Modeling and dynamical analysis of cortical network activity in semantic priming

Auteur : Dejace, Caroline
Promoteur(s) : Sacré, Pierre
Faculté : Faculté des Sciences appliquées
Diplôme : Master en ingénieur civil biomédical, à finalité spécialisée
Année académique : 2022-2023
URI/URL : http://hdl.handle.net/2268.2/18065

Modeling and dynamical analysis of cortical network activity in semantic priming

Master thesis carried out with the aim of obtaining the degree of Master in Biomedical Engineering by

Caroline Dejace

Promoter: Professor Pierre Sacré

University of Liège
Faculty of Applied Sciences
in partnership with PsyNCog
Academic Year 2022 - 2023

Abstract

The semantic priming paradigm, involved in language comprehension, refers to the facilitated processing and retrieval of a word (known as target) following the previous processing of another semantically related word (known as prime). Literature on semantic priming reveals a vivid debate about the nature of priming: it can be associative (e.g. *afraid-scared*), semantic, that is "true relations of meanings", (e.g. *sheep-goat*) or a combination of both (e.g. *cat-dog*). This debate impacts then how the semantic memory, coding words' meaning, is modeled and how the priming is thought to occur.

Brunel and Lavigne (2009) designed a network model that studies semantic priming as a function of a set of parameters. In addition, they used an input-output relationship that is mathematically good-looking but rather difficult to manipulate numerically. This master thesis thus focuses on assessing whether using a more standard and a more numerically stable input-output relationship, such as a sigmoid function, would give a qualitatively similar dynamic behavior to the original model. Furthermore, the thesis investigates the parameter sensitivity. To these ends, the network model is simplified into a one-dimensional model and the dynamic behavior is investigated for both input-output relationships. The modified model is then tested with experimental-like stimuli to mimic real psychology experiments and to understand semantic memory functioning.

Dynamical analysis performed on the derived one-dimensional model reveals that the dynamic behavior remains qualitatively the same when using one or the other input-output relationship. Results also suggest parameter sensitivity of the original model. The modified model with experimental-like stimuli suggests that the semantic memory system should be in a bistable regime to observe semantic priming. Activation of a word in semantic memory depends then on the amplitude and the duration of the stimulus. Extension to higher-order dimensions is also discussed.

Keywords: semantic priming, network model, rate model, phase portrait, bifurcation, psychology experiments

Acknowledgements

What a roller coaster of a last year it has been. Here I am finally and it would not have been possible without the following people.

I would like to deeply and immensely thank my promoter **Professor Pierre Sacré**, first for proposing this very interesting topic that is a crossover between engineering and psychology fields, but especially for his great availability, his guidance, his ideas and his reassuring support when my emotions became overwhelming at some point. I am very honored to have been able to work under his supervision and I am grateful that this collaboration will hopefully continue.

I would also like to thank **Professor Alessio Franci** for his suggestions and his help with the bifurcation diagrams, **Professor Steve Majerus** for his help in understanding the more psychological side of this master thesis, and **Professor Guillaume Drion** for the time and the attention he will give to the reading of this master thesis.

I wish to thank as well **Benjamin Kowialiewski**, F.R.S.−FNRS researcher, for his help in understanding the more psychological side of this work. Thanks also to **Anaëlle De Worm**, PhD candidate and teaching assistant, for her help with graphs commands in the *Julia* programming language.

Finally, last but not least, I would like to deeply and immensely thank my **family**. My parents who always supported me and took care of me mentally and physically over these last five years and especially during this crazy adventure of the Erasmus Exchange. A special thank to my father for his countless free hugs whenever needed. Thanks also to my elder sister especially for our mutual sisterhood support.

Liège, August 10, 2023

Caroline DEJACE

To my grandfather **Victor**,
I wish you were still alive to enjoy the finish line.

To my parents **Olivier** and **Anne**,
I would not have come this far without you.

Contents

List of Figures

List of Tables

Symbols and Acronyms

$\dot{x} = \frac{dx}{dt}$	Time derivative of variable $x(t)$
$\ddot{x} = \frac{d^2x}{dt^2}$	Second time derivative of variable $x(t)$
1D	One-dimensional
N-D	N-dimensional
E	Excitatory
I	Inhibitory
AP	Action Potential
B&L	Brunel and Lavigne
DSM	Distributional Semantic Models
FP	Fixed Point, Attractor, Equilibrium
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
PF	Pitchfork
RT	Reaction, Response, Recognition Time
SN	Saddle-Node
SOA	Stimulus Onset Asynchrony
SP	Semantic Priming
TC	Transcritical
WC	Wilson-Cowan

Introduction

Motivations

Language is a key tool for human beings' real-time interactions and verbal communication with each other. Verbal communication implies a flow of information between two individuals: the speaker sends information whereas the listener receives that information, and potentially responds to it. In order to speak coherent sentences or to understand these sentences, the knowledge or *meaning* of each word should be retrieved from semantic memory since it contains general knowledge such as vocabulary, facts, ..., shared by a large number of individuals (Schacter 2000). However, it is still unclear how semantic memory exactly represents the knowledge of words or other concepts such as pictures (Hutchison 2003; Kumar 2021; Sperber et al. 1979).

In addition, it happens that the processing of a word is facilitated by the previous processing of another semantically related word. For example, if one transiently sees the word *cat*, the processing of that word facilitates the processing of the following word *dog* in the sense that the word *dog* is already pre-activated in memory before seeing explicitly the word *dog*. This facilitation occurs because the words *cat* and *dog* are related. This phenomenon is known as semantic priming. Literature on semantic priming highlights a vivid debate on the nature of semantic priming: associative, semantic or both (the term "semantic" in "semantic priming" encompasses all types). The nature of priming, in turn, has an impact on the semantic memory modeling (Hutchison 2003; Kumar 2021). Brunel and Lavigne 2009 designed a network model that studies semantic priming effects as a function of several key parameters. However, Brunel and Lavigne 2009 use a rather large number of parameters and their assigned values seem to be very specific to that model, suggesting that the model is potentially sensitive to changes in parameter values. Moreover, their input-output relationship (also known as transfer function) is mathematically good-looking but seems rather unpractical to use and manipulate numerically. Also, the physiological interpretation of that transfer function remains unclear.

A closer look at this model and its transfer function seems appropriate to investigate the dynamics of semantic memory and semantic priming.

Contributions of this master thesis

My master thesis revisits the model of Brunel and Lavigne 2009 and attempts to determine whether the same model with another more standard transfer function could be used, that is, the modified model would give the same qualitative dynamic behavior as the original model. In other words, a first contribution of my master thesis is to assess the Brunel and Lavigne's model equivalence. If this equivalence is successful, then the modified model would allow similar results with a greater numerical stability. To assess model equivalence, a simplified one-dimensional version of the original model is used and the full dynamic behavior is explored and explained using engineering tools.

Another contribution of my master thesis is the evaluation of the sensitivity of the model's parameters. My master thesis explores whether parameter values could be changed without changing the global behavior and properties of the model and if so, to what extent they can be varied.

Finally, stimuli similar to those used in real psychology experiments are applied to the simplified model to observe how the latter responds to them. This final step also explores the condition(s) to observe semantic priming, constituting then the last contribution of my thesis.

Code availability

The whole computational study has been performed using the *Julia* programming language (v1.9.0; https://julialang.org/) within a *Jupyter Notebook* environment (v6.5.2 from Anaconda https://www.anaconda.com/). All the source codes can be found on the following GitHub repository: https://github.com/CarolineDejace/ATFE0016-1-master-thesis.

Structure of this thesis

This master thesis is divided into four main parts:

Part I is composed of three chapters and reviews the theory on dynamical system modeling. Chapter 1 defines the concepts that are the basis for mathematical modeling. It explains what are signals, what are their properties, how can they be transformed and how to extend the concepts to systems. Chapter 2 reviews the complete analysis of a one-dimensional model. It addresses the topics of phase portrait analysis and bifurcation analysis, two common engineering tools that are used in this master thesis. Chapter 3 talks about network models, also often referred to as rate models. Different topologies

are discussed notably those of a recurrent network model and a Wilson-Cowan network model, that are of great importance in this master thesis.

Part II addresses the more psychological side of this work. It includes a review on memory classification (Chapter 4) and the link between language and memory is also discussed. Chapter 5 reviews in turn literature on semantic priming: how to observe it, how to measure it, what are the possible semantic relationships and how can semantic priming be modeled.

Part III is made of two chapters and outlines the computational study that was performed. Chapter 6 focuses on the assessments of the model equivalence and the parameter sensitivity. It uses phase portrait and bifurcation analyses to investigate in details the dynamic behavior of the model using one or the other transfer function. Chapter 7 continues the study with an application of the model to experimental-like stimuli.

A last chapter then concludes the work with a summary of the study and the results. Perspectives on potential improvements and extensions to higher-order dimensions are discussed as well.

Part I

Modeling Background

Chapter 1

Signal and System Basics

This chapter aims at understanding the very basic concepts and terminology of the commonly called dynamical system modeling.

1.1 What is *dynamical system modeling*?

To understand what is dynamical system modeling, the concepts of *signal*, *system*, *dynamics* and *modeling* should be properly understood first.

Palani 2022a defines a *signal* as any "physical phenomenon that carries some information or data". The signal is in general time-dependent, that is, a function of variable t where t does not depend on any other variable and usually encodes for the time. The variable t is then said to be independent. Such signals are for instance temperature, chemical concentration, number of individuals in a population, ...

Palani 2022a then defines a *system* as a "set of interconnected objects or components with a definite relationship between objects and attributes". For example, a mass m attached to a spring with stiffness k that is itself attached to a fixed wall (Figure 1.1) constitutes with the spring a system. The objects are the mass and the spring, and their attributes are their respective displacements and speeds. A definite relationship links the objects and the time evolution of their attributes (see hereafter).

In simpler terms, a system can be seen as a black box that transforms or converts an input signal into an output signal (Drion 2019-2020). Put differently, a system receives a signal, acts on it and outputs a modified version of the input signal.

The third concept to understand properly is the *dynamics*. When a system evolves with time, it is said to be dynamic (Strogatz 1994). The system then outputs a signal that depends on the input signal evaluated at the current time instant t^* but can also depend on the input signal evaluated at past and/or future time instants (Drion 2019-2020).

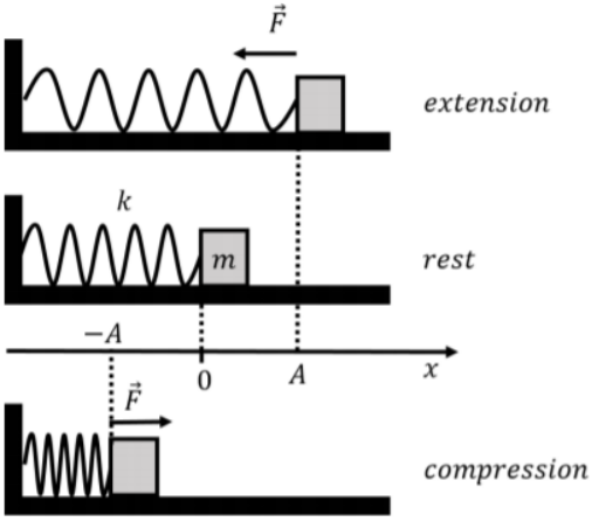


Figure 1.1: Horizontal single mass m-spring (stiffness k) system oscillating around the zero-displacement position ($x = 0$) with an amplitude $A > 0$. The force $\overrightarrow{F}$ generates this oscillation. Extracted from Marting and Ng n.d.

