Lecture Notes

Stochastic modelling of neuronal data

Adeline Leclercq Samson

Université Grenoble Alpes and ENSIMAG
Master of Science in Industrial and Applied Math (MSIAM)

Contents

Chapter 1

Introduction to neuronal systems

Our perception of the world is constructed out of the raw data sent to our brains by our sensory nerves. In each situation, these data come in the same standard form: as sequences of identical voltage pulses called action potentials or *spikes*. All the myriad tasks our brains perform in the processing of incoming sensory signals begin with these sequences of spikes. Spike sequences are the language of brain: when it comes time to act on the results of the stimulation, the brain sends out sequences of spikes to the motor neurons. The purpose of neurosciences is to provide a dictionary of that language of the brain, to understand the structure of this dictionary.

The neurons are the elementary processing units in the central nervous system, interconnected with intricate patterns. Neurons and there connections (long wire-like extensions) are packed into a dense network. Their estimated number in the human brain is around 10^{12}, and several kilometers of 'wires' per cubic millimeter. We focus on neurons, also various other cells are present in the cortex (like glia cells), and especially on spiking neurons (neurons which emit an output signal).

1.1 Main properties of a neuron

1.1.1 A spiking neuron

A typical neuron can be divided into three functionally distinct parts (see Figure 1.1), called dendrites, soma and axon:

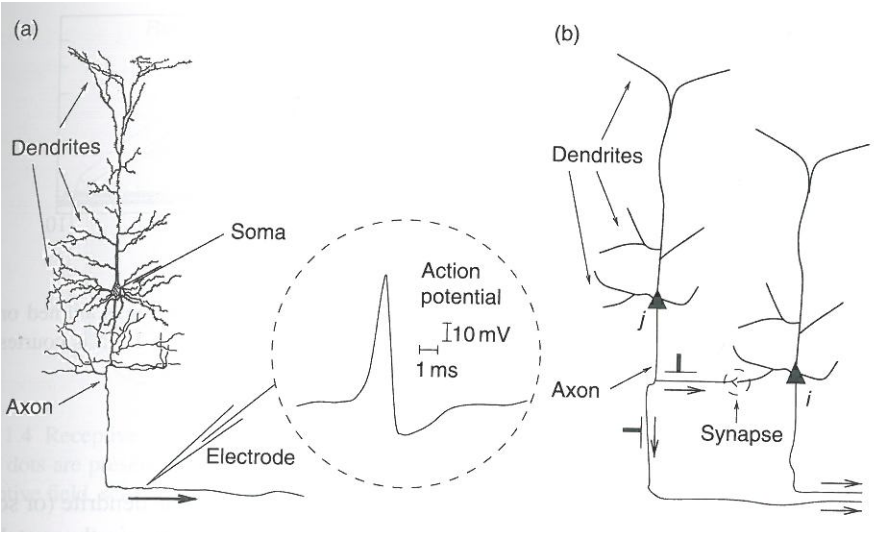

Figure 1.1: From Gerstner et al. (2014) (a) Single neuron with Dendrites, soma and axon. The inset shows an example of a neuronal action potential. (b) Signal transmission from a presynaptic neuron j to a postsynaptic neuron i.

- The **dendrites** play the role of the input device, that collects signals from other neurons and transmits them to the soma.
- The **soma** is the central processing unit, that performs an important non-linear processing step. If the total input exceeds a certain threshold, then an output signal is generated
- The **axon**, the output device, takes over the output signal and delivers it to other neurons.

The junction between two neurons is called a **synapse**. A neuron sends a signal across a **synapse**. The sending neuron is called the **presynaptic cell** and the receiving neuron is called the **postsynaptic cell**.

A single neuron in vertebrate cortex often connects to more than 10^4 postsynaptic neurons. Most of its axial branches end in the direct neighborhood of the neuron, but the axon can also stretch over several centimeters so as to reach to neurons in other areas of the brain.

1.1.2 Spike trains

Neurons are just as other cells enclosed by a membrane which separates the interior of the cell from the extracellular space. Inside the cell the

concentration of ions is different from that in the surrounding liquid. The difference in concentration is called the **membrane potential**. A huge membrane potential generates an electrical potential, a short electrical pulse, which plays an important role in neuronal dynamics. The pulses, so-called **action potentials or spikes**, have an amplitude of about 100 mV and typically a duration of 1-2 ms. The form of the pulse does not change as the action potential propagates along the axon.

A chain of action potentials emitted by a single neuron is called a **spike train**, a sequence of stereotyped events which occur at regular or irregular intervals. The action potential is the elementary unit of signal transmission.

Action potentials in a spike train are usually well separated. Even with very strong input, it is impossible to excite a second spike during or immediately after a first one. The minimal distance between two spikes defines the absolute **refractory period**. The absolute refractory period is followed by a phase of relative refractoriness where it is difficult, but not impossible to excite an action potential.

1.1.3 Synapse

The most common type of synapse in the vertebrate brain is a chemical synapse. At a chemical synapse, the axon terminal comes very close to the postsynaptic neuron, leaving only a tiny gap between pre- and postsynaptic cell membrane.

When an action potential arrives at a synapse, it triggers a complex chain of bio-chemical processing steps that lead to a release of neurotransmitter from the presynaptic terminal into the synapse. As soon as transmitter molecules have reached the postsynaptic side, they are detected by specialized receptors in the postsynaptic cell membrane and open specific **ion channels** so that ions from the extracellular fluid flow into the cell. The ion influx leads to a change of the membrane potential at the postsynaptic site so that the chemical signal is translated into an electrical response.

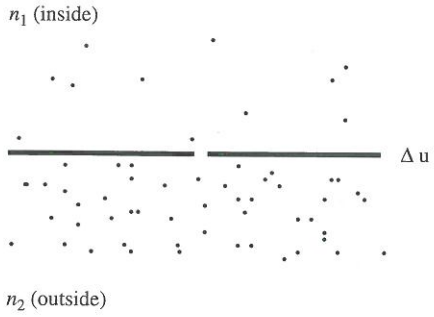

Figure 1.2: The solid line indicates the cell membrane, ions can pass though the gap. A difference in ion concentration generates an electrical potential. The concentration n_2 inside the neuron is different from the concentration n_1 of the surround. The resulting potential is the Nernst potential.

1.1.4 Equilibrium potential

The cell membrane is a nearly perfect electrical insulator. However, embedded in the membrane are specific proteins which act as ion gates. There are two types of **ion gates**: **ion pumps and ion channels**. Ions can pass through the ion channels. Ion pumps actively transport ions from one side to the other.

The concentration n_2 inside the neuron (in the intra-cellular liquid) is different from the concentration n_1 of the surround. The resulting potential is called the **Nernst potential** (see Figure 1.2).

For example, the sodium concentration inside the cell ($\equiv$ 60 mM/l) is lower than that in the extracellular liquid ($\equiv$ 440 mM/l). At equilibrium, this difference in concentration causes an Nernst potential $V_{Na} \equiv +50$ mV. That is, the interior of the cell has a positive potential with respect to the surround. If the voltage difference (membrane potential) is smaller that the equilibrium potential, more Na^+ ions flow into the cell (through ion channels) so as to decrease the concentration difference. Thus the direction of the current is reversed when the voltage passes V_{Na}. For this reason, V_{Na} is called the **reversal potential**.

The potassium concentration inside is higher ($\equiv$ 400 mM/l) that in the surround ($\equiv$ 20 mM/l). The reversal potential for K^+ ions is negative ($V_k \equiv -77$ mV).

Both sodium and potassium ions are present and contribute to the

voltage across the membrane. It is found experimentally that the resting potential of the membrane is about $\equiv -65$ mV. Since $V_K < -65 < V_{Na}$, potassium ions will flow out of the cell while sodium ions flow into the cell. The active ion pumps balance this flow and transport just as many ions back as pass through the channels.

1.1.5 Dynamics of a neuron

A difference in the voltage (the membrane potential) generates a difference in ion density. Similarly, a difference in ion concentration generates an electrical potential.

Without any spike input, the neuron is at rest corresponding to a constant membrane potential (the difference in concentration between the interior and the exterior of the cell). After the arrival of a spike, the potential changes and finally decays back to the resting potential. If the change is positive, the synapse is said to be **excitatory**. If the change is negative, the synapse is **inhibitory**.

At rest, the cell membrane has already a strong negative polarization of about -65mV (the resting potential). An input at an excitatory synapse reduces the negative polarization of the membrane and is called **depolarizing**. An input that increases the negative polarization of the membrane is called **hyperpolarizing**.

1.2 Experimental data

Our understanding of how the sensory world is represented in the electrical activity of the sensory nerves is limited, first and foremost, by our ability to record this activity. Indeed, the history of experiments on the electrical activity of nerves is intertwined with the history of electrical measurements more generally. What can be measured depends on whether we want to observe a single neuron or several neurons.

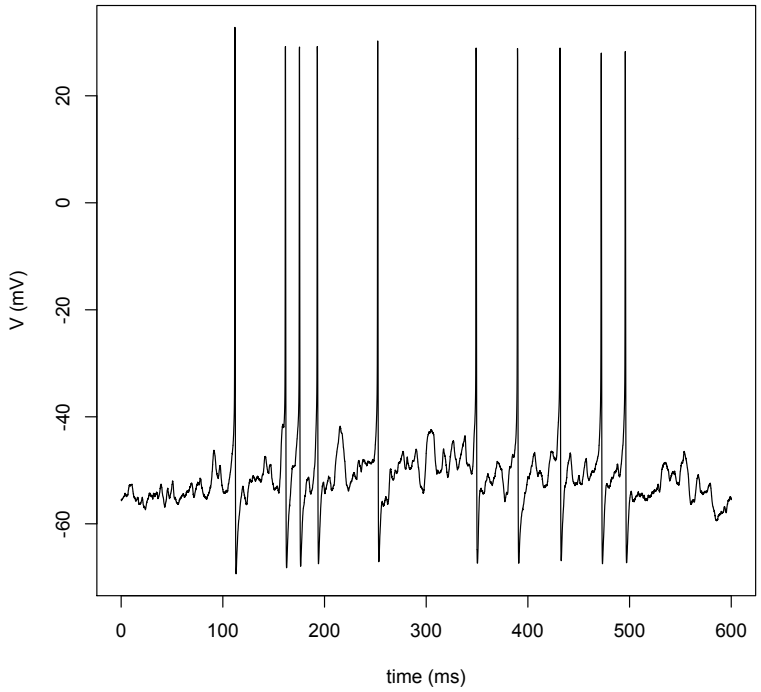

Figure 1.3: Intra-cellular neuronal data: membrane potential function of time

1.2.1 Single neuron observation

An electrode can be placed in the soma of a single neuron. Then the voltage at this electrode is measured relative to that at a reference electrode placed in the body fluids. What is measured at discrete times is the **membrane potential**, that is the difference in voltage between the interior and the exterior of the cell. We call these measurements **intra-cellular recordings**. An example of intra-cellular recordings is plotted in Figure 1.3.

Although the trace is noisy, there are clear, stereotyped event, which are the *action potentials* or *spikes* produced by this neuron and seen from outside the cell. The activity observed below the threshold that produces a spike is called the *sub-threshold activity*.

1.2.2 Several neurons observation

Several neurons can be observed by **extra-cellular recordings**, with 16 electrodes placed in the neuronal network. They measure simultaneously

the activity of several neurons. Only spike trains can be measured (after filtering), the sub-threshold activity can not be measured.

To exploit extra-cellular recordings, we must first

- Find out how many neurons are recorded
- For each neuron estimate some features like the spike waveform, the discharge statistics, etc
- For each detected event find the probability with which each neuron could have generated it.
- Find an automatic method to answer these questions.

This is what we call the **spike sorting**.

Efficient spike sorting requires:

1. Events detection followed by events space dimension reduction.
2. A clustering stage. This can be partially or fully automatized depending on the data.
3. Events classification.

Then the main issue is to analyze spike trains.

1.3 Fundamental issues

1. characterizing the sub-threshold activity and understand how is produced an action potential
2. describing the evolution of a neuron's firing rate across time,
3. assessing time-varying correlation between 2 neurons.

In this lecture, we focus on the first issue.

Chapter 2

Modelling intra-cellular recordings with deterministic models

From a biophysical point of view, action potentials are the result of currents that pass through ion channels in the cell membrane. In an extensive series of experiments, Hodgkin and Huxley succeeded to measure these currents and to describe their dynamics in terms of differential equations. The HH equations are the starting point for detailed neuron models which account for numerous ion channels, different types of synapse.

2.1 Conductance based models

2.1.1 Hodgkin-Huxley model

Hodgkin and Huxley (1952a, 1952b, 1952c) analyzed the electrical dynamics of the cell membrane in the giant axon of squid, and showed that these dynamics could be described with relatively simple phenomenological models of conductances that depend on voltage and are selective for different ions. When these local, active elements are assembled into a long cable, such as the axon, the nonlinear dynamics of the conductances select a stereotyped pulse which can propagate at constant velocity, while all other voltage changes eventually decay.

More precisely, Specific voltage-dependent ion channels, one for sodium, and another one for potassium, control the flow of those ions through the cell membrane. The leak current takes care of other channel types

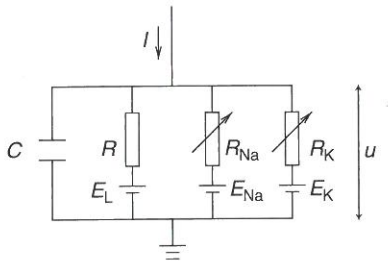

Figure 2.1: Schematic diagram for the Hodgkin-Huxley model

which are not described explicitly.

The membrane separates the interior of the cell from the extracelullar liquid and acts as a capacitor (see Figure 2.1). The conservation of electric charge on a piece of membrane implies that the applied current $I(t)$ may be split in a capacitive current I_C which charges the capacitor C and further components I_k which pass the ion channels.

Thus

$$I(t) = I_C(t) + \sum_k I_k(t)$$

where the sum runs over all ion channels. There are only three different types of ion currents involved in HH model: a sodium channel Na, a potassium channel K and an unspecific leakage channel with resistance R (that consists mainly of Cl^- ions). From the definition of a capacity, $C = Q(t)/V(t)$ where $Q(t)$ is a charge and $V(t)$ the voltage across the capacitor. Thus

$$I_C = CdV(t)/dt.$$

Hence

$$C\frac{dV(t)}{dt} = -\sum_k I_k(t) + I(t)$$

The leakage channel is described by a voltage-independent conductance $g_L = \frac{1}{R}$.

The conductance of the other ion channels is voltage and time dependent. If all channels are open, they transmit currents with a maximal conductance g_{Na} or g_K, respectively. But some of the channels are blocked. The probability that a channel is open is described by addi-

tional functions m, n, and h. The combined action of m and h controls the Na^+ channels. The K^+ gates are controlled by n. Specifically, we have

$$\sum_k I_k(t) = g_{Na} m^3(t) h(t)(V(t) - V_{Na}) + g_K n^4(t)(V(t) - V_K) + g_L(V(t) - V_L)$$

The parameters V_{Na}, V_K, V_L are the resting/reversal potentials. The three functions m, n, and h are called the gating variables. They evolve according to the differential equations

$$\begin{aligned}
\frac{dm(t)}{dt} &= \alpha_m(V(t))\,(1 - m(t)) - \beta_m(V(t))\,m(t) \\
\frac{dn(t)}{dt} &= \alpha_n(V(t))\,(1 - n(t)) - \beta_n(V(t))\,n(t) \\
\frac{dh(t)}{dt} &= \alpha_h(V(t))\,(1 - h(t)) - \beta_h(V(t))\,h(t)
\end{aligned}$$

The functions $\alpha(v)$ and $\beta(v)$ are non-linear functions that have been adjusted by HH for the giant axon of the squid.

One can rewrite the three equations in the form, for $x = m, n$ or h:

$$\frac{dx(t)}{dt} = -\frac{1}{\tau_x(V(t))}(x(t) - x_0(V(t)))$$

where $\tau_x(v)$ is the time constant, and $x_0(v)$ is the value reached by the process x for a fixed value v: $x_0(v) = \alpha_x(v)/(\alpha_x(v) + \beta_x(v))$ and $\tau_x(v) = 1/(\alpha_x(v) + \beta_x(v))$.

Final model

$$\begin{aligned}
C\frac{dV(t)}{dt} &= -\big(g_{Na} m^3(t) h(t)(V(t) - V_{Na}) + g_K n^4(t)(V(t) - V_K) + g_L(V(t) - V_L) + \\
\frac{dm(t)}{dt} &= -\frac{1}{\tau_m(V(t))}(m(t) - m_0(V(t))) \\
\frac{dn(t)}{dt} &= -\frac{1}{\tau_n(V(t))}(n(t) - n_0(V(t))) \\
\frac{dh(t)}{dt} &= -\frac{1}{\tau_h(V(t))}(h(t) - h_0(V(t)))
\end{aligned}$$

Dynamics of HH model

- If some external input (for example short current pulse of 1ms duration) causes the membrane voltage to rise, the conductance of sodium channels ($m(t)$) increases. Positive sodium ions flow into the cell and raise the membrane potential further. If this positive feedback is large enough, a spike is generated. High values of $V(t)$ decreases the sodium conductance because of $h(t)$. But the time constant τ_h is always larger than τ_m. The the variable h closing the channels reacts more slowly to the voltage increase than the variable m which opens the channel.
- A constant input $I(t) = I_0$ with I_0 larger than a critical value, induces regular spiking. Then, usually, the quantity of interest is the firing rate $\nu = 1/T$ where T is the inter-spike interval (ISI).
- There is no unique current threshold for spike generation. This can be seen with a Heavyside step current. (see page 48 of Gerstner's book).
- A time-dependent input may generate spikes that occur at irregular intervals.

2.1.2 Morris-Lecar model

The behavior of high dimensional nonlinear differential equations is difficult to analyze. Reduction of the four-dimensional Hodgkin-Huxley model is therefore useful.

The idea is the following.

1. The dynamics of the gating function m is much faster than that of the functions n, h, and V. We assume that m is an instantaneous function and approximate m with a steady-state function $m_\infty(V(t))$.
2. The time constants $\tau_h(V)$ and $\tau_n(V)$ are roughly the same, whatever the value of V, and $n_0(V)$ is closed to $1 - h_0(V)$. We approximate the two functions n and h by a single function u. Set $U = b - h = an$ for some constants b and a.

Then, the model becomes

$$\begin{aligned} C\frac{dV(t)}{dt} &= -\left(g_{Na}(m_\infty(t))^3\,(b-U(t))\,(V(t)-V_{Na}) + g_K\left(\frac{U(t)}{a}\right)^4(V(t)-V_K) + g\right. \\ \frac{dU(t)}{dt} &= -\frac{1}{\tau_u(V(t))}(U(t)-u_0(V(t))) \end{aligned}$$

Morris and Lecar (1981) proposed to simplify the previous model with fast and slow conductances:

$$\begin{aligned} C\frac{dV(t)}{dt} &= -\left(g_{fast}\,m_\infty(t)\,(V(t)-V_{fast}) + g_{slow}U(t)(V(t)-V_{slow}) + g_L(V(t)-V_L)\right. \\ \frac{dU(t)}{dt} &= -\frac{1}{\tau_u(V(t))}(U(t)-u_0(V(t))) = (\alpha(V(t))(1-U(t)) - \beta(V(t))U(t)) \end{aligned}$$

with

$$\begin{aligned} m_\infty(v) &= \frac{1}{2}\left(1+\tanh(\frac{v-V_1}{V_2})\right) \\ u_0(v) &= \frac{1}{2}\left(1+\tanh(\frac{v-V_3}{V_4})\right) \\ \tau_u(v) &= \frac{\tau_u}{\cosh\left(\frac{v-V_3}{V_4}\right)} \\ \alpha(v) &= \frac{1}{2}\phi\cosh\left(\frac{v-V_3}{2V_4}\right)\left(1+\tanh\left(\frac{v-V_3}{V_4}\right)\right), \\ \beta(v) &= \frac{1}{2}\phi\cosh\left(\frac{v-V_3}{2V_4}\right)\left(1-\tanh\left(\frac{v-V_3}{V_4}\right)\right) \end{aligned}$$

2.1.3 FitzHugh-Nagumo model

FitzHugh and Nagumo propose an oscillation model defined by

$$\begin{aligned} C\frac{dV(t)}{dt} &= \frac{1}{\varepsilon}\left(V(t)-V^3(t)-U(t)-I\right) \\ \frac{dU(t)}{dt} &= \gamma V(t) - U(t) + \beta \end{aligned}$$

2.2 Leaky-Integrate and Fire model

Detailed conductance-based neuron models reproduce electrophysiological behavior with a high accuracy, but are difficult to analyze, estimate from data. Simple phenomenological spiking neuron models are thus popular. These models are threshold models: Spikes are generated whenever the membrane potential V crosses some threshold S from below. Spikes being stereotyped events, they are fully characterized by their firing time. The best-know example of firing model is the leaky integrate-and-fire model.

