x - J_i(x, t), t) \\ &\quad \times \det \frac{\partial(J_{1i}, \ldots, J_{ji})}{\partial(x_1, \ldots, x_j)} - \lambda_i(x) p(x, t) \Big] dx, \end{aligned} \tag{56}$$

where $\det \frac{\partial(J_{1i}, \ldots, J_{ji})}{\partial(x_1, \ldots, x_j)}$ denotes the Jacobi determinant of the entries of the column vector J_i. Combining all of these integrals (including the sums) and comparing with Eq. (45), we can therefore read off

[9] Loosely speaking, a martingale is a sequence of random variables, whose conditional expectation in the next time step is equal to the value of the random variable at the current time step.

[10] Here, we make the assumption that the density $p(x, t)$ and all its derivatives vanish at infinity.

the form of the adjoint operator:

$$\begin{aligned}\mathcal{A}^\dagger p(x,t) = &\sum_{i=1}^{n} \partial_{x_i}\Big[f_i(x,t)p(x,t)\Big] \\ &+ \frac{1}{2}\sum_{i,j=1}^{n} \partial_{x_i}\partial_{x_j}\Big[\big(GG^\top(x,t)\big)_{ij}p(x,t)\Big] \\ &+ \sum_{i=1}^{k}\Big[\,\lambda_i(x - J_i(x,t))p\big(x - J_i(x,t),t\big) \\ &\times \det\frac{\partial(J_{1i},\ldots,J_{ji})}{\partial(x_1,\ldots,x_j)} - \lambda_i(x)p(x,t)\,\Big]\,.\end{aligned} \tag{57}$$

Using Eq. (45) once more, we find the evolution equation for the density $p(x,t)$:

$$\partial_t p(x,t) = \mathcal{A}^\dagger p(x,t). \tag{58}$$

If we leave out the jump terms, this is called the **Fokker–Planck equation** or Kolmogorov forward equation. With the jump terms, it is often referred to as the **Master equation**.

3.2. Observation model

In the previous section, we have encountered various signal models, which are the processes we want to infer. The knowledge about how these processes evolve in time, formally given by the generator $\mathcal{A}$, serves as prior knowledge to the inference task. Equally important, we need measurements, or observations, to update this prior knowledge. In particular, an observation model describes how the signal gets corrupted during measurement. This may comprise both a lossy transformation (e.g. only certain components of a vector-valued process are observed), and some stochastic additive noise that randomly corrupts the measurements. Roughly speaking, the measurements Y_t are given by

$$Y_t = h(X_t) + \text{noise},$$

but we will need to be careful about the precise way in which the noise is added in order to make sense in continuous time.

In the following, we will consider two types of noise: Gaussian and Poisson. The simplicity of noise of these two noise models greatly simplifies the formal treatment of the filtering problem, and while the two types of noise seem very different, there is a common structure that will emerge.

When considering more general noise models than the ones below, the technique of Section 4 (change of measure) can be applied whenever the observation noise (whatever is added to the deterministic transformation) is additive and independent of the hidden state.

3.2.1. Continuous-time Gaussian noise

The simplest noise model is often white Gaussian noise. For continuous-time observations, however, one cannot simply take an observation model $Y_t = h(X_t) + \eta_t$ with independent Gaussian η_t because for a reasonably well-behaved process X_t, an integration of Y_t over a finite time interval would completely average out the noise and therefore allow one to perfectly recover the transformed signal $h(X_t)$.[11] The filtering problem would therefore be reduced to simply inverting h.

One way of resolving the problem of finding a (nontrivial) model of white Gaussian noise is to switch to a differential form and use increments of the Wiener process as a noise term. One therefore obtains an SDE for the **observation process** Y_t:

$$dY_t = h(X_t,t)\,dt + \Sigma_y(t)^{1/2}\,dV_t. \tag{59}$$

Here, $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^l$ is a vector-valued function that links the hidden state (and time, if time-dependence is explicit) with the deterministic drift of the observations. Further, $V_t \in \mathbb{R}^l$ is a vector Brownian motion process and $\Sigma_y(t) : \mathbb{R} \to \mathbb{R}^{l\times l}$ is the time-dependent observation noise covariance.

In the standard literature, one usually finds the special case $\Sigma_y = \mathbb{I}^{l\times l}$, which is equivalent to Eq. (59) if the increment of the observation process Y_t is rescaled accordingly:

$$d\tilde{Y}_t = \Sigma_y(t)^{-1/2}dY_t = \tilde{h}(X_t,t)\,dt + dV_t, \tag{60}$$

where $\tilde{h}(x,t) = \Sigma_y(t)^{-1/2}h(x,t)$ is the rescaled observation function.

3.2.2. Poisson noise

In many fields, observations come in the form of a series of events. Examples include neuroscience (neural spike trains), geoscience (earthquakes, storms), financial transactions, etc. This suggests a point process (whose output is a series of event times) or counting process (which counts the number of events) observation model. A simple, but versatile, model for events is a Poisson process N_t with time-varying and possibly history-dependent intensity. As an observation model, this doubly-stochastic Poisson process (also known as Cox process, Cox 1955) has an intensity that depends on its own history as well as the hidden state. We can informally write this as

$$dN_t^i \sim \text{Poisson}\left(R_t^i\,dt\right), \quad i = 1,\ldots,l \tag{61}$$

where the intensity processes R_t^i are nonnegative processes that can be computed from the current value of X_t and the history of observations $N_{0:s}$ for $s < t$.

To keep the notation simple, we will assume that the vector R_t of intensities is given by a function of the hidden state,

$$R_t = h(X_t), \tag{62}$$

but history-dependence in the form $R_t = h(X_t, N_{0:t-})$ does not significantly increase the difficulty of the filtering (given the observation, any history-dependence of the intensity is deterministic and can be factored out of the conditional expectation of the intensity).

4. The answer to life, the universe and (not quite) everything: the filtering equations

> *"I've just been created. I'm completely new to the Universe in all respects.*
> *Is there anything you can tell me?"*
>
> [Douglas Adams]

The filtering problem is to compute the posterior (or conditional) density of the hidden state conditioned on the whole sequence of observations up to time t, $Y_{0:t}$, or equivalently, to compute the posterior expectation (if it exists) of any real-valued measurable function $\phi : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}_{\mathbb{P}}\left[\phi(X_t)|Y_{0:t}\right] = \int_{-\infty}^{\infty} p(x|Y_{0:t})\phi(x)\,dx =: \langle\phi_t\rangle_{\mathbb{P}}\,, \tag{63}$$

where we use subscript $\mathbb{P}$ to indicate expectations with respect to the original probability measure $\mathbb{P}$.

That this is not an easy problem should be clear by now, because already the discrete-time filtering task (e.g. in Eq. (13)) involved a computation of as many integrals as there are time

[11] If the observations are made at discrete times $t_1, t_2, \ldots$, this is not problematic. Filtering of a continuous-time hidden process with discrete-time observations is reviewed in Jazwinski (1970, p. 163ff). If the observation model in Eq. (59) is discretized, one gets back to a discrete-time observation model with Gaussian noise.

steps. In continuous time, this would amount to an infinite number of integrals. This continuous-time problem has already been recognized and tackled by mathematicians in the 60s and 70s of the last century, providing formal solutions for the posterior density in terms of stochastic partial differential equations (Kushner, 1962; Zakai, 1969). In the following, we will derive these equations, using what we have been using in the previous sections as our ultimate "Point of View Gun for nonlinear filtering"[12]: the change of probability measure method.[13]

4.1. Changes of probability measure — once again

Let us once more revisit the change of probability measure in the context of filtering. The goal is to pass from the original probability measure $\mathbb{P}$ (under which the processes behave as our signal and observation model dictates), to an equivalent measure $\mathbb{Q}$, called *reference measure*, under which the observation process becomes simpler and decouples from the signal process. Here, we will finally generalize our introductory treatment from Section 2.1 to stochastic processes. Unsurprisingly, the calculations are quite similar.

If $\mathbb{P}$ is a probability measure and we have a collection of processes (X_t and Y_t), the measure $\mathbb{P}_t$ is the restriction of $\mathbb{P}$ to all events that can be described in terms of the behavior of X_s and Y_s for $0 \leq s \leq t$. If $\mathbb{P}$ and $\mathbb{Q}$ are equivalent, also their restrictions $\mathbb{P}_t$ and $\mathbb{Q}_t$ are equivalent.[14] The **Radon–Nikodym theorem** (Klebaner, 2005, Theorem 10.6, p. 272ff) then states that a random variable L_t exists, such that for all functions ϕ

$$\mathbb{E}_{\mathbb{P}}\left[\phi(X_t)\right] = \mathbb{E}_{\mathbb{Q}}\left[L_t \cdot \phi(X_t)\right], \tag{64}$$

where $L_t = \frac{d\mathbb{P}_t}{d\mathbb{Q}_t}$ is called the Radon–Nikodym derivative or density of $\mathbb{P}_t$ with respect to $\mathbb{Q}_t$. This is the generalization of Eq. (2) in Section 2.1.

In analogy to Eq. (64), also the conditional expectations can then be rewritten in terms of a reference probability measure $\mathbb{Q}$:

$$\mathbb{E}_{\mathbb{P}}\left[\phi_t | Y_{0:t}\right] = \frac{\mathbb{E}_{\mathbb{Q}}[\phi_t\, L_t | Y_{0:t}]}{\mathbb{E}_{\mathbb{Q}}[L_t | Y_{0:t}]} = \frac{1}{Z_t}\left\langle \phi_t L_t \right\rangle_{\mathbb{Q}}. \tag{65}$$

Eq. (65) is known as a Bayes' formula for stochastic processes (compare Eq. (5)) or **Kallianpur–Striebel formula**. Here, we require a time-dependent normalization $Z_t := \mathbb{E}_{\mathbb{Q}}[L_t|Y_{0:t}]$, and $\langle \phi_t L_t \rangle_{\mathbb{Q}} := \mathbb{E}_{\mathbb{Q}}[\phi_t\, L_t | Y_{0:t}]$ was introduced to keep the notation concise. This generalizes Eq. (5).

But wait: what exactly does the Radon–Nikodym derivative L_t look like? This really depends on the measure change we are about to perform, but it helps to recall that in a discrete-time (or actually already static) setting the equivalent of the Radon–Nikodym derivative is nothing else than the ratio between two probability densities. For the filtering problem below, we will choose a reference measure $\mathbb{Q}$ such that the path of the observations $Y_{0:t}$ (or equivalently the set of the increments $dY_{0:t}$) becomes independent of the path of the state process $X_{0:t}$, i.e. $q(X_{0:t}, dY_{0:t}) = p(X_{0:t})q(dY_{0:t})$. This is very convenient, as this allows us to compute expectations with respect to the statistics of the state process (and we know how to do that). Eq. (6) then directly tells us what the likelihood ratio has to look like for this measure change:

$$L(x_{0:t}, dy_{0:t}) = \frac{p(dy_{0:t}|x_{0:t})}{q(dy_{0:t})} = \frac{\prod_{s=0}^{t} p(dy_s|x_s)}{q(dy_{0:t})}, \tag{66}$$

where now $x_{0:t}$ and $dy_{0:t}$ are variables reflecting the whole path of the random variable X_t and the set of infinitesimal increments $dY_{0:t}$. Importantly, this particular measure change is agnostic to how the hidden state variable X_t evolves in time, but just takes into account how the observations are generated via $p(dy_t|x_t)$.

Let us first consider Gaussian observation noise, as encountered in Section 3.2.1. From Eq. (59), we know that $dY_t \sim \mathcal{N}(dY_t; h(X_t)dt, \Sigma_y dt)$. Further, we choose $q(dY_t) = \mathcal{N}(dY_t; 0, \Sigma_y dt)$ under the reference measure $\mathbb{Q}$. Thus, the Radon–Nikodym derivative $L_t = L(X_{0:t}, dY_{0:t})$ can be written as

$$\begin{aligned} L_t &= \frac{\prod_{s=0}^{t} p(dY_s|X_s)}{\prod_{s=0}^{t} q(dY_s)} = \prod_{s=0}^{t} \frac{\mathcal{N}(dY_s; h(X_s)ds, \Sigma_y ds)}{\mathcal{N}(dY_s; 0, \Sigma_y ds)} \\ &= \prod_{s=0}^{t} \exp\left[h(X_s)^\top \Sigma_y^{-1} dY_s - \frac{1}{2} h(X_s)^\top \Sigma_y^{-1} h(X_s) ds \right] \\ &\overset{\lim_{dt\to 0}}{=} \exp\left[\int_0^t h(X_s)^\top \Sigma_y^{-1} dY_s - \frac{1}{2} h(X_s)^\top \Sigma_y^{-1} h(X_s) ds \right], \end{aligned} \tag{67}$$

where in the last step we took the continuum limit $\lim_{dt\to 0}$. Consistently, we would have obtained this result if we had simply (and mindlessly) applied a theorem called Girsanov's theorem, choosing the reference measure $\mathbb{Q}$ under which the rescaled observations process $\tilde{Y}_t$ is a Brownian motion process (Klebaner, 2005, Chapter 10.3, on p. 274, in particular Remark 10.3).