Finally, *modeling* can be seen as the process of finding a mathematical law that describes the temporal behavior (or evolution) of the system (Drion 2019-2020). It thus consists actually in finding the definite relationship between the objects and their attributes. Taking again the example with the mass-spring system, this definite relationship is given by Newton's law stating that the mass-spring system will be in equilibrium when the sum of all external forces acting on it is equal to zero (Drion 2019-2020; Strogatz 1994):

$$m\frac{d^2x}{dt^2} + k \cdot x = 0 \tag{1.1}$$

with $x = x(t)$, the time-dependent variable coding for the displacement of the mass and the spring, and $\frac{d^2(.)}{dt^2}$, the second time derivative operator or in other words, differentiation of the variable $x(t)$ with respect to the independent variable t has been applied twice. The first term in (1.1) accounts for the acceleration or gravity force whereas the second term accounts for the restoring force from the spring.

Combining all the above-mentioned concepts, dynamical system modeling is thus the process of finding a set of a finite number n of equations describing the time evolution of each individual variable as relationships between the n time-dependent variables and the r inputs ($r \leq n$).

1.2 Signal properties

Signals can have several of the following properties simultaneously. Here below are the most common properties that can be found in signals. The list is thus non-exhaustive.

1.2.1 Continuous Time VS Discrete Time

A signal $x(t)$ is said to be a *continuous time* signal if and only if $x(t)$ is defined for any value of the independent variable t (coding for time usually). On the contrary, a signal $x[n]$, where n is an integer, is said to be a *discrete time* signal if and only if it is defined at discrete values of time only. These discrete times are then indexed by an integer. A discrete time signal can be seen as a sampled version of a continuous time signal (Drion 2019-2020; Palani 2022a).

This master thesis will deal with continuous time signals only.

1.2.2 Periodic VS Non-periodic

A continuous time signal $x(t)$ is periodic if it satisfies

$$x(t + nT) = x(t) \quad \text{for all } t \tag{1.2}$$

with n an integer and $T > 0$ the period, that is, the signal repeats itself after a time period T. In case $x(t)$ does not satisfy Eq. (1.2), the signal is said to be *aperiodic* or *non-periodic*. A similar definition can be applied to discrete time signals (Drion 2019-2020; Palani 2022a).

1.2.3 Even VS Odd

A continuous time signal $x(t)$ is even if it satisfies

$$x(-t) = x(t) \quad \text{for all } t \tag{1.3}$$

that is, the signal is symmetric with respect to the y-axis (ordinate axis). On the contrary, the signal $x(t)$ is odd if it satisfies

$$x(-t) = -x(t) \quad \text{for all } t \tag{1.4}$$

that is, the signal $x(t)$ is a central symmetry about the time origin. A similar definition can be applied to discrete time signals. It should be noted then that any signal can be expressed as the sum of its even and odd parts (Drion 2019-2020; Palani 2022a).

$$x(t) = x_{even}(t) + x_{odd}(t) = \underbrace{\frac{1}{2}[x(t) + x(-t)]}_{even} + \underbrace{\frac{1}{2}[x(t) - x(-t)]}_{odd} \tag{1.5}$$

1.3 Signal transformation

A signal can undergo various transformations. Here below is a non-exhaustive list of these possible transformations for continuous time signals but these transformations can be transposed to discrete time signals as well.

1.3.1 Time shifting

The transformation $t \rightarrow t - t_0$, where t_0 is a real constant, results in a *linear shift* of a signal $u(t)$ by t_0 time units. This shift will be to the right (time delay) if t_0 is strictly positive, to the left (time advance) if t_0 is strictly negative and no shift if $t_0 = 0$ (Drion 2019-2020; Palani 2022a).

1.3.2 Time scaling

The transformation $t \rightarrow a \cdot t$, where a is a strictly positive real constant, results in a contraction of the signal $u(t)$ if a is greater than one, and in a dilatation (or expansion) if a is smaller than one. The amplitude of the signal does not change but the signal is accelerated ($a > 1$) or slowed down ($a < 1$) over time (Drion 2019-2020; Palani 2022a).

1.3.3 Folding

A folded (or reflected) signal $u(-t)$ is obtained when the signal $u(t)$ is reflected on the y-axis (vertical axis or ordinate axis), that is, the y-axis acts as a mirror between the positive and the negative times t (Drion 2019-2020; Palani 2022a).

1.3.4 Inversion

An inverted signal $-u(t)$ is obtained when the amplitude of the signal $u(t)$ is inverted. For this transformation, the x-axis (horizontal axis or abscissa axis) acts as a mirror between the positive and the negative values of $u(t)$ for all t (Drion 2019-2020; Palani 2022a).

1.3.5 Combined transformations

When several transformations are applied simultaneously to a signal $u(t)$, the order of the different transformations matters! If not carefully followed, the resulting transformed signal might be very different from the expected signal. The order to follow is the reverse order of the priority of basic arithmetic operations. Put differently, the order to follow for signal transformations is

1. Addition and subtraction: $+$ and $-$

2. Multiplication and division: $*$ and /
3. Powers and exponents
4. Parentheses

For example, the transformed signal $y(t) = u(\frac{-t-t_0}{a})$ (adapted from Palani 2022a p.26) is obtained by

1. Time scaling the signal $u(t)$ by a factor a: $u_1(t) = u(\frac{t}{a})$
2. Time shifting the signal $u_1(t)$ by t_0 time units to the right: $u_2(t) = u_1(t - t_0) = u(\frac{t-t_0}{a})$
3. Folding the signal $u_2(t)$: $y(t) = u_2(-t) = u(\frac{-t-t_0}{a})$

In this case, another order could have been applied (*i.e.* 1. folding, 2. time scaling by a, 3. time shifting by t_0 units to the left) (Drion 2019-2020; Palani 2022a).

1.4 System properties

Similarly to signals, (dynamical) systems can also have several properties simultaneously.

1.4.1 Continuous VS Discrete

A system is continuous if it deals with continuous input and output signals. Similarly, a system is discrete if it deals with discrete input and output signals. If a system deals with discrete (continuous) input signals and continuous (discrete) output signals, then the system is said to be hybrid (Drion 2019-2020).

Dynamical systems are modeled using *differential equations* when the systems are continuous, and modeled with *difference equations* when they are discrete (Strogatz 1994). Also, in continuous systems, depending on whether there exist one or more independent variables, the systems are modeled using *ordinary* differential equations (ODE) (one independent variable only and so ordinary derivatives only) or *partial* differential equations (PDE) (more than one independent variable and so partial derivatives). For example, Eq. (1.1) describing the single mass-spring system is an ODE because the time t is the only independent variable, hence ordinary derivatives are used. However, if the heat equation is considered

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{1.6}$$

with u the heat, then Eq. (1.6) is a PDE because both time t and space x are independent variables (Strogatz 1994).

This master thesis will deal with continuous systems with one independent variable (time t) only. As a result, differential equations will be used. A general framework for this master thesis can thus be given by the system (or set) of ODEs

$$\begin{cases} \dot{x}_1 &= f_1(x_1, \ldots, x_n, u_1, \ldots, u_r) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \ldots, x_n, u_1, \ldots, u_r) \end{cases} \tag{1.7}$$

where $\dot{x}$ is equivalent to $\frac{dx}{dt}$, $(x_1, \ldots, x_n)$ are the time-dependent variables (or state variables), $(u_1, \ldots, u_r)$ $(r \leq n)$ are the time-dependent external input signals, and the functions $f_1, \ldots, f_n$ are the definite relationships linking all (state) variables and input signals together. The functions are determined according to the problem that is considered (Strogatz 1994). The integer n describes the number of variables and consequently corresponds to the *dimension* or the *order* of the system (Drion 2019-2020; Strogatz 1994).

1.4.2 Linear VS Non-linear

The system (1.7) is said to be *linear* when all functions $f_1, \ldots, f_n$ are *linear* functions of the variables x_i and inputs u_j $(i = 1, \ldots, n;\ j = 1, \ldots, r)$. Put differently, the system (1.7) is linear if all its equations satisfy the additivity and the homogeneity properties (Drion 2019-2020; Palani 2022b). The additivity property is given by

$$\begin{gathered} z_1(t) = S\{y_1(t)\} \\ z_2(t) = S\{y_2(t)\} \\ z(t) = S\{y_1(t) + y_2(t)\} = S\{y_1(t)\} + S\{y_2(t)\} = z_1(t) + z_2(t) \end{gathered} \tag{1.8}$$

with $S\{.\}$ the system function, y_i the input signals and z_i the output signals.

The additivity property of a linear system states that the total output signal resulting from the system function applied to the sum of each individual input signal is equal to the sum of each individual output signal resulting from the system function applied to each individual input signal (Drion 2019-2020).

The homogeneity property is given by

$$\begin{gathered} z_1(t) = S\{y_1(t)\} \\ z(t) = S\{a \cdot y_1(t)\} = a \cdot S\{y_1(t)\} = a \cdot z_1(t) \end{gathered} \tag{1.9}$$

with a a scaling factor, $S\{.\}$ the system function, y_i the input signals and z_i the output signals.

The homogeneity property of a linear system states that the output signal resulting

from the system function applied to an input signal scaled by a factor a is equal to a scaled version by a factor a of the output signal resulting from the system function applied to the individual input signal (Drion 2019-2020).

If at least one equation of (1.7) is non-linear, then the whole system is non-linear. Non-linearities are for example products ($x_i \cdot x_j$, $i \neq j$), powers (x_i^2) and non-linear functions ($\exp(-x_i)$) of the variables (Strogatz 1994). Non-linear systems are very difficult and often impossible to solve analytically ("by hand"). Moreover, even when a solution of the non-linear system can be found, it gives little meaningful information (Drion 2019-2020). The usual way to proceed is to find numerically an approximate solution and/or to adopt a geometric/graphical perspective to have a global and qualitative behavior of the system, and/or to consider an operating point of the system and linearize the system around it, that is, considering a small neighbourhood around that operating point where the system will *locally* behave as a linear one and can therefore be solved easily using the additivity and the homogeneity properties (Strogatz 1994).

1.4.3 Time Varying VS Time Invariant

Palani 2022b defines a continuous time invariant system as a system where "the parameters of the system do not change with time". This definition also implies that the input-output relationship does not change with time, that is, the input-output relationship is independent of the time origin: a time shift applied in the input signal results in the same time shift in the output signal given by the system. Mathematically, the time invariance property is expressed as follows

$$\begin{aligned} z(t) &= S\{y(t)\} \\ z(t-\tau) &= S\{y(t-\tau)\} \end{aligned} \tag{1.10}$$

with τ a time shift, $S\{.\}$ the system function, y the input signal and z the output signal.

Systems can also have other properties such as causality, stability, invertibility, ... (Palani 2022b) but these are more involved in a more advanced signal and image processing topic, which is not the point in this background. These properties will not therefore be further investigated.

Chapter 2

One-dimensional system analysis

This chapter aims at explaining in details the dynamic behavior of a one-dimensional (1D) system. This chapter also focuses on the mathematical and graphical (or geometric) approaches to analyze such a system. Basic concepts of dynamical system analysis such as trajectory, vector field, fixed point, stability, bifurcation, ... are explained. This chapter is highly based on Strogatz 1994 Chapter 2 pp 15 – 35 and Chapter 3 pp 44 – 78.