The basic circuit of an integrate-and-fire model consists of a capacitor C in parallel with a resistor R driven by a current $I(t)$. The driving current $I(t)$ is the sum of two currents: the resistive current I_R which passes through the linear resistor R: $I_R(t) = V(t)/R$ (Ohm's law) and the capacitive current which charges the capacitor $C = q/V(t)$ where q is the charge: $CdV(t)/dt$. Thus

$$I(t) = \frac{V(t)}{R} + C\frac{dV(t)}{dt}$$

Introducing the time constant $\tau = RC$ of the 'leaky integrator' (time constant of the neuron), we obtain the leaky integrate-and-fire (LIF) model

$$\frac{dV(t)}{dt} = -\frac{V(t)}{\tau} + C\ I(t), \quad V(0) = V_0$$

In LIF models, the form of a spike is not described explicitly and is considered deterministic. The membrane voltage increases until it reaches a constant threshold S. Spikes are characterized by a *firing time* t_s, defined by a threshold criterion

$$t_s = \inf_{t \geq 0} \{V(t) = S\}$$

Immediately after t_s, a spike occurs and the potential is reset to its initial resting value $V(t_s) = V_0$, to start again to evolve. The combination of leaky integration and rest defined the basic LIF model.

Chapter 3

Modelling intra-cellular recordings with stochastic models

3.1 Stochastic Processes

The previous neuronal models are deterministic. However, it has been proved experimentally that similar input and similar experimental conditions lead to different sequences of spikes generation. It is now popular to deal with stochastic version of neuronal models to mimic the randomness observed in experiments.

We recall that a stochastic process $\{Y(t) : t \geq 0\}$ is a collection of random variables on a common probability space $(\Omega, \mathcal{A}, P)$ indexed by a parameter $t \in T$, which we usually interpret as time.

3.1.1 Definitions

We first give some definitions.

Definition 1. *A stochastic process* $\{Y(t) : t \geq 0\}$ *is*

- **stationary**, *if for all* $t_1 < t_2 < \ldots < t_n$, *and* $h > 0$, *the random n-vectors* $(Y(t_1), Y(t_2), \ldots, Y(t_n))$ *and* $(Y(t_1+h), Y(t_2+h), \ldots, Y(t_n+h))$ *are identically distributed*
- **Gaussian** *if for all* $t_1 < t_2 < \ldots < t_n$, *the n-vector* $(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n})$ *is multivariate normally distributed*
- **Markovian** *if for all* $t_1 < t_2 < \ldots < t_n$, $\mathbb{P}(Y(t_n) \leq y | Y(t_1), Y(t_2), \ldots, Y(t_{n-1})) = \mathbb{P}(Y(t_n) \leq y | Y(t_{n-1}))$

A process $\{Y(t) : t \geq 0\}$ is said to have **independent increments** if for all $t_1 < t_2 < \ldots < t_n$, the n random variables $Y(t_1) - Y(t_0), \ldots, Y(t_n) - Y(t_{n-1})$ are independent. This condition implies that $\{Y(t) : t \geq 0\}$ is Markovian, but not conversely.
The increments are said to be **stationary** if for any $t > s$, and $h > 0$ the distribution of $Y(t+h) - Y(s+h)$ is the same as the distribution of $Y(t) - Y(s)$.

Definition 2. *A stochastic process* $\{B(t) : t \geq 0\}$ *is a* **Brownian motion** *if*

1. $B(0) = 0$
2. *it has independent increments, i.e.* $B(t_1)$, $B(t_2) - B(t_1), \ldots, B(t_k) - B(t_{k-1})$ *are independent random variables for all* $0 \leq t_1 < t_2 < \ldots < t_k$.
3. $B(t+s) - B(s) \sim \mathcal{N}(0, t)$ *for each* $t > 0$.

It follows that $\{B(t) : t \geq 0\}$ is Gaussian, with mean zero $\mathbb{E}(B(t)) = 0$ and variance proportional to the elapsed time $Var(B(t)) = t$. We can also prove that $Cov(B(s), B(t)) = \min(s, t)$. Indeed if $s < t$, then

$$\begin{aligned} Cov(B(s), B(t)) &= \mathbb{E}(B(t)B(s)) \\ &= \mathbb{E}((B(t) - B(s))B(s)) + \mathbb{E}(B(s)B(s)) \\ &= \mathbb{E}(B(s)B(s)) = s \end{aligned}$$

because increments $(B(t) - B(s))$ and $B(s) - B(0)$ are independent.

Almost all sample paths of Brownian motion are everywhere continuous but nowhere differentiable. The Brownian motion is everywhere of unbounded variations

$$V_a^b(B) = \sup \sum_{k=1}^{n} |B(t_k) - B(t_{k-1})| = \infty$$

where the supremum is taken over all finite partitions $a \leq t_0 < \cdots < t_n \leq b$ of $[a, b]$.

3.1.2 Stochastic differential equations

An ordinary differential equation is defined by

$$\frac{dX(t)}{dt} = b(t, X(t)) \quad \text{or } dX(t) = b(t, X(t))dt, \quad X(0) = x_0$$

The integral form of the equation is

$$X(t) = X(0) + \int_0^t b(s, X(s))ds$$

Some regularity assumption is usually made on b, such as Lipschitz continuity, to ensure the existence of a unique solution $X(t)$ for each initial condition x_0.

A stochastic differential equation is written in the form

$$dX(t) = b(t, X(t))dt + \sigma(t, X(t))dB(t) \tag{3.1}$$

where $(B(t))$ is a Brownian motion, $b(t, x)$ is the drift function and $\sigma(t, x)$ is the diffusion coefficient. The integral form is then

$$X(t) = X(0) + \int_0^t b(s, X(s))ds + \int_0^t \sigma(s, X(s))dB(s)$$

The second integral is not an ordinary Riemann or Lebesgue integral, because the sample paths of a Brownian process are not of bounded variation on any bounded time interval. We need to define a stochastic integral in the sense of Itô. For any function f, we want to define the following integral

$$I(f) = \int_0^T f(s)dB(s)$$

Let us consider a piecewise function f, such that $f(t) = f_j$ on $t_j \leq t < t_{j+1}$ for $j = 1, 2, \ldots, n$, with $0 = t_1 < t_2 \ldots < t_n = T$. Then we want to define, with probability 1, that

$$I(f) = \sum_{j=1}^{n} f_j(B(t_{j+1}) - B(t_j))$$

This is a random variable with zero mean since it is the sum of random

variables with zero mean. We can compute the second moment:

$$\begin{aligned}\mathbb{E}(I(f)^2) &= \sum_{j=1}^{n} \mathbb{E}\left(f_j^2 \mathbb{E}\left(|B(t_{j+1}) - B(t_j)|^2\right)\right) \\ &= \sum_{j=1}^{n} \mathbb{E}(f_j^2)(t_{j+1} - t_j)\end{aligned}$$

For a general function f, we shall define $I(f)$ as the limit of integrals $I(f^{(n)})$ of step functions $f^{(n)}$ converging to f. More precisely, define a partition of the interval $[0, T]$ by $t_0 = t_1 < t_2 < \cdots < t_{n+1} = T$ where $|\delta_n| = \max\{|t_{j+1} - t_j| : j = 1, \ldots, n\}$ is the norm of the partition, and approximate

$$f(t) \approx f(t_j^*) \quad \text{for} \quad t_j \leq t < t_{j+1}$$

where the point t_j^* belongs to the interval $[t_j, t_{j+1}]$. Then we define

$$I(f) = \int_{t_0}^{T} f(s)\, dB_s = \lim_{|\delta_n| \to 0} \sum_{j=1}^{n} f(t_j^*)\left(B(t_{j+1}) - B(t_j)\right),$$

where the limit is considered with the mean-square convergence. When $f(t)$ is stochastic it turns out that - unlike ordinary integrals - it makes a difference how t_j^* is chosen! Two useful and common choices are the following:

- *The Itô integral:* $t_j^* = t_j$, the left end point.
- *The Stratonovich integral:* $t_j^* = (t_j + t_{j+1})/2$, the mid point.

We call $I(f)$ the **Itô stochastic integral** of f, resulting in a random variable

$$X(t) = \int_{t_0}^{t} f(s) dB(s)$$

which is mean-square integrable with zero mean $\mathbb{E}(X(t)) = 0$ and

$$\mathbb{E}(X(t)^2) = \int_{t_0}^{t} \mathbb{E}(f^2(s)) ds.$$

$X(t)$ inherits the linearity property from the sums.

Stratonovich SDE

In some applications (especially to study the so-called hypoelliptic models, see below), it is more appropriate to formulate SDE in terms of Stratonovich rather than Ito stochastic integrals. We call such an equation a **Stratonovich SDE** writing it in differential form as

$$dX(t) = \underline{b}(t, X(t))dt + \sigma(t, X(t)) \circ dB(t).$$

The $\circ$ notation denotes the use of Stratonovich calculus. As already defined, the Stratonovich integral

$$\int_0^t f(s, X(s)) \circ dB(s)$$

is defined as the mean-square limit of the sums

$$S_n = \sum_{j=1}^{n} f(t_j, \frac{1}{2}(X(t_j) + X(t_{j+1})))(B(t_{j+1}) - B(t_j))$$

for partitions $0 = t_1 < t_2 \ldots < t_{n+1} = t$ and $\delta_n = \max |t_{j+1} - t_j| \to \infty$.

It turns out that the solutions of the Stratonovich SDE also satisfy an Itô SDE with the same diffusion coefficient σ but with the modified drift coefficient

$$\underline{b}(t, x) = b(t, x) - \frac{1}{2}\sigma(t, x)\frac{\partial \sigma}{\partial x}(t, x)$$

Existence and uniqueness

To ensure the existence of a solution to Eq. 4.1 for $0 \leq t \leq T$ where T is fixed, the following is sufficient:

$$|b(t, x)| + |\sigma(t, x)| \ \leq \ C(1 + |x|)$$

for some constant C Kloeden and Platen (1992); Oksendal (2007). This ensures that $\{X(t)\}_{t \geq 0}$ does not explode, i.e that $\{|X(t)|\}_{t \geq 0}$ does not tend to ∞ in finite time. To ensure uniqueness of a solution the Lipschitz condition is sufficient:

$$|b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \ \leq \ D|x - y|$$

for some constant D. Note that only *sufficient* conditions are stated, and in many biological applications these are too strict, and weaker conditions can be found. We will not treat these here, though. In (Chapter Jacobsen) conditions on the functions b and σ to ensure that the process stays away from the boundaries without assuming the Lipschitz condition are discussed in detail. Note also that the above conditions are fulfilled for three of the processes described above.

Itôs formula

Stochastic differentials do not obey the ordinary chain rule as we know it from classical calculus Oksendal (2007). An additional term appears because $(dB(t))^2$ behaves like dt. We have

Proposition 1. Itô's formula. *Let $\{X(t)\}_{t\geq 0}$ be an Itô process given by*

$$dX(t) \;=\; b(t, X(t))dt + \sigma(t, X(t))\, dB(t)$$

and let $f(t, x)$ be a twice continuously differentiable function in x and once continuously differentiable function in t. Then

$$dY(t) \;=\; f(t, X(t))$$

is also an Itô process, and

$$dY(t) \;=\; \frac{\partial f}{\partial t}(t, X(t))dt + \frac{\partial f}{\partial x}(t, X(t))dX(t) + \frac{1}{2}\sigma^2(t, X(t))\frac{\partial^2 f}{\partial x^2}(t, X(t))dt \quad (3)$$

Note that the first two terms on the right hand side correspond to the chain rule we know from classical calculus, but an extra term appears in stochastic calculus because the Wiener process is of unbounded variation, and thus the quadratic variation comes into play.

Example 1. *Let us calculate the integral*

$$\int_0^t B(s)dB(s).$$

From classical calculus we expect a term like $\frac{1}{2}B(t)^2$ in the solution. Thus, we choose $f(t, x) = \frac{1}{2}x^2$ and $X(t) = B(t)$ and apply Itôs formula

to

$$Y(t) \;=\; f(t, B(t)) \;=\; \frac{1}{2}B(t)^2.$$

We obtain

$$dY(t) \;=\; \frac{\partial f}{\partial t}(t, B(t))dt + \frac{\partial f}{\partial x}(t, B(t))dB(t) + \frac{1}{2}\sigma^2(t, B(t))\frac{\partial^2 f}{\partial x^2}(t, B(t))dt \;=\; 0 + B($$

because $\sigma^2(t, B(t)) = 1$. *Hence*

$$Y(t) \;=\; \frac{1}{2}B(t)^2 \;=\; \int_0^t B(s)dB(s) + \frac{1}{2}\int_0^t ds \;=\; \int_0^t B(s)dB(s) + \frac{1}{2}t$$

and finally

$$\int_0^t B(s)dB(s) \;=\; \frac{1}{2}B(t)^2 - \frac{1}{2}t.$$

□

Example 2. *Let us find the solution* $\{X(t)\}_{t\geq 0}$ *to the Geometric Brownian motion*

$$dX(t) \;=\; \mu X(t)\, dt + \sigma X(t)\, dB(t).$$

Rewrite the equation as

$$\frac{dX(t)}{X(t)} \;=\; \mu\, dt + \sigma\, dB(t).$$

Thus, we have

$$\int_0^t \frac{dX(s)}{X(s)} \;=\; \mu\, t + \sigma\, B(t) \tag{3.3}$$

which suggests to apply Itôs formula on $f(t, x) = \log x$. *We obtain*

$$\begin{aligned} dY(t) \,=\, d(\log X(t)) \;&=\; \frac{\partial f}{\partial t}(t, X(t))dt + \frac{\partial f}{\partial x}(t, X(t))dX(t) + \frac{1}{2}\sigma^2(t, X(t))\frac{\partial^2 f}{\partial x^2}(t, X \\ &=\; 0 + \frac{1}{X(t)}dX(t) + \frac{1}{2}\sigma^2 X(t)^2\left(-\frac{1}{X(t)^2}\right)dt \;=\; \frac{dX(t)}{X(t)} - \frac{1}{2}\sigma^2 \end{aligned}$$

and thus

$$\frac{dX(t)}{X(t)} \;=\; d(\log X(t)) + \frac{1}{2}\sigma^2 dt. \tag{3.4}$$

Integrating Eq. 3.4 and using Eq. 3.3 we finally obtain

$$\log \frac{X(t)}{X_0} \;=\; \int_0^t \frac{dX(s)}{X(s)} - \frac{1}{2}\sigma^2 t \;=\; \mu\, t + \sigma\, B(t) - \frac{1}{2}\sigma^2 t$$

and so

$$X(t) \;=\; X_0 \exp\left\{\left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma\, B(t)\right\}.$$

Note that it is simply the exponential of a Wiener process with drift. □

Quadratic variation

It can be seen that the quadratic variation of a diffusion process solution to (4.1) is given by

$$< X, X >_t = \int_0^t \sigma^2(u, X(u))du$$

Infinitesimal generator of a diffusion process

Given a diffusion process X solution to (4.1) , a differential operator $\mathcal{L}$ of the form

$$(\mathcal{L}f)(x) = \frac{\sigma^2(x)}{2} f''(x) + b(x) f'(x)$$

with f two times differentiable is called the *infinitesimal generator* of the diffusion process X.

Transition density

From the Markovian property of the diffusion, it is also possible to define the transition density from value x at time s to value y at time t by $p(t, y|s, x)$. The transition density satisfies the Kolmogorov forward equation

$$\frac{\partial}{\partial t} p(t, y|s, x) = -\frac{\partial}{\partial y}(b(y) p(t, y|s, x)) + \frac{1}{2}\frac{\partial^2}{\partial y^2}(\sigma^2(y) p(t, y|s, x))$$

and the Kolmogorov backward equation

$$-\frac{\partial}{\partial s} p(t, y|s, x) = b(x)\frac{\partial}{\partial x}(p(t, y|s, x)) + \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}(p(t, y|s, x)).$$

Letting $t \to -\infty$ in the Kolmogorov forward equation, one obtain

$$\frac{d^2}{dx^2}(\sigma^2(x)\pi(x)) = 2\frac{d}{dx}(b(x)\pi(x))$$

where $\pi(x)$ is the stationary density.

Discrete time approximation

The solution of an Itô process is rarely explicit. When no explicit solution is available we can approximate different characteristics of the process by simulation, such as sample paths, moments, qualitative behavior etc. Usually such simulation methods are based on discrete approximations of the continuous solution to a stochastic differential equation Iacus (2008); Kloeden and Platen (1992). Different schemes are available depending on how good we want the approximation to be, which comes at a price of computer time. Assume we want to approximate a solution to Eq. (4.1) in the time interval $[0, T]$. Consider the time discretization

$$0 \;=\; t_0 < t_1 < \cdots < t_j < \cdots < t_N = T$$

and denote the time steps by $\Delta_j = t_{j+1} - t_j$ and the increments of the Wiener process by $\Delta B_j = B(t_{j+1}) - B(t_j)$. Then $\Delta B_j \sim N(0, \Delta_j)$, which we can use to construct approximations by drawing normally distributed numbers from a random number generator. For simplicity assume that the process is time-homogenous.

The simplest scheme is the stochastic analogue of the deterministic Euler scheme, and is called **the Euler-Maruyama scheme**. Approximate the process $(X(t))$ at the discrete time-points $t_j, 1 \le j \le N$ by the recursion

$$Y(t_{j+1}) \;=\; Y(t_j) + \mu(Y(t_j))\Delta_j + \sigma(Y(t_j))\Delta B_j \;\;;\;\; Y(t_0) = x_0 \quad (3.5)$$

where $\Delta B_j = \sqrt{\Delta_j} \cdot Z_j$, with Z_j being standard normal variables with mean 0 and variance 1 for all j. This procedure approximates the drift and diffusion functions by constants between time steps, so obviously the approximation improves for smaller time steps. To evaluate the convergence things are more complicated for stochastic processes, and

we operate with two criteria of optimality: the *strong* and the *weak* orders of convergence Bally and Talay (1996, 1995); Iacus (2008); Kloeden and Platen (1992).

Consider the expectation of the absolute error at the final time instant T of the Euler-Maruyama scheme. It can be shown that there exist constants $K > 0$ and $\delta_0 > 0$ such that

$$E(|X(T) - Y(t_N)|) \;\leq\; K\delta^{0.5}$$

for any time discretization with maximum step size $\delta \in (0, \delta_0)$. We say that the approximating process Y *converges in the strong sense* with order 0.5. This is similar to how approximations are evaluated in deterministic systems, only here we take expectations, since $X(T)$ and $Y(t_N)$ are random variables. Compare with the Euler scheme for an ordinary differential equation which has order of convergence 1. Sometimes we do not need a close *pathwise* approximation, but only some function of the value at a given final time T (e.g. $E(X(T))$, $E(X(T)^2)$ or generally $E(g(X(T)))$). In this case we have that there exist constants $K > 0$ and $\delta_0 > 0$ such that for any polynomial g

$$|E(g(X(T)) - E(g(Y(t_N))))| \;\leq\; K\delta$$

for any time discretization with maximum step size $\delta \in (0, \delta_0)$. We say that the approximating process Y *converges in the weak sense* with order 1.

