u^b

b
UNIVERSITÄT
BERN

Graduate School for Cellular and Biomedical Sciences
University of Bern

Towards a Functional Understanding of Short-term Plasticity and the Development of a Statistical Model of *in vivo* Neuronal Dynamics

PhD thesis submitted by:
Simone Carlo Surace
Burgistein, BE

for the degree of

PhD in Neuroscience

Supervisor:
Prof. Dr. Jean-Pascal Pfister
Department of Physiology
Faculty of Medicine of the University of Bern
New address:
Prof. Dr. Jean-Pascal Pfister
Institute of Neuroinformatics
University of Zürich and ETH Zürich, Switzerland

Co-advisor:
Máté Lengyel, PhD
Computational and Biological Learning Lab
Engineering Department
University of Cambridge UK

Accepted by the Faculty of Medicine, the Faculty of Science and the Vetsuisse Faculty of the University of Bern at the request of the Graduate School for Cellular and Biomedical Sciences

Bern, Dean of the Faculty of Medicine

Bern, Dean of the Faculty of Science

Bern, Dean of the Vetsuisse Faculty Bern

This work is dedicated to

Carolina
my wonderful grandmother

and

Gustav
a good friend and teacher

Preface

It is with great pleasure and relief that I now present my doctoral thesis which is the product of 3.5 years of intense research, sweat and tears. It encompasses two main projects, both of which gave me great reward and at times immense frustration. And as happens often in research, things didn't turn out as planned, surprises turned up, and the struggle for publication sometimes overshadowed opportunities for new ideas to develop.

First and foremost, this is the right place to express my gratitude towards every person who helped me along the way, both before and after starting my thesis. My parents Graziella and Saverio, who made it possible to get a good education, my sisters Serena and Sabina, my friends (whom I cannot mention all for the risk of forgetting one is too big – you know who you are) who were able to cheer me up and offer advice during the most difficult times, my professors of Physics and Mathematics in Bern who gave inspiring and great lectures and set the bar high for me. The members of Walter Senn's lab in Bern, most notably Johanni Brea, for many past and ongoing interesting discussions. And, last but not least, my supervisor Jean-Pascal Pfister, who was very approachable and generous. I also especially thank my piano teacher and good friend Gustav Gertsch, to whom this thesis is especially dedicated. There is no other person that had bigger and more positive influence on my life.

It is well known and I have been warned before starting my PhD thesis that this wouldn't be an easy ride. Of course doing a PhD requires competence on fundamental hard skills on which I was trained – Physics and Mathematics. But far greater is the requirement for learning quickly, managing time, setting the right goals, and asking the right questions. No one is fully prepared for these challenges, and therefore doing this PhD was a great piece of training that I am very happy for.

I also learned that the forefront of science, and perhaps especially the field of theoretical neuroscience, is still a wild forest, a treacherous desert, and a vast ocean of possibilities. This leads to circumstances as those I encountered – not fully understanding where we were headed, not sure who would be interested in the research and would benefit from it. These insecurities haunt many people in many different fields. Fields so specialized that sometimes recognition and appreciation is rare or superficial – or both. Nevertheless, the influence of biology is strong. It leads us to think that we are forced to sell our work to people who are not equipped or willing to understand it, making us devalue the essence of what we are doing. The most important lesson in that respect is that you have to be fully convinced about the value of your work, as that conviction will provide the energy to persist against the odds of submission, revision, and publication of your work.

Let me end my rant and end this preface on a more positive note. The wild forest is a place in which the opportunity for discovery presents itself unexpectedly. Discovering something so interesting that it makes you forget all the rest. Fortunately, I made a couple such discoveries on the way.

Simone Carlo Surace
Bern
May 2015

Abstract

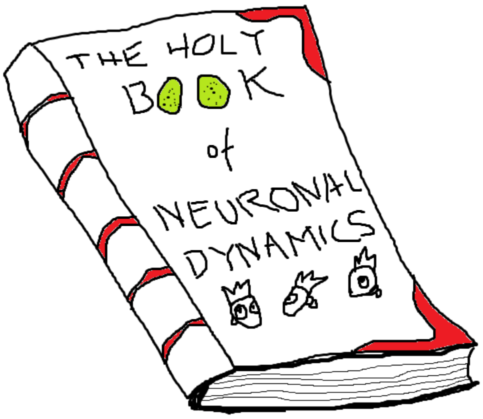
We consider a recently proposed hypothesis for the functional role of short-term synaptic plasticity by Pfister et al., which shows the remarkable similarity of short-term synaptic dynamics and the dynamics of a non-linear filter estimating the presynaptic membrane potential from the spike train. We build on this theory by extending it in various ways and making new predictions. The extensions allow the theory to be formulated for a greater class of stochastic processes, namely multivariate diffusions and Gaussian processes for the membrane potential dynamics and adaptive (uni- or multivariate) point processes for the observed spike train. The new predictions are that presynaptic adaptation is linked to short-term facilitation, and that certain types of short-term depression are linked with a form of presynaptic adaptation which changes the coupling parameter of the gain function.

The second part of the thesis deals with a statistical model for intracellular, *in vivo* recordings of single-neuron activity which does not rely upon an input. Our model generalizes linear-nonlinear-Poisson and generalized linear models by providing the following: 1) a Gaussian process model for the presynaptic membrane potential, and 2) a spike shape kernel for the stereotypic components of the action potential. By using suitable approximations and optimization algorithms, the model can be fitted to *in vivo* data despite the model's non-convexity. The model is shown to perform well on different datasets from different animals and conditions.

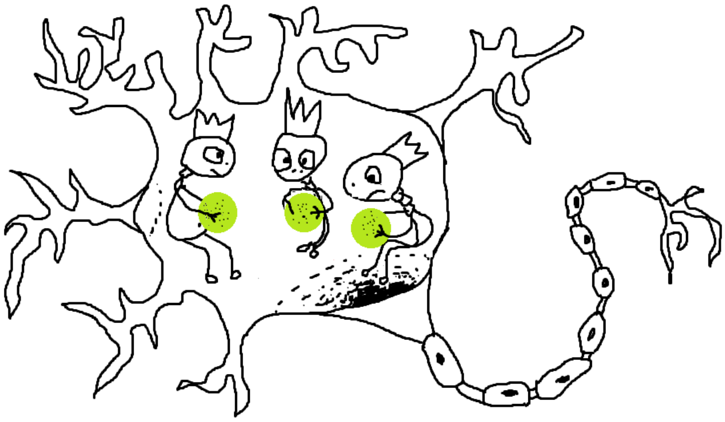
Together, the two parts of the thesis provide all the necessary theoretical tools required to test the theory of short-term plasticity of Pfister et al. By fitting presynaptic activity with the statistical model from the second part, and using the inference methods from the first part, one can derive predictions for the properties of downstream synapses. These predictions can be tested by performing short-term plasticity experiments *in vitro*.

A synaptic fairy-tale

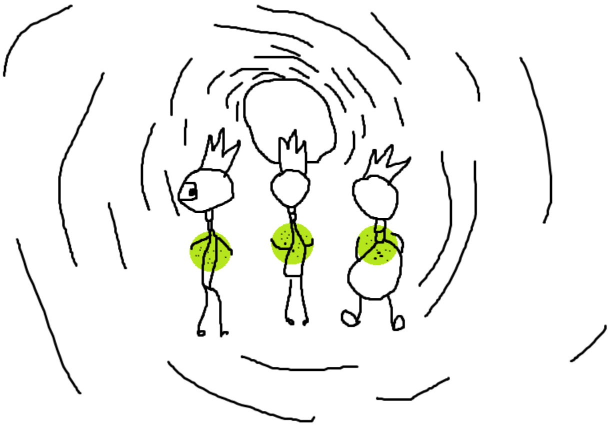
Drawings by Gino Caspari
Text by Simone Surace



Let me take you on a journey to a dark past, whose only written record can be found in The Holy Book of Neuronal Dynamics.

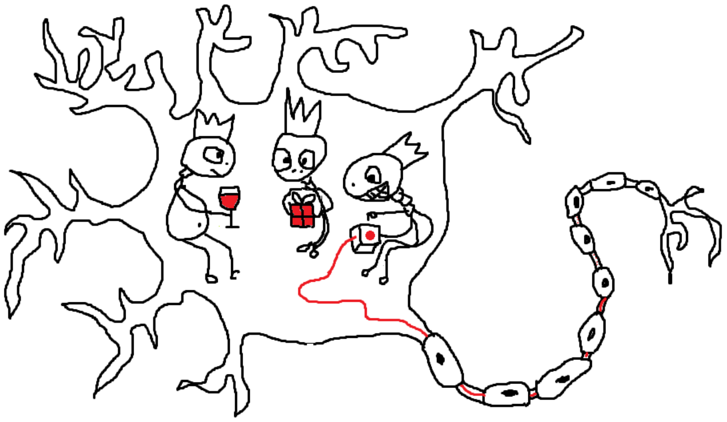


Back then, many moons ago, a single neuron was inhabited by three kings. Their job was to listen to dendritic input and discuss its information content. Each of them would craft a synaptic vesicle and fill it with precious neurotransmitter, shown in green.

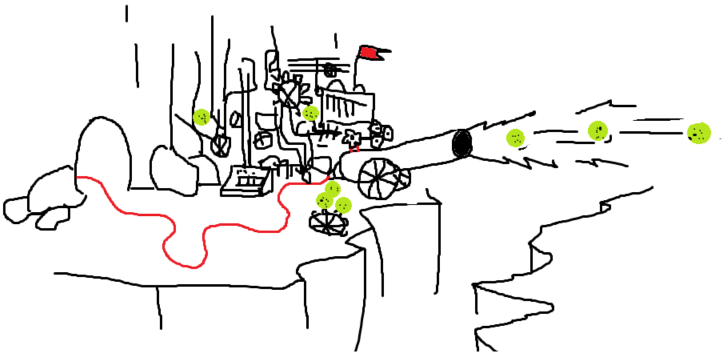


When the time was right, the kings would embark onto a long journey into a dark tunnel, which was how the axon used to look like back then.

After a tedious march, the kings reached the synaptic cleft, a deep crevice that separated them from the neighboring neuron kingdom. They flung their presents over and then hurried back home. The kings had to be quick or else they would miss out on too much information from the voices of the dendritic trees, which they instinctively knew were not to be ignored. Understandably, the kings were growing tired of the stress and of having to do the same thing every day.



The council of kings convened and came to the conclusion that this analog means of communication was inefficient, unreliable and slow. They decided to liberate all kings from their burden by introducing an action potential generator, shown as a small box with red button and corresponding cable. The axon was filled up with conducting material, and from then on the kings had an easy life and a lot of free time which they spent drinking and having fun. They could leisurely listen to the sounds of the dendritic input and press the button when the need arose.



Meanwhile, a complex machine had been installed at the synaptic cleft. It was composed of thousands of proteins, designed to infer what the kings were thinking and to deliver the right amount of neurotransmitter. This story of how synapses became so complex has been told many times, but people have always wondered whether this was all true...

Contents

1 **Introduction** 1

1.1 Phenomenology and Biophysics of STP 2
1.2 The Markram-Tsodyks model of STP 3
1.3 Functional Accounts of STP 4
1.4 The 'Know Thy Neighbour' theory of STP 5

2 **A Statistical Model for *in vivo* Neuronal Dynamics** 9

2.1 Introduction 9
2.2 Results 11
  2.2.1 Definition of the AGAPE model 11
  2.2.2 The model has a rich dynamical repertoire 13
  2.2.3 The model can be learned efficiently 14
  2.2.4 Validation with synthetic data 17
  2.2.5 The model can fit *in vivo* data 18
  2.2.6 The model does not overfit *in vivo* data 20
2.3 Discussion 22
  2.3.1 Flexibility and tractability of the model 23
  2.3.2 Comparison with existing models 23
  2.3.3 Extensions and future directions 24
2.4 Materials and Methods 26
  2.4.1 Description of the datasets used 26
  2.4.2 Preprocessing 26
  2.4.3 Parametrizations and initializations 27
  2.4.4 Model validation 28
2.5 Acknowledgments 29
2.6 Supplementary Text 29
  2.6.1 Discrete Fourier Transform 29
  2.6.2 Circulant matrices 29
  2.6.3 Circulant approximation 29
  2.6.4 Sampling using FFTs 31
  2.6.5 Optimization method 31
  2.6.6 Gradient 31
  2.6.7 Hessian 33

3 **Extensions of a Functional Theory of Short-term Plasticity** 39

3.1 Mathematical Reformulation of the 'Know Thy Neighbour' Theory 39

3.1.1 The filtering problem within the KTN theory 40
3.1.2 Formal Solution of the Filtering Problem 41
3.1.3 The approximate Gaussian filter of Pfister et al. (2009) 44
3.2 Extensions I 54
3.2.1 Adaptive point process observations 54
3.2.2 Multivariate diffusion prior 55
3.2.3 Multivariate point process observations 56
3.3 Extensions II – Gaussian Process Prior 58
3.3.1 Definition and properties 58
3.3.2 Spectral approximation scheme 59
3.3.3 CAR approximation of the squared-exponential GP 60
3.3.4 Filtering with approximate prior 61
3.4 Analysis of the Dynamics and Predictions 64
3.4.1 Adaptation can lead to facilitation 64
3.4.2 The problem of time-constants 67
3.5 Discussion and Future Directions 70
3.5.1 Suggested experiments 70
3.5.2 Open problems 71

A Appendix **75**
A.1 A brief Introduction to SDEs and Stochastic Filtering Theory 75
A.1.1 The Wiener process, stochastic integration, stochastic calculus . . 75
A.1.2 The Fokker-Planck equation 78
A.1.3 Stochastic calculus for counting processes 78
A.1.4 Stochastic filtering 80
A.2 Discrete and Continuous Autoregressive Processes 80
A.3 Proof of the Normalization of the Adaptive Point Process 84

Bibliography **87**

List of Mathematical Symbols **95**

Introduction 1

The brain is a mystifying organ, evolved over hundreds of millions of years, studied only for a few centuries. Its complexity, setting it well apart from even the most complex systems understood so far, is both quantitative and qualitative. The aspects relating to quantity, i.e. numbers, are easily appreciated – ten billion neurons, many more glial cells, a quadrillion synapses. But there is more: each cell is complex system in itself, being of complicated morphology, containing thousands of proteins, and employing many biochemical means of communication.

The other, more qualitative aspects, are even more important. Brains are highly interconnected, come in a large variety, and are plastic and ever-changing. Due to evolutionary pressures, brains of different animals are adapted to specific environments, extracting relevant information from the senses, processing and transforming them into motor output. Moreover, plasticity allows the brain to adapt further throughout the lifetime of the animal.

While there are now recording techniques that enable us to see hundreds of thousands of nerve cells 'in action' while the animal is engaged in behavior, the vast amounts of data that is thereby collected is very hard to analyze. A neuroscientist is therefore challenged to find the right way to look at the data and to ask good questions.

This thesis has two parts (Chapters 2 and 3) which are connected by a common idea. The brain is plastic in many different ways. One of the best-studied is synaptic plasticity, the ability of synapses (i.e. chemical connections between neurons) to change their strength. This plasticity occurs on many different time-scales. To enable the learning of complex skills and languages, the synaptic and structural changes are likely to last for years. On the other end of the spectrum is short-time plasticity, which occurs on the very short time-scale of a few tens to hundreds of milliseconds. It is the idea that this short-term form of synaptic plasticity is involved in learning and estimation, and the connection to uncertainty, stochastic dynamics, and inference that inspires this thesis.

The theory proposed by Pfister et al. (2009), which is the starting point of this thesis, looks upon the synapse as an estimator, decoder, or filter. This filter extracts information about the presynaptic membrane time-course from the spike train arriving at the synapse. In order to perform this task near-optimally, the synapse needs to incorporate some prior knowledge about the presynaptic neuron's statistics in its biophysical machinery – how the

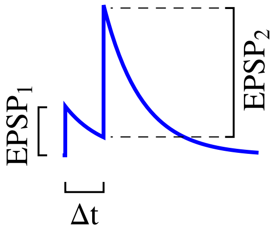


Figure 1.1: The paired-pulse ratio (PPR) is the ratio between the second and the first of two subsequent postsynaptic potentials (PSP, shown here are idealized depictions of excitatory PSPs, EPSPs) which occur as a response to a presynaptic stimulation. It varies as a function of the inter-pulse interval Δt. The depicted pair of pulses shows strong facilitation.

presynaptic membrane potential fluctuations look like and how spikes are generated. Chapter 2 of this thesis investigates the stochastic modeling of the presynaptic quantities, while Chapter 3 deals with the estimation or filtering task itself. The theory requires work on both fronts, and this is what this thesis accomplishes.

In the remaining sections of the Introduction, we will set the stage for the coming chapters by briefly reviewing the basic phenomenology of short-term plasticity and the functional roles proposed for it, including the theory by Pfister et al. (2009) which was mentioned above.

1.1 Phenomenology and Biophysics of STP

Chemical synapses constitute the predominant form of connection between nerve cells of the vertebrate central nervous system. They transmit information from one cell to the next by the release of neurotransmitter molecules from the presynaptic axon terminal, which can be detected via specialized receptors on the postsynaptic side, leading to the opening of ion channels on the postsynaptic membrane and a subsequent change in the postsynaptic local membrane potential (Kandel et al., 2012). The efficacy of synapses, which is commonly measured as the magnitude of change in postsynaptic membrane potential associated with a single presynaptic action potential, varies widely, both over space and time, and the changes over time can occur on various time-scales, reaching from 10 ms up to the life-time of the animal. Those changes are referred to as *synaptic plasticity*, and *short-term synaptic plasticity* (STP) is a summary term for the subset of temporary changes in synaptic efficacy which occur on time-scales between 10 ms and a few seconds.

The first measurements of STP date back to the 1940s (see the review by Zucker and Regehr (2002) for a comprehensive list of references) and study the changes of synaptic efficacy as a function of the history of presynaptic activity. The simplest protocol of presynaptic stimulation is the paired-pulse stimulation and the associated paired-pulse ratio (PPR), the ratio between the amplitudes of the second and the first PSPs (see Fig. 1.1). The PPR usually varies as a function of the inter-pulse interval and is close to unity for large separations of the two pulses. However, the PPR is a very simple readout of synaptic dynamics, and does not tell the whole truth about them. More advanced protocols, such as pulse trains with subsequent recovery pulse, or indeed randomized protocols (Costa et al., 2013), serve to better constrain the various parameters of STP. Finally, pharmacological manipulations are needed in order to isolate individual mechanisms.

Roughly speaking, there are two types of short-term plasticity, namely short-term *facilitation* and short-term *depression*. They correspond to distinct and specific biophysical mechanisms; facilitation occurs because of calcium influx into the presynaptic terminal following an action potential, which increases the release probability of neurotransmitters. Depression is caused mainly by the depletion of neurotransmitter vesicle pools. Both mechanisms are usually at play, and their relative magnitude determines whether the synapse is dominated by facilitation or depression. A facilitation-dominated synapse usually shows PPRs above one, and responses to pulse trains show a monotonic increase of the amplitude of PSPs. A recovery period of a few hundreds of milliseconds allows the calcium levels in the presynaptic terminal to decrease to the original level through calcium buffers or removal from the cell. A depression-dominated synapses show the inverse, i.e. PPRs below one and in response to pulse trains, steadily declining PSP amplitudes. There are synapses where facilitation and depression have similar strength, leading to responses which are a mixture of the facilitation- and depression-dominated ones, e.g. the PPR can have values greater and smaller than one for different lengths of the inter-pulse interval.

1.2 The Markram-Tsodyks model of STP

Due to the great complexity of the biophysical apparatus behind STP, a simple phenomenological model is very useful. For an overview of different approaches, see Hennig (2013). One of the better-known phenomenological models of STP is the Markram-Tsodyks (MT) which was introduced in Tsodyks et al. (1998) and Markram et al. (1998), and its various extensions. The simplest form of the MT model which is able to describe both facilitation and depression is defined by the system of ordinary differential equations (ODE)

$$\begin{aligned}\dot{v}(t) &= \frac{v_0 - v(t)}{\tau_v} + Jx(t)y(t)s(t),\\ \dot{x}(t) &= \frac{1 - x(t)}{\tau_x} - x(t)y(t)s(t),\\ \dot{y}(t) &= \frac{Y - y(t)}{\tau_y} + F(1 - y(t))s(t).\end{aligned} \tag{1.1}$$

Here, $s(t)$ is the presynaptic spike train[1], and the variable $v(t)$ denotes the postsynaptic membrane potential at time t, which has a resting value v_0 and a membrane time-constant τ_m. After a spike, it increases by an amount proportional to the vesicle occupancy or resource variable $x(t)$ times the release probability $y(t)$. Those two variables, in turn, have their own dynamics. Upon spike arrival, the resource variable is depleted by a fraction given by the release probability $y(t)$ and replenishes (decays back to unity) with a time-constant τ_x. The release probability $y(t)$, on the other hand, increases at a spike, and decays back to the baseline Y with its own time-constant.

---

[1] The terms on the right hand side containing the spike train $s(t)$, which is a sum of Dirac deltas, is understood to be non-anticipating, meaning that e.g. the value of v at the time t_{sp} of a spike changes from $v(t_{\mathrm{sp}}^-)$ to $v(t_{\mathrm{sp}}^-) + Jx(t_{\mathrm{sp}}^-)y(t_{\mathrm{sp}}^-)$. In a slightly different version of the model, the spiking term in the second equation reads $-x(t_{\mathrm{sp}}^-)y(t_{\mathrm{sp}}^+)$ with the interpretation that the increase in release probability $y(t)$ is mediated by Calcium influx, which occurs before vesicle release

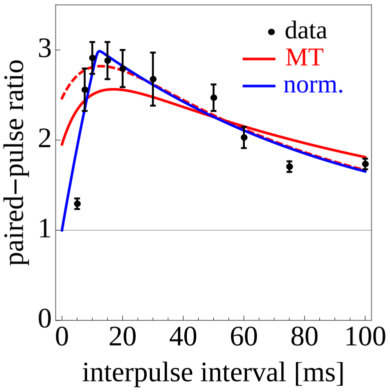


Figure 1.2: An example of very short-term facilitation data by Dobrunz et al. (1997) (black) is badly explained by the Markram-Tsodyks model of STP in Eq. (1.1) (red line), even when the first data point is left out (red, dashed line). An extension of the normative theory of STP (blue line, see Section 3.4 for details) is shown for comparison.

This simple three-variable model with seven parameters can account for many experimentally observed forms of STP, but there are notable cases, e.g. the one of very short-term facilitation by Dobrunz et al. (1997), where it fails (see Fig. 1.2). Motivated by the phenomenology and accompanied by advances in biophysical understanding, there have been a large number of extensions of the basic MT model, some of which are listed and explained in the review by Hennig (2013). Nevertheless, in this thesis, the MT model in Eq. (1.1) will serve as a base-line for comparison to the normative theory of short-term plasticity.

1.3 Functional Accounts of STP

People have wondered about the functional significance of STP for information processing in the brain since the first observations of the phenomenon. While it is difficult to imagine a synaptic apparatus which does not, in some way or another, express STP, the richness of different STP types and the apparent non-randomness of its deployment suggest that there might be functional benefits associated with it.

It is easily appreciated that STP makes the postsynaptic response dependent on the history of presynaptic activity. This history-dependence occurs on a time-scale which coincides with time-scales relevant for many types of behaviors and processes in the environment. It has been thought that low-frequency information is favored by a depressing synapse, whereas high-frequency information can be communicated more efficiently with a facilitating synapse. Computational studies investigating the information transfer across dynamic synapses found a broadband response in the steady-state regime, independently of the degree of facilitation or depression (Lindner et al., 2009). However, it has recently been shown that dynamic synapses improve information transfer compared to static ones (Rotman et al., 2011), and that stochastic release of neurotransmitters has important effects on information transmission (Rosenbaum et al., 2012). Information transfer was investigated also in Scott (2005) and Scott et al. (2012). For more references on the temporal filtering aspects of STP, see also Tsodyks and Wu (2013).

The temporal filtering properties of dynamic synapses can lead to several functional consequences, both for the neuron-to-neuron communication and network-wide computations. First of all, automatic gain control (Abbott et al., 1997) can be achieved by a neuron which receives synaptic inputs from synapses with similar degree of short-term depression, but different presynaptic activity patterns. Because of the low-pass filter properties of depression, low firing rate inputs will have a stronger influence on the postsynaptic neuron than high firing rate inputs. A single input spike train with temporal correlation between spikes is decorrelated by short-term depression (Goldman et al., 2002). This occurs because depression is strongest for the short inter-spike intervals for which correlation is also strongest.

Dynamic synapses can have profound effects on network dynamics, e.g. stability (Tsodyks et al., 1998; Katori et al., 2013; Torres and Kappen, 2013) and response to external inputs (Barak and Tsodyks, 2007), and can subserve different functions related to network-wide computations, e.g. working memory (Mongillo et al., 2008).

There have also been attempts to study STP from a Bayesian perspective, proposing it as a way to achieve adaptation to internal changes in the brain(Stevenson et al., 2010), or optimal inference (Pfister et al., 2009) (see below).

1.4 The 'Know Thy Neighbour' theory of STP

We now want to give a more detailed exposition of the normative theory of STP proposed by Pfister et al. (2009) on which Chapter 2 is based. The main tenet of this theory, which we will call 'Know Thy Neighbour' (KTN) theory from now on, is that a neuron receives many synaptic inputs, contributing to subthreshold membrane potential fluctuations. Those fluctuations lead to spiking output, and the analog information contained in the subthreshold fluctuations is therefore converted to a digital signal which travels down the axon to other neurons. The central hypothesis of that work is that the synapse has the task to decode the spike train and estimate the presynaptic subthreshold membrane potential.

In a strict interpretation, this hypothesis requires that computations occur exclusively in the dendrite, and that the combination of action potential generation and synaptic transmission serves to transmit the result of that computation to the postsynaptic neurons. This raises the question why spikes are needed at all, and whether the encoding scheme that is employed is also optimal. A tentative explanation of the usefulness of spikes is the following: Firstly, information has to be transmitted through the brain from one neuron to the next via either an analog (e.g. electrical synapse) or a digital (e.g. chemical synapse via action potential generation) means. Secondly, analog signals deteriorate with distance whereas digital signal transmission is essentially loss-free. By converting inputs to action potentials, a neuron incurs an initial loss due to the analog-to-digital conversions, but transmitting the digital signal does not lead to further losses, even for distances that are very large. In contrast, by not converting inputs to action potentials, transmission of the information is limited to short-range targets, as the signal quality rapidly decays with the distance to the target. Therefore it can be stipulated that there is a critical distance below which projections can be analog and above which it is beneficial to convert the signal to action potentials first (with associated energy costs) in order to reliably transmit it. We will come back to the foundations of the KTN theory in Section 3.5.

The KTN theory is formalized mathematically as a Bayesian inference task (which may also be called a stochastic filtering problem, see Section 3.1) in a stochastic generative model of neuronal activity, which holds that the membrane potential U_t follows an Ornstein-Uhlenbeck (OU) process

$$dU_t = -\theta(U_t - u_r) + b dW_t, \tag{1.2}$$

and that spiking activity follows an inhomogeneous Poisson process with an instantaneous rate equal to the exponential of U_t

$$N_t - N_0 \sim \text{Poisson}\left[\int_0^t g_0 \exp(\beta U_s) ds\right]. \tag{1.3}$$

The authors then proceed to attack the problem of finding the conditional distribution of the membrane potential conditioned on the spiking history,

$$P\left(U_T | \left\{N_t, 0 \leq t \leq T\right\}\right). \tag{1.4}$$

In a sequence of derivations which we do not repeat here (we will show an alternative, more general derivation in Section 3.1), they discretize time and solve the problem approximately by doing recursive Bayesian estimation using assumed density filtering. After taking the continuum limit, they find the following system of ODEs for the approximate posterior mean $\mu(t)$ and the variance $\sigma^2(t)$:

$$\begin{aligned} \dot{\mu}(t) &= -\theta(\mu(t) - u_r) + \beta\sigma^2(t)(s(t) - \gamma(t)), \\ \dot{\sigma}^2(t) &= -2\theta(\sigma^2(t) - \sigma^2_{\text{OU}}) - \beta^2\sigma^4(t)\gamma(t), \end{aligned} \tag{1.5}$$

where $\sigma^2_{\text{OU}} = \frac{b^2}{2\theta}$ is the stationary (prior) variance of the OU process and

$$\gamma(t) = g_0 \exp\left[\beta\mu(t) + \tfrac{1}{2}\beta^2\sigma^2(t)\right] \tag{1.6}$$

is the posterior expectation of the firing rate $g(U_t)$.

Pfister et al. (2009) establish a link between this approximate estimator and STP, for which they bring forward three arguments:

1. The increment of the posterior mean $\mu(t)$ at the time of a presynaptic spike, given by $\beta\sigma^2(t_{\text{sp}})$, is dynamic. When $s(t)$ is a regular spike train of frequency f, the increment approaches a stationary value after enough spikes have elapsed. This stationary increment is a decreasing function of the presynaptic stimulation rate f, and the shape of this function is similar to the one observed for the stationary excitatory postsynaptic potential (EPSP) amplitude found in *in vitro* studies of STP.

2. By properly tuning the parameters (v_0, τ_v, J, and τ_x) of the depressive MT model (defined by having a fixed release probability, i.e. $y(t) = Y$, $F = 0$), the dynamics of $v(t)$ of Eqs. (1.1) can be matched very closely to the dynamics of the posterior mean $\mu(t)$ in Eqs. (1.5). Therefore, the postsynaptic potential of a properly tuned synapse can perform the estimation task which was stated in the hypothesis.

3. In the low stimulation frequency limit, an analytic link can be established between the dynamics of Eqs. (1.5) and (1.1).

The theory was later generalized to a model of subthreshold fluctuations which includes up- and down-states, see Pfister et al. (2010).

The main prediction of the KTN theory is that for the synapse to perform the estimation task, its STP parameters have to match the statistics of the presynaptic neuron (see point 2 above). More precisely, if the presynaptic cell's statistics under *in vivo* conditions is well characterized by the OU process and inhomogeneous Poisson spiking, the downstream synapses are predicted to be depressing, and to have metrics of STP consistent with those derived from Eqs. (1.5). In order to test these predictions, *in vivo* intracellular data of a neuron and STP data from a downstream synapse has to be available. The subsequent validation would entail 1) fitting the presynaptic neuron's intracellular recording with a statistical model (see Chapter 2), 2) performing the calculations for the dynamics of the optimal estimator, similar to Eqs. (1.5) (see Chapter 3), and 3) calculating STP predictions for the protocols used to record STP data and comparing them to the data.

Let us clarify the use of the word 'theory' in relation to the Pfister et al. (2009) and Pfister et al. (2010) studies and this thesis. It is clear that this does not constitute a well-established theory which has been repeatedly confirmed by experiments, but rather a hypothesis. The word 'theory' is used to describe the mathematical framework which allows quantitative predictions, and we will also describe the extensions presented in this thesis as 'extensions of the theory', despite the fact that they have not yet been experimentally confirmed.

A Statistical Model for *in vivo* Neuronal Dynamics 2

(∗ This chapter has been published in Surace and Pfister (2015) ∗)

Single neuron models have a long tradition in computational neuroscience. Detailed biophysical models such as the Hodgkin-Huxley model as well as simplified neuron models such as the class of integrate-and-fire models relate the input current to the membrane potential of the neuron. Those types of models have been extensively fitted to *in vitro* data where the input current is controlled. Those models are however of little use when it comes to characterize intracellular *in vivo* recordings since the input to the neuron is not known. Here we propose a novel single neuron model that characterizes the statistical properties of *in vivo* recordings. More specifically, we propose a stochastic process where the subthreshold membrane potential follows a Gaussian process and the spike emission intensity depends nonlinearly on the membrane potential as well as the spiking history. We first show that the model has a rich dynamical repertoire since it can capture arbitrary subthreshold autocovariance functions, firing-rate adaptations as well as arbitrary shapes of the action potential. We then show that this model can be efficiently fitted to data without overfitting. Finally, we show that this model can be used to characterize and therefore precisely compare various intracellular *in vivo* recordings from different animals and experimental conditions.

2.1 Introduction

During the last decade, there has been an increasing number of studies providing intracellular *in vivo* recordings. From the first intracellular recordings performed in awake cats (Woody and Gruen, 1978; Baranyi et al., 1993) to more recent recording in cats (Steriade et al., 2001), monkeys (Matsumura et al., 1988), mice (Poulet and Petersen, 2008), and even in freely behaving rats (Lee et al., 2006), it has been shown that the membrane potential displays large fluctuations and is very rarely at the resting potential. Some recent findings in the cat visual cortex have also suggested that the statistical properties of spontaneous activity is comparable to the neuronal dynamics when the animal is exposed to natural images (El Boustani et al., 2009). Similar results have been found in extracellular recordings in the ferret (Berkes et al., 2011). Those data are typically characterized by simple quantifications such as the firing rate or the mean subthreshold membrane potential (Poulet and Petersen, 2008), but a more comprehensive quantification is often missing. So the increasing amount of intracellular data of awake animals as well as the need to compare in a rigorous way the data under various recording conditions call for a model of spontaneous activity in single neurons.

Single neuron models have been studied for more than a century. Simple models such as the integrate-and-fire model (Lapicque, 1907; Stein, 1967) and its more recent nonlinear versions (Latham et al., 2000; Fourcaud-Trocmé et al., 2003; Brette and Gerstner, 2005) describe the relationship between the input current and the membrane potential in terms of a small number of parameters and are therefore convenient for analytical treatment, but do not provide much insight about the underlying biophysical processes. On the other end of the spectrum, biophysical models such as the Hodgkin-Huxley model (Hille, 2001; Hodgkin and Huxley, 1952) relate the input current to the membrane potential through a detailed description of the various transmembrane ion channels, but estimating the model parameters remains challenging (Gerstner and Naud, 2009; Druckmann et al., 2007). Despite the success of those types of models, none of them can be directly applied to intracellular *in vivo* recordings for the simple reason that the input current is not known.

Another reason why a precise model of spontaneous activity is needed is that there are several theories that have been proposed that critically depend on statistical properties of spontaneous activity. For example Berkes et al. validate their Bayesian treatment of the visual system by comparing the spontaneous activity and the averaged evoked activity (Berkes et al., 2011). Another Bayesian theory proposed the idea that short-term plasticity acts as a Bayesian estimator of the presynaptic membrane potential (Pfister et al., 2010). To validate this theory, it is also necessary to characterize spontaneous activity with a statistical model that describes the subthreshold as well as the spiking dynamics.

The last motivation for a model that describes both the subthreshold and the suprathreshold dynamics is the possibility to separate those two dynamics in a principled way. Indeed, it is interesting to know from the recordings what reflects the input dynamics and what aspect comes from the neuron itself (or rather what is associated with the spiking dynamics). Of course a simple voltage threshold can separate the sub- and a suprathreshold dynamics, but the value of the threshold is somewhat arbitrary and could lead to undesirable artifacts. Therefore a computationally sound model that decides itself what belongs to the subthreshold and what belongs to the suprathreshold dynamics is highly desirable.

Here, we propose a single neuron model that describes intracellular *in vivo* recordings as a sum of a sub- and suprathreshold dynamics. This model is flexible enough in order to capture the large diversity of neuronal dynamics while remaining tractable, i.e. the model can be fitted to data in a reasonable time. More precisely, we propose a stochastic process where the subthreshold membrane potential follows a Gaussian process and the firing intensity is expressed as a non-linear function of the membrane potential. Since we further include refractoriness and adaptation mechanisms, our model, which we call the Adaptive Gaussian Point Emission process (AGAPE), can be seen as an extension of both the log Gaussian Cox process (Møller et al., 1998) and the generalized linear model (Truccolo et al., 2005; Pillow et al., 2008; Paninski et al., 2009).

2.2 Results

Here we present a statistical model of the subthreshold membrane potential and firing pattern of a single neuron *in vivo*. See Fig. 2.1A for such an *in vivo* membrane potential recording. We first provide a formal definition of the model and then show a range of different results. 1) The model is flexible and supports arbitrary autocorrelation structures and adaptation kernels. Therefore, the range of possible statistical features is very large. 2) The model is efficiently fittable and the learning procedure is validated on synthetic data. 3) The model can be fitted to *in vivo* datasets. 4) All the features included in the model are required to provide a good description of in vivo data.