2.1 1D model

As mentioned in section 1.4.1, the general framework for this master thesis is given by Eq. (1.7). As a first step, the simplest case where $n = 1$ is considered. The set of equations (1.7) then reduces to

$$\dot{x} = f(x, u) \tag{2.1}$$

with $x = x(t)$ a real-valued function of independent time variable t, $u = u(t)$ the input signal and $f(x, u)$ a smooth real-valued function of variable x and input u. The function f can be linear or non-linear (Drion 2019-2020; Strogatz 1994).

Equation (2.1) is called a *one-dimensional* or *first-order* system[1] (or model). The dimension associated to this model is thus one. A solution to Eq. (2.1) is given by the time evolution of variable x (*i.e.* $x(t)$) and is called a *trajectory*. Trajectories allow to know the value of each variable at each time instant (Drion 2019-2020; Strogatz 1994).

[1]The word "system" means a *dynamical* system in this context, not a set of equations as used in the classical sense. It should therefore be noted that a single equation can model a system.

2.2 Graphical approach

A practical example is always better than words to understand concepts. The following differential equation is considered:

$$\dot{x} = \cos x \tag{2.2}$$

with $x = x(t)$ a time-dependent variable.

This equation is non-linear since $f(.) = \cos(.)$ is non-linear. Moreover, Eq. (2.2) possesses a closed form, that is, an explicit relationship between x and t where x is a function of t only (and *vice-versa*). The analytical solution to Eq. (2.2) is given by (see Appendix B.1 for calculation details)

$$t = \frac{1}{2} \ln \left(\frac{|1 + \sin x|}{|1 - \sin x|} \cdot \frac{|1 - \sin x_0|}{|1 + \sin x_0|} \right) \tag{2.3}$$

with $x(t = 0) = x_0$ an initial condition.

The result given by Eq. (2.3) is exact but does not give meaningful or at least clear information. For instance, what is the (qualitative) behavior of the trajectory $x(t)$ as time t tends towards infinity for any arbitrary initial condition x_0 or even for a fixed x_0? Based on the analytical solution alone, it is not easy to answer that question (Strogatz 1994).

On the other hand, by adopting a geometric point of view and thus using a graphical analysis, the interpretation of Eq. (2.2) becomes clear, simple and easy. If one thinks of t as time and x as the position of some object moving along the real line (*i.e.* one direction only), then $\dot{x}$ corresponds to the velocity of that object. In other words, $\dot{x}$ indicates how fast or how slow the variable x evolves with time and in which direction (increases/decreases) the position will evolve. For a 1D model, the graphical analysis consists in plotting $\dot{x}$ VS x, also known as the *phase portrait* or the *phase plane*. Equation (2.2) can then be interpreted as a *vector field* on the line, that is, $\dot{x}$ indicates the velocity vector at each x (Strogatz 1994). Arrows can thus be drawn on the line (or x-axis) to illustrate these velocity vectors at each x. When $\dot{x} > 0$ ($\dot{x} < 0$), the arrows point to the right (left) indicating that the variable x will increase (decrease) if it is in that state (Figure 2.1). At points where $\dot{x} = 0$, no evolution of the variable x (and so the corresponding system) will be observed because the velocity vector is the null vector. These points are therefore called *fixed points* (FP), *equilibria* or *attractors* if a psychological context is used (Brunel and Lavigne 2009; Drion 2019-2020; Strogatz 1994).

In Figure 2.1, two kinds of FPs can be observed. Since the vector field points towards them, blue dots attract trajectories whose initial condition lies in the region bounded by the two red dots on each side of a blue dot. These blue dots are then called *stable*

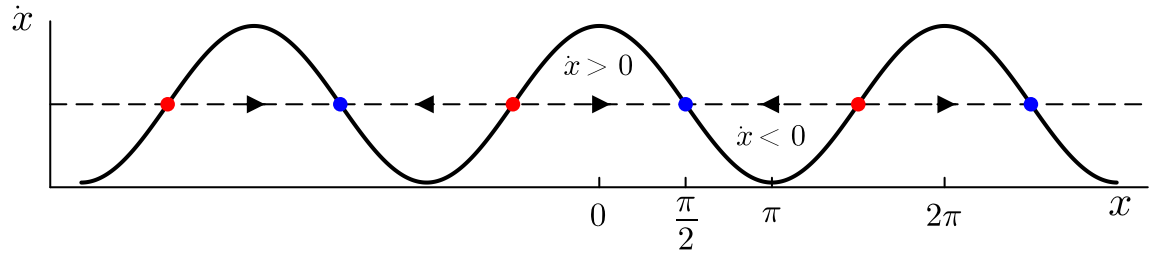


Figure 2.1: Phase portrait $\dot{x}$ VS x for the 1D model $\dot{x} = \cos x$. Horizontal black dashed line represents the stationary state $\dot{x} = 0$. Vector field of the model is illustrated by black arrows. Fixed points (*i.e.* $\cos x = 0$) are blue ($\rightarrow$ stable FP) and red ($\rightarrow$ unstable FP) dots. Inspired from Strogatz 1994 Ch2 p17.

fixed points. On the contrary, since the vector field points away from them, the red dots repel any trajectory whose initial condition stands nearby them. These red dots are then *unstable* fixed points (Strogatz 1994). More details about stability of fixed points will be discussed in the next section but it should be noted that the stability of the fixed points can already be determined graphically even though no formula for the fixed points are available (Strogatz 1994).

The graphical information given by the phase portrait (Figure 2.1) of the 1D model $\dot{x} = \cos x$ allows one to easily understand the qualitative behavior of the solution(s) of the model for any initial condition x_0. For an initial condition x_0 such that $\dot{x} < 0$, the trajectory $x(t)$ will decrease more or less fast (depending on the exact location of x_0). The movement of x in Figure 2.1 is thus to the left. As time goes by, $x(t)$ asymptotically approaches and reaches the nearest stable FP from a greater position (or right if one considers Figure 2.1). Similarly, if $\dot{x} > 0$ initially, $x(t)$ will increase and the movement of x in Figure 2.1 is to the right. The solution $x(t)$ asymptotically approaches and reaches the nearest stable FP from a lower position (or left if one considers Figure 2.1). When $\dot{x} = 0$, the system does not evolve anymore and $x(t)$ remains constant (Drion 2019-2020; Strogatz 1994).

The graphical technique developed in this section can be applied to any 1D system $\dot{x} = f(x)$. All that is needed is the graph of $f(x)$ that can then be used to deduce and sketch the vector field on the x-axis (Strogatz 1994).

2.2.1 Limitations

Drawing the phase portrait is an easy way to get meaningful and clear information about the dynamical system. However, three major limitations of this graphical technique can be identified.

1. The phase portrait can only give *qualitative* information and not *quantitative* information.

   The information given by the phase portrait is very qualitative in the sense that it helps to have an *idea* of the time evolution of a trajectory but no figures are given. The quantitative information such as the time at which the FP is reached, the time at which the speed $|\dot{x}|$ is the greatest or the rate of growth/decay can not be obtained from the phase portrait (Drion 2019-2020; Strogatz 1994).

2. The phase portrait is a *global* point of view.

   One looks at the evolution of the system for all possible values of the variable x when using the phase portrait. It does not give local information around an operating point for instance (Drion 2019-2020).

3. When uniqueness of the solution $x(t)$ (if it exists) fails, the geometric/graphical approach does not help any longer because (infinitely) many solutions (and so trajectories) could start from the same initial condition but give different behaviors. Consequently, the graphical approach with such an initial condition would not help in determining which behavior has been taken (Strogatz 1994).

2.3 Stability analysis

As explained in the previous section, fixed points can be obtained graphically by drawing the phase portrait of the 1D model and finding the intersection(s), if any, with the x-axis. Mathematically, the fixed points are thus points x^* defined by

$$\dot{x} = f(x^*) = 0 \tag{2.4}$$

These points correspond then to steady states of the system; no time evolution of the system can be observed in these states and $x(t)$ remains constant. In addition, as seen in Figure 2.1, the fixed points can be of different kinds which can also be determined using the phase portrait: if the vector field points towards (away from) a fixed point then the latter is stable (unstable). Expressed differently, a fixed point is defined to be stable if *all sufficiently small* disturbances away from it decay in time. On the contrary, FPs or equilibria are defined to be unstable when disturbances grow in time (Drion 2019-2020; Strogatz 1994).

An important note to make is that stability can be local or global. Thus far, stability was based on *small* disturbances but it happens that in some cases, certain large disturbances may fail to damp out or could change the fixed point towards which a trajectory $x(t)$ eventually converges. For example, in Figure 2.2, if the disturbances are contained

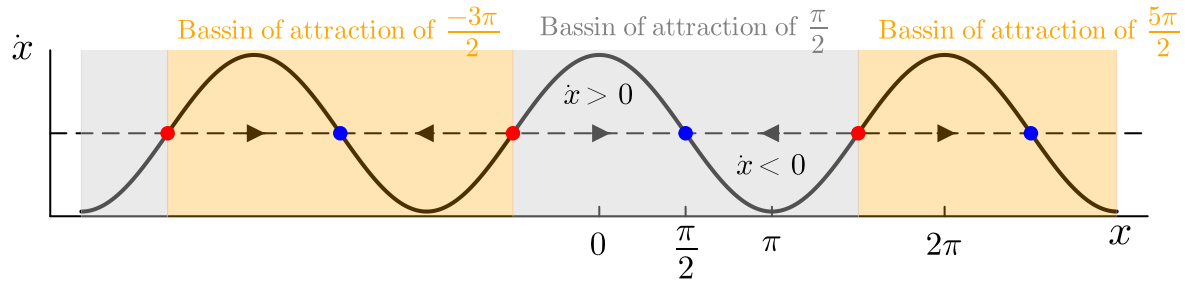


Figure 2.2: Phase portrait $\dot{x}$ VS x for the 1D model $\dot{x} = \cos x$. Horizontal black dashed line represents the stationary state $\dot{x} = 0$. Vector field of the model is illustrated by black arrows. Stable fixed points are represented by blue dots whereas unstable fixed points are represented by red dots. Their respective basins of attraction are the gray-shaded and orange-shaded area.

in a bounded region, called the *basin of attraction*, around a FP ($\frac{\pi}{2}$ for instance), then these disturbances damp out and any trajectory starting in this basin indeed converges to $\frac{\pi}{2}$. However, if the disturbances are large enough to push an initial condition or even the value of $x(t)$ at some time instant t from the basin of attraction of $\frac{\pi}{2}$ into the basin of attraction of $\frac{-3\pi}{2}$, then $x(t)$ eventually converges to a FP other than that initially planned without disturbances. The stable fixed points of the 1D model $\dot{x} = \cos x$ are therefore *locally* stable but not *globally* stable. Using another example such as $\dot{x} = -x + 1$, one immediately sees that the graph of $f(x)$ (Figure 2.3) is a straight line with a negative slope. Thus, the associated vector field has a single FP at $x^* = 1$. Since the vector field points towards it, the FP is stable. More than that, since the FP is reached from all initial conditions, this FP is *globally* stable. Any trajectory *monotonically* converges to the FP with a speed decreasing linearly as stated by $\dot{x} = -x + 1$ (Strogatz 1994).