3.2 Stochastic neuronal models

The previous neuronal models are deterministic. However, it has been proved experimentally that similar input and similar experimental conditions lead to different sequences of spikes generation. It is now popular to deal with stochastic version of neuronal models to mimic the randomness observed in experiments.

3.2.1 Stochastic LIF models

In stochastic LIF models, the membrane potential is described as a stochastic process, whereas the spike generation is due to the crossing of a threshold from below by the process.

Wiener process

Gerstein and Mandelbrot describe the time evolution of the sub threshold membrane potential through a Wiener process $V(t)$. Their model was motivated by experimental observations of the ISIs exhibiting histograms typical of stable distributions.

$$dV(t) = Idt + \sigma dB(t), \quad V(0) = V_0 \tag{3.6}$$

where $B(t)$ is a standard Brownian motion.

To mimic the spiking times, a constant absorbing boundary S is introduced. The spike times are then identified with the first passage time T of the Wiener process originated at $V(0) = V_0$ through the boundary. To obtain the renewal property, the process is instantaneously reset to V_0 after each spike.

Properties of the Wiener process

Proposition 2. *The moments of* $V(t)$ *are* $\mathbb{E}(V(t)) = V_0 + It$ *and* $Var(V(t)) = \sigma^2 t$. *The transition density function of* $V(t)$ *is Gaussian:*

$$p(v, t|v_0, 0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(v - v_0 - I\,t)^2}{2\sigma^2 t}\right)$$

Proof. $V(t)$ can be written in an integral form:

$$V(t) = V_0 + \int_0^t I ds + \sigma \int_0^t dB(s) = V_0 + It + \sigma \int_0^t dB(s)$$

As $\mathbb{E}(\int_0^t dB(s)) = 0$, we have $\mathbb{E}(V(t)) = V_0 + It$. Then $Var(V(t)) = \sigma^2 Var(\int_0^t dB(s)) = \sigma^2 \int_0^s ds = \sigma^2 t$. □

Ornstein-Uhlenbeck process

The most common stochastic LIF model is the Ornstein-Uhlenbeck process, described by the following stochastic differential equation (SDE):

$$dV(t) = \left(-\frac{V(t) - V_0}{\tau} + I \right) dt + \sigma \, dB(t), \quad V(0) = V_0 \tag{3.7}$$

where σ is the diffusion coefficient, $B(t)$ is standard Brownian motion.

Generation of the action potentials is not a part of process (4.8). To make a cell fire, a firing threshold is imposed at level $S > V_0$. the first time the process reaches the boundary level, an action potential is emitted and the membrane potential is instantaneously reset to V_0. Then the evolution restarts anew according to the same law.

Properties of the OU process

Proposition 3. *The conditional moments of $V(t)$ are*

$$\begin{aligned}
\mathbb{E}(V(t)|V(0) = V_0) &= \left(\frac{V_0}{\tau} + I \right) \tau (1 - e^{-t/\tau}) + V_0 e^{-t/\tau} \\
Var(V(t)|V(0) = V_0) &= \frac{\sigma^2 \tau}{2} (1 - e^{-2t/\tau}) \\
Cov(V(t), V(s)|V(0) = V_0) &= \frac{\sigma^2 \tau}{2} (e^{-|s-t|/\tau} - e^{-(s+t)/\tau})
\end{aligned}$$

Proof. Define $Y(t) = e^{t/\tau} V(t)$ and applying the Ito's formula with $f(t, v) = e^{t/\tau} v$ yields

$$dY(t) = \left(\frac{1}{\tau} e^{t/\tau} V(t) - \frac{1}{\tau} e^{t/\tau} V(t) + \frac{V_0}{\tau} e^{t/\tau} + e^{t/\tau} I \right) dt + \sigma e^{t/\tau} dB(t)$$

Thus

$$\begin{aligned} Y(t) &= \left(\frac{V_0}{\tau} + I\right) \int_0^t e^{s/\tau} ds + \sigma \int_0^t e^{s/\tau} dB(s) \\ &= \left(\frac{V_0}{\tau} + I\right) \tau(e^{t/\tau} - 1) + \sigma \int_0^t e^{s/\tau} dB(s) \end{aligned}$$

Finally

$$V(t) = \left(\frac{V_0}{\tau} + I\right) \tau(1 - e^{-t/\tau}) + \sigma \int_0^t e^{-(t-s)/\tau} dB(s)$$

The conditional moments can then be directly computed.□

We can deduce that the transition density function of $V(t)$ is Gaussian:

$$p(v, t|v_0, 0) = \frac{1}{\sqrt{\pi\sigma^2\tau(1 - e^{-2t/\tau})}} \exp\left(-\frac{(v - v_0 - \left(\frac{V_0}{\tau} + I\right)\tau(1 - e^{-t/\tau}) - V_0 e^{-t/\tau})^2}{\sigma^2\tau(1 - e^{-2t/\tau})}\right)$$

Feller process

In the OU process, the membrane potential is not limited from below. It may happen that $V(t)$ reaches unrealistic low values. The Feller process includes a lower bound of the depolarization

$$dV(t) = \left(-\frac{V(t) - V_0}{\tau} + I\right) dt + \sigma\sqrt{V(t) - V_I}\, dB(t), \quad V(0) = V_0 \quad (3.8)$$

where V_I is a new parameter, the minimum value allowed of the membrane potential. It is called the inhibitory reversal potential. This process is also called Cox-Ingersoll-Ross diffusion.

Generation of the action potentials is not a part of process (3.8) and a firing threshold S is introduced, as for the OU process.

Properties of the Feller process

Let us introduce $X(t) = V(t) - V_I$, and the parameters $X_0 = V_0 - V_I$, $\mu = I + X_0/\tau$. Then the Feller process $X(t)$ is the solution of the stochastic differential equation

$$dX(t) = \left(-\frac{X(t)}{\tau} + \mu\right) dt + \sigma\sqrt{X(t)} dB(t), \quad X(0) = X_0 \qquad (3.9)$$

where $X_0 > 0$ is the resetting point after emitting a spike, and the inhibitory reversal potential is shifted to the zero level. We denote $S_X = S - V_I$ the firing threshold.

For $2\mu \geq \sigma^2$, the zero value is an inaccessible barrier. The process is time-homogeneous.

Proposition 4. *The conditional expectation of process (3.9) is*

$$\mathbb{E}(X(t)|X_0, \mu) = X_0 e^{-t/\tau} + \mu\tau(1 - e^{-t/\tau})$$

The conditional variance and covariance are

$$Var(X(t)|X_0, \mu, \sigma^2) = \sigma^2 \frac{\tau}{2}(1 - e^{-t/\tau})(\mu\tau(1 - e^{-t/\tau}) + 2X_0 e^{-t/\tau})$$

and for $s < t$,

$$Cov(X(t), X(s)X_0, \mu, \sigma^2) = e^{-(t-s)/\tau} Var(X(s)|X_0, \mu, \sigma^2)$$

Proof. The Feller process has the explicit solution

$$X(t) = (X_0 - \mu\tau)e^{-t/\tau} + \sigma e^{-t/\tau} \int_0^t e^{u/\tau} \sqrt{X(u)} dB(u).$$

Under the hypothesis $2\mu > \sigma^2$, the process is stationary. The conditional transition density exists in explicit form. The process $Y(t) = 2cX(t)$ with $c = 2/(\tau\sigma^2(1-e^{-t/\tau}))$ has a conditional distribution $Y(t)|Y(0) = y_0$ which follows the law of the non central chi-squared distribution with $\nu = 4\mu/\sigma^2$ degrees of freedom and non centrality parameter $y_0 e^{-t/\tau}$. The transition density of $X(t)$ can be easily obtain. □

The transition density function is

$$p(x, t|x_0) = ce^{-r-s} \left(\frac{s}{r}\right)^{q/2} I_q(2\sqrt{rs})$$

where

$$c = \frac{2}{\tau\sigma^2(1 - e^{-t/\tau})},\ q = \frac{2\mu}{\sigma^2} - 1,\ r = cx_0 e^{-t/\tau},\ s = cx$$

and $I_q(\cdot)$ is the modified Bessel function.

3.2.2 Elliptic stochastic conductance based models

We study two types of stochastic conductance based models, elliptic or hypoelliptic. We start with the elliptic case. Elliptic stochastic neuronal models are diffusions with a non-degenerated diffusion coefficient.

FitzHugh-Nagumo model We introduce noise on two coordinates . Then the elliptic FitzHugh-Nagumo model is

$$\begin{aligned} dV(t) &= \frac{1}{\varepsilon}\left(V(t) - V(t)^3 - U(t) - I\right)dt + \sigma_1(V(t), U(t))dB_1(t) \\ dU(t) &= (\gamma V(t) - U(t) - \beta)\,dt + \sigma_2(V(t), U(t))dB_2(t) \end{aligned}$$

where $B_1(t)$ and $B_2(t)$ are two independent Brownian motions.

Let us detail some properties of this model. First remark that the conductance $U(t)$ is not autonomous, but depends on $V(t)$. The transition density of the couple $(V(t), U(t))$ is not explicit because of the non-linearities in the drifts and the diffusion coefficients. The transition density of $(U(t))$ given $V(t)$ is not explicit neither.

Synaptic inhibitory-excitatory model

We consider that the two conductances are stochastic and that the input is stochastic. This yields to the following elliptic system

$$\begin{aligned} dV(t) &= -\left(g_{inh}(t)(V(t) - V_{inh}) + g_{exc}(t)(V(t) - V_{exc}) + I(t)\right)dt + \gamma dB(t) \\ dg_{inh}(t) &= -\frac{1}{\tau_{inh}}(g_{inh}(t) - \bar{g}_{inh})dt + \gamma dB_i(t) \\ dg_{exc}(t) &= -\frac{1}{\tau_{exc}}(g_{exc}(t) - \bar{g}_{exc})dt + \gamma dB_e(t) \end{aligned}$$

where $B(t)$, $B_i(t)$ and $B_e(t)$ are three independent Brownian motions.

Let us detail some properties of this model. First remark that two conductances $g_{inh}(t)$ and $g_{exc}(t)$ are autonomous. They do not depend on $V(t)$. They are time homogeneous and they are described by a Ornstein-Uhlenbeck process. The transition densities are Gaussian, with known expectation and variances. For all $i \geq 1$, with step size Δ, we have for

both conductances:

$$g(t_{i+1}) = g(t_i)e^{-\Delta/\tau} + \bar{g}(1 - e^{-\Delta/\tau}) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \frac{\sigma^2\tau}{2}(1 - e^{-2\Delta/\tau}))$$

Morris-Lecar model We consider both the stochastic conductances and the stochastic input I. Then the elliptic Morris-Lecar model is

$$\begin{aligned} dV(t) &= -\left(g_{fast}\, m_\infty(t)\,(V(t) - V_{fast}) + g_{slow}U(t)(V(t) - V_{slow}) + g_L(V(t) - V_L) + \right. \\ dU(t) &= (\alpha(V(t))(1 - U(t)) - \beta(V(t))U(t))\, dt + \sigma(V(t), U(t))dB(t) \end{aligned}$$

where $\tilde{B}(t)$ and $B(t)$ are two independent Brownian motions.

Let us detail some properties of this model. First remark that the conductance $U(t)$ is not autonomous, but depends on $V(t)$. The transition density of the couple $(V(t), U(t))$ is not explicit because of the non-linearities in the drifts and the diffusion coefficients. The transition density of $(U(t))$ given $V(t)$ is not explicit neither. Fixing the value of $V(t) = v$, Ditlevsen and Greenwood (2013) study the process $U(t)$ showing that it stays between 0 and 1.

3.2.3 Hypoelliptic stochastic conductance based models

Different sources of noise have been introduced in neuronal models and the diffusion coefficient could be degenerate.
FitzHugh Nagump model Let us start with the FitzHugh Nagumo neuronal model:

$$\begin{aligned} dV(t) &= \frac{1}{\varepsilon}\left(V(t) - V(t)^3 - U(t) - I\right) dt \\ dU(t) &= (\gamma V(t) - U(t) - \beta)\, dt \end{aligned}$$

Assuming only the second coordinate noisy yields to

$$\begin{aligned} dV(t) &= \frac{1}{\varepsilon}\left(V(t) - V(t)^3 - U(t) - I\right) dt && (3.10) \\ dU(t) &= (\gamma V(t) - U(t) - \beta)\, dt + \sigma_2(V(t), U(t))dB_2(t) && (3.11) \end{aligned}$$

The diffusion coefficient is degenerated because there is no noise on the first component.

One can prove that the noise from the last component generates $\mathbb{R}^2$. To prove this hypoelliptic property, we need first to write the system in the Stratonovich form.

The Stratonovich form of (3.10) is

$$\left\{ \begin{array}{rcl} dV(t) & = & b_1(V(t),U(t))dt \\ dU(t) & = & \underline{b_2}(V(t),U(t))dt + \sigma_2(V(t),U(t)) \circ dB_2(t) \end{array} \right. \tag{3.12}$$

with $\underline{b_2}(v,u) := b_2(v,u) - \frac{1}{2}\sigma_2(u)\partial_u\sigma_2(u)$.

Then we writing the coefficients of (3.12) as vector fields

$$A_0(v,u) = \left(\begin{array}{c} b_1(v,u) \\ \underline{b}_2(v,u) \end{array} \right) \quad \text{and} \quad A_1(v,u) = \left(\begin{array}{c} 0 \\ \sigma_2(v,u) \end{array} \right)$$

We need to define the Lie bracket

Definition 3. *The Lie bracket $[X,Y] : M \to \mathbb{R}^n$ of the two vectors $X = \sum_{i=1}^n X^i e_i$, $Y = \sum_{i=1}^n Y^i e_i$ is given by*

$$[X,Y] := J_Y X - J_X Y = \sum_{i=1}^n \left(\sum_{j=1}^n (X^j \partial_j Y^i - Y^j \partial_j X^i) \right) e_i$$

where J_Y and J_X are the Jacobian matrices of Y and X, respectively.

The Lie bracket of (3.12) leads to

$$[A_0, A_1] = \left(\begin{array}{c} -\sigma_2(u,v)\partial_u \underline{b}_2 \\ \underline{b}_1 \partial_v \sigma_2 + \underline{b}_2 \partial_u \sigma_2 - \sigma_2 \partial_u \underline{b}_2 \end{array} \right)$$

Definition 4. *A system is called hypoelliptic in the sense of stochastic calculus of variations (Nualart, 2006) if the vector fields*

$$A_1(v,u) = \left(\begin{array}{c} 0 \\ \sigma_2(v,u) \end{array} \right)$$

and the Lie Bracket $[A_0, A_1]$ generate $\mathbb{R}^2$.

SIE model

Let us start with the SIE neuronal model assuming the capacitance $C = 1$.

$$
\begin{aligned}
dV(t) &= -\left(g_{inh}(t)(V(t)-V_{inh}) + g_{exc}(t)(V(t)-V_{exc}) + I(t)\right)dt \\
dg_{inh}(t) &= -\frac{1}{\tau_{inh}}(g_{inh}(t)-\bar{g}_{inh})dt \\
dg_{exc}(t) &= -\frac{1}{\tau_{exc}}(g_{exc}(t)-\bar{g}_{exc})dt
\end{aligned}
$$

Assuming the inhibitory and excitatory conductances are noisy yields to

$$
\begin{aligned}
dV(t) &= -\left(g_{inh}(t)(V(t)-V_{inh}) + g_{exc}(t)(V(t)-V_{exc}) + I(t)\right)dt && (3.13)\\
dg_{inh}(t) &= -\frac{1}{\tau_{inh}}(g_{inh}(t)-\bar{g}_{inh})dt + \gamma_i(g_{inh})dB_i(t) && (3.14)\\
dg_{exc}(t) &= -\frac{1}{\tau_{exc}}(g_{exc}(t)-\bar{g}_{exc})dt + \gamma_e(g_{exc})dB_e(t) && (3.15)\\
& && (3.16)
\end{aligned}
$$

where $B_i(t)$ and $B_e(t)$ are two independent Brownian motions and γ_i, γ_e two diffusion coefficients. The diffusion coefficient is degenerated because there is no noise on the first component.

First remark that two conductances $g_{inh}(t)$ and $g_{exc}(t)$ are autonomous. They do not depend on $V(t)$. They are time homogeneous.

One can prove that the noise from the two last components generates $\mathbb{R}^3$. To prove this hypoelliptic property, we need first to write the system in the Stratonovich form.

The Stratonovich form of (3.13) is

$$
\left\{
\begin{aligned}
dV(t) &= f(V(t), g_{inh}(t), g_{exc}(t))dt \\
dg_{inh}(t) &= \underline{b_i}(V(t), g_{inh}(t), g_{exc}(t))dt + \gamma_i(g_{inh}(t)) \circ dB_i(t) \\
dg_{exc}(t) &= \underline{b_e}(V(t), g_{inh}(t), g_{exc}(t))dt + \gamma_e(g_{exc}(t)) \circ dB_e(t)
\end{aligned}
\right. \qquad (3.17)
$$

with $\underline{b}(v, g_i, g_e) := b(v, g_i, g_e) - \frac{1}{2}\gamma(g)\partial_g\gamma(g)$.

Then we writing the coefficients of (3.17) as vector fields

$$
A_0(v, g_i, g_e) = \begin{pmatrix} f(v, g_i, g_e) \\ \underline{b_i}(v, g_i, g_e) \\ \underline{b_e}(v, g_i, g_e) \end{pmatrix} \quad \text{and} \quad A_1(v, g_i, g_e) = \begin{pmatrix} 0 \\ \gamma_i(g_i) \\ \gamma_e(g_e) \end{pmatrix}
$$

Under conditions

1. $-\gamma_i(g_i)\partial_{g_i}\underline{b}_i - \gamma_e(g_e)\partial_{g_e}\underline{b}_e \neq 0$
2. the functions $\underline{b}_i\partial_{g_i}\gamma_i - \gamma_i\partial_{g_i}\underline{b}_i$ and $\underline{b}_e\partial_{g_e}\gamma_e - \gamma_e\partial_{g_e}\underline{b}_e$ are not proportional

the vectors A_1 and $[A_0, A_1]$ generate $\mathbb{R}^3$ and system (3.17) is hypoelliptic.

A non hypoelliptic model Assuming the input function $I(t)$ is noisy or random $(I(t) + \sigma\eta(t))$, yields to the following stochastic model

$$\begin{aligned} CdV(t) &= -\left(g_{inh}(t)(V(t) - V_{inh}) + g_{exc}(t)(V(t) - V_{exc}) + I(t)\right)dt + \sigma dB(t) \\ dg_{inh}(t) &= -\frac{1}{\tau_{inh}}(g_{inh}(t) - \bar{g}_{inh})dt \\ dg_{exc}(t) &= -\frac{1}{\tau_{exc}}(g_{exc}(t) - \bar{g}_{exc})dt \end{aligned}$$

where $B(t)$ is a Brownian motion. The diffusion coefficient is degenerated but the noise from the first component can not distribute $\mathbb{R}^3$. This model is not hypoelliptic, and will not be considered thereafter.