Similarly, we can compute the Radon–Nikodym derivative for observations corrupted by Poisson noise as in Eq. (61). Here, we choose $\mathbb{Q}$ such that $q(dN_{0:t})$ is Poisson with a constant reference rate λ_0. The corresponding density reads

$$\begin{aligned} L_t &= \prod_{s=0}^{t} \prod_{i=1}^{l} \frac{p(dN_s^i|X_s)}{q(dN_s^i)} = \prod_{s,i} \frac{\text{Poisson}(dN_s^i; h_i(X_s)ds)}{\text{Poisson}(dN_s^i; \lambda_0 ds)} \\ &= \prod_{s,i} \exp\left[\lambda_0 ds - h_i(X_s)\, ds + \log \frac{h_i(X_s)}{\lambda_0} dN_s^i \right] \qquad (68) \\ &\overset{\lim_{dt\to 0}}{=} \prod_{i=1}^{l} \exp\left[\int_0^t (\lambda_0 - h_i(X_s))\, ds + \log \frac{h_i(X_s)}{\lambda_0} dN_s^i \right]. \qquad (69) \end{aligned}$$

Again, the same result could have been obtained with a Girsanov theorem (see Bremaud, 1981, Chapter VI, Theorems T2 and T3).

4.2. Filtering equations for observations corrupted by Gaussian noise

We are now equipped with the necessary tools to tackle the derivation of the filtering equations. Here, the derivation will be briefly outlined (for a more detailed and formal derivation, see Bain & Crisan, 2009; Van Handel, 2007).

As we stated in the beginning of this Section, we want to find a *convenient* reference measure which decouples the signal and observations and at the same time turns the observations into something simple. Recall that the Radon–Nikodym derivative (expressed for a rescaled observations process $\tilde{Y}_t = \Sigma_y^{-1/2} Y_t$ with $\tilde{h}(x) = \Sigma_y^{-1/2} h(x)$) then takes the form

$$L_t = \frac{d\mathbb{P}_t}{d\mathbb{Q}_t} = \exp\left[\int_0^t \tilde{h}(X_s)^\top d\tilde{Y}_s - \frac{1}{2} \int_0^t \left\| \tilde{h}(X_s) \right\|^2 ds \right]. \tag{70}$$

which evolves according to the following SDE (see Appendix A.1 for calculation steps):

$$dL_t = L_t \tilde{h}(X_t)^\top d\tilde{Y}_t. \tag{71}$$

[12] The Point of View Gun is a weapon that causes its target to see things from the side of the shooter. Actually, it never appeared in any of Douglas Adams' novels, but it was featured in the 2005 movie.

[13] There are other methods to arrive at the same equations, for instance the *innovations approach* (Bain & Crisan, 2009, Chpt. 3.7) or the more heuristic continuum limit approach originally taken by Kushner (1962).

[14] For stochastic processes, equivalence implies having the same noise covariance.

Under $\mathbb{Q}$, the stochastic differential can be taken inside the expectation (see Van Handel, 2007, Chapter 7, Lemma 7.2.7), and we therefore obtain using Itô's lemma[15]

$$\begin{aligned} &d\mathbb{E}_{\mathbb{P}}\left[\phi_t|Y_{0:t}\right] \\ &= d\left(\frac{1}{Z_t}\left\langle\phi_t L_t\right\rangle_{\mathbb{Q}}\right) \\ &= \frac{1}{Z_t}\, d\left\langle\phi_t L_t\right\rangle_{\mathbb{Q}} + \left\langle\phi_t L_t\right\rangle_{\mathbb{Q}}\, d\left(\frac{1}{Z_t}\right) + d\left\langle\phi_t L_t\right\rangle_{\mathbb{Q}}\, d\left(\frac{1}{Z_t}\right) \\ &= \frac{1}{Z_t}\,\left\langle d(\phi_t L_t)\right\rangle_{\mathbb{Q}} + \left\langle\phi_t L_t\right\rangle_{\mathbb{Q}}\, d\left(\frac{1}{Z_t}\right) + \left\langle d(\phi_t L_t)\right\rangle_{\mathbb{Q}}\, d\left(\frac{1}{Z_t}\right), \end{aligned} \tag{72}$$

where introduced the short-hand notation $\mathbb{E}_{\mathbb{Q}}[\cdot|Y_{0:t}] = \langle\cdot\rangle_{\mathbb{Q}}$ for the conditional expectation. We recall from Section 3 that for both of the signal models, we may write the time evolution of $\phi_t = \phi(X_t)$ as

$$d\phi_t = \mathcal{A}\phi_t dt + dM_t^{\phi}, \tag{73}$$

where we denote $\mathcal{A}\phi_t = \mathcal{A}\phi(X_t)$. M_t^{ϕ} is a martingale that is independent of the observations under $\mathbb{Q}$, and thus $\left\langle dM_t^{\phi}\right\rangle_{\mathbb{Q}} = 0$ as well as $\left\langle L_t\, dM_t^{\phi}\right\rangle_{\mathbb{Q}} = 0$. Therefore, we only retain the dt term under the conditional expectation. The first term in Eq. (72) can then be computed using Eq. (71):

$$\begin{aligned} \left\langle d(\phi_t L_t)\right\rangle_{\mathbb{Q}} &= \left\langle (d\phi_t) L_t + \phi_t\,(dL_t) + (d\phi_t)\,(dL_t)\right\rangle_{\mathbb{Q}}. \\ &= \left\langle L_t \mathcal{A}\phi_t\right\rangle_{\mathbb{Q}} dt + \left\langle \phi_t L_t \tilde{h}(X_t)^{\top} d\tilde{Y}_t\right\rangle_{\mathbb{Q}}, \end{aligned} \tag{74}$$

which is the SDE of the *unnormalized* posterior expectation. Here we further used that $\langle (d\phi_t)(dL_t)\rangle = 0$, because the noise in state and observations are independent.

Note that the evolution equation of the normalization constant Z_t in Eq. (65), $dZ_t = d\left\langle L_t\right\rangle_{\mathbb{Q}}$, corresponds to Eq. (74) with the constant function $\phi = 1$. The differential dZ_t^{-1} is obtained by consecutive application of Itô's lemma (Eq. (47)). By plugging Eq. (74) and dZ_t^{-1} into Eq. (72) and rewriting everything in terms of expectations under $\mathbb{P}$ using Eq. (65), one finally obtains the evolution equation for the posterior expectation, the so-called **Kushner–Stratonovich equation** (KSE, Bain & Crisan, 2009, p. 68, Theorem 3.30, cf. Appendix A.2 for calculation steps):

$$d\left\langle\phi_t\right\rangle_{\mathbb{P}} = \left\langle\mathcal{A}\phi_t\right\rangle_{\mathbb{P}} dt + \mathrm{cov}_{\mathbb{P}}(\phi_t, \tilde{h}(X_t)^{\top})(d\tilde{Y}_t - \langle\tilde{h}(X_t)\rangle_{\mathbb{P}}\, dt). \tag{75}$$

Equivalently, it can straightforwardly be expressed in terms of the original observation process in Eq. (59):

$$d\left\langle\phi_t\right\rangle_{\mathbb{P}} = \left\langle\mathcal{A}\phi_t\right\rangle_{\mathbb{P}} dt + \mathrm{cov}_{\mathbb{P}}(\phi_t, h(X_t)^{\top})\Sigma_y^{-1}(dY_t - \langle h(X_t)\rangle_{\mathbb{P}}\, dt). \tag{76}$$

In analogy to the calculations in Section 3, one may also pass from an evolution equation for the expectations to an adjoint equation for the conditional probability density,

$$\begin{aligned} dp(x|Y_{0:t}) &= \mathcal{A}^{\dagger} p(x|Y_{0:t})\, dt \\ &\quad + p(x|Y_{0:t})(h(x) - \langle h(X_t)\rangle)^{\top}\Sigma_y^{-1}(dY_t - \langle h(X_t)\rangle\, dt) \end{aligned} \tag{77}$$

Writing Eq. (75) and (77) in terms of the (adjoint of the) infinitesimal generator of the signal process, allows us to use any signal process for which $\mathcal{A}$ is known. For instance, if the signal process is a Markov chain on a finite set S, the expression $p(x|Y_{0:t})$ can be interpreted as the vector of posterior probabilities $\hat{p}(t)$, with entries $\hat{p}_i(t)$ denoting the probability to be in state i given all observations up to time t. The generator $\mathcal{A}^{\dagger}$ is then represented by the matrix $A^{\top}$ that has appeared in the evolution equation for the prior density, Eq. (41). Specifically, $\hat{p}_i(t)$, evolves as

$$\begin{aligned} d\hat{p}_i(t) &= \sum_{j=1}^{n} A_{ij}^{\top}\hat{p}_j(t)\, dt \\ &\quad + \hat{p}_i(t)(h_i - h\hat{p}(t))^{\top}\Sigma_y^{-1}(dY_t - h\hat{p}(t)\, dt), \end{aligned} \tag{78}$$

where $h_i = h(i) \in \mathbb{R}^l$, $i = 1, \ldots, n$ and h is an $l \times n$-matrix whose columns are the h_i's. Eq. (78) is known as the Wonham filter (Wonham, 1964), and it is a finite-dimensional SDE that completely solves the filtering problem.

Eq. (77) is a stochastic integro-differential equation, known as **Kushner equation** (KSE) (Kushner, 1962; Stratonovich, 1960), and its solution is in general infinite-dimensional. This fundamental problem is easily illustrated by considering the time evolution of the first moment, using $\phi(x) = x$:

$$d\langle X_t\rangle = \langle f(X_t)\rangle\, dt + \mathrm{cov}_{\mathbb{P}}(X_t, h(X_t)^{\top})\Sigma_y^{-1}(dY_t - \langle h(X_t)\rangle_{\mathbb{P}}\, dt). \tag{79}$$

For non-trivial (i.e. non-constant) observation functions h, any moment equation will depend on higher-order moments due to the posterior covariance between the observation function h and the function ϕ, which effectively amounts to a closure problem when $f(x)$ is nonlinear. This is not surprising; even the Fokker–Planck equation (58) (the evolution equation for the prior distribution) presents such a closure problem. In some cases (e.g. when using kernel or Galerkin methods), it is more convenient to use the evolution equation of the *unnormalized* posterior density $\varrho(X_t|Y_{0:t})$:

$$d\varrho(x|Y_{0:t}) = \mathcal{A}^{\dagger}\varrho(x|Y_{0:t})\, dt + \varrho(x|Y_{0:t})h(x)^{\top}\Sigma_y^{-1}dY_t, \tag{80}$$

which is a linear stochastic partial differential equation (SPDE), the **Zakai equation** (named after Zakai, 1969).

In very rare cases, under specific signal and observation models, the moment equations close, e.g. in the Kalman–Bucy filter (Kalman & Bucy, 1961, see Section 4.3 below) or the Beneš filter (Benes, 1981). Other finite-dimensional filters include the Daum filter (Daum, 1986) for continuous-time processes and discrete-time measurements. However, in most cases that occur in practice, the KSE needs to be approximated using a finite-dimensional realization. For instance, one could use the KSE as a starting point for these approximations, e.g. Markov-chain approximation methods (Kushner & Dupuis, 2001) or projection onto a finite-dimensional manifold (Brigo, Hanzon, & Le Gland, 1999; Brigo, Hanzon, & LeGland, 1998), which can be shown to be equivalent to assumed density filtering (ADF), or Galerkin-type methods with specific metrics and manifolds (see Armstrong & Brigo, 2013). Other numerical algorithms associated with overcoming the numerical burden of solving the Kushner or Zakai equation rely on a Fourier approximation of the involved densities, and the fact that convolutions correspond to simple products in Fourier space (Brunn, Sawo, & Hanebeck, 2006; Jia & Xin, 2010; Mikulevicius & Rozovskii, 2000).

4.3. A closed-form solution for a linear model: Kalman–Bucy filter

In models where the hidden drift function $f(X)$ and the observation function $h(X)$ are linear, i.e. $f(x) = Ax$ and $h(x) = Bx$, and the initial distribution is Gaussian, there exists a closed-form solution to the filtering problem. In this case, the KSE (Eq. (75)) closes with the second posterior moment Σ_t, i.e. the evolution equation for Σ_t becomes independent of the observations process, and the posterior density corresponds to a Gaussian with time-varying mean μ_t and variance Σ_t. The dynamics of these

[15] Recall that this corresponds to a Taylor expansion up to second order (i.e. first order in dt, since $\mathcal{O}(dW_t) = dt^{1/2}$) for diffusion processes, which is why we have to consider the product of differentials (product rule for stochastic differentials).

parameters are given by the **Kalman–Bucy filter** (KBF, Kalman & Bucy, 1961) and form a set of coupled SDEs:

$$d\mu_t = A\mu_t\,dt + \Sigma_t B^\top \Sigma_y^{-1}\big(dY_t - B\mu_t\,dt\big), \tag{81}$$

$$d\Sigma_t = \Big(A\Sigma_t + \Sigma_t A^\top + \Sigma_x - \Sigma_t B^\top \Sigma_y^{-1} B\Sigma_t\Big)\,dt. \tag{82}$$

The posterior variance follows a differential Riccati equation and, due to its independence from observations as well as from the posterior mean, it can be solved offline.