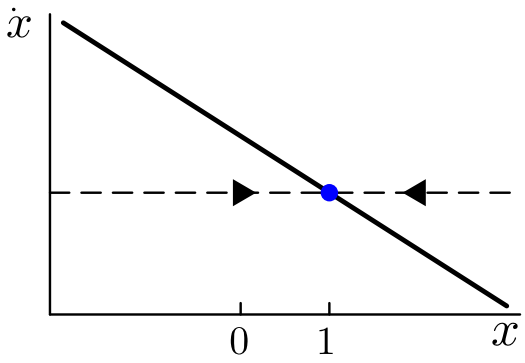


Figure 2.3: Phase portrait $\dot{x}$ VS x for the 1D model $\dot{x} = -x + 1$. Horizontal black dashed line represents the stationary state $\dot{x} = 0$. Vector field of the model is illustrated by black arrows. Stable fixed point is represented by blue dot.

2.3.1 Linear stability analysis

An advantage of the graphical method is the possibility to determine the stability of a fixed point without having to calculate anything. Conversely, a drawback of this method is that it does not give any quantitative information. However, it is common to want a quantitative measure of stability such as the rate of decay to a stable fixed point (Strogatz 1994). Using a more mathematical approach, called the *small perturbation study*, this quantitative information can be obtained by considering an operating point and *linearizing* the system about it. Since one wants to know about the stability of the fixed points, the operating points to choose should be the corresponding fixed points.

If one finds x^* to be a fixed point, that is $\dot{x}|_{x=x^*} = f(x^*) = 0$, and considers

$$\eta(t) = x(t) - x^* \tag{2.5}$$

as a small perturbation away from x^*, the small perturbation study consists in determining whether the perturbation will grow or decay in time. In order to answer this question, one needs to find a differential equation for $\eta(t)$ describing the time evolution of the perturbation (Drion 2019-2020; Strogatz 1994). Differentiation of Eq. (2.5) on both sides yields

$$\dot{\eta} = \frac{d(x - x^*)}{dt} = \dot{x} - \dot{x^*} = \dot{x} \tag{2.6}$$

using the definition of a fixed point (x^* is a constant). Also, one has

$$\dot{\eta} = \dot{x} = f(x) = f(x^* + \eta) \tag{2.7}$$

Now one can use the Taylor's expansion to approximate Eq. (2.7). Taylor's expansion (see Appendix B.2 for a summary of this formula) gives

$$\dot{\eta} = f(x^* + \eta) \approx f(x^*) + \eta f'(x^*) + \mathcal{O}(\eta^2) = \eta f'(x^*) + \mathcal{O}(\eta^2) \tag{2.8}$$

with $f'(x^*)$ the first derivative, with respect to variable x, of function f evaluated at the fixed point x^*. $\mathcal{O}(\eta^2)$ denotes quadratic terms in η (Drion 2019-2020; Strogatz 1994).

If $f'(x^*) \neq 0$ in Eq. (2.8), then the $\mathcal{O}(\eta^2)$ terms are negligible. Indeed, since η is a small perturbation, its squared value will be even smaller and close to zero. One can thus approximate

$$\dot{\eta} \approx \eta f'(x^*) \tag{2.9}$$

Equation (2.9) is now *linear* in variable η hence the name *linearization about* x^* (Drion 2019-2020; Strogatz 1994). The solution of Eq. (2.9) is given by

$$\eta(t) = \eta_0 \exp\left(f'(x^*) \cdot t\right) \tag{2.10}$$

Thus, thanks to Eq. (2.10), one can easily determine the time evolution of the small perturbation η:

- If $f'(x^*) < 0$, then the perturbation decays exponentially fast and the fixed point x^* is, at least, locally stable.
- If $f'(x^*) > 0$, then the perturbation grows exponentially fast and the fixed point x^* is unstable.
- If $f'(x^*) = 0$, then $\dot{\eta} \approx \eta f'(x^*) + \mathcal{O}(\eta^2) = \mathcal{O}(\eta^2)$ so that the quadratic terms are not negligible anymore and a non-linear stability analysis is needed.

The key message in this subsection is that the slope $f'(x^*) = \frac{d\dot{x}}{dx}|_{x=x^*}$ at the fixed point x^* gives the stability of x^*. The *sign* of $f'(x^*)$ is thus of crucial importance and can be immediately identified using the graphical approach, by looking at the slope of function f at the fixed point. With the small perturbation technique, one can have access to a measure of *how* (un)stable a fixed point is in addition to the sign of $f'(x^*)$. This degree of stability is given by the *magnitude* of $f'(x^*)$: the larger $|f'(x^*)|$, the more (un)stable the fixed point is and the faster $\eta(t)$ and $x(t)$ evolve with time (Drion 2019-2020; Strogatz 1994).

The computation of $f'(x^*) = \frac{d\dot{x}}{dx}|_{x=x^*}$ actually amounts to compute the *eigenvalue* of the system at the fixed point. The sign of the eigenvalue gives the stability and the magnitude gives the degree of stability.

A limitation of the small perturbation technique should still be noticed. The linearization about a fixed point is only valid for a small neighborhood or perturbation around the fixed point. This technique can therefore be applied *locally* only (Drion 2019-2020).

2.4 Bifurcations

For a 1D model, the dynamics of vector fields on the (real) line is very limited: the trajectory either converges monotonically to a stable fixed point (or equilibrium) or diverges monotonically to $\pm\infty$ if the initial condition is not contained in any basin of attraction of a finite-value fixed point (Strogatz 1994). As a result, the interesting part about 1D models is not their vector fields but rather the *dependence* of the solution of the model *on model parameters*. Indeed, it happens that by varying the values of the model parameters the structure or behavior of the solution changes completely. The most common examples are the creation or destruction of fixed points and the change in stability of a fixed point. These changes in behavior are called *bifurcations* and the parameter values for which bifurcation happens are referred to as *bifurcation points* (Strogatz 1994). In other words, a

"bifurcation point creates a frontier between two different behaviors" (Drion 2021-2022). At bifurcation point(s), the following two conditions should be met (Strogatz 1994)

$$\begin{cases} \dot{x}|_{x=x^*} = 0 & x^* \text{ is a fixed point} \\ \frac{d\dot{x}}{dx}|_{x=x^*} = 0 & \text{the eigenvalue at the fixed point is zero} \end{cases} \tag{2.11}$$

From a geometric point of view, the second condition means that the graph of $\dot{x} = f(x)$ is tangent to the x-axis.

Bifurcations are thus of crucial importance because they allow to understand any behavior a model could have as *control* parameters vary. More importantly, bifurcations give information about *when* these changes in behavior occur.

Here below is a summary for each of the three basic bifurcations that can be encountered for 1D models and that can even be extended to higher-order models.

2.4.1 Saddle-node bifurcation

Creation or destruction of fixed points happens through what is called a *saddle-node bifurcation* (SN) or *fold bifurcation*. The basic mechanism is that as a parameter is varied, two fixed points move towards each other, merge and disappear. The canonical (or *normal*) forms[2] of the SN bifurcation for a 1D model are given by

$$\dot{x} = \alpha + x^2 \tag{2.12}$$

and

$$\dot{x} = \alpha - x^2 \tag{2.13}$$

where α is a real parameter (Drion 2021-2022; Strogatz 1994).

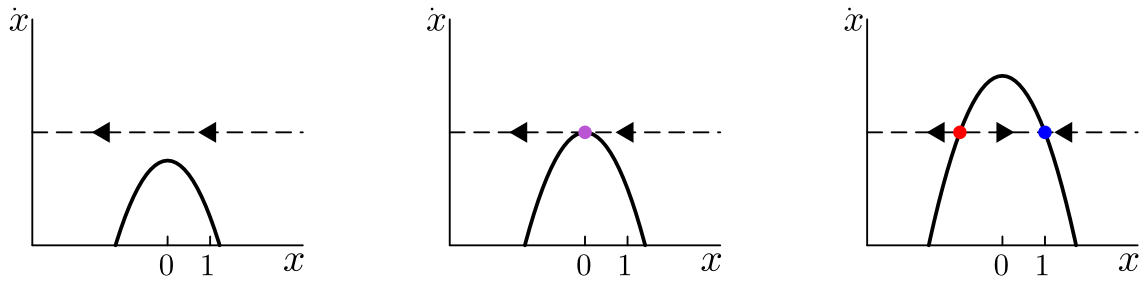
By varying the value of α, three case studies can be observed. The parameter α is then called a *bifurcation parameter* (Drion 2021-2022; Strogatz 1994). Considering the normal form $\dot{x} = \alpha - x^2$, these three case studies are the following (Figure 2.4)

- If $\alpha < 0$ (Figure 2.4a), then no real fixed points can be observed. Indeed, the fixed points are given by $x^* = \pm i\sqrt{-\alpha}$ with i the imaginary unit[3]. Moreover, since $\dot{x}$ is always negative for all x, the associated vector field is unidirectional (and points to the decreasing x's).

- As α increases, the parabola (*i.e.* $-x^2$) moves up. If $\alpha = 0$ (Figure 2.4b), then the parabola is tangent to $\dot{x} = 0$ (dashed line) and a fixed point appears at $x = 0$. This

[2]The SN bifurcation admits two normal forms because fixed points have two different ways to be created and destroyed: the vector field is unidirectional but has two possible directions (right or left).

[3]The imaginary unit i is defined by $i^2 = -1$.



(a) $\alpha < 0$ **(b)** $\alpha = 0$ **(c)** $\alpha > 0$

Figure 2.4: Phase portrait of the saddle-node bifurcation $\dot{x} = \alpha - x^2$ for all possible values of the real parameter α. Horizontal black dashed line indicates the states $\dot{x} = 0$. Vector field is represented by black arrows. Stable fixed point is illustrated by the blue dot, unstable fixed point is shown with the red dot and the purple dot stands for the saddle-node. Inspired from Strogatz 1994 Ch3 p45.

FP is attractive from one side and repulsive from the other side as observed with the vector field. This is the reason why this FP is called a saddle-node.

- As α still increases (Figure 2.4c), two separate fixed points appear, one stable (blue dot) and one unstable (red dot). The vector field is thus not unidirectional anymore.

It is said that a bifurcation occurred at $x = 0$ because the vector fields for $\alpha < 0$ and $\alpha > 0$ are different (Strogatz 1994).

In order to have a better view of the effect of parameter changes on the fixed point values, a *bifurcation diagram* can be drawn. This bifurcation diagram consists in plotting x^* VS α, that is the fixed point values against the parameter values (Figure 2.5) (Drion 2021-2022; Strogatz 1994).

If the other normal form were used, then the vector field of the phase portrait (Figure 2.4) would change direction so that fixed points exchange their stability. Figure 2.5 would undergo a central symmetry with respect to $(0, 0)$ (*i.e.* $\alpha \to -\alpha$ and $x^* \to -x^*$).

2.4.2 Transcritical bifurcation

It happens that a fixed point always exists whatever the value of a model parameter, and can never be destroyed. Nevertheless, such a fixed point may still change its stability as the parameter varies (Strogatz 1994). This mechanism is called a *transcritical bifurcation*.