Morris-Lecar model Let us consider the deterministicMorris-Lecar model assuming the capacitance $C = 1$.

$$\begin{aligned} dV(t) &= -\left(g_{fast}\, m_\infty(t)\,(V(t) - V_{fast}) + g_{slow}U(t)(V(t) - V_{slow}) + g_L(V(t) - V_L) + \right. \\ dU(t) &= \left(\alpha(V(t))(1 - U(t)) - \beta(V(t))U(t)\right)dt \end{aligned}$$

Assuming the slow conductance is random (from the Piecewise deterministic Markov Process), then the stochastic ML model is defined by

$$\left\{ \begin{array}{rcl} dV(t) &=& f(V(t), U(t))dt \\ dU(t) &=& b(V(t), U(t))dt + \sigma(V(t), U(t))dB(t) \end{array} \right. \tag{3.18}$$

where $\sigma(V(t), U(t))$ is the diffusion coefficient modeling the channel or conductance noise. We consider the following function that ensures that $U(t)$ stays bounded in the unit interval is $\sigma \leq 1$

$$\sigma(V(t), U(t)) = \sigma\sqrt{2\frac{\alpha(V(t))\beta(V(t))}{\alpha(V(t)) + \beta(V(t))}U(t)(1 - U(t))}.$$

Now we study the property of the system (3.18).

Assume that the following condition holds

(C1) $\forall (v, uz) \in \mathbb{R} \times \mathbb{R}, \quad \partial_u f(v,u) \neq 0$

Under assumption **(C1)**, system (3.18) is hypoelliptic in the sense of stochastic calculus of variations (Nualart, 2006). Indeed, the Stratonovich form of (3.18) is

$$\left\{ \begin{array}{rcl} dV(t) & = & f(V(t),U(t))dt \\ dU(t) & = & \tilde{b}(V(t),U(t))dt + \sigma(V(t),U(t)) \circ dB(t) \end{array} \right. \tag{3.19}$$

with $\tilde{b}(v,u) := b(v,u) - \frac{1}{2}\sigma(v,u)\partial_u\sigma(v,u)$. Writing the coefficients of (3.19) as vector fields

$$A_0(v,u) = \begin{pmatrix} f(v,u) \\ \tilde{b}(v,u) \end{pmatrix} \quad \text{and} \quad A_1(v,u) = \begin{pmatrix} 0 \\ \sigma(v,u) \end{pmatrix}$$

and computing their Lie bracket leads to

$$[A_0, A_1] = \begin{pmatrix} \sigma(v,u)\partial_u f(v,u) \\ g(v,u) \end{pmatrix}$$

The form of g is explicit but not detailed here. Under condition **(C1)** the vectors A_1 and $[A_0, A_1]$ generate $\mathbb{R}^2$ and system (3.18) is hypoelliptic.

Chapter 4

Estimating intra-cellular recordings with one-dimensional SDE

4.1 Assumptions

We assume that X is a process satisfying the following SDE

$$dX(t) = b_\theta(X(t))dt + \sigma_\theta(X(t))dB(t). \tag{4.1}$$

We assume that the process is observed at discrete times $t_i = i\Delta$, $i = 1, \ldots, n$ and $T = n\Delta$. We denote $X_i = X_{i\Delta}$. We denote $\mathcal{F}_n = \sigma\{X_i, i \leq n\}$ the σ-field generated by the first n observations.

We want to estimate the parameters θ from the discrete observations $(X_i)_{1\leq i\leq n}$. In the following, we assume that

1. **Linear growth assumption** There exists a constant K independent of θ such that for all x

$$|b_\theta(x)| + |\sigma_\theta(x)| \leq K(1 + |x|)$$

2. **Global Lipschitz assumption** There exists a constant K independent of θ such that

$$|b_\theta(x) - b_\theta(y)| + |\sigma_\theta(x) - \sigma_\theta(y)| \leq K|x - y|$$

3. **Positiveness of diffusion coefficient**

$$\inf_x \sigma_\theta^2(x) > 0$$

4. **Bounded moments** For all $k > 0$, all moments of order k of the diffusion process exist and are such that

$$\sup_t \mathbb{E}|X(t)|^k < \infty$$

4.2 Maximum likelihood estimation

Using the Markov property of X, the likelihood is defined by

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i|X_{i-1}, \Delta) p_\theta(X_0)$$

where $p_\theta(X_i|X_{i-1}, \Delta)$ is the transition density of X for a time step Δ.

We denote by $\ell_n(\theta) = \log L_n(\theta)$ the log-likelihood function

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(X_i|X_{i-1}, \Delta) + \log(p_\theta(X_0))$$

Then we define the maximum likelihood estimator as

$$\hat{\theta}_n = \arg\max \ell_n(\theta).$$

We can prove the following theorem (Kessler, 1997)

Theorem 1. *Suppose assumptions 1-4 are satisfied. If $\theta = (\theta_1, \theta_2) \in \Theta$, Θ a compact subset of $\mathbb{R}^p$, $b_\theta(x) = b_{\theta_1}(x)$ and $\sigma_\theta(x) = \sigma_{\theta_2}(x)$, then the maximum likelihood estimators obtained on the conditional likelihood $\ell_n(\theta)$ are consistent and asymptotically normal:*

$$\begin{pmatrix} \sqrt{n\Delta}(\hat{\theta}_{n,1} - \theta_{1,0}) \\ \sqrt{n}(\hat{\theta}_{n,2} - \theta_{2,0}) \end{pmatrix} \to_d \mathcal{N}(0, \mathcal{I}^{-1}(\theta_0)) \tag{4.2}$$

with

$$\mathcal{I}(\theta_0)^{-1} = \begin{pmatrix} \left(\int \left(\frac{\partial_{\theta_1} b(x,\theta_{1,0})}{\sigma(x,\theta_{2,0})} \right)^2 \pi(dx) \right)^{-1} & 0 \\ 0 & 2 \left(\int \left(\frac{\partial_{\theta_2} \sigma(x,\theta_{2,0})}{\sigma^2(x,\theta_{2,0})} \right)^2 \pi(dx) \right)^{-1} \end{pmatrix} \tag{4.3}$$

where $\theta_{1,0}$ and $\theta_{2,0}$ are the true values of the parameter and $\pi(\cdot)$ is the invariant density of the diffusion process.

Example 3. *Let us consider the Wiener process*

$$dV(t) = Idt + \sigma dB(t), \quad V(0) = V_0. \tag{4.4}$$

The parameters to be estimated are $\theta = (I, \sigma)$. The log-likelihood is

$$\ell_n(\theta) = -\frac{n}{2}\log(2\pi\sigma^2\Delta) - \frac{1}{2}\sum_{i=1}^{n}\frac{(V_i - V_{i-1} - I\Delta)^2}{\Delta\sigma^2}$$

Deriving the likelihood, we obtain the score function

$$\frac{\partial\ell_n(\theta)}{\partial I} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta), \quad \frac{\partial\ell_n(\theta)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\Delta\sigma^4}\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta)^2.$$

Thus the maximum likelihood estimators are

$$\hat{I} = \frac{1}{n\Delta}\sum_{i=1}^{n}(V_i - V_{i-1}), \quad \hat{\sigma}^2 = \frac{1}{n\Delta}\sum_{i=1}^{n}(V_i - V_{i-1} - \hat{I}\Delta)^2$$

Applying the law of large numbers and the central limit theorem, we obtain

$$\frac{1}{\sqrt{n\Delta}}\frac{\partial\ell_n(\theta)}{\partial I} = \frac{1}{\sigma}\frac{n}{\sqrt{n}}\frac{\frac{1}{n}\sum_{i=1}^{n}(V_i - V_{i-1}) - I\Delta}{\sqrt{\Delta}\sigma} = \frac{1}{\sigma}\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}(V_i - V_{i-1}) - I\Delta}{\sqrt{\Delta}\sigma}$$

Noticing that

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}(V_i - V_{i-1}) - I\Delta}{\sqrt{\Delta}\sigma} \longrightarrow_{\mathcal{L}} \mathcal{N}(0,1)$$

We obtain

$$\frac{1}{\sqrt{n\Delta}}\frac{\partial\ell_n(\theta)}{\partial I} \longrightarrow_{\mathcal{L}} \mathcal{N}(0, \frac{1}{\sigma^2})$$

Similarly, we can show that

$$\frac{1}{\sqrt{n}}\frac{\partial\ell_n(\theta)}{\partial\sigma^2} = \frac{1}{\sqrt{n}}n\frac{\frac{1}{n}\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta)^2 - \Delta\sigma^2}{2\Delta\sigma^4} = \frac{\sqrt{n}}{\sqrt{2}\sigma^2}\frac{\frac{1}{n}\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta)^2}{\sqrt{2}\Delta\sigma^2}$$

Using the fact that $var(X^2) = 2var(X)^2$ for a centered Gaussian variable, we prove that

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta)^2 - \Delta\sigma^2}{\sqrt{2}\Delta\sigma^2} \longrightarrow_{\mathcal{L}} \mathcal{N}(0,1)$$

and we get

$$\frac{1}{\sqrt{n}}\frac{\partial \ell_n(\theta)}{\partial \sigma^2} \longrightarrow_{\mathcal{L}} \mathcal{N}(0, \frac{1}{2\sigma^4})$$

Finally,

$$\begin{pmatrix} \frac{1}{\sqrt{n\Delta}}\frac{\partial}{\partial I}\ell_n(\theta) \\ \frac{1}{\sqrt{n}}\frac{\partial}{\partial \sigma^2}\ell_n(\theta) \end{pmatrix} \longrightarrow \mathcal{N}_2(0, \mathcal{I}(\theta))$$

where $\mathcal{I}(\theta)$ is the Fisher information matrix

$$\mathcal{I}(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \tag{4.5}$$

Computing the second derivatives, we obtain

$$\frac{\partial^2 \ell_n}{\partial I^2} = -\frac{n\Delta}{\sigma^2}, \quad \frac{\partial^2 \ell_n}{\partial I \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta), \quad \frac{\partial^2 \ell_n}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6\Delta}\sum_{i=1}^{n}(V_i - V_{i-1}$$

We have:

$$-\frac{1}{n\Delta}\frac{\partial^2 \ell_n}{\partial I^2} = \frac{1}{\sigma^2}$$

For the second term, we use the fact that $\mathbb{E}(V_i - V_{i-1} - I\Delta) = 0$ and $\frac{1}{n^2}\mathbb{E}((\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta))^2) = \frac{n\sigma^2\Delta}{n^2} \to 0$ to prove that

$$-\frac{1}{n}\frac{1}{\sigma^4}\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta) \longrightarrow_P 0$$

For the last term, we use $\mathbb{E}((V_i - V_{i-1} - I\Delta)^2) = \Delta\sigma^2$, thus

$$\frac{1}{n}\left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6\Delta}\sum_{i=1}^{n}(V_i - V_{i-1} - I\Delta)^2\right) = -\frac{1}{2\sigma^4}$$

Finally the matrix of the Hessian converges to the Fisher Information matrix:

$$-\begin{pmatrix} \frac{1}{n\Delta}\frac{\partial^2}{\partial I^2}\ell_n(\theta) & \frac{1}{n\Delta}\frac{\partial^2}{\partial I \partial \sigma^2}\ell_n(\theta) \\ \frac{1}{n}\frac{\partial^2}{\partial I \partial \sigma^2}\ell_n(\theta) & \frac{1}{n}\frac{\partial^2}{\partial \sigma^2 \partial \sigma^2}\ell_n(\theta) \end{pmatrix} \longrightarrow_P \mathcal{I}(\theta)$$

Then, we note that

$$-\partial \ell_n(\theta_0) = (\hat{\theta} - \theta_0)'\partial^2 \ell_n(\theta_0) + O(1)$$

to prove

$$\begin{pmatrix} \sqrt{n\Delta}(\hat{I}_n - I_0) \\ \sqrt{n}(\hat{\sigma}_n^2 - \sigma_0^2) \end{pmatrix} = - \begin{pmatrix} \sqrt{n\Delta} \\ \sqrt{n} \end{pmatrix} \partial \ell_n(\theta_0)(\partial^2 \ell_n(\theta_0))^{-1} = - \begin{pmatrix} \frac{1}{\sqrt{n\Delta}} \partial_I \ell_n(\theta_0) \\ \frac{1}{\sqrt{n}} \partial_\sigma \ell_n(\theta_0) \end{pmatrix} \begin{pmatrix} \frac{1}{n\Delta} \frac{\partial^2}{\partial I^2} \\ \frac{1}{n} \frac{\partial^2}{\partial I \partial \sigma^2} \end{pmatrix} \tag{4.6}$$

Using Slutsky lemma and the previous results, we can prove that

$$\begin{pmatrix} \sqrt{n\Delta}(\hat{I}_n - I_0) \\ \sqrt{n}(\hat{\sigma}_n^2 - \sigma_0^2) \end{pmatrix} \to_d \mathcal{N}(0, \mathcal{I}^{-1}(\theta_0)) \tag{4.7}$$

The definition of $\mathcal{I}(\theta)$ in theorem 1 leads to the same formula.

Example 4. *We consider the Ornstein-Uhlenbeck process*

$$dV(t) = - \left(\frac{V(t) - \alpha}{\tau} \right) dt + \sigma dB(t), \quad V(0) = V_0. \tag{4.8}$$

Eq. 4.8 provides an explicit expression of V_{i+1} as a function of V_i and $\theta = (\tau, \alpha, \sigma)$:

$$V_{i+1} = V_i e^{-\Delta/\tau} + \alpha(1 - e^{-\Delta/\tau}) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \frac{\sigma^2 \tau}{2}(1 - e^{-2\Delta/\tau})) \tag{4.9}$$

The /og likelihood is thus explicit and equal to

$$\ell_n(\theta) = \sum_{i=1}^{n} \log \varphi \left(V_{i+1}; V_i e^{-\frac{\Delta}{\tau}} + \alpha(1 - e^{-\frac{\Delta}{\tau}}), \frac{\sigma^2 \tau}{2}(1 - e^{-\frac{2\Delta}{\tau}}) \right),$$

where $\varphi(x; \mu, \sigma^2)$ denotes the density of a Gaussian variable with mean μ and variance σ^2. The unique maximum of the likelihood function provides the MLE $\hat{\theta} = (\hat{\tau}, \hat{\alpha}, \hat{\sigma}^2)$, given by two equations. Let $X = ((1 \dots 1)' V_{1:n-1})$, $\rho = e^{-\Delta/\tau}, \beta = \alpha(1 - \rho)$, then

$$\begin{aligned} (\hat{\beta} \quad \hat{\rho})' &= (X'X)^{-1} X' V_{2:n} \\ \hat{\alpha} &= \hat{\beta}/(1 - \hat{\rho}) \\ \hat{\sigma}^2 &= \frac{2 \sum_{i=1}^{n} (V_{i+1} - \hat{\alpha} - (V_i - \hat{\alpha}) e^{-\triangle/\hat{\tau}})^2}{n(1 - e^{-2\triangle/\hat{\tau}})\hat{\tau}} \end{aligned}$$

It requires that $\sum_{j=1}^{n} (V_{i+1} - \hat{\alpha})(V_i - \hat{\alpha}) > 0$. Otherwise there is no solution. □

4.3 Least squares

Usually, the transition density is unknown so maximum likelihood estimation on the true likelihood of the process is a rare case. When the moments are known, least squares estimators can be used. This is the case of the Feller process.

Example 5. *We consider the Feller process*

$$dV(t) = \left(-\frac{V(t)}{\tau} + \mu\right) dt + \sigma\sqrt{V(t) - V_I}\, dB(t), \quad V(0) = V_0. \quad (4.10)$$

A spike is emitted when the membrane voltage $V(t)$ hits the threshold S.

Several estimation methods have been proposed for the Feller model (see Bibbona et al., 2010, for a review). We study the least-squares and the conditional least-squares methods. We focus on the estimation of μ and σ. We assume that the membrane potential is observed at discrete times $t_i = i\Delta$ where h is the step size.

4.3.1 Least-squares method

The least-squares estimator of μ is defined as the minimizer of the function

$$\begin{aligned} LS_1(\mu) &= \sum_{i=1}^{n} (V_i - \mathbb{E}(V_i|V_0, \mu))^2 \\ &= \sum_{i=1}^{n} \left(V_i - V_0 e^{-i\Delta/\tau} + \mu\tau(1 - e^{-i\Delta/\tau})\right)^2 \end{aligned}$$

with respect to μ. The solution is

$$\hat{\mu}_{LS} = \frac{\sum_{i=1}^{n} (V_i - V_0 e^{-i\Delta/\tau})(1 - e^{-i\Delta/\tau})}{\tau \sum_{i=1}^{n} (1 - e^{-i\Delta/\tau})^2}$$

We can easily prove that $\mathbb{E}(\hat{\mu}_{LS}) = \mu$. The variance of $\hat{\mu}_{LS}$ is

$$Var(\hat{\mu}_{LS}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (1 - e^{-j\Delta/\tau})(1 - e^{-i\Delta/\tau}) Cov(V_i, V_j|V_0, \mu, \sigma^2)}{\tau^2 \left[\sum_{k=1}^{n} (1 - e^{-k\Delta/\tau})^2\right]^2}.$$

Then, we estimate σ^2 as the minimizer of the function

$$LS_2(\sigma^2) = \sum_{i=1}^{n} \left[(V_i - \mathbb{E}(V_i|V_0,\mu))^2 - \mathbb{E}(V_i - \mathbb{E}(V_i|V_0,\mu))^2 \right]^2$$

We replace μ by its estimator $\hat{\mu}_{LS}$ in this criterion. Then, we obtain the estimator

$$\hat{\sigma}^2_{LS} = \frac{\sum_{i=1}^{n}(V_i - \mathbb{E}(V_i|V_0,\hat{\mu}_{LS}))^2 \frac{\tau}{2}(1-e^{-i\Delta/\tau})(\hat{\mu}_{LS}\tau(1-e^{-i\Delta/\tau}) + 2V_0 e^{-i\Delta/\tau})}{\sum_{i=1}^{n}(\frac{\tau}{2}(1-e^{-i\Delta/\tau})(\hat{\mu}_{LS}\tau(1-e^{-i\Delta/\tau}) + 2V_0 e^{-i\Delta/\tau}))^2}$$

The estimator $\hat{\sigma}^2_{LS}$ is biased because we use the value $\hat{\mu}_{LS}$ instead of μ.

4.3.2 Conditional least-squares method

The conditional least-squares method used the conditional expectation and variance. The criterion function is thus

$$\begin{aligned} CLS_1(\mu) &= \sum_{i=1}^{n}(V_i - \mathbb{E}(V_i|V_{i-1},\mu))^2 \\ &= \sum_{i=1}^{n}\left(V_i - V_{i-1}e^{-\Delta/\tau}\right)^2 \end{aligned}$$

with respect to μ. The solution is

$$\hat{\mu}_{CLS} = \frac{\sum_{i=1}^{n}(V_i - V_{i-1}e^{-\Delta/\tau})}{n\tau(1-e^{-\Delta/\tau})}$$

The estimator $\hat{\mu}_{CLS}$ is unbiased. The variance is

$$Var(\hat{\mu}_{CLS}) = \sigma^2 \frac{e^{-\Delta/\tau}(1-e^{-n\Delta/\tau})(V_0 - \mu\tau) + \frac{n\mu\tau}{2}(1-e^{-2\Delta/\tau})}{n^2\tau(1-e^{-\Delta/\tau})^2}$$

Then, we estimate σ^2 as the minimizer of the function

$$CLS_2(\sigma^2) = \sum_{i=1}^{n}\left[(V_i - \mathbb{E}(V_i|V_{i-1},\mu))^2 - \mathbb{E}((V_i - \mathbb{E}(V_i|V_{i-1},\mu))^2|V_{i-1})\right]^2$$

We replace μ by its estimator $\hat{\mu}_{CLS}$ in this criterion. Then, we obtain the estimator

$$\hat{\sigma}^2_{CLS} = \frac{n}{n-1}\frac{\sum_{i=1}^{n}(V_i - \mathbb{E}(V_i|V_{i-1},\hat{\mu}_{CLS}))^2 \frac{\tau}{2}(1-e^{-\Delta/\tau})(\hat{\mu}_{LS}\tau(1-e^{-\Delta/\tau}) + 2V_{i-1}e^{-\Delta}}{\sum_{i=1}^{n}(\frac{\tau}{2}(1-e^{-\Delta/\tau})(\hat{\mu}_{CLS}\tau(1-e^{-\Delta/\tau}) + 2V_{i-1}e^{-\Delta/\tau}))^2}$$

4.4 Pseudo-likelihood estimation

When the transition density and the moments of the process are unknown, the Euler scheme can be used to approximate the likelihood. This leads to a pseudo-likelihood approach.