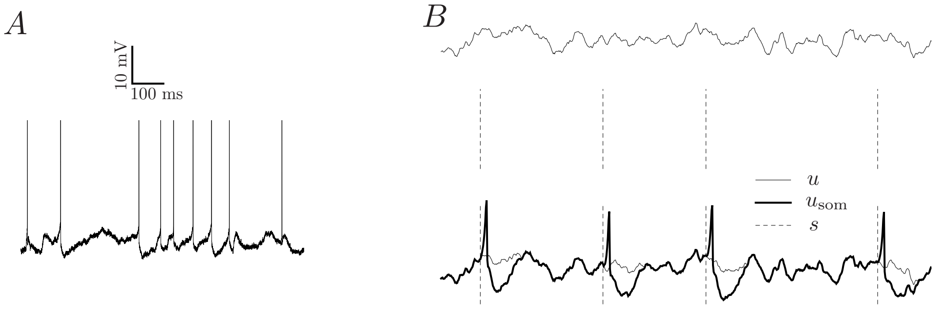


FIGURE 2.1: (A) A sample *in vivo* membrane potential trace from an intracellular recording of a neuron in HVC of a Zebra Finch. (B) The generative AGAPE model can generate a trace of subthreshold membrane potential u (top trace). Based on this potential, a spike train s is generated (middle, dashed vertical lines). Finally, a stereotypic spike-related kernel is convolved with the spike train and added to u, giving rise to u_{som} (bottom, thick line). This quantity is the synthetic analog of the recorded, preprocessed *in vivo* membrane potential.

2.2.1 Definition of the AGAPE model

The AGAPE model is a single neuron model where the input to the neuron is not known, which is typically the case under *in vivo* conditions. The acronym AGAPE stands for Adaptive GAussian Point Emission process since the subthreshold membrane potential follows a Gaussian process and since the spike emission process is adaptive.

More formally, the AGAPE process defines a probability distribution p(u_{som}, s) over the somatic membrane voltage trace $u_{\text{som}}(t)$ and the spike train $s(t) = \sum_{i=1}^{n_s} \delta(t-\hat{t}_i)$ where $\hat{t}_i$, $i = 1, ..., n_s$ are the nominal spike times (decision times), which occur a certain fixed time period $\delta > 0$ before the peak of the action potential. From this probability distribution (or generative model) it is possible to draw samples that look like intracellular *in vivo* activity (for practical purposes, the samples will be compared to the preprocessed recordings, see explanations below). The AGAPE model assumes that the somatic membrane voltage as a function of time $u_{\text{som}}(t)$ is given by (see Fig. 2.1)

$$u_{\text{som}}(t) = u_r + u(t) + u_{\text{spike}}(t), \tag{2.1}$$

where u_r is a constant (the reference potential), $u(t)$ describes the subthreshold membrane potential as a stochastic function drawn from a stationary Gaussian process (GP) (Rasmussen and Williams, 2006)

$$u \sim \mathcal{GP}\left[0, k(t-t')\right] \tag{2.2}$$

with covariance function $k(t-t')$ (which can be parametrized by a weighted sum of exponential decays with weights σ_i^2 and inverse time constants θ_i, see Materials and Methods). For small values of δ (e.g. 1-3 ms), $u(t)$ can be seen as the net contribution from the unobserved synaptic inputs and $u_{\text{spike}}(t)$ is the spike-related contribution (see Fig. 2.1B) which consists of the causal convolution of a stereotypical spike-related kernel α with the spike train $s(t)$, i.e.

$$u_{\text{spike}}(t) = \int_0^\infty \alpha(t')s(t-t')dt'. \tag{2.3}$$

where α can be parametrized by a weighted sum of basis functions with weights a_i, see Materials and Methods. Here, we have made a separation of subthreshold and suprathreshold layers, in that whatever is stereotypic and triggered by the point-like spikes $s(t)$ is attributed to $u_{\text{spike}}(t)$, and the rest belongs to the fluctuating signal $u(t)$. This separation need not correspond to the biophysical distinction between synaptic inputs and active processes of the recorded cell (i.e. the positive feedback loop of the spiking mechanism). Indeed, especially for a choice of large δ (e.g. $\sim$ 20 ms), $u_{\text{spike}}(t)$ also contains large depolarizations due to strong synaptic input which cannot be explained by the GP signal $u(t)$.

Note that this model could easily be extended by including an additional term in Eq. (1) which depends on an external input, e.g. a linear filter of the input (see also Discussion). However, since this input current was not accessible in our recordings, its contribution was assumed to be part of $u(t)$ or $u_{\text{spike}}(t)$.

Now we proceed to the coupling between the subthreshold potential $u(t)$ and the spiking output, as well as adaptive effects associated with spike generation. These effects are summarized by an instantaneous firing rate $r(t)$ – as in the generalized linear model (GLM) (Truccolo et al., 2005; Pillow et al., 2008; Paninski et al., 2009) or escape-rate models (Gerstner and Kistler, 2002) – which is computed from the value of the subthreshold membrane potential at time t, $u(t)$, and the spike history as

$$r(t) = g\left[A(t) + \beta u(t)\right], \quad A(t) = \int_0^\infty \eta(t')s(t-t')dt', \tag{2.4}$$

where $\beta \geq 0$ is the coupling strength between u and the spikes, and $A(t)$ is the adaptation variable which is the convolution of an adaptation kernel η (which can be parametrized by a weighted sum of basis functions with weights w_i, see Materials and Methods) with the past spike train. Also note that we choose not to model adaptation currents explicitly, since they would simultaneously impact the membrane potential and the firing probability (see Discussion). The function g is called gain function, and here we use an exponential one, i.e. $g\left[A(t) + \beta u(t)\right] = e^{\log r_0 + A(t) + \beta u(t)}$. Other functional forms such as rectified linear or sigmoidal could be used depending on the structure of the data. However, this choice has important implications on the efficiency of learning of the model parameters (Paninski, 2004). We define the probability density for s on an interval $[0, T]$ conditioned on u as

$$\begin{aligned} p(s|u) &= \exp\left(-\int_0^T r(t)\,dt\right)\prod_{i=1}^{n_s} r(\hat{t}_i), \\ s(t) &= \sum_{i=1}^{n_s}\delta(t-\hat{t}_i), \quad 0 \leq \hat{t}_1 < \ldots < \hat{t}_{n_s} \leq T, \quad n_s \in \mathbb{N}_0. \end{aligned} \tag{2.5}$$

The parameter β connects the subthreshold membrane potential u to the rate fluctuations. The magnitude of the rate fluctuations depend on the variance σ^2 of u, and therefore we use $\beta\sigma$ as a measure of the effective coupling strength. When $\beta > 0$ the quantity $\theta(t) = -A(t)/\beta$ can be regarded as a soft threshold variable which is modulated after a spike, and $u(t) - \theta(t)$ is the effective membrane potential relevant for the spike generation. This spiking process is a point process which generalizes the log Gaussian Cox process. Indeed, when $A = 0$, Eq. (2.5) describes an inhomogeneous Poisson process with rate $g\left[\beta u(t)\right]$.

Practically, if we want to draw a sample from the AGAPE process, we first draw a sample u from the Gaussian Process (see S1 Text2.6 for how to do this efficiently), then for each time t we draw spikes $s(t)$ with probability density $r(t)$ and update the adaptation variable $A(t)$. Finally, the somatic membrane potential is calculated using Eq. 2.1.

It is important to emphasize at this point that while the model may be directly fitted to the raw membrane potential u_{raw} as recorded by an intracellular electrode, we median filter the data in order to avoid artifacts and downsample for computational efficiency (see 'Materials and Methods'). In this study the model is always fitted to the preprocessed recordings u^*_{som} and this is reflected e.g. in the shape of α which is most strongly affected by the preprocessing. It is important to keep in mind this point while interpreting the results of model fitting. The details of the preprocessing steps which were used are given in the 'Materials and Methods' section.

2.2.2 The model has a rich dynamical repertoire

The AGAPE provides a flexible framework which can be adjusted in complexity to model a wide range of dynamics. While for the datasets presented here a covariance function was used which consists of a sum of Ornstein-Uhlenbeck (OU) kernels, the Gaussian Process (GP) allows for arbitrary covariance functions to be used. This includes simple exponential decay (as produced by a leaky integrate-and-fire neuron driven by white noise current), but it can produce also more interesting covariance functions such as power-law covariances, which are reported in Pozzorini et al. (2013) and El Boustani et al. (2009), or subthreshold oscillations, as reported in Buzsáki (2002).

The model is also able to reproduce a wide range of firing statistics. A common measure of firing irregularity is the coefficient of variation (C_V, i.e. the ratio of standard deviation and mean) of the inter-spike interval distribution. In the absence of adaptation, the AGAPE is a Cox process and therefore has a coefficient of variation $C_V \geq 1$ (Shinomoto and Tsubo, 2001). The precise value of the C_V is a function of the coupling strength ($\beta\sigma$) as well as the autocorrelation of the GP. To illustrate this, we sampled synthetic data from a simple version of the AGAPE where the subthreshold potential u is an OU process with time-constant τ. As shown in Fig. 2.2A, the C_V is an increasing function of the membrane time-constant τ, baseline firing rate r_0, and dimensionless coupling parameter between membrane potential and firing rate $\beta\sigma$. Moreover, the range of the C_V extends from 1 to ≈ 8 within a range of $\beta\sigma \in [0, 2]$ and $r_0\tau \in [2^{-2}, 2^8]$. In the presence of adaptation, firing statistics are markedly different and can produce values of $C_V < 1$ (Gerstner and Kistler, 2002; Lindner et al., 2002). To illustrate this point, we considered an exponential adaptation kernel, i.e. $\eta(t) = -\eta_0 e^{-t/\tau_r}$. While the C_V increases as a function of $\beta\sigma$ and $r_0\tau$ as before, the range of values of the C_V now also covers the interval $(0, 1)$ which is not accessible by the Cox process but which is observed in many neurons across the brain (Softky and Koch, 1993). In order to study the influence of the parameters of the adaptation mechanism, we fix $\beta\sigma = r_0\tau = 1$ and plot the C_V as a function of $r_0\tau_r$ and η_0 (see Fig. 2.2B). Within the parameter region explored in Fig. 2.2B, the C_V spans values from 0.1 up to 1.6.

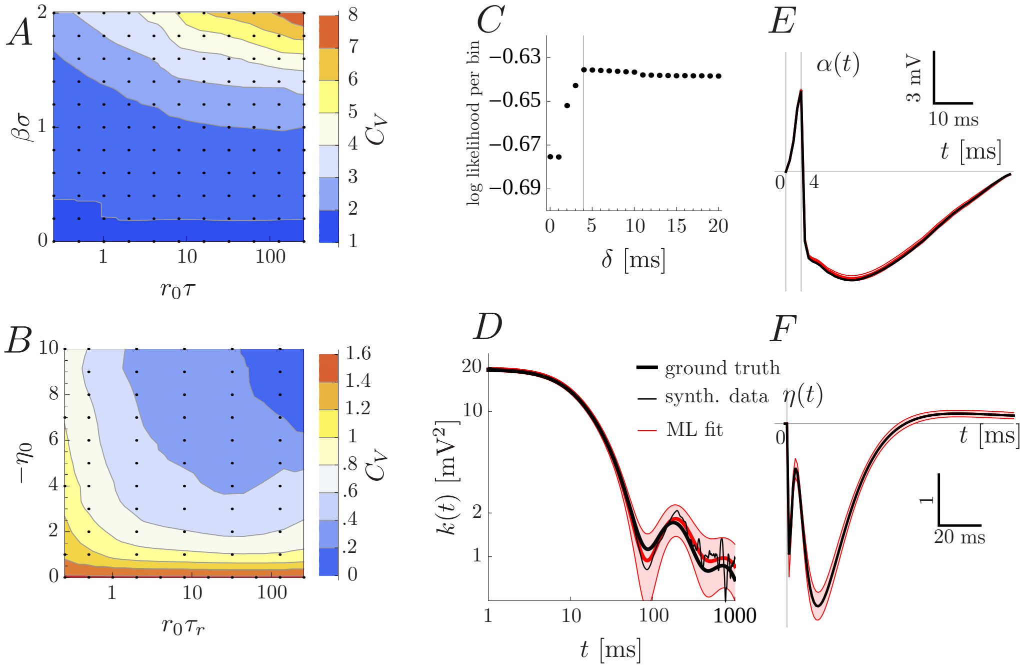


FIGURE 2.2: The model has a rich dynamical repertoire (A,B) and can be correctly fitted to synthetic data (C-F). (A,B) The coefficient of variation (C_V) of the inter-spike interval distribution is computed for parameter values shown as black dots and then linearly interpolated. (A) The C_V of a simple version of the AGAPE ($k(t) = \sigma^2 e^{-t/\tau}$, $\alpha = \eta = 0$) as a function of the model parameters (membrane time-constant τ, baseline firing rate r_0 and coupling strength $\beta\sigma$). (B) C_V of the AGAPE model with an exponentially adaptive process with fixed membrane time-constant, firing rate and coupling ($\beta\sigma = r_0\tau = 1$) as a function of the parameters describing adaptation, namely adaptation strength η_0 and time-constant τ_r. (C,D,E,F) Synthetic data is sampled from the AGAPE model with GP (D), spike-related (E), and adaptation (F) kernels as depicted in black, and $\delta = 4$ ms, $r_0 = 4.15$ Hz, $\beta = 0.374$ mV^{-1}. Then the AGAPE is fitted to the synthetic data by maximum likelihood (ML). (C) The maximum log likelihood per bin as a function of the parameter δ has its maximum at the ground truth value $\delta = 4$ ms. (D,E,F) The ML estimates (red) of the GP, spike-related and adaptation kernels lie within two standard deviations (red shaded regions, estimated by means of the observed Fisher information) from the ground truth.

2.2.3 The model can be learned efficiently

The parameters of the AGAPE model are learned through a maximum likelihood approach. More precisely, we fit the model to an *in vivo* sample (highlighted by a '$*$') of preprocessed somatic membrane potential u^*_{som} and spike train $s^{*,\delta}$ by maximizing the log likelihood applied to the joint data set $(u^*_{\text{som}}, s^{*,\delta})$ over the parameter space of the model (i.e. $u_r, \log r_0, \beta$, the coefficients of the kernels k, η, and α, and the delay parameter δ). The empirical spike train $s^{*,\delta}$ depends on the parameter δ because the formal spike times $\hat{t}_i$ are assigned to be a time period δ before the recorded peak of the action potential. The joint probability of the data can be expressed as a product

$$\begin{aligned} p(u^*_{\text{som}}, s^{*,\delta}) &= \int p(u)p(s^{*,\delta}|u)p(u^*_{\text{som}}|u, s^{*,\delta})\mathcal{D}u \\ &= \int p(u)p(s^{*,\delta}|u)\delta(u^*_{\text{som}} - u - u_r - u_{\text{spike}})\mathcal{D}u \\ &= p(u = u^*_{\text{som}} - u_r - u_{\text{spike}})p(s^{*,\delta}|u = u^*_{\text{som}} - u_r - u_{\text{spike}}) \\ &\equiv p_{s^{*,\delta}}(u^*_{\text{som}})p(s^{*,\delta}|u^*_{\text{som}}). \end{aligned} \tag{2.6}$$

The subscript $s^{*,\delta}$ of the first factor denotes the explicit dependence on the spike train. The individual terms on the r.h.s. will be given below. The function we are optimizing is the logarithm of the above joint probability which we can write as

$$\begin{aligned} \mathcal{L}(u_r, k, \alpha, \log r_0, \beta, \eta, \delta) = &\log p_{s^{*,\delta}}(u^*_{\text{som}}; u_r, k, \alpha) \\ &+ \log p(s^{*,\delta}|u^*_{\text{som}}; u_r, \alpha, \log r_0, \beta, \eta). \end{aligned} \tag{2.7}$$

It should be noted that the presence of the spike-related kernel α in both terms produces a trade-off situation: removing the spike-related trajectory improves the Gaussianity of the membrane potential u (and therefore boosts the first term) at the cost of the of the second term by removing the short upward fluctuation that leads to the spike. This trade-off situation makes maximum likelihood parameter estimation a non-concave optimization problem. Moreover, the evaluation of the GP likelihood of n samples, where $n = \mathcal{O}(10^5)$, comes at a high computational cost. Two important techniques make the parameter learning both tractable and fast: the first is the use of the circulant approximation of the GP covariance matrix which makes the evaluation of the likelihood function fast. The second is the use of an alternating fitting algorithm which (under an appropriate parametrization, see 'Materials and Methods') replaces the non-concave optimization in the full parameter space with two concave optimizations and a non-concave one in suitable parameter subspaces. Those two techniques are further described in the next section.

2.2.3.1 Efficient likelihood computation

The log-likelihood function is evaluated in its discrete-time form with n time points separated by a time-step Δt. The GP variable u (which leads to u_{som} through Eq. (2.1)) is multivariate Gaussian distributed with a covariance matrix $K_{ij} = k(t_i - t_j)$, where $t_i = i\Delta t$. The matrix K is symmetric and, by virtue of stationarity, Toeplitz. Evaluation of the GP likelihood requires inversion of K, which is computationally expensive (the time required to invert a matrix typically scales with n^3). For this reason we approximate this Toeplitz matrix by the circulant matrix C which minimizes the Kullback-Leibler divergence (see (Katsaggelos and Lay, 1991; Bach and Jordan, 2004; Gray, 2006) and S1 Text2.6)

$$C = \underset{D \text{ circulant}}{\text{argmin}} \; \mathcal{D}_{\text{KL}} \left[\mathcal{N}(m, K) || \mathcal{N}(m, D)\right] \tag{2.8}$$

between the two multivariate Gaussian distributions with the same mean but different covariance matrices. This optimization problem can be solved by calculating the derivative of $\mathcal{D}_{\text{KL}}\left[\mathcal{N}(m,K)||\mathcal{N}(m,D)\right]$ with respect to D and using the diagonalization of D by a Fourier transform matrix (Gray, 2006). After a bit of algebra (see S1 Text2.6), denoting $k_i = K_{1i}$ and $k_{n+1} \equiv 0$, the optimal circulant matrix can be written as $C_{ij} = c_{(i-j \mod n)+1}$, where $i, j = 1, ..., n$ and

$$c_i = \frac{1}{n}\left[(n-i+1)k_i + (i-1)k_{n-i+2}\right]. \tag{2.9}$$

The replacement of K by C is equivalent to having periodic boundary conditions on u, which has a small effect under the assumption that the time interval spanned by the data is much longer than the largest temporal autocorrelation length of k. So the first term on the r.h.s. of Eq. (2.6) is a multivariate Gaussian density $\mathcal{N}(0, C)$. The determinant of the covariance matrix C is the product of eigenvalues, which for a circulant matrix are conveniently given by the entries of $\hat{c}$, the discrete Fourier transform of c (see the S1 Text2.6 for our conventions regarding discrete Fourier transforms). Also the scalar product $u^T C^{-1} u$ can be written in terms of $\hat{c}$. Together, the first term on the r.h.s. of Eq. (2.6) takes the simple form

$$\log p_{s^{*,\delta}}(u^*_{\text{som}}) = -\frac{1}{2}\sum_{i=1}^{n}\left(\log(2\pi\hat{c}_i) + \frac{1}{n}\frac{|\hat{u}_i|^2}{\hat{c}_i}\right), \tag{2.10}$$

where $\hat{u}_i$ are the components of the discrete Fourier transform of u^*. The Gaussian component of the membrane potential u is implicitly given by the discretized somatic voltage modified by a discrete-time version of the spike-related kernel convolution,

$$u^*_i = u^*_{\text{som},i} - u_r - \sum_{j=1}^{i-1}\alpha_j s^{*,\delta}_{i-j}, \tag{2.11}$$

where $s^{*,\delta}_i$ is the binned spike train (see below), α_i is a discretized version of the spike-related kernel. The time required to compute $\log p_s(u^*_{\text{som}})$ is determined by the complexity of the Fourier transform, which is of the order of $n \log n$. This dramatic reduction in complexity (compared to n^3) allows a fast evaluation of the log-likelihood.

The spiking distribution $p(s^{*,\delta}|u^*_{\text{som}})$ is approximated by a Poisson distribution with constant rate within one time bin. For each bin, $s^{*,\delta}_i$ counts the number of spikes that occur in that bin, and the conditional likelihood of the spikes therefore reads

$$\log p(s^{*,\delta}|u^*_{\text{som}}) = \sum_{i=1}^{n}\left\{s^{*,\delta}_i \log\left[r_i\Delta t\right] - r_i\Delta t - \log\left[s^{*,\delta}_i!\right]\right\}, \tag{2.12}$$

where $r_i = g[\beta u^*_i + \sum_{j=1}^{i-1}\eta_j s^{*,\delta}_{i-j}]$ and u^*_i as defined in Eq. (2.11). If $s^{*,\delta}_i$ contains only zeros and ones (which can be accomplished given small enough bins), the last term $\log s^{*,\delta}_i!$ vanishes.

2.2.3.2 Efficient parameter estimation

Except for the parameter δ, which takes discrete values of multiples of the discretization step Δt, it is possible to analytically calculate the first and second partial derivatives of the likelihood function defined in Eq. (2.6) with respect to the model parameters (u_r, k, α, $\log r_0$,

β, η) (see S1 Text2.6) to facilitate the use of gradient ascent, Newton, and pseudo-Newton optimization algorithms. A desirable feature of an optimization problem is concavity of the objective function (in our case, the log-likelihood function). Even though the problem of finding optimal parameters for the AGAPE process is not concave, the optimization can be done in three alternating subspaces (see Fig. 2.3). The full set of parameters Θ is divided into three parts: θ_{GP} for the GP parameters (u_r, parameters of k), $\theta_{\text{spike kernel}}$ for the spike-related kernel parameters, and θ_{spiking} for the parameters controlling spike emission ($\log r_0, \beta$ and parameters of η). The optimization then proceeds according to the following cycle: (1) the GP parameters are learned, (2) the spike-related kernel parameters are learned, and lastly (3) the spiking parameters are learned. In each step the remaining parameters are held fixed. The cycle is repeated until the parameters reach a region where the log likelihood is locally concave in the full parameter space, after which the optimization can be run in the full parameter space until it converges. Joint concavity of the log likelihood holds if all the eigenvalues of the Hessian matrix are strictly negative. As shown in Paninski (2004), step (3) is concave for a certain class of gain functions g, including the exponential function, and linear parametrizations of the adaptation kernel. The same holds for the spiking term of the log-likelihood in step (2). The voltage term of the log likelihood of step (2) is concave by numerical inspection in the cases we considered. To summarize, steps (2) and (3) are concave and Newton's method can be used in these steps as well as for the final concave optimization in the full space. Step (1) is non-concave and therefore a simple gradient ascent algorithm is used.

The optimization over $(u_r, k, \alpha, \log r_0, \beta, \eta)$ is repeated for every $\delta = 0, \Delta t, 2\Delta t, ..., \delta_{\text{max}}$ in order to select the one δ that maximizes the log-likelihood $\mathcal{L}(u_r, k, \alpha, \log r_0, \beta, \eta, \delta)$. The value of δ_{max} is chosen such that it is less than the least upper bound of the support of the basis of the spike-related kernel α. Since the parameters $u_r, k, \alpha, \log r_0, \beta, \eta$ are expected to change only a little when going from one δ to the next, $\delta + \Delta t$, learned parameters for δ can be used as initial guesses for nearby $\delta + \Delta t$ or $\delta - \Delta t$. We thus get two different initializations, which we can exploit by starting e.g. with $\delta = 0$, ascending through the sequence of candidate δ's up to the maximum δ, and descending back to zero.

2.2.4 Validation with synthetic data

Despite this improvement in speed and tractability, the optimization is still riddled with multiple local minima which require the use of multiple random initializations. In order to demonstrate the validity of the fitting method, synthetic data of length 270.112 seconds ($n = 270112$, the same as *in vivo* dataset $\mathcal{D}_1$, see below) was generated with known parameters ($\delta = 4$ ms, $r_0 = 4.15$ Hz, $\beta = 0.374$ mV^{-1} and GP, spike-related kernel and adaptation kernels as depicted in Fig. 2.2D-F). The learning algorithm was initialized with least-squares estimates of the covariance function parameters σ_i^2 based on the empirical autocorrelation function of u_{som} and spike-related kernel and adaptation kernels set to zero. The true underlying δ can be recovered from the synthetic data (Fig. 2.2C). Moreover, the algorithm converges after a few dozen iterations (taking only three minutes on an ordinary portable computer) and – with δ set to 4 ms – recovers the correct GP, spike-related, and adaptation kernels (Fig. 2.2D-F). All ML estimates lie within a region of two standard deviations around the ground truth, where standard deviations are estimated from the observed Fisher information (Efron and Hinkley, 1978).

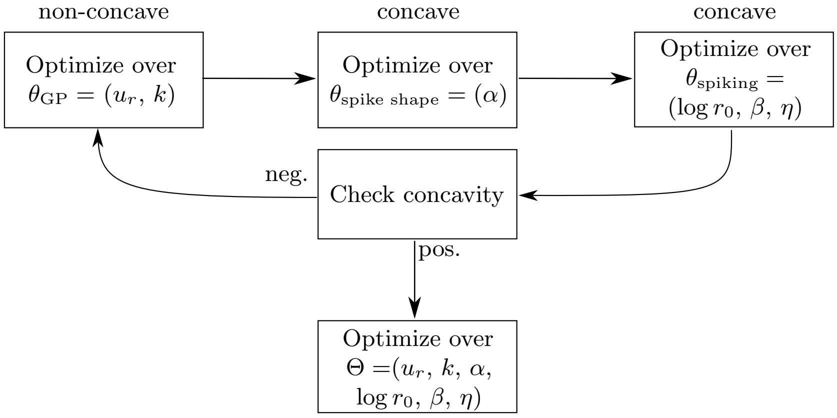

non-concave
concave
concave
Optimize over $\theta_{\mathrm{GP}} = (u_r,\, k)$
Optimize over $\theta_{\text{spike shape}} = (\alpha)$
Optimize over $\theta_{\text{spiking}} = (\log r_0,\, \beta,\, \eta)$
neg.
Check concavity
pos.
Optimize over $\Theta = (u_r,\, k,\, \alpha,\, \log r_0,\, \beta,\, \eta)$

Figure 2.3: This schematic shows the optimization scheme that is used to learn the parameters of the AGAPE model when it is fitted to the data (for a given δ). As long as the current parameter estimate sits in a non-concave region of the likelihood function, the top cycle optimizes over different subspaces of the parameter space. If and when a concave point is reached, the optimization proceeds in the full parameter space. This whole scheme is repeated for each value of δ in order to find the optimal one.

2.2.5 The model can fit *in vivo* data

We fitted the model to a number of *in vivo* traces from different animals and conditions (see 'Materials and Methods' for a detailed description of the data sets). We would like to remind the reader at this point that the model is never fitted to the raw membrane potential, but to a preprocessed, i.e. median-filtered and downsampled dataset (see Materials and Methods). Because of this preprocessing stage, the model only sees the truncated action potentials which emerge from the median filter. This is reflected in the extracted spike-related kernel α, which is characterized by a smaller amplitude than the original action potential in the raw membrane potential data.

We show the detailed results of the model fitting for the example songbird HVC dataset $\mathcal{D}_1$. The optimal value of δ for this dataset was $\delta = 18$ ms (see S3 Fig), with which the model captures the subthreshold and suprathreshold statistics (smaller values of δ compromise both the subthreshold and suprathreshold description because the large upward fluctuations which preceed spikes in this dataset are unlikely to arise from a GP). In particular, the stationary distribution of the membrane potential u is well approximated by a Gaussian (Fig. 2.4B) and pronounced after-hyperpolarization is seen in the spike-related kernel (Fig. 2.4D). The subthreshold autocorrelation structure is well reproduced by the parametric autocorrelation function k (Fig. 2.4C). The adaptation kernel reveals an interesting structure in the way the spiking statistics deviates from a Poisson process (Fig. 2.4E). This feature of the spiking statistics is also reflected in the inter-spike interval (ISI) distribution (Fig. 2.4F). Both the data and the fitted model first show an increased, and then a significantly decreased probability density when compared to a pure Poisson process. The remaining parameters are listed in Tab. 2.1 (errors denote two standard deviations, estimated from Fisher information, see Materials and Methods). The model can be used to generate synthetic data, which is shown in Fig. 2.4H.

In order to show the generality of the model, we fitted the model on two more datasets,

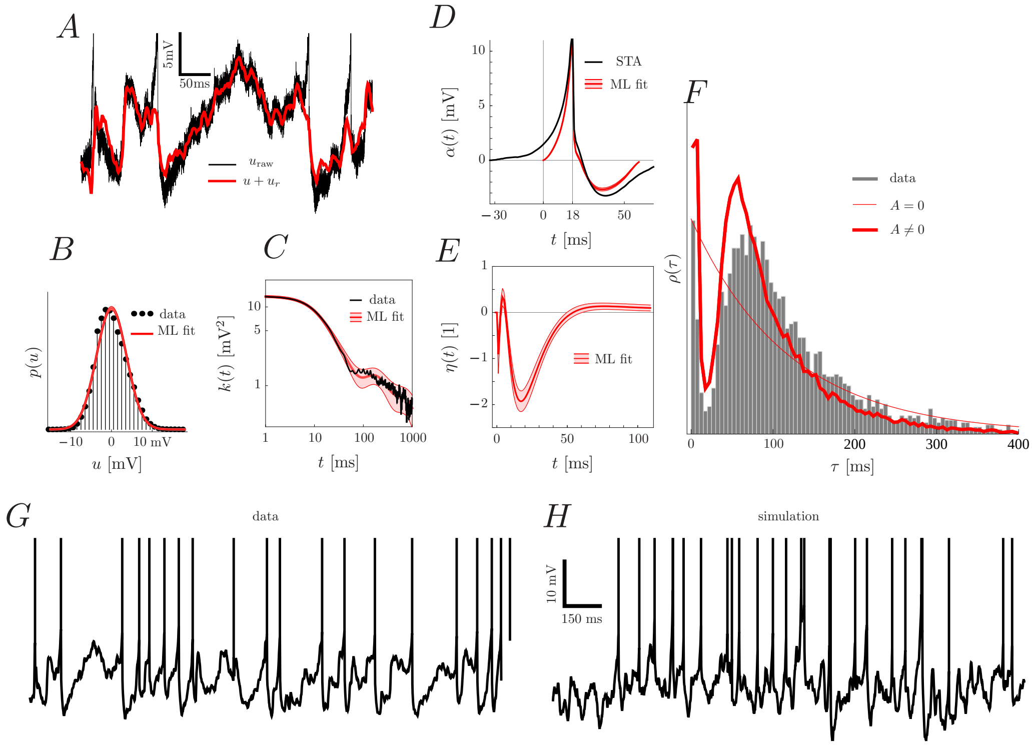


FIGURE 2.4: The results of maximum likelihood (ML) parameter fitting to dataset $\mathcal{D}_1$. After fitting, we see (A) the removal of the spike-related kernel through the difference between the recorded trace u^*_{som} and the subthreshold membrane potential $u + u_r$; (B) the match of the stationary distribution of the subthreshold potential u and a Gaussian. We also observe that (C) the autocorrelation function of the data, Eq. (2.14), is well reproduced by $k(t)$ in Eq. (2.15); (D) the spike-related kernel $\alpha(t)$ starts at $-\delta = -18$ ms relative to the peak of the action potential. The difference between the spike-triggered average (STA) and the spike-related kernel is attributed to the GP; and (E) that the adaptation kernel $\eta(t)$ shows distinct modulation of firing rate which produces firing statistics significantly different from a Poisson process. This is also reflected in the inter-spike interval density $\rho(\tau)$ (F) of the data, which shows good qualitative agreement with a simulated AGAPE with adaptive kernel as in (E) (thick red line), but not by a non-adaptive (i.e. Poisson) process (thin red line). After fitting, a two second sample of synthetic data (H) looks similar as the *in vivo* data (G). In (G,H) vertical lines are drawn at the spiking times. All red shaded regions denote ± 2 standard deviations, estimated from the observed Fisher information.

Dataset	$\mathcal{D}_1$	$\mathcal{D}_3$	$\mathcal{D}_4$
δ [ms]	18	12	32
u_r [mV]	-52.9±0.2	-66.6±0.1	-51.5±0.5
r_0 [Hz]	11.7±0.6	71±5	0.15±0.05
β [mV^{-1}]	0.12±0.01	0.24±0.02	0.46±0.05
$\beta\sigma$ [1]	0.45±0.03	0.67±0.04	1.3±0.1

Table 2.1: The values (p.m. two standard deviations, estimated from the observed Fisher information) of the fitted parameters not shown in Fig. 2.5 for the *in vivo* datasets described in the main text. The last row shows the effective coupling strength between the membrane potential and the firing rate, given by β times the standard deviation σ of the membrane potential.

$\mathcal{D}_3$ from another HVC neuron and $\mathcal{D}_4$ from mouse visual cortex. The parameter δ was found to take the optimal value of 12 ms for $\mathcal{D}_3$ and 32 ms for $\mathcal{D}_4$ (to see how fitted parameters change as a function of δ, see S4 Fig and S5 Fig). The comparison of the GP, spike-related and adaptation kernels is shown in Fig. 2.5, and the remaining parameters are listed in Tab. 2.1. The three cells show pronounced differences in autocorrelation structure, spike-related kernel and spike-history effects. In particular the two datasets $\mathcal{D}_1$ and $\mathcal{D}_4$ show rather long autocorrelation lengths of the membrane potential and asymmetric spike-related kernels, whereas the cell in $\mathcal{D}_3$ has comparatively short autocorrelation length and very pronounced hyperpolarization. Adaptation is much stronger in $\mathcal{D}_3$ than in $\mathcal{D}_1$, balancing the much higher baseline firing rate r_0, see Tab. 2.1. The error bars on the adaptation kernel are small for datasets $\mathcal{D}_1$ and $\mathcal{D}_3$ due to the abundance of spikes. On the other hand, the adaptation kernel of dataset $\mathcal{D}_4$ is poorly constrained by the available data. This is due to the fact that dataset $\mathcal{D}_4$ consists of very short trials with very few spikes. Despite this fact, good agreement is achieved between the distribution of inter-spike intervals of the *in vivo* data and ISI statistics sampled from the AGAPE (see Fig. 2.5, bottom row) for all datasets.

2.2.6 The model does not overfit *in vivo* data

The AGAPE process has a fairly large number of parameters. Therefore it is important to check whether the model overfits the data, compromising its generalization performance. In short, when a model has too many parameters, it tends to be poorly constrained by the data and therefore when the model is first trained on one part of the data and then tested on another part on which it is not trained, the test performance will be significantly worse than the training performance.

Here, we use cross-validation to perform a factorial model comparison on an exemplary dataset in order to validate the different structural parts of the model. The procedure is described in detail in the Materials and Methods.