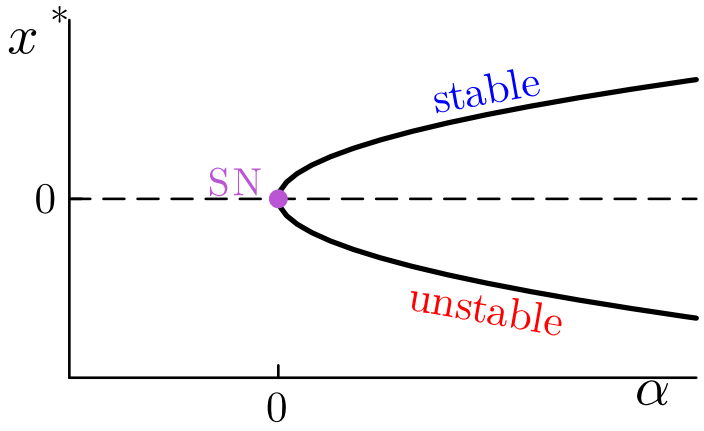


Figure 2.5: Bifurcation diagram of the saddle-node bifurcation using the normal form $\dot{x} = \alpha - x^2$. Horizontal black dashed line stands for $x^* = 0$. The upper branch ($x^* = \sqrt{\alpha}$) of the bifurcation diagram represents stable fixed points whereas the lower branch ($x^* = -\sqrt{\alpha}$) represents unstable fixed points. Saddle-node point is illustrated by the purple dot.

The transcritical bifurcation normal form is given by

$$\dot{x} = \alpha x - x^2 \tag{2.14}$$

with α a real-valued parameter.

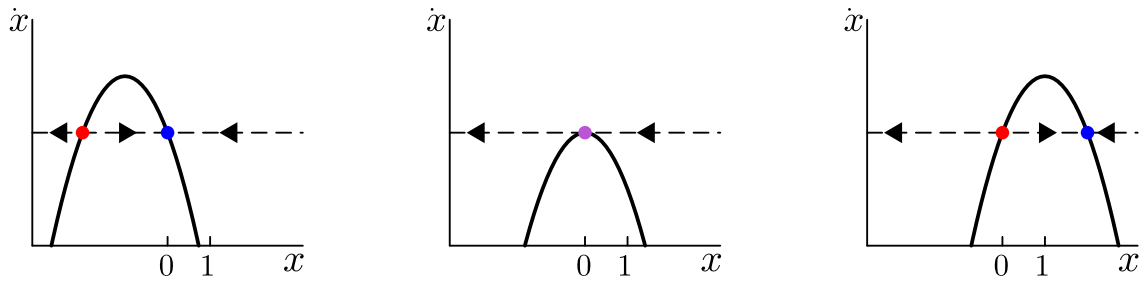
One can immediately see that $x^* = 0$ is always a fixed point whatever the value for α. By varying the value of α, three setups can be observed (Drion 2021-2022; Strogatz 1994):

- If $\alpha < 0$ (Figure 2.6a), then two real fixed points can be observed. Based on the vector field, the leftmost FP (*i.e.* $x^* = \alpha$ with $\alpha < 0$; red dot) is unstable whereas the rightmost FP (*i.e.* $x^* = 0$; blue dot) is stable.

- As α increases, the unstable FP approaches the origin. If $\alpha = 0$ (Figure 2.6b), then only one fixed point is left at $x = 0$ (purple dot). This FP is attractive from one side and repulsive from the other side as observed in the SN bifurcation. The vector field is unidirectional and points to the decreasing x's.

- As α still increases (Figure 2.6c), an exchange of stability has occurred: the leftmost FP (*i.e.* $x^* = 0$; red dot) is unstable while the rightmost FP (*i.e.* $x^* = \alpha$ with $\alpha > 0$; blue dot) is stable.

Again, one can also draw the associated bifurcation diagram (Figure 2.7) to determine the fixed point values for all possible parameter values (Drion 2021-2022; Strogatz 1994).



(a) $\alpha < 0$ **(b)** $\alpha = 0$ **(c)** $\alpha > 0$

Figure 2.6: Phase portrait of the transcritical bifurcation $\dot{x} = \alpha x - x^2$ for all possible values of the real-valued parameter α. Horizontal black dashed line indicates the states $\dot{x} = 0$. Vector field is represented by black arrows. Stable fixed points are illustrated by the blue dots, unstable fixed points are shown with the red dots and the purple dot stands for the transcritical point. Inspired from Strogatz 1994 Ch3 p50.

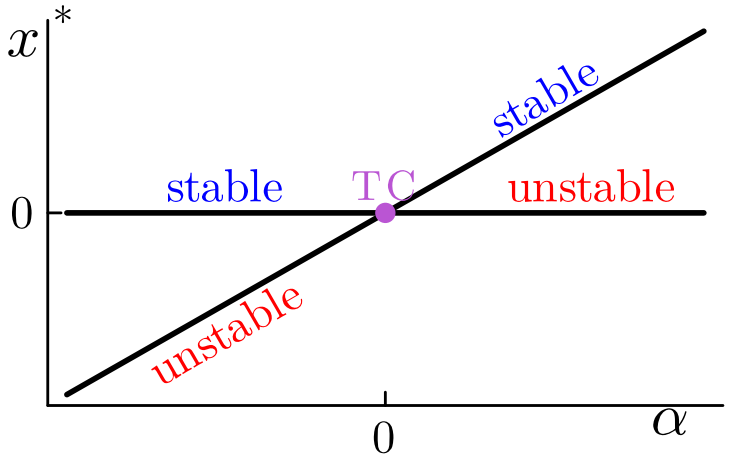


Figure 2.7: Bifurcation diagram of the transcritical bifurcation whose normal form is $\dot{x} = \alpha x - x^2$. Stability of branches ($x^* = 0$ and $x^* = \alpha$) is indicated by the colored strings. Transcritical point (TC) is illustrated by the purple dot.

2.4.3 Pitchfork bifurcation

When the considered problem has symmetry, this symmetry may be reflected in the bifurcation diagram. More precisely, fixed points are created and destroyed in symmetrical pairs (Strogatz 1994). Two types of pitchfork bifurcation can be discussed.

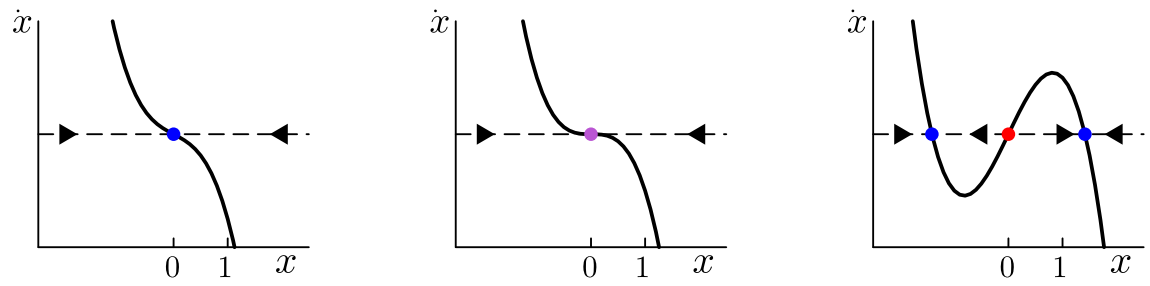
Supercritical pitchfork bifurcation

The normal form of this first type of pitchfork bifurcation is given by

$$\dot{x} = \alpha x - x^3 \tag{2.15}$$

The symmetry feature can be assessed by checking invariance of Eq. (2.15) under the transformation $x \to -x$. If one gets back Eq. (2.15) after this transformation, then the vector field (Figure 2.8) is said to be *equivariant* (Strogatz 1994). Again, by varying the parameter α, the phase portrait adopts three different configurations:

- If $\alpha < 0$ (Figure 2.8a), then only one FP ($x^* = 0$; blue dot) can be observed. Based on the vector field, this FP is stable.
- If $\alpha = 0$ (Figure 2.8b), then the slope at the origin becomes really close to zero. The FP at $x = 0$ is still stable based on the vector field but the sensitivity to variations in α has become very large.
- As α still increases (Figure 2.8c), the FP at the origin becomes unstable (red dot) and two new fixed points ($x^* = \pm\sqrt{\alpha}$) appear symmetrically about the origin. These new FPs (blue dots) are both stable.



(a) $\alpha < 0$ **(b)** $\alpha = 0$ **(c)** $\alpha > 0$

Figure 2.8: Phase portrait of the supercritical pitchfork bifurcation $\dot{x} = \alpha x - x^3$ for all possible values of the real-valued parameter α. Horizontal black dashed line indicates the states $\dot{x} = 0$. Vector field is represented by black arrows. Stable fixed points are illustrated by the blue dots, unstable fixed point is shown with the red dot and the purple dot stands for the pitchfork point. Inspired from Strogatz 1994 Ch3 p56.

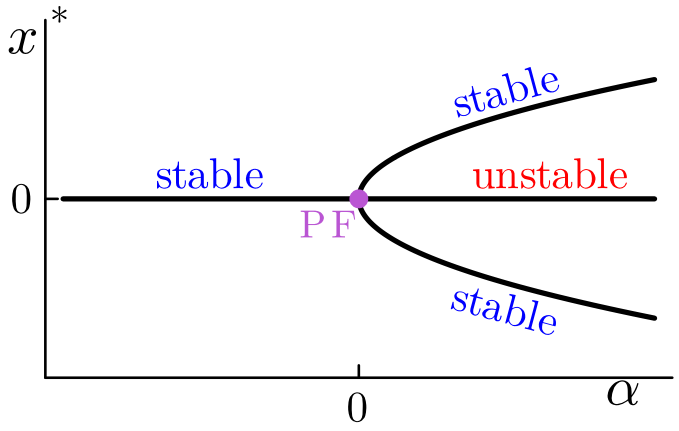


Figure 2.9: Bifurcation diagram of the supercritical pitchfork bifurcation whose normal form is $\dot{x} = \alpha x - x^3$. Stability of branches ($x^* = 0$ and $x^* = \pm\sqrt{\alpha}$) is indicated by the colored strings. Pitchfork point (PF) is illustrated by the purple dot.

The word "pitchfork" takes on its full meaning when looking at the bifurcation diagram (Figure 2.9).

Subcritical pitchfork bifurcation

The cubic term in the supercritical pitchfork (*i.e.* $\dot{x} = \alpha x - x^3$) is *stabilizing* in the sense that as $x(t)$ increases and becomes very large, the cubic term becomes dominant with respect to αx so that $\dot{x} < 0$. As a result, $x(t)$ decays and is "pulled back" towards $x = 0$; the cubic term exerts a *negative feedback*. (Drion 2021-2022; Strogatz 1994). The *subcritical pitchfork bifurcation* uses a *destabilizing* or *positive feedback* cubic term:

$$\dot{x} = \alpha x + x^3 \tag{2.16}$$

The same reasoning as for the supercritical pitchfork can be applied and one would find that all the results are inverted: stable fixed points become unstable (or *vice-versa)* and exist *below* the bifurcation point, and the bifurcation diagram is inverted as well (Figure 2.10) (Drion 2021-2022; Strogatz 1994).

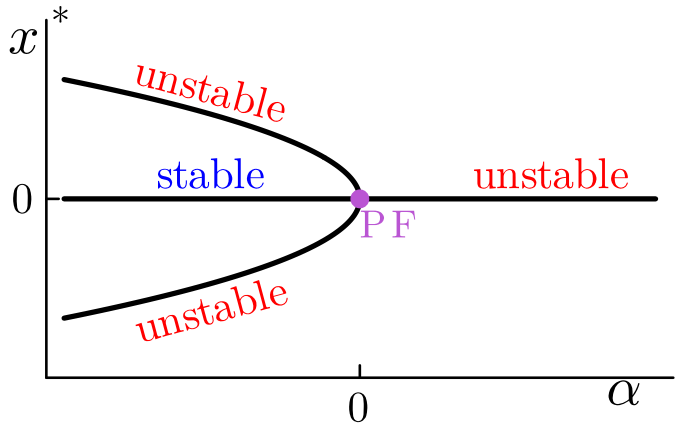


Figure 2.10: Bifurcation diagram of the subcritical pitchfork bifurcation whose normal form is $\dot{x} = \alpha x + x^3$. Stability of branches ($x^* = 0$ and $x^* = \pm\sqrt{-\alpha}$) is indicated by the colored strings. Pitchfork point (PF) is illustrated by the purple dot.