4.4. Filtering equations for observations corrupted by Poisson noise

In analogy to the previous section, the formal solution to the filtering problem with observations corrupted by Poisson noise (Eq. (61)) can also be derived with the change of probability measure method. We will very briefly outline the derivation, referring to similarities to continuous-time derivations.[16]

The idea is again to make use of the Kallianpur–Striebel formula (Eq. (65)) and to rewrite the posterior expectation under the original measure $\mathbb{E}_{\mathbb{P}}[\phi_t|N_{0:t}]$ in terms of an expectation under a reference measure $\mathbb{Q}$, under which hidden process X_t and observation process N_t are decoupled. Using a Girsanov-type theorem for point processes (see Bremaud, 1981, Chapter VI, Theorems T2 and T3), the measure is changed to the reference measure $\mathbb{Q}$ under which all point processes have a constant rate λ_0. The Radon–Nikodym derivative reads (cf. Eq. 69)

$$L_t = \prod_{i=1}^{l} \exp\left(\int_0^t \log\frac{h_i(X_s)}{\lambda_0} dN_s^i + \int_0^t (\lambda_0 - h_i(X_s))\,ds\right), \tag{83}$$

which solves the SDE

$$dL_t = L_t \cdot \sum_{i=1}^{l}\left(\frac{h_i(X_t)}{\lambda_0} - 1\right)(dN_t^i - \lambda_0 dt). \tag{84}$$

We can now repeat the calculations of the previous section. First, we obtain

$$\begin{aligned}\langle d(\phi_t L_t)\rangle_{\mathbb{Q}} &= \langle (d\phi_t)\,L_t + \phi_t\,(dL_t) + (d\phi_t)(dL_t)\rangle_{\mathbb{Q}} &(85)\\ &= \langle \mathcal{A}\phi_t\,L_t\rangle_{\mathbb{Q}}\,dt \\ &\quad + \left\langle \phi_t L_t \cdot \sum_{i=1}^{l}\left(\frac{h_i(X_t)}{\lambda_0} - 1\right)\right\rangle_{\mathbb{Q}} (dN_t^i - \lambda_0 dt). &(86)\end{aligned}$$

Here, we used again that under $\mathbb{Q}$, differentiation and expectation can be interchanged.

Using $\phi_t = 1$ gives us the evolution of the time-dependent normalization Z_t in the Kallianpur–Striebel formula (65). We can use these, together with Itô's lemma (Eq. (47)) to compute $dZ_t^{-1} = d\,\langle L_t\rangle_{\mathbb{Q}}^{-1}$, to obtain a point-process observations analogue to the KSE for the normalized posterior estimate:[17]

$$\begin{aligned} d\,\langle\phi_t\rangle_{\mathbb{P}} &= d\left(Z_t^{-1}\,\langle L_t\phi_t\rangle_{\mathbb{Q}}\right) \\ &= \frac{1}{Z_t}\,\langle d(\phi_t L_t)\rangle_{\mathbb{Q}} + \langle\phi_t L_t\rangle_{\mathbb{Q}}\,d\left(\frac{1}{Z_t}\right) + \langle d(\phi_t L_t)\rangle_{\mathbb{Q}}\,d\left(\frac{1}{Z_t}\right) \\ &= \langle\mathcal{A}\phi_t\rangle_{\mathbb{P}}\,dt + \sum_{i=1}^{l}\frac{\mathrm{cov}_{\mathbb{P}}(\phi_t, h_i(X_t))}{\langle h_i(X_t)\rangle_{\mathbb{P}}}\left(dN_t^i - \langle h_i(X_t)\rangle_{\mathbb{P}}\,dt\right) \\ &= \langle\mathcal{A}\phi_t\rangle_{\mathbb{P}}\,dt \\ &\quad + \mathrm{cov}_{\mathbb{P}}(\phi_t, h(X_t)^\top)\,\mathrm{diag}(\langle h(X_t)\rangle_{\mathbb{P}})^{-1} \\ &\quad \times\,(dN_t - \langle h(X_t)\rangle_{\mathbb{P}}\,dt)\,, \end{aligned} \tag{87}$$

where $\mathrm{diag}(x)$ denotes a diagonal matrix with diagonal entries given by the vector x. The adjoint form of this equation, i.e. the evolution equation for the posterior density $p(x|N_{0:t})$, reads:

$$\begin{aligned} dp(x|N_{0:t}) &= \mathcal{A}^\dagger p(x|N_{0:t})\,dt + \\ &\quad p(x|N_{0:t})\sum_{i=1}^{l}\frac{1}{\langle h_i(X_t)\rangle}\left(h_i(x) - \langle h_i(X_t)\rangle\right) \\ &\quad \times\,(dN_t^i - \langle h_i(X_t)\rangle\,dt) \\ &= \mathcal{A}^\dagger p(x|N_{0:t})\,dt + \\ &\quad p(x|N_{0:t})\,(h(x) - \langle h(X_t)\rangle)^\top\,\mathrm{diag}\,(\langle h(X_t)\rangle)^{-1} \\ &\quad \times\,(dN_t - \langle h(X_t)\rangle\,dt). \end{aligned} \tag{88}$$

Note the structural similarity to the Kushner equation (Eq. (77)): it also relies on a Fokker–Planck term denoting the prediction, and a correction term that is proportional to the posterior density, the *innovation* $dN_t - \langle h(X_t)\rangle\,dt$, as well as a local correction $h(x) - \langle h(X_t)\rangle$. The difference is that the observation noise covariance Σ_y in the Kushner equation has been replaced by a diagonal matrix whose components are proportional to the rate function in each observable dimension. Considering that the observations are Poisson processes, this is not surprising: for Poisson processes, the variance is proportional to the instantaneous rate, and thus, analogously, the correction term in this equation has a similar proportionality.

Similarly, we find for the unnormalized posterior density $\varrho(x|N_{0:t})$:

$$d\varrho(x|N_{0:t}) = \mathcal{A}^\dagger\varrho(x|N_{0:t})\,dt + \varrho(x|N_{0:t})\frac{1}{\lambda_0}\,(h(x) - \lambda_0)^T\,(dN_t - \lambda_0 dt). \tag{89}$$

Analogously to Eqs. (77) and (80), these equations are obtained by integrating the equation for the unnormalized posterior estimate (Eq. (86)) and the normalized posterior estimate (Eq. (87)) twice.

4.5. Down to Earth - an example from decision making

In this section, we have seen that the filtering equations can be derived by changing from the original measure to a measure under which the signal and observation processes are independent. Interestingly, what we have seen here also gives us a recipe of how to treat a filtering problem in general: all we need is a characterization of the state dynamics in terms of the infinitesimal generator $\mathcal{A}$. Further, any information about the observations is carried by the Radon–Nikodym derivative.

To illustrate this, let us consider the following example from dynamical decision making. In order to make a decision that relies on an unobserved, but dynamic, variable, animals have to be able to integrate the (noisy) observations about this variable in a way that discounts older evidence in favor of more recent evidence, depending on the dynamics of the variable. In other words: they have to do filtering. In Piet et al. (2018) an experimental paradigm is considered where the hidden variable can be in one of two states, i.e. $X_t \in \{S_1, S_2\}$, which switches states with a hazard rate a and which influences the click rates r_+ and r_- of two speakers. More precisely, whenever the system is in state S_1, speaker 1 will have click rate r_+ and speaker 2 will have click rate r_-, and vice versa if the system is in state S_2 (Fig. 4a). Presented with clicks from these speakers, rats have to make a decision about the state the environment is in at the end of the trial. The optimal solution to the problem is then the filtering distribution conditioned on the clicks at the end of the trial, and the optimal decision corresponds to the state with the highest

[16] A very detailed derivation is offered in Bremaud (1981, p. 170ff.) or, more intuitively, in Bobrowski, Meir, and Eldar (2009, SI) and in Surace (2015, p. 41ff).

[17] See Appendix A.3 for detailed derivation steps.

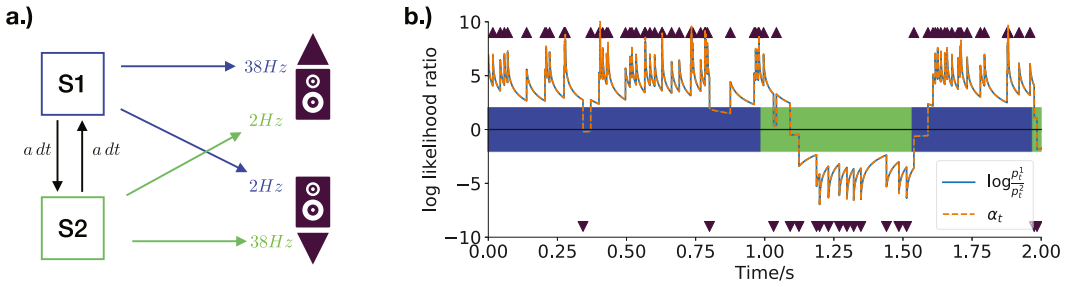

Fig. 4. A hidden Markov model with point process observations as an example for a dynamical decision making paradigm. (a) Schematic of the HMM underlying the experiments in Piet et al. (2018). (b) Logarithm of the likelihood ratio α_t, which serves as the decision variable, as a function of time. The sign of the decision variable at the end of the trial denotes the optimal decision.

posterior probability. This study has shown that rats are able to perform optimal evidence discounting.[18]

To come up with the solution for the filtering task, Piet et al. (2018) consider the evolution of the log-likelihood as the decision variable (based on Veliz-Cuba et al., 2016) and derive this by taking the continuum-limit of the discrete-time process. In principle, this approach is perfectly equivalent to a change of measure in the log domain. Here, we will re-derive their result for the optimal solution of the filtering task by directly applying our 'recipe'.

Let us first consider the state process, which is a discrete-state Markov model. Without observations, the probabilities of the hidden states evolve as $d\tilde{p}_t = \mathcal{A}^\dagger \tilde{p}_t\, dt$, where $\tilde{p}_t^{(i)} = p(X_t = S_i)$. It is easy to check that for this model the adjoint generator matrix is given by

$$\mathcal{A}^\dagger = \begin{pmatrix} -a & a \\ a & -a \end{pmatrix}, \tag{90}$$

where a denotes the switching rate, i.e. $a\,dt$ is the probability of switching the state.

The observation model is a two-dimensional point process, and N_t^j denotes the number of clicks in speaker j up to time t. Let $h_{ji} = h_j(X = S_i)$ be the rate of speaker j if the hidden variable is in state S_i. Thus, the evolution of the posterior probability p_t is given by (cf. Eq. (88)):

$$dp_t^{(i)} = (\mathcal{A}^\dagger p_t)_i\, dt + p_t^{(i)} \sum_{j=1}^{2} (h_{ji} - \langle h_j \rangle)\langle h_j \rangle^{-1} (dN_t^j - \langle h_j \rangle\, dt), \tag{91}$$

where $\langle h_j \rangle = \sum_i h_{ji} p_t^{(i)}$. Since this particular system is 2-dimensional, i.e. $p_t^{(2)} = 1 - p_t^{(1)}$, we can substitute this in the expression for the first component of the posterior and get a one-dimensional equation for $p_t^{(1)}$.

$$\begin{aligned} dp_t^{(1)} &= a(1 - 2p_t^{(1)})\, dt \\ &\quad + p_t^{(1)} \sum_{j=1}^{2} (h_{j1} - h_{j2})(1 - p_t^{(1)}) \left(1/\langle h_j \rangle\, dN_t^j - dt \right), \end{aligned} \tag{92}$$

where $\langle h_j \rangle = h_{j1} p_t^{(1)} + h_{j2}(1 - p_t^{(1)})$. At the end of the trial, S_1 is the optimal choice whenever $p_t^{(1)} > 1/2$.

For several reasons, it might be desirable to define the decision variable as the log likelihood ratio of being in state S_1 as opposed to being in S_2. Let $\alpha_t = \log \frac{p_t^{(1)}}{p_t^{(2)}} = \log \frac{p_t^{(1)}}{1 - p_t^{(1)}}$. In order to derive its evolution equation, we can directly apply Itô's lemma for point processes to Eq. (92) with $\phi(x) = \frac{x}{1-x}$ and after straightforward, but tedious, algebra arrive at the desired SDE.

$$\begin{aligned} d\alpha_t &= -2a \sinh \alpha_t\, dt \\ &\quad + \sum_{j=1}^{2} \left[(h_{j2} - h_{j1})\, dt + \log \frac{h_{j1}}{h_{j2}}\, dN_t^j \right]. \end{aligned} \tag{93}$$

Note that if $h_{11} - h_{12} = h_{22} - h_{21} = r_+ - r_-$ (which is the experimental setting in Piet et al., 2018), this equation becomes very simple:

$$d\alpha_t = -2a \sinh \alpha_t\, dt + \log \frac{r_+}{r_-} (dN_t^{(1)} - dN_t^{(2)}), \tag{94}$$

and resembles Eq. (9) in Piet et al. (2018). Without this symmetry, there is a drift term modulated by the difference in click rates, indicating that the *absence* of clicks is informative for estimating the current state. In Fig. 4, we plotted the log likelihood ratio, both computed from the posterior probabilities as in Eq. (91), and directly from running Eq. (94). Unsurprisingly, both plots lie exactly on top of each other.

Note that this result was obtained by simply plugging in a model, i.e. the signal and the observation process, into the solution to the filtering equation, and making use of the fact that for a finite state space, the Kushner equation for point processes, Eq. (88), becomes finite-dimensional. Unlike in Piet et al. (2018) or in Veliz-Cuba et al. (2016), we did not have to explicitly carry out the continuum limit - in fact, this is implicitly taken care of by using the appropriate Radon–Nikodym derivative for this observation model. This allows for much more flexibility when the models and/or experimental settings become more complicated, for instance if we want to increase the number of states, modify the state dynamics or modify the properties of the speakers.