Model comparison is performed on the dataset $\mathcal{D}_2$ and the results are shown in Fig. 2.6, where the mean difference of per-bin log-likelihood (see ‘Materials and Methods’)

$$\Delta p_i^{\text{valid}} = \langle \Delta p_{ij}^{\text{valid}} \rangle_j, \quad \Delta p_i^{\text{test}} = \langle \Delta p_{ij}^{\text{test}} \rangle_j, \qquad \Delta p_{ij}^{\text{valid,test}} = p_{i,j}^{\text{valid,test}} - p_{G\alpha\beta\eta,j}^{\text{valid,test}} \quad (2.13)$$

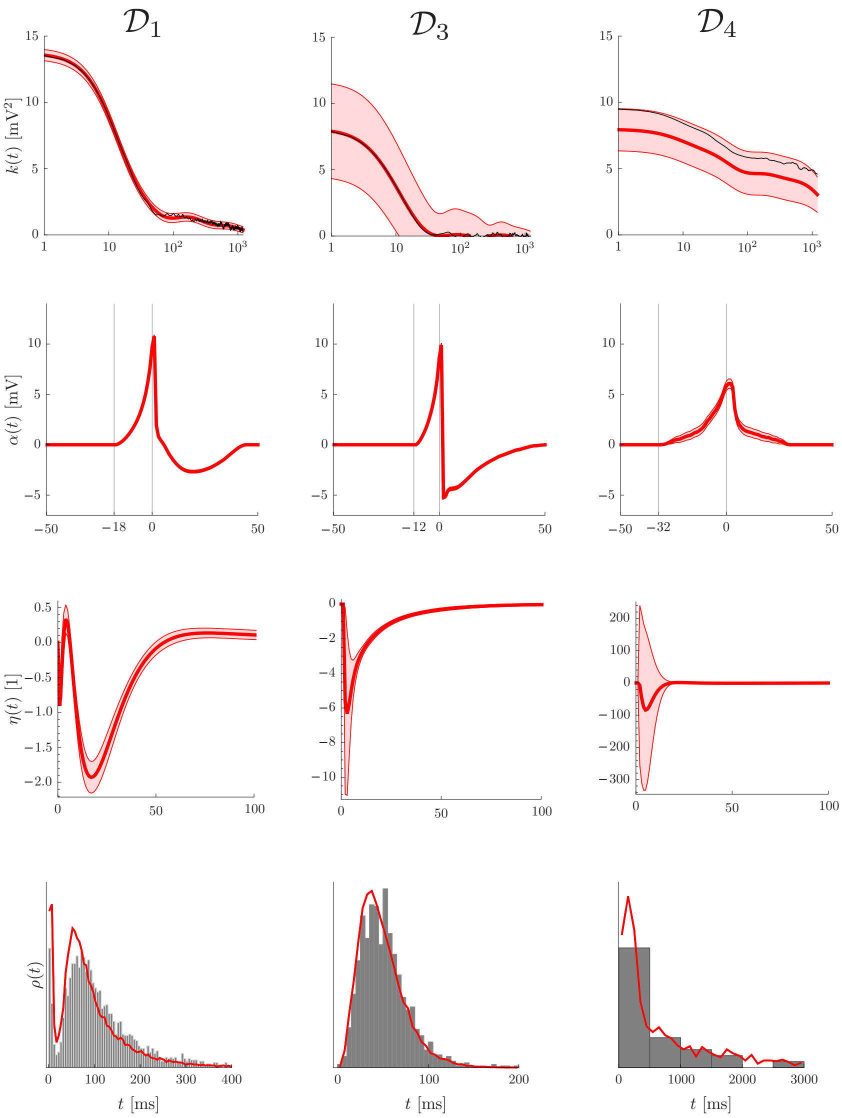


Figure 2.5: Fitting results for three different datasets. Dataset $\mathcal{D}_1$ is the same as in Fig. 2.4, i.e. an HVC neuron from anesthetized Zebra Finch. $\mathcal{D}_3$ is from HVC in awake Zebra Finch, and $\mathcal{D}_4$ is from mouse visual cortex in awake mouse. The different panels show the results after fitting; in the first line the GP covariance function $k(t)$ (red) and the empirical autocorrelation (black), Eq. (2.14), in the second line the spike-related kernel $\alpha(t)$, in the third line the adaptation kernel $\eta(t)$, and in the fourth line the inter-spike interval density $\rho(t)$ (data ISI histogram in gray, simulated ISI distribution from AGAPE in red). There are pronounced differences between datasets in all three kernels, showing the flexibility of the AGAPE model in describing a wide range of statistics. All red shaded regions denote ± 2 standard deviations, estimated from the observed Fisher information.

is shown for all models $i \in \{0, ..., G\alpha\beta\eta\}$ (here, $\langle\cdot\rangle_j$ denote averages over chunks j of the cross-validation). The results are very similar for both validation data (which was left-out during training, but appeared in other training runs) and the test data which was never seen during training. The most complex model ($\mathcal{M}_{G\alpha\beta\eta}$) performs significantly better than any one of the simpler models on validation data except $\mathcal{M}_{G\alpha\eta}$ where the difference is too small and lies inside a region of two standard errors of the mean. This confirms that most of the model features are required to provide an accurate description of the experimental data.

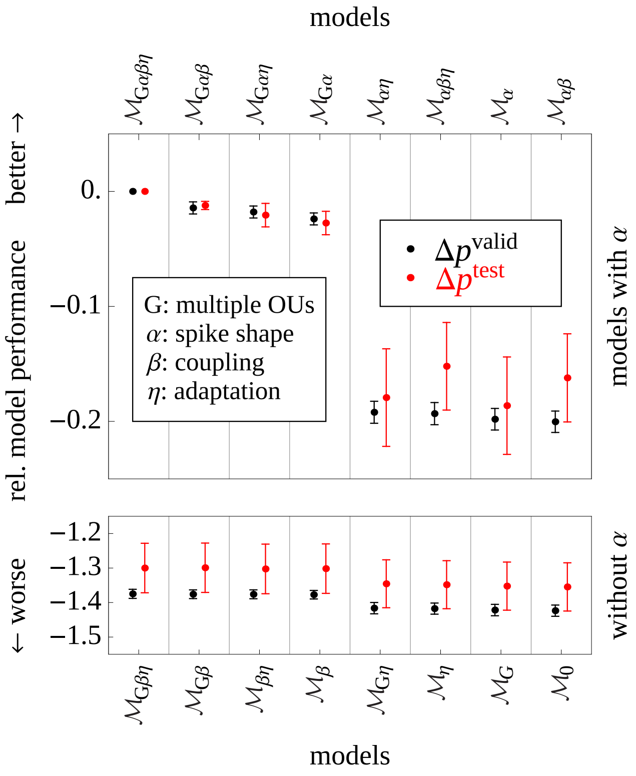


Figure 2.6: Comparison of the different models on dataset $\mathcal{D}_2$. The relative measure of model performance, i.e. the per-bin log-likelihood Δp (see Eq. (2.13)) between any model and the most complex model ($\mathcal{M}_{G\alpha\beta\eta}$) are significantly negative (with exception of $\mathcal{M}_{G\alpha\eta}$, and trivially $\mathcal{M}_{G\alpha\beta\eta}$) , implying that the added complexity improves the model fit without overfitting. This holds for both validation scores Δp^{valid} (black) and scores from unseen test data Δp^{test} (red). Error bars denote one standard error of the mean (S.E.M.). The biggest improvement of fit quality is achieved by including the spike-related kernel (upper vs. lower part of the figure).

2.3 Discussion

In this study, we introduced the AGAPE generative model for single-neuron statistics in order to describe the spontaneous dynamics of the somatic potential without reference to an input current. We showed that this model has a rich dynamical repertoire and can be fitted to data efficiently. By fitting a heterogeneous set of data, we finally demonstrated that the

AGAPE model can be used for the systematic characterization and comparison of in vivo intracellular recordings.

2.3.1 Flexibility and tractability of the model

The AGAPE model provides a unified description of intracellular dynamics, offering a large degree of flexibility in accounting for the distinct statistical features of a neuron. As the example datasets demonstrate, the model readily teases apart the differences in the statistics which exist between different cells in different animals (see Fig. 2.5). This shows that the model is sensitive enough to distinguish between datasets which are in fact very similar.

We used a set of approximations and techniques to make the model fitting tractable, despite the non-concavity of the log likelihood function. It is still the case that multiple local maxima of the likelihood function can make the fitting somewhat hard, especially if the quantity of data available for fitting is quite low. However, since one run of the fitting itself takes only a few minutes even on a portable computer, multiple initializations can be tried out in a relatively short amount of time.

2.3.2 Comparison with existing models

From an operational perspective, existing spiking neuron models can be divided into three main categories: stimulus-driven, current-driven and input-free spiking neurons. The first category contains phenomenological models that relate sensory stimuli to the spiking output of the neuron. The linear-nonlinear-Poisson model (LNP) (Chichilnisky, 2001), the generalized linear model (GLM) (Truccolo et al., 2005; Pillow et al., 2008; Paninski et al., 2009) or the GLM with additional latent variables (Vidne et al., 2012) are typical examples in this category. Even though the spike generation of the AGAPE shares some similarities with those models, there is an important distinction to make. In those models the convolved input (i.e. the output of the 'L' step of the LNP or the input filter of the GLM) is an internal variable that does not need to be mapped to the somatic membrane potential whereas in our case, the detailed modeling of the membrane potential dynamics is an important part of the AGAPE. Consequently, those phenomenological models are descriptions of extracellular spiking recordings whereas the AGAPE models the dynamics of the full membrane potential accessible with intracellular methods.

The second class of spiking models aims at bridging the gap between the input current and the spiking output. The rather simple integrate-and-fire types of models such as the exponential integrate-and-fire (Brette and Gerstner, 2005) or the spike-response model (Gerstner and Kistler, 2002; Jolivet et al., 2006) as well as the more biophysical models such as the Hodgkin-Huxley model (Hodgkin and Huxley, 1952) fall within this category. In contrast to those models where the action potentials are caused by the input current, the AGAPE produces a fluctuating membrane potential and stochastic spikes without a reference to an input current.

The last category of models aims at producing spontaneous spiking activity without an explicit dependence to a given input (Cunningham et al., 2007; Pfister et al., 2010; Macke et al., 2011). For example, Cunningham et al. propose a doubly stochastic process where the spiking generation is given by a gamma interval process and the firing intensity by a rectified Gaussian process, which provides a flexible description of the firing statistics (Cunningham et al., 2007). However, the membrane potential dynamics is not modeled. In opposition, the neuronal dynamics assumed by Pfister et al. (Pfister et al., 2010) models explicitly the membrane potential (as a simple Ornstein-Uhlenbeck process) but is not flexible enough to capture the dynamics of *in vivo* recordings. Also any of the current-driven spiking neuron models mentioned above can be turned into an input-independent model by assuming some additional input noise. So why is there a need to go beyond stochastic versions of those models? An integrate-and-fire model with additive Gaussian white noise is certainly fittable, but does not have the flexibility to model arbitrary autocorrelation for the membrane potential. At the other end of the spectrum, a Hodgkin-Huxley model with some colored noise would certainly be able to model a richer dynamical repertoire, but the fitting of it remains challenging (Gerstner and Naud, 2009) (but see (Druckmann et al., 2007)). The main advantage of the AGAPE is that it is at the same time very flexible and easily fittable. The flexibility mostly comes from the fact that any covariance function can be assumed for the GP process. The relative ease of fitting comes from the circulant approximation as well as from the presence of concave subspaces in the full parameter space.

Another distinct feature of our model with respect to other existing models is the explicit modeling of the spike-related trajectory instead of the spike-triggered average (as e.g. in Mensi et al. (2012)). Even though both concepts share similarities - both would capture a sudden and strong input that lead to a spike - there is an important distinction. The spike-triggered average also captures the (possibly smaller) upward fluctuations of the membrane potential which causes the spike while the spike-related kernel α precisely avoids capturing those fluctuations, letting the GP kernel explain them.

So if we removed the spike-triggered average e.g. in synthetic data where the true coupling parameter β is large, we would also remove the characteristic upward fluctuation of the membrane potential which causes the spike. By doing so, the fitting procedure would not find the correct relation between the values of the membrane potential and the observed spike patterns and therefore choose a β close to zero. Thus, if something has to be removed around an action potential (and our model comparison, Fig. 2.6, demonstrates convincingly that this is necessary), the formulation of the model demands that it is parametrically adjustable. This is the main reason why in our model framework the spike-triggered average has to be rejected as a viable extraction method. Note that if the true coupling parameter β is close to zero, the spike-triggered average is close to the extracted spike-related kernel α. For data where the action potential shape shows considerable variability, the model could be generalized to include a stochastic or a history-dependent spike-related kernel.

2.3.3 Extensions and future directions

Despite the focus of the present work on single-neuron spontaneous dynamics, the AGAPE model admits a straightforward inclusion of both stimulus-driven input and recurrent input. The inclusion of stimulus-driven input is similar as for the GLM model and allows the model to capture the neuronal correlate of stimulus-specific computation. The recurrent input makes the framework adaptable to multi-neuron recordings *in vivo*. While intracellular recordings from many neurons *in vivo* are very hard to perform, the rapid development of new recording techniques (e.g. voltage-sensitive dyes) makes the future availability of subthreshold data with sufficient time-resolution at least conceivable. The full-fledged model would allow questions regarding the relative importance of background activity, recurrent activity due to computation in the circuit, and activity directly evoked by sensory stimuli to be answered in a systematic way. In this setup, the contribution of the GP-distributed membrane potential to the overall fluctuations would be reduced (since it has to capture less unrecorded neurons) while the contribution of the recorded neurons would increase. This modified model can be seen as a generalization of the stochastic spike-response model (Gerstner and Kistler, 2002) or a generalization of the GLM (if the internal variable of the GLM is interpreted as the membrane potential).

So far, we assumed that weak synaptic inputs are captured by the Gaussian process while the strong inputs that lead to the spikes are captured by the spike-related kernel α. A straightforward extension of the model would be to consider additional intermediate inputs that cannot be captured by the GP nor by the spike-related kernel α but that can drive the neuron to emit (with a given probability) an action potential. Those intermediate input could be modeled as filtered Poisson events. The inclusion of those latent events would increase the complexity of the model and at the same time change some of the fitted parameters. In particular, we expect that it would increase the coupling β between the membrane potential and the firing rate and reduce the optimal delay δ between the decision time and the peak of the action potential. This could also provide a better way to separate the subthreshold dynamics (which depends on the input activity) from the suprathreshold dynamics (which would depend only on the neuron dynamics, and not on the strong inputs that it receives, as it is the case now).

A central assumption of our model is that of a Gaussian marginal distribution of the subthreshold potential. Although it is remarkably valid for the dataset considered here (i.e. the HVC dataset $\mathcal{D}_1$ see also Fig. 2.4B), datasets characterized by a distinctly non-Gaussian voltage distribution even after spike-related kernel removal are beyond the scope of the current model. In order to address this limitation, the Gaussian process could be extended to a different stochastic process, e.g. a nonlinear diffusion process, permitting non-Gaussian and in fact arbitrary marginal distributions. Moreover, a reset behavior similar to the one exhibited by an integrate-and-fire model (Brette and Gerstner, 2005) could be achieved with a non-stationary GP which features a mean which is reset after a spike. Both modifications would have a severe impact on the technical difficulty of model fitting. Therefore, the Gaussian assumption can be regarded as a useful compromise which is preferable over a perfect account for the skewness of the marginal distribution.

The spike-related kernel method to separate subthreshold and suprathreshold dynamics is an important feature of the model which is used to rid the membrane potential recording of stereotypic waveforms associated with a spike. The spike related kernel as modeled in the AGAPE has no bearing on the probability of the spikes, whereas the adaptation kernel η which modulates the firing rate after a spike is not visible in the somatic membrane potential dynamics. A simple extension of the model could include spike-triggered adaptation currents which affect both the somatic membrane potential as well as the firing intensity. Another possible extension is to allow the firing probability to depend on a filtered version of the subthreshold potential u instead of the instantaneous value of u at a time δ before the peak of the action potential. Both of the mentioned extensions would improve the biophysical interpretability of the AGAPE, but they would also vastly increase the number of parameters. Therefore, a model comparison would be required to determine what level of model complexity is required in order to characterize the statistics of the recording.

In the present study, the AGAPE was fit to different datasets of two different animals and brain regions. A systematic fitting to *in vivo* intracellular data from a wide range of animals and brain regions would constitute a classification scheme which does not only complement existing classifications of neurons which are based on electrophysiological, morphological, histological, and biochemical data; such as the one in Markram et al. (2004), but which is in direct relationship with the computational tasks the brain is facing *in vivo*.

Another application of the AGAPE could be in the context of a normative theory of short-term plasticity. Indeed, it has been recently hypothesized that short-term plasticity performs Bayesian inference of the presynaptic membrane potential based on the observed spike-timing (Pfister et al., 2010, 2009). According to this theory, short-term plasticity properties have to match the *in vivo* statistics of the presynaptic neuron. Since the AGAPE provides a realistic generative model of presynaptic activity under which inference is supposedly performed, our model can be used to make testable predictions on the dynamical properties of downstream synapses.

2.4 Materials and Methods

2.4.1 Description of the datasets used

1. Dataset $\mathcal{D}_1$ is a recording from a HVC neuron of an anesthetized Zebra Finch (Ondracek and Hahnloser, unpublished recordings). The recording has a total length of 270 seconds at 32 kHz (see Fig. 2.1A for a snippet of this recording) and contains 2281 action potentials.

2. Dataset $\mathcal{D}_2$ is another recording from a projection cell in HVC of Zebra Finch, but this time the animal is awake (Vallentin and Long, unpublished recordings). It consists of 6 individual recordings which together have a length of 152.5 seconds at 40 kHz. This dataset is used for model comparison (see below).

3. Dataset $\mathcal{D}_3$ is from similar conditions as $\mathcal{D}_2$ (Vallentin and Long, unpublished recordings, see (Long et al., 2010; Hamaguchi et al., 2014; Vallentin and Long, 2015) for similar recordings) and has a length of 60 seconds.

4. Dataset $\mathcal{D}_4$ consists of 19 individual trials of 4.95s duration at 20 kHz. The recording was obtained from a pyramidal neuron in layer 2/3 of awake mouse visual cortex (Haider et al., 2013).

2.4.2 Preprocessing

Intracellular voltage traces are often recorded at a rate between 20 and 40 kHz. This allows the action potentials to be resolved very clearly and precise spike timings to be extracted. However, for the study of the subthreshold regime, this high sampling rate is not required, and therefore the data may be down-sampled to roughly 1 kHz after obtaining the precise spike timings. Prior to down-sampling, we smooth with a median filter of 1ms width in order to truncate the sharp action potential peaks and avoid artifacts (see details below).

We define the spike peak times $\hat{t}_{i,\text{peak}}$ operationally as the time where the local maximum of the action potential is reached. This means that $\hat{t}_{i,\text{peak}}$ occurs after action potential onset, and hence the spike-related kernel has to extend to the past of $\hat{t}_{i,\text{peak}}$. The spike-related kernel starts at the nominal spike time $\hat{t}_i$ which is shifted from the peak time by a fixed amount δ, i.e. $\hat{t}_{i,\text{peak}} = \hat{t}_i + \delta$. The nominal spike times $\hat{t}_i$ are then binned to 1 ms, yielding a binary spike train $s_i = 0, 1$.

For $u_{\text{som}}(t)$ we use a preprocessed version of the recorded trace which has been median-filtered with a width of the filter of 1 ms and then down-sampled to 1 kHz, making it the same length as the binary spike train. This procedure preserves the relevant correlation structure of the membrane potential while reducing the computational demands of fitting as much as possible. In the data we examined, the median-filtered membrane potential has a dip after $\hat{t}_{i,\text{peak}}$, but unless down-sampling is done carefully, this dip sometimes occurs one timestep after $\hat{t}_{i,\text{peak}}$ and sometimes right at $\hat{t}_{i,\text{peak}}$ in the downsampled u_{som}. Since this dip will have to be captured by the spike-related kernel which has a fixed shape for all action potentials, the down-sampling procedure has to ensure that the dip occurs always in the first time-step. We solved this problem by setting the down-sampled value of u_{som} at $\hat{t}_{i,\text{peak}}$ (rounded to 1 ms) to the value of u_{som} at $\hat{t}_{i,\text{peak}}$ before down-sampling.

While applying the model to the raw recording u_{raw} directly (without first filtering and downsampling it) is possible in principle, it comes at a massively increased computational cost. In the interest of time required to fit the model and amount of data having to be handled, it is therefore sensible to include that pre-processing stage.

2.4.3 Parametrizations and initializations

We already introduced the parameters u_r, r_0 and β. Additional parameters are needed to describe the autocorrelation $k(t)$, the spike-related kernel $\alpha(t)$ and the adaptation kernel $\eta(t)$.

The covariance function of the GP has to be parametrized such that it can explain the autocorrelation structure of the data. Therefore, an initial examination of the empirical autocovariance of u_{som}, i.e.

$$\begin{aligned} k^{\text{emp}}(j\Delta t) = & \frac{1}{n-j-1} \sum_{i=1}^{n-j} \left(u_{\text{som},i} - \frac{1}{n-j} \sum_{k=1}^{n-j} u_{\text{som},k} \right) \\ & \times \left(u_{\text{som},i+j} - \frac{1}{n-j} \sum_{k=1}^{n-j} u_{\text{som},k+j} \right), \end{aligned} \tag{2.14}$$

for $j = 0, ..., j_{\text{max}}$, is done in order to determine a suitable basis. Here, we used a sum of Ornstein-Uhlenbeck (OU) kernels, i.e.

$$k(t) = \sum_{i=1}^{n_k} \sigma_i^2 e^{-\theta_i |t|}, \tag{2.15}$$

where $n_k = 10$ and $\theta_i = 2^{-i}$ ms^{-1}. The autocovariance has to remain positive definite. This induces the following linear constraints:

$$\hat{c}_i = \sum_{j=1}^{n_k} \sigma_j^2 \hat{c}_i^{(j)} > 0, \quad \forall i = 1, ..., n, \tag{2.16}$$

on σ_i^2, where $\hat{c}_i^{(j)}$ are the discrete Fourier transforms of the circulant basis vectors. The optimization problem is non-concave in the subspace of σ_i^2 and multiple local maxima and saddle points can occur. Therefore, multiple initializations have to be made in order to find a potential global optimum. In general, the least-squares fit of $k(t)$ to the empirical autocovariance function (2.14) yields a good starting point for the optimization.

The spike-rate adaptation kernel is chosen to be a linear combination of ten different alpha shapes

$$\eta(t) = \begin{cases} \sum_{i=1}^{n_\eta} w_i \left[\exp(-\nu_i t) - \exp(-\omega_i t)\right], & t > 0, \\ 0, & t \leq 0, \end{cases} \tag{2.17}$$

where we chose $n_\eta = 10$, $\nu_i = 2\omega_i$ and $\nu_i = 2^{-i}$ ms^{-1}.

Since the median filter time constant is short, the voltage change around the spike can be fast, requiring flexible spike-related kernel basis. Most of this flexibility is required around $t = \delta$. Because δ is adapted, we choose a discrete parametrization which has equal flexibility from $t = 0$ up to a maximum t. In our case, this maximum is at $t = 60$ ms, and therefore our parametrization of the spike-related kernel reads

$$\alpha(t) = \begin{cases} a_i, & if \quad t \in [i\Delta t, (i+1)\Delta t) \\ 0, & else \end{cases} \tag{2.18}$$

where $a_i \in \mathbb{R}$, $i = 1, ..., 60$ are the free parameters. Since the spike-related kernel fitting is concave, the large number of parameters does not lead to a dramatic increase of computational time. It also does not lead to overfitting, as is evidenced by the smoothness of the fitted kernels (see updated Figs.4,5 in the main text and the new S3-S5) and by the new model comparison results (see updated Fig.6 in the main text).

2.4.4 Model validation

We performed a factorial model comparison (see Fig. 2.6) where the four factors were the presence/absence of each of the following: multiple OU components in the GP autocorrelation function (see Eq. (2.15), as opposed to only one OU kernel with variable time-constant), the spike-related kernel α, coupling between u and s (through β) and adaptation η, which gives a total of 16 different models. We use the nomenclature that $\mathcal{M}_0$ is the simplest model, e.g. $\alpha = \beta = \eta = 0$ and only one OU component, having only four parameters (u_r, θ, σ and r_0). A subscript G (for GP) indicates that we use the multiple OU basis and any other subscript indicates that the corresponding parameter is adjustable in addition to the parameters already present in $\mathcal{M}_0$ and the parameters that are associated with the subscribed ones. E.g. $\mathcal{M}_{G\alpha}$ indicates that we use the multiple OU basis and allow a non-zero spike-related kernel and that there are now 73 parameters (δ, u_r, θ_i, a_i for $i = 1, ..., 60$, and $\log r_0$). The parameter δ is optimized only for the 12 out of 16 models which depend on this parameter, i.e. that have at least $\beta \neq 0$ or $\alpha \neq 0$.

For each of the models $\mathcal{M} \in \{\mathcal{M}_0, ..., \mathcal{M}_{G\alpha\beta\eta}\}$, we performed eight-fold cross-validation (Arlot and Celisse, 2010) on dataset $\mathcal{D}_2$ in order to assess the models' generalization performance. The entire dataset was cut into eight equally-sized chunks d_j, where $j = 1, ..., 8$, each of length 15s ($n = 15000$), and six chunks of 3s d'_j, $j = 1, ..., 6$ set aside as a test set ($n' = 3000$). Each model was then trained on seven out of eight chunks (treating them as independent samples) giving an optimal set of parameters $\Theta^i_j = \text{argmax}_\Theta\, p(\{d_k, k \neq j\} \,|\mathcal{M}_i, \Theta)$ and training per-bin log-likelihood $p^{\text{train}}_{ij} = \frac{1}{7n} \log p(\{d_k, k \neq j\} \,|\mathcal{M}_i, \Theta^i_j)$. Then the validation likelihood $p^{\text{valid}}_{ij} = \frac{1}{n} \log p(d_j|\mathcal{M}_i, \Theta^i_j)$ of the left-out chunk #j was evaluated. The unseen data d'_j is used for a final benchmark of model performance, where the best set of parameters is selected for each model, i.e. $p^{\text{test}}_{ij} = \frac{1}{n'} \max_{k=1,...,8} \log p(d'_j|\mathcal{M}_i, \Theta^i_k)$.

2.5 Acknowledgments

We would like to thank Janie Ondracek and Richard Hahnloser in Zürich, Switzerland, Daniela Vallentin and Michael Long at NYU, Bilal Haider and Matteo Carandini at UCL for kindly providing the data for this study. We also thank Máté Lengyel, Christian Pozzorini and Johanni Brea for helpful discussions.

2.6 Supplementary Text

2.6.1 Discrete Fourier Transform

In the following and in the main text, we denote discrete Fourier transforms of vectors of length n by a hat. The Fourier transformed vector is again of length n and can be formally expressed as

$$\hat{v} = \mathbb{F}v, \quad (\mathbb{F}_n)_{ij} = e^{\frac{2\pi I(i-1)(j-1)}{n}}, \quad i,j = 1,...,n, \tag{2.19}$$

where I denotes the imaginary unit. In practice, discrete Fourier transforms are not actually computed by matrix multiplication, but by means of a Fast Fourier Transform (FFT) algorithm.

2.6.2 Circulant matrices

In order to reduce the computational complexity of the likelihood estimation, we approximate the autocovariance matrix K (which is a Toeplitz matrix K) with a circulant matrix C. By definition a circulant matrix can be expressed as

$$C_{ij} = c_{(i-j \bmod n)+1} \tag{2.20}$$

we write $C = C_n(c)$. All circulant matrices of dimension n can be diagonalized by the unitary discrete Fourier transform matrix $\mathbb{U} = \frac{1}{\sqrt{n}}\mathbb{F}_n$:

$$C_n(c) = \mathbb{U}_n^\dagger \text{diag}(\mathbb{F}_n c)\mathbb{U}_n = \mathbb{U}_n^\dagger \text{diag}(\hat{c})\mathbb{U}_n, \tag{2.21}$$

where $\dagger$ is the conjugate transpose. This implies that $\hat{c}$ is the vector of eigenvalues of C, and it is a vector with real entries. This makes calculation of inverse and determinant of C extremely cheap, as is multiplication of C^{-1} by a vector $x \in \mathbb{R}^n$, which simplifies to

$$C_n^{-1}(c)x = \frac{1}{n}\mathbb{F}_n^\dagger \left(\frac{\hat{x}}{\hat{c}}\right) \tag{2.22}$$

where the vector in brackets is the component-wise quotient of the vectors $\mathbb{U}_n x$ and $\mathbb{F}_n c$.

2.6.3 Circulant approximation

The task is now to find a circulant matrix which is as close as possible to the covariance matrix K. This can be formalized as the following minimization problem:

$$C = \underset{D \text{ circulant}}{\text{argmin}}\ D_{\text{KL}}\left(\mathcal{N}(m,K)||\mathcal{N}(m,D)\right), \quad \forall m \tag{2.23}$$

where $\mathcal{N}$ denotes a multivariate Gaussian with specified mean vector and covariance matrix. This problem has the unique solution

$$c_i = \frac{1}{n}\left\{(n-i+1)k_i + (i-1)k_{n-i+2}\right\}, \quad 1 \le i \le n, \quad k_{n+1} \equiv 0 \tag{2.24}$$

Proof: The Kullback-Leibler divergence between two Gaussians is given by

$$D_{\mathrm{KL}}\left(\mathcal{N}(m,K)||\mathcal{N}(m,C)\right) = \frac{1}{2}\left[\operatorname{tr}(C^{-1}K) - \log\det(C^{-1}K)\right] - \frac{n}{2} \tag{2.25}$$

We have

$$C^{-1} = \mathbb{U}_n^\dagger \operatorname{diag}\left(\frac{1}{\hat{c}}\right)\mathbb{U}_n, \qquad \det C^{-1} = \prod_{i=1}^{n}\frac{1}{\hat{c}_i} \tag{2.26}$$

and hence

$$D_{\mathrm{KL}}\left(\mathcal{N}(m,K)||\mathcal{N}(m,C)\right) = \frac{1}{2}\sum_{i=1}^{n}\left[\frac{(\mathbb{U}_nK\mathbb{U}_n^\dagger)_{ii}}{\hat{c}_i} + \log\hat{c}_i\right] + \text{const.} \tag{2.27}$$

where the constant does not depend on c. We obtain the derivative

$$\frac{\partial}{\partial c_i}D_{\mathrm{KL}}\left(\mathcal{N}(m,K)||\mathcal{N}(m,C)\right) = \frac{1}{2\hat{c}_i}\left[1 - \frac{(\mathbb{U}_nK\mathbb{U}_n^\dagger)_{ii}}{\hat{c}_i}\right] = 0 \tag{2.28}$$

and therefore, at the stationary point we have

$$\begin{aligned}
\hat{c}_i &= (\mathbb{U}_nK\mathbb{U}_n^\dagger)_{ii} \\
c_i &= \frac{1}{n^2}\sum_{j,l,m=1}^{n}(\mathbb{F}_n^\dagger)_{ij}(\mathbb{F}_n)_{jl}K_{lm}(\mathbb{F}_n^\dagger)_{mj} \\
&= \frac{1}{n^2}\sum_{j,l,m=1}^{n}K_{lm}\exp\left[\frac{2\pi I}{n}\left(-(i-1)(j-1) + (j-1)(l-1) - (m-1)(j-1)\right)\right] \\
&= \frac{1}{n^2}\sum_{j,l,m=1}^{n}k_{|l-m|+1}\exp\left[\frac{2\pi I}{n}(j-1)(l-m+1-i)\right]
\end{aligned} \tag{2.29}$$

The sum of roots of unity over j only gives a non-zero value if the integer $q = l-m+1-i$ is a multiple of n. Since q has a maximum of $q = n-1$ when $l = n, m = i = 1$ and a minimum of $q = 2-2n$ when $l = 1, m = i = n$, only $q = 0$ and $q = -n$ are eligible. Hence,

$$\begin{aligned}
c_i &= \frac{1}{n^2}\sum_{l,m=1}^{n}nk_{|l-m|+1}\left(\delta_{0,l-m+1-i} + \delta_{-n,l-m+1-i}\right) \\
&= \frac{1}{n}\sum_{r=-n+1}^{n-1}(n-|r|)k_{|r|+1}\left(\delta_{i-1,r} + \delta_{i-1,n+r}\right) \\
&= \frac{1}{n}\left\{(n-i+1)k_i + (i-1)k_{n-i+2}\right\}, \quad k_{n+1} \equiv 0
\end{aligned} \tag{2.30}$$

The second equality is obtained by reparametrizing $r = l-m$. □

2.6.4 Sampling using FFTs

In order generate a sample u of length n from a multivariate Gaussian with mean vector m and circulant covariance matrix $C = C_n(c)$, one generates a zero-mean white noise vector x with unit variance and then uses FFTs to compute u, i.e.

$$u = m + \frac{1}{n}\mathbb{F}_n^{\dagger}\left[(\hat{c})^{1/2}\,\hat{x}\right] = m + \frac{1}{n}C_n\left(\mathbb{F}_n^{\dagger}\,(\hat{c})^{1/2}\right)x \tag{2.31}$$

as the following calculation shows, the covariance comes out correctly

$$\begin{aligned}\left\langle (u-m)(u-m)^T\right\rangle &= \frac{1}{n^2}C_n\left(\mathbb{F}_n^{\dagger}\,(\hat{c})^{1/2}\right)\left\langle xx^T\right\rangle C_n\left(\mathbb{F}_n^{\dagger}\,(\hat{c})^{1/2}\right)^T\\ &= \frac{1}{n}C_n\left(\mathbb{F}_n^{\dagger}\hat{c}\right) = C_n(c) = C\end{aligned} \tag{2.32}$$

2.6.5 Optimization method

The optimization scheme used is a quasi-Newton method, where the vector of parameters Θ is updated according to

$$\Theta^{(k+1)} = \Theta^{(k)} - B^{(k)}\nabla f\left(\Theta^{(k)}\right) \tag{2.33}$$

where f is the function to be minimized (e.g. $-\log p(u_{\text{som}}, s)$), ∇f is the gradient, and the matrix B is chosen to be

$$B^{(k)} = \begin{cases} H_f^{-1}\left(\Theta^{(k)}\right), & \text{if } H_f\left(\Theta^{(k)}\right) \text{ positive definite}\\ \gamma^{(k)}G^{-1}, & \text{else}\end{cases} \tag{2.34}$$

Where H_f is the Hessian of f, $\gamma^{(k)}$ denotes a learning rate, and G is a metric tensor on the parameter space which is used to rescale the parameters to lie in similar ranges. The learning rate $\gamma^{(k)} < 0$ is increased when the previous step was successful (typically, by 10 percent), and reduced when the optimizer either runs into boundaries of the admissible parameter region or increases the value of the function (we used a reduction by a factor of 2).

Below, we derive the formulae for the gradient and Hessian required for the optimization. The derivations hold for the case where the GP covariance function k is parametrized arbitrarily by θ_i and the spike-shape kernel and adaptation kernel are given by linear combinations of basis functions $\alpha^{(k)}$, $k = 1, ..., n_a$ and $\eta^{(k)}$, $k = 1, ..., n_w$ respectively.