Application

In decision-making, supercritical pitchfork bifurcations are used to illustrate *flexible representations*, that is, showing the transition from a faithful (or analog) representation of a stimulus, to a categorical (or digital; high or low for instance) representation of the same stimulus (Franci 2023a). A faithful representation will therefore be associated with a regime with a single stable fixed point whereas categorical representations will be associated with regimes with multiple stable fixed points of the system.

In the categorical part of the supercritical pitchfork, that is values of α above the bifurcation point, two branches of stable equilibria coexist; in this region the system is said to be *bistable*. The initial condition and the external inputs/stimulus (if any) determine together which of the two stable steady states will be reached. A drawback of the supercritical pitchfork is that even tiny values of external inputs/stimulus may be mapped to one of the two stable branches in the bistable region; there is no possibility to remain somehow uncertain for a while (Franci 2023a). However, it can be seen even in human behavior that remaining uncertain is a desirable feature depending on the context. A way to model this possibility of remaining uncertain is to add a stable "neutral" state coexisting with the stable upper and lower branches of the supercritical pitchfork (Franci 2023a). One needs then to use a subcritical pitchfork with a stabilizing higher-order term that opposes the destabilizing effect of the positive feedback due to the cubic term. Thus, the normal form of such a flexible system is given by

$$\dot{x} = \alpha x + x^3 - x^5 \tag{2.17}$$

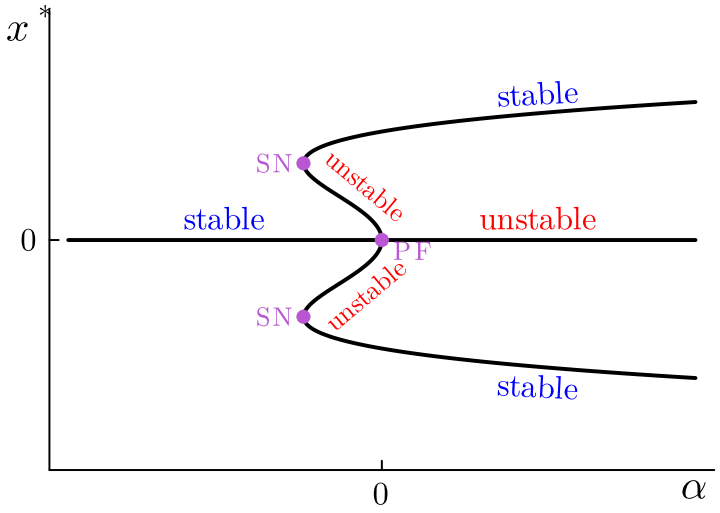


Figure 2.11: Bifurcation diagram of the combination of subcritical and supercritical pitchfork bifurcations: $\dot{x} = \alpha x + x^3 - x^5$. Stability of the different branches is indicated by the colored strings. Pitchfork point (PF) and saddle-node points (SN) are shown with the purple dots.

where the higher-order term x^5 is used to keep symmetry of the system under the transformation $x \to -x$ (Strogatz 1994). The associated bifurcation diagram is given in Figure 2.11.

A last note about bistability is the possible absence of reversibility, also known as *hysteresis* (Strogatz 1994). Hysteresis means that the path followed by a trajectory is different as the parameter varies in one direction or the other (Franci 2023a). Hysteresis also implies *jumps* in the bifurcation diagram as the parameter is varied (Strogatz 1994). For example (Figure 2.12), assuming the system starts in the state $x^* = 0$ with $\alpha < 0$, the system remains at $x^* = 0$ as α increases slightly. When α becomes greater than PF by a tiny bit, the system jumps to a stable branch (the upper branch for example) because the state $x^* = 0$ has become unstable and a tiny noise is assumed. If the parameter α starts decreasing, then the system does not jump back but rather follow the path of the upper stable branch. It is only when α becomes smaller than the SN value that the system jumps back to $x^* = 0$.

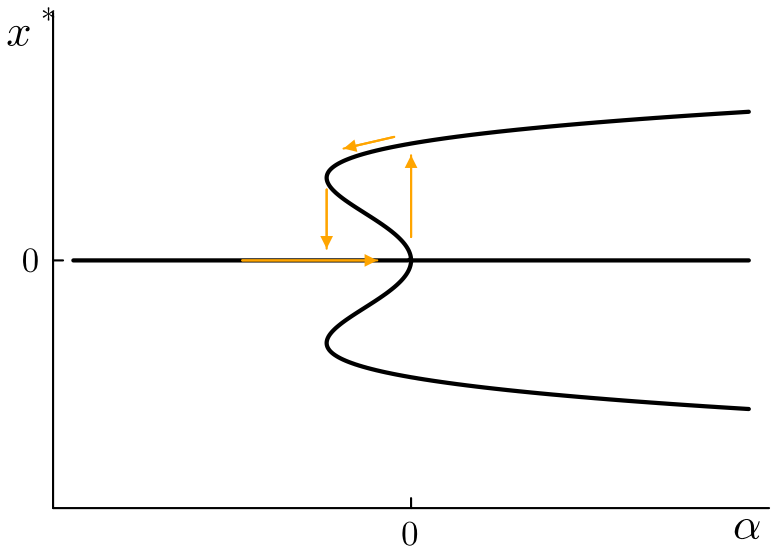


Figure 2.12: Hysteresis phenomenon happening along the bifurcation diagram of $\dot{x} = \alpha x + x^3 - x^5$. Evolution of the system state as parameter α varies is illustrated by the orange arrows. Inspired from Strogatz 1994 Ch3 p60.

Chapter 3

Network models

This chapter aims at reviewing the class of models used in this master thesis, namely the *network models* also referred to as *firing-rate models* or simply *rate models*. Since this master thesis takes place in a neuroscience-oriented context, a quick summary of the underlying neurobiology and neurophysiology is provided as a beginning. The goal is not to go deep into the details but rather to give the basic concepts in order to understand how the network models operate.

3.1 Context

The elementary unit of (human) brain is called the *neuron*. Neurons are *morphologically* and *functionally specialized cells* (Vandewalle 2020-2021) made of three parts (Dayan and Laurence F. Abbott 2001a; Vandewalle 2020-2021):

1. A *dentritic tree* or simply *dendrites* receiving information from other neurons, brain areas, ...
2. A *cell body*, also known as *soma*, containing notably the nucleus with the genetic inheritance (*i.e.* the so-called *DNA*).
3. An *axon* conveying electrical signals to other neurons, brain areas, ...

In addition, like any other cells, neurons have a *cell membrane* separating the intracellular medium from the extracellular medium. Moreover, this membrane has a *membrane potential* due to the difference in both the concentration and the number of charged molecules, called *ions*, between the interior of the neuron and the surrounding extracellular medium (Dayan and Laurence F. Abbott 2001a; Vandewalle 2020-2021). Under resting conditions, the difference of potential between the inside and the outside of a neuron is about $-70\ mV$. The neuron is then said to be *polarized* (Dayan and Laurence F. Abbott 2001a; Vandewalle 2020-2021).

Neurons are *excitable* because they respond to input stimuli (chemical, mechanical, ...) by generating (or *firing*) characteristic electrical pulses called *action potential* (AP) or simply *spikes* (Dayan and Laurence F. Abbott 2001a). This generation of AP is made through a transient *depolarization* (*i.e.* increase) of the membrane potential (Dayan and Laurence F. Abbott 2001a; Vandewalle 2020-2021). Neurons then represent and transmit information by firing sequences of spikes in various temporal patterns. It should be noted that the AP generation may depend on the recent history of firing of a neuron. Spikes are essential for the communication between neurons because they are the only form of membrane potential fluctuation that can propagate over large distances (Dayan and Laurence F. Abbott 2001a).

Once generated, spikes are propagated unidirectionally along the axon from the soma to the *axon terminals* of a neuron. Transmission of information between two neurons effectively occur at *synapses* (Dayan and Laurence F. Abbott 2001a; Vandewalle 2020-2021). Synapses couple one (or sometimes several) axon terminal(s) of one (or several) neuron(s) sending information, called *presynaptic* neuron(s), to a dendrite of a neuron receiving that information, called the *postsynaptic* neuron (Dayan and Laurence F. Abbott 2001a; Vandewalle 2020-2021). When the AP of a presynaptic neuron arrives at the axon terminal, it usually triggers the release of molecules, called *neurotransmitters*, into the *synaptic cleft*, the tiny space between the presynaptic axon terminal and a dendrite of a postsynaptic neuron. These neurotransmitters diffuse and then bind to proteins, called *receptors*, in the membrane of a postsynaptic dendrite. The bond between a neurotransmitter and its associated receptor results in an excitatory or inhibitory postsynaptic current (or potential by using Ohm's law), the so-called EPSC (EPSP) and IPSC (IPSP), respectively (Gerstner et al. 2014 Ch 3). Depending on the nature and the strength of this postsynaptic signal, the postsynaptic neuron will generate or not a spike and propagate it to another neuron (Dayan and Laurence F. Abbott 2001a; Vandewalle 2020-2021).

3.2 Network models

Network models allow one to explore the computing power of complex connectivity schemes between several neurons or even populations of neurons. Moreover this exploration can be done using both mathematical analysis and simulations on computers (Dayan and Laurence F. Abbott 2001b). When building or designing a network model, a few questions about the *architecture* (or *topology*) and the *components* of the model should be answered first because the dynamics produced by one specific architecture with specific components will be different from another architecture with other components (Brunel 2021).

Regarding the architecture of the model, three typical questions arise (Brunel 2021):

1. How *many* neurons or populations are considered?
2. How many *types* of neurons? → One (excitatory E or inhibitory I), two (E and I), more than two (subtypes, different layers, ...), ...
3. *How* are the neurons or populations *connected*? → Fully connected (*i.e. all-to-all*), randomly connected with fixed connection probability or with fixed number of connections per neuron/population (Gerstner et al. 2014 Ch 12), connected with a temporal structure imposed by learning (*i.e. synaptic plasticity*), ...

For the components of the network, three questions arise as well (Brunel 2021):

1. What are the *external inputs*? → Constant, stochastic (e.g. Poisson process[4], ...), time-dependent, space-dependent, ...
2. Which *neuron/population model* should be used? → Binary model, rate model (output = rate), spiking model (output = membrane potential), ...
3. Which *synapse model* should be used? → Constant number (*i.e.* synaptic weight; → usually for binary or rate models), temporal kernel (*i.e.* a time-dependent function specifying how synaptic currents are triggered by a presynaptic AP and then evolve with time; → usually for spiking models), current-based or conductance-based (*i.e.* voltage-dependent), ...

3.3 Network of single neurons

The most direct way to model and simulate neural networks (or systems) is to gather data on the currents in each individual neurons, that is, to synaptically connect models of spiking neurons. The information on the currents is then used to compute the membrane potential of each neuron in the network. However, this approach represents a great challenge and the resulting model is difficult to analyze and understand. Instead, neural activity can be described using *firing-rates* of the neurons (L. F. Abbott 1991; Dayan and Laurence F. Abbott 2001b).