The Euler scheme produces the following discretization of SDE (4.1).

$$X(t+\Delta) = X(t) + \Delta b_\theta(X(t)) + \sqrt{\Delta}\sigma_\theta(X(t))\eta,$$

where $\eta \sim \mathcal{N}(0,1)$. Therefore the transition density of the process can be written as

$$p_\theta(y|x,\Delta) = \frac{1}{\sqrt{2\pi\Delta\sigma_\theta^2(x)}} \exp\left\{-\frac{1}{2}\frac{(y-x-b_\theta(x)\Delta)^2}{\Delta\sigma_\theta^2(x)}\right\}$$

This approximation is good if Δ is small. Otherwise some bias is introduced.

The log-likelihood of the discretized process, also called the **locally Gaussian approximation** or the **pseudo-likelihood**, is

$$\ell_n(\theta) = -\frac{1}{2}\left\{\sum_{i=1}^{n}\frac{(X_i - X_{i-1} - \Delta b_\theta(X_{i-1}))^2}{\Delta\sigma_\theta^2(X_{i-1})} + \sum_i \log(2\pi\Delta\sigma_\theta^2(X_{i-1}))\right\}$$

Note that if the diffusion coefficient is constant $\sigma(x) = \sigma > 0$, then we want to estimate the parameters σ and θ. Given that σ^2 is constant, the maximization of the pseudo-likelihood is equivalent to the maximization of the function

$$\sum_{i=1}^{n}(X_i - X_{i-1})b_\theta(X_{i-1}) - \frac{\Delta}{2}\sum_{i=1}^{n} b_\theta^2(X_{i-1})$$

We can prove the following theorem (Kessler, 1997)

Theorem 2. *Suppose assumptions 1-4 are satisfied. If $\theta = (\theta_1, \theta_2) \in \Theta$, Θ a compact subset of $\mathbb{R}^p$, $b_\theta(x) = b_{\theta_1}(x)$ and $\sigma_\theta(x) = \sigma_{\theta_2}(x)$, then the pseudo likelihood estimators obtained on the conditional pseudo likelihood $\ell_n(\theta)$ are consistent and asymptotically normal if $n \to \infty$, $n\Delta \to \infty$ and $n\Delta^2 \to 0$:*

$$\begin{pmatrix} \sqrt{n\Delta}(\hat{\theta}_{n,1} - \theta_{1,0}) \\ \sqrt{n}(\hat{\theta}_{n,2} - \theta_{2,0}) \end{pmatrix} \to_d \mathcal{N}(0, \mathcal{I}_0^{-1}) \tag{4.11}$$

with

$$\mathcal{I}_0^{-1} = \begin{pmatrix} \left(\int \left(\frac{\partial_{\theta_1} b(x,\theta_{1,0})}{\sigma(x,\theta_{2,0})} \right)^2 \pi(dx) \right)^{-1} & 0 \\ 0 & 2 \left(\int \left(\frac{\partial_{\theta_2} \sigma(x,\theta_{2,0})}{\sigma^2(x,\theta_{2,0})} \right)^2 \pi(dx) \right)^{-1} \end{pmatrix} \quad (4.12)$$

where $\theta_{1,0}$ and $\theta_{2,0}$ are the true values of the parameter and $\pi(\cdot)$ is the invariant density of the diffusion process.

Chapter 5

Estimation for multidimensionnal elliptic SDE

We consider in this section the case of multi dimensional elliptic SDE. We first treat the ideal case of complete observations. In practice, only the first coordinate $V(t)$ is observed at discrete times. The other components are hidden. Two cases have to be distinguished depending on whether the hidden components are autonomous or not. When they are autonomous, then the observations can be considered as observations of a hidden Markov model (HMM). If not, the HMM is degenerated and more complex estimation methods have to be considered.

5.1 Ideal case of complete observations

Let us start with the ideal case of complete observations of a multidimensional neuronal SDE. For example, we consider the simple model:

$$\begin{aligned} dV(t) &= -\left(g_{inh}(t)(V(t)-V_{inh}) + g_{exc}(t)(V(t)-V_{exc}) + I\right)dt + \gamma_v dB(t) \quad (5.1)\\ dg_{inh}(t) &= -\frac{1}{\tau_{inh}}(g_{inh}(t)-\bar{g}_{inh})dt + \gamma_i dB_i(t) \quad (5.2)\\ dg_{exc}(t) &= -\frac{1}{\tau_{exc}}(g_{exc}(t)-\bar{g}_{exc})dt + \gamma_e dB_e(t) \quad (5.3) \end{aligned}$$

where $B(t)$, $B_i(t)$ and $B_e(t)$ are three independent Brownian motions.

The transition density and the moments of this multidimensional process are unknown because of the non-linear drift. A multidimensional version of the Euler scheme can be used to approximate the like-

lihood. This leads to a pseudo-likelihood approach, as seen in the one-dimensional case.

Let us denote $Z(t) = (V(t), g_{inh}(t), g_{exc}(t))$. The Euler scheme produces the following discretization of SDE (5.1).

$$Z(t+\Delta) = Z(t) + \Delta b(Z(t),\theta) + \sqrt{\Delta}\Gamma\eta,$$

where $\eta \sim \mathcal{N}(0, I_3)$, $\Gamma = diag(\gamma_v, \gamma_i, \gamma_e)$ and $b(Z(t),\theta) = (b_v(Z(t),\theta), b_i(Z(t),\theta), b_e(Z$
The log-likelihood of the discretized process is

$$\ell_n(\theta) = -\frac{1}{2}\left\{\sum_{i=1}^{n}(Z_i - Z_{i-1} - \Delta b(Z_{i-1},\theta))'(\Delta\Gamma\Gamma')^{-1}(Z_i - Z_{i-1} - \Delta b(Z_{i-1},\theta)) + n\log\right.$$

Because the diffusion coefficient is diagonal, this is equivalent to

$$\ell_n(\theta) = -\frac{1}{2}\sum_{k=v,i,e}\left\{\sum_{i=1}^{n}\frac{(Z_{k,i} - Z_{k,i-1} - \Delta b_k(Z_{k,i-1},\theta))^2}{\Delta\gamma_k^2} + n\log(2\pi\Delta\gamma_k^2)\right\}$$

The estimator is then defined as

$$\hat{\theta}_n = \arg\max \ell_n(\theta)$$

Theoretical results obtained for one-dimensional SDE can be generalized to the multidimensional case. In the neuronal models, the drift is linear in the parameters. Therefore, the optimization of the contrast is explicit.

5.2 Incomplete observations with autonomous hidden components

We consider the following elliptic system

$$dV(t) = -\left(g_{inh}(t)(V(t) - V_{inh}) + g_{exc}(t)(V(t) - V_{exc}) + I\right)dt + \gamma_v dB(t) \quad (5.4)$$

$$dg_{inh}(t) = -\frac{1}{\tau_{inh}}(g_{inh}(t) - \bar{g}_{inh})dt + \gamma_i dB_i(t) \quad (5.5)$$

$$dg_{exc}(t) = -\frac{1}{\tau_{exc}}(g_{exc}(t) - \bar{g}_{exc})dt + \gamma_e dB_e(t) \quad (5.6)$$

where $B(t)$, $B_i(t)$ and $B_e(t)$ are three independent Brownian motions.

The two hidden components are autonomous. Their solution is explicit (and thus their transition densities). For all $i \geq 1$, with step size Δ, we

have for both conductances:

$$g(t_{i+1}) = g(t_i)e^{-\Delta/\tau} + \bar{g}(1 - e^{-\Delta/\tau}) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \frac{\sigma^2\tau}{2}(1 - e^{-2\Delta/\tau})) \tag{5.7}$$

5.2.1 A Hidden Markov Model

Thus the discretization of the system leads to a hidden Markov model. Let us denote $X_i = (g_{inh}(t_i), g_{exc}(t_i))$ and $Y_i = V(t_i)$. (X_i) is a bidimensional Markov process given its explicit solution (5.7).

The standard definition of a HMM is the following:

Definition 5. *A hidden Markov model is a bivariate discrete time process $(Y_i, X_i)_{i\geq 0}$ where $(X_i)_{i\geq 0}$ is a Markov chain and conditional on (X_i), (Y_i) is a sequence of independent random variables such that the conditional distribution of Y_i only depends on X_i.*

This definition is sufficient in most cases (for example in genetics). But this is not the case for the neuronal system, because conditional on (X_i), Y_i is not a sequence of independent random variables. More precisely, the conditional distribution of Y_i depends on X_{i-1}, Y_{i-1}.

Thus we need a more general definition of HMM, given by Cappe, Moulines and Ryden.

Definition 6. *A HMM with not countable state space is defined as a bivariate Markov chain (X_i, Y_i) with only partial observations Y_i, whose transition kernel has a special structure: both the joint process (X_i, Y_i) and the marginal hidden chain (X_i) are Markovian.*

Y_i is not Markovian itself, but (Y_i, X_i) is Markovian. Thus (X_i, Y_i) is a HMM with this definition.

5.2.2 Likelihood function

We want to estimate the parameter θ by maximum likelihood of the approximate model, with likelihood

$$p_\Delta(V_{0:n}; \theta) = \int p(V_0, X_0; \theta) \prod_{i=1}^{n} p_\Delta(V_i, X_i | V_{i-1}, X_{i-1}; \theta) dX_{0:n}. \tag{5.8}$$

It corresponds to a pseudo-likelihood for the exact diffusion. The multiple integrals of equation (5.8) are difficult to handle and it is not possible to maximize the pseudo-likelihood directly.

A solution is to consider the statistical model as an incomplete data model. The observable vector $V_{0:n}$ is then part of a so-called complete vector $(V_{0:n}, X_{0:n})$, where $X_{0:n}$ has to be imputed. To maximize the likelihood of the complete data vector $(V_{0:n}, X_{0:n})$, we propose to use a stochastic version of the EM algorithm, namely the SAEM algorithm (Delyon et al, 1999).

5.2.3 EM algorithm

The EM algorithm (Dempster et al, 1977) is useful in situations where the direct maximization of the marginal likelihood $\theta \to p_\Delta(V_{0:n}\ ;\theta)$ is more difficult than the maximization of the conditional expectation of the complete likelihood $Q(\theta|\theta') = \mathbb{E}_\Delta\left[\log p_\Delta(V_{0:n}, X_{0:n};\theta)|V_{0:n};\theta'\right]$, where $p_\Delta(V_{0:n}, X_{0:n};\theta)$ is the likelihood of the complete data $(V_{0:n}, X_{0:n})$ of the approximate model and the expectation is under the conditional distribution of $X_{0:n}$ given $V_{0:n}$ with density $p_\Delta(X_{0:n}|V_{0:n};\theta')$. The EM algorithm is an iterative procedure: at the mth iteration, given the current value $\widehat{\theta}_{m-1}$,

[EM algorithm]

- *Iteration* 0*:* initialization of $\widehat{\theta}_0$
- *Iteration* $m \geq 1$*:*

E-Step: evaluation of $Q_m(\theta) = Q(\theta \,|\, \widehat{\theta}_{m-1})$

M-Step: update of $\widehat{\theta}_m$ by $\widehat{\theta}_m = \arg\max_{\theta\in\Theta} Q_m(\theta)$.

Let us describe the principle of the EM algorithm.

Proposition 5. *The sequence $(\widehat{\theta}_m)$ produced by the EM algorithm is such that for all $m \geq 0$, $L(\widehat{\theta}_m) \geq L(\widehat{\theta}_{m-1})$.*

Proof. Write the likelihood as

$$p(V_{0:n};\theta) = \frac{p(V_{0:n}, X_{0:n};\theta)}{p(X_{0:n}|V_{0:n};\theta)}$$

Taking the log, we obtain

$$\log p(V_{0:n};\theta) = \log p(V_{0:n}, X_{0:n};\theta) - \log p(X_{0:n}|V_{0:n};\theta)$$

Taking the conditional expectation, we have

$$\begin{aligned}
\mathbb{E}(\log p(V_{0:n};\theta)|V_{0:n},\theta') &= \mathbb{E}(\log p(V_{0:n}, X_{0:n};\theta)|V_{0:n},\theta') - \mathbb{E}(\log p(X_{0:n}|V_{0:n};\theta)|V_{0:n},\\
L_n(\theta) = \log p(V_{0:n};\theta) &= Q(\theta|\theta') - H(\theta|\theta')
\end{aligned}$$

Consider two successive values $\widehat{\theta}_{m-1}$ and $\widehat{\theta}_m$. We have

$$L_n(\widehat{\theta}_m) - L_n(\widehat{\theta}_{m-1}) = Q(\widehat{\theta}_m|\widehat{\theta}_{m-1}) - Q(\widehat{\theta}_{m-1}|\widehat{\theta}_{m-1}) - (H(\widehat{\theta}_m|\widehat{\theta}_{m-1}) - H(\widehat{\theta}_{m-1}|\widehat{\theta}_{m-1}))$$

By definition of the M-step, we have $Q(\widehat{\theta}_m|\widehat{\theta}_{m-1}) - Q(\widehat{\theta}_{m-1}|\widehat{\theta}_{m-1}) \geq 0$.
Then, we study the last term:

$$\begin{aligned}
H(\widehat{\theta}_m|\widehat{\theta}_{m-1}) - H(\widehat{\theta}_{m-1}|\widehat{\theta}_{m-1}) &= \int \log \frac{p(X_{0:n}|V_{0:n};\widehat{\theta}_m)}{p(X_{0:n}|V_{0:n};\widehat{\theta}_{m-1})} p(X_{0:n}|V_{0:n};\widehat{\theta}_{m-1}) dX_{0:n}\\
&\leq \log \int \frac{p(X_{0:n}|V_{0:n};\widehat{\theta}_m)}{p(X_{0:n}|V_{0:n};\widehat{\theta}_{m-1})} p(X_{0:n}|V_{0:n};\widehat{\theta}_{m-1}) dX_{0:n}\\
&= \log \int p(X_{0:n}|V_{0:n};\widehat{\theta}_m) dX_{0:n} = \log 1 = 0
\end{aligned}$$

using the Jensen inequality for a concave function φ: $\int \varphi(g(x))dx \leq \varphi(\int g(x)dx)$. Finally,

$$L_n(\widehat{\theta}_m) - L_n(\widehat{\theta}_{m-1}) \geq 0$$

□

At each iteration of the EM algorithm, the likelihood increases. It does not mean that the sequence converges to the maximum argument of the likelihood.

The convergence of the algorithm can be proved for an exponential family. More precisely, we assume:

(M1) The parameter space Θ is an open subset of $\mathbb{R}^p$. The complete likelihood $p_\Delta(V_{0:n}, X_{0:n};\theta)$ belongs to a curved exponential family,

i.e. $\log p_\Delta(V_{0:n}, X_{0:n}; \theta) = -\psi(\theta) + \langle S(V_{0:n}, X_{0:n}), \nu(\theta)\rangle$, where ψ and ν are two functions of θ, $S(V_{0:n}, X_{0:n})$ is known as the minimal sufficient statistic of the complete model, taking its value in a subset $\mathcal{S}$ of $\mathbb{R}^d$, and $\langle\cdot,\cdot\rangle$ is the scalar product on $\mathbb{R}^d$.

Assumption **(M1)** is satisfied by the Euler approximation of system (5.4). Indeed, we have

$$\log p(V_{1:n}, X_{1:n}) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(V_i - V_{i-1} + \Delta(g_{inh,i-1}(V_{i-1} - V_{inh}) + g_{exc,i-1}(V_{i-1} - V_{ex}}{\gamma_v^2}$$
$$-\frac{1}{2}\sum_{i=1}^{n} \frac{(g_{inh,i} - g_{inh,i-1} + \Delta(g_{inh,i-1} - \bar{g}_{inh}))^2}{\gamma_i^2} - \frac{1}{2}\sum_{i=1}^{n} \frac{(g_{exc,i} - g_{exc,i-1} + \Delta(g_{exc,i-1}}{\gamma_e^2}$$
$$-n\log 2\pi\gamma_V^2\gamma_i^2\gamma_e^2$$

Then, the sufficient statistics are

$$\sum_{i=1}^{n}(V_i - V_{i-1} + \Delta g_{inh,i-1}V_{i-1} + g_{exc,i-1}V_{i-1})^2, \sum_{i=1}^{n} g_{inh,i-1}^2, \sum_{i=1}^{n} g_{exc,i-1}^2$$
$$\sum_{i=1}^{n}(V_i - V_{i-1} + \Delta g_{inh,i-1}V_{i-1} + g_{exc,i-1}V_{i-1})g_{inh,i-1}, \sum_{i=1}^{n}(V_i - V_{i-1} + \Delta g_{inh,i-1}V_{i-1} +$$
$$\sum_{i=1}^{n}(g_{inh,i} - g_{inh,i-1} + \Delta g_{inh,i-1})^2, \sum_{i=1}^{n}(g_{inh,i} - g_{inh,i-1} + \Delta g_{inh,i-1}),$$
$$\sum_{i=1}^{n}(g_{exc,i} - g_{exc,i-1} + \Delta g_{exc,i-1})^2, \sum_{i=1}^{n}(g_{exc,i} - g_{exc,i-1} + \Delta g_{exc,i-1})$$

Under assumption (M1), the E-step reduces to the computation of $\mathbb{E}_\Delta\left[S(V_{0:n}, X_{0:n})|V\right.$

[EM algorithm for exponential family]

- *Iteration* 0*:* initialization of $\widehat{\theta}_0$
- *Iteration* $m \geq 1$*:*

E-Step: evaluation of $s_m = \mathbb{E}_\Delta\left[S(V_{0:n}, X_{0:n})|V_{0:n}; \widehat{\theta}_{m-1}\right]$

M-Step: update of $\widehat{\theta}_m$ by $\widehat{\theta}_m = \arg\max_{\theta\in\Theta}\left(-\psi(\theta) + \langle s_m, \nu(\theta)\rangle\right).$

Theorem 3 (Wu, 1983)**.** *Assume the complete data model belongs to the exponential family. Then all the limit points of the sequence $\widehat{\theta}_m$ generated by the EM algorithm are stationary points of $L_n(\theta)$ and $L(\widehat{\theta}_m)$ converges monodically to some value $L^* = L(\theta^*)$ for some stationary point θ^*.*

Furthermore, if $||\widehat{\theta}_m - \widehat{\theta}_{m-1}|| \to 0$ as $m \to \infty$, then $\widehat{\theta}_m$ converges to θ^* where θ^* is one of the stationary point of $L_n(\theta)$.

To implement the EM algorithm, we need to compute the conditional expectation of $X_{0:n}$ given $V_{0:n}$. This conditional distribution $p(X_{0:n}|V_{0:n})$ is called the filtering or smoothing distribution. For a linear Gaussian model, the filtering distribution can be exactly computed with the Kalman filter. Otherwise, the filtering distribution is generally not explicit and has to be approximated. This is the purpose of new particle filter techniques.

5.2.4 Kalman Filter

In this section, we will consider the case of a linear Gaussian model (which is not the case of model (5.4)). We consider here the Gaussian linear state-space model

$$\begin{aligned} Y_k &= B_k X_k + S V_k \\ X_{k+1} &= A_k X_k + R U_k \end{aligned}$$

where $(U_k)_{k\geq 1}$ and $(V_k)_{k\geq 1}$ are two independent Gaussian sequences such that $U_k \sim \mathcal{N}(0, I)$ and $V_k \sim \mathcal{N}(0, I)$. The matrices A_k, B_k, S, R are of appropriate dimensions and contain the parameters. In this section we assume the parameters are fixed to the current value of the EM algorithm.