2.6.6 Gradient

The likelihood function has the form

$$\begin{aligned}\log p(u_{\text{som}}, s) = \sum_{i=1}^{n}\Bigg[&-\frac{1}{2}\log(2\pi\hat{c}_i) - \frac{1}{2n}\frac{|\hat{u}_i|^2}{\hat{c}_i}\\ &+ s_i\log q_i + (1-s_i)\log\left[1-q_i\right]\Bigg],\end{aligned} \tag{2.35}$$

where

$$q_i = \Delta t e^{\beta u_i + A_i + \log r_0} \tag{2.36}$$

is the probability of a spike in bin i and depends on all parameters except the ones that parametrize the covariance function k. The membrane potential is given implicitly by

$$u = u_{\text{som}} - u_r - \alpha * s. \tag{2.37}$$

It is worthwhile to write the derivatives of $\log p$ in the following form:

$$\begin{aligned} d\log p(u_{\text{som}}, s) = \sum_{i=1}^{n} \Bigg[& -\frac{1}{2}\left(\frac{1}{\hat{c}_i} - \frac{1}{n}\frac{|\hat{u}_i|^2}{\hat{c}_i^2}\right) d\hat{c}_i \\ & - \frac{1}{n\hat{c}_i}\Re\left\{\hat{u}_i^* d\hat{u}_i\right\} + \frac{s_i - q_i}{q_i(1-q_i)} dq_i \Bigg], \end{aligned} \tag{2.38}$$

Now, let us evaluate all the terms one by one. The Fourier transform of the circulant covariance c only depends on the GP kernel parameters θ, i.e.

$$d\hat{c}_i = \frac{\partial \hat{c}_i}{\partial \theta_k} d\theta_k = \sum_{k=1}^{n_k} \left(\widehat{\frac{\partial c}{\partial \theta_k}}\right)_i d\theta_k. \tag{2.39}$$

Let us turn to the q terms next. Their differential is

$$\begin{aligned} dq_i &= \frac{\partial q_i}{\partial u_i} du_i + \frac{\partial q_i}{\partial A_i} dA_i + \frac{\partial q_i}{\partial r_0} dr_0 + \frac{\partial q_i}{\partial \beta} d\beta \\ &= q_i(\beta du_i + dA_i + d\log r_0 + u_i d\beta), \end{aligned} \tag{2.40}$$

where by (2.2) and using the fact that α is a linear combination of basis kernels $\sum_{k=1}^{n_\alpha} a_k \alpha^{(k)}$

$$du_i = -du_r - \sum_{k=1}^{n_\alpha} S_i^{(k)} da_k, \quad S_i^{(k)} = (\alpha^{(k)} * s)_i. \tag{2.41}$$

Moreover, A_i is also a linear combination, so

$$dA_i = \sum_{k=1}^{n_\eta} A_i^{(k)} dw_k, \quad A_i^{(k)} = (\eta^{(k)} * s)_i. \tag{2.42}$$

Therefore (2.40) can be written as

$$dq_i = q_i \left(-\beta du_r - \beta \sum_{k=1}^{n_\alpha} S_i^{(k)} da_k + \sum_{k=1}^{n_\eta} A_i^{(k)} dw_k + d\log r_0 + u_i d\beta \right), \tag{2.43}$$

Lastly, by (2.41) we also have

$$d\hat{u}_i = -n\delta_{1i} du_r - \sum_{k=1}^{n_\alpha} \hat{S}_i^{(k)} da_k. \tag{2.44}$$

Using

$$v_i = \frac{s_i - q_i}{1 - q_i}, \tag{2.45}$$

all the previous results can be regrouped to yield

$$\begin{aligned} d\log p(u_{\mathrm{som}}, s) = \sum_{i=1}^{n}\Bigg[& -\frac{1}{2}\left(\frac{1}{\hat{c}_i} - \frac{1}{n}\frac{|\hat{u}_i|^2}{\hat{c}_i^2}\right)\sum_{k=1}^{n_k}\left(\widehat{\frac{\partial \hat{c}_i}{\partial \theta_k}}\right)_i d\theta_k \\ & + \sum_{k=1}^{n_\alpha}\left(\frac{1}{n\hat{c}_i}\Re\left\{\hat{u}_i^*\hat{S}_i^{(k)}\right\} - \beta S_i^{(k)} v_i\right) da_k \\ & + \sum_{k=1}^{n_\eta} A_i^{(k)} v_i dw_k \\ & + v_i d\log r_0 \\ & + u_i v_i d\beta \\ & + \left(\frac{\delta_{1i}}{\hat{c}_i}\Re\left\{\hat{u}_i\right\} - \beta v_i\right) du_r \Bigg], \end{aligned} \tag{2.46}$$

2.6.7 Hessian

For the Hessian, we mainly need the following

$$-\frac{1}{2}d\left(\frac{1}{\hat{c}_i} - \frac{1}{n}\frac{|\hat{u}_i|^2}{\hat{c}_i^2}\right) = -\frac{1}{2}\left(\frac{1}{\hat{c}_i^2} - \frac{2}{n}\frac{|\hat{u}_i|^2}{\hat{c}_i^3}\right)d\hat{c}_i + \frac{1}{n\hat{c}_i^2}\Re\left\{\hat{u}_i^* d\hat{u}_i\right\}, \tag{2.47}$$

$$d\left(\frac{1}{n\hat{c}_i}\Re\left\{\hat{u}_i^*\hat{S}_i^{(k)}\right\}\right) = -\frac{1}{n\hat{c}_i^2}\Re\left\{\hat{u}_i^*\hat{S}_i^{(k)}\right\}d\hat{c}_i + \frac{1}{n\hat{c}_i}\Re\left\{\hat{S}_i^{*(k)}d\hat{u}_i\right\}, \tag{2.48}$$

$$dv_i = d\left(\frac{s_i - q_i}{1 - q_i}\right) = \frac{s_i - 1}{(1 - q_i)^2}dq_i = \frac{\xi_i}{q_i}dq_i, \tag{2.49}$$

where

$$\frac{\xi_i}{q_i} = \frac{s_i - 1}{(1 - q_i)^2}. \tag{2.50}$$

The components of the Hessian matrix are computed as follows

$$\begin{aligned} \frac{\partial^2 \log p(u_{\mathrm{som}}, s)}{\partial\theta_k\partial\theta_l} = & -\frac{1}{2}\sum_{i=1}^{n}\Bigg\{\frac{\partial^2\hat{c}_i}{\partial\theta_k\partial\theta_l}\left[\frac{1}{\hat{c}_i} - \frac{1}{n}\left|\frac{\hat{u}_i}{\hat{c}_i}\right|^2\right] \\ & - \frac{\partial\hat{c}_i}{\partial\theta_k}\frac{\partial\hat{c}_i}{\partial\theta_l}\left[\frac{1}{\hat{c}_i^2} - \frac{2}{n}\frac{|\hat{u}_i|^2}{\hat{c}_i^3}\right]\Bigg\}, \end{aligned} \tag{2.51}$$

$$\frac{\partial^2 \log p(u_{\mathrm{som}}, s)}{\partial\theta_k\partial a_l} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{c}_i}{\partial\theta_k}\Re\left\{\frac{\hat{u}_i^*\hat{S}_i^{(l)}}{\hat{c}_i^2}\right\}, \tag{2.52}$$

$$\frac{\partial^2 \log p(u_{\mathrm{som}}, s)}{\partial\theta_k\partial u_r} = -\frac{\partial\hat{c}_1}{\partial\theta_k}\Re\left\{\frac{\hat{u}_1}{\hat{c}_1^2}\right\}, \tag{2.53}$$

$$\frac{\partial^2 \log p(u_{\mathrm{som}}, s)}{\partial\theta_k\partial w_k} = \frac{\partial^2 \log p(u_{\mathrm{som}}, s)}{\partial\theta_k\partial r_0} = \frac{\partial^2 \log p(u_{\mathrm{som}}, s)}{\partial\theta_k\partial\beta} = 0, \tag{2.54}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial a_k \partial a_l} = \sum_{i=1}^{n} \left[-\frac{1}{n} \Re \left\{ \frac{\hat{S}_i^{*(k)} \hat{S}_i^{(l)}}{\hat{c}_i} \right\} + \beta^2 S_i^{(k)} S_i^{(l)} \xi_i \right], \tag{2.55}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial a_k \partial w_l} = -\beta \sum_{i=1}^{n} S_i^{(k)} A_i^{(l)} \xi_i, \tag{2.56}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial a_k \partial u_r} = -\Re \left\{ \frac{\hat{S}_1^{(k)}}{\hat{c}_1} \right\} + \beta^2 \sum_{i=1}^{n} S_i^{(k)} \xi_i, \tag{2.57}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial a_k \partial \log r_0} = -\beta \sum_{i=1}^{n} S_i^{(k)} \xi_i, \tag{2.58}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial a_k \partial \beta} = -\beta \sum_{i=1}^{n} S_i^{(k)} \xi_i u_i, \tag{2.59}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial w_k \partial w_l} = \sum_{i=1}^{n} A_i^{(k)} A_i^{(l)} \xi_i, \tag{2.60}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial w_k \partial u_r} = -\beta \sum_{i=1}^{n} A_i^{(k)} \xi_i, \tag{2.61}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial w_k \partial \log r_0} = \sum_{i=1}^{n} A_i^{(k)} \xi_i, \tag{2.62}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial w_k \partial \beta} = \sum_{i=1}^{n} A_i^{(k)} \xi_i u_i, \tag{2.63}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial u_r^2} = -\frac{n}{\hat{c}_1} + \beta^2 \sum_{i=1}^{n} \xi_i, \tag{2.64}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial u_r \partial \log r_0} = -\beta \sum_{i=1}^{n} \xi_i, \tag{2.65}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial u_r \partial \beta} = -\sum_{i=1}^{n} \left[v_i + \beta u_i \xi_i \right], \tag{2.66}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial (\log r_0)^2} = \sum_{i=1}^{n} \xi_i, \tag{2.67}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial \log r_0 \partial \beta} = \sum_{i=1}^{n} u_i \xi_i, \tag{2.68}$$

$$\frac{\partial^2 \log p(u_{\text{som}}, s)}{\partial \beta^2} = \sum_{i=1}^{n} u_i^2 \xi_i. \tag{2.69}$$

The components as given by equations (33-51) are then combined to form the Hessian matrix H_f, which is used in Eq. (2.34).

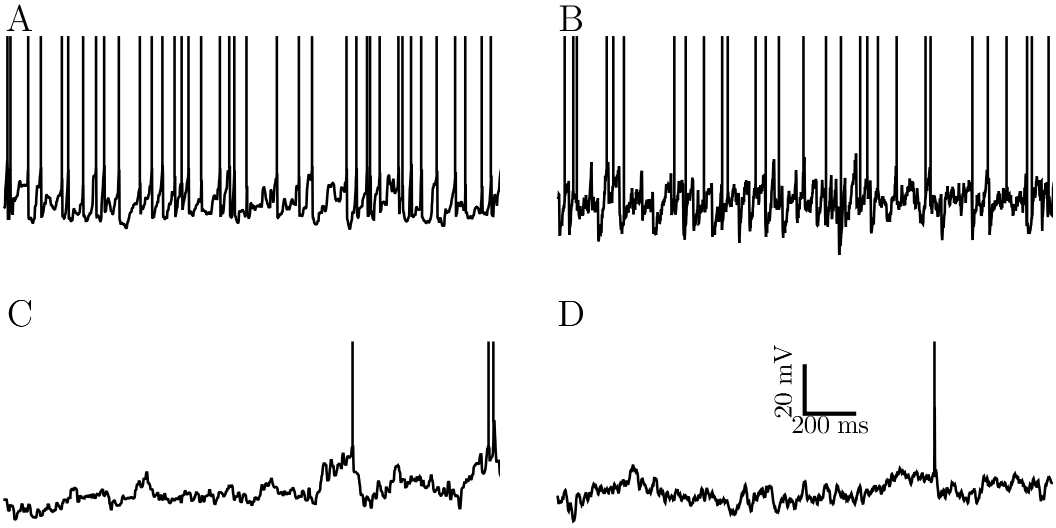


Figure 2.7: **S2 Fig. Supplementary Figure.** Comparison of *in vivo* and artificial data snippets for datasets $\mathcal{D}_3$ and $\mathcal{D}_4$, analogous to Fig. 3G,H. The scale (shown on panel D) is the same for all four panels. Vertical lines are drawn at the spiking times. (A) A 2-second sample of *in vivo* activity from dataset $\mathcal{D}_3$ (Zebra Finch HVC). (B) Artificial data sampled from AGAPE with parameters learned from dataset $\mathcal{D}_3$. (C) A 2-second sample of *in vivo* activity from dataset $\mathcal{D}_4$ (mouse visual cortex). (D) Artificial data sampled from AGAPE with parameters learned from dataset $\mathcal{D}_4$.

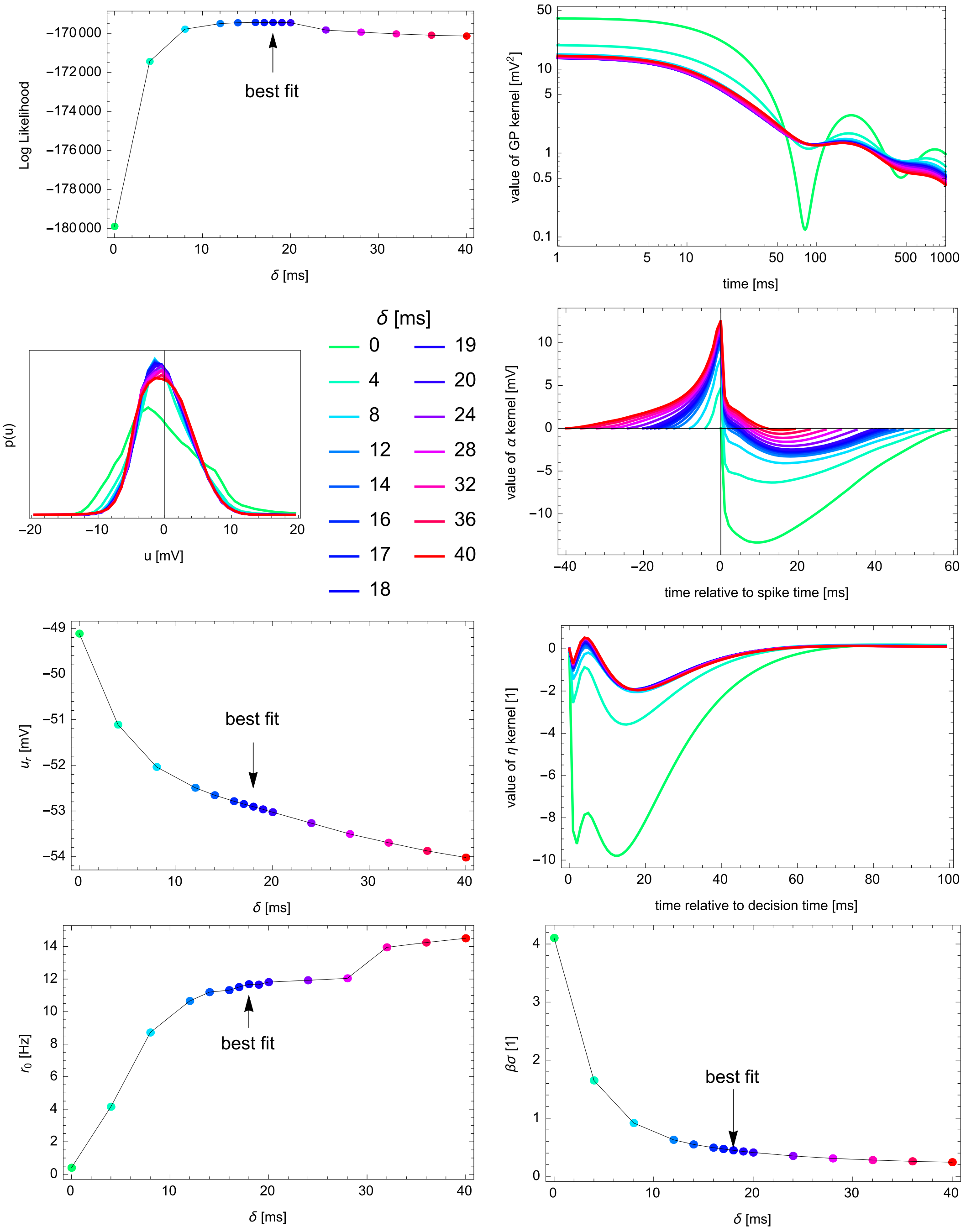


Figure 2.8: **S3 Fig. Supplementary Figure.** The fitting result as a function of the parameter δ for dataset $\mathcal{D}_1$, see color code next to the plot of the marginal distribution of u in the second row of the left column. The top left panel shows that the log likelihood peaks at $\delta = 18$ ms, and the bottom right panel shows the decrease of the effective coupling stength as δ increases.

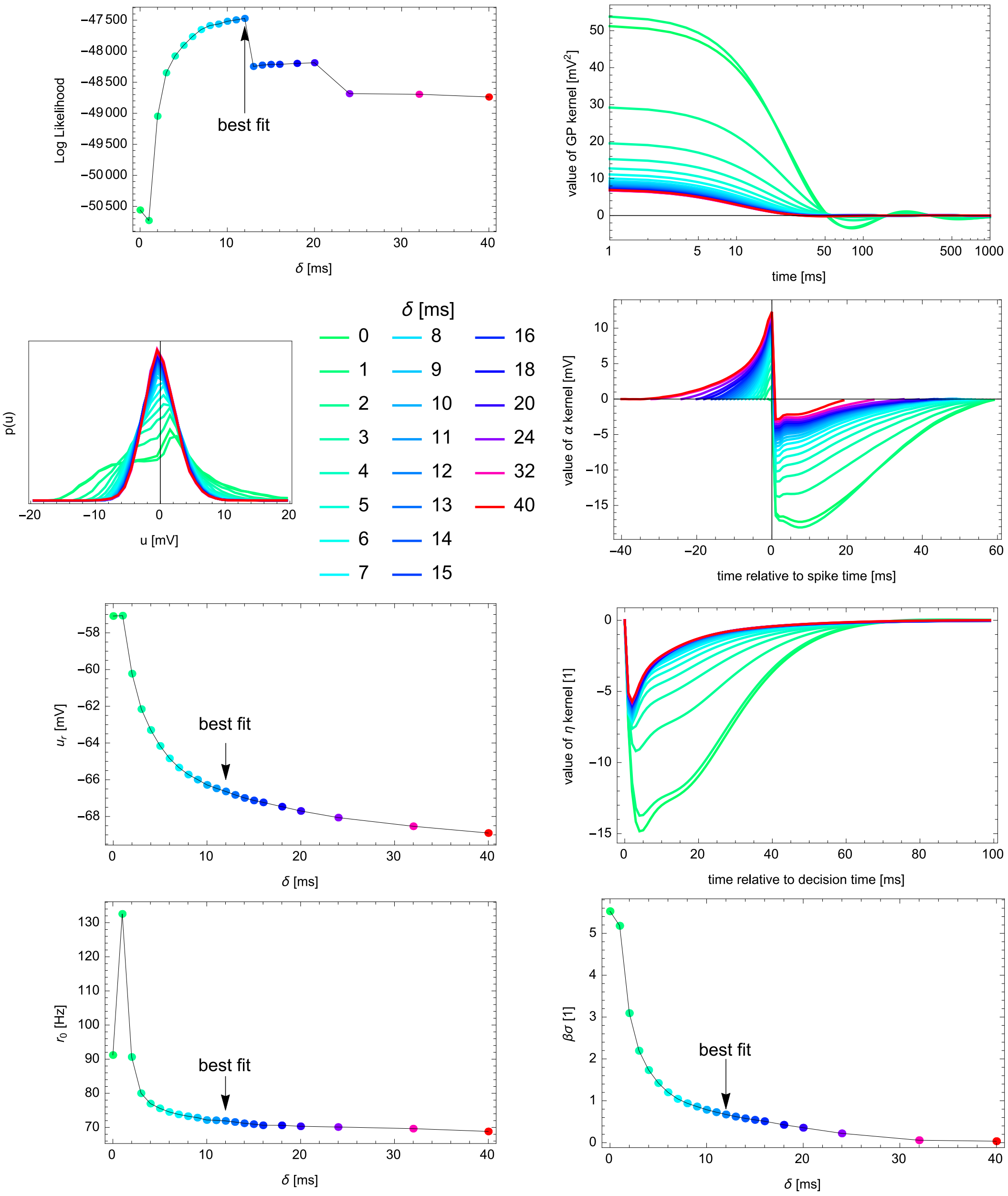


FIGURE 2.9: **S4 Fig. Supplementary Figure.** The fitting result as a function of the parameter δ for dataset $\mathcal{D}_3$, see color code next to the plot of the marginal distribution of u in the second row of the left column. The top left panel shows that the log likelihood peaks at $\delta = 12$ ms, and the bottom right panel shows the decrease of the effective coupling stength as δ increases.

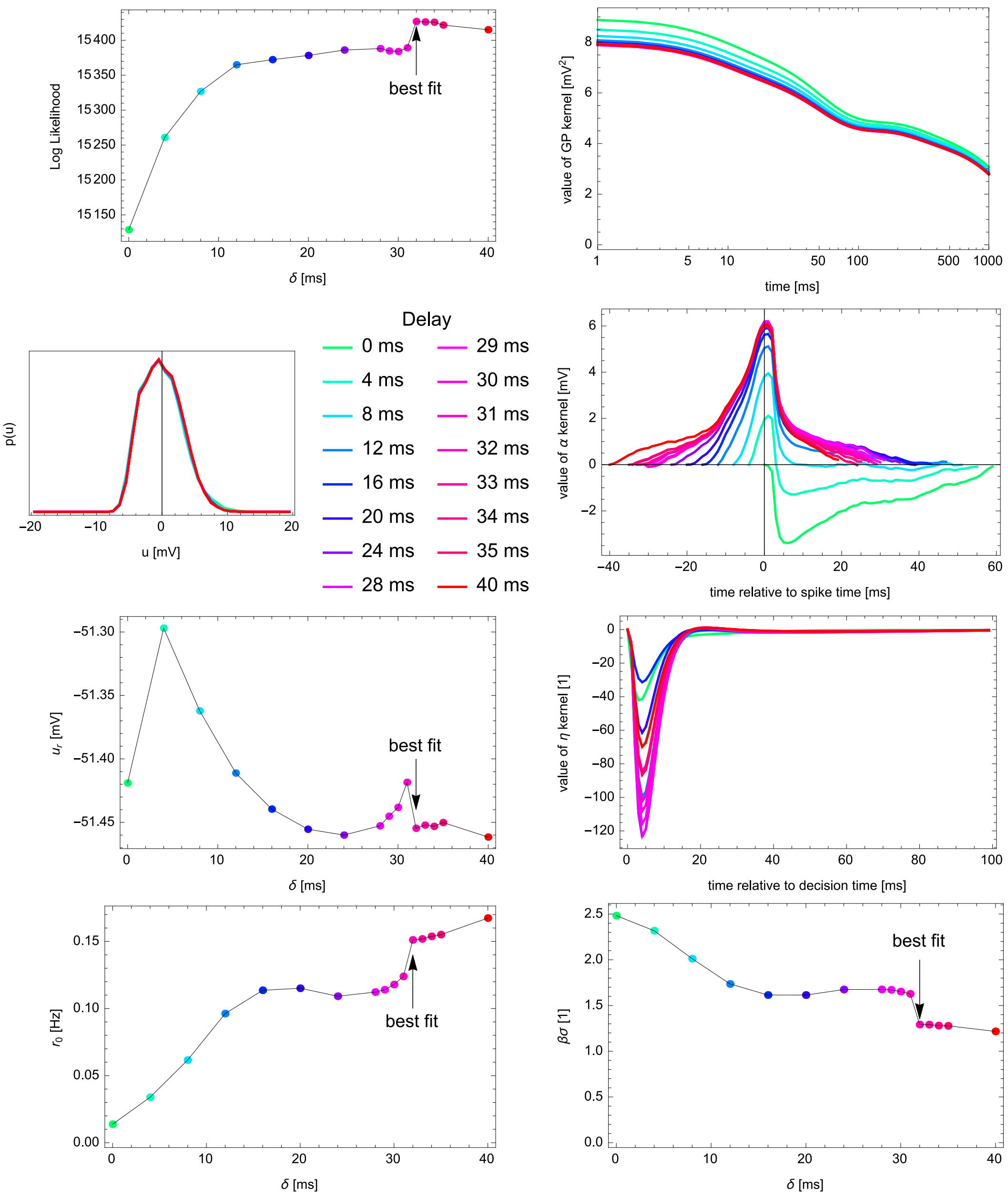


Figure 2.10: **S5 Fig. Supplementary Figure.** The fitting result as a function of the parameter δ for dataset $\mathcal{D}_4$, see color code next to the plot of the marginal distribution of u in the second row of the left column. The top left panel shows that the log likelihood peaks at $\delta = 32$ ms, and the bottom right panel shows the decrease of the effective coupling stength as δ increases.

Extensions of a Functional Theory of Short-term Plasticity

3

This chapter contains the contributions and extensions to the theory of short-term plasticity of Pfister et al. (2009). We have reviewed the basic phenomenology and biophysics of short-term synaptic plasticity in the Introduction, and we have summarized the main functional roles which have been proposed for it, including the original theory of Pfister et al. (2009). Here, we present a mathematical reformulation in terms of stochastic filtering theory which sheds new light on the previous results. This leads to important extensions of the theory and new predictions.

The mathematical parts of this chapter make heavy use of stochastic calculus and stochastic filtering theory. Although very powerful and elegant, they may be scarcely accessible to a lot of readers. We therefore encourage readers which are unfamiliar with these topics to take a look at Appendix A.1, where we offer a brief review and references to relevant literature.

3.1 Mathematical Reformulation of the 'Know Thy Neighbour' Theory

The 'Know Thy Neighbour' (KTN) theory of short-term plasticity which we introduced in Section 1.4 was originally formulated as recursive Bayesian inference problem in Pfister et al. (2009), and the inference calculations were carried out in an approximate fashion and in discrete time. In the end, the timestep was taken to zero in a delicate limit procedure, leading to the main results. Here, we reformulate the mathematics of the theory using filtering theory. This allows us to carry out the inference in continuous time and in an exact way, leading to a stochastic partial differential equation for the posterior probability density. Indeed, this method is by no means new, but goes back to a standard result of filtering theory for diffusion process observations by Zakai (1969). The corresponding result for point process observations was derived by Snyder (1972). Using this powerful formalism, we will recover the results from the 2009 paper by using a Gaussian ansatz to solve the filtering equation approximately. In addition, we obtain new insights into the filtering problem and the estimation error which is incurred by using a Gaussian approximation.

3.1.1 The filtering problem within the KTN theory

The KTN theory fits remarkably well into the setting of filtering theory. In this context, the signal of interest is the membrane potential of the presynaptic neuron and we model it as a diffusion process X_t. The measurements are the spike train of the presynaptic neuron, which we model as a counting process N_t (counting the number of spikes which have occured since time zero), which depends on the membrane potential. The synapse – having only access to the presynaptic spikes, i.e. the measurement process N_t – has the task of extracting information about the presynaptic membrane potential X_t. We will derive the optimal Bayesian solution to this filtering problem, and then find a possible synaptic implementation of this solution.

Let us now formalize this idea. As already noted, the presynaptic membrane potential is modeled as a (one-dimensional) diffusion process, which is a solution to the Itô stochastic differential equation (SDE)

$$dX_t = a(X_t)dt + b(X_t)dW_t. \tag{3.1}$$

Here, W_t denotes a standard Wiener process and the functions a and b are called *drift* and *diffusion* terms. As a diffusion process, its one-time probability density $\pi(x,t)$ satisfies the Fokker-Planck equation (FPE)

$$\partial_t \pi(x,t) = \mathcal{L}^\dagger \pi(x,t), \tag{3.2}$$

where

$$\mathcal{L}^\dagger = -\partial_x \left[a(x)\cdot\right] + \frac{1}{2}\partial_x^2 \left[b^2(x)\cdot\right] \tag{3.3}$$

is the Fokker-Planck operator of the diffusion process.

The spiking process is an inhomogeneous Poisson process– also called Cox process, see Cox (1955) – the rate of which depends on the value of the membrane potential. The value of the counting process N_t counts the numbers of spikes which have occured up to and including time t and may be written as

$$N_t = \sum_{n=1}^{\infty} H(t - T_n), \tag{3.4}$$

where H is the Heaviside function ($H(0) = 0$) and $T_i \geq 0$ are the spike arrival times. Let $g : \mathbb{R} \to \mathbb{R}$ be a nonnegative function, called *gain function*. The probability of a certain sequence of spikes reads

$$N_t - N_{t'} \sim \text{Poisson}\left[\int_{t'}^{t} g(X_s)ds\right], \tag{3.5}$$

which has to hold for all $t > t'$.

The subsequent derivations of the solution to the filtering problem will not depend on the particular choice of the Fokker-Planck operator $\mathcal{L}^\dagger$ and the function g and are therefore much more general than the derivations in Pfister et al. (2009), which relied on the assumptions in eqs. (3.30) and (3.31).

3.1.2 Formal Solution of the Filtering Problem

As already mentioned, a full probabilistic solution of the filtering problem is one which allows us to compute the posterior expectation, i.e. the expectation at time t conditioned on the observations up to time t, of any function φ,

$$p_t\left[\varphi\right] = \mathbb{E}\left[\varphi(X_t)|N_{[0,t]}\right]. \tag{3.6}$$

If the posterior measure p_t has a density $p(x,t)$, all posterior expectations can be calculated from it as

$$p_t\left[\varphi\right] = \int_{-\infty}^{\infty} p(x,t)\varphi(x)dx. \tag{3.7}$$

The goal of this section[1] is to derive an equation for the time-evolution of the posterior density $p(x,t)$, in a similar way as the FPE (3.2) is a time-evolution equation for the *prior* probability density $\pi(x,t)$. The equation we are looking for will depend on the measurement process N_t, which provides additional information which is not included in the definition of the signal process.

We begin by writing down the joint law[2] of the process (X_t, N_t) in a time interval $[0,t]$ (we use the notation $X_{[0,t]}$ to denote the trajectory of X on the interval $[0,t]$), which reads

$$\begin{aligned} P\left(X_{[0,t]}, N_{[0,t]}\right) &= P\left(X_{[0,t]}\right) P\left(N_{[0,t]}|X_{[0,t]}\right) \\ &= P\left(X_{[0,t]}\right) \exp\left[-\int_0^t g(X_s)ds\right] \prod_{0\le T_n\le t} g(X_{t_n}). \end{aligned} \tag{3.8}$$

Following the reference measure approach to nonlinear filtering (see App. A.1), we want to change the measure to an alternate probability Q, which preserves the marginal distribution of the process X_t, but transforms the distribution of N_t into a homogeneous Poisson process with rate $g_0 > 0$. The new joint probability reads

$$\begin{aligned} Q\left(X_{[0,t]}, N_{[0,t]}\right) &= Q\left(X_{[0,t]}\right) Q\left(N_{[0,t]}\right) \\ &= Q\left(X_{[0,t]}\right) \exp\left[-g_0 t\right] \prod_{0\le T_n\le t} g_0. \end{aligned} \tag{3.9}$$

The likelihood-ratio of the two joint distributions – also known as the Radon-Nikodym derivative of the change of measure from Q to P – reads

$$\begin{aligned} L_t &\doteq \frac{P\left(X_{[0,t]}, N_{[0,t]}\right)}{Q\left(X_{[0,t]}, N_{[0,t]}\right)} = \frac{P\left(N_{[0,t]}|X_{[0,t]}\right)}{Q\left(N_{[0,t]}\right)} \\ &= \exp\left[\int_0^t \left(g_0 - g(X_s)\right) ds\right] \prod_{0\le T_n\le t} \frac{g(X_{T_n})}{g_0}. \end{aligned} \tag{3.10}$$

The variable L_t has jumps at the arrival times T_n which are given by multiplication of L_{t-} with $g(X_{T_n})/g_0$. In-between arrival times, L_t solves an ordinary differential equation $\dot{L}_t = \left(g_0 - g(X_t)\right) L_t$. This can be summarized in terms of an SDE for L_t as

$$dL_t = \left(\frac{g(X_t)}{g_0} - 1\right) L_t(dN_t - g_0 dt), \quad L_0 = 1. \tag{3.11}$$

---

[1] We are not asking for an explicit, closed-form solution for the posterior probability density $p(x,t)$, which would be much too hard. Note that even the FPE rarely admits a closed-form solution.

[2] We use the term ‘law’ loosely to mean the probability density of the trajectory of the process.

The change of measure from P to Q is convenient because it allows us to derive an evolution equation for the unnormalized measure

$$\rho_t\left[\varphi\right] = \mathbb{E}_Q\left[\varphi(X_t)L_t|N_{[0,t]}\right] \tag{3.12}$$

from which the normalized measure can be recovered through the so-called Kallianpur-Striebel formula, a version of Bayes' theorem

$$p_t\left[\varphi\right] = \frac{\rho_t\left[\varphi\right]}{\rho_t\left[1\right]}. \tag{3.13}$$

In order to compute the stochastic derivative of ρ_t, we note that the two factors under the expectation do not share any stochastic terms dN_t or dW_t, and that therefore[3]

$$d\rho_t\left[\varphi\right] = \mathbb{E}_Q\left[d\varphi(X_t)L_t\middle|N_{[0,t]}\right] + \mathbb{E}_Q\left[\varphi(X_t)dL_t\middle|N_{[0,t]}\right]. \tag{3.14}$$

The first term can be simplified using Itô's lemma,

$$\begin{aligned} d\varphi(X_t) &= \varphi'(X_t)dX_t + \frac{1}{2}\varphi''(X_t)dX_t^2 \\ &= a(X_t)\varphi'(X_t)dt + \frac{1}{2}b^2(X_t)\varphi''(X_t)dt + \mathcal{O}\left(dW_t\right) \\ &= (\mathcal{L}\varphi)(X_t)dt + \mathcal{O}\left(dW_t\right). \end{aligned} \tag{3.15}$$

In the third line, we recognized the combined operator acting on φ to be the generator $\mathcal{L}$ of the process X,

$$\mathcal{L} = a(x)\partial_x + \frac{1}{2}b^2(x)\partial_x^2, \tag{3.16}$$

the adjoint of the Fokker-Planck operator $\mathcal{L}^\dagger$ of X, see Eq. (3.3). Under the measure Q the dW_t term does not contribute under the expectation because it is independent of $N_{[0,t]}$. Therefore, we have

$$\mathbb{E}_Q\left[d\varphi(X_t)L_t\middle|N_{[0,t]}\right] = \rho_t\left[\mathcal{L}\varphi\right]dt. \tag{3.17}$$

The second term of Eq. (3.14) can be calculated by substituting (3.11):

$$\begin{aligned} \mathbb{E}_Q\left[\varphi(X_t)dL_t\middle|N_{[0,t]}\right] &= \mathbb{E}_Q\left[\varphi(X_t)\left(\frac{g(X_t)}{g_0}-1\right)L_t(dN_t - g_0dt)\middle|N_{[0,t]}\right] \\ &= \mathbb{E}_Q\left[\varphi(X_t)\left(\frac{g(X_t)}{g_0}-1\right)L_t\middle|N_{[0,t]}\right](dN_t - g_0dt) \\ &= \left(\frac{1}{g_0}\rho_t\left[g\varphi\right] - \rho_t\left[\varphi\right]\right)(dN_t - g_0dt). \end{aligned} \tag{3.18}$$

In the second line, we recognized that N_t is a Poisson process of rate g_0 under the measure Q and that therefore dN_t is independent of $N_{[0,t]}$. In the third line, we expressed the expectations using Eq. (3.12). In summary, we can rewrite Eq. (3.14) as

$$d\rho_t\left[\varphi\right] = \rho_t\left[\mathcal{L}\varphi\right]dt + \left(\frac{1}{g_0}\rho_t\left[g\varphi\right] - \rho_t\left[\varphi\right]\right)(dN_t - g_0dt). \tag{3.19}$$

---

[3] We have $\mathbb{E}_Q\left[\varphi(X_t)L_t|N_{[0,t]}\right] = \mathbb{E}_Q\left[\varphi(X_t)L_t|N_{[0,T]}\right]$ for some $T > t$ (see Xu (2011), Proposition 2.7), which means that the stochastic differential can be taken inside the expectation. Note that our derivation of the filtering equation is merely heuristic, for a rigorous proof see Xu (2011), Theorem 2.9.

Now, we may introduce an unnormalized posterior density $\rho(x,t)$ with the property that

$$\rho_t\left[\varphi\right] = \int_{-\infty}^{\infty} \rho(x,t)\varphi(x)dx. \tag{3.20}$$

Using the same techniques as for the Fokker-Planck equation (see App. A.1.2), namely by expressing all the unnormalized expectations in Eq. (3.19) as integrals of the form of Eq. (3.20), integrating by parts and dropping the arbitrary function φ, we obtain a stochastic partial differential equation (SPDE) for the unnormalized posterior density, which reads

$$d\rho(x,t) = \mathcal{L}^{\dagger}\rho(x,t)dt + \left(\frac{g(x)}{g_0} - 1\right)\rho(x,t)\left(dN_t - g_0 dt\right). \tag{3.21}$$

In order to obtain equations for the normalized density $p(x,t)$ and measure $p_t[\varphi]$, we introduce a normalization constant Z_t,

$$Z_t \doteq \rho_t[1] = \int_{-\infty}^{\infty} \rho(x,t)dx, \tag{3.22}$$

which according to Eq. (3.19) satisfies the SDE

$$dZ_t = \left(\frac{1}{g_0}\rho_t\left[g\right] - \rho_t\left[1\right]\right)\left(dN_t - g_0 dt\right) \tag{3.23}$$

The first term in the last line vanishes because the integrand is a total differential in x and $\rho(x,t)$ and its derivatives vanish at $\pm\infty$. By defining the *posterior firing rate*

$$\gamma_t = p_t\left[g\right] = \frac{\rho_t\left[g\right]}{\rho_t\left[1\right]} \tag{3.24}$$

we can write the SDE for Z_t as

$$dZ_t = \left(\frac{\gamma_t}{g_0} - 1\right)Z_t\left(dN_t - g_0 dt\right). \tag{3.25}$$

From this follows that the SDE for the normalized measure reads

$$dp_t\left[\varphi\right] = p_t\left[\mathcal{L}\varphi\right]dt + \left(p_t\left[g\varphi\right] - p_t\left[\varphi\right]\right)\left(dN_t - p_t\left[g\right]dt\right), \tag{3.26}$$

and that therefore the SPDE for the normalized density takes the form

$$dp(x,t) = \mathcal{L}^{\dagger}p(x,t)dt + \left(\frac{g(x)}{\gamma_t} - 1\right)p(x,t)\left(dN_t - \gamma_t dt\right). \tag{3.27}$$

We observe that the arbitrary firing rate g_0, which was chosen for the process N_t under the measure Q and which appeared in the equations for the unnormalized measure and density, has now disappeared. The stochastic integro-differential equation for the conditional density (3.27) is comparable to the Kushner (1962) equation, which gives an analogous result for the case of a diffusive measurement process.