A fully annotated code for this example is available in our github repository (Kutschireiter, 2019).

5. Don't panic: Approximate closed-form solutions

> *"It is a mistake to think you can solve any major problems just with potatoes".*
>
> [Douglas Adams]

If the signal model is a jump–diffusion, the KSE (Eqs. (75), (87)) is infinite-dimensional, a fact known as 'closure problem'. In other words, except for some important exceptions for very specific models, such as the KBF or the Beneš filter (Benes, 1981), solutions to the general filtering problems are not analytically accessible. Furthermore, unlike for observations following a diffusion process, no closed-form filter for point-process observations is known. However, there exist important *approximate* closed-form solutions, which address the closure problem by approximating the posterior density in terms of a set number of sufficient statistics.

[18] We are glad they did not use mice, as these animals, according to the Hitchhiker's guide, are the most intelligent species on planet Earth and as such would surely have outperformed the optimal solution. Rats are close enough, though.

Here, we will briefly outline some important examples: first, the Extended Kalman–Bucy Filter and related methods for point-process observations that rely on a series expansion of the functions in the generative model, such that the posterior is approximated by a Gaussian density. We will further describe assumed density filters, that choose a specific form of the posterior and propagate the KSE according to this approximation.

5.1. The extended Kalman–Bucy filter and related approaches

Based on the Kalman–Bucy filter (Section 4.3), the **extended Kalman–Bucy filter (EKBF)** is an approximation scheme for nonlinear generative models of the form

$$dX_t = f(X_t)dt + \Sigma_x^{1/2} dW_t$$
$$dY_t = h(X_t)dt + \Sigma_y^{1/2} dV_t.$$

The EKBF approximates the posterior by a Gaussian with mean μ_t and variance Σ_t, whose dynamics are derived by local linearization (around the mean) of the nonlinearities in the model (Jazwinski, 1970, p. 338, Example 9.1):

$$d\mu_t = f(\mu_t)\,dt + \Sigma_t H^\top(\mu_t)\Sigma_y^{-1}\big(dY_t - h(\mu_t)\,dt\big), \tag{95}$$

$$d\Sigma_t = \Big(F(\mu_t)\Sigma_t + \Sigma_t F(\mu_t)^\top + \Sigma_x - \Sigma_t H^\top(\mu_t)\Sigma_y^{-1} H(\mu_t)\Sigma_t \Big)\, dt, \tag{96}$$

where $F_{ij} = \frac{\partial f_i}{\partial x_j}$ and $H_{ij} = \frac{\partial h_i}{\partial x_j}$ denote the Jacobian of the hidden drift function and the observation function, respectively. For models with multimodal posteriors, this approximation often breaks down: e.g. if the noise covariance Σ_y is large, the mean of the EKBF tends to 'get stuck' in one of the modes.

Similar approximations exist for point-process observations. One way to achieve this would be to simply construct an EKBF by assuming *Gaussian* noise in the observations, together with the appropriate linearization (see paragraph below Eq. (17) in Eden, 2007). Another way that allows the point-process observations to directly enter the expressions for mean and variance relies on a Taylor expansion in the log domain of the approximated posterior up to second order (see Eden, Frank, Barbieri, Solo, & Brown, 2004 for discrete-time and Eden & Brown, 2008 for the continuous-time models). The continuous-time approximate filter for point processes can be seen as the **point-process analogue of the EKBF** (cf. Eden & Brown, 2008, extended by nonlinearity in the hidden state process):

$$d\mu_t = f(\mu_t)\,dt + \Sigma_t \sum_{i=1}^{l} (\nabla_x \log h_i(x))|_{x=\mu_t} \big(dN_t^i - h_i(\mu_t)\,dt\big), \tag{97}$$

$$d\Sigma_t = \Big(F(\mu_t)\Sigma_t + \Sigma_t F(\mu_t)^\top + \Sigma_x\Big)\,dt - \Sigma_t \sum_{i=1}^{l} \left(\frac{\partial^2 h_i(x)}{\partial x \partial x^\top}\bigg|_{x=\mu_t} dt + S_i\, dN_t^i \right) \Sigma_t, \tag{98}$$

with

$$S_i = \begin{cases} \left(\Sigma_t - \left(\frac{\partial^2 \log h_i(x)}{\partial x \partial x^\top}\Big|_{x=\mu_t} \right)^{-1} \right)^{-1} & \text{if } \frac{\partial^2 \log h_i(x)}{\partial x \partial x^\top}\Big|_{x=\mu_t} \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{99}$$

5.2. Assumed density filtering

The idea of **assumed density filters** (ADF) is to specify a set of sufficient statistics, which is supposed to approximate the posterior density, derive evolution equations from the KSE, i.e. from Eqs. (75) and (87), and approximate expectations within these evolution equations under the initial assumptions. To be less abstract, consider approximating the posterior density by a Gaussian density. Then it suffices to derive evolution equations for mean μ_t and variance Σ_t of the approximated Gaussian posterior. In these evolution equations, higher-order moments will enter, which in turn can be expressed in terms of mean and variance for a Gaussian.

As a concrete example, let us consider a Gaussian ADF for point-process observations (the treatment for diffusion-observations is completely analogous). Consider the SDEs for the first two moments of the posterior (cf. Eq. (87), detailed derivation in Appendix A.4):

$$d\mu_t = \langle f(X_t)\rangle\, dt + \mathrm{cov}(X_t, h(X_t)^\top)\,\mathrm{diag}\,(\langle h(X_t)\rangle)^{-1} \times\ (dN_t - \langle h_t\rangle dt)\,, \tag{100}$$

$$d\Sigma_t = \Big(\mathrm{cov}(f(X_t), X_t^\top) + \mathrm{cov}(X_t, f(X_t)^\top) + \Sigma_x\Big)\,dt + \sum_{i=1}^{l} \frac{1}{\langle h_i(X_t)\rangle} \Big[\mathrm{cov}(h_i(X_t), X_t X_t^\top) - \mathrm{cov}(h_i(X_t), X_t)\mu_t^\top - \mu_t \mathrm{cov}(h_i(X_t), X_t^\top)\Big] \times \big(dN_t^i - \langle h_i(X_t)\rangle dt\big) - \sum_{i=1}^{l} \frac{1}{\langle h_i(X_t)\rangle^2} \mathrm{cov}(h_i(X_t), X_t)\mathrm{cov}(h_i(X_t), X_t^\top) dN_t^i. \tag{101}$$

The effective realization of the ADF will crucially depend on the specifics of the signal model defined by Eq. (43) and the observation model (61), respectively. For example, Pfister, Dayan, and Lengyel (2009) consider an exponential rate function $h(x) \propto \exp(\beta x)$, which leads to a variance update term that is independent of the spiking process. Of particular interest for decoding tasks in neuroscience are ADFs with Gaussian-shaped rate function (e.g. Harel, Meir, & Opper, 2018).

However, for some models ADFs cannot be computed in closed form. Consider for simple example a rate function that is a soft rectification of the hidden process, e.g. $h(x) = \log(\exp(x) + 1)$, which, when taking expectations with respect to a Gaussian, does not admit a closed-form expression in terms of mean μ_t and variance Σ_t.

6. Approximations without infinite improbability drive: Particle methods

> *"Each particle of the computer, each speck of dust*
> *held within itself, faintly and weakly, the pattern of the whole".*
>
> [Douglas Adams]

Particle filtering (PF) is a numerical technique to approximate solutions to the filtering problem by a finite number of samples, or 'particles', from the posterior. Thus, they serve as a finite-dimensional approximation of the KSE, overcoming the closure problem. The true posterior is approximated by the empirical distribution formed by the particle states $X_t^{(i)}$, i.e. a sum of Dirac-delta functions at the particle positions $\delta(x - X_t^{(i)})$, and, if it is a weighted PF, weighted by their corresponding importance weights $w_t^{(i)}$,

$$p(x|Y_{0:t}) \approx \sum_{i=1}^{M} w_t^{(i)} \delta(x - X_t^{(i)}), \tag{102}$$

with $\sum_i w_t^{(i)} = 1$ ensuring normalization. Consequently,

$$\mathbb{E}_{\mathbb{P}}\left[\phi(X_t)|Y_{0:t}\right] \approx \sum_{i=1}^{M} w_t^{(i)} \phi(X_t^{(i)}). \tag{103}$$

The rationale is based on a similar idea as using the Euler–Maruyama scheme to numerically solve the Fokker–Planck equation and its associated equation for the posterior moments. As a numerical recipe (for instance provided by Doucet et al., 2000; Doucet & Johansen, 2009, for discrete-time models), it is easily accessible, because in principle no knowledge of the Fokker–Planck equation, nonlinear filtering theory or numerical methods for solving partial differential equations is needed.

In this section, weighted particle filters will be introduced from a continuous-time perspective based on the change of probability measure formalism (roughly following Bain & Crisan, 2009, Chapt. 9.1, and extending this to point-process observations). From this formalism, we derive dynamics for the weights and link these to the 'curse of dimensionality'. Finally, to give context for readers more familiar with discrete-time particle filtering, the continuous-time perspective will be linked to the 'standard' particle filter (PF).

6.1. Particle filtering in continuous time

Based on sequential importance sampling, samples (or 'particles') $X_t^{(i)}$ as well as their respective weights are propagated through time. As we have seen before in Section 2.1.1, importance sampling amounts to a change of measure from the original measure $\mathbb{P}$ to a reference measure $\mathbb{Q}$, from which sampling is feasible. Here, the idea is to change to a measure under which the observation processes are independent of the hidden process, effectively enabling us to sample from the hidden process. In other words, the particle positions evolve as specified by the infinitesimal generator $\mathcal{A}$ of the hidden process (e.g. Eq. (43) if the hidden process is a jump–diffusion process).

Why this should be the case is rather intuitive when recalling the Kallianpur–Striebel formula (65):

$$\mathbb{E}_{\mathbb{P}}[\phi_t|Y_{0:t}] = \frac{1}{Z_t}\mathbb{E}_{\mathbb{Q}}[\phi_t\, L_t|Y_{0:t}].$$

If we want to approximate the left-hand side of this equation with empirical samples, it would require us to have access to samples from the real posterior distribution, which is usually not the case. However, since under the measure $\mathbb{Q}$ on the right-hand side the hidden state and observations are decoupled, the estimate is approximated by empirical samples that correspond to realizations of the hidden process:

$$\frac{1}{Z_t}\mathbb{E}_{\mathbb{Q}}[\phi_t\, L_t|Y_{0:t}] \approx \frac{1}{\bar{Z}_t}\sum_{i=1}^{M}\phi(X_t^{(i)})L_t(X_t^{(i)}). \tag{104}$$

$\bar{Z}_t = \sum_{i=1}^{M} L_t(X_t^{(i)})$ is an empirical estimate of the normalization constant.

Thus, we just need to evaluate the Radon–Nikodym derivative at the particle states $X_t^{(i)}$, giving us the importance weight $w_t^{(i)}$ of particle i at time t. For observation corrupted by Gaussian noise (cf. Eq. (70)), this reads:

$$w_t^{(i)} = \frac{1}{\bar{Z}_t}L_t(X_t^{(i)}) \tag{105}$$

$$= \frac{1}{\bar{Z}_t}\exp\left[\int_0^t h(X_s^{(i)})^\top \Sigma_y^{-1} dY_s - \frac{1}{2}\int_0^t h(X_s^{(i)})^\top \Sigma_y^{-1} h(X_s^{(i)})\, ds\right]. \tag{106}$$

Similarly, we find for point-process observations (cf. Eq. (83))

$$w_t^{(i)} = \frac{1}{\bar{Z}_t}\prod_{j=1}^{l}\exp\left(\int_0^t \log\frac{h_j(X_s^{(i)})}{\lambda_0}dN_s^j + \int_0^t(\lambda_0 - h_j(X_s^{(i)}))\, ds\right). \tag{107}$$

6.1.1. Weight dynamics in continuous time

If one is interested how the weight of particle i changes over time, it is possible to derive an evolution equation for the particle weights. Using Itô's lemma, we find:

$$dw_t^{(i)} = d\left(\frac{L_t(X_t^{(i)})}{\bar{Z}_t}\right) = \bar{Z}_t^{-1}dL_t^{(i)} + L_t^{(i)}d\bar{Z}_t^{-1} + dL_t^{(i)}\, d\bar{Z}_t^{-1}, \tag{108}$$

For continuous-time observations, (cf. Eq. (71)) yields

$$dL_t^{(i)} = L_t^{(i)}(h(X_t^{(i)}))^\top \Sigma_y^{-1} dY_t, \tag{109}$$

$$d\bar{Z}_t = \sum_{i=1}^{M} dL_t^{(i)} = \bar{Z}_t\,(\bar{h}_t)^\top \Sigma_y^{-1} dY_t, \tag{110}$$

where $\bar{h}_t := \sum_i w_t^{(i)} h(X_t^{(i)}) = \bar{Z}_t^{-1}\sum_i L_t^{(i)} h(X_t^{(i)})$ is the weighted estimate of the observation function h_t (i.e. under the original measure $\mathbb{P}$). Applying Itô's lemma on Eq. (110) to obtain $d\bar{Z}_t^{-1}$, we find for the dynamics of the weights

$$dw_t^{(i)} = w_t^{(i)}\left(h(X_t^{(i)}) - \bar{h}_t\right)^\top \Sigma_y^{-1}(dY_t - \bar{h}_t dt) \tag{111}$$

Similarly, with Eq. (84) we find for point-process observations:

$$dL_t^{(i)} = L_t^{(i)}\sum_{j=1}^{l}\frac{1}{\lambda_0}\left(h_j(X_t^{(i)}) - \lambda_0\right)\left(dN_t^j - \lambda_0\, dt\right), \tag{112}$$

$$d\bar{Z}_t = \bar{Z}_t\sum_{j=1}^{l}\frac{1}{\lambda_0}(\bar{h}_j - \lambda_0)\left(dN_t^j - \lambda_0\, dt\right), \tag{113}$$

and thus, using Itô's lemma for point processes to obtain $d\bar{Z}_t^{-1}$, with Eq. (108):

$$dw_t^{(i)} = w_t^{(i)}\sum_{j=1}^{l}\frac{1}{\bar{h}_{j,t}}\left(h_j(X_t^{(i)}) - \bar{h}_{j,t}\right)\left(dN_t^j - \bar{h}_{j,t}dt\right) \tag{114}$$

$$= w_t^{(i)}\left(h(X_t^{(i)}) - \bar{h}_t\right)^\top \mathrm{diag}(\bar{h}_t)^{-1}\left(dN_t - \bar{h}_t dt\right). \tag{115}$$

Interestingly, there is a striking similarity to the dynamics of the importance weights and the Kushner equation (77) and the point-process observation equivalent of the Kushner equation (Eq. (88)), respectively. The weight dynamics seem to directly correspond to the dynamics of the correction step (with true posterior estimates replaced by their empirical counterparts). This is rather intuitive: since we chose a change of measure under which the particles follow the hidden dynamics, serving as the prediction step, the observation dynamics have to be fully accounted for by the weight dynamics, in a way to be consistent with the Kushner equation.