We are interesting in computing the filtered distribution $p(X_k|Y_{1:k})$ for $k = 1, \ldots, n$, especially $\hat{X}_{k|1:k} = \mathbb{E}(X_k|Y_{1:k})$; and the predictive distribution $p(X_k|Y_{1:k-1})$ for $k = 1, \ldots, n$, especially $\hat{X}_{k|1:k-1} = \mathbb{E}(X_k|Y_{1:k-1})$. The following elementary lemma is instrumental in computing the predictive and filtered state estimator.

Lemma 1. *Let X and V be two independent Gaussian random vectors with $\mathbb{E}(X) = \mu_X$, $Cov(X) = \Sigma_X$, $\mathbb{E}(V) = 0$, and $Cov(V) = \Sigma_V$. Con-*

sider the model

$$Y = BX + V$$

where B is a deterministic matrix of appropriate dimensions. Assume that $B\Sigma_X B' + \Sigma_V$ is a full rank matrix. Then

$$\begin{aligned}\mathbb{E}(X|Y) &= \mathbb{E}(X) + Cov(X,Y)(Cov(Y))^{-1}(Y - \mathbb{E}(Y)) \\ &= \mu_X + \Sigma_X B'(B\Sigma_X B' + \Sigma_V)^{-1}(Y - B\mu_X)\end{aligned}$$

and

$$\begin{aligned}Cov(X|Y) &= Cov(X - \mathbb{E}(X|Y)) \\ &= \Sigma_X - \Sigma_X B'(B\Sigma_X B' + \Sigma_V)^{-1}B\Sigma_X\end{aligned}$$

Proof. Denote by $\hat{X} = \mathbb{E}(X) + Cov(X,Y)(Cov(Y))^{-1}(Y - \mathbb{E}(Y))$. We have

$$X - \hat{X} = X - \mathbb{E}(X) - Cov(X,Y)(Cov(Y))^{-1}(Y - \mathbb{E}(Y))$$

which implies that

$$Cov(X - \hat{X}, Y) = Cov(X,Y) - Cov(X,Y)(Cov(Y))^{-1}Cov(Y) = 0$$

The random vectors Y and $X - \hat{X}$ are jointly Gaussian (as linear transformation of a Gaussian multivariate random vector) and uncorrelated. Hence Y and $X - \hat{X}$ are also independent. Writing $X = \hat{X} + (X - \hat{X})$, $X - \hat{X}$ is independent of Y, $\hat{X}$ is $\sigma(Y)$-mesurable (as a linear combination of the components of Y), thus $\hat{X} = \mathbb{E}(X|Y)$. Moreover because $X - \hat{X}$ is independent of Y

$$\begin{aligned}Cov(X|Y) &= Cov((X - \hat{X})(X - \hat{X})'|Y) = Cov(X - \hat{X}) \\ &= \mathbb{E}((X - \hat{X})(X - \hat{X})') = \mathbb{E}((X - \hat{X})X') \\ &= \mathbb{E}((X - \mathbb{E}(X) - Cov(X,Y)(Cov(Y))^{-1}(Y - \mathbb{E}(Y))X')) \\ &= Cov(X) - Cov(X,Y)(Cov(Y))^{-1}\mathbb{E}((Y - \mathbb{E}(Y))X') \\ &= Cov(X) - Cov(X,Y)(Cov(Y))^{-1}Cov(X,Y)' \\ &= Cov(X) - Cov(X,Y)(Cov(Y))^{-1}Cov(X,Y)'\end{aligned}$$

□

Applying the formula of the lemma to the Gaussian linear morel, we obtain the following proposition.

Proposition 6. *The filtered mean and covariance matrices* $\hat{X}_{k|1:k} = \mathbb{E}(X_k|Y_{1:k})$ *and* $\Sigma_{k|k1:} = Cov(X_k|Y_{1:k})$*; and the predictive mean and covariance matrices* $\hat{X}_{k|1:k-1} = \mathbb{E}(X_k|Y_{1:k-1})$ *and* $\Sigma_{k|1:k-1} = Cov(X_k|Y_{1:k-1})$ *can be computed as follows, for* $k \geq 0$*:*

Filtering

$$\begin{aligned}\hat{X}_{k|1:k} &= \hat{X}_{k|1:k-1} + \Sigma_{k|1:k-1}B_k'(B_k\Sigma_{k|1:k-1}B_k' + SS')^{-1}(Y_k - B_k\hat{X}_{k|1:k-1}) \\ \Sigma_{k|1:k} &= \Sigma_{k|1:k-1} - \Sigma_{k|1:k-1}B_k'(B_k\Sigma_{k|1:k-1}B_k' + SS')^{-1}B_k\Sigma_{k|1:k-1}\end{aligned}$$

Prediction

$$\begin{aligned}\hat{X}_{k+1|1:k} &= A_k\hat{X}_{k|1:k} \\ \Sigma_{k+1|1:k} &= 1_k\Sigma_{k|1:k}1_k' + RR'\end{aligned}$$

Proof. To compute the filtered distribution, we apply the Lemma to

$$Y_k = B_kX_k + V_k$$

with $X_k \sim \mathcal{N}(\hat{X}_{k|1:k-1}, \Sigma_{k|1:k-1})$. The predictive moments correspond to the moments of

$$X_{k+1} = AX_k + RU_k$$

considering $X_k \sim \mathcal{N}(\hat{X}_{k|1:k}, \Sigma_{k|1:k})$ and $U_k \sim \mathcal{N}(0, I)$. □

Using Proposition 6, one can compute the conditional expectation of the sufficient statistics in the EM algorithm.

When the model is non-linear (as the neuronal model), it does not enter the family of linear Gaussian state space model. Extended versions of the Kalman filter have been proposed, where the drift and the diffusion coefficient are linearized with respect to X and V. But there is no guarantee on the accuracy of the estimators.

Therefore, stochastic approximation of the filtering distribution has to be considered.

5.2.5 Monte Carlo integration

In this section, we consider sampling the unknown sequence of states $X_0, \ldots, X_n$, conditionally on the observed sequence $Y_0, \ldots, Y_n$.

We want to compute an expectation, which is an integral. When the analytic evaluation is not possible, numerical integration can become time consuming (increase exponentially with the dimension of the problem).

Thus it is useful to consider other methods for evaluating integrals. There are methods that do not surfer so directly from the curse of dimensionality, as Monte-Carlo integration.

The idea is the following. The strong law of large numbers says that if $\xi^1, \xi^2, \ldots, \xi^N$ is a sequence of iid X-valued random variable with common probability distribution π, then the estimator

$$\hat{\pi}_N(f) = \frac{1}{N} \sum_{k=1}^{N} f(\xi^k)$$

converges almost surely to $\pi(f)$ for all π-integrable functions f.

Increasing the number N can render the approximation error arbitrarily small. If

$$\pi(|f|^2) = \int |f(x)|^2 \pi(dx) < \infty,$$

the central limit theorem shows that

$$\sqrt{N}(\hat{\pi}_N(f) - \pi(f)) \to^{\mathcal{L}}_{N\to\infty} \mathcal{N}(0, \sigma^2_N(\pi(f)))$$

where $\sigma^2_N(\pi, f) = \frac{1}{N} \sum_{i=1}^{N} (f(\xi^k) - \hat{\pi}_N(f))^2$.

This MC integration approximates the expectation under the distribution π, but requires to be able to simulate under π. This is not the case for the conditional distribution $\pi(X_{0:n}) = p(X_{0:n}|V_{0:n})$. Indeed, we have

$$\pi(X_{0:n}) = p(X_{0:n}|V_{0:n}) = \frac{p(X_{0:n}, V_{0:n})}{p(V_{0:n})}$$

where $p(X_{0:n}, V_{0:n}) = \prod_{i=1}^{n} p(X_i, V_i|X_{i-1}, V_{i-1})$ is known analytically but where the normalizing constant $p(V_{0:n} = \int p(X_{0:n}, V_{0:n}) dX_{0:n}$ is unknown. Thus, we need a method to simulate under any distribution π, even if π is not analytically known.

A first standard method is the Markov Chain Monte Carlo methodology.

5.2.6 Markov Chain Monte Carlo

This class of MCMC method relies on Markov-dependent simulations. This Markov has nothing to do with the hidden Markov model. Two main algorithms have been proposed Metropolis-Hastings and Gibbs. The idea is the following. Let $(\xi^k)_{k\geq 1}$ be a Markov-dependent sequence with stationary distribution π (meaning that if ξ^k is distributed with π, ξ^{k+1} is distributed with π). The ergodic theorem for Markov chains asserts that under suitable conditions

$$\hat{\pi}_N(f) = \frac{1}{N}\sum_{k=1}^{N} f(\xi^k)$$

is an estimate of $\mathbb{E}_\pi(f) = \int f(x)\pi(x)dx$.

We now present the Metropolis-Hastings algorithm. We use what is called a proposal distribution q.

[Metropolis-Hastings algorithm] Given ξ^k,

1. Generate $\xi \sim q(\xi^k; \cdot)$
2. Set

$$\xi^{k+1} = \begin{cases} \xi & \text{with probability } \alpha(\xi^k, \xi) = \min\left(1, \frac{\pi(\xi)q(\xi,\xi^k)}{\pi(\xi^k)q(\xi^k,\xi)}\right) \\ \xi^k \end{cases} \tag{5.9}$$

The quantity $\alpha(\xi^k, \xi)$ is often called the acceptance-ratio of the MH algorithm.

Proposition 7. *The chain $(\xi^k)_{k\geq 0}$ generated by the Metropolis-Hastings algorithm has π as stationary probability density function.*

Remarks

1. The simulation is realized under the proposal distribution q. So we don't need to know how to simulate under π.
2. The proposal distribution has to be chosen such that the candidate ξ is more likely to be accepted.

3. In the acceptance ratio, we do not need to know explicitly π but the ratio of $\pi(\xi)/\pi(\xi^k)$. This allows the algorithm to be used without knowing the normalizing constant.

This algorithm may be applied to hidden Markov models for simulating from the posterior distribution of $X_{0:n}$ given the observations $V_{0:n}$ because the conditional distribution $\pi(X_{0:n}) = p(X_{0:n}|V_{0:n})$ is known up to the normalizing factor $p(V_{0:n})$. Then, we have to chose a proposal distribution to generate a candidate, i.e. a new trajectory $\xi = X_{0:n}$

Two important classes of MH algorithms are now presented.

- The independent Metropolis-Hastings algorithm uses a proposal distribution q which is independent of the current value ξ^k. In the case of the filtering distribution of the neuronal model, one could consider $q(X^k_{0:n}, X_{0:n}) = p(X_{0:n})$. Then the acceptance ratio is equal to

$$\begin{aligned}
\alpha(X^k_{0:n}, X_{0:n}) &= \min\left(1, \frac{\pi(X_{0:n})q(X_{0:n}, X^k_{0:n})}{\pi(X^k_{0:n})q(X^k_{0:n}, X_{0:n})}\right) \\
&= \min\left(1, \frac{p(V_{0:n}|X_{0:n})p(X_{0:n})p(X^k_{0:n})}{p(V_{0:n}|X^k_{0:n})p(X^k_{0:n})p(X_{0:n})}\right) \\
&= \min\left(1, \frac{p(V_{0:n}|X_{0:n})}{p(V_{0:n}|X^k_{0:n})}\right)
\end{aligned}$$

  Intuitively, the transition in the MH algorithm is accomplished by generating an independent trajectory from the distribution of X (not knowing the V's), and then accepting this new trajectory by comparing the likelihood of this new trajectory comparing the ratio $\frac{p(V_{0:n}|X_{0:n})}{p(V_{0:n}|X^k_{0:n})}$). If the new trajectory is more likely given the observations V, then we accept it with a high probability. The main drawback of this algorithm is that the entire new trajectory has to be more likely to be accepted with probability 1. The acceptance rate may thus be very slow.

- The random walk Metropolis-Hastings algorithm proposes another option for the choice of the proposal of $q(\xi^k, \cdot)$: a local move around ξ. The idea is that by successive small jumps, the Markov chain

will actually explore the whole range of the target distribution. The most natural proposal is the Gaussian random walk proposal

$$q(X^k_{0:n}, X_{0:n}) = \mathcal{N}(X^k_{0:n}, \Sigma)$$

where Σ is a matrix to be chosen. Thus $q(X^k_{0:n}, X_{0:n}) = q(X_{0:n}, X^k_{0:n})$ and the acceptance probability is

$$\alpha(X^k_{0:n}, X_{0:n}) \;=\; \min\left(1, \frac{p(V_{0:n}|X_{0:n})p(X_{0:n}))}{p(V_{0:n}|X^k_{0:n})p(X^k_{0:n})}\right)$$

Depending on which scale Σ is chosen, the Markov chain may be very slow to converge either because it moves too cautiously (if the scale is too small) or too widely (if the scale is too large). The main drawback of this algorithm is that, as for independent MH algorithm, the the entire new trajectory has to be more likely to be accepted. The acceptance rate may thus be very slow.

When the distribution of interest π is multivariate, the MH algorithm may be slow to converge. We may prefer an algorithm allowing to accept only one component at each iteration, because "improving" one coordinate is more easy that improving the whole trajectory. This is the purpose of Gibbs sampling.

[Gibbs sampling] Starting from an initial arbitrary state $X^0_{0:n}$, update the current state $X^k_{0:n}$ to a new trajectory $X^{k+1}_{0:n}$ as follows

1. Simulate X^{k+1}_0 from $\pi_0(\cdot|X^k_{1:n}, V_{0:n})$
2. Simulate X^{k+1}_1 from $\pi_1(\cdot|X^{k+1}_0, X^k_{2:n}, V_{0:n})$
3. ...
4. Simulate X^{k+1}_i from $\pi_i(\cdot|X^{k+1}_{0:i-1}, X^k_{i+1:n}, V_{0:n})$
5. ...

where π_i is the i-th marginal distribution. Each iteration can be performed using a Metropolis-Hasting algorithm if the marginal distribution is not explicit.

Proposition 8. *Each of the n individual steps of the Gibbs sampler admits π as a stationary probability density function.*

In the case of the HMM, we have

[MH within Gibbs sampling for HMM] Starting from an initial arbitrary state $X_{0:n}^0$, update the current state $X_{0:n}^k$ to a new trajectory $X_{0:n}^{k+1}$ as follows

1. Simulate X_0^{k+1} from $\pi_0(\cdot|X_1^k, V_{0:1})$ with Metropolis-Hastings
2. ...
3. Simulate X_i^{k+1} from $\pi_i(\cdot|X_{i-1}^{k+1}, X_{i+1}^k, V_{i-1:i+1})$ with Metropolis-Hastings
4. ...

Even if this MH within Gibbs sampling has an acceptance rate which is greater than for the standard MH algorithm, the convergence may be very slow. Especially, it is very difficult to calibrate the n proposals for each Metropolis-Hastings algorithm. The mixing of the Markov chain can be very poor.

An alternative exists, which is specific to the case of time series and the filtering distribution. This is the particle filter, which is a stochastic version of the filtering. This algorithm is based on the importance sampling, which is now explained.

5.2.7 Importance sampling

Importance sampling aims at approximating integrals of the form $\pi(f) = \int f(x)\pi(x)dx$. The Monte-Carlo approach consists in drawing an iid sample $\xi^1, \ldots, \xi^N$ from the probability measure π and then evaluating the sample mean $\frac{1}{N}\sum_{k=1}^N f(\xi^k)$. This technique is applicable only when it is possible to sample from the target distribution π.

Importance sampling is based on the idea that it may be more easy to sample from an **instrumental distribution** q, and then to weight the samples using appropriate importance weights. More formally, the idea

is the following. Assume π is absolutely continuous with respect to any instrumental distribution q from which sampling is easily feasible. Then

$$\pi(f) = \int f(x)\pi(x)dx = \int f(x)\frac{\pi(x)}{q(x)}q(x)dx$$

So if $\xi^1, \ldots, \xi^N$ is an iid sample from q, the importance sampler is an estimator of $\pi(f)$:

$$\hat{\pi}^{IS}_{q,N}(f) = \frac{1}{N}\sum_{k=1}^{N} f(\xi^k)\frac{\pi(\xi^k)}{q(\xi^k)}$$

The strong law of large number implies that $\hat{\pi}_{q,N}(f)$ converges to $\pi(f)$ almost surely as N tends to infinity.

However, the target distribution π or the instrumental distribution q are generally known only up to a normalizing factor. This is the case for HMM and the filtering problem. It is possible to use the importance sampling approach by adopting a self-normalized from of the importance sampling estimator:

$$\tilde{\pi}_{q,N}(f) = \frac{\sum_{k=1}^{N} f(\xi^k)\frac{\pi(\xi^k)}{q(\xi^k)}}{\sum_{k=1}^{N}\frac{\pi(\xi^k)}{q(\xi^k)}}$$

This estimator is the ratio of the sample means of the functions $f_1 = f\pi/q$ and $f_2 = \pi/q$. The strong law of large numbers implies that $\frac{1}{N}\sum_{k=1}^{N} f_1(\xi^k)$ and $\frac{1}{N}\sum_{k=1}^{N} f_2(\xi^k)$ converges almost surely to $\pi(f_1)$ and $q(\pi/q) = 1$. Thus $\tilde{\pi}_{q,N}(f)$ is a consistent estimator of $\pi(f)$.

An extension to importance sampling is the Sampling Importance Sampling method, which allows to obtain a sample from the (approximate) distribution π. SIR is a two-stage procedure in which importance sampling is followed by an additional random sampling step.

In IS, an iid sample $(\tilde{\xi}^1, \ldots, \tilde{\xi}^N)$ is drawn from the instrumental distribution q and associated to normalized importance weights

$$\omega^k = \frac{\frac{\pi(\tilde{\xi}^k)}{q(\tilde{\xi}^k)}}{\sum_{j=1}^{N}\frac{\pi(\tilde{\xi}^j)}{q(\tilde{\xi}^j)}}$$

In SIR, the idea is that points $\tilde{\xi}^k$ for which ω^k is large are most likely under the target distribution π and should be selected with higher probability when one wants to draw a sample under the (approximate) distribution π. Therefore, the resampling stage consists of drawing a sample of size M denoted $(\xi^1, \ldots, \xi^M)$ from the intermediate set of points $(\tilde{\xi}^1, \ldots, \tilde{\xi}^N)$ with replacement with probability of sampling $\tilde{\xi}^k$ equal to the importance weight ω^k.

[Sampling Importance Resampling]

1. Sampling: Draw a sample $\tilde{\xi}^1, \ldots, \tilde{\xi}^N$ from the instrumental distribution q
2. Weighting: Compute the normalized importance weights

$$\omega^k = \frac{\frac{\pi(\tilde{\xi}^k)}{q(\tilde{\xi}^k)}}{\sum_{j=1}^N \frac{\pi(\tilde{\xi}^j)}{q(\tilde{\xi}^j)}}$$

3. Resampling: Draw M discrete random variables $(I^1, \ldots, I^M)$ with a multinomial distribution with probability $(\omega^1, \ldots, \omega^M)$, i.e.

$$P(I^1 = j) = \omega^j, \quad j = 1, \ldots, M$$

Set, for $j = 1, \ldots, M$, $\xi^j = \tilde{\xi}^{I_j}$.

The resampling step might be seen as a means to transform the weighted importance sampling estimate $\hat{\pi}_{q,N}^{IS}(f)$ into an unweighted sample average. If N^j is the number of times the element $\tilde{\xi}^{I_j}$ is resampled, we have

$$\hat{\pi}_{q,M}^{SIR} = \frac{1}{M} \sum_{j=1}^{M} f(\xi^j) = \sum_{j=1}^{N} \frac{N^j}{N} f(\tilde{\xi}^j)$$

It is easily seen that $\hat{\pi}_{q,M}^{SIR}$ is, conditionally on $(\tilde{\xi}^1, \ldots, \tilde{\xi}^N)$, equal to the importance sampling estimator $\hat{\pi}_{q,N}^{IS}(f)$

$$\mathbb{E}(\hat{\pi}_{q,N}^{SIR}(f) | \tilde{\xi}^1, \ldots, \tilde{\xi}^N) = \hat{\pi}_{q,N}^{IS}(f)$$

As a consequence, the SIR estimator is an unbiased estimator of $\pi(f)$.