Remarks

The first published solution of the filtering problem with point process observations appears to be due to Snyder (1972), who derived an evolution equation for the posterior characteristic function

$$\chi(k,t) \doteq \int_{-\infty}^{\infty} p(x,t) e^{ikx} dx. \tag{3.28}$$

The derivation shown here uses the change of measure approach and therefore closely follows the pioneering work of Zakai (1969) on a diffusion process signal and measurement, which is also covered in standard books on filtering theory, such as Bain and Crişan (2009). Applications of the method for point process observations are found e.g. in Kliemann et al. (1990), Plienpanich (2007), Bobrowski et al. (2009) (which performs the calculation for a finite-state signal process); Gertner (1978), Xu (2011) and Ceci and Colaneri (2012) (combining point process observations with diffusive observations). See also the recent review by Venugopal et al. (2014). Nowadays, the different cases are all under the umbrella of filtering for semimartingales.

The filtering problem discussed here is remarkable for having been 'reinvented' again and again in different applied fields, pointing to a lack of exchange of ideas between the fields of stochastic analysis, signal processing etc. and applied fields such as machine learning and neuroscience. Unfortunately, this sort of communication problem is by no means unique to the case of filtering theory. Notable cases who do not seem to be aware of (or choose to ignore) the full extent of the mathematical literature on the subject include Eden (2007), Eden and Brown (2008), Pfister et al. (2009), Ujfalussy and Lengyel (2011a).

3.1.3 The approximate Gaussian filter of Pfister et al. (2009)

The formal solution to the filtering problem in Eq. (3.27) is a very powerful result but hard to work with and implement because it is in essence infinite-dimensional; for each value of x, Eq. (3.27) gives an SDE for the value of the posterior density at that point. Finite-dimensional filters can be constructed e.g. by discretizing the state space – such as in Bobrowski et al. (2009). As an alternative, one can try to find choices of the signal and measurement process parameters for which exact or (good) approximate finite-dimensional filters exist.

Here, we want to construct an approximate two-dimensional filter for a special filtering problem considered in Pfister et al. (2009). They used an Ornstein-Uhlenbeck (OU) process,

$$dX_t = -\theta X_t dt + b dW_t. \tag{3.29}$$

The parameter θ is the inverse time-constant of the OU process. We also note the generator and Fokker-Planck operator of the OU process

$$\mathcal{L} = -\theta x \partial_x + \frac{b^2}{2}\partial_x^2, \quad \mathcal{L}^\dagger = \theta + \theta x \partial_x + \frac{b^2}{2}\partial_x^2. \tag{3.30}$$

Moreover, they assumed an exponential gain function

$$g(x) = g_0 \exp\left[\beta x\right]. \tag{3.31}$$

Pfister et al. (2009) found an (approximate) Gaussian solution to the filtering problem which we would like to re-derive from the filtering equation (3.27). We therefore attempt a Gaussian ansatz for the posterior probability density,

$$\tilde{p}(x,t) = \mathcal{N}(x; \tilde{\mu}_t, \tilde{\sigma}_t^2) \equiv \frac{1}{\sqrt{2\pi}\tilde{\sigma}_t} \exp\left[-\frac{(x-\tilde{\mu}_t)^2}{2\tilde{\sigma}_t^2}\right]. \tag{3.32}$$

In the following, we will show three different derivations of the equations for $\tilde{\mu}_t$ and $\tilde{\sigma}_t^2$: the assumed density filter in continuous time, a variational scheme, and a geometrical scheme. All methods lead to the same results, but looking at the problem from different angles is potentially interesting.

Continuous-time Gaussian assumed density filtering

Assumed density filtering (ADF, see e.g. Kushner (1967)) makes use of the evolution equations of the conditional moments, which are readily deduced from Eq. (3.26) by setting $\varphi(x) = x^n$. The equation for the conditional moment of order n, $m_t^{(n)} \doteq p_t[X_t^n]$, reads

$$dm_t^{(n)} = p_t[\mathcal{L}X_t^n]dt + \left(\frac{p_t[g(X_t)X_t^n]}{p_t[g(X_t)]} - m_t^{(n)}\right)(dN_t - p_t[g(X_t)]dt). \tag{3.33}$$

By assuming that the measure at time t is Gaussian and evaluating the right-hand sides of these equations with $p_t = \tilde{p}_t$, where $\tilde{p}_t$ is a Gaussian measure, one finds evolution equations for the first two moments, which can then be translated into evolution equations for the mean and variance of the Gaussian. The modified mean and variance are then used for the next time, where the measure is reset to a Gaussian measure with the evolved mean and variance. For the specific Fokker-Planck operator and gain function above, we find

$$p_t[\mathcal{L}X_t^n] = -n\theta m_t^{(n)} + \frac{b^2 n(n-1)}{2}\theta m_t^{(n-2)} \tag{3.34}$$

and

$$\tilde{p}_t[g(X_t)] = \tilde{\gamma}_t = g_0 e^{\beta\tilde{\mu}_t + \frac{1}{2}\beta^2\tilde{\sigma}_t^2}, \tag{3.35}$$

$$\tilde{p}_t[g(X_t)X_t] = \frac{\partial}{\partial\beta}\tilde{\gamma}_t = \left(\tilde{\mu}_t + \beta\tilde{\sigma}_t^2\right)\tilde{\gamma}_t, \tag{3.36}$$

$$\tilde{p}_t[g(X_t)X_t^2] = \frac{\partial^2}{\partial\beta^2}\tilde{\gamma}_t = \left[\tilde{\sigma}_t^2 + \left(\tilde{\mu}_t + \beta\tilde{\sigma}_t^2\right)^2\right]\tilde{\gamma}_t, \tag{3.37}$$

$$\tilde{p}_t[g(X_t)X_t^3] = \frac{\partial^3}{\partial\beta^3}\tilde{\gamma}_t = \left[3\tilde{\sigma}_t^2\left(\tilde{\mu}_t + \beta\tilde{\sigma}_t^2\right) + \left(\tilde{\mu}_t + \beta\tilde{\sigma}_t^2\right)^3\right]\tilde{\gamma}_t, \tag{3.38}$$

where $\tilde{\mu}_t$ is the mean and $\tilde{\sigma}_t^2$ is the variance under the Gaussian measure. Using Eq. (3.33) we therefore obtain

$$dm_t^{(1)} = -\theta\tilde{\mu}_t dt + \beta\tilde{\sigma}_t^2(dN_t - \tilde{\gamma}_t dt), \tag{3.39}$$

$$dm_t^{(2)} = \left(b^2 - 2\theta\tilde{m}_t^{(2)}\right)dt + \left(\tilde{\mu}_t + \beta\tilde{\sigma}_t^2\right)^2(dN_t - \tilde{\gamma}_t dt). \tag{3.40}$$

The first of the above equations can be directly used for the time-evolution of the mean by replacing the left-hand side by $d\tilde{\mu}_t$. The second equation has to be transformed into an equation for the second centered moment,

$$C_t^{(2)} = m_t^{(2)} - \left(m_t^{(1)}\right)^2, \tag{3.41}$$

by noting (using $dN_t^2 = dN_t$ and $dt\, dN_t = 0$) that

$$\begin{aligned} d\left(m_t^{(1)}\right)^2 &= 2m_t^{(1)} dm_t^{(1)} + (dm_t^{(1)})^2 \\ &= -2\theta\tilde{\mu}_t^2 dt + 2\beta\tilde{\mu}_t\tilde{\sigma}_t^2 \left(dN_t - \tilde{\gamma}_t dt\right) + \beta^2\tilde{\sigma}_t^4 dN_t, \end{aligned} \tag{3.42}$$

such that we obtain (the dN_t terms cancel out)

$$\dot{C}_t^{(2)} = -2\theta\tilde{\sigma}_t^2 + b^2 - \beta^2\tilde{\sigma}_t^4\tilde{\gamma}_t. \tag{3.43}$$

We can therefore take

$$d\tilde{\mu}_t = -\theta\tilde{\mu}_t dt + \beta\tilde{\sigma}_t^2 \left(dN_t - \tilde{\gamma}_t dt\right), \tag{3.44}$$

$$\dot{\tilde{\sigma}}_t^2 = -2\theta\tilde{\sigma}_t^2 + b^2 - \beta^2\tilde{\sigma}_t^4\tilde{\gamma}_t, \tag{3.45}$$

as evolution equations for the mean and variance of the Gaussian density. These equations are the same as those derived in Pfister et al. (2009).

The problem with ADF is its lack of self-consistency. The above method of calculating the right-hand side of the moment equation by assuming a Gaussian density and then assigning the result as a moment equation of the corresponding Gaussian moment breaks down for the centered moment of order three,

$$C_t^{(3)} = m_t^{(3)} - 3m_t^{(2)} C_t^{(2)} - \left(m_t^{(1)}\right)^3, \tag{3.46}$$

for which one obtains the non-vanishing evolution equation

$$dC_t^{(3)} = \left[-3\theta C_t^{(3)} + C_t^{(3)}\tilde{\gamma}(t) - \beta^3\tilde{\sigma}^6(t)\tilde{\gamma}(t)\right] dt - C_t^{(3)} dN_t, \tag{3.47}$$

which is in direct contradiction to the identity

$$\tilde{C}_t^{(3)} = \tilde{p}_t \left[(X_t - \tilde{\mu}_t)^3\right] = 0 \tag{3.48}$$

for the centered moment of order three of a Gaussian. This contradiction makes the ADF method unacceptable from a rigorous point of view[4], but it also raises the question in what sense the Gaussian moment equations obtained by the ADF method stand out from other moment equations, i.e. what makes them special. In the following two approaches, we try to answer this question.

---

[4] Lack of rigor does not imply lack of value as a heuristic method. As we will show later, the ADF here is equivalent to a projection filter, which is fully rigorous.

A variational approach

Another method – which avoids the inconsistencies of the ADF – is to use a variational approximation of the posterior measure. Since we suspect that the Gaussian ansatz in Eq. (3.32) does not solve Eq. (3.27), we introduce an error term ϵ which measures the degree to which the filtering equation is violated:

$$d\tilde{p}(x,t) = \mathcal{L}^\dagger \tilde{p}(x,t)dt + \left(\frac{g(x)}{\tilde{\gamma}_t} - 1\right)\tilde{p}(x,t)\left(dN_t - \tilde{\gamma}_t dt\right) + \epsilon(x,t)\tilde{p}(x,t), \tag{3.49}$$

where $\tilde{\gamma}_t$ is the posterior firing rate under the Gaussian ansatz,

$$\tilde{\gamma}_t = \int_{-\infty}^{\infty} g(x)\tilde{p}(x,t)dx = g_0 \exp\left[\beta\tilde{\mu}_t + \frac{1}{2}\beta^2\tilde{\sigma}_t^2\right]. \tag{3.50}$$

In order to obtain the time-evolution of the mean and variance of the Gaussian part, which have to take the form

$$d\tilde{\mu}_t = B_t^{11}dt + B_t^{12}dN_t, \tag{3.51}$$

$$d\tilde{\sigma}_t^2 = B_t^{21}dt + B_t^{22}dN_t, \tag{3.52}$$

we plug the ansatz for $\tilde{p}$ into (3.49) and try to match the coefficients of dN_t and dt of both sides (and for all x) while making $\epsilon(x,t)$ as small as possible. Let us first consider what happens at the arrival time of a spike. Equation (3.49) demands that the posterior density is multiplied by $g(x)/\tilde{\gamma}_t$ when a spike occurs, i.e.

$$\tilde{p}(x,t^+) = \frac{g(x)}{\tilde{\gamma}_{t^-}}\tilde{p}(x,t^-). \tag{3.53}$$

Writing this out, we see that the Gaussian form is preserved,

$$\begin{aligned}\mathcal{N}(x;\tilde{\mu}_t + B_t^{12}, \tilde{\sigma}_t^2 + B_t^{22}) &\overset{!}{=} \exp\left[\beta(x-\tilde{\mu}_t) - \frac{1}{2}\beta^2\tilde{\sigma}_t^2\right]\mathcal{N}(x;\tilde{\mu}_t,\tilde{\sigma}_t^2)\\ &= \mathcal{N}(x;\tilde{\mu}_t + \beta\tilde{\sigma}_t^2, \tilde{\sigma}_t^2),\end{aligned} \tag{3.54}$$

and therefore we can match the mean and variance by picking $B_t^{12} = \beta\tilde{\sigma}_t^2$ and $B_t^{22} = 0$. Substituting these coefficients back into Eq. (3.49) and solving for $\epsilon(x,t)$, all the dN_t cancel out and we obtain

$$\begin{aligned}\epsilon(x,t) &= \frac{d\tilde{p}(x,t) - \mathcal{L}^\dagger\tilde{p}(x,t)dt}{\tilde{p}(x,t)} + g(x)dt - \tilde{\gamma}_t dt\\ &= \left\{\frac{x-\tilde{\mu}_t}{\tilde{\sigma}_t^2}(B_t^{11} + \theta x) + \frac{(x-\tilde{\mu}_t)^2 - \sigma_t^2}{2\tilde{\sigma}_t^4}(B_t^{21} - b^2) - \theta\right.\\ &\qquad \left. + g_0\exp\left[\beta x\right] - g_0\exp\left[\beta\tilde{\mu}_t + \frac{1}{2}\beta^2\tilde{\sigma}_t^2\right]\right\}dt\\ &\equiv f(x,t)dt.\end{aligned} \tag{3.55}$$

Since B_t^{11} and B_t^{21} cannot depend on x, they cannot be chosen such that $\epsilon(x,t) = 0$ for all x. Instead, we ask that the expectation of the square of $f(x,t)$ under the Gaussian distribution,

$$\mathcal{E}_t \equiv \int_{-\infty}^{\infty} f^2(x,t)\tilde{p}(x,t)dx, \tag{3.56}$$

be minimized. Imposing this minimization criterion leads to the following result:

$$B_t^{11} = -\theta\tilde{\mu}_t - \beta\tilde{\sigma}_t^2\tilde{\gamma}_t, \quad B_t^{21} = b^2 - 2\theta\tilde{\sigma}_t^2 - \beta^2\tilde{\sigma}_t^4\tilde{\gamma}_t. \tag{3.57}$$

When we combine these results with the dN_t terms found earlier, the equations for the time-evolution of the parameters of $\tilde{p}(x,t)$ read

$$d\tilde{\mu}_t = -\theta\tilde{\mu}_t dt + \beta\tilde{\sigma}_t^2(dN_t - \tilde{\gamma}_t dt), \tag{3.58}$$

$$d\tilde{\sigma}_t^2 = \left(b^2 - 2\theta\tilde{\sigma}_t^2 - \beta^2\tilde{\sigma}_t^4\tilde{\gamma}_t\right) dt. \tag{3.59}$$

Thus Eqs. (3.58,3.59) are the same as found by Pfister et al. (2009), as well as by the ADF method above. They minimize the square of the error term in equation (3.49), which takes a minimal value of order β^3

$$\begin{aligned} \epsilon(x,t) &= \left\{ g(x) - \tilde{\gamma}_t \left[1 + \beta(x - \tilde{\mu}_t) + \frac{1}{2}\beta^2 \left((x - \tilde{\mu}_t) - \tilde{\sigma}_t^2 \right) \right] \right\} dt \\ &= \frac{1}{6} g_0 dt \left[(x - \tilde{\mu}_t)^3 - 3\tilde{\sigma}_t^2 \left(x - \tilde{\mu}_t \right) \right] \beta^3 + \mathcal{O}\left(\beta^4 dt \right). \end{aligned} \tag{3.60}$$

The minimized function in Eq. (3.56) takes the form

$$\mathcal{E}_t = \tilde{\gamma}_t^2 \left(e^{\beta^2\tilde{\sigma}_t^2} - 1 - \beta^2\tilde{\sigma}_t^2 - \frac{1}{2}\beta^4\tilde{\sigma}_t^4 \right) = \tilde{\gamma}_t^2 \sum_{n=3}^{\infty} \frac{\left(\beta^2\tilde{\sigma}_t^2\right)^n}{n!} \tag{3.61}$$

A geometrical picture

Even though the variational approach leads to the same results as the ADF method, the minimization criterion (3.56) remains somewhat arbitrary. As has been pointed out in Hanzon and Hut (1991), ADFs with a Gaussian assumption can in some cases be regarded as projection filters, which were introduced in Brigo et al. (1998) and Brigo et al. (1999). We will briefly explain the geometric framework of projection filters which not only improves our understanding of the Gaussian approximation and approximation errors, but has the potential of being generalized to exponential families in order to obtain better approximations.

Consider the manifold $\mathcal{M}$ of density functions with respect to the Lebesgue measure dx on the state space $\mathcal{X}$ of the signal process. The filtering equation (3.27) is an equation for a piecewise-smooth curve $c : \mathbb{R} \to \mathcal{M}$ through the manifold. It has the form $dp = V(p)dt + (\phi(p) - p)dN$ and can therefore be characterized by the pair (V, ϕ), where V is the vector field

$$\begin{aligned} V :\ & \mathcal{M} \to T\mathcal{M}, \\ & p \mapsto (\mathcal{L}^\dagger + \gamma[p] - g)p, \quad \gamma[p] = \int_{\mathcal{X}} g(x)p(x)dx, \end{aligned} \tag{3.62}$$

for the time evolution in-between observation events, where c is smooth and tangent to V, i.e. $c'(t) = V(c(t))$. At the arrival time of an observation, the curve has a discontinuity and jumps from $c(t^-)$ to $c(t^+) = \phi(c(t^-))$, where ϕ is the map defined by

$$\begin{aligned} \phi :\ & \mathcal{M} \to \mathcal{M}, \\ & p \mapsto \frac{g(\cdot)p(\cdot)}{\int_{\mathcal{X}} g(x)p(x)dx}. \end{aligned} \tag{3.63}$$

The manifold $\mathcal{M}$ is infinite-dimensional, but we can consider a finite-dimensional submanifold $\mathcal{S} \subset \mathcal{M}$, e.g. the manifold of Gaussian densities on $\mathcal{X}$, or – more generally – an exponential family of probability distributions. Given this submanifold, we attempt to reduce the equation for the curve in $\mathcal{M}$ to an equation for a curve in $\mathcal{S}$. Suppose the curve starts in $\mathcal{S}$, i.e. $p_0 = c(0) \in \mathcal{S}$. Then in general, the vector $V(p_0)$ is not an element of $T_{p_0}\mathcal{S}$, and $\phi(p_0)$ lies outside of $\mathcal{S}$. In order to project the two quantities onto $\mathcal{S}$ we therefore require a metric D on $\mathcal{M}$. It induces a Riemannian metric G, in terms of which we can define a projection $\Pi_p : T_p\mathcal{M} \to T_p\mathcal{S}$ for each $p \in \mathcal{S}$ and therefore a projected vector field

$$\begin{aligned} \tilde{V}:\ &\mathcal{S} \to T\mathcal{S}, \\ &p \mapsto \Pi_p(V(p)). \end{aligned} \tag{3.64}$$

By using the metric D, we can also define what is the closest point to $\phi(p_0)$ within $\mathcal{S}$ and therefore a map

$$\begin{aligned} \tilde{\phi}:\ &\mathcal{S} \to \mathcal{S}, \\ &p \mapsto \operatorname*{argmin}_{q \in \mathcal{S}} D(\phi(p), q). \end{aligned} \tag{3.65}$$

The equation for the curve inside the submanifold $\mathcal{S}$ is then determined by the pair $(\tilde{V}, \tilde{\phi})$. Let us make things more concrete by specifying a distance measure in $\mathcal{M}$. We will work with the *Hellinger distance*

$$\begin{aligned} D:\ &\mathcal{M} \times \mathcal{M} \to \mathbb{R}, \\ &D(p, q) = \int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx. \end{aligned} \tag{3.66}$$

In order to construct a Riemannian metric on the tangent space, it is necessary to go to the space $\mathcal{M}^{1/2} = \{p^{1/2} | p \in \mathcal{M}\}$ of the square roots of the densities in $\mathcal{M}$, which is a subset of L_2. The tangent space $T_p\mathcal{M}^{1/2}$ will therefore be a vector subspace of L_2 and we can use the L_2 scalar product as a Riemannian metric:

$$G_p(X, Y) = \int_{\mathcal{X}} X(x)Y(x)dx, \quad X, Y \in T_p\mathcal{M}^{1/2}. \tag{3.67}$$

The difference between $\mathcal{M}$ and $\mathcal{M}^{1/2}$ is but a change of coordinates and the Riemannian metric can also be expressed in the coordinates of $T_p\mathcal{M}$ as

$$G_p(X, Y) = \frac{1}{4} \int_{\mathcal{X}} \frac{X(x)Y(x)}{p(x)} dx, \quad X, Y \in T_p\mathcal{M}, \tag{3.68}$$

or in the coordinates of $\log \mathcal{M} = \{\log p | p \in \mathcal{M}\}$ as

$$\begin{aligned} G_p(X, Y) &= \frac{1}{4} \int_{\mathcal{X}} p(x)X(x)Y(x)dx \\ &= \frac{1}{4} \mathbb{E}_p\left[XY\right], \quad X, Y \in T_p \log \mathcal{M}. \end{aligned} \tag{3.69}$$

This representation in terms of the expectation operator is the most convenient one for calculations, but of course all the representations lead to the same result. Let $\theta = (\theta^1, ..., \theta^m)$

be a parametrization of $\mathcal{S}$, $p(\cdot,\theta) \in \mathcal{S}$ and $E_1, ..., E_m$ the associated coordinate basis of $T_p\mathcal{S}$ given by the functions

$$E_i(x) = \frac{\partial p(x,\theta)}{\partial \theta^i}, \quad i = 1, ..., m. \tag{3.70}$$

The Riemannian metric restricted on the submanifold yields a metric tensor

$$\begin{aligned} G_{ij} = G_p(E_i, E_j) &= \frac{1}{4}\int_{\mathcal{X}} \frac{1}{p(x)} \frac{\partial p(x,\theta)}{\partial \theta^i} \frac{\partial p(x,\theta)}{\partial \theta^j} dx \\ &= \frac{1}{4}\int_{\mathcal{X}} p(x) \frac{\partial \log p(x,\theta)}{\partial \theta^i} \frac{\partial \log p(x,\theta)}{\partial \theta^j} dx, \end{aligned} \tag{3.71}$$

which is proportional to the Fisher information matrix. The projection operator onto $T_p\mathcal{S}$ can be written in that basis as

$$\Pi_p(X) = G^{ij} G_p(X, E_i) E_j, \quad X \in T_p\mathcal{M}, \tag{3.72}$$

where G^{ij} is the inverse of the metric tensor, i.e. $G^{ij}G_{jk} = \delta^i_k$ (we use the Einstein summation convention for repeated indices). The quality of the projection filter can be measured by two quantities. The first is the *projection residual* $V(p) - \tilde{V}(p)$ and its squared norm

$$\mathcal{E}^V(p) = \left\| V(p) - \tilde{V}(p) \right\|^2 = G_p\left(V(p) - \tilde{V}(p), V(p) - \tilde{V}(p)\right), \quad p \in \mathcal{S} \tag{3.73}$$

where G_p denotes the Riemannian metric at p. We will see that this is the quantity which was minimized in the last section. The second quantification of the filter quality is the *jump error*

$$\mathcal{E}^\phi(p) = D\left(\phi(p), \tilde{\phi}(p)\right), \quad p \in \mathcal{S}. \tag{3.74}$$

The approximate Gaussian filter as a projection filter

Let us now calculate the projection filter for the case in which

$$\mathcal{S} = \left\{\mathcal{N}(\mu, \sigma^2) | \mu \in \mathbb{R}, \sigma^2 > 0\right\} \tag{3.75}$$

is the two-dimensional manifold of Gaussian distributions on $\mathbb{R}$. In this case, as we noted earlier, the submanifold is mapped onto itself, i.e. $\phi(\mathcal{S}) = \mathcal{S}$, and therefore we can set $\tilde{\phi} = \phi$, making the jump error vanish. We then only have to consider the smooth part of the filter. Let us start at a point $p = \mathcal{N}(\mu, \sigma^2) \in \mathcal{S}$. The vector that is to be projected reads

$$\begin{aligned} V(x) &= (\mathcal{L}^\dagger + \gamma[p] - g(x))\mathcal{N}(x; \mu, \sigma^2) \\ &= -\theta x E_1(x) + b^2 E_2(x) + \left[\theta + g_0 e^{\beta\mu + \frac{1}{2}\beta^2\sigma^2} - g_0 e^{\beta x}\right] \mathcal{N}(x; \mu, \sigma^2), \end{aligned} \tag{3.76}$$

where we introduced the two basis vectors of $T_p\mathcal{S}$

$$\begin{aligned} E_1(x) &= \frac{\partial \mathcal{N}(x; \mu, \sigma^2)}{\partial \mu} = \frac{x - \mu}{\sigma^2} \mathcal{N}(x; \mu, \sigma^2), \\ E_2(x) &= \frac{\partial \mathcal{N}(x; \mu, \sigma^2)}{\partial \sigma^2} = \frac{(x-\mu)^2 - \sigma^2}{2\sigma^4} \mathcal{N}(x; \mu, \sigma^2), \end{aligned} \tag{3.77}$$

in the chosen (μ, σ^2) parametrization of $\mathcal{S}$. The Riemannian metric tensor can be calculated from Eq. (3.68)

$$\begin{aligned} G_{11} = G_p(E_1, E_1) &= \frac{1}{4}\int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sigma^4}\mathcal{N}(x;\mu,\sigma^2)dx = \frac{1}{4\sigma^2}, \\ G_{12} = G_{21} = G_p(E_1, E_2) &= \frac{1}{4}\int_{-\infty}^{\infty} \frac{(x-\mu)^3 - \sigma^2(x-\mu)}{\sigma^6}\mathcal{N}(x;\mu,\sigma^2)dx = 0, \\ G_{22} = G_p(E_2, E_2) &= \frac{1}{4}\int_{-\infty}^{\infty} \left[\frac{(x-\mu)^2 - \sigma^2}{2\sigma^4}\right]^2 \mathcal{N}(x;\mu,\sigma^2)dx = \frac{1}{8\sigma^4}, \end{aligned} \tag{3.78}$$

and as we expected the metric tensor is – up to a factor of $1/4$ – equal to the Fisher information matrix. We can now calculate the projection of V onto the basis vectors and find

$$\begin{aligned} \tilde{V} &= G^{11}G_p(V, E_1)E_1 + G^{22}G_p(V, E_2)E_2 + \\ &= \left(-\theta\mu - \beta\sigma^2\gamma\right)E_1 + \left(b^2 - 2\theta\sigma^2 - \beta^2\sigma^4\gamma\right)E_2, \end{aligned} \tag{3.79}$$

which is consistent with the equations found earlier. The projection residual $V - \tilde{V}$ reads

$$V(x) - \tilde{V}(x) = -\mathcal{N}(x;\mu,\sigma^2)\left\{g(x) - \gamma\left[1 + \beta(x-\mu) + \frac{1}{2}\beta^2\left((x-\mu) - \sigma^2\right)\right]\right\}, \tag{3.80}$$

where the bracketed expression is the same as the one computed in (3.60). This means that the function $\mathcal{E}_t$ in Eq. (3.56) indeed was the squared norm of the projection residual.

Numerical solution of the filtering equation

In order to obtain the exact moments, one has to solve the unnormalized filtering equation (3.21), which in the absence of spikes and for the present case reduces to the following partial differential equation (PDE)

$$\partial_t\rho(x,t) = \left(\theta + g_0 - g_0e^{\beta x} + \theta x\partial_x + \frac{b^2}{2}\partial_x^2\right)\rho(x,t). \tag{3.81}$$

In order to solve this PDE numerically, it is convenient to transform the spatial domain to lie in a bounded interval, i.e. to do a transformation ϕ defined e.g. by[5]

$$\phi : (-1, 1) \to \mathbb{R}, \quad y \mapsto \phi(y) = \log\left[\frac{1/\beta}{1-y}\right] - \log\left[\frac{1/\beta}{1+y}\right], \tag{3.82}$$

such that the unnormalized density may be described in terms of the transformed density

$$v(\cdot, t) : (-1, 1) \to \mathbb{R}, \quad y \mapsto v(y, t) = \phi'(y)\rho(\phi(y), t). \tag{3.83}$$

Using the chain rule, one can derive a PDE for $v(y,t)$ which takes the form

$$\begin{aligned} \partial_t v(y,t) = \Bigg[&\theta + g_0 - g_0e^{\beta\phi(y)} - \frac{\theta\phi(y)\phi''(y)}{\phi'(y)^2} - \frac{\theta b^2\left(\phi'''(y)\phi'(y) - 3\phi''(y)^2\right)}{\phi'(y)^4} \\ &+ \theta\frac{\phi(y)\phi'(y)^2 - 3b^2\phi''(y)}{\phi'(y)^3}\partial_y + \frac{\theta b^2}{\phi'(y)^2}\partial_y^2\Bigg]v(y,t), \end{aligned} \tag{3.84}$$

---

[5] Other transformations may be chosen, but they must satisfy $\phi'(y) > 0, \forall y$, and they must be tested for stability. The transformation given in Eq. (3.82) seems to lead to stable behavior of the numerical integration.

and the boundary conditions read $v(-1,t) = v(1,t) = 0$, $\forall t \geq 0$. Equation (3.84) can be integrated with standard PDE solvers (e.g. as built into Wolfram Mathematica). Spike arrivals are implemented by stopping the numerical integration at the spike arrival time T, multiplying the function $v(y,T)$ by the gain function $g_0 e^{\beta\phi(y)}$ and restarting the numerical integration at T with the new initial condition. Posterior expectations can be obtained via the formula

$$p_t[\varphi] = \frac{1}{Z_t}\int_{-1}^{1} \varphi(\phi(y))v(y,t)dy, \quad Z_t = \int_{-1}^{1} v(y,t)dy, \tag{3.85}$$

where these integrals are done numerically by using the discretized representation of $v(y,t)$. In order to check whether this PDE method yields the correct result, one can compare it to the results from a sequential Monte-Carlo (SMC) method (i.e. a particle filter).

Deviation of the Gaussian approximation from the exact posterior distribution

The above derivations show that the Gaussian distribution only solves the filtering equation up to an error term ϵ given in Eq. (3.60), which can be geometrically interpreted as a projection residual. We can also investigate the deviation of the exact posterior moments to the moments of the Gaussian approximation. Here, we will look at the first three moments, expressed as the mean μ_t, the variance σ_t^2 and the third centered moment (or third cumulant) $C_t^{(3)}$. Both the stationary values of these moments (i.e. in the absence of spikes) and the values of the moments immediately after a spike (with no spikes beforehand) deviate from the approximated Gaussian solutions. The deviations increase with β, as suggested by the error term in Eq. (3.60).

We computed the first three moments using the exact, SMC, and approximate[6] Gaussian methods and show them on Fig. 3.1 as a function of β (choosing g_0 such as to keep the expected firing rate constant). The most important difference between the exact moments and the approximate ones is that the variance σ^2 of the Gaussian approximation, which does not have a dN_t term according to Eq. 3.59, does not change after a spike, whereas the variance of the exact posterior jumps upon a spike.

---

[6] The approximate moments are obtained from Eqs. (3.58,3.59) for the mean and variance, and Eq. (3.47) for the third cumulant. The third cumulant of the Gaussian is of course zero, but Eq. (3.47) provides a hint that the inconsistency is of the same order of magnitude as the third cumulant of the exact posterior distribution.

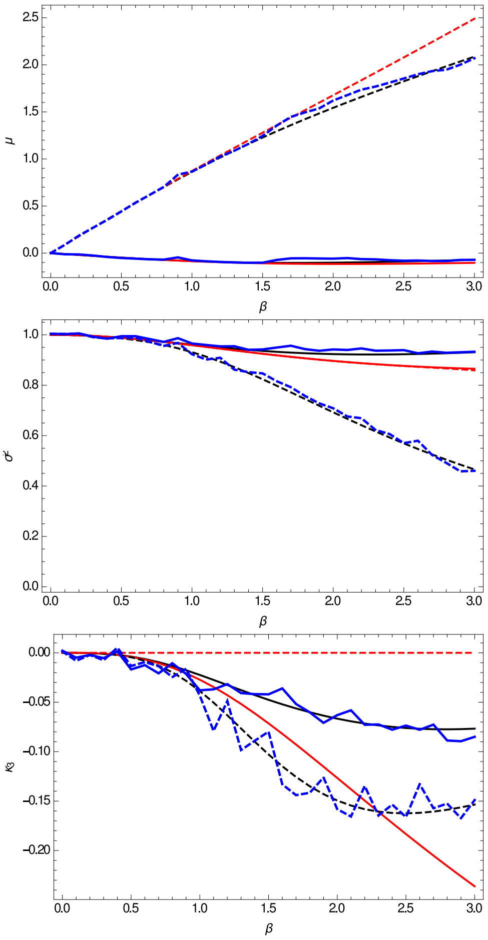


Figure 3.1: Approximation errors of the approximate Gaussian filter in calculating the mean μ_t (top panel), variance σ_t^2 (center panel) and third cumulant $C_t^{(3)}$ (bottom panel) of the posterior distribution. Solid lines: stationary value in the absence of spikes. Dashed lines: value immediately after a spike, with no spikes beforehand. Black lines: Numerical solution of the PDE (3.81). Blue lines: Particle filter. Red lines: Gaussian filter from Eqs. (3.58,3.59), and the ‘wrong’ third cumulant of Eq. (3.47). The particle filter agrees with the PDE results. The Gaussian approximations shows deviations from the exact moments which increase with β. Most notable is the fact that according to the exact solution, σ^2 changes after a spike (i.e. the solid and the dashed lines in the center panel differ for the exact solution shown in black), whereas it does not change according to the Gaussian approximation (i.e. the red solid and dashed lines of the center panel coincide). Parameter values: $\theta = 0.1$, $b^2 = 0.2$, $g_0 = 0.01e^{-\beta^2/2}$.

3.2 Extensions I

The derivations of the filtering equation in Section 3.1.2 were carried out for a one-dimensional diffusion signal and an inhomogeneous Poisson process measurement, but there is a number of extensions which can be made to that model such that the filtering problem remains tractable. In this section, we are going to look at two such extensions.

The inclusion of adaptive mechanisms in the point process observation model is an important extension of the theory. Fortunately, the filtering problem does not become much harder even though the resulting filter has new properties. In the context of neuroscience, the extension is motivated from a simple observation, namely that neurons show refractory and adaptive effects in their firing rates and that therefore firing properties are not adequately described by a Poisson process (see Chapter 3, which offers a detailed analysis of this problem). The extension leads – for a certain range of parameters – to a prediction of short-term facilitation in the downstream synapse in order to compensate for the expected reduction in the presynaptic firing rate following a spike. We will discuss the predictions due to this mechanism in Section 3.4.

Another extension which we will present here is the generalization of the signal process to a multivariate diffusion process and of the observation to a multivariate point process, both of which are straightforward in the framework of filtering theory. These extensions allow the KTN theory to be formulated for multiple presynaptic inputs, multi-component membrane potentials (e.g. with fast and slow timescales), and higher-order Gauss-Markov processes, which approximate certain Gaussian processes. Through these mathematical tools, predictions can be made based on the neuron model presented in Chapter 3, and the problem of target-cell specificity can be addressed (see Section 3.5.2 for further ideas on this subject).

3.2.1 Adaptive point process observations

What happens if we want to make the gain function g adaptive, i.e. depend on the previous history of spikes? We also include an explicit time dependence, i.e.

$$g(X_t) \longrightarrow g\left(X_t, N_{[0,t]}, t\right), \tag{3.86}$$

such that the number of spikes observed in a certain time interval is no longer independent (conditioned on X_t) of the number of spikes observed in any previous interval

$$N_t - N_0 \sim \text{Poisson}\left[\int_0^t g(X_s, N_{[0,s]}, s)ds\right]. \tag{3.87}$$

The resulting process is no longer a renewal process, therefore we call it *adaptive point process*[7]. The expression for the conditional law of N_t reads[8]

$$P\left(N_{[0,t]} \text{ jumps at } t_1, ..., t_n | X_{[0,t]}\right) = \exp\left[-\int_0^t g(X_s, \{t_i \le s\}, s)ds\right] \times \prod_{0 \le t_n \le t} g(X_{t_n}, \{t_i \le t_n\}, t_n). \tag{3.88}$$

---

[7] This sort of process is sometimes called self-exciting point process or Hawkes process, see e.g. Hawkes (1971), Gerencsér et al. (2008), and Chen and Hall (2015) and references therein.