It is important to note that the weight dynamics inevitably lead to a system in which all but one weight equals zero. This is called weight degeneracy. In this degenerate state, the particle system cannot represent the posterior sufficiently, and hence has to be avoided numerically, e.g. by resampling the particles from the weight distribution and resetting the weights to $1/M$. The time scale on which this weight degeneracy happens depends on the number of observable dimensions, in other words it is accelerated as the dimensionality of the system is increased (Surace, Kutschireiter, & Pfister, 2019b). This is a form of the so-called 'curse of dimensionality', a common nuisance in weighted particle filters.

6.1.2. Equivalence between continuous-time particle filtering and bootstrap particle filter

The practitioner who is using PF algorithms in their numerical implementations might usually be more familiar with the

discrete-time formulation. Further, since the continuous-time formulation of the particle filter based on the measure change formalism seems to be so different from the discrete-time formulation, one might rightfully ask how these two concepts are related and whether they are equivalent. Indeed, we will now quickly show that the continuous-time PF in Section 6.1 corresponds to the Bootstrap PF in a continuous-time limit. More precisely, if we apply the Bootstrap PF to a time-discretized version of our hidden state process model and observation model, and then take the continuum limit, we will regain the equations for the weights as in Section 6.1.

Irrespectively of the generator $\mathcal{A}$ of the hidden process, it is straightforward to write the hidden process in terms of a transition density, with $t-dt$ corresponding to the previous time step. This acts as the proposal density $\pi(X_t|X^{(i)}_{0:t-dt}, Y_{0:t}) = p(X_t|X^{(i)}_{t-dt})$. Consider for example a drift–diffusion process (Eq. (43) with $J=0$). Then the particles are sampled from the time-discretized transition density, $X^{(i)}_t \sim p(X_t|X^{(i)}_{t-dt})$, which is given by[19]:

$$p(X_t|X^{(i)}_{t-dt}) = \mathcal{N}\Big(X_t; X^{(i)}_{t-dt} + f(X^{(i)}_{t-dt})\,dt,\, \Sigma_x\,dt\Big). \tag{116}$$

For observations corrupted by Gaussian noise, the emission likelihood is given by the emission probability for the *instantaneous increments* dY_t in Eq. (43), i.e.

$$p(dY_t|X^{(i)}_t) = \mathcal{N}\Big(dY_t; h(X^{(i)}_t)\,dt,\, \Sigma_y\,dt\Big), \tag{117}$$

such that the unnormalized weights are given by

$$\tilde{w}^{(i)}_t = \tilde{w}^{(i)}_{t-dt}\,\mathcal{N}\Big(dY_t; h(X^{(i)}_t)\,dt,\, \Sigma_y\,dt\Big). \tag{118}$$

It is evident that the proposal of continuous-time particle filtering and that of the BPF match, and it remains to show that the same holds for the importance weights. In other words, when taking the continuum limit of Eq. (118), we should be able to recover Eq. (106). Keeping only terms up to $\mathcal{O}(dt)$, we find

$$\begin{aligned} &p(dY_t|X^{(i)}_t) \\ &\propto \exp\left(-\frac{1}{2}(dY_t - h(X^{(i)}_t)dt)^\top(\Sigma_y dt)^{-1}(dY_t - h(X^{(i)}_t)dt)\right) \\ &\propto \exp\left(h(X^{(i)}_t)^\top\Sigma_y^{-1}dY_t - \frac{1}{2}h(X^{(i)}_t)^\top\Sigma_y^{-1}h(X^{(i)}_t)dt\right), \end{aligned} \tag{119}$$

where the term $\propto (dY_t)^2$ was absorbed in the normalization because it is independent of the particle positions. Thus, the continuous-time limit $dt \to 0$ of Eq. (35) reads

$$\begin{aligned} \tilde{w}^{(i)}_t \;\propto\; \prod_{s=0}^{t}\tilde{w}^{(i)}_s &= \prod_{s=0}^{t}\exp\Bigg((h(X^{(i)}_s))^\top\Sigma_y^{-1}dY_s \\ &\quad - \frac{1}{2}h(X^{(i)}_s)^\top\Sigma_y^{-1}h(X^{(i)}_s)ds\Bigg) \\ &\to \exp\left(\int_0^t h(X^{(i)}_s)^\top\Sigma_y^{-1}dY_s\right. \\ &\quad \left. - \frac{1}{2}\int_0^t h(X^{(i)}_s)^\top\Sigma_y^{-1}h(X^{(i)}_s)ds\right), \end{aligned} \tag{120}$$

which, up to the normalization constant $\bar{Z}_t$, is equivalent to Eq. (106).

For point-process observations, the emission likelihood is given by $p(dN_t|X_t)$, which is defined by the Poisson density in Eq. (61). Neglecting the term that is independent of the particle positions (which is absorbed in the normalization), it can be rewritten as:

$$\begin{aligned} p(dN_t|X^{(i)}_t) &= \prod_j \text{Poisson}(dN^j_t; h_j(X^{(i)}_t)dt) \\ &= \prod_j \frac{1}{dN^j_t!}\exp\Big(-h_j(X^{(i)}_t)dt + dN^j_t\log(h_j(X^{(i)}_t)dt)\Big) \\ &\propto \prod_j \exp\Big(\log h_j(X^{(i)}_t)dN^j_t - h_j(X^{(i)}_t)dt\Big). \end{aligned} \tag{121}$$

Again, since $\tilde{w}^{(i)}_t \propto \prod_s p(dN_s|X^{(i)}_s)$, the continuous-time limit of the unnormalized importance weight is

$$\tilde{w}^{(i)}_t \to \prod_j \exp\left(\int_0^t \log h_j(X^{(i)}_s)dN^j_s - h_j(X^{(i)}_s)ds\right). \tag{122}$$

The explicit dependence of Eq. (107) on the reference rate λ_0 can be absorbed in the normalization constant, yielding equivalent expressions for the normalized weights $w^{(i)}_t$.

6.2. The feedback particle filter

In contrast to weighted particle filtering approaches, unweighted approaches for particle filtering exist, for example the Ensemble Kalman (Bucy) Filter (Bergemann & Reich, 2012; Evensen, 1994), the Feedback Particle Filter (FBPF, Yang, Blom, & Mehta, 2014; Yang, Mehta, & Meyn, 2013), the (stochastic) particle flow filter (Daum, Huang, & Noushin, 2010; de Melo, Maskell, Fasiolo, & Daum, 2015) or the point-process analogue to the FBPF (Surace, Kutschireiter, & Pfister, 2019a). Since these methods do not rely on importance weights in the first place, there is no weight degeneracy. Unweighted particle filters therefore hold the promise of avoiding the curse of dimensionality (see Surace et al., 2019b).

All of these methods have in common that the posterior is approximated by equally weighted particles, i.e.:

$$p(x|Y_{0:t}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(x - X^{(i)}_t). \tag{123}$$

Consequently, the observations have to enter the particle dynamics in such a way that the particles are moved towards regions of high posterior density. As an example, we will outline how this is achieved in the Feedback particle filter.

In the FBPF, the observations directly enter the particle dynamics, which evolve according to the Itô SDE:

$$\begin{aligned} dX^{(i)}_t &= \Big(f(X^{(i)}_t, t) + \Omega(X^{(i)}_t, t)\Big)\,dt + G(X^{(i)}_t, t)^{1/2}dB^{(i)}_t \\ &\quad + K(X^{(i)}_t, t)\Sigma_y^{-1}\left[dY_t - \frac{1}{2}\Big(h(X^{(i)}_t) + \bar{h}_t\Big)\,dt\right], \end{aligned} \tag{124}$$

where $B^{(i)}_t$ are uncorrelated vector Brownian motion processes, $K(X^{(i)}_t, t)$ is the gain matrix, and $\bar{h}_t = \frac{1}{N}\sum_{i=1}^{N}h(X^{(i)}_t)$ denotes the particle estimate of the observation function. The components of the additional vector-valued drift function $\Omega(X^{(i)}_t, t)$ are given by (Yang, Laugesen, Mehta, & Meyn, 2016)

$$\Omega_l(x,t) = \frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{m}K_{jk}(x,t)\frac{\partial K_{lk}}{\partial x_k}(x,t). \tag{125}$$

The gain K is the solution of a boundary value problem that emerges from an optimal control problem (Yang et al., 2016, Eqs. 4, 5). It is chosen such that it minimizes the Kullback–Leibler divergence between the particle distribution and the posterior

[19] Remark: The so-called Euler–Maruyama scheme for numerical implementation of diffusion processes is based on the very same discretization.

filtering distribution (conditioned on proper initialization), which leads to the following conditions:

$$\nabla \cdot \left(p(x,t|Y_{0:t})\nabla\psi_j(x,t)\right) = -\left(h_j - \bar{h}_j\right) p(x,t|Y_{0:t}) \tag{126}$$

$$\int \psi_j(x,t)p(x,t|Y_{0:t})dx = 0, \tag{127}$$

with

$$K_{ij} = \frac{\partial \psi_j}{\partial x_i}. \tag{128}$$

In general, the gain matrix $K(x,t)$ cannot be solved for in closed form, and in practical implementations relies on a numerical solution of the Euler–Lagrange boundary value problem (Taghvaei & Mehta, 2016). For instance, one way to approximate the gain $K(x)$ is using a Galerkin approximation. In particular, choosing the coordinate functions as basis functions, the so-called constant gain approximation reads (Yang et al., 2016, Eq. (20)):

$$K(x,t) = \frac{1}{N}\sum_{i=1}^{N} X_t^{(i)} \left(h(X_t^{(i)}) - \bar{h}_t\right)^\top = K(t). \tag{129}$$

In this approximation, the gain is constant with respect to the particle positions, i.e. each particle has the same gain, but still changes as a function of time. In this approximation, the additional drift function Ω in Eq. (124) is zero.

For a linear state space model the FBPF with constant-gain approximation becomes exact and is identical[20] to the ensemble Kalman–Bucy filter (EnKBF, Bergemann & Reich, 2012; Taghvaei, de Wiljes, Mehta, & Reich, 2017), which can be shown to be asymptotically exact (Künsch, 2013). More precisely, for a linear model with $f(x) = Ax$, $G(x) = \Sigma_x^{1/2}$ and $h(x) = Bx$, the gain $K(x)$ can be solved for in closed form, using the knowledge that the posterior is Gaussian at all times, and is given by $K = \hat{\Sigma}_t B$. The particles are thus propagated according to (cf. Eq. (124))

$$dX_t^{(i)} = AX_t^{(i)}\,dt + \Sigma_x dB_t^{(i)} + \hat{\Sigma}_t B \Sigma_y^{-1}\left[dY_t - \frac{1}{2}B\left(X_t^{(i)} + \hat{\mu}_t\right)\right] \tag{130}$$

where $\hat{\mu}_t$ and $\hat{\Sigma}_t$ denote the posterior mean and variance as estimated from the particle positions.

6.3. Particle filters in action

Here, we have seen how weighted particle filters can be constructed for nonlinear, continuous time-filtering. Further, we introduced unweighted particle filters as an alternative, which, for some systems (for instance with a high-dimensional observation model), may be advantageous over standard particle filtering. However, unweighted particle filters come at the cost of having to compute the gain function, which can be numerically expensive. Here, we want to demonstrate with an example how these algorithms can be applied as numerical algorithms to a simple nonlinear filtering problem.