Firing-rate models have several advantages:

- They are simple and easy to analyze and simulate (L. F. Abbott 1991) because they work at a longer time scale than action potentials (Dayan and Laurence F. Abbott 2001b).

[4] A Poisson process generates a sequence of spikes where the probability of firing a spike at the current time instant is independent of all previous spikes. All the spikes in the sequence are then statistically independent (Dayan and Laurence F. Abbott 2001a).

- The number of free parameters is usually smaller for the firing-rate models than for the spiking models. Setting these free parameters becomes then easier (Dayan and Laurence F. Abbott 2001b).

- Spiking models can accurately predict the spike sequences of a neuron only if all the inputs to that neuron are known, which is usually not the case especially when the network is complex. This precision about spike times may not be realizable in practice. On the other hand, firing-rate models can be used to generate stochastic spike sequences with a constant firing rate (Dayan and Laurence F. Abbott 2001b).

- When considering populations of neurons, firing-rate models at the neuron-level can be easily adapted to population-level by using a *mean-field* approach (L. F. Abbott 1991), that is, by simply averaging the firing rates of the neurons constituting the populations. This mean-field approach works as long as the populations have a large number of neurons constituting them and are *homogeneous*, that is, the populations are constituted of neurons with similar properties, that respond similarly to a given stimulus and that receive the same inputs (Gerstner et al. 2014 Ch 12). If the responses of neurons were spikes, then how should these spikes be averaged to have a signal adapted to the population-level? Can these spikes even be averaged? (Dayan and Laurence F. Abbott 2001b).

Since any model is never perfect, rate models also have their limitations:

- The major drawback of rate models is that voltage-dependent quantities (such as the membrane potential) are not computed (L. F. Abbott 1991).

- Rate models determine how *often* and not *when* APs are fired. Rate models cannot therefore account for precise spike timing and/or spike correlations (L. F. Abbott 1991; Dayan and Laurence F. Abbott 2001b).

- For a rate model to work, the number of inputs that a neuron in the network receives should be high so that the firing rate of each network unit approximates well the effect of actual spike sequences of each network unit on the network's dynamic behavior (L. F. Abbott 1991; Dayan and Laurence F. Abbott 2001b).

- Rate models are thus restricted to cases where the firing from one neuron or another is uncorrelated, is asynchronous or at least little synchronous, and where precise sequences of spikes are not important (L. F. Abbott 1991; Dayan and Laurence F. Abbott 2001b).

3.3.1 What does "firing rate" mean?

As explained in section 3.1, neurons represent and transmit information using temporal sequences of spikes. If the spikes are treated as instantaneous, identical and idealized events, then these sequences can be completely characterized by what is called the *neural response function* defined as

$$\rho(t) = \sum_{i=1}^{n} \delta(t - t_i) \tag{3.1}$$

with $0 \leq t_i \leq T$ the time of the ith spike ($i = 1, 2, \ldots, n$), T the total duration of the sequence and $\delta(t)$ the Dirac function, that is

$$\delta(t) = \begin{cases} 0 & \text{for all } t \neq 0 \\ \int_{-\infty}^{+\infty} \delta(t)dt = 1 & \text{for } t = 0 \end{cases}$$

In other words, the neural response function is a binary list of the n spike times (Dayan and Laurence F. Abbott 2001a).

In rate models, the exact time course of $\rho(t)$ is replaced by the approximate description given by the firing rate (Dayan and Laurence F. Abbott 2001b). However, what is implied by the terms "firing rate"? Unfortunately, there is no unique definition. The terms "firing rate" can actually have at least four different definitions depending on the averaging procedure that is used (Gerstner et al. 2014 Ch 7).

Rate as a spike count (average over time)

The most commonly used definition for the firing rate is the temporal average. For a given neuron generating a sequence of spikes in a given trial k, the *spike-count rate* r is the total number of spikes fired by that neuron in trial k, divided by the total duration T of trial k (Eq. (3.2)). Put differently, the spike-count rate is the time average of $\rho(t)$ over the duration of trial k (single neuron, single trial) (Dayan and Laurence F. Abbott 2001a; Gerstner et al. 2014).

$$r = \frac{n}{T} = \frac{1}{T} \int_0^T \rho(t)dt \tag{3.2}$$

A drawback of this computation is that all the time resolution in the time course of the trial is lost (Dayan and Laurence F. Abbott 2001a).

Rate as a firing probability (average over trials)

A time-dependent (or instantaneous) firing rate can be defined as "the average number of spikes (averaged over trials) appearing during a short interval between t and $t + \Delta t$, divided by the duration of the interval" (Dayan and Laurence F. Abbott 2001b). Also,

the average number of spikes over trials occurring during Δt is obtained by the integral of the trial-averaged neural response function. The firing rate $r(t)$ is then given by the following equation (single neuron, repeated trials)

$$r(t) = \frac{1}{\Delta t} \int_t^{t+\Delta t} \langle \rho(s) \rangle \, ds = \frac{1}{K \Delta t} \int_t^{t+\Delta t} \sum_{k=1}^{K} \rho_k(s) \, ds \tag{3.3}$$

with K the total number of trials and ρ_k the sequence of spikes in trial k (Dayan and Laurence F. Abbott 2001b; Gerstner et al. 2014).

If Δt is sufficiently small, then the number of spikes in any sequence k over an interval Δt will be at most one so that $\int_t^{t+\Delta t} \rho_k(s) \, ds$ is either equal to zero or one. Consequently, $r(t)\Delta t$ is the probability of firing a spike during an interval Δt. Similarly, $r(t)\Delta t$ is the proportion of trials in which a spike occurred between t and $t + \Delta t$. Due to this relationship between this fraction and the firing rate, it is allowed to replace the trial-averaged neural response function $\langle \rho(s) \rangle$ with the firing rate $r(t)$ within any integral (Dayan and Laurence F. Abbott 2001b).

When no additional specifications are explicitly stated, the terms "firing rate" refer to this meaning.

Rate as an average firing rate (average over time and trials)

If the spike-count measure is performed on repeated trials, then the value might vary from one trial to the next due to the stochastic nature of neurons. Their associated spike sequences might vary from one trial to the next as well (Gerstner et al. 2014). An average firing rate can then be obtained by averaging the spike-count rate over the total number of trials (Eq. (3.4)) (Dayan and Laurence F. Abbott 2001b). The averaging procedure looks like a "mean of means".

$$\langle r \rangle = \frac{\langle n \rangle}{T} = \frac{1}{T} \int_0^T \langle \rho(t) \rangle \, dt = \frac{1}{T} \int_0^T r(t) dt \tag{3.4}$$

Based on Eq. (3.4), one can observe that the average firing rate corresponds to the time average of the instantaneous firing rate $r(t)$.

Rate as a population activity (average over several neurons)

When considering populations of neurons, the firing rate can be seen as an average over the neurons in the population (Gerstner et al. 2014):

$$\bar{r}(t) = \frac{1}{N \Delta t} \int_t^{t+\Delta t} \sum_{j=1}^{N} \rho_j(s) \, ds \tag{3.5}$$

with N the size of the population (*i.e.* the number of neurons within a population) and Δt a short time interval around time t. One can actually see that formulations for $r(t)$ and $\bar{r}(t)$ are the same; the only difference lies in the context of the formula.

3.3.2 Rate models derivation

To construct a rate model, one needs to follow two steps (L. F. Abbott 1991; Dayan and Laurence F. Abbott 2001b):

1. First, one needs to determine how the total synaptic input current (or voltage (Gerstner et al. 2014)) to a neuron depends on the firing rates of its presynaptic afferents. This is the step where $r(t)$ approximates $\rho(t)$ (Dayan and Laurence F. Abbott 2001b). In other words, the relation $r_{pre}(t) \rightarrow I_{syn}(t)$ is modeled.

2. Secondly, the firing rates of the postsynaptic neurons as a function of their total synaptic currents are determined (L. F. Abbott 1991; Dayan and Laurence F. Abbott 2001b). Expressed differently, it amounts to find the relation $I_{syn}(t) \rightarrow r_{post}(t)$.

In order to better visualize the derivation of a rate model, a simple network with P presynaptic neurons (*i.e.* presynaptic inputs) and a single postsynaptic neuron (*i.e.* postsynaptic output) can be found in Figure 3.1. Presynaptic neurons make *feedforward* connections with the postsynaptic neuron, that is, the connections bring inputs to a neuron from neurons located at an earlier stage (Dayan and Laurence F. Abbott 2001b). The firing rate of input p ($p = 1, 2, \ldots, P$) can be denoted by $r_{pre,p}(t)$ [5]. The firing rates of all inputs together can be denoted by the input vector $\mathbf{r}_{pre}$ having P components. Also, an input neuron makes connection onto the output neuron with a *synaptic strength* (or *synaptic weight* or simply *weight*) w_p. All synaptic weights can be collectively denoted by the synaptic weight vector $\mathbf{w}$ (Dayan and Laurence F. Abbott 2001b).

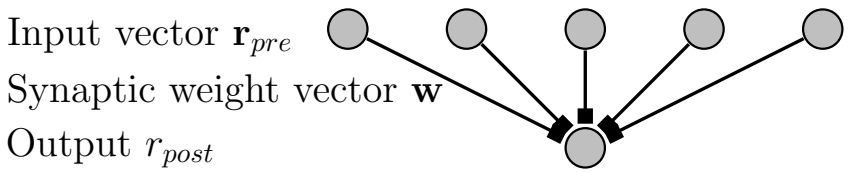


Figure 3.1: Feedforward network with P input (or presynaptic) rates $\mathbf{r}_{pre}$, 1 output (or postsynaptic) rate r_{post}, and a feedforward synaptic weight vector $\mathbf{w}$. Squares represent connections from inputs to output. The connections can be either excitatory or inhibitory. Inspired from Dayan and Laurence F. Abbott 2001b.

[5] The firing rates that are used hereafter are the time-dependent firing rate defined in section 3.3.1. However, to avoid an overload of notations, the notation r will often be used. This notation should not be understood as the spike-count rate! It should rather be understood as an abuse of language to simplify notations.

As a reminder, the first step consists in modeling how the total synaptic current I_{syn} received at the postsynaptic neuron depends on the presynaptic firing rates $\mathbf{r}_{pre}$. If a spike arrives at the presynaptic neuron p at time $t = 0$, then the synaptic current received by the postsynaptic neuron at time t is modeled as

$$I_{syn,p}(t) \approx w_p K_{syn}(t) \tag{3.6}$$

with w_p the weight as described above and $K_{syn}(t)$ the synaptic kernel (Dayan and Laurence F. Abbott 2001b) or filter (Gerstner et al. 2014 Ch 15).

The synaptic kernel $K_{syn}(t) \geq 0$ represents the time course, that is the time evolution, of the synaptic current of an input neuron in response to a presynaptic spike occurring at time $t = 0$. Moreover, the kernel is assumed to be the same for all input neurons and normalized to 1 for all positive times, that is

$$\int_{-\infty}^{+\infty} K_{syn}(t)\ dt = \int_{0}^{+\infty} K_{syn}(t)\ dt = 1$$

(Dayan and Laurence F. Abbott 2001b).

The amplitude and the sign of the synaptic current $I_{syn,p}$ are then determined by the weight w_p. If the coupling is excitatory (inhibitory), then $w_p > 0$ ($w_p < 0$) (Dayan and Laurence F. Abbott 2001b; Gerstner et al. 2014; Gjorgjieva et al. 2021b; Wilson and Cowan 1972).