[8] One might wonder whether this expression is normalized. We prove in Appendix A.3 that it is.

After having checked that the expression for the probability of a sequence of spikes in Eq. (3.88) looks the same as for the previously used renewal case, we can verify that the derivation of the filtering equation remains basically unaltered as well. The equation for the Radon-Nikodym derivative in Eq. (3.11), for instance, is the same and therefore all the steps that follow it can be carried out as before. We therefore obtain the following filtering equation for the adaptive point process measurements

$$dp(x,t) = \mathcal{L}^{\dagger} p(x,t)dt + \left(\frac{g\left(x, N_{[0,t]}, t\right)}{\gamma_t} - 1 \right) p(x,t) \left(dN_t - \gamma_t dt\right), \tag{3.89}$$

where now γ_t depends on the previous observations,

$$\gamma_t = \int_{-\infty}^{\infty} g\left(x, N_{[0,t]}, t\right) p(x,t) dx. \tag{3.90}$$

Also the Gaussian approximate filter for the OU signal and exponential gain function can be easily extended to the adaptive point process. If we parametrize the adaptive effect on the gain function as

$$g\left(x, N_{[0,t]}, t\right) = g_0 \exp\left[\beta_t x + A_t\right], \tag{3.91}$$

where A_t and β_t depend on the spike history $N_{[0,t]}$, we can write the equations for the parameters of the Gaussian approximation as

$$d\tilde{\mu}_t = -\theta \tilde{\mu}_t dt + \beta_t \tilde{\sigma}_t^2 (dN_t - \tilde{\gamma}_t dt), \tag{3.92}$$

$$d\tilde{\sigma}_t^2 = \left(b^2 - 2\theta \tilde{\sigma}_t^2 - \beta_t^2 \tilde{\sigma}_t^4 \tilde{\gamma}_t\right) dt, \tag{3.93}$$

which are identical to eqs. (3.58) and (3.59), except for the fact that the posterior expected firing rate depends on A,

$$\tilde{\gamma}_t = g_0 \exp\left[\beta_t \tilde{\mu}_t + \frac{1}{2}\beta_t^2 \tilde{\sigma}_t^2 + A_t\right]. \tag{3.94}$$

We call A_t, β_t *adaptation variables.*

3.2.2 Multivariate diffusion prior

We want to generalize the filtering equation to a multivariate diffusion signal

$$d\boldsymbol{X}_t = \boldsymbol{a}(\boldsymbol{X}_t)dt + b(\boldsymbol{X}_t)d\boldsymbol{W}_t, \tag{3.95}$$

where $\boldsymbol{X}_t$ is an $\mathbb{R}^n$-valued random variable, $\boldsymbol{W}_t$ is a n-vector-valued standard Wiener process, $\boldsymbol{a}$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^n$, and b a map from $\mathbb{R}^n$ into the space of $n \times n$-matrices. The gain function is now a map from $\mathbb{R}^n$ to $\mathbb{R}^+$, potentially including adaptation effects (see previous section). All the steps in Section 3.1.2 apply for the multivariate case by replacing scalars with vectors and derivatives by partial derivatives. Carrying out all the steps, one finds that the filtering equation reads

$$dp(\boldsymbol{x},t) = \mathcal{L}^{\dagger} p(\boldsymbol{x},t)dt + \left(\frac{g\left(\boldsymbol{x}, N_{[0,t]}, t\right)}{\gamma_t} - 1 \right) p(\boldsymbol{x},t) \left(dN_t - \gamma_t dt\right), \tag{3.96}$$

where

$$\gamma_t = \int_{\mathbb{R}^n} g\left(\boldsymbol{x}, N_{[0,t]}, t\right) p(\boldsymbol{x}, t) d\boldsymbol{x}, \tag{3.97}$$

and $\mathcal{L}^\dagger$ denotes the Fokker-Planck operator of the multivariate diffusion,

$$\mathcal{L}^\dagger = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left[\boldsymbol{a}_i(\boldsymbol{x})\cdot\right] + \frac{1}{2} \sum_{i,j,k=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j} \left[b_{ik}(\boldsymbol{x}) b_{jk}(\boldsymbol{x})\cdot\right]. \tag{3.98}$$

For the special case of a multivariate OU process,

$$d\boldsymbol{X}_t = -\mathcal{A}\boldsymbol{X}_t dt + \mathcal{B} d\boldsymbol{W}_t, \tag{3.99}$$

where $\mathcal{A}, \mathcal{B}$ are $n \times n$-matrices, and an exponential gain function

$$g\left(\boldsymbol{x}, N_{[0,t]}, t\right) = g_0 \exp\left[\boldsymbol{\beta}_t^\top \boldsymbol{x} + A_t\right], \tag{3.100}$$

we can find a multivariate Gaussian approximation of the posterior density

$$\tilde{p}(\boldsymbol{x}, t) = \mathcal{N}(\boldsymbol{x}; \tilde{\boldsymbol{\mu}}_t, \tilde{\Sigma}_t) \equiv \left(\det(2\pi\tilde{\Sigma}_t)\right)^{-1/2} \exp\left[-(\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_t)^\top \tilde{\Sigma}_t^{-1} (\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_t)\right], \tag{3.101}$$

which solves the diffusion equation up to a term ϵ

$$d\tilde{p}(\boldsymbol{x}, t) = \mathcal{L}^\dagger \tilde{p}(\boldsymbol{x}, t) dt + \left(\frac{g\left(\boldsymbol{x}, N_{[0,t]}, t\right)}{\tilde{\gamma}_t} - 1\right) \tilde{p}(\boldsymbol{x}, t) \left(dN_t - \tilde{\gamma}_t dt\right) + \epsilon(\boldsymbol{x}, t) \tilde{p}(\boldsymbol{x}, t). \tag{3.102}$$

As in the corresponding scalar case (see Section 3.1.3), the dN_t term of ϵ vanishes because the Gaussian is preserved by multiplication of an exponential. By minimizing a function analogous to the one in Eq. (3.56), we find the time-evolution of the parameters of $\tilde{p}$ to be given by

$$d\tilde{\boldsymbol{\mu}}_t = -\mathcal{A}\tilde{\boldsymbol{\mu}}_t dt + \tilde{\Sigma}_t \boldsymbol{\beta}_t (dN_t - \tilde{\gamma}_t dt), \tag{3.103}$$

$$d\tilde{\Sigma}_t = \left(\mathcal{B}\mathcal{B}^\top - \mathcal{A}\tilde{\Sigma}_t - \tilde{\Sigma}_t \mathcal{A}^\top - \tilde{\gamma}_t \tilde{\Sigma}_t \boldsymbol{\beta}_t \boldsymbol{\beta}_t^\top \tilde{\Sigma}_t\right) dt, \tag{3.104}$$

where the posterior firing rate is given by

$$\tilde{\gamma}_t = g_0 \exp\left[\boldsymbol{\beta}_t^\top \tilde{\boldsymbol{\mu}}_t + \frac{1}{2} \boldsymbol{\beta}_t^\top \tilde{\Sigma}_t \boldsymbol{\beta}_t + A_t\right]. \tag{3.105}$$

3.2.3 Multivariate point process observations

The scalar or multivariate diffusion signal may be observed through multiple – say, m – conditionally independent point process measurements $N^{(i)}$, each coupled to the signal through a gain function g_i. In neuroscience, this could apply in the context of spike inputs from multiple presynaptic neurons converging in the postsynaptic dendrite. Fortunately, the derivation in Section 3.1.2 is easily generalizable to this scenario.

We begin by noting that the joint probability of the signal and observation takes the form,

$$\begin{aligned} P\left(\boldsymbol{X}_{[0,t]},\boldsymbol{N}_{[0,t]}\right) &= P\left(\boldsymbol{X}_{[0,t]}\right)P\left(\boldsymbol{N}_{[0,t]}|\boldsymbol{X}_{[0,t]}\right) \\ &= P\left(\boldsymbol{X}_{[0,t]}\right)\prod_{i=1}^{m}P\left(N^{(i)}_{[0,t]}|\boldsymbol{X}_{[0,t]}\right) \\ &= P\left(\boldsymbol{X}_{[0,t]}\right)\prod_{i=1}^{m}\exp\left[-\int_0^t g_i(\boldsymbol{X}_s)ds\right]\prod_{0\le t_n^{(i)}\le t} g_i\left(\boldsymbol{X}_{t_n^{(i)}}\right). \end{aligned} \tag{3.106}$$

We omit the adaptive notation $g_i(\boldsymbol{X}_t, N^{(i)}_{[0,t]}, t)$ in the understanding that the derivation is not modified by adding adaptation. The Radon-Nikodym derivative of going to a measure Q where all processes are Poisson with rate g_0 reads

$$L_t \equiv \frac{P\left(\boldsymbol{X}_{[0,t]},\boldsymbol{N}_{[0,t]}\right)}{Q\left(\boldsymbol{X}_{[0,t]},\boldsymbol{N}_{[0,t]}\right)} = \prod_{i=1}^{m}\exp\left[\int_0^t\left(g_0 - g_i(\boldsymbol{X}_s)\right)ds\right]\prod_{0\le t_n^{(i)}\le t}\frac{g_i\left(\boldsymbol{X}_{t_n^{(i)}}\right)}{g_0}, \tag{3.107}$$

but it now satisfies a modified SDE

$$dL_t = \sum_{i=1}^{m}\left(\frac{g_i(\boldsymbol{X}_t)}{g_0} - 1\right)L_t(dN_t^{(i)} - g_0 dt), \quad L_0 = 1. \tag{3.108}$$

From then on, all steps can be carried out as in Section 3.1.2, finally giving us a filtering equation that reads

$$dp(\boldsymbol{x},t) = \mathcal{L}^\dagger p(\boldsymbol{x},t)dt + \sum_{i=1}^{m}\left(\frac{g_i\left(\boldsymbol{x},N_{[0,t]},t\right)}{\gamma_t^{(i)}} - 1\right)p(\boldsymbol{x},t)\left(dN_t^{(i)} - \gamma_t^{(i)}dt\right), \tag{3.109}$$

where

$$\gamma_t^{(i)} = \int_{\mathbb{R}^n} g\left(\boldsymbol{x},N_{[0,t]},t\right)p(\boldsymbol{x},t)d\boldsymbol{x}. \tag{3.110}$$

The simple additive structure of the filtering equation is not accidental. In fact, any number of measurement processes which are independent when conditioned on the signal process lead to independent terms in the filtering equation, i.e. each of the terms is the same as if this measurement were the only one available. Filtering results exhibiting this nice structure can be found e.g. in Frey et al. (2013) for the combination of point process and diffusive observations. In Ujfalussy and Lengyel (2011b), a special case of the one presented here was treated.

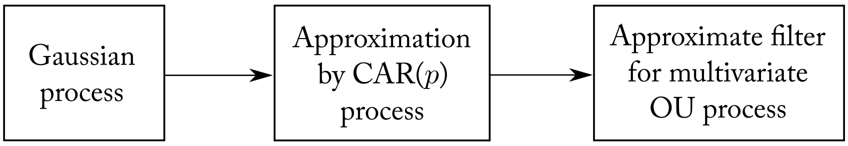

Gaussian process
Approximation by CAR(p) process
Approximate filter for multivariate OU process

Figure 3.2: The proposed scheme for filtering a Gaussian process with point process observations: The general Gaussian process is first approximated by a continuous autoregressive (CAR) process of a certain order p. For this class of processes the filtering problem is tractable and we can adopt the approximate Gaussian filter of Section 3.2.2.

3.3 Extensions II – Gaussian Process Prior

Here, we want to present a scheme which makes approximate filtering possible for a general Gaussian process (GP). The scheme is illustrated in Fig. 3.2. The basic idea is to approximate the Gaussian process spectral density with the one of a suitable Gaussian diffusion process which we call *continuous autoregressive* process or CAR process. Their relation to discrete-time autoregressive processes is briefly treated in Appendix A.2. The filtering problem for CAR processes is a special case of the multivariate OU discussed in the previous section, and therefore admits an approximate Gaussian filter.

In the following, we define the CAR process and list a number of properties, in particular the representation of the spectral density function. Afterwards, we will use that representation to derive a spectral approximation scheme.

3.3.1 Definition and properties

A certain subclass of Gaussian diffusion processes can be regarded as solutions to a linear stochastic differential equation of order p of the form

$$U_t^{(p)} + a_{p-1}U_t^{(p-1)} + ... + a_0 U_t = b\zeta_t, \tag{3.111}$$

where ζ_t is white noise, and $U_t^{(i)}$ stands for the derivative of order i of U_t. This higher-order SDE can be written as an Itô SDE of the form of Eq. (3.99), where

$$\boldsymbol{X}_t = \begin{pmatrix} U_t \\ U_t^{(1)} \\ \vdots \\ U_t^{(p-1)} \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} 0 & -1 & 0 & \dots & 0 \\ 0 & 0 & -1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 0 & -1 \\ a_0 & a_1 & \dots & \dots & a_{p-1} \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} 0 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & b \end{pmatrix}. \tag{3.112}$$

We call such a process continuous autoregressive of order p, or CAR(p). Our interest lies in the marginal distribution of the first component of $\boldsymbol{X}_t$, U_t. It is not possible to give a general closed-form expression for the autocovariance function of U_t,

$$k(\tau) = \mathbb{E}\left[U_t U_{t+\tau}\right], \tag{3.113}$$

nor of the autocovariance matrix of $\boldsymbol{X}_t$. But we can give a closed formula for the spectral density matrix of $\boldsymbol{X}_t$ which holds for any multivariate OU process,

$$\mathcal{S}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbb{E}\left[\boldsymbol{X}_t \boldsymbol{X}_{t+\tau}^\top\right] e^{-i\omega\tau} d\tau = \frac{1}{2\pi}(\mathcal{A} + i\omega\mathbb{1})^{-1}\mathcal{B}\mathcal{B}^\top(\mathcal{A}^\top - i\omega\mathbb{1})^{-1}, \tag{3.114}$$

c.f. Gardiner (1985). By using the special forms of the matrices $\mathcal{A}, \mathcal{B}$ in Eq. (3.112), we can write the spectral density of U_t as

$$s(\omega) \doteq \frac{1}{2\pi} \int_{-\infty}^{\infty} k(\tau) e^{-i\omega\tau} d\tau = \frac{b^2}{2\pi\chi(\omega)}, \tag{3.115}$$

where

$$\chi(\omega) = \left| \sum_{j=0}^{p} a_j \cdot (i\omega)^j \right|^2, \quad a_p = 1, \tag{3.116}$$

is the *fundamental polynomial* associated with the CAR(p) process.

3.3.2 Spectral approximation scheme

Given a stationary Gaussian process (GP) with covariance function $k_{\mathrm{GP}}(\tau)$, we want to find the parameters $a_0, ..., a_{p-1}, b$ of a CAR(p) process, such that the latter approximates the former. This can be done numerically, e.g. by minimizing the L_2 norm of the difference of the two spectral density functions. Here, we want to give an alternative, semi-analytical method.

First, we have to establish which order p of the CAR process is suitable. This depends on the smoothness of the GP which we wish to approximate. The smoothness properties of a GP are expressed in terms of the covariance function $k_{\mathrm{GP}}(\tau)$, c.f. Rasmussen and Williams (2006): The GP is $(p-1)$-times mean-square (MS) differentiable iff the covariance function is $(2p-2)$-times continuously differentiable at $t = 0$. If a given covariance function has $q = 2p-2$ continuous derivatives at $t = 0$ but its $(q+1)$st derivative is either discontinuous or singular at $t = 0$, then the Fourier transform of the $(q+1)$st derivative is not integrable. Thus we write the condition for the GP to be $(p-1)$-times MS differentiable as

$$\left| \int s_{\mathrm{GP}}(\omega) \omega^{2p-2} d\omega \right| < \infty. \tag{3.117}$$

where $s_{\mathrm{GP}}(\omega)$ is the spectral density of the GP,

$$s_{\mathrm{GP}}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\omega\tau} k_{\mathrm{GP}}(\tau) d\tau. \tag{3.118}$$

On the other hand, a CAR(p) process is by definition $(p-1)$-times mean square (MS) differentiable. Therefore, if the GP we are considering satisfies the condition (3.117), we approximate it with a CAR process whose order is at most p.

Let us proceed to the approximation scheme. Given the freedom of choosing $p+1$ parameters $a_0, ..., a_{p-1}, b$, we demand that the spectral function of the GP coincide with the spectral function of the CAR(p) process at $p+1$ frequencies $0 \leq \omega_0 < \omega_1 < ... < \omega_p < \infty$,

$$s_{\mathrm{GP}}(\omega_i) = s(\omega_i), \quad i = 0, 1, ..., p. \tag{3.119}$$

Note that this does not solve the problem *per se*, it merely re-parametrizes the problem such that instead of having to specify $a_0, ..., a_{p-1}$ and b whose meaning is not very intuitive, we get to choose $p+1$ 'important' or matching frequencies $\omega_0 < ... < \omega_p$ based on the structure of $s_{\mathrm{GP}}(\omega)$.

Let's choose $\omega_0 = 0$ for convenience. By using Eq. (3.115), we can rewrite (3.119) as

$$f(\omega_i) \equiv \chi(\omega_i) s_{\mathrm{GP}}(\omega_i) = \frac{b^2}{2\pi}, \quad i = 0, 1, ..., p. \tag{3.120}$$

or

$$\frac{b^2}{2\pi} = f(\omega_0) = f(\omega_1) = f(\omega_2) = \cdots = f(\omega_p), \tag{3.121}$$

which is a system of $p+1$ polynomial equations in $a_0, ..., a_{p-1}, b$. This system of equations has in general many solutions, but the right one to pick is the one where all a_i's are real and where the eigenvalues of $\mathcal{A}$ have positive real part. With this solution, the spectral density takes the form

$$s(\omega) = \frac{f(\omega_i)}{\chi(\omega)} \tag{3.122}$$

3.3.3 CAR approximation of the squared-exponential GP

We consider a GP with squared exponential covariance function

$$k_{\mathrm{GP}}(\tau) = k_0 \exp\left(-\tfrac{1}{2}\theta^2\tau^2\right), \tag{3.123}$$

which is smooth at $\tau = 0$ and therefore yields a process with smooth sample paths. For this GP a CAR process of any order is suitable. Let us illustrate the above scheme by calculating the CAR(2) spectral approximation. The spectral density function of the GP reads

$$s_{\mathrm{GP}}(\omega) = \frac{k_0}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{\omega^2}{2\theta^2}\right), \tag{3.124}$$

and the fundamental polynomial (3.116) of the CAR(2) process takes the form

$$\chi(\omega) = a_0^2 + (a_1^2 - 2a_0)\omega^2 + \omega^4. \tag{3.125}$$

Let us choose $\omega_0 = 0$, and write down the system of equations in (3.121)

$$\frac{b^2}{2\pi} = a_0^2 = e^{-\frac{\omega_1^2}{2\theta^2}}\left(a_0^2 + (a_1^2 - 2a_0)\omega_1^2 + \omega_1^4\right) = e^{-\frac{\omega_2^2}{2\theta^2}}\left(a_0^2 + (a_1^2 - 2a_0)\omega_2^2 + \omega_2^4\right). \tag{3.126}$$

This is a quadratic system of equations in a_0, a_1, b and it has the positive solution

$$b = \sqrt{2\pi} a_0, \tag{3.127}$$

$$a_0 = \left[-\frac{\omega_1^2 - \omega_2^2}{c_1 - c_2}\right]^{1/2}, \tag{3.128}$$

$$a_1 = \left[2a_0 - \tfrac{1}{2}(c_1 + c_2)a_0^2 - \tfrac{1}{2}(\omega_1^2 + \omega_2^2)\right]^{1/2}, \tag{3.129}$$

where

$$c_i = \frac{1}{\omega_i^2}\left[1 - \exp\left(\frac{\omega_i^2}{2\theta^2}\right)\right]. \tag{3.130}$$

We are left with choosing two non-zero matching frequencies ω_1, ω_2. By visually inspecting the spectral density function of the GP, we find that $(\omega_1, \omega_2) = (\theta, 2\theta)$ is a good choice. Incidentally, we find numerically that this is close to the choice of matching frequency that minimizes the L^2-norm between s and s_{GP}. For this choice of matching frequencies, we can write the CAR(2) spectral density as

$$s(\omega) = \frac{\sqrt{\frac{72}{\pi}} k_0/\theta}{\left(3 - 4e^{1/2} + e^2\right)(\omega/\theta)^4 - \left(15 - 16e^{1/2} + e^2\right)(\omega/\theta)^2 + 12}, \tag{3.131}$$

and we display it, along with the GP spectral density function, in Fig. 3.3.

The CAR(2) process is but a crude approximation to a smooth GP. The samples from the CAR(2) process have similar correlation structure, but they are less smooth on small time-scales. In fact, the CAR(2) is exactly once MS differentiable, so its second derivative is white noise (as can be inferred from the SDE). In Fig. 3.4 we show sample traces and their derivatives from the squared exponential GP and from the fitted CAR(2).

3.3.4 Filtering with approximate prior

Since the CAR(2) process is a special case of the multivariate OU process (see Section 3.2.2), we can use the approximation of the squared-exponential GP by a CAR(2) process in order to filter the GP from point process observations. Interestingly, despite this crude approximation and the associated model mismatch during filtering, using the CAR(2) process as a prior for inference when the ground truth of the data is sampled from a squared-exponential GP yields a filtering performance which is indistinguishable from the case where the ground truth is a CAR(2) process. This is shown in Fig. 3.5 to hold even for higher orders of the CAR(p) process, and the performance saturates already for $p = 2$. Therefore, the filtering performance seems to be limited by factors *other* than the model mismatch. Possible factors include the limited bandwidth of the spiking process and the approximate nature of the filter itself (recall that the filter that is used for the multivariate OU process is not exact, but rather a Gaussian ADF).

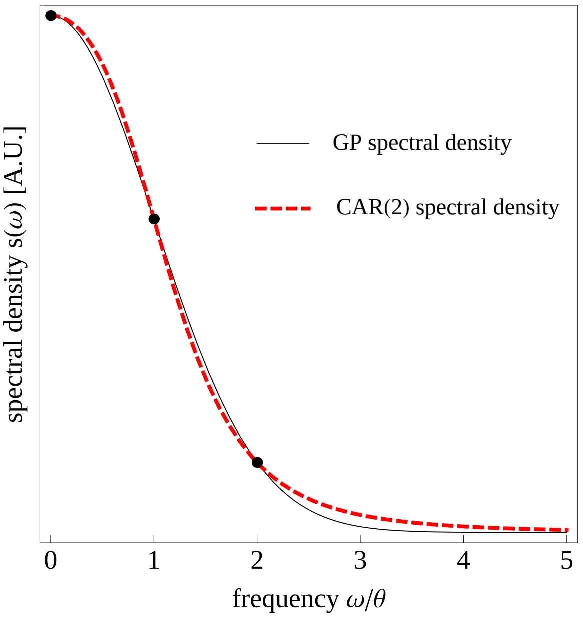


Figure 3.3: The spectral densities of the original squared exponential GP (black) and the fitted CAR(2) process (red, dashed). The matching frequencies (black dots) are close to those that would minimize the $\mathcal{L}^2$-norm between the two curves under the constraint that they cross at $\omega = 0$.

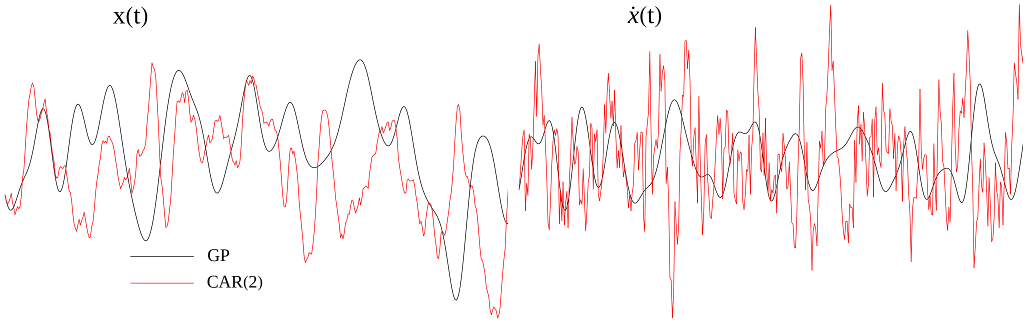


Figure 3.4: A comparison of sample traces and their derivatives from the squared exponential GP and from the fitted CAR(2). While all the derivatives of the GP are smooth, the first derivative of the CAR(2) process looks (on a small scale) like the Ornstein-Uhlenbeck process, i.e. it is MS continuous but not MS differentiable.

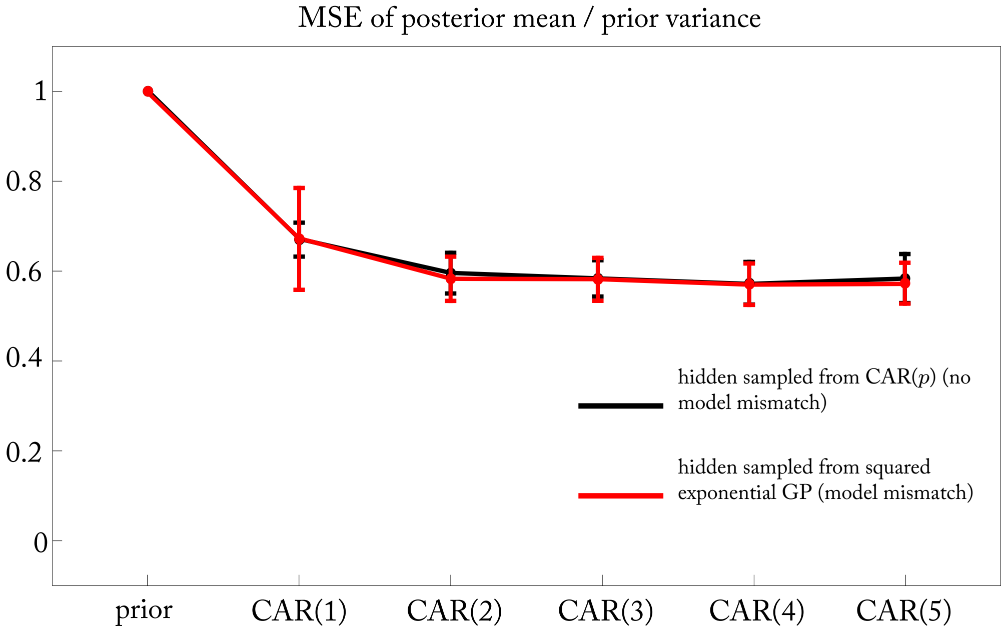


Figure 3.5: Comparison of the filter performance in terms of the mean-squared error in units of the prior variance. Black line: The signal is sampled from a CAR(p) process, which means that the model under which filtering is performed coincides with the ground truth. Red line: The signal is sampled from a squared-exponential Gaussian process, but filtered according to its CAR(p) approximation. The performance in the two cases is equivalent, which shows that the filtering performance is not limited by the approximation of the signal process. Length of the time window: $T = 300$. Parameters of the GP: $k_0 = \theta = 1$. Parameters of the spike emission: $g_0 = \beta = 1$. Matching frequencies for the CAR(p) approximation: $\omega_0 = 0$, $\omega_i = 2^{i-(p+1)/2}$, $1 \leq i \leq p$.

3.4 Analysis of the Dynamics and Predictions

Here, we want to analyse the dynamics of the posterior for the adaptive point process from section 3.2.1. Recall the equations for mean, variance, and the expected firing rate,

$$d\mu_t = -\theta\mu_t dt + \beta_t\sigma_t^2(dN_t - \gamma_t dt), \tag{3.132}$$

$$d\sigma_t^2 = \left(b^2 - 2\theta\sigma_t^2 - \beta_t^2\sigma_t^4\gamma_t\right) dt, \tag{3.133}$$

$$\gamma_t = g_0 \exp\left[\beta_t\mu_t + \frac{1}{2}\beta_t^2\sigma_t^2 + A_t\right], \tag{3.134}$$

where we omitted tildes for notational convenience. The fact that these equations describe an approximate posterior is not relevant for the following discussion, we simply assume that the approximation error is small and that they estimate the presynaptic membrane potential with an accuracy which is sufficient in a biological setting. Also recall that the interpretation of these equations in the context of the KTN theory is that the posterior mean μ_t corresponds to the post-synaptic potential and the posterior variance σ_t^2 is some kind of resource variable.

Establishing a biological interpretation of the variables A_t and β_t in the context of STP is an open problem. In the generative model, they correspond to the presynaptic adaptation variables which determine the spiking output of the presynaptic cell. In order to allow for the estimation of the presynaptic membrane potential, the synapse has to have exact copies of these processes. In the following subsections, we will present various analyses of how these presynaptic mechanisms alter the dynamics of STP. Therefore we assume that the synapse has access to its own instantiations of A_t and β_t, and even though we call them adaptation variables, we think of them as being local quantities of the synapse akin to μ_t and σ_t^2.

First, we consider a model where A_t is a simple process and β_t is a constant. This corresponds to the case where the presynaptic cell has firing-rate adaptation. We show that this can require the synapse to exhibit short-term facilitation for certain parameter regimes, and that the parameter space is structured in the same way as the parameter space of the three-variable Markram-Tsodyks (MT) model. Following that, in Subsection 3.4.2 we suggest a synaptic interpretation of A_t, which however leads to a problem of time-constants. We make an attempt to solve this problem by introducing a time-dependent β_t. We then show that the resulting four-variable normative STP model is richer than the MT model by fitting it to very short-term plasticity which the MT model fails to explain (see Fig. 1.2).

3.4.1 Adaptation can lead to facilitation

For this section, we let $\beta_t = \beta > 0$ constant and assume a simple dynamics for the adaptation variable, i.e.

$$dA_t = -\theta_r A_t dt - \eta_0 dN_t, \tag{3.135}$$

which means that A_t can be expressed as a convolution of an exponential kernel with the previous spikes,

$$A_t = A_0 + \int_0^t \eta(t-s)dN_s, \quad \eta(t) = -\eta_0 e^{-\theta_r t}. \tag{3.136}$$

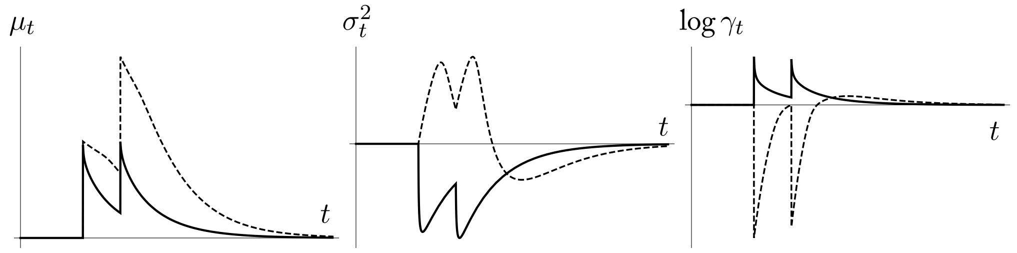


FIGURE 3.6: The qualitative dynamics of μ_t, σ_t^2, and γ_t of Eqs. (3.132-3.134) in response to a pair of spikes. The solid black line shows the response in the absence of adaptation ($A_t = 0$): γ_t increases after the first spike, leading to a decrease in σ_t^2 and a smaller second EPSP. The dashed line shows the response with adaptation ($\eta_0 > 0$): γ_t now decreases after the first spike, leading to an increase in σ_t^2 and a larger second EPSP. Also note that for the chosen parameter values σ_t^2 experiences a rebound, and therefore the second EPSP could be smaller for a certain range of inter-pulse intervals, despite the fact that $\eta_0 < 0$. See also Fig. 3.7.

We call $\eta(t)$ *adaptation kernel*, and $\tau_r = 1/\theta_r$ the adaptation time-constant. Here, we assume $\eta_0 \geq 0$ in order to make the process N_t non-explosive.

We consider the paired-pulse ratio (PPR) as a function of the inter-pulse interval Δt, as predicted by the STP model in Eqs. (3.132-3.134). More precisely, we calculate the ratio of presynaptic responses to a sequence of two pulses at $t = 0$ and $t = \Delta t$, i.e.

$$N_t = H(t) + H(t - \Delta t), \tag{3.137}$$

where H is the Heaviside function. The postsynaptic potential jump in response to a spike equals the dN_t term in Eq. (3.132) and is proportional to the value of σ_t^2 at the time of the spike. Therefore, the PPR reads

$$\mathrm{PPR}(\Delta t) = \frac{\sigma_{\Delta t}^2}{\sigma_0^2}. \tag{3.138}$$

In accordance with how PPR measurements are usually conducted *in vitro*, namely with no spikes occurring in the few seconds before the first spike of the pulse pair, we choose initial conditions for μ_0, σ_0^2, and A_0 that are equal to the stationary values in the absence of spikes. This implies that $A_0 = 0$ and that μ_0, σ_0^2 are the solutions to the nonlinear equations (obtained by setting dN_t and the left-hand sides of Eqs. (3.132,3.133) to zero)

$$0 = -\theta\mu_0 - \beta\sigma_0^2\gamma_0, \quad 0 = b^2 - 2\theta\sigma_0^2 - \beta^2\sigma_0^4\gamma_0. \tag{3.139}$$

The numerator of Eq. (3.138) is then obtained by integrating the system of ordinary differential equations in Eqs. (3.132-3.134) (note that $dN_t = 0$ between the two pulses and that the second pulse does influence the value of $\sigma_{\Delta t}^2$) from 0 to Δt.

Inspection of Eqs. (3.132-3.134) already reveals that the presence of a non-zero adaptation variable A_t has the potential of altering the dynamics quite significantly:

- With $A_t = 0$ (see Fig. 3.6, solid line), the posterior firing rate γ_t increases instantaneously after the first spike due to the sudden increase (EPSP) of μ_t from μ_0 to $\mu_0 + \beta\sigma_0^2$. Therefore, σ_t^2 decreases, and the subsequent spike will lead to a jump in μ_t from $\mu_{\Delta t}$ to $\mu_{\Delta t} + \beta\sigma_{\Delta t}^2$ which is smaller than the previous one. Therefore, in the absence of adaptation, the STP model (3.132-3.134) leads to depression, i.e. the PPR(Δt) is negative for all values of the inter-pulse interval Δt.

- With $A_t < 0$ (see Fig. 3.6, dashed line), i.e. if $\eta_0 > 0$, the increase in μ from μ_0 to $\mu_0 + \beta\sigma_0^2$ at the time of the first spike is counter-acted by a decrease in A_t from 0 to $-\eta_0$. If $\eta_0 < \beta\sigma_0^2$, the firing rate is still increased, albeit less than in the previous case. However, if $\eta_0 > \beta\sigma_0^2$, the firing rate is decreased, leading to an increase in σ_t^2 which makes it (at least temporarily) bigger than σ_0^2. Therefore, in a certain time window, a second spike will lead to a jump in μ_t which is bigger than the first one.

The condition for the occurrence of PPRs bigger than unity for certain values of the inter-pulse interval Δt is

$$\eta_0 > \beta\sigma_0^2, \tag{3.140}$$

but the PPR can become smaller than unity again for bigger values of Δt, leading to a PPR curve which has values above and below zero. Conversely, even if the above condition is violated, the PPR can become bigger than unity for large values of Δt. We therefore distinguish four different qualitative behaviors of the posterior dynamics: 1) purely facilitating, 2) mixed, with facilitation preceding depression, 3) mixed, with depression preceding facilitation, and 4) purely depressing dynamics.

We analyzed the shape of the PPR curve for a large range of parameter values (removing redundancies by using dimensionless quantities). We find that the parameter space for $\eta_0 < 0$ can be neatly partitioned into four regions (see Fig. 3.7A) according to two dimensionless numbers $\mathfrak{r}, \mathfrak{s}$ as follows: 1) $\mathfrak{r}, \mathfrak{s} > 1$, 2) $\mathfrak{r} > 1, \mathfrak{s} < 1$, 3) $\mathfrak{r} < 1, \mathfrak{s} > 1$, and 4) $\mathfrak{r}, \mathfrak{s} < 1$. The relevant quantities read

$$\mathfrak{r} = \frac{\eta_0}{\beta\sigma_0^2}, \qquad \mathfrak{s} = \frac{\tau_r}{\tau_m}, \tag{3.141}$$

where $\tau_m = 1/\theta$ is the membrane time-constant appearing in Eqs. (3.132,3.133).