Let us consider a hidden process with drift function $f(x) = -4x(x^2 - 1)$ and diffusion constant $g(x) = \Sigma_x^{-1/2}$. The corresponding stationary probability density of this nonlinear model is a bimodal distribution, with peaks at $x = \pm 1$. First, we use observations corrupted by Gaussian noise, e.g. with $h(x) = x$ (note that the model is still nonlinear due to the nonlinearity in the state transition). This filtering problem cannot be solved for in closed form and thus we *have to* use a finite-dimensional approximation, such as the particle filter.

[20] Up to small numerical difference when computing the gain

With Eqs. (116) and (118), the particle transition and weight dynamics for a standard particle filter (i.e. the Bootstrap particle filter) for this model is given by

$$p(X_t|X_{t-dt}^{(i)}) = \mathcal{N}\left(X_t; X_{t-dt}^{(i)} - 4X_{t-dt}^{(i)}((X_{t-dt}^{(i)})^2 - 1)\,dt,\, \Sigma_x\,dt\right), \tag{131}$$

$$\tilde{w}_t^{(i)} = \tilde{w}_{t-dt}^{(i)}\,\mathcal{N}\left(dY_t; X_t^{(i)}\,dt,\, \Sigma_y\,dt\right). \tag{132}$$

After each iterative step, the weights need to be normalized according to $w_t^{(i)} = \tilde{w}_t^{(i)}/\sum_j \tilde{w}_t^{(j)}$. An example tracking simulation is shown in Fig. 5a and b. For comparison, we also consider an extended Kalman–Bucy filter (see Section 5.1) and a feedback particle filter with constant gain approximation (Fig. 5a,c).

Similarly, we might also consider point-process observations with intensity $g(X_t)$. For example, let us use a Gaussian-shaped rate function $g(x) = g_0 \exp(\frac{x-m_0}{2s_0^2})$ for two sensors with peaks at $m_0 = \pm 1$ and width s_0. Fig. 5d shows that a particle filter is able to track the hidden state reasonably well based only on the events elicited from these two sensors. Note also the similarity of this model to the two-state HMM model example we used earlier in Section 4.5.

A fully annotated code for this example, which also explains the technical details of this simulation, is available in our github repository (Kutschireiter, 2019).

7. The restaurant at the end of the universe: Take-away messages

"For where he had expected to find nothing,
there was instead a continuous stream of data".

[Douglas Adams]

In this tutorial, we reviewed some of the theory of nonlinear filtering theory. In fact, we wanted to emphasize that the change of probability measure approach can be used as a universal tool in this context. Not only does it help to derive the corresponding equations such as the filtering equations, but also leads to a deeper understanding of particle filters.

Let us once more summarize the main take-away messages of this tutorial:

1. The change of measure method comes in handy whenever expectations are easier to evaluate under a different measure. For example, when computing conditional expectations, it is often easier to compute under a measure in which the random variables are independent. The Radon–Nikodym derivative acts as the 'conversion factor' between the expectations under the different measures.
2. The filtering equations can be derived by changing to a reference measure under which signal process and observation process are independent. Since this measure change acts on the observation process and leaves the signal process untouched, the filtering equations have the same structure independently of the signal process (which enters in terms of its infinitesimal generator $\mathcal{A}$). Further, all the information about the observations is carried by the Radon–Nikodym derivative.
3. For a general continuous-time and continuous state space nonlinear filtering problem the filtering equations suffer from a closure problem. Suitable approximations are based on a finite-dimensional representation of the filtering density, e.g. in terms of a finite number of statistics (such as the EKF or ADFs) or a finite number of samples (such as PFs).

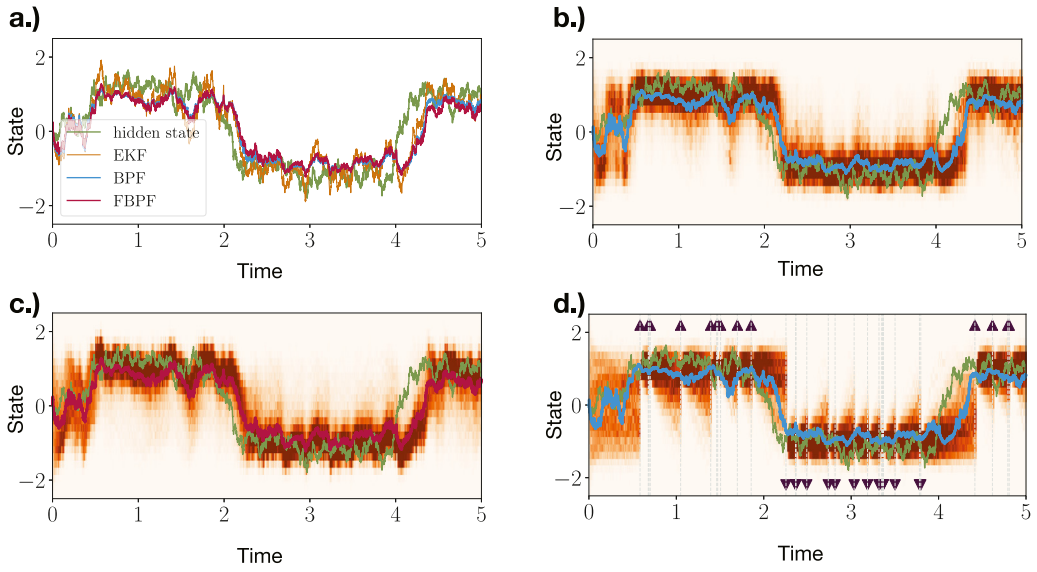

Fig. 5. Nonlinear filtering for a hidden state with state dynamics given by $f(x) = -4x(x^2 - 1)$ and $\Sigma_x = 2$. (a) Estimated first posterior moment $\hat{\mu}_t$ from observations corrupted by Gaussian noise with $h(x) = 1$ and $\Sigma_y = 0.1$. (b) Full posterior from BPF, corresponding to a weighted histogram of the particle positions. (c) Same as (b), but for FBPF. (d) Filtering with the BPF from point-process observations. Here, we consider two sensors with Gaussian-shaped rate functions $g(x)$ and $g_0 = 50$, $s_0 = 0.05$ and $m_0 = \pm 1$. EKF: Kalman–Bucy filter, BPF: Bootstrap particle filter, FBPF: Feedback particle filter with constant-gain approximation.

4. Bootstrap particle filtering can be derived by again changing to a reference measure under which the signal and observation processes are decoupled (which is the very same that was used for deriving the filtering equations), and evaluating the expectations empirically. The importance weights correspond to the Radon–Nikodym derivative of this measure change, evaluated at the particle positions. Thus, the contribution of the signal process (to the solution of the filtering problem) enters via the positions of the particles ('prediction'), while the contribution of the observations enters via the importance weights ('update').

So long, and thanks for all the fish: acknowledgments

The authors would like to express their gratitude towards Christian Horvat, Jannes Jegminat, Luke Rast, Hazem Toutounji, and the anonymous reviewers for taking a test ride and providing constructive feedback. Further, we thank Johannes Bill and Jan Drugowitsch for helpful discussions.

Funding

AK, SCS and JPP were supported by the Swiss National Science Foundation, Switzerland grants PP00P3_179060 and PP00P3_150637 (AK, SCS, JPP) and P2ZHP2_184213 (AK).

Appendix A. Mostly harmless: detailed derivation steps

"You ask this of me who have contemplated the very vectors of the atoms of the Big Bang itself?
Molest me not with this pocket calculator stuff".

[Douglas Adams]

Here, we provide some additional steps that we used in our derivations and left out to keep the main text concise.

A.1. Evolution equation of Radon–Nikodym derivative L_t (Eq. (71))

In Section 4.2, we took the form of the Radon–Nikodym derivative in Eq. (70), and from this form deduced the evolution equation in Eq. (71), i.e., Eq. (70) is the solution of Eq. (71). Showing this is a nice application of Itô's lemma and is therefore outlined here in more detail.

Define

$$\begin{aligned} \Lambda_t &:= \log L_t \\ &= \left[\int_0^t \left(\Sigma_y^{-1/2} h_s\right)^\top d\bar{Y}_s - \frac{1}{2} \int_0^t \left\| \Sigma_y^{-1/2} h_s \right\|^2 ds \right]. \end{aligned} \tag{A.1}$$

$$d\Lambda_t = \left(\Sigma_y^{-1/2} h_t\right)^\top d\bar{Y}_t - \frac{1}{2} \left\| \Sigma_y^{-1/2} h_t \right\|^2 dt. \tag{A.2}$$

Then, with Itô's lemma (recall that for diffusion processes this amounts to a Taylor expansion up to second order in the differential):

$$\begin{aligned} dL_t &:= d(\exp \Lambda_t) \\ &= \exp \Lambda_t d\Lambda_t + \frac{1}{2} \exp \Lambda_t (d\Lambda_t)^2 \\ &= \exp \Lambda_t d\Lambda_t + \frac{1}{2} \exp \Lambda_t \left\| \Sigma_y^{-1/2} h_t \right\|^2 dt \\ &= L_t \left[\left(\Sigma_y^{-1/2} h_t\right)^\top d\bar{Y}_t - \frac{1}{2} \left\| \Sigma_y^{-1/2} h_t \right\|^2 dt \right] \\ &\quad + \frac{1}{2} L_t \left\| \Sigma_y^{-1/2} h_t \right\|^2 dt \\ &= L_t h_t^\top \Sigma_y^{-1/2} d\bar{Y}_t \qquad \text{(A.3)} \\ &= L_t h_t^\top \Sigma_y^{-1} dY_t. \qquad \text{(A.4)} \end{aligned}$$

A.2. Kushner–Stratonovich equation (Eq. (75))

The Kushner–Stratonovich equation describes the SDE for the posterior estimate under the original measure $\mathbb{P}$. With the Kallianpur–Striebel formula (65), we first write the posterior estimate under the reference measure $\mathbb{Q}$. This is given by Eq. (72),

revisited here for convenience:

$$d\mathbb{E}_{\mathbb{P}}[\phi_t|Y_t] = \frac{1}{Z_t}\,\langle d(\phi_t L_t)\rangle_{\mathbb{Q}} + \langle \phi_t L_t\rangle_{\mathbb{Q}}\; d\left(\frac{1}{Z_t}\right) + \langle d(\phi_t L_t)\rangle_{\mathbb{Q}}\; d\left(\frac{1}{Z_t}\right). \tag{72 revisited}$$

Let us evaluate these terms separately. The first term can be expressed as

$$\begin{aligned}\frac{1}{Z_t}\,\langle d(\phi_t L_t)\rangle_{\mathbb{Q}} &= \frac{1}{Z_t}\left\langle L_t d\phi_t + \phi_t dL_t + \underbrace{d\phi_t dL_t}_{=0}\right\rangle_{\mathbb{Q}} \\ &= \frac{1}{Z_t}\left(\langle L_t \mathcal{A}\phi_t\rangle_{\mathbb{Q}}\, dt + \left\langle \phi_t L_t (\Sigma_y^{-1/2} h_t)^\top\right\rangle_{\mathbb{Q}} d\bar{Y}_t\right) \\ &= \langle \mathcal{A}\phi_t\rangle_{\mathbb{P}}\, dt + \left\langle \phi_t h_t^\top\right\rangle_{\mathbb{P}} \Sigma_y^{-1} dY_t, \end{aligned} \tag{A.5}$$

where we first used the Itô lemma for products, and then made use of $\langle dW_t = 0\rangle_{\mathbb{Q}}$. The term $d\phi_t dL_t$ equals zero because their noise components are independent. Further, we used the Kallianpur–Striebel formula to rewrite the expressions in terms of expectations under $\mathbb{P}$. This equation is basically the unnormalized measure in Eq. (74) multiplied by Z_t^{-1}, which we could have used directly.