But for both the IS and SIR estimators, the normalizing constants has to be known. Moreover the proposal q for the whole trajectory is difficult to choose. We now present the case of filtering.

5.2.8 Filtering problem

We consider the HMM defined by the neuronal model. We want to define the filtering and smoothing distribution. We have

$$\pi_{0|0}(f) = \mathbb{E}(f(X_0)|V_0) = \frac{\int f(x_0)p(y_0|x_0)p(x_0)dx_0}{\int p(y_0|x_0)p(x_0)dx_0}$$

Then recursively,

$$\pi_n f \quad = \quad = \frac{\int p(X_0)\prod_{i=1}^n p_\Delta(V_i, X_i|V_{i-1}, X_{i-1})f(U_n)dX_{0:n}}{\int p(X_0)\prod_{i=1}^n p_\Delta(V_i, X_i|V_{i-1}, X_{i-1})dX_{0:n}}. \quad (5.10)$$

We introduce for $i = 1, \ldots, n$ the kernels H_i from $\mathbb{R}$ into itself by

$$H_i f(x) = \int p_\Delta(V_i, y|V_{i-1}, x)f(y)dy. \quad (5.11)$$

Then π_n can be expressed recursively by

$$\pi_n f \quad = \quad \frac{\pi_{n-1}H_n f}{\pi_{n-1}H_n 1} \quad (5.12)$$

Note that the denominator of (5.18) is $\mu H_1 \cdots H_n 1 = p_\Delta(V_{0:n})$, which is different from 0 since it has support the real line. Thus, the filtering problem is well-posed.

5.2.9 Particle Filter

We call particle filter or sequential Monte Carlo (SMC) any stochastic method which approximates a filter distribution.

The SMC algorithm provides a set of K particles $(X_{0:n}^{(k)})_{k=1\ldots K}$ and weights $(W_{0:n}^{(k)})_{k=1\ldots K}$ approximating the conditional smoothing distribution $p_\Delta(X_{0:n}|V_{0:n})dX_{0:n}$ (see Doucet et al., 2001). The SMC method relies on proposal distributions $q(X_i|V_i, V_{i-1}, X_{i-1})$ to sample what we call particles from these distributions. We write $V_{0:i} = (V_0, \ldots, V_i)$ and likewise for $X_{1:i}$.

[SMC algorithm]

- *At time* $i = 0$*:* $\forall\, k = 1, \ldots, K$
  1. sample $X_0^{(k)}$ from $p(X_0|V_0)$
  2. compute and normalize the weights: $w_0\left(X_0^{(k)}\right) = p\left(V_0, X_0^{(k)}\right)$, $W_0\left(X_0^{(k)}\right) = \frac{w_0\left(X_0^{(k)}\right)}{\sum_{k=1}^K w_0\left(X_0^{(k)}\right)}$
- *At time* $i = 1, \ldots, n$*:* $\forall\, k = 1, \ldots, K$
  1. Sample the indices $A_{i-1}^{(k)} \sim r(\cdot|W_{i-1}(X_{1:i-1}))$ and set $X_{1:i-1}^{'(k)} = X_{1:i-1}^{(A_{i-1}^{(k)})}$
  2. sample $X_i^{(k)} \sim q\left(\cdot|X_{i-1}^{'(k)}\right)$ and set $X_{0:i}^{(k)} = (X_{1:i-1}^{'(k)}, X_i^{(k)})$
  3. compute and normalize the weights $W_i(X_{0:i}^{(k)}) = \frac{w_i\left(X_{0:i}^{(k)}\right)}{\sum_{k=1}^K w_i\left(X_{0:i}^{(k)}\right)}$ with $w_i\left(X_{0:i}^{(k)}\right) = \frac{p_\Delta\left(V_i, X_i^{(k)}|V_{0:i-1}, X_{0:i-1}^{(k)}\right)}{q\left(X_i^{(k)}|X_{i-1}^{'(k)}\right)} = \frac{p_\Delta\left(V_i|V_{0:i-1}, X_{0:i-1}^{(k)}\right) p_\Delta\left(X_i^{(k)}|X_{i-1}^{(k)}\right)}{q\left(X_i^{(k)}|X_{i-1}^{'(k)}\right)}$

Thus the SMC is a Sampling Importance Resampling approach, with iterative simulation along time. It can be proved that the sample $(X_n^{(k)})_{k\geq 0}$ is approximately distributed with $p_\Delta(X_n|V_{0:n})$.

The choice of the kernel q is very important to allow a "quick" convergence of the algorithm. The transition density, also called the prior distribution $p(X_i|X_{i-1})$, is a natural proposal. However, the transition density may not be optimal, if the observations are unlikely with the hidden states. The proposal should be close to the optimal distribution $p(X_i|X_{i-1}, V_{i-1})$.

The resampling step allows to avoid the problem of weight degeneracy. If there are too many ineffective particles, the particle approximation becomes both computationally and statistically inefficient: most of the computing effort is put on updating particles and weights that do not contribute significantly to the estimator.

The resampling allows to reduce the degeneracy of the importance weights. The trajectories with small importance weights are eliminated, whereas those with large importance weights are duplicated.

5.2.10 Deviation inequality

For a bounded Borel function f, denote $\Psi_n^K f = \sum_{k=1}^K f(X_n^{(k)}) W_{n,\theta}(X_{0:n}^{(k)})$, the conditional expectation of f under the empirical measure $\Psi_{n,\theta}^K$ obtained by the SMC algorithm for a given value of θ. We have:

Proposition 9. *Under assumption (SMC3), for any $\varepsilon > 0$, and for any bounded Borel function f on $\mathbb{R}$, there exist constants C_1 and C_2, independent of θ, such that*

$$\mathbb{P}\left(\left|\Psi_{n,\theta}^K f - \pi_{n,\theta} f\right| \geq \varepsilon\right) \;\leq\; C_1 \exp\left(-K \frac{\varepsilon^2}{C_2 \|f\|^2}\right) \tag{5.13}$$

where $\|f\|$ is the sup-norm of f.

The proof is provided in Appendix **??**. A similar result can be obtained with respect to the exact smoothing distribution of the exact diffusion model, under assumptions on the number of particles and the step size of the Euler approximation.

5.2.11 SAEM algorithm

When the conditional expectation of the EM algorithm has no closed form, we simulate a sample from the (approximated) distribution $X_{0:n}|V_{0:n}$ and then approximate the expectation by an empirical mean (Importance sampling estimator or MCMC).

This is the idea of the Stochastic Approximation EM algorithm (SAEM) replacing the E-step by a stochastic approximation of $Q_m(\theta)$. The E-step is then divided into a simulation step (S-step) of the non-observed data $(X_{0:n}^{(m)})$ with the conditional density $p_\Delta(X_{0:n}\,|V_{0:n}; \widehat{\theta}_{m-1})$ and a stochastic approximation step (SA-step) of $\mathbb{E}_\Delta\left[S(V_{0:n}, X_{0:n})|V_{0:n}; \widehat{\theta}_{m-1}\right]$ with a sequence of positive numbers $(a_m)_{m\in\mathbb{N}}$ decreasing to zero. We write s_m for the approximation of this expectation. Iterations of the SAEM algorithm are written as follows:

[SAEM algorithm]

- *Iteration* 0*:* initialization of $\widehat{\theta}_0$ and set $s_0 = 0$.
- *Iteration* $m \geq 1$*:*

S-Step: simulation of the non-observed data $(X_{0:n}^{(m)})$ targeting the distribution $p_\Delta(X_{0:n}|V_{0:n};\widehat{\theta}_{m-1})dX_{0:n}$.

SA-Step: update s_{m-1} using the stochastic approximation:

$$s_m = s_{m-1} + a_{m-1}\left[S(V_{0:n}, X_{0:n}^{(m)}) - s_{m-1}\right] \tag{5.14}$$

M-Step: update of $\widehat{\theta}_m$ by $\widehat{\theta}_m = \underset{\theta\in\Theta}{\arg\max}\left(-\psi(\theta) + \langle s_m, \nu(\theta)\rangle\right).$

Convergence of SAEM We introduce a set of convergence assumptions which are the classic ones for the SAEM algorithm (Delyon et al., 1999).

(M2) The functions $\psi(\theta)$ and $\nu(\theta)$ are twice continuously differentiable on Θ.

(M3) The function $\bar{s} : \Theta \longrightarrow \mathcal{S}$ defined by $\bar{s}(\theta) = \int S(v,x)p_\Delta(x|v;\theta)dv\,dx$ is continuously differentiable on Θ.

(M4) The function $\ell_\Delta(\theta) = \log p_\Delta(v,x,\theta)$ is continuously differentiable on Θ and $\partial_\theta \int p_\Delta(v,x;\theta)dv\,dx = \int \partial_\theta p_\Delta(v,x;\theta)dv\,dx$.

(M5) Define $L : \mathcal{S}\times\Theta \to \mathbb{R}$ by $L(s,\theta) = -\psi(\theta) + \langle s, \nu(\theta)\rangle$. There exists a function $\hat{\theta} : \mathcal{S} \to \Theta$ such that $\forall\theta\in\Theta$, $\forall s \in \mathcal{S}$, $L(s,\hat{\theta}(s)) \geq L(s,\theta)$.

(SAEM1) The positive decreasing sequence of the stochastic approximation $(a_m)_{m\geq 1}$ is such that $\sum_m a_m = \infty$ and $\sum_m a_m^2 < \infty$.

(SAEM2) $\ell_\Delta : \Theta \to \mathbb{R}$ and $\hat{\theta} : \mathcal{S} \to \Theta$ are d times differentiable, where d is the dimension of $S(v,x)$.

(SAEM3) For all $\theta\in\Theta$, $\int ||S(v,x)||^2 p_\Delta(x|v;\theta)dx < \infty$ and the function $\Gamma(\theta) = Cov_\theta(S(\cdot, X_{0:n}))$ is continuous, where the covariance is under the conditional distribution $p_\Delta(X_{0:n}|V_{0:n};\theta)$.

(SAEM4) Let $\{\mathcal{F}_m\}$ be the increasing family of σ-algebras generated by the random variables $s_0, X_{0:n}^{(1)}, X_{0:n}^{(2)}, \ldots, X_{0:n}^{(m)}$. For any positive Borel function f, $\mathbb{E}_\Delta(f(X_{0:n}^{(m+1)})|\mathcal{F}_m) = \int f(x)p_\Delta(x|v, \widehat{\theta}_m)dx$.

Assumptions (M1)-(M5) ensure the convergence of the EM algorithm when the E-step is exact (Delyon et al., 1999). Assumptions (M1)-(M5) and (SAEM1)-(SAEM4) together with the additional assumption that $(s_m)_{m\geq 0}$ takes its values in a compact subset of $\mathcal{S}$ ensure the convergence of the SAEM estimates to a stationary point of the observed likelihood $p_\Delta(V_{0:n};\theta)$ when the simulation step is exact (Delyon et al., 1999).

Theorem 4. *Assume that (M1)-(M5), (SAEM1)-(SAEM3) hold. Then, with probability 1,* $\lim_{m\to\infty} d(\widehat{\theta}_m, \mathcal{L}) = 0$ *where* $\mathcal{L} = \{\theta \in \Theta, \partial_\theta \ell_\Delta(\theta) = 0\}$ *is the set of stationary points of the log-likelihood* $\ell_\Delta(\theta) = \log p_\Delta(V_{0:n};\theta)$.

We have seen two versions to realize the simulation step: either by MCMC or by SMC. The convergence proof is not the same.

SAEM and MCMC When the simulation of the non-observed vector $X_{0:n}$ cannot be directly performed, Kuhn and Lavielle Kuhn and Lavielle (2005) propose to combine this algorithm with a Markov Chain Monte-Carlo (MCMC) procedure. The convergence of the SAEM-MCMC algorithm is ensured under the following additional assumption:

Assumption (**MCMC1**):

1. For any compact subset V of Θ, there exists a real constant L such that for any (θ, θ') in V^2
$$\sup_{\{X_{0:n}, X'_{0:n}\}\in\mathcal{E}} |\Pi_\theta(X'_{0:n}|X_{0:n}) - \Pi_{\theta'}(X'_{0:n}|X_{0:n})| \leq L\|\theta - \theta'\|.$$
2. The transition probability Π_θ supplies an uniformly ergodic chain whose invariant probability is the conditional distribution $p(X_{0:n}|V_{0:n};\theta)$, i.e.
$$\exists K_\theta \in \mathbb{R}^+, \quad \exists \rho_\theta \in]0,1[\quad | \quad \forall \ell \in \mathbb{N} \quad \|\Pi_\theta^\ell(\cdot|X_{0:n}) - p(\cdot,\cdot|V_{0:n};\theta)\|_{TV} \leq C_\theta \rho_\theta^\ell,$$
where $\|\cdot\|_{TV}$ is the total variation norm. Furthermore,
$$C = \sup_{\theta\in\Theta} C_\theta < \infty \quad \text{and} \quad \rho = \sup_{\theta\in\Theta} \rho_\theta < 1.$$

3. Function S is bound on $\mathcal{E}$.

Theorem 5. *Assume that (M1)-(M5), (SAEM1)-(SAEM3) and (MCMC1) hold. Then, with probability 1,* $\lim_{m\to\infty} d(\widehat{\theta}_m, \mathcal{L}) = 0$ *where* $\mathcal{L} = \{\theta \in \Theta, \partial_\theta \ell_\Delta(\theta) = 0\}$ *is the set of stationary points of the log-likelihood* $\ell_\Delta(\theta) = \log p_\Delta(V_{0:n}; \theta)$.

SAEM and SMC When we combine the SAEM algorithm with SMC, we need some additional assumptions because the simulation step is not exact. We have three additional assumptions on the SMC algorithm to bound the error induced by this algorithm and prove the convergence of the SAEM-SMC algorithm.

(SMC1) The number of particles K used at each iteration of the SAEM algorithm varies along the iteration: there exists a function $g(m) \to \infty$ when $m \to \infty$ such that $K(m) \geq g(m)\log(m)$.

(SMC2) The function S is bounded uniformly in u.

(SMC3) The functions $p_\Delta(V_i|X_i, X_{i-1}, X_{i-1}; \theta)$ are bounded uniformly in θ.

Theorem 6. *Assume that (M1)-(M5), (SAEM1)-(SAEM3) and (SMC1)-(SMC3) hold. Then, with probability 1,* $\lim_{m\to\infty} d(\widehat{\theta}_m, \mathcal{L}) = 0$ *where* $\mathcal{L} = \{\theta \in \Theta, \partial_\theta \ell_\Delta(\theta) = 0\}$ *is the set of stationary points of the log-likelihood* $\ell_\Delta(\theta) = \log p_\Delta(V_{0:n}; \theta)$.

5.3 Incomplete observations with non autonomous hidden components

We consider the elliptic Morris-Lecar model

$$\begin{aligned} dV(t) &= -\left(g_{fast}\, m_\infty(t)\,(V(t)-V_{fast}) + g_{slow}U(t)(V(t)-V_{slow}) + g_L(V(t)-V_L) + \right. \\ dU(t) &= (\alpha(V(t))(1-U(t)) - \beta(V(t))U(t))\,dt + \sigma(V(t),U(t))dB(t) \end{aligned}$$

where $\tilde{B}(t)$ and $B(t)$ are two independent Brownian motions.

The aim is to estimate θ by maximum likelihood. However, this likelihood is intractable, as the transition density of model (**??**) is not explicit. Let Δ denote the step size between two observation times, which we for simplicity assume does not depend on i. The extension to unequally spaced observation times is straightforward. The Euler-Maruyama approximation of model (**??**) leads to a discretized model defined as follows

$$\begin{aligned} V_{i+1} &= V_i + \Delta f(V_i,U_i) + \sqrt{\Delta}\,\gamma\,\tilde{\eta}_i, \\ U_{i+1} &= U_i + \Delta b(V_i,U_i) + \sqrt{\Delta}\sigma(V_i,U_i)\eta_i, \end{aligned} \tag{5.15}$$

where $(\tilde{\eta}_i)$ and (η_i) are independent centered Gaussian variables. To ease readability the same notation (V_i,U_i) is used for the original and the approximated processes. This should not lead to confusion, as long as the transition densities are distinguished, as done below.

5.3.1 Property of the observation model

The observation model is a degenerate HMM. Let us recall the definition proposed by Cappé et al. (2005): A HMM with not countable state space is defined as a bivariate Markov chain (X_i,Y_i) with only partial observations Y_i, whose transition kernel has a special structure: both the joint process (X_i,Y_i) *and* the marginal hidden chain (X_i) are Markovian.

In our model, (U_i) is not Markovian, only (V_i,U_i) is Markovian. So set $X_i=(V_i,U_i)$, with Markov kernel $R(X_{i-1},dX_i) = p_\Delta(dV_i,dU_i|V_{i-1},X_{i-1})$, the transition density of model (5.15), and $Y_i = X_i^{(1)}$, the first coordinate of X_i with transition kernel $F(X,dY) = \mathbf{1}_{Y=X^{(1)}}$. Here, $\mathbf{1}_x$ is the Dirac measure in x. Thus, the kernel F is zero almost everywhere and

the HMM is degenerate. This leads to an intrinsic degeneracy of the particle filter used in the standard HMM toolbox, as explained below.

Therefore we consider the observation model as a bivariate Markov chain (V_i, U_i) with only partial observations V_i whose hidden coordinate U_i is *not* Markovian. It is not a HMM but a general dynamic model as considered by Andrieu et al. (2001). The hidden process U_i is distributed as

$$U_0 \sim \mu(dU_0),\ U_i|(U_{0:i-1}, V_{0:i-1}) \sim K(dU_i|U_{0:i-1}, V_{0:i-1})$$

for some conditional distribution function K and the observed process V_i is distributed as

$$V_i|(U_{0:i}, V_{0:i-1}) \sim G(dV_i|U_{0:i}, V_{0:i-1})$$

for some distribution function G. Given the Markovian structure of the pair (V_i, U_i), we have $K(dU_i|U_{0:i-1}, V_{0:i-1}) = K(dU_i|X_{i-1}, V_{i-1})$ and $G(dV_i|\ U_{0:i}, V_{0:i-1}) = G(dV_i|U_{i-1:i}, V_{i-1})$. To simplify, we use the same notation for random variables and their realizations and assume that $G(dV_i|U_{0:i},\ V_{0:i-1}) = G(V_i|U_{0:i}, V_{0:i-1})dV_i$.

5.3.2 Likelihood function

We want to estimate the parameter θ by maximum likelihood of the approximate model, with likelihood

$$p_\Delta(V_{0:n};\theta) = \int p(V_0, U_0;\theta) \prod_{i=1}^{n} p_\Delta(V_i, U_i|V_{i-1}, U_{i-1};\theta) dU_{0:n}. \qquad (5.16)$$

It corresponds to a pseudo-likelihood for the exact diffusion. The multiple integrals of equation (5.16) are difficult to handle and it is not possible to maximize the pseudo-likelihood directly.

To maximize the likelihood of the complete data vector $(V_{0:n}, U_{0:n})$, we propose to use a stochastic version of the EM algorithm, namely the SAEM algorithm (Delyon et al., 1999). Simulation under the smoothing distribution $p_\Delta(U_{0:n}\ |V_{0:n};\theta)dU_{0:n}$ is performed with a SMC algorithm, also known as Particle Filtering. We have adapted this algorithm to handle a coupled two-dimensional SDE, i.e. the unobserved coordinate is non-autonomous and non-Markovian.