A similar behavior is seen in the phenomenological Markram-Tsodyks (MT) model, see Section 1.2 and Eq. 1.1. Due to the simplicity of those equations, we can calculate the PPR analytically,

$$\mathrm{PPR}_{\mathrm{MT}}(\Delta t) = \frac{1}{Y}\left(1 - Ye^{-\frac{\Delta t}{\tau_x}}\right)\left(Y + F(1-Y)e^{-\frac{\Delta t}{\tau_y}}\right). \tag{3.142}$$

We find that the two dimensionless numbers

$$\mathfrak{r}_{\mathrm{MT}} = \frac{F(1-Y)^2}{Y^2}, \qquad \mathfrak{s}_{\mathrm{MT}} = \frac{\tau_y}{\tau_x}, \tag{3.143}$$

allow the same type of classification as we did for the normative model above, see Fig. 3.7B.

There are also important differences between the MT model and the normative model. The first is that all the PPR curves start at unity for the normative model, but they have values different from one at $\Delta t = 0$ in the MT model. This follows from the fact that the resource variable x of the MT model sees an instantaneous increase upon a spike, see Eq. 1.1. In contrast, the normative model's resource variable σ_t^2 does not increase immediately, but changes only as a second-order effect due to the change in γ_t (see the analysis above). Related to this is the second distinction, namely that the MT PPR curve has at most one local extremum, whereas the normative model can have more than one.

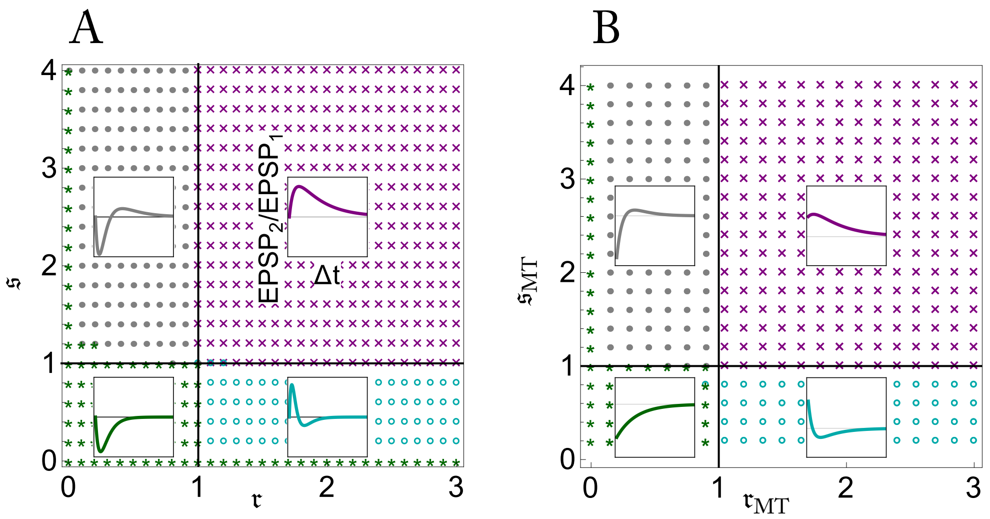


Figure 3.7: Distinct dynamics are seen in different regions of the parameter space of two STP models: (A) In the normative STP theory, described by Eqs. (3.132-3.134), the regions are defined by the values of the dimensionless strength of adaptation, $\mathfrak{r} = \eta_0/\beta\sigma_0^2$, and the ratio of refractory time-constant to the membrane time-constant $\mathfrak{s} = \frac{\tau_r}{\tau_m}$. (B) Analogously, in the Markram-Tsodyks model, the boundaries are determined by $\mathfrak{r}_{\mathrm{MT}} = F(1-Y)^2/Y^2$ and the ratio of time-constants $\mathfrak{s}_{\mathrm{MT}} = \tau_y/\tau_x$.

The new prediction of this section is that there is an intimate link between the presynaptic firing-rate adaptation and the degree of short-term facilitation in the synapses directly downstream. More precisely, the KTN theory predicts that the magnitude of change of the presynaptic adaptation variable A_t in response to a spike determines whether facilitation can be observed in the synapse, and the time-constant of the presynaptic adaptation variable changes the time-course of this facilitation.

3.4.2 The problem of time-constants

The previous analysis of PPR dynamics has revealed a striking similarity between the normative STP model (with adaptation variable A_t) and the MT model. Crucially, whereas the normative STP model produces only short-term depression when $A_t = 0$, as in the original paper Pfister et al. (2009), the extension to adaptive point process dynamics for the presynaptic neuron produces, in the very same general framework, predictions of short-term facilitation. The similarities between the KTN dynamics and the MT dynamics go even further, extending to the precise partitioning of the parameter space shown in Fig. 3.7.

The analysis above suggests that the adaptation variable A_t (or rather, its synaptic instantiation) be mapped to the release probability variable Y_t of the MT model. Indeed, the jump in Y_t after a spike is completely analogous to the (negative) jump in A_t, and it is the size of this jump relative to other parameters which determines whether facilitation occurs. However, there is an important distinction between the two models, namely the number of time-constant parameters. The MT model has three time-constants τ_v, τ_x, τ_y, whereas the normative model so far has two time-constants, τ_m and τ_r.[9]

---

[9] It could be argued that the inverse of g_0 has units of time as well and therefore constitutes a time-constant. However, it does not appear as the time-constant in a dynamical equation, and therefore does not count as such.

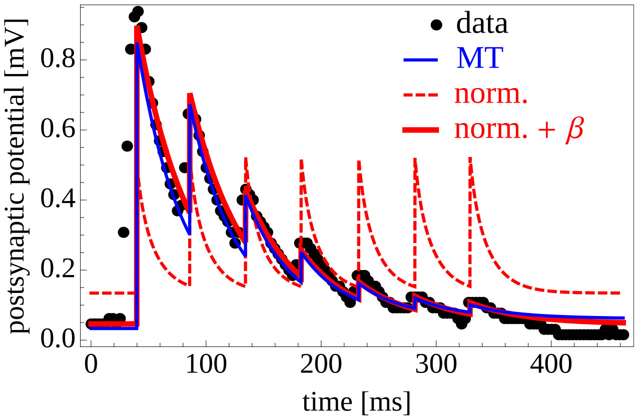


Figure 3.8: Short-term depression data from Markram and Tsodyks (1996) (black dots) is fitted with the Markram-Tsodyks model (blue line) and the KTN model (dashed red line). Whereas the MT model does fairly well, the best estimate for the KTN model parameters yields a static synapse. By introducing a time-dependent β_t, the fit improves considerably (solid red line).

The problem of time-constants has the consequence that some widespread types of short-term depression, namely those where the recovery from depression is much slower than the membrane time-constants, remains outside the scope of the present KTN model. Indeed, the recovery from depression in this model (we can set $A_t = 0$ for this purpose) is governed by the dynamics of σ_t^2, which by Eq. (3.133) has a time-constant of $\tau_m/2$. Therefore, when we fit the KTN model to an example STP dataset (Markram and Tsodyks, 1996), the best-fit parameters are $\tau_m = 33.9$ ms, $b = 0.21$ mV/ms$^{1/2}$, $g_0 = 121$ Hz, and $\beta = 0.974941$ mV^{-1}. This choice of parameters produces a response which looks like a static synapse (see Fig. 3.8). In contrast, the MT yields a reasonable fit. The MT on the other hand gives a reasonable fit with best-fit parameters $J = 1.40$ mV, $Y = 0.61$, $F = 0.64$, $\tau_v = 46.5$ ms, $\tau_x = 2.2$ s, and $\tau_y = 1.04$ ms). Note that the optimization problems are non-convex, so despite the use of random searches the existence of slightly better fits cannot be excluded.

An additional time-constant can be introduced by making the gain parameter β time-dependent. More precisely, we introduce simple dynamics for β_t, namely

$$d\beta_t = \frac{\beta_0 - \beta_t}{\tau_\beta} + B\beta_t dN_t, \quad B \geq -1, \tag{3.144}$$

where the constraint for B prevents β from becoming negative. The effect of this extension is that the presynaptic firing probability becomes more or less sensitive to the value of the membrane potential after a spike (see Fig. 3.9). This effect decays after a characteristic time-constant of τ_β.

Equipped with a time-dependent β and a new time-constant parameter, the model can now reasonably fit the data from Markram and Tsodyks (1996), see Fig. 3.8. The fitted time-constant τ_β is 4.7 seconds, whereas the membrane time-constant is $\tau_m = 40$ ms. The remaining best fit parameters read $b = 0.24$ mV/ms$^{1/2}$, $g_0 = 0.66$ Hz, $\beta_0 = 0.72$ mV^{-1}, and $B = -0.55$. Moreover, this STP-data motivated extension of the model predicts that STP data showing a pronounced difference of time-constants (i.e. between the depression

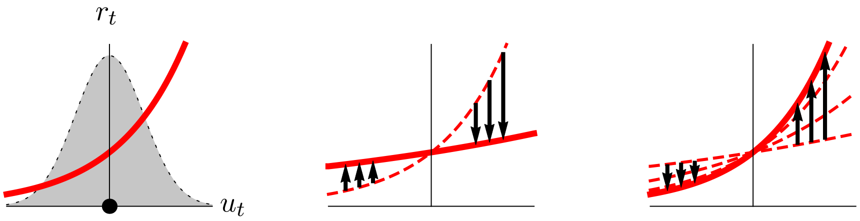


FIGURE 3.9: The effect of the dynamic β_t variable on the presynaptic firing rate r_t through the gain function $r_t = g_0 e^{\beta_t u_t}$ shown in red. The membrane potential u_t follows a Gaussian distribution, and higher values lead to higher firing rate (left panel). If β_t decreases instantaneously upon a spike, the gain function (red) becomes shallower (center panel), reducing the firing rate for high values of the membrane potential and increasing it if the membrane potential is low. Since the membrane potential is likely to be high when a spike occurs, this often leads to a decrease in firing rate. After the spike, β_t decays back (right panel), restoring the shape of the gain function.

time-constant and the membrane time-constant), must have pronounced and long-lasting presynaptic adaptation effects best described by the dynamics of β_t in Eq. (3.144).

Making β time-dependent changes the behavior of the posterior and therefore the conditions for and the properties of STP. Essentially, the analysis presented in Section 3.4.1 for constant β has to be redone (see Open Problems below).

3.5 Discussion and Future Directions

In this chapter, we reformulated the theory in a more general and more easily extendable framework. In particular, inference calculations required for predicting postsynaptic responses are now possible for a large class of generative models, including those where the membrane potential is a multivariate Ornstein-Uhlenbeck process or indeed a Gaussian process instead of a simple OU process, and the spikes follow an adaptive point process instead of a simple inhomogeneous Poisson process. We looked at one extension in detail, namely the addition of presynaptic adaptation to the spiking mechanism, and showed its link to short-term facilitation. We also showed how to match the phenomenological model of STP, the Markram-Tsodyks model and proposed one possible resolution of the problem of time-constants.

By performing the extensions to the KTN theory that are mentioned above, we are carrying out the ideas which were sketched in the original paper. Indeed, in Pfister et al. (2009) it was said that

> "Our assumption about the prior distribution of presynaptic membrane potential dynamics is highly restrictive. A broader scheme that has previously been explored is that it follow a Gaussian process model [...] with a more general covariance function. Recursive estimation is often a reasonable approximation in such cases, even for those covariance functions, for instance enforcing smoothness, for which it cannot be exact."

To wit, this program was carried out in Section 3.3. Along with the results of Section 3.2.1, it allows STP predictions to be derived for the full AGAPE model proposed in Chapter 3 as a generative model for the presynaptic dynamics.

3.5.1 Suggested experiments

Despite the fact that the KTN theory, both in its original form and with the extensions presented in this section and the last section, provides a number of specific predictions for how STP types and dynamics should be allocated in the brain, a direct experimental test of these ideas is still missing. Here, we sketch a few ideas for potential experiments.

Testing the general link between presynaptic statistics and STP

The KTN theory, both in its original and extended forms (including adaptation), predicts that in order to perform the estimation task, the short-term synaptic properties should be tuned to the statistics of the presynaptic neurons. In the mathematical formulation of the theory, this is evidenced by the fact that the presynaptic parameters appear in the equations governing the synaptic dynamics.

In order to test this hypothesis, intracellular *in vivo* recordings have to be obtained from a specific neuron. Then, a suitable presynaptic model has to be fitted to the recorded data (see Chapter 3). The presynaptic model has to be in a class which allows STP predictions to be derived according to Section 3.2, e.g. a Gaussian process for the membrane potential (signal) and an adaptive point process for the spikes (observations). STP predictions can be done for artificial stimuli (e.g. pulse trains or paired pulses) or for naturalistic ones (i.e. as sampled from the presynaptic model). *In vitro* STP experiments on a synapse which is downstream to the previously recorded neurons will then measure the postsynaptic responses (preferably, in the dendrite close to the synapse) to presynaptic stimulation according to the protocols for which predictions have been made. Comparison of the measured and predicted responses will establish whether that particular synapse is attuned to the presynaptic neuron's statistics in a way required to estimate the presynaptic membrane potential.

Besides the technical challenges of such an experiment, there are a number of theoretical concerns with its interpretability. First of all, even if we take the position that the KTN theory is true and that certain synapses indeed perform estimation of presynaptic quantities, we do not expect a random synapse to exhibit this behavior. Therefore, a negative outcome of the above experiment disproves the blunt statement that all synapses perform estimation of the presynaptic membrane potential, but it does not constrain a weaker form of the theory which states that *certain* synapses perform estimation of the presynaptic membrane potential, or that they perform estimation of other presynaptic quantities, such as nonlinear functions of the presynaptic potential. Only a sufficiently large number of experiments will allow one to establish an estimate of the fraction of synapses invested in this specific estimation task. As for the alternative hypotheses, we believe that they need more careful formulation (see the discussion in Section 3.5).

While we haven't been able to arrange for this experiment to be performed, we have assembled all the theoretical tools required to predict the outcome, which is an important achievement of this thesis.

Testing specific predictions related to presynaptic adaptation and STP

The extensions of the KTN theory have produced two new predictions. The first is that sufficiently strong (either by magnitude or time-constant) presynaptic adaptation requires short-term facilitation in the downstream synapse for the estimation of the presynaptic membrane potential. The second is that under the assumptions of the KTN theory, one requires a very specific type of presynaptic adaptation (namely of the coupling parameter β in the gain function) in order to produce a specific type of short-term depression which has a long depression time-constant relative to the membrane time-constant.

The testing of these hypotheses requires a good knowledge of the synapses exhibiting the estimation of the presynaptic membrane potential in order to reduce the risk of 'fishing in the dark'. Given a population of candidate synapses, one can then (for example) select a subset which is strongly facilitating and measure presynaptic dynamics. Comparing models of presynaptic activity with and without adaptation can in principle establish whether there is a connection.

3.5.2 Open problems

There are a number of open problems which were not addressed in detail. Some of them are indeed now possible to address, given the advances in theoretical tools that have been provided with this thesis.

Optimality of the encoder

In Chapter 1, we mentioned the possible benefit of the analog-to-digital-to-analog conversion between somatic membrane potential of the presynaptic cell and the dendrite of the postsynaptic cell in the context of long-distance projections. According to the KTN theory, STP serves to optimally decode the digital signal transmitted through the axon. So far it assumes one type of encoding (exponential gain function with or without adaptation) and asks for the corresponding optimal decoder (filter). Since the performance of the filter is constrained by the bandwidth of the encoder, i.e. the filter error is directly linked to the mutual information between the presynaptic membrane potential and the spikes (Guo et al., 2008), the system is not working optimally unless encoding maximizes the mutual information within some constraints (e.g. maximum firing rate). This raises the question what this optimal encoder looks like, and how it affects the relation between presynaptic and postsynaptic quantities.

Target-cell specificity problem

Foremost among the challenges to the KTN theory is the target-cell specificity problem. The main assertions of the KTN theory are massively challenged by both the possibility that multiple synapses from the same cell can have different STP properties and the cases where the type of STP can be predicted by knowing the cell type the synapse is projecting to (Reyes et al., 1998; Markram et al., 1998; Blackman et al., 2013). In order to resolve the problem, one has to make the theory's claims more case-specific. Specifying where in the brain estimation (in the sense described by the theory) can be helpful, and which quantities should be estimated, is the next important step towards that goal. Then, using the general framework developed in Section 3.1, one can develop filters of various presynaptic quantities, and predict the range of STP properties they produce according to the theory. These predictions can be compared to experimental findings.

There are two immediate possibilities to address the target-cell specificity problem. The first is to assume that the different synapses estimate different components of the presynaptic membrane potential. This can be formalized as a filtering problem where a multivariate Ornstein-Uhlenbeck process is observed through a spike train. The multivariate OU describes different signals present within the presynaptic membrane potential, and each synapse has the task of filtering out one of these components and communicate it to the postsynaptic cell. The resulting synapse population will have a range of STP properties.

The second option that is accessible within the mathematical framework presented in this thesis is to allow for different synapses to estimate different, possibly nonlinear functions of the presynaptic membrane potential. Again, this will require different STP dynamics for different synapses.

Both approaches could potentially explain away the variabilities seen in target-cell specific STP and make them fit the KTN theory with the modification that the target of estimation is target-cell specific, but estimation *per se* is not. In these scenarios, the target-cell determines the component or function to be estimated, which in turn determines the type of STP in that synapse. This reconciles the idea of the KTN theory that estimation is performed by synapses with the observation that the STP properties can be told from the postsynaptic cell-type. Indeed, if a certain postsynaptic cell-type is involved in a computation requiring a specific signal or transformation from the presynaptic cell, the theory shows how this may be made possible by equipping the synapse with a suitable STP dynamics.

Further research is necessary in order to establish the degree to which this explanation fits STP data, or whether it is too general to remain falsifiable.

Parameter learning

The KTN theory states that the STP dynamics are attuned to the presynaptic dynamics, embodying some kind of prior knowledge of the presynaptic statistics of membrane potential and spikes. This prior knowledge could be present either by genetic predetermination or by adaptive processes which allow it to be learned from the statistics of spikes. At any rate, neither the original KTN theory nor the extended version of this thesis offer any explanation how this learning could be achieved. Let us give a brief outline or plan how this gap could be filled.

Mathematically speaking, the problem of learning the presynaptic parameters can be phrased as recursive identification of a partially observed system. This is an old problem in stochastic filtering and control, and there is a correspondingly large body of literature about it (see Kantas et al. (2014) and references therein). A recursive maximum likelihood approach can be used to derive a recursive update rule for the parameter θ (any of the presynaptic parameters at play), which in discrete time takes the form

$$\theta_{t+1} = \theta_t + \gamma_{t+1} \nabla p_{\theta_{0:t}}(y_t | y_{0:t-1}), \tag{3.145}$$

where

$$\nabla p_{\theta_{0:t}}(y_t | y_{0:t-1}) = \nabla p_{\theta_{0:t}}(y_{0:t}) - \nabla p_{\theta_{0:t}}(y_{0:t-1}). \tag{3.146}$$

While this approach provides the mathematical form of the learning rules, the harder problem is to interpret these rules in a biological setting. For example, the parameter β, which is the coupling parameter of the presynaptic firing rate to the membrane potential fluctuations (see Eq. (3.31)), acquires the role of a long-term synaptic strength parameter in Eqs. (3.58,3.59) for the posterior. The recursive parameter update rule will therefore be a synaptic plasticity rule from that perspective. The interpretation is less clear for the other parameters of the model.

Dynamic analysis for time-dependent β

In Section 3.4.1 we investigated the dynamics of the equations for the posterior distribution for constant β and discovered a nice structure of the parameter space which can be mapped to the Markram-Tsodyks model. This analysis would change for the extended model with a time-dependent β, and we haven't had the time to redo it for this case. The parameter space becomes higher-dimensional due to the introduction of τ_β and B. We can expect the parameter space to be still divided into different regions of STP properties, but there might now be more than four classes and possibly nonlinear boundaries. Therefore, an analysis similar to the one in Section 3.4.1 is likely to be more complex.

Stationarity of the presynaptic neuron

So far, the KTN theory assumes that the presynaptic cell's dynamics is stationary, such that it is possible for the synapse to learn the appropriate generative model, either from the statistics themselves or by being genetically preset. This assumption can be questioned on the grounds that a lot of neurons display a set of completely disparate behavior depending on such factors as neuromodulation or the source and type of synaptic input and associated non-linear gating effects.

A presynaptic neuron which has even just two distinct modes of operation requires a multi-state or switching model, and the synapse has to infer, in addition to the membrane potential, the state that its presynaptic partner is currently in. Such a switching model has been considered in Pfister et al. (2010) for up- and down-states.

Given the switching of neuronal states, the synapse could have the strategy of having different estimation targets for the different states, or being in the 'estimation' state only for certain states of the presynaptic neuron. It could be conceivable that the STP properties are attuned to one state or behavior of the presynaptic neuron, allowing the synapse to infer the presynaptic membrane potential when the presynaptic neuron is in that state, and doing a different computation in another state. However, although this is a plausible possibility, such an arrangement would be even harder to detect experimentally.

Multisynaptic generalization

The KTN theory focuses on one single synapse and the computations it can perform by having the 'right' STP dynamics, i.e. those attuned to the presynaptic neuron's statistics. However, an average neuron receives thousands of synaptic inputs, and it is the combination of all these inputs determining the neuron's output. Therefore, it could be argued that the computational function assigned to a single synapse within a neuron should have a correspondence to the computational function of the neuron within a local circuit, and the circuit's function within a brain region or indeed the brain.

The mathematical framework presented in this thesis allows in principle to treat the case where multiple spike trains are observed and used to infer the joint state of multiple presynaptic neurons. This could serve to infer a specified function of all the presynaptic inputs. However, the biological mechanisms which could support such a filter are difficult to find, and it would also be difficult to validate such a theory without knowing the function the system is trying to compute. However, such a theory might have merits as a conceptual tool.

Appendix A

A.1 A brief Introduction to SDEs and Stochastic Filtering Theory

Here, we review the material necessary to understand the filtering theory aspects of this thesis. We will introduce the theory of continuous-time Markov processes from the perspective of Itô stochastic calculus and stochastic differential equations (SDEs). Then, we will introduce the key concepts of stochastic filtering theory. We assume that the reader is somewhat familiar with the basic concepts of probability theory.

This brief introduction by no means replaces a thorough study of the subject, but it should give the reader a basic idea of the relevant topics. For a more in-depth study, we suggest the following textbooks. For a general introduction to stochastic processes, see Klebaner (2005) or Gardiner (1985). For a treatment of stochastic filtering theory specifically, see Bain and Crişan (2009) or Jazwinski (1970).

A.1.1 The Wiener process, stochastic integration, stochastic calculus

In order to introduce SDEs, we make use of the Wiener process, defined as the stochastic process $\{W_t, t \geq 0\}$ with the following properties (Klebaner, 2005)

1. $W_0 = 0$,
2. Its increments are independent, i.e. $W_s - W_t$, $s > t$ is independent of W_u, $0 \leq u \leq t$,
3. Its increments are normal, i.e. $W_s - W_t \sim \mathcal{N}(0, s-t)$, $s > t$,
4. Its paths are continuous.

Given the Wiener process, one can define stochastic integration. This can be properly done by first defining the integral for simple processes which are piecewise-constant, and then taking the limit of approximations of a more general process by simple processes (Klebaner, 2005). Here, however, we want to give a simpler definition (Gardiner, 1985). The stochastic integral of the process $X = \{X_t\}$ with respect to the Wiener process is defined as

$$\int_0^t X_s dW_s \doteq \lim_{n\to\infty} \sum_{i=1}^n X_{t_{i-1}}(W_{t_i} - W_{t_{i-1}}). \tag{A.1}$$

In Eq. (A.1), for any $n \in \mathbb{N}$, $0 = t_0 < t_1 < \ldots < t_n = t$ is a partition of the interval $[0, t]$ and the limit is taken in the mean-square (MS) sense.[1] The convention that the process X is evaluated at the beginning of each subinterval is the Itô convention, and the integral as defined above is called the Itô integral. In the other common convention, called the Stratonovich convention, the process X_t is taken mid-interval.

For which class of processes X does the integral above make sense? Here and later in filtering theory, filtrations are important, so it is convenient to introduce them here. A filtration $\mathcal{F} = \{\mathcal{F}_t\}$ is an increasing (i.e. $\mathcal{F}_t \subset \mathcal{F}_s$ for $s > t$) family of σ-algebras, with the algebra $\mathcal{F}_t$ containing the information up to time t. In the case that the Wiener process is the only source of information, the filtration is generated by the values of the Wiener process, which we write as $\mathcal{F}_t = \sigma(\{W_s, 0 \le s \le t\})$. Moreover, we say that a process X is adapted to the filtration $\mathcal{F}$ if X_t is $\mathcal{F}_t$-measureable, which means that it is possible to decide based on information contained in $\mathcal{F}_t$ whether an event for X_t (e.g. $X_t > 0$) has occurred. In other words, the process X is non-anticipating; it does not see into the future.

If X is continuous and adapted to the filtration $\mathcal{F}_t = \sigma(\{W_s, 0 \le s \le t\})$, then the Itô integral in Eq. (A.1) is well-defined. Moreover, the process Y defined by $Y_t = \int_0^t X_s dW_s$ is also adapted, and if X satisfies

$$\int_0^t \mathbb{E}\left[X_s^2\right] ds < \infty, \tag{A.2}$$

then Y_t has zero mean and variance given by the left-hand side of Eq. (A.2), and is a martingale, i.e.

$$\mathbb{E}\left[Y_t|\mathcal{F}_s\right] = Y_s, \qquad s \le t. \tag{A.3}$$

However, if condition (A.2) is not satisfied, Y_t is only a local martingale.

Can we build other adapted processes from the Wiener process and an adapted process X? One can show that

$$\int_0^t X_s dW_s^2 \doteq \lim_{n\to\infty} \sum_{i=1}^n X_{t_{i-1}}(W_{t_i} - W_{t_{i-1}})^2 = \int_0^t X_s ds, \tag{A.4}$$

$$\int_0^t X_s dW_s^n \doteq \lim_{n\to\infty} \sum_{i=1}^n X_{t_{i-1}}(W_{t_i} - W_{t_{i-1}})^n = 0, \quad n > 2, \tag{A.5}$$

meaning that integrals involving expressions beyond dW_s do not make sense as they either vanish or reduce to an ordinary integral. Moreover, expressions involving combinations of Wiener process and time increments vanish as well, such as

$$\int_0^t X_s dW_s dt \doteq \lim_{n\to\infty} \sum_{i=1}^n X_{t_{i-1}}(W_{t_i} - W_{t_{i-1}})(t_i - t_{i-1}) = 0. \tag{A.6}$$

---

[1] A sequence $(V_n, n \in \mathbb{N})$ of random variables converges to V in the MS sense if $\mathbb{E}\left[(V_n - V)^2\right] \to 0$ for $n \to \infty$.

Thus, from the Wiener process and adapted process X, one can construct the following adapted processes:

$$F(X_t, W_t, t), \qquad \int_0^t G(X_s, W_s, s) dW_s, \qquad \int_0^t H(X_s, W_s, s) ds, \tag{A.7}$$

where F, G, H are continuous functions. Given these building blocks, an Itô equation for X_t is a relation of the form

$$X_t = X_0 + \int_0^t a(X_s, s) ds + \int_0^t b(X_s, s) dW_s. \tag{A.8}$$

This is a stochastic integral equation which describes a relation between X_t and the previous values of the process. We often write this equation in the differential form

$$dX_t = a(X_t, t) dt + b(X_t, t) dW_t, \tag{A.9}$$

which is called an Itô stochastic differential equation (SDE). It is important to remember that the SDE obtains its meaning in terms of the integral equation Eq. (A.8) only.

Itô calculus consists in using the differential representation and the rules $dW_t^2 = dt$, $dW_t^n = 0$, $n > 2$, etc., to do calculations. For instance, let X_t satisfy the SDE (A.9) and let $Y_t = f(X_t, t)$ be a nonlinear time-dependent transformation of the process X. Then, provided that f is twice differentiable in its first argument and once in its second, there is a stochastic chain rule that lets us write an SDE for Y_t. Since terms containing $dt^p dW_t^q$ vanish under the integral if $p > 1$ or $q > 2$ or $p = q = 1$, we can drop all terms higher than dX_t^2 in the Taylor expansion of f. We may therefore write

$$\begin{aligned} dY_t &= \partial_t f(X_t, t) dt + \partial_x f(X_t, t) dX_t + \frac{1}{2} \partial_x^2 f(X_t, t) dX_t^2 + ... \\ &= \left(\partial_t f(X_t, t) + a(X_t, t) \partial_x f(X_t, t) + \frac{1}{2} b^2(X_t, t) \partial_x^2 f(X_t, t) \right) dt \\ &\quad + \left(b(X_t, t) \partial_x f(X_t, t) \right) dW_t. \end{aligned} \tag{A.10}$$

In the second line we used $dW_t^2 = dt$ and $dW_t dt = 0$ in evaluating dX_t^2. Equation (A.10) is called Itô's formula or Itô's lemma. Taking expectations, we can drop the dW_t term (under the assumptions given above for the dW_t integral), and we find that

$$\frac{d\mathbb{E}\left[Y_t\right]}{dt} = \mathbb{E}\left[\partial_t f(X_t, t) + a(X_t, t) \partial_x f(X_t, t) + \frac{1}{2} b^2(X_t, t) \partial_x^2 f(X_t, t) \right]. \tag{A.11}$$

By introducing the operator $\mathcal{L} = a(X_t, t)\partial_x + \frac{1}{2} b^2(X_t, t) \partial_x^2$, called infinitesimal generator of the process X, we can write this as

$$\frac{d\mathbb{E}\left[Y_t\right]}{dt} = \mathbb{E}\left[\partial_t f(X_t, t) + \mathcal{L} f(X_t, t) \right]. \tag{A.12}$$

Equation (A.12) is the key to calculating time-evolution equations for the moments of the process X.

A.1.2 The Fokker-Planck equation

If the Itô SDE (A.9) has a solution with an initial condition X_0, it defines a continuous Markov process $\{X_t, 0 \leq t \leq T\}$ (if it is non-explosive, on the entire non-negative real line). We assume that the conditional probability density $p(x,t|x_0,0)$ exists. The conditional expectation of $\varphi(X_t)$, which can be expressed by virtue of the density as

$$\mathbb{E}\left[\varphi(X_t)|X_0 = x_0\right] = \int_{-\infty}^{\infty} \varphi(x)p(x,t|x_0,0)dx, \tag{A.13}$$

satisfies the differential equation

$$\frac{d\mathbb{E}\left[\varphi(X_t)|X_0 = x_0\right]}{dt} = \mathbb{E}\left[\mathcal{L}\varphi(X_t)|X_0 = x_0\right], \tag{A.14}$$

which follows from Eq. (A.12). By expressing both sides of Eq. (A.14) in terms of the density, we find

$$\begin{aligned}\frac{d\mathbb{E}\left[\varphi(X_t)|X_0 = x_0\right]}{dt} &= \int_{-\infty}^{\infty} \varphi(x)\partial_t p(x,t|x_0,0)dx \\ &= \int_{-\infty}^{\infty} \mathcal{L}\varphi(x)p(x,t|x_0,0)dx = \mathbb{E}\left[\mathcal{L}\varphi(X_t)|X_0 = x_0\right].\end{aligned} \tag{A.15}$$

If the transition density and its first derivative vanish at infinity, we can integrate the second integral by parts and obtain

$$\int_{-\infty}^{\infty} \varphi(x)\partial_t p(x,t|x_0,0)dx = \int_{-\infty}^{\infty} \varphi(x)\mathcal{L}^{\dagger}p(x,t|x_0,0)dx, \tag{A.16}$$

where

$$\mathcal{L}^{\dagger} = \partial_x\left(a(\cdot,t)\cdot\right) + \frac{1}{2}\partial_x^2\left(b^2(\cdot,t)\cdot\right) \tag{A.17}$$

is the adjoint of $\mathcal{L}$, called Fokker-Planck operator. Since this holds for arbitrary φ, we deduce that the transition probability density must satisfy the partial differential equation

$$\partial_t p(x,t|x_0,0) = \mathcal{L}^{\dagger}p(x,t|x_0,0), \tag{A.18}$$

called the *Fokker-Planck equation* (FPE). The transition probability density is the solution to the FPE with the singular initial condition $p(x,0|x_0,0) = \delta(x - x_0)$. If the initial state X_0 has a density $p_0(x)$, the one-time probability density can be calculated as

$$p(x,t) = \int_{-\infty}^{\infty} p_0(y)p(x,t|y,0)dy, \tag{A.19}$$

and also satisfies the FPE.

A.1.3 Stochastic calculus for counting processes

A counting process $N = \{N_t, t \geq 0\}$ is a pure jump process with jumps of size one (also called a simple point process). It can be represented as

$$N_t = \sum_{n=1}^{\infty} H(t - T_n), \tag{A.20}$$

where H is the Heaviside function ($H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ otherwise) and $(T_n \geq 0)$ is the random sequence of arrival times. Therefore, N is right-continuous with left limits (càdlàg: continue à droite, limite à gauche). Stochastic calculus for counting processes falls under the umbrella of stochastic calculus for semimartingales. However, for the purposes of this thesis we do not need the full machinery of that calculus. We simply define the integral of an adapted process X with respect to a counting process as

$$\int_0^t X_s dN_s = \sum_{n=1}^{\infty} X_{T_n^-} H(t - T_n), \tag{A.21}$$

where $T_n^- = \lim_{t \nearrow T_n} t$ denotes the left limit. The limit is required to ensure that the function to be integrated is predictable, i.e. that its value is known at the time of the arrival of the jump in N. This is not a problem for continuous integrands, but crucial for integrands containing jumps. For example, if $X_t = N_t$, the value of N increases by one at the arrival time T, so the contribution to the integral must come from the value of N prior to the jump.

Expressions involving dN_t are always understood to be evaluated with the left limit as discussed above. One can use the counting process to drive the evolution of a jump process:

$$X_t = X_0 + \int_0^t a(X_s, s)ds + \int_0^t b(X_s, s)dN_s. \tag{A.22}$$

This process will evolve deterministically (and be differentiable) between arrival times, and have a jump of size $b(X_{T^-}, T^-)$ at an arrival time T. The above integral equation can also be written in the differential form

$$dX_t = a(X_t, t)dt + b(X_t, t)dN_t. \tag{A.23}$$

By combining dN_t and dW_t terms, one can write SDEs for jump-diffusion processes, but we do not use them in this thesis. For a process X satisfying Eq. (A.23), we can write a stochastic chain rule. Let Y be the process defined by $Y_t = f(X_t, t)$. By observing that between jumps, the differential of Y_t obeys the classical chain rule and that at a jump arrival time, the jump in Y_t is equal to the nonlinear transformation of the jump in X_t (i.e. the difference between the values of f after and before the jump), we end up with

$$dY_t = [\partial_t f(X_t, t) + a(X_t, t)\partial_x f(X_t, t)]\, dt + [f(X_t + b(X_t, t), t) - Y_t]\, dN_t. \tag{A.24}$$

One may wish to formulate an SDE for the evolution of the counting process. This is not strictly necessary, since the process may be defined by explicitly stating the probability of a sequence of arrival times or of the distribution of increments. For example, an inhomogeneous Poisson process N with the stochastic intensity $g(X_t)$ which depends on a diffusion process X is characterized by the properties

1. $N_0 = 0$,
2. Non-overlapping increments are independent, i.e. for $t > s > t' > s'$, $N_t - N_s$ and $N_{t'} - N_{s'}$ are independent,
3. Each increment has a Poisson distribution, i.e. $N_t - N_{t'} \sim$ Poisson $\left(\int_{t'}^t g(X_s)ds\right)$, $t > t'$.

A similar definition is possible for the adaptive point processes discussed in this thesis. If one wishes to write an SDE, one can make use of the Doob-Meyer decomposition for semimartingales. For a counting process with a stochastic intensity λ_t (which may depend on a latent process and on the counting process itself), this decomposition takes the form

$$dN_t = \lambda_t dt + dM_t, \tag{A.25}$$

where M_t is a martingale. The process $\Lambda_t = \int_0^t \lambda_s ds$ is called the compensator of N. It is the predictable projection of the process N.