To obtain dZ_t^{-1}, we need to apply Itô's lemma to the SDE of the normalization constant Z_t in Eq. (65). $dZ_t = d\langle L_t\rangle_{\mathbb{Q}}$ is given by Eq. (74) with $\phi = 1$.

$$dZ_t = d\langle L_t\rangle_{\mathbb{Q}} = \left\langle L_t (\Sigma_y^{-1/2} h_t)^\top d\bar{Y}_t\right\rangle_{\mathbb{Q}}. \tag{A.6}$$

Using Itô's lemma, we find:

$$\begin{aligned} dZ_t^{-1} &= -Z_t^{-2} dZ_t + Z_t^{-3}\,(dZ_t)^2 \\ &= -Z_t^{-2}\left\langle L_t (\Sigma_y^{-1/2} h_t)^\top\right\rangle_{\mathbb{Q}} d\bar{Y}_t + Z_t^{-3}\left\langle L_t \,\middle\|\, \Sigma_y^{-1/2} h_t\right\|_{\mathbb{Q}}^2 dt \\ &= -Z_t^{-1}\left\langle (\Sigma_y^{-1/2} h_t)^\top\right\rangle_{\mathbb{P}} d\bar{Y}_t + Z_t^{-1}\left\langle \left\| \Sigma_y^{-1/2} h_t\right\|\right\rangle_{\mathbb{P}}^2 dt \\ &= -Z_t^{-1}\,\langle h_t\rangle_{\mathbb{P}}^\top \Sigma_y^{-1}(dY - \langle h_t\rangle_{\mathbb{P}}\, dt) \end{aligned} \tag{A.7}$$

Thus, the second term in Eq. (72) reads:

$$\langle \phi_t L_t\rangle_{\mathbb{Q}}\; d\left(\frac{1}{Z_t}\right) = -\langle \phi_t\rangle_{\mathbb{P}}\,\langle h_t\rangle_{\mathbb{P}}^\top \Sigma_y^{-1}(dY - \langle h_t\rangle_{\mathbb{P}}\, dt), \tag{A.8}$$

Finally, the third term uses the SDE of the unnormalized posterior expectation in Eq. (74) and the SDE in Eq. (A.7), keeping only terms up to $\mathcal{O}(dt)$.

$$\langle d(\phi_t L_t)\rangle_{\mathbb{Q}}\; (d\frac{1}{Z_t}) = -\left\langle \phi_t h_t^\top\right\rangle_{\mathbb{P}} \Sigma_y^{-1}\,\langle h_t\rangle_{\mathbb{P}}\; dt. \tag{A.9}$$

Adding up and rearranging the terms, we end up with

$$\begin{aligned} d\mathbb{E}_{\mathbb{P}}[\phi_t|Y_t] &= \langle \mathcal{A}\phi_t\rangle_{\mathbb{P}}\, dt + \left(\left\langle \phi_t h_t^\top\right\rangle_{\mathbb{P}} - \langle \phi_t\rangle_{\mathbb{P}}\,\langle h_t\rangle_{\mathbb{P}}^\top\right) \\ &\quad\times \Sigma_y^{-1}\,(dY_t - \langle h_t\rangle_{\mathbb{P}})\, dt \\ &= \langle \mathcal{A}\phi_t\rangle_{\mathbb{P}}\, dt + \mathrm{cov}_{\mathbb{P}}(\phi_t, h_t^\top) \\ &\quad\times \Sigma_y^{-1}\,(dY_t - \langle h_t\rangle_{\mathbb{P}}\, dt). \end{aligned} \tag{A.10}$$

A.3. Kushner–Stratonovich equation for point-process observations (Eq. (87))

The steps are analogous to those taken in the previous section, with a little caveat: here, SDEs are governed by a point process due to the observation process N_t, so whenever we apply Itô's lemma, we need to consider an *infinite-dimensional Taylor expansion* in the differential, since $dN_t^n = dN$. Also, for simplicity, the following derivation is done for a one-dimensional observation process. However, it is straightforward to be generalized to l dimensions by considering that the observations in each dimension are independent conditioned on all observations up to t, which lead to the product in Eq. (83) and the sums in Eq. (83ff).

First, using the Kallianpur–Striebel formula, we compute the SDE for the normalized posterior expectation by expressing it in terms of the unnormalized posterior SDE:

$$\begin{aligned} d\,\langle \phi_t\rangle_{\mathbb{P}} &= d\left(Z_t^{-1}\,\langle L_t\phi_t\rangle_{\mathbb{Q}}\right) \\ &= \frac{1}{Z_t}\,\langle d(\phi_t L_t)\rangle_{\mathbb{Q}} + \langle \phi_t L_t\rangle_{\mathbb{Q}}\; d\left(\frac{1}{Z_t}\right) \\ &\quad + \langle d(\phi_t L_t)\rangle_{\mathbb{Q}}\; d\left(\frac{1}{Z_t}\right), \end{aligned} \tag{A.11}$$

where we used Itô's lemma for products and the fact that under $\mathbb{Q}$, we can interchange expectation and differentiation.

Again, we compute the terms separately. Using the evolution equation of the unnormalized measure Eq. (86), we find:

$$\begin{aligned} \frac{1}{Z_t}\,\langle d(\phi_t L_t)\rangle_{\mathbb{Q}} &= Z_t^{-1}\,\langle \mathcal{A}\phi_t\, L_t\rangle_{\mathbb{Q}}\, dt \\ &\quad + Z_t^{-1}\left\langle \phi_t L_t \cdot \left(\frac{h_t}{\lambda_0} - 1\right)\right\rangle_{\mathbb{Q}} (dN_t - \lambda_0 dt) \\ &= \langle \mathcal{A}\phi_t\rangle_{\mathbb{P}}\, dt + \frac{1}{\lambda_0}\left(\langle \phi_t L_t\rangle_{\mathbb{P}} - \langle \phi_t\rangle_{\mathbb{P}}\,\lambda_0\right) \\ &\quad\times (dN_t - \lambda_0 dt). \end{aligned} \tag{A.12}$$

For the second term, we again write down the SDE for the normalization constant Z_t and its inverse. From Eq. (86) with $\phi_t = 1$ we find:

$$\begin{aligned} dZ_t = d\,\langle L_t\rangle_{\mathbb{Q}} &= \left\langle L_t\left(\frac{h_t}{\lambda_0} - 1\right)\right\rangle_{\mathbb{Q}} (dN_t - \lambda_0 dt) \\ &= \left(\frac{\langle h_t\rangle_{\mathbb{P}}}{\lambda_0} - 1\right)(dN_t - \lambda_0 dt). \end{aligned} \tag{A.13}$$

The SDE for its inverse is obtained by using Itô's lemma for point processes (Eq. (47)):

$$\begin{aligned} dZ_t^{-1} &= -Z_t^{-1}(\lambda_0 - \langle h_t\rangle_{\mathbb{P}})\, dt \\ &\quad + \left[Z_t + Z_t\left(\frac{\langle h_t\rangle_{\mathbb{P}}}{\lambda_0} - 1\right)^{-1} - Z_t^{-1}\right] dN_t \\ &= Z_t^{-1}\,(\lambda_0 - \langle h_t\rangle_{\mathbb{P}})\,\frac{1}{\langle h_t\rangle_{\mathbb{P}}}\,(dN_t - \langle h_t\rangle_{\mathbb{P}}\, dt). \end{aligned} \tag{A.14}$$

Now we can write

$$\begin{aligned} \langle \phi_t L_t\rangle_{\mathbb{Q}}\; d\frac{1}{Z_t} &= Z_t^{-1}\,\langle \phi_t L_t\rangle_{\mathbb{Q}}\,(\lambda_0 - \langle h_t\rangle_{\mathbb{P}})\,\frac{1}{\langle h_t\rangle_{\mathbb{P}}}\,(dN_t - \langle h_t\rangle_{\mathbb{P}}\, dt) \\ &= \langle \phi_t\rangle_{\mathbb{P}}\,(\lambda_0 - \langle h_t\rangle_{\mathbb{P}})\,\frac{1}{\langle h_t\rangle_{\mathbb{P}}}\,(dN_t - \langle h_t\rangle_{\mathbb{P}}\, dt). \end{aligned} \tag{A.15}$$

Finally, for the last term, we only keep terms of $\mathcal{O}(dN)$:

$$\langle d(\phi_t L_t)\rangle_{\mathbb{Q}}\; (d\frac{1}{Z_t}) = \frac{1}{\lambda_0}\left(\langle \phi_t L_t\rangle_{\mathbb{P}} - \langle \phi_t\rangle_{\mathbb{P}}\,\lambda_0\right)(\lambda_0 - \langle h_t\rangle_{\mathbb{P}})\,\frac{1}{\langle h_t\rangle_{\mathbb{P}}}\, dN_t. \tag{A.16}$$

Adding up and rearranging the terms, we end up with

$$\begin{aligned} &d\,\langle \phi_t\rangle_{\mathbb{P}} \\ &= \langle \mathcal{A}\phi_t\rangle_{\mathbb{P}}\, dt - \left(\langle \phi_t h_t\rangle_{\mathbb{P}} + \langle \phi_t\rangle_{\mathbb{P}}\right) dt + \left(\langle \phi_t h_t\rangle_{\mathbb{P}} + \langle \phi_t\rangle_{\mathbb{P}}\right) dN_t \\ &= \langle \mathcal{A}\phi_t\rangle_{\mathbb{P}}\, dt + \frac{1}{\langle h_t\rangle_{\mathbb{P}}}\mathrm{cov}_{\mathbb{P}}(\phi_t, h_t)(dN_t - \langle h_t\rangle_{\mathbb{P}}\, dt). \end{aligned} \tag{A.17}$$

A generalization to multivariate N_t and h_t, respectively, is done by treating all observable dimensions separately and summing up.

A.4. ADF for point-process observations (Eqs. (100) and (101))

For the ADF in Section 5.2, we need the SDEs for the first two posterior moments. These can be obtained with the KSE for point processes (Eq. (87)). For the mean in Eq. (100), we use $\phi(x) = x$:

$$\begin{aligned} d\mu_t &= \langle \mathcal{A}X_t \rangle \, dt + \sum_{i=1}^{l} \frac{\mathrm{cov}(h_{i,t}, X_t)}{\langle h_{i,t} \rangle} \left(dN_t^i - \langle h_{i,t} \rangle \, dt \right) \\ &= \langle f_t \rangle \, dt + \sum_{i=1}^{l} \frac{\mathrm{cov}(h_{i,t}, X_t)}{\langle h_{i,t} \rangle} \left(dN_t^i - \langle h_{i,t} \rangle \, dt \right) \\ &= \langle f_t \rangle \, dt + \mathrm{cov}(X_t, h_t^\top) \mathrm{diag}(\langle h_t \rangle)^{-1} \left(dN_t - \langle h_t \rangle \, dt \right) \end{aligned} \tag{A.18}$$

To compute the SDE for the posterior variance, we use Itô's lemma:

$$\begin{aligned} d\Sigma_t = d(\langle X_t X_t^\top \rangle - \mu_t \mu_t^\top) = d\langle X_t X_t^\top \rangle - \mu_t d\mu_t^\top \\ - (d\mu_t)\, \mu_t^\top - (d\mu_t)(d\mu_t)^\top, \end{aligned} \tag{A.19}$$

with

$$\begin{aligned} d(\langle X_t X_t^\top \rangle) = \left\langle f_t X_t^\top + X_t f_t^\top \right\rangle dt + \Sigma_x \, dt \\ + \sum_{i=1}^{l} \frac{\mathrm{cov}(h_{i,t}, X_t X_t^\top)}{\langle h_{i,t} \rangle} \left(dN_t^i - \langle h_{i,t} \rangle \, dt \right), \end{aligned} \tag{A.20}$$

$$\mu_t d\mu_t^\top = \mu_t \, \langle f_t \rangle^\top \, dt + \sum_{i=1}^{l} \frac{\mu_t \mathrm{cov}(h_{i,t}, X_t^\top)}{\langle h_{i,t} \rangle} \left(dN_t^i - \langle h_{i,t} \rangle \, dt \right), \tag{A.21}$$

$$(d\mu_t)\mu_t^\top = \langle f_t \rangle \, \mu_t^\top \, dt + \sum_{i=1}^{l} \frac{\mathrm{cov}(h_{i,t}, X_t)\mu_t^\top}{\langle h_{i,t} \rangle} \left(dN_t^i - \langle h_{i,t} \rangle \, dt \right), \tag{A.22}$$

$$(d\mu_t)(d\mu_t)^\top = \sum_{i=1}^{l} \frac{1}{\langle h_{i,t} \rangle^2} \mathrm{cov}(h_{i,t}, X_t) \mathrm{cov}(h_{i,t}, X_t)^\top dN_t^i. \tag{A.23}$$

Adding these up gives us Eq. (101).

References

Armstrong, J., & Brigo, D. (2013). Stochastic filtering via L2 projection on mixture manifolds with computer algorithms and numerical examples. http://dx.doi.org/10.1007/s00498-015-0154-1, arXiv.

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*, 174–188. http://dx.doi.org/10.1109/78.978374, URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=978374.

Bain, A., & Crisan, D. (2009). *Fundamentals of stochastic filtering*. New York: Springer, http://dx.doi.org/10.1007/978-0-387-76896-0.

Benes, V. E. (1981). Exact finite-dimensional filters for certain diffusions with nonlinear drift. *Stochastics*, *5*, 65–92. http://dx.doi.org/10.1080/17442508108833174.

Bergemann, K., & Reich, S. (2012). An ensemble Kalman–Bucy filter for continuous data assimilation. *Meteorologische Zeitschrift*, *21*, 213–219. http://dx.doi.org/10.1127/0941-2948/2012/0307.

Bishop, C. M. (2006). *Pattern recognition and machine learning, volume 4*. Berlin, Heidelberg: Springer-Verlag.

Bobrowski, O., Meir, R., & Eldar, Y. (2009). Bayesian filtering in spiking neural networks: Noise, adaptation, and multisensory integration. *Neural Computation*, 1–39, URL: http://www.mitpressjournals.org/doi/abs/10.1162/neco.2008.01-08-692.

Bremaud, P. (1981). Point processes and queues: Martingale dynamics. In *Springer series in statistics*. Springer New York, URL: https://books.google.ch/books?id=k5jarX-BssUC.

Brigo, D., & Hanzon, B. (1998). On some filtering problems arising in mathematical finance. *Insurance: Mathematics & Economics*, *22*, 53–64. http://dx.doi.org/10.1016/S0167-66879800008-0.

Brigo, D., Hanzon, B., & Le Gland, F. (1999). Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, *5*, 495–534. http://dx.doi.org/10.2307/3318714, URL: http://projecteuclid.org/euclid.bj/1172617201.

Brigo, D., Hanzon, B., & LeGland, F. (1998). A differential geometric approach to nonlinear filtering: the projection filter. *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control*, *43*, 247–252. http://dx.doi.org/10.1109/9.661075, URL: http://ieeexplore.ieee.org/document/661075/.