5.3.3 The filtering problem and the SMC algorithm

For any bounded Borel function $f : \mathbb{R} \mapsto \mathbb{R}$, we denote $\pi_{n,\theta} f = \mathbb{E}_\Delta\left(f(U_n)|V_{0:n};\theta\right)$, the conditional expectation under the exact smoothing distribution $p_\Delta(U_{0:n}|V_{0:n};\theta)$ of the approximate model. The aim is to approximate this distribution for a fixed value of θ. When included in the stochastic EM algorithm, this value will be the current value $\widehat{\theta}_m$ at the given iteration. For notational simplicity, θ is omitted in the rest of this Section.

We now argue why the HMM point of view is ill-posed for the filtering problem. Considering the model as a HMM, $X_i = (V_i, U_i)$ is the hidden Markov chain and $Y_i = X_i^{(1)}$. But then the filtering problem $\pi_n f$ is the ratio of $\int \mu(dU_0)R(X_0, dX_1)F(X_0;Y_1)\cdots R(X_{n-1}, dX_n)F(X_{n-1};Y_n)f(X_n)$ and $\int \mu(dU_0)\, R(X_0, dX_1)\, F(X_0;Y_1)\cdots R(X_{n-1}, dX_n)\, F(X_{n-1};Y_n)$. Since $F(X_{n-1};Y_n) = \mathbf{1}_{Y_n = X_{n-1}^{(1)}}$ and the state space is continuous, the denominator is zero almost surely and the filtering problem is ill-posed.

Now consider the model in a more general framework with the hidden state U_i not Markovian, and introduce for $i = 1,\ldots,n$ the kernels H_i from $\mathbb{R}$ into itself by

$$H_i f(u) = \int K(dz|V_{i-1},u)G(V_i|u,V_{i-1},z)f(z)dz = \int p_\Delta(V_i,y|V_{i-1},u)f(z)dz. \tag{5.17}$$

Then π_n can be expressed recursively by

$$\pi_n f \;=\; \frac{\pi_{n-1}H_n f}{\pi_{n-1}H_n 1} = \frac{\int \mu(U_0)\prod_{i=1}^n p_\Delta(V_i,U_i|V_{i-1},U_{i-1})f(U_n)dU_{0:n}}{\int \mu(U_0)\prod_{i=1}^n p_\Delta(V_i,U_i|V_{i-1},U_{i-1})dU_{0:n}} \tag{5.18}$$

Note that the denominator of (5.18) is $\mu H_1\cdots H_n 1 = p_\Delta(V_{0:n})$, which is different from 0 since it has support the real line. Thus, the filtering problem is well-posed.

The kernels H_i are extensions of the kernels considered by Del Moral et al. (2001) in the context of two-dimensional SDEs with hidden coordinate U_t autonomous (and thus Markovian). We do not extend their particle filter since it is based on simulation of both V_i and U_i with transition kernel $p_\Delta(V_i,U_i|V_{i-1},U_{i-1})$. They avoid the degeneracy of the weights by introducing an instrumental function ψ and the weights are computed as $\psi(V_i^{(k)} - V_i)$. The choice of this instrumental function may influence the numerical properties of the filter. Therefore, we adopt the

general filter proposed by Andrieu et al. (2001) for more general dynamic system, that we recall here.

The SMC is the following.

[SMC algorithm]

- *At time* $i = 0$*:* $\forall\, k = 1, \ldots, K$
  1. sample $U_0^{(k)}$ from $p(U_0|V_0)$
  2. compute and normalize the weights: $w_0\left(X_{0:k}\right) = p\left(V_0, X_{0:k}\right)$, $W_0\left(X_{0:k}\right) = \frac{w_0(X_{0:k})}{\sum_{k=1}^K w_0(X_{0:k})}$
- *At time* $i = 1, \ldots, n$*:* $\forall\, k = 1, \ldots, K$
  1. Sample the indices $A_{i-1}^{(k)} \sim r(\cdot|W_{i-1}(U_{0:i-1}))$ and set $U_{0:i-1}^{'(k)} = U_{0:i-1}^{(A_{i-1}^{(k)})}$
  2. sample $U_i^{(k)} \sim q\left(\cdot|V_{i-1:i}, U_{i-1}^{'(k)}\right)$ and set $U_{0:i}^{(k)} = (U_{0:i-1}^{'(k)}, U_i^{(k)})$
  3. compute and normalize the weights $W_i(U_{0:i}^{(k)}) = \frac{w_i\left(U_{0:i}^{(k)}\right)}{\sum_{k=1}^K w_i\left(U_{0:i}^{(k)}\right)}$ with $w_i\left(U_{0:i}^{(k)}\right) = \frac{p_\Delta\left(V_{0:i}, U_{0:i}^{(k)}\right)}{p_\Delta\left(V_{0:i-1}, U_{0:i-1}^{'(k)}\right) q\left(U_i^{(k)}|V_{i-1:i}, U_{0:i-1}^{'(k)}\right)}$

Finally, the SMC algorithm provides an empirical measure $\Psi_n^K = \sum_{k=1}^K W_n(U_{0:n}^{(k)}) \mathbf{1}_{U_{0:n}^{(k)}}$ which is an approximation to the smoothing distribution $p_\Delta(U_{0:n}|V_{0:n})dU_{0:n}$. A draw from this distribution can be obtained by sampling an index k from a multinomial distribution with probabilities W_n and setting the draw $U_{0:n}$ equal to $U_{0:n} = X_{0:n}^{(k)}$.

The variable $A_{i-1}^{(k)}$ plays an important role to discard the samples with small weights and multiply those with large weights (Gordon et al., 1993). It generates a number of offspring $N_{i-1}^{(\ell)}$, $\ell = 1, \ldots, K$, such that $\sum_{\ell=1}^K N_{i-1}^{(\ell)} = K$ and $\mathbb{E}(N_{i-1}^{(\ell)}) = K\, W_{i-1}(U_{0:i-1}^{(l)})$. Many schemes r have been presented in the literature, including multinomial sampling (Gordon et al., 1993), residual sampling (Liu and Chen, 1998) or stratified resampling (Doucet et al., 2000). They differ in terms of $var(N_{i-1}^{(\ell)})$ (see

Doucet et al., 2000). The key property that we need in order to prove the deviation inequality is that $\mathbb{E}(\mathbf{1}_{A_{i-1}^{(k)}=\ell}) = W_{i-1}(U_{0:i-1}^{(l)})$.

Since our model is not a HMM, the weights $w_i\left(U_{0:i}^{(k)}\right)$ cannot be written in terms of a Markov transition kernel of the hidden path as is usually done. It follows that the proposal q, which is crucial to ensure good convergence properties, has to depend on V_i. The first classical choice of q is $q(U_i|V_{i-1:i}, U_{i-1}) = p_\Delta(U_i|V_{i-1}, U_{i-1})$, i.e. the transition density. In this case, the weight reduces to $w_i\left(U_{0:i}^{(k)}\right) = p_\Delta(V_i|V_{i-1}, U_{0:i}^{(k)})$. A second choice for the proposal is $q(U_i|V_{i-1:i}, U_{i-1}) = p_\Delta(U_i|V_{i-1:i}, U_{i-1})$, i.e. the conditional distribution. In this case, the weight reduces to $w_i\left(U_{0:i}^{(k)}\right) = p_\Delta(V_i|V_{i-1}, U_{0:i-1}^{(k)})$. Transition densities and conditional distributions are detailed in Appendix **??**. When the two Brownian motions are independent, as we assume, the two choices are equivalent.

5.3.4 Deviation inequality

In the literature, deviation inequalities for SMC algorithms only appear for HMM. To our knowledge, this is the first non-asymptotic result proposed for a SMC applied to a non-Markovian hidden path. The only result of this type with SDEs has been obtained by Del Moral et al. (2001), with autonomous second coordinate. Here, we generalize their deviation inequality to a non-autonomous hidden path.

For a bounded Borel function f, denote $\Psi_n^K f = \sum_{k=1}^K f(U_n^{(k)}) W_{n,\theta}(U_{0:n}^{(k)})$, the conditional expectation of f under the empirical measure $\Psi_{n,\theta}^K$ obtained by the SMC algorithm for a given value of θ. We have:

Proposition 10. *Under assumption (SMC3), for any $\varepsilon > 0$, and for any bounded Borel function f on $\mathbb{R}$, there exist constants C_1 and C_2, independent of θ, such that*

$$\mathbb{P}\left(\left|\Psi_{n,\theta}^K f - \pi_{n,\theta} f\right| \geq \varepsilon\right) \;\leq\; C_1 \exp\left(-K \frac{\varepsilon^2}{C_2 \|f\|^2}\right) \tag{5.19}$$

where $\|f\|$ is the sup-norm of f.

The proof is provided in Appendix **??**. A similar result can be obtained with respect to the exact smoothing distribution of the exact diffusion

model, under assumptions on the number of particles and the step size of the Euler approximation.

5.3.5 Estimation method

We now detail the SAEM algorithm.

The approximate Morris-Lecar model (5.15) satisfies the exponential family assumption **(M1)** when the scaling parameters V_1, V_2, V_3 and V_4 are known. The sufficient statistics of the approximate model (5.15) are the following. Consider the $n \times 6$-matrix

$$X = \left(-V_{0:(n-1)}, -m_\infty(V_{0:(n-1)})V_{0:(n-1)}, -U_{0:(n-1)}V_{0:(n-1)}, U_{0:(n-1)}, \mathbf{1}, m_\infty(V_{0:(n-1)})\right)$$

where $\mathbf{1}$ is the vector of 1's of size n. Then the vector

$$S_1(V_{0:(n-1)}, U_{0:(n-1)}) = (X'X)^{-1}\, X'\, (V_{1:n} - V_{0:(n-1)})$$

is the sufficient statistic vector corresponding to the parameters $\nu_1(\theta) = (g_L, g_{Ca}, g_K, g_K V_K, g_L V_L + I, g_{Ca} V_{Ca})$, where $'$ denotes transposition.

The sufficient statistics corresponding to $\nu_2(\theta) = 1/\gamma^2$ are

$$\sum_{i=1}^{n} (V_i - V_{i-1})\, U_{i-1}, \quad \sum_{i=1}^{n} U_{i-1}^2, \quad \sum_{i=1}^{n} (V_i - V_{i-1})\, V_{i-1} m_\infty(V_{i-1}),$$

$$\sum_{i=1}^{n} (V_i - V_{i-1})\, U_{i-1} V_{i-1}, \quad \sum_{i=1}^{n} U_{i-1}^2 V_{i-1}^2.$$

The sufficient statistics corresponding to ϕ is also explicit but more complex and not detailed here.

We write s_m for the approximation of the expectation $\mathbb{E}_\Delta\left[S(V_{0:n}, U_{0:n})|V_{0:n}; \widehat{\theta}_{m-1}\right]$. At the S-step, the simulation under the smoothing distribution is done by SMC. We call this algorithm the SAEM-SMC algorithm. Iterations of the SAEM-SMC algorithm are written as follows:

[SAEM-SMC algorithm]

- *Iteration* 0*:* initialization of $\widehat{\theta}_0$ and set $s_0 = 0$.
- *Iteration* $m \geq 1$*:*

S-Step: simulation of the non-observed data $(U_{0:n}^{(m)})$ with SMC targeting the distribution $p_\Delta(U_{0:n}|V_{0:n}; \widehat{\theta}_{m-1})dU_{0:n}$.

SA-Step: update s_{m-1} using the stochastic approximation:

$$s_m = s_{m-1} + a_{m-1}\left[S(V_{0:n}, U_{0:n}^{(m)}) - s_{m-1}\right] \tag{5.20}$$

M-Step: update of $\widehat{\theta}_m$ by $\widehat{\theta}_m = \arg\max_{\theta\in\Theta} \left(-\psi(\theta) + \langle s_m, \nu(\theta)\rangle\right)$.

5.3.6 Convergence of the SAEM-SMC algorithm

The SAEM algorithm we propose in this paper is based on an approximate simulation step performed with an SMC algorithm. We prove that even if this simulation is not exact, the SAEM algorithm still converges towards the maximum of the likelihood of the approximated diffusion (5.15). This is true because the SMC algorithm has good convergence properties.

Let us be more precise. We introduce a set of convergence assumptions which are the classic ones for the SAEM algorithm (Delyon et al., 1999).

(M2) The functions $\psi(\theta)$ and $\nu(\theta)$ are twice continuously differentiable on Θ.

(M3) The function $\bar{s} : \Theta \longrightarrow \mathcal{S}$ defined by $\bar{s}(\theta) = \int S(v,u)p_\Delta(u|v;\theta)dv\,du$ is continuously differentiable on Θ.

(M4) The function $\ell_\Delta(\theta) = \log p_\Delta(v,u,\theta)$ is continuously differentiable on Θ and $\partial_\theta \int p_\Delta(v,u;\theta)dv\,du = \int \partial_\theta p_\Delta(v,u;\theta)dv\,du$.

(M5) Define $L : \mathcal{S}\times\Theta \to \mathbb{R}$ by $L(s,\theta) = -\psi(\theta) + \langle s, \nu(\theta)\rangle$. There exists a function $\hat{\theta} : \mathcal{S} \to \Theta$ such that $\forall\theta\in\Theta$, $\forall s \in \mathcal{S}$, $L(s,\hat{\theta}(s)) \geq L(s,\theta)$.

(SAEM1) The positive decreasing sequence of the stochastic approximation $(a_m)_{m\geq 1}$ is such that $\sum_m a_m = \infty$ and $\sum_m a_m^2 < \infty$.

(SAEM2) $\ell_\Delta : \Theta \to \mathbb{R}$ and $\hat{\theta} : \mathcal{S} \to \Theta$ are d times differentiable, where d is the dimension of $S(v,u)$.

(SAEM3) For all $\theta \in \Theta$, $\int ||S(v,u)||^2 p_\Delta(u|v;\theta)du < \infty$ and the function $\Gamma(\theta) = Cov_\theta(S(\cdot, U_{0:n}))$ is continuous, where the covariance is under the conditional distribution $p_\Delta(U_{0:n}|V_{0:n};\theta)$.

(SAEM4) Let $\{\mathcal{F}_m\}$ be the increasing family of σ-algebras generated by the random variables $s_0, U_{0:n}^{(1)}, U_{0:n}^{(2)}, \ldots, U_{0:n}^{(m)}$. For any positive Borel function f, $\mathbb{E}_\Delta(f(U_{0:n}^{(m+1)})|\mathcal{F}_m) = \int f(u)p_\Delta(u|v, \widehat{\theta}_m)du$.

Assumptions (M1)-(M5) ensure the convergence of the EM algorithm when the E-step is exact (Delyon et al., 1999). Assumptions (M1)-(M5) and (SAEM1)-(SAEM4) together with the additional assumption that $(s_m)_{m\geq 0}$ takes its values in a compact subset of $\mathcal{S}$ ensure the convergence of the SAEM estimates to a stationary point of the observed likelihood $p_\Delta(V_{0:n};\theta)$ when the simulation step is exact (Delyon et al., 1999).

Here the simulation step is not exact and we have three additional assumptions on the SMC algorithm to bound the error induced by this algorithm and prove the convergence of the SAEM-SMC algorithm.

(SMC1) The number of particles K used at each iteration of the SAEM algorithm varies along the iteration: there exists a function $g(m) \to \infty$ when $m \to \infty$ such that $K(m) \geq g(m)\log(m)$.

(SMC2) The function S is bounded uniformly in u.

(SMC3) The functions $p_\Delta(V_i|U_i, V_{i-1}, U_{i-1};\theta)$ are bounded uniformly in θ.

Theorem 7. *Assume that (M1)-(M5), (SAEM1)-(SAEM3), and (SMC1)-(SMC3) hold. Then, with probability 1,* $\lim_{m\to\infty} d(\widehat{\theta}_m, \mathcal{L}) = 0$ *where* $\mathcal{L} = \{\theta \in \Theta, \partial_\theta \ell_\Delta(\theta) = 0\}$ *is the set of stationary points of the log-likelihood* $\ell_\Delta(\theta) = \log p_\Delta(V_{0:n};\theta)$.

Theorem 7 is proved in Appendix **??**. Note that assumption (SAEM4) is not needed thanks to the conditional independence of the particles generated by the SMC algorithm, as detailed in the proof. Similarly, the additional assumption that $(s_m)_{m\geq 0}$ takes its values in a compact subset of $\mathcal{S}$ is not needed, as it is directly satisfied under assumption (SMC2).

We deduce that the SAEM algorithm converges to a (local) maximum of the likelihood under standard additional assumptions (LOC1)-(LOC3) proposed by Delyon et al. (1999) on the regularity of the log-likelihood $\ell_\Delta(V_{0:n};\theta)$ that we do not recall here.

Corollary 1. *Under the assumptions of Theorem 7 and additional assumptions (LOC1)-(LOC3), the sequence $\widehat{\theta}_m$ converges with probability 1 to a (local) maximum of the likelihood $p_\Delta(V_{0:n};\theta)$.*

The classical assumptions (M1)-(M5) are usually satisfied. Assumption (SAEM1) is easily satisfied by choosing properly the sequence (a_m). Assumptions (SAEM2) and (SAEM3) depend on the regularity of the model. They are satisfied for the approximate Morris-Lecar model.

In practice, the SAEM algorithm is implemented with an increasing number equal to the iteration number, which satisfies Assumption (SMC1). Assumption (SMC2) is satisfied for the approximate Morris-Lecar model because the variables U are bounded between 0 and 1 and the variables V are fixed at their observed values. This would not have been the case with the filter of Del Moral et al. (2001), which resimulates the variables V at each iteration. Assumption (SMC3) is satisfied if we require that γ is strictly bounded away from zero; $\gamma \geq \epsilon > 0$.

5.4 Hypoelliptic SDE

Bibliography

Andrieu, C., Doucet, A., and Punskaya, E. (2001). Sequential Monte Carlo methods for optimal filtering. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pages 79–95. Springer, New York.

Bally, V. and Talay, D. (1995). The law of the Euler Scheme for Stochastic Differential Equations: I. Convergence Rate of the Density. *Probability Theory and Related Fields*, 104(2675):43–60.

Bally, V. and Talay, D. (1996). The law of the Euler scheme for stochastic differential equations I. Convergence rate of the distribution function. *Monte Carlo Methods and Applications*, 2:93–128.

Bibbona, E., Lansky, P., and Sirovich, R. (2010). Estimating input parameters from intracellular recordings in the Feller neuronal model. *Physical Review E*, 81(3, 1). **Feller model for subthreshold activity. Nice review of estimates of drift parameters (least squares, conditional least squares, martingale estimating functions, Gauss-Markov method, optimal estimating function, MLE**.

Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag New York, USA.

Del Moral, P., Jacod, J., and Protter, P. (2001). The Monte-Carlo method for filtering with discrete-time observations. *Probab. Theory Related Fields*, 120(3):346–368.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27:94–128.

Ditlevsen, S. and Greenwood, P. (2013). The Morris-Lecar neuron model embeds a leaky integrate-and-fire model. *J. Math. Biol.*, 67(2):239–259.

Doucet, A., de Freitas, N., and Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, Stat. Eng. Inf. Sci., pages 3–14. Springer, New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for bayesian filterin. *Stat. Comput.*, 10:197–208.

Gerstner, W., Kistler, W., Naud, R., and Paninski, L. (2014). *Neuronal dynamics: from single neurons to nnetwork and models of cognition*. Cambridge.

Gordon, N., D., S., and A., S. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings-F*, 140:107–113.

Iacus, S. (2008). *Simulation and Inference for stochastic difdifferent equations*. Springer-Verlag New York.

Kloeden, P. E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag Berlin.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.*, 49:1020–1038.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, 93(443):1032–1044.

Nualart, D. (2006). *The Malliavin calculus and related topics*. Springer-Verlag New York.

Oksendal, B. (2007). *Stochastic differential equations: an introduction with applications*. Springer-Verlag, Berlin-Heidelberg.