A.1.4 Stochastic filtering

In many applications, such as this thesis, one is interested in a stochastic process X which is not directly observable (called the signal or state process). Instead, one observes a related stochastic process Y, called the measurement, observation, or emission process. The goal of filtering is to estimate the state based on the observations, i.e. to calculate conditional expectations

$$p_t\left[\varphi\right] \doteq \mathbb{E}\left[\varphi(X_t)|\mathcal{Y}_t\right], \tag{A.26}$$

where $\mathcal{Y}_t = \sigma(\{Y_s, 0 \leq s \leq t\})$ is the natural filtration of the observation process. The assumptions for the processes X, Y may be that both are of diffusion type, such as in the following problem

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t, \tag{A.27}$$

$$dY_t = c(X_t, t)dt + dV_t, \tag{A.28}$$

where W, V are independent Wiener processes. This filtering problem is the classical one studied by Kushner (1962). In this thesis, we are dealing with counting process observations, i.e. $Y = N$, where

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t, \tag{A.29}$$

$$dN_t = g(X_t, t)dt + dM_t, \tag{A.30}$$

with M being a martingale independent of W.

Filtering theory addresses this problem by deriving an SDE for the posterior (conditional) measure p_t, called filtering equation. There are two main approaches to achieve this, 1) the innovations approach (Fujisaki et al., 1972), and 2) the reference measure approach (Zakai, 1969). The first approach uses the innovation process and martingale theory to find the equation for p_t. The second approach uses a change of probability measure to make the observations process trivial, and then simplifying the problem by first deriving an equation for an unnormalized measure, from which p_t can be recovered using the Kallianpur-Striebel formula. In Section 3.1.2 of this thesis, we used the reference measure approach in order to rederive a filtering equation for the filtering problem with point process observations.

A.2 Discrete and Continuous Autoregressive Processes

In section 3.3.1, we introduced the continuous-time autoregressive (CAR) process of order p, which is the solution to an order-p SDE, written formally as

$$U_t^{(p)} + a_{p-1}U_t^{(p-1)} + ... + a_0 U_t = b\zeta_t, \tag{A.31}$$

where $\zeta(t)$ is a white-noise process, i.e.

$$\mathbb{E}\left[\zeta_t\right] = 0, \qquad \mathbb{E}\left[\zeta_t \zeta_{t+\tau}\right] = \delta(\tau). \tag{A.32}$$

On the other hand, the discrete-time autoregressive process of order p is a countable collection of random variables satisfying the recursive law

$$V_n = \sum_{i=1}^{p} \alpha_i V_{n-i} + \varphi_n, \quad n \in \mathbb{Z} \tag{A.33}$$

where $\varphi_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma_\varphi^2)$. The goal of this section is to elucidate the relation between AR and CAR processes of matching orders. In particular, we want to give conditions under which the continuum limit of an AR process yields a CAR process and how they are related, i.e. how the parameters $\alpha_i, \sigma_\varphi^2$ of the AR process translate to parameters a_i, b of the CAR process and vice versa.

An example: $p = 1$

The general theme of the continuum limit or any time-scale transformation is that the parameters of the discrete-time models become functions of the time-discretization step Δt with certain asymptotics. In the case of an AR(1) process,

$$V_t = \alpha_1 V_{t-1} + \varphi_t, \tag{A.34}$$

it is required that $\alpha_1 = 1 - a_0 \Delta t + \mathcal{O}\left(\Delta t^2\right)$ and $\sigma_\varphi^2 = b^2 \Delta t + \mathcal{O}\left(\Delta t^2\right)$ in order for the continuum limit to be a non-trivial process. Under the condition that these asymptotics hold, the limit is a CAR(1) or OU process with parameters a_0, b^2. In the case that the higher-order terms vanish, the AR(1) corresponds *exactly* to the Euler-Maruyama discretization, but there is an infinite class of other discretizations which all converge to the OU process. It is not the case however that any AR(1) process – with arbitrary functions $\alpha_1(\Delta t)$ and $\sigma_\varphi^2(\Delta t)$ – is a discretization of the CAR(1) process.

While the condition that the limit be non-trivial can only give the linear terms in of the Δt dependence, the assumption of time-scale invariance can give terms of arbitrary order. The argument is as follows. The AR(1) recursion law can be used to write a law spanning two timesteps, i.e.

$$V_{t+1} = \alpha_1 V_t + \varphi_{t+1}, \tag{A.35}$$

$$\begin{aligned} V_{t+2} &= \alpha_1 V_{t+1} + \varphi_{t+2} \\ &= \alpha_1^2 V_t + \alpha_1 \varphi_{t+1} + \varphi_{t+2}. \end{aligned} \tag{A.36}$$

If we expand the functions α_1 and σ_φ^2 to second order in Δt,

$$\alpha_1(\Delta t) = A_0 + A_1 \Delta t + \frac{1}{2} A_2 \Delta t^2 + \mathcal{O}\left(\Delta t^3\right), \tag{A.37}$$

$$\sigma_\varphi^2(\Delta t) = B_0 + B_1 \Delta t + \frac{1}{2} B_2 \Delta t^2 + \mathcal{O}\left(\Delta t^3\right) \tag{A.38}$$

and demand that $V_{t+2} = \alpha_1(2\Delta t) + \varphi'_{t+2}$ with $\mathrm{Var}(\varphi'_{t+2}) = \sigma^2_\varphi(2\Delta t)$, we obtain conditions on the coefficients which read as follows

$$\begin{aligned} A_0^2 &= A_0, && \text{(A.39)}\\ A_0A_1 &= A_1, && \text{(A.40)}\\ A_1^2 + A_0A_2 &= 2A_2, && \text{(A.41)}\\ A_0^2B_0 &= 0, && \text{(A.42)}\\ (1 + A_0^2)B_1 + 2A_0A_1B_0 &= 2B_1, && \text{(A.43)}\\ 2A_0A_1B_1 + \frac{1}{2}(A_0^2 + 1)B_2 + (A_1^2 + A_0A_2)B_0 &= 2B_2. && \text{(A.44)} \end{aligned}$$

This system of nonlinear equations has two classes of solutions. The first reads

$$\begin{aligned} \alpha_1(\Delta t) &= 1 + A_1\Delta t + \frac{1}{2}A_1^2\Delta t^2 + \mathcal{O}\left(\Delta t^3\right), && \text{(A.45)}\\ \sigma^2_\varphi(\Delta t) &= B_1\Delta t + A_1B_1\Delta t^2 + \mathcal{O}\left(\Delta t^3\right), && \text{(A.46)} \end{aligned}$$

for some arbitrary nonzero A_1, B_1. The second solution reads,

$$\begin{aligned} \alpha_1(\Delta t) &= A_1\Delta t + \frac{1}{4}A_1^2\Delta t^2 + \mathcal{O}\left(\Delta t^3\right), && \text{(A.47)}\\ \sigma^2_\varphi(\Delta t) &= B_0 + \frac{1}{3}A_1^2B_0\Delta t^2 + \mathcal{O}\left(\Delta t^3\right), && \text{(A.48)} \end{aligned}$$

for some A_1, B_0. Both solutions are time-scale invariant in discrete-time (to second order in Δt), but whereas the first solution yields the OU process in the continuum limit, the continuum limit of the second solution is ill-defined unless $B_0 = 0$. We thus see that time-scale invariance even to second order is an insufficient condition for the correct asymptotics. It is however the case that a set of coefficients with the correct asymptotics yields time-scale invariance to first order in Δt.

Another example: $p = 2$

Now we would like to make a similar argument to find the asymptotics for $p = 2$. The AR(2) process

$$V_t = \alpha_1V_{t-1} + \alpha_2V_{t-2} + \varphi_t, \qquad \text{(A.49)}$$

can be put in a vector form

$$\begin{pmatrix} V_t \\ V_{t-1} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} V_{t-1} \\ V_{t-2} \end{pmatrix} + \begin{pmatrix} \varphi_t \\ 0 \end{pmatrix}. \qquad \text{(A.50)}$$

We now define a finite-difference operator T_2 and a finite-difference state vector $(V_t\ \dot{V}_t)^\top$:

$$T_2 = \begin{pmatrix} 1 & 0 \\ \Delta t^{-1} & -\Delta t^{-1} \end{pmatrix}, \quad \begin{pmatrix} V_t \\ \dot{V}_t \end{pmatrix} = T_2\begin{pmatrix} V_t \\ V_{t-1} \end{pmatrix}, \qquad \text{(A.51)}$$

whereupon we obtain

$$\begin{aligned} \Delta t^{-1}\begin{pmatrix} V_t - V_{t-1} \\ \dot{V}_t - \dot{V}_{t-1} \end{pmatrix} &= \Delta t^{-1}\left[T_2\begin{pmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{pmatrix}T_2^{-1} - \mathbb{1}\right]\begin{pmatrix} V_{t-1} \\ \dot{V}_{t-1} \end{pmatrix} + \Delta t^{-1}T_2\begin{pmatrix} \varphi_t \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \Delta t^{-1}(\alpha_1 + \alpha_2 - 1) & -\alpha_2 \\ \Delta t^{-2}(\alpha_1 + \alpha_2 - 1) & -\Delta t^{-1}(1 + \alpha_2) \end{pmatrix}\begin{pmatrix} V_{t-1} \\ \dot{V}_{t-1} \end{pmatrix} + \begin{pmatrix} \Delta t^{-1}\varphi_t \\ \Delta t^{-2}\varphi_t \end{pmatrix}. \end{aligned} \qquad \text{(A.52)}$$

In the continuum limit the LHS reads $(\dot{V}_t\ \ddot{V}_t)^\top$, so in accordance with a CAR(2) process, we should have

$$\begin{pmatrix} \Delta t^{-1}(\alpha_1+\alpha_2-1) & -\alpha_2 \\ \Delta t^{-2}(\alpha_1+\alpha_2-1) & -\Delta t^{-1}(1+\alpha_2) \end{pmatrix} \overset{\Delta t\to 0}{\longrightarrow} \begin{pmatrix} 0 & 1 \\ -a_0 & -a_1 \end{pmatrix}, \tag{A.53}$$

as well as

$$\Delta t \operatorname{Var}(\Delta t^{-2}\varphi_t) \overset{\Delta t\to 0}{\longrightarrow} \sigma^2. \tag{A.54}$$

From (A.54) we deduce without further ado that

$$\sigma_\varphi^2 = \sigma^2\Delta t^3 + \mathcal{O}\left(\Delta t^4\right), \tag{A.55}$$

and we also see, by looking at the lower row of (A.53) that

$$\alpha_1+\alpha_2-1 = -a_0\Delta t^2 + \mathcal{O}\left(\Delta t^3\right), \tag{A.56}$$

$$1+\alpha_2 = a_1\Delta t + \mathcal{O}\left(\Delta t^2\right), \tag{A.57}$$

from which we can conclude that

$$\begin{aligned} \alpha_1 &= 2 - a_1\Delta t - \xi a_0\Delta t^2 + \mathcal{O}\left(\Delta t^3\right), \\ \alpha_2 &= -1 + a_1\Delta t - (1-\xi)a_0\Delta t^2 + \mathcal{O}\left(\Delta t^3\right), \end{aligned} \tag{A.58}$$

for some $\xi \in \mathbb{R}$. Thus for any CAR(2), there is a one-parameter family of AR(2) processes which converge to that CAR(2) process in the continuum. The higher-order terms can be fixed by time-scale invariance. Indeed, a calculation analogous to what we did in the previous section for$p = 1$ shows that time-scale invariance fixes

$$\xi = 1 - \frac{a_1^2}{2a_0}. \tag{A.59}$$

Arbitrary p

For a general p, we can use the same framework in order to obtain a mapping from CAR(p) to AR(p) processes. The parameters of the discrete-time process read, as a function of the parameters of the continuous-time counter-part[2]

$$\sigma_\varphi^2 = \sigma^2\Delta t^{2p-1} + \mathcal{O}\left(\Delta t^{2p}\right), \tag{A.60}$$

$$\alpha_i = (-1)^{i-1}\left[\binom{p}{i} - \sum_{j=1}^{p}\binom{p-j}{i-1}a_{p-j}\Delta t^j\right] + \mathcal{O}\left(\Delta t^{p+1}\right), \tag{A.61}$$

which for $p = 2$ yields the solution (A.58) with $\alpha = 1$. For $p > 2$ the symmetry which gives multiple solutions seems to be absent.

How big is the class of AR(p) processes which converge to one given CAR(p) process? To answer this question, we need to restrict the infinite-dimensional space of functions $\mathbb{R}_{\geq 0} \to \mathbb{R}^p$ describing how the parameters of the AR(p) evolve as a function of Δt (we can restrict ourselves to the class of AR(p) processes where the variance is a monomial in Δt, such that it has no free parameters due to (A.60)). If we further restrict the parameters α_i to be polynomials of order $q \geq p$ in Δt, we have $p(q+1)$ parameters to fix. Out of those, $p(p+1)$ are given by (A.61). The remaining $p(q-p)$ parameters define the manifold of solutions.

---

[2] The convention is that $\binom{i}{j} = 0$ if $j > i$.

A.3 Proof of the Normalization of the Adaptive Point Process

We want to prove that the expression for the probability of a sequence of events of the adaptive point process reads

$$\begin{aligned} P\left(N_{[0,t]} = \{t_1, ..., t_n\}\,|X_{[0,t]}\right) = \exp\left[-\int_0^t g(X_s, \{t_i \leq s\}, s)ds\right] \\ \times \prod_{0\leq t_n \leq t} g(X_{t_n}, \{t_i \leq t_n\}, t_n). \end{aligned} \tag{A.62}$$

This is equivalent to asking whether the expression is normalized. We define the probability $p(n,t)$ of having n spikes within the interval $[0,t]$ as

$$\begin{aligned} &p(0,t) = P\left(N_{[0,t]} = \{\}\,|X_{[0,t]}\right), \\ &p(n,t) = \int_0^t dt_n \int_0^{t_n} dt_{n-1} ... \int_0^{t_2} dt_1\, P\left(N_{[0,t]} = \{t_1, ..., t_n\}\,|X_{[0,t]}\right), \quad n \geq 1. \end{aligned} \tag{A.63}$$

We want to prove that

$$\mathcal{N}(t) = \sum_{n=0}^{\infty} p(n,t) = 1, \quad T \geq 0 \tag{A.64}$$

We have

$$p(0,0) = 1, \quad p(n,0) = 0,\ n \geq 1 \quad \Rightarrow \quad \mathcal{N}(0) = 1 \tag{A.65}$$

and thus the normalization is correct for $t = 0$. To complete the proof, we show that the derivative of $\mathcal{N}$ vanishes. The derivative of the first term in the sum reads

$$\begin{aligned} \frac{\partial}{\partial t} p(0,t) &= \frac{\partial}{\partial t} \exp\left[-\int_0^t g(X_s, \{\}, s)ds\right] \\ &= -g(X_t, \{\}, t) \exp\left[-\int_0^t g(X_s, \{\}, s)ds\right] \end{aligned} \tag{A.66}$$

For $n \geq 1$, we use the following property

$$g(t) = \int_0^t f(t,t')dt' \quad \Rightarrow \quad g'(t) = f(t,t) + \int_0^t \frac{\partial f(t,t')}{\partial t} dt' \tag{A.67}$$

and compute

$$\begin{aligned} \frac{\partial}{\partial t} p(n,t) = &\int_0^t dt_{n-1} ... \int_0^{t_2} dt_1\, g(X_t, \{t_1, ..., t_{n-1}\}, t) \\ &\times P\left(N_{[0,t]} = \{t_1, ..., t_{n-1}\}\,|X_{[0,t]}\right) \\ &- \int_0^t dt_n ... \int_0^{t_2} dt_1\, g(X_t, \{t_1, ..., t_{n-1}, t_n\}, t) \\ &\times P\left(N_{[0,t]} = \{t_1, ..., t_n\}\,|X_{[0,t]}\right), \quad n \geq 1. \end{aligned} \tag{A.68}$$

We define a quantity $R_n(t)$ as

$$\begin{aligned} R_n(t) = &\int_0^t dt_n ... \int_0^{t_2} dt_1\, g(X_t, \{t_1, ..., t_{n-1}, t_n\}, t) \\ &\times P\left(N_{[0,t]} = \{t_1, ..., t_n\}\,|X_{[0,t]}\right) \end{aligned} \tag{A.69}$$

which allows us to write

$$\frac{\partial}{\partial t}p(n,t) = R_{n-1}(t) - R_n(t), \quad n \geq 1, \qquad \frac{\partial}{\partial t}p(0,t) = -R_0(t) \tag{A.70}$$

Under the assumption that $R(t) = \sum_{n=0}^{\infty} R_n(t)$ (the expectation of the firing rate at time t, given all possible spike histories with n spikes) is finite, we can use re-ordering of the series to prove that the derivative of the normalization vanishes:

$$\begin{aligned}\frac{\partial}{\partial t}\mathcal{N}(t) &= \frac{\partial}{\partial t}\sum_{n=0}^{\infty} p(n,t) \\ &= \sum_{n=0}^{\infty} \frac{\partial}{\partial t} p(n,t) \\ &= -R_0(t) + \sum_{n=1}^{\infty} (R_{n-1}(T) - R_n(t)) = 0.\end{aligned} \tag{A.71}$$

Bibliography

L. F. Abbott, J. A. Varela, K. Sen, and S. B. Nelson. Synaptic depression and cortical gain control. *Science*, 275(5297):220–224, Jan. 1997.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

F. R. Bach and M. I. Jordan. Learning graphical models for stationary time series. *IEEE Transactions on Signal Processing*, 52(8):2189–2199, 2004.

A. Bain and D. Crişan. *Fundamentals of stochastic filtering*, volume 60 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2009.

O. Barak and M. Tsodyks. Persistent Activity in Neural Networks with Dynamic Synapses. *PLoS Computational Biology*, 3(2):e35, 2007.

A. Baranyi, M. B. Szente, and C. D. Woody. Electrophysiological characterization of different types of neurons recorded in vivo in the motor cortex of the cat. II. Membrane parameters, action potentials, current-induced voltage responses and electrotonic structures. *Journal of Neurophysiology*, 69(6):1865–1879, June 1993.

P. Berkes, G. Orbán, M. Lengyel, and J. Fiser. Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science*, 331(6013):83–87, Jan. 2011.

A. V. Blackman, T. Abrahamsson, R. P. Costa, T. Lalanne, and P. J. Sjöström. Target-cell-specific short-term plasticity in local circuits. *Frontiers in Synaptic Neuroscience*, 5:11, 2013.

O. Bobrowski, R. Meir, and Y. C. Eldar. Bayesian filtering in spiking neural networks: noise, adaptation, and multisensory integration. *Neural Computation*, 21(5):1277–1320, May 2009.

R. Brette and W. Gerstner. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology*, 94(5):3637–3642, Nov. 2005.

D. Brigo, B. Hanzon, and F. LeGland. A differential geometric approach to nonlinear filtering: the projection filter. *IEEE Transactions on Automatic Control*, 43(2):247–252, 1998.

D. Brigo, B. Hanzon, and F. Le Gland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, 5(3):495–534, 1999.

G. Buzsáki. Theta oscillations in the hippocampus. *Neuron*, 33(3):325–340, 2002.

C. Ceci and K. Colaneri. Nonlinear filtering for jump diffusion observations. *Advances in Applied Probability*, 44(3):678–701, 2012.

F. Chen and P. Hall. Nonparametric Estimation for Self-Exciting Point Processes—A Parsimonious Approach. *Journal of Computational and Graphical Statistics*, Feb. 2015.

E. J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213, May 2001.

R. P. Costa, P. J. Sjöström, and M. van Rossum. Probabilistic Inference of Short-Term Synaptic Plasticity in Neocortical Microcircuits. *Frontiers in Computational Neuroscience*, June 2013.

D. R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–164, 1955.

J. P. Cunningham, M. Y. Byron, K. V. Shenoy, and M. Sahani. Inferring Neural Firing Rates from Spike Trains Using Gaussian Processes. In *Advances in neural information processing systems*, pages 329–336, 2007.

L. E. Dobrunz, E. P. Huang, and C. F. Stevens. Very short-term plasticity in hippocampal synapses. *Proceedings of the National Academy of Sciences*, 94(26):14843–14847, Dec. 1997.

S. Druckmann, Y. Banitt, A. Gidon, F. Schürmann, H. Markram, and I. Segev. A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data. *Frontiers in Neuroscience*, 1(1):7–18, Nov. 2007.

U. T. Eden. Point process adaptive filters for neural data analysis: Theory and applications. In *Decision and Control, 2007 46th IEEE Conference on*, pages 5818–5825, 2007.

U. T. Eden and E. N. Brown. Continuous-time filters for state estimation from point process models of neural data. *Statistica Sinica*, 18(4):1293–1310, 2008.

B. Efron and D. V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, 65(3):457–487, 1978.

S. El Boustani, O. Marre, S. Béhuret, P. Baudot, P. Yger, T. Bal, A. Destexhe, and Y. Frégnac. Network-State Modulation of Power-Law Frequency-Scaling in Visual Cortical Neurons. *PLoS Computational Biology*, 5(9):e1000519, Sept. 2009.

N. Fourcaud-Trocmé, D. Hansel, C. van Vreeswijk, and N. Brunel. How spike generation mechanisms determine the neuronal response to fluctuating inputs. *The Journal of Neuroscience*, 23(37):11628–11640, Dec. 2003.

R. Frey, T. Schmidt, and L. Xu. On Galerkin Approximations for the Zakai Equation with Diffusive and Point Process Observations. *SIAM Journal on Numerical Analysis*, 51(4): 2036–2062, Jan. 2013.

M. Fujisaki, G. Kallianpur, and H. Kunita. Stochastic differential equations for the non linear filtering problem. *Osaka Journal of Mathematics*, 9:19–40, 1972.

C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry, and the natural sciences*. Springer Verlag, 1985.

L. Gerencsér, C. Matias, Z. Vágó, and B. Torma. Self-exciting point processes with applications in finance and medicine. *18th International symposium on Mathematical Theory of Networks and Systems*, pages 1–10, 2008.

W. Gerstner and W. M. Kistler. *Spiking Neuron Models*. Single Neurons, Populations, Plasticity. Cambridge University Press, Aug. 2002.

W. Gerstner and R. Naud. How Good Are Neuron Models? *Science*, 326(5951):379–380, Oct. 2009.

I. Gertner. An alternative approach to nonlinear filtering. *Stochastic Processes and their Applications*, 7(3):231–246, 1978.

M. S. Goldman, P. Maldonado, and L. F. Abbott. Redundancy reduction and sustained firing with stochastic depressing synapses. *The Journal of Neuroscience*, 22(2):584–591, Jan. 2002.

R. M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

D. Guo, S. Shamai, and S. Verdu. Mutual Information and Conditional Mean Estimation in Poisson Channels. *IEEE Transactions on Information Theory*, 54(5):1837–1849, 2008.

B. Haider, M. Häusser, and M. Carandini. Inhibition dominates sensory responses in the awake cortex. *Nature*, 493(7430):97–100, Jan. 2013.

K. Hamaguchi, K. A. Tschida, I. Yoon, B. R. Donald, and R. Mooney. Auditory synapses to song premotor neurons are gated off during vocalization in zebra finches. *eLife*, 3: e01833, 2014.

B. Hanzon and R. Hut. New results on the projection filter. *Research Memorandum*, 1991: 23, 1991.

A. G. Hawkes. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1):83–90, Apr. 1971.

M. H. Hennig. Theoretical models of synaptic short term plasticity. *Frontiers in Computational Neuroscience*, 7:45, 2013.

B. Hille. *Ion Channels of Excitable Membranes*. Sinauer Associates Incorporated, Jan. 2001.

A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500, 1952.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, Jan. 1970.

R. Jolivet, A. Rauch, H.-R. Lüscher, and W. Gerstner. Predicting spike timing of neocortical pyramidal neurons by simple threshold models. *Journal of Computational Neuroscience*, 21(1):35–49, Apr. 2006.

E. Kandel, J. Schwartz, T. Jessell, S. Siegelbaum, and A. J. Hudspeth. *Principles of Neural Science, Fourth Edition*. McGraw-Hill Professional, Sept. 2012.

N. Kantas, A. Doucet, S. S. Singh, J. M. Maciejowski, and N. Chopin. On Particle Methods for Parameter Estimation in State-Space Models. *arXiv.org*, Dec. 2014.

Y. Katori, Y. Otsubo, M. Okada, and K. Aihara. Stability analysis of associative memory network composed of stochastic neurons and dynamic synapses. *Frontiers in Computational Neuroscience*, 7:6, 2013.

A. K. Katsaggelos and K. T. Lay. Maximum likelihood blur identification and image restoration using the EM algorithm. *IEEE Transactions on Signal Processing*, 39(3):729–733, 1991.

F. C. Klebaner. *Introduction to Stochastic Calculus with Applications*. Imperial College Press, Jan. 2005.

W. H. Kliemann, G. Koch, and F. Marchetti. On the unnormalized solution of the filtering problem with counting process observations. *IEEE Transactions on Information Theory*, 36(6):1415–1425, 1990.

H. J. Kushner. On the differential equations satisfied by conditional probablitity densities of Markov processes, with applications. *Journal of the Society for Industrial & Applied Mathematics, Series A: Control*, 2(1):106–119, 1962.

H. J. Kushner. Approximations to optimal nonlinear filters. *IEEE Transactions on Automatic Control*, 12(5):546–556, 1967.

L. Lapicque. Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation. *Journal de Physiologie et de Pathologie Générale*, 9(1):620–635, 1907.

P. E. Latham, B. J. Richmond, P. G. Nelson, and S. Nirenberg. Intrinsic dynamics in neuronal networks. I. Theory. *Journal of Neurophysiology*, 83(2):808–827, Feb. 2000.

A. K. Lee, I. D. Manns, B. Sakmann, and M. Brecht. Whole-Cell Recordings in Freely Moving Rats. *Neuron*, 51(4):399–407, Aug. 2006.

B. Lindner, L. Schimansky-Geier, and A. Longtin. Maximizing spike train coherence or incoherence in the leaky integrate-and-fire model. *Physical Review E*, 66(3):031916, Sept. 2002.

B. Lindner, D. Gangloff, A. Longtin, and J. E. Lewis. Broadband Coding with Dynamic Synapses. *The Journal of Neuroscience*, 29(7):2076–2087, Feb. 2009.

M. A. Long, D. Z. Jin, and M. S. Fee. Support for a synaptic chain model of neuronal sequence generation. *Nature*, 468(7322):394–399, Nov. 2010.

J. H. Macke, L. Buesing, and J. P. Cunningham. Empirical models of spiking in neural populations. *Advances in Neural Information Processing Systems*, 24:1350–1358, 2011.

H. Markram and M. Tsodyks. Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature*, 382(6594):807–810, 1996.

H. Markram, Y. Wang, and M. Tsodyks. Differential signaling via the same axon of neocortical pyramidal neurons. *Proceedings of the National Academy of Sciences*, 95(9):5323–5328, Apr. 1998.

H. Markram, M. Toledo-Rodriguez, Y. Wang, A. Gupta, G. Silberberg, and C. Wu. Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience*, 5(10):793–807, Oct. 2004.

M. Matsumura, T. Cope, and E. E. Fetz. Sustained excitatory synaptic input to motor cortex neurons in awake animals revealed by intracellular recording of membrane potentials. *Experimental Brain Research*, 70(3):463–469, 1988.

S. Mensi, R. Naud, C. Pozzorini, M. Avermann, C. C. H. Petersen, and W. Gerstner. Parameter extraction and classification of three cortical neuron types reveals two distinct adaptation mechanisms. *Journal of Neurophysiology*, 107(6):1756–1775, Mar. 2012.

J. Møller, A. Syversveen, and R. Waagepetersen. Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, Aug. 1998.

G. Mongillo, O. Barak, and M. Tsodyks. Synaptic Theory of Working Memory. *Science*, 319(5869):1543–1546, Mar. 2008.

L. Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, Nov. 2004.

L. Paninski, Y. Ahmadian, D. G. Ferreira, S. Koyama, K. Rahnama Rad, M. Vidne, J. Vogelstein, and W. Wu. A new look at state-space models for neural data. *Journal of Computational Neuroscience*, 29(1-2):107–126, Aug. 2009.

J.-P. Pfister, P. Dayan, and M. Lengyel. Know thy neighbour: A normative theory of synaptic depression. *Advances in Neural Information Processing Systems*, 22:1464–1472, 2009.

J.-P. Pfister, P. Dayan, and M. Lengyel. Synapses with short-term plasticity are optimal estimators of presynaptic membrane potentials. *Nature Neuroscience*, 13(10):1271–1275, 2010.

J. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, July 2008.

T. Plienpanich. *On some problems of stochastic filtering applied to finance*. PhD thesis, Suranaree University of Technology, 2007.

J. F. A. Poulet and C. C. H. Petersen. Internal brain state regulates membrane potential synchrony in barrel cortex of behaving mice. *Nature*, 454(7206):881–885, July 2008.

C. Pozzorini, R. Naud, S. Mensi, and W. Gerstner. Temporal whitening by power-law adaptation in neocortical neurons. *Nature Neuroscience*, 16(7):942–948, June 2013.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

A. Reyes, R. Lujan, A. Rozov, N. Burnashev, P. Somogyi, and B. Sakmann. Target-cell-specific facilitation and depression in neocortical circuits. *Nature Neuroscience*, 1(4):279–285, Aug. 1998.

R. Rosenbaum, J. Rubin, and B. Doiron. Short Term Synaptic Depression Imposes a Frequency Dependent Filter on Synaptic Information Transfer. *PLoS Computational Biology*, 8(6):e1002557, June 2012.

Z. Rotman, P. Y. Deng, and V. A. Klyachko. Short-Term Plasticity Optimizes Synaptic Information Transmission. *The Journal of Neuroscience*, 31(41):14800–14809, Oct. 2011.

P. Scott, A. I. Cowan, and C. Stricker. Quantifying impacts of short-term plasticity on neuronal information transfer. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 85(4 Pt 1):041921, Apr. 2012.

P. C. Scott. *Information transfer at dynamic synapses: effects of short-term plasticity*. PhD thesis, The Australian National University, 2005.

S. Shinomoto and Y. Tsubo. Modeling spiking behavior of neurons with time-dependent Poisson processes. *Physical Review E*, 64(4):041910, Sept. 2001.

D. L. Snyder. Filtering and detection for doubly stochastic Poisson processes. *IEEE Transactions on Information Theory*, 18(1):91–102, 1972.

W. R. Softky and C. Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *The Journal of Neuroscience*, 13(1):334–350, Jan. 1993.

R. B. Stein. The information capacity of nerve cells using a frequency code. *Biophysical journal*, 7(6):797–826, Nov. 1967.

M. Steriade, I. Timofeev, and F. Grenier. Natural waking and sleep states: a view from inside neocortical neurons. *Journal of Neurophysiology*, 85(5):1969–1985, 2001.

I. H. Stevenson, B. Cronin, M. Sur, and K. P. Kording. Sensory Adaptation and Short Term Plasticity as Bayesian Correction for a Changing Brain. *PLoS ONE*, 5(8):e12436, Aug. 2010.

S. C. Surace and J.-P. Pfister. A Statistical Model for In Vivo Neuronal Dynamics. *PLoS ONE*, 10(11):e0142435, Nov. 2015.

J. J. Torres and H. J. Kappen. Emerging phenomena in neural networks with dynamic synapses and their computational implications. *Frontiers in Computational Neuroscience*, 7:30, 2013.

W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93(2):1074–1089, Feb. 2005.

M. Tsodyks and S. Wu. Short-term synaptic plasticity. *Scholarpedia*, 8(10):3153, 2013.

M. Tsodyks, K. Pawelzik, and H. Markram. Neural networks with dynamic synapses. *Neural Computation*, 10(4):821–835, May 1998.

B. Ujfalussy and M. Lengyel. Active dendrites: adaptation to spike-based communication. *NIPS*, pages 1–7, Oct. 2011a.

B. Ujfalussy and M. Lengyel. Active dendrites: adaptation to spike-based communication. *Advances in Neural Information …*, 2011b.

D. Vallentin and M. A. Long. Motor Origin of Precise Synaptic Inputs onto Forebrain Neurons Driving a Skilled Behavior. *The Journal of Neuroscience*, 35(1):299–307, Jan. 2015.

M. Venugopal, R. M. Vasu, and D. Roy. An ensemble Kushner-Stratonovich-Poisson filter. *arXiv preprint arXiv:1407.2192*, 2014.

M. Vidne, Y. Ahmadian, J. Shlens, J. Pillow, J. E. Kulkarni, A. M. Litke, E. J. Chichilnisky, E. Simoncelli, and L. Paninski. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *Journal of Computational Neuroscience*, 33(1): 97–121, July 2012.

C. D. Woody and E. Gruen. Characterization of electrophysiological properties of intracellularly recorded neurons in the neocortex of awake cats: a comparison of the response to injected current in spike overshoot and undershoot neurons. *Brain Research*, 158(2): 343–357, Dec. 1978.

L. Xu. *On Galerkin approximations for the Zakai equation with diffusive and point process observations*. PhD thesis, Universität Leipzig, 2011.

M. Zakai. On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11(3):230–243, 1969.

R. S. Zucker and W. G. Regehr. Short-term synaptic plasticity. *Annual review of physiology*, 64:355–405, 2002.

List of Mathematical Symbols

$\doteq$	Is defined to be equal to
$\sim$	Is distributed according to
$\approx$	Is approximately equal to
$\mathbb{E}\left[X\right]$	Expectation of X
$\mathbb{E}\left[X_t \vert \mathcal{F}\right]$	Conditional expectation of X_t under the filtration $\mathcal{F}_t$
$\mathcal{F}_t$	Filtration
dX_t	Stochastic differential of the process X_t
∂_x	Partial differential operator with respect to the variable x
W_t	Value of the standard Wiener process at time $t \geq 0$
δ	Dirac delta distribution
H	Heaviside function
$\mathcal{L}$	Generator of a diffusion process
$\mathcal{L}^\dagger$	Fokker-Planck operator of a diffusion process, adjoint of $\mathcal{L}$

Curriculum vitae

Education

September 2011	MSc in Physics, University of Bern, Switzerland
September 2009	BSc in Physics, University of Bern, Switzerland

Teaching

September 2013	Substitute teacher in High School mathematics, Solothurn
April 2013	Substitute teacher in High School mathematics, Solothurn
2011-2014	Tutor at a private school, Interlink GmbH, Bern
2010-2011	Academic tutor, Institute for Theoretical Physics, University of Bern
May 2010	Substitute teacher in High School mathematics, Gymnasium Thun-Schadau
2008-2010	Substitute teacher in High School physics, Gymnasium Neufeld, Bern

Language skills

German	native speaker
Italian	native speaker
English	fluent
French	good knowledge

Computer skills

Operating systems	extensive knowledge of Windows, Mac OS, Linux
Programming	Python, Mathematica
Other	Adobe Photoshop

List of publications

Peer-reviewed journal papers
S. C. Surace and J.-P. Pfister. A Statistical Model for in vivo Neuronal Dynamics. *PLoS One*. To be announced.

Preprints
A. Kutschireiter, S. C. Surace, H. Sprekeler, J.-P. Pfister. A Neural Implementation for Nonlinear Filtering. *arXiv.org*. 2015.

Conference posters
Cosyne 2013, Salt Lake City, USA. *Adaptive Gaussian Poisson process: a model for in vivo neuronal dynamics.*
Cosyne 2014, Salt Lake City, USA. *Short-term facilitation as a normative consequence of presynaptic spike-rate adaptation.*
Cosyne 2015, Salt Lake City, USA. *A flexible and tractable statistical model for in vivo neuronal dynamics.*

Declaration of originality

Last name, first name: Surace, Simone Carlo
Matriculation number: 06-122-404

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations. All data, tables, figures and text citations which have been reproduced from any other source, including the internet, have been explicitly acknowledged as such. I am aware that in case of non-compliance, the Senate is entitled to withdraw the doctorate degree awarded to me on the basis of the present thesis, in accordance with the "Statut der Universität Bern (Universitätsstatut; UniSt)", Art.69, of 7 June 2011.

Bern, October 28, 2015

S. Surace

Colophon

This document was typeset using the XeTeX typesetting system created by the Non-Roman Script Initiative and the memoir class created by Peter Wilson. The body text is set 10pt with Adobe Caslon Pro. Other fonts include `Envy Code R` and Optima Regular.