Brillinger, D. (1988). Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological Cybernetics*, *59*, 189–200. http://dx.doi.org/10.1007/BF00318010.

Brunn, D., Sawo, F., & Hanebeck, U. D. (2006). Efficient nonlinear Bayesian estimation based on fourier densities. In *IEEE international conference on multisensor fusion and integration for intelligent systems* (pp. 317–322). http://dx.doi.org/10.1109/MFI.2006.265642.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *17*, 129–164. http://dx.doi.org/10.2307/2983950.

Daum, F. (1986). Exact finite-dimensional nonlinear filters. *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control*, *31*, 616–622. http://dx.doi.org/10.1109/TAC.1986.1104344.

Daum, F., Huang, J., & Noushin, A. (2010). Exact particle flow for nonlinear filters. In *Signal processing, sensor fusion, and target recognition XIX*: *vol. 7697*, (pp. 769704–769719). http://dx.doi.org/10.1117/12.839590, URL: http://link.aip.org/link/?PSI/7697/769704/1.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, *10*, 197–208.

Doucet, A., & Johansen, A. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. In *Handbook of nonlinear filtering* (pp. 4–6). URL: http://automatica.dei.unipd.it/tl_files/utenti/lucaschenato/Classes/PSC10_11/Tutorial_PF_doucet_johansen.pdf.

Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E., & Pouget, A. (2014). Optimal multisensory decision-making in a reaction-time task. *eLife*, *3*, 748–756. http://dx.doi.org/10.7554/eLife.03005, URL: https://elifesciences.org/articles/03005.

Eden, U. T. (2007). Point process adaptive filters for neural data analysis: Theory and applications. In *Decision and control, 2007 46th IEEE conference on* (pp. 5818–5825). http://dx.doi.org/10.1109/CDC.2007.4434708, URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4434708.

Eden, U. T., & Brown, E. N. (2008). Continuous-time filters for state estimation from point-process models of neural data. *Statistica Sinica*, *18*, 1293–1310. http://dx.doi.org/10.1038/nrm2621, URL: http://www.nature.com/doifinder/10.1038/nrm2621, http://www.ncbi.nlm.nih.gov/pubmed/22065511, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3208353, http://www.nature.com/articles/nrm2621.

Eden, U. T., Frank, L. M., Barbieri, R., Solo, V., & Brown, E. N. (2004). Dynamic analysis of neural encoding by point process adaptive filtering. *Neural Computation*, *16*, 971–998. http://dx.doi.org/10.1162/089976604773135069, URL: http://www.ncbi.nlm.nih.gov/pubmed/15070506.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, *99*(10143), http://dx.doi.org/10.1029/94JC00572.

Gardiner, C. W. (2009). *Handbook of stochastic methods* (fourth ed.). Heidelberg: Springer.

Glaze, C. M., Kable, J. W., & Gold, J. I. (2015). Normative evidence accumulation in unpredictable environments. *eLife*, *4*, http://dx.doi.org/10.7554/eLife.08825.

Gordon, N., Salmond, D., & Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, *140*, 107. http://dx.doi.org/10.1049/ip-f-2.1993.0015, URL: http://digital-library.theiet.org/content/journals/10.1049/ip-f-2.1993.0015, http://www3.nd.edu/lemmon/courses/ee67033/pubs/GordonSalmondSmith93.pdf.

Harel, Y., Meir, R., & Opper, M. (2018). Optimal decoding of dynamic stimuli by heterogeneous populations of spiking neurons: A closed-form approximation. *Neural Computation*, *30*, 2056–2112. http://dx.doi.org/10.1162/neco_a_01105.

Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. New York: Academic Press.

Jia, B., & Xin, M. (2010). A new nonlinear filtering algorithm via Fourier series. In *Proceedings of the 2010 American control conference, ACC 2010* (pp. 606–6070). http://dx.doi.org/10.1109/acc.2010.5531300.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME. Series D, Journal of Basic Engineering*, *82*, 35–45. http://dx.doi.org/10.1115/1.3662552, URL: http://fluidsengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1430402.

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the ASME. Series D, Journal of Basic Engineering*, *83*(95), http://dx.doi.org/10.1115/1.3658902, URL: http://fluidsengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1430804.

Klebaner, F. C. (2005). *Introduction to stochastic calculus with applications* (second ed.). Imperial College Press.

Körding, K. P., Tenenbaum, J. B., & Shadmehr, R. (2007). The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nature Neuroscience*, *10*, 779–786. http://dx.doi.org/10.1038/nn1901, URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2551734&tool=pmcentrez&rendertype=abstract, http://www.nature.com/doifinder/10.1038/nn1901.

Koyama, S., Eden, U. T., Brown, E. N., & Kass, R. E. (2010). Bayesian Decoding of neural spike trains. *Annals of the Institute of Statistical Mathematics*, *62*, 37–59. http://dx.doi.org/10.1007/s10463-009-0249-x.

Künsch, H. R. (2013). Particle filters. *Bernoulli*, *19*, 1391–1403. http://dx.doi.org/10.3150/12-BEJSP07, URL: http://projecteuclid.org/euclid.bj/1377612857.

Kushner, H. (1962). On the differential equations satisfied by conditional probability densities of Markov processes, with applications. *Journal of the Society for Industrial & Applied Mathematics, Control*, *2*, URL: http://epubs.siam.org/doi/pdf/10.1137/0302009.

Kushner, H. J., & Dupuis (2001). Numerical methods for stochastic control problems in continuous time. In *Stochastic modelling and applied probability*: *Vol. 24*, New York, NY: Springer New York, http://dx.doi.org/10.1007/978-1-4613-0007-6, URL: http://link.springer.com/10.1007/978-1-4613-0007-6.

Kutschireiter, A. (2019). Code for: The Hitchhiker's guide to nonlinear filtering. http://dx.doi.org/10.17605/OSF.IO/4C9H.

Macke, J. H., Buesing, L., & Sahani, M. (2011). Estimating state and parameters in state space models of spike trains. In Z. Chen (Ed.), *Advanced state space methods for neural and clinical data* (pp. 137–159). Cambridge: Cambridge University Press, http://dx.doi.org/10.1017/CBO9781139941433.007, URL: http://ebooks.cambridge.org/ref/id/CBO9781139941433A054.

de Melo, F. E., Maskell, S., Fasiolo, M., & Daum, F. (2015). *Stochastic particle flow for nonlinear high-dimensional filtering problems* (pp. 1–61). ArXiv.

Mikulevicius, R., & Rozovskii, B. (2000). Fourier-Hermite Expansions for nonlinear filtering. *Theory of Probability and Its Applications*, *44*, 606–612. http://dx.doi.org/10.1137/S0040585X97977835, URL: http://epubs.siam.org/doi/10.1137/S0040585X97977835.

Ostwald, D., Kirilina, E., Starke, L., & Blankenburg, F. (2014). A tutorial on variational Bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, *60*, 1–19. http://dx.doi.org/10.1016/j.jmp.2014.04.003, URL: http://dx.doi.org/10.1016/j.jmp.2014.04.003.

Pfister, J. -P., Dayan, P., & Lengyel, M. (2009). Know thy neighbour: A normative theory of synaptic depression. *Advances in Neural Information Processing Systems*, 1464–1472, URL: http://papers.nips.cc/paper/3841-know-thy-neighbour-a-normative-theory-of-synaptic-depression.pdf.

Piet, A. T., Hady, A. E., Brody, C. D., Hady, A. El, & Brody, C. D. (2018). Rats adopt the optimal timescale for evidence integration in a dynamic environment. *Nature Communications*, *9*, 1–12. http://dx.doi.org/10.1038/s41467-018-06561-y, URL: http://arxiv.org/abs/1710.05945.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286. http://dx.doi.org/10.1109/5.18626, URL: http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/tutorialonhmmandapplications.pdf, http://ieeexplore.ieee.org/document/18626/.

Radillo, A. E., Veliz-Cuba, A., Josić, K., & Kilpatrick, Z. (2017). Evidence accumulation and change rate inference in dynamic environments. *Neural Computation*, *29*, 1561–1610. http://dx.doi.org/10.1162/NECO_a_00957, URL: http://arxiv.org/abs/1803.01446, http://www.mitpressjournals.org/doi/abs/10.1162/NECO_a_00957.

Speekenbrink, M. (2016). A tutorial on particle filters. *Journal of Mathematical Psychology*, *73*, 140–152. http://dx.doi.org/10.1016/j.jmp.2016.05.006.

Stratonovich, R. L. (1960). Conditional Markov processes. *Theory of Probability and its Applications*, *5*, 3–6.

Surace, S. C. (2015). *Towards a functional understanding of short-term plasticity and the development of a statistical model of in vivo neuronal dynamics*. (Ph.D. thesis).

Surace, S. C., Kutschireiter, A., & Pfister, J. -P. (2019a). Asymptotically exact unweighted particle filter for manifold-valued hidden states and point process observations. *IEEE Control Systems Letters*, *2*, 1. http://dx.doi.org/10.1109/LCSYS.2019.2951093, URL: http://arxiv.org/abs/1907.10143, https://ieeexplore.ieee.org/document/8890670/.

Surace, S. C., Kutschireiter, A., & Pfister, J. -P. (2019b). How to avoid the curse of dimensionality: Scalability of particle filters with and without importance weights. *SIAM Review*, *61*, 79–91. http://dx.doi.org/10.1137/17M1125340, URL: http://arxiv.org/abs/1703.07879, https://epubs.siam.org/doi/10.1137/17M1125340.

Taghvaei, A., & Mehta, P. G. (2016). Gain function approximation in the feedback particle filter. In *2016 IEEE 55th Conference on Decision and Control, CDC* (pp. 5446–5452). IEEE, http://dx.doi.org/10.1109/CDC.2016.7799105, URL: https://arxiv.org/pdf/1603.05496v1.pdf, http://ieeexplore.ieee.org/document/7799105/.

Taghvaei, A., de Wiljes, J., Mehta, P. G., & Reich, S. (2017). Kalman filter and its modern extensions for the continuous-time nonlinear filtering problem. *Transactions of the ASME. Journal of Dynamic Systems, Measurement and Control*, http://dx.doi.org/10.1115/1.4037780, URL: http://arxiv.org/abs/1702.07241, https://asmedigitalcollection.asme.org/dynamicsystems/article/doi/10.1115/1.4037780/474492/Kalman-Filter-and-Its-Modern-Extensions-for-the.

Truccolo, W. (2004). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, *93*, 1074–1089. http://dx.doi.org/10.1152/jn.00697.2004, URL: http://jn.physiology.org/cgi/doi/101152/jn.006972004.

Van Handel, R. (2007). *Stochastic calculus, filtering, and stochastic control*. Caltech, URL: http://www.princeton.edu/rvan/acm217/ACM217.pdfcalculus, Filtering,andStochasticcontrol.pdf.

Van Leeuwen, P. J. (2010). Nonlinear data assimilation in geosciences: An extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, *136*, 1991–1999. http://dx.doi.org/10.1002/qj.699, URL: http://onlinelibrary.wiley.com/store/10.1002/qj.699/asset/699_ftp.pdf?v=1&t=iyznovi1&s=ebd583b403e12f513bcce80d3b7de4473987cac0.

Veliz-Cuba, A., Kilpatrick, Z. P., & Josić, K. (2016). Stochastic models of evidence accumulation in changing environments. *SIAM Review*, *58*, 264–289. http://dx.doi.org/10.1137/15M1028443, URL: http://arxiv.org/abs/1505.04195.

Visser, I. (2011). Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. *Journal of Mathematical Psychology*, *55*, 403–415. http://dx.doi.org/10.1016/j.jmp.2011.08.002.

Wolpert, D., Ghahramani, Z., & Jordan, M. (1995). An internal model for sensorimotor integration. *Science*, *269*, 1880–1882. http://dx.doi.org/10.1126/science.7569931.

Wonham, W. M. (1964). Some applications of stochastic differential equations to optimal nonlinear filtering. *Journal of the Society for Industrial and Applied Mathematics Series A Control*, *2*, 347–369. http://dx.doi.org/10.1137/0302028.

Yang, T., Blom, H. A. P., & Mehta, P. G. (2014). The continuous-discrete time feedback particle filter. In *Proceedings of the American control conference* (pp. 648–653). http://dx.doi.org/10.1109/ACC.2014.6859259.

Yang, T., Laugesen, R. S., Mehta, P. G., & Meyn, S. (2016). Multivariable feedback particle filter. *Automatica*, *71*, 10–23. http://dx.doi.org/10.1016/j.automatica.2016.04.019, URL: https://arxiv.org/pdf/1303.1205v1.pdf, http://dx.doi.org/10.1016/j.automatica.2016.04.019, http://ac.els-cdn.com/S000510981630142X/1-s2.0-S000510981630142X-main.pdf?_tid=86f73940-c6b8-11e6-9fcb-00000aacb361&acdnat=1482240832_857bbc5eb2ba02e92c0598e0a8403835.

Yang, T., Mehta, P. G., & Meyn, S. (2013). Feedback particle filter. *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control*, *58*, 2465–2480. http://dx.doi.org/10.1109/TAC.2013.2258825.

Zakai, M. (1969). On the optimal filtering of diffusion processes. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *11*, 230–243, URL: http://link.springer.com/article/10.1007/BF00536382. http://www.springerlink.com/index/Q730671QUQR1L044.pdf.