Assuming that the effects of the spikes in the spike sequence of input neuron p can be summed linearly, the total synaptic current received by the postsynaptic neuron at time t due to the spikes that occurred at times $t_i < t$ in input neuron p is given by

$$I^{tot}_{syn,p}(t) = w_p \sum_{t_i<t} K_{syn}(t - t_i) = w_p \int_{-\infty}^{t} K_{syn}(t - \tau)\rho_p(\tau)\ d\tau \tag{3.7}$$

with $\rho_p(\tau)$ the neural response function of input neuron p (Dayan and Laurence F. Abbott 2001b). Moreover, if no non-linear interactions between synaptic currents from different input neurons exist, then the total synaptic current received by the postsynaptic neuron at time t from all input (or presynaptic) neurons is obtained by summing over the input neurons,

$$I_{syn}(t) = \sum_{p=1}^{P} w_p \int_{-\infty}^{t} K_{syn}(t - \tau) \cdot \rho_p(\tau)\ d\tau \tag{3.8}$$

Since the neural response function $\rho_p(\tau)$ can be approximated by the firing rate $r_{pre,p}(\tau)$, the total synaptic current can be expressed using the firing rates of the presynaptic neurons

as

$$I_{syn}(t) = \sum_{p=1}^{P} w_p \int_{-\infty}^{t} K_{syn}(t-\tau) \cdot r_{pre,p}(\tau)\ d\tau \tag{3.9}$$

(Dayan and Laurence F. Abbott 2001b).

For rate models, the synaptic kernel is usually a decaying exponential

$$K_{syn}(t) = \frac{1}{\tau_{syn}} \exp\left(-\frac{t}{\tau_{syn}}\right)$$

with τ_{syn} the synaptic time constant (Brunel 2021; Dayan and Laurence F. Abbott 2001b; Gerstner et al. 2014; Wilson and Cowan 1972). Using this kernel, the synaptic current I_{syn} at the postsynaptic neuron can be described with a differential equation (by differentiating Eq. (3.9) with respect to time on both sides)

$$\tau_{syn} \frac{dI_{syn}(t)}{dt} = -I_{syn} + \sum_{p=1}^{P} w_p r_{pre,p} = -I_{syn}(t) + \mathbf{w} \cdot \mathbf{r}_{pre} \tag{3.10}$$

with $\mathbf{w} \cdot \mathbf{r}_{pre}$ the dot product between weight and input vectors (Dayan and Laurence F. Abbott 2001b).

The second step for building a rate model consists in determining the output rate of the postsynaptic neuron based on the knowledge of I_{syn}.

For constant synaptic current ($I_{syn}(t) = I_{syn},\ \forall t$), the firing rate of the postsynaptic neuron is expressed as a function of this current, that is

$$r_{post} = \Phi(I_{syn}) \tag{3.11}$$

with $\Phi(.)$ the *activation function* (Dayan and Laurence F. Abbott 2001b) or static *transfer function* (Brunel 2021; Brunel and Lavigne 2009) or again *gain function* (Gerstner et al. 2014) or $f-I$ *curve* (Brunel and Lavigne 2009; Gerstner et al. 2014). The transfer function is usually a sigmoid function (Dayan and Laurence F. Abbott 2001b; Gjorgjieva et al. 2021a; Gjorgjieva et al. 2021b; Wilson and Cowan 1972) but it can be (threshold) linear (L. F. Abbott 1991; Gjorgjieva et al. 2021a), a hyperbolic tangent (Franci 2023a; Gjorgjieva et al. 2021a) or others (Gjorgjieva et al. 2021a). The tranfer function indicates how a neuron (or a population) integrates its presynaptic inputs in order to determine how it should respond; what decision should be made (Gjorgjieva et al. 2021a; Gjorgjieva et al. 2021b).

For a time-dependent synaptic current $I_{syn}(t)$, the firing rate of the postsynaptic

neuron is modeled as a low-pass filtered version of the steady state firing rate $r_{post} = \Phi(I_{syn}(t))$. In simpler terms, the output firing rate evolves as

$$\tau \frac{dr_{post}(t)}{dt} = -r_{post}(t) + \Phi(I_{syn}(t)) \tag{3.12}$$

with τ the time constant determining how rapidly r_{post} averages $\Phi(I_{syn}(t))$ (Dayan and Laurence F. Abbott 2001b).

Thus, as a summary, the rate model is given by

$$\begin{cases} \tau_{syn} \frac{dI_{syn}(t)}{dt} & = -I_{syn}(t) + \mathbf{w} \cdot \mathbf{r}_{pre} \\ \tau \frac{dr_{post}(t)}{dt} & = -r_{post}(t) + \Phi(I_{syn}(t)) \end{cases} \tag{3.13}$$

If $\tau \ll \tau_{syn}$, then the dynamics of r_{post} is extremely fast so that r_{post} can be considered to be instantaneously equal to its steady state value, *i.e.* $r_{post} = \Phi(I_{syn}(t))$. The model (3.13) reduces to

$$\tau_{syn} \frac{dI_{syn}(t)}{dt} = -I_{syn}(t) + \mathbf{w} \cdot \mathbf{r}_{pre} \qquad \text{with} \qquad r_{post} = \Phi(I_{syn}) \tag{3.14}$$

On the other hand, if $\tau \gg \tau_{syn}$, which is usually the case, then the dynamics of $I_{syn}(t)$ is extremely fast so that $I_{syn}(t)$ can be considered to be instantaneously equal to its steady state value, *i.e.* $I_{syn}(t) = \mathbf{w} \cdot \mathbf{r}_{pre}$. The model (3.13) reduces to

$$\tau \frac{dr_{post}(t)}{dt} = -r_{post}(t) + \Phi(\mathbf{w} \cdot \mathbf{r}_{pre}) \tag{3.15}$$

(Dayan and Laurence F. Abbott 2001b).

When referring to rate models, one usually refers to Eq. (3.15).

3.3.3 Feedforward networks

What happens if the network has Q output neurons as in Figure 3.2?

In that case, the scalar output becomes an output vector $\mathbf{r}_{post}$ and the vector of weights becomes a $Q \times P$ matrix $\mathbf{W}$ whose element w_{qp} represents the connection strength from input neuron p to output neuron q. The rate model (Eq. (3.15)) then becomes

$$\tau \frac{d\mathbf{r}_{post}}{dt} = -\mathbf{r}_{post} + \Phi(\mathbf{W} \cdot \mathbf{r}_{pre}) \qquad \text{or} \qquad \tau \frac{dr_{post,q}}{dt} = -r_{post,q} + \Phi\left(\sum_{p=1}^{P} w_{qp} r_{pre,p}\right) \tag{3.16}$$

with $p = 1, 2, \ldots, P$ and $q = 1, 2, \ldots, Q$ (Dayan and Laurence F. Abbott 2001b; Gjorgjieva et al. 2021a; Gjorgjieva et al. 2021b).

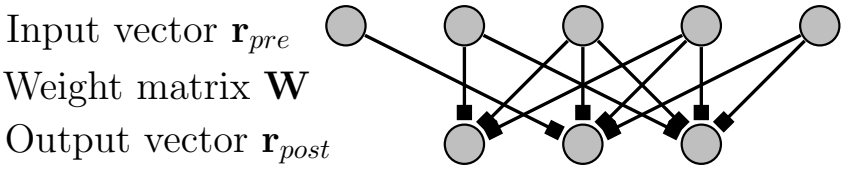


Figure 3.2: Feedforward network with P input (or presynaptic) rates $\mathbf{r}_{pre}$, Q output (or postsynaptic) rates $\mathbf{r}_{post}$, and a feedforward synaptic weight matrix $\mathbf{W}$. Squares represent connections from inputs to outputs. The connections can be either excitatory or inhibitory. Inspired from Dayan and Laurence F. Abbott 2001b.

Feedforward models are extensively used to study synaptic plasticity and learning but they can be used to model and compute coordinate transformations involved in reaching tasks (Dayan and Laurence F. Abbott 2001b).

3.3.4 Recurrent networks

Similarly to feedforward networks, recurrent networks have input neurons making feedforward connections to output neurons but, in addition, output (or input) neurons make *recurrent connections* with each other. In other words, recurrent connections interconnect neurons that are considered to be at the same stage (Dayan and Laurence F. Abbott 2001b). An example is provided in Figure 3.3. The rate model then becomes

$$\tau \frac{d\mathbf{r}_{post}}{dt} = -\mathbf{r}_{post} + \Phi(\mathbf{W}\cdot\mathbf{r}_{pre} + \mathbf{M}\cdot\mathbf{r}_{post}) = -\mathbf{r}_{post} + \Phi(\mathbf{M}\cdot\mathbf{r}_{post} + \mathbf{I}) = -\mathbf{r}_{post} + \Phi(\tilde{\mathbf{W}}\cdot\mathbf{r}) \quad (3.17)$$

with $\mathbf{M}$ the weight matrix of the recurrent connections, $\mathbf{I} = \mathbf{W}\cdot\mathbf{r}_{pre}$ the total feedforward input to each output neuron in the network, $\tilde{\mathbf{W}} = (\mathbf{W}, \mathbf{M})$ the "agglomerate" weight matrix and $\mathbf{r} = (\mathbf{r}_{pre}, \mathbf{r}_{post})^T$ the vector of all neurons.

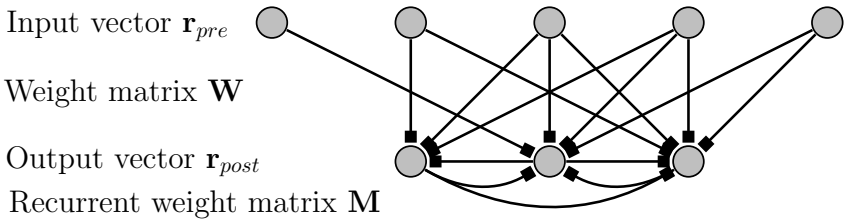


Figure 3.3: Recurrent network with P input (or presynaptic) rates $\mathbf{r}_{pre}$, Q output (or postsynaptic) rates $\mathbf{r}_{post}$, a feedforward synaptic weight matrix $\mathbf{W}$, and a recurrent weight matrix $\mathbf{M}$. Squares represent connections from inputs to outputs and from outputs to other outputs. The connections can be either excitatory or inhibitory. Inspired from Dayan and Laurence F. Abbott 2001b.

Recurrent networks offer more complex dynamics than feedforward networks, but in return they are more difficult to analyze. Advantages of recurrent networks are that they can perform selective amplification of a stimulus feature along eigendirection(s) of the network or they can be used to model memory (Dayan and Laurence F. Abbott 2001b).

3.3.5 Excitatory-Inhibitory networks

Neurons are usually categorized as either excitatory or inhibitory because their effect on the postsynaptic neurons is excitatory or inhibitory. In that case, the output vector is typically partitioned into two vectors: $\mathbf{v}_E$, the firing-rate vector of the excitatory neurons and $\mathbf{v}_I$, the firing-rate vector of the inhibitory neurons. Their dynamics is then modeled separately as well but the equations remain nevertheless coupled

$$\begin{cases} \tau_E \frac{d\mathbf{v}_E}{dt} = -\mathbf{v}_E + \Phi_E(\mathbf{M}_{EE} \cdot \mathbf{v}_E + \mathbf{M}_{EI} \cdot \mathbf{v}_I + \mathbf{I}_E) \\ \tau_I \frac{d\mathbf{v}_I}{dt} = -\mathbf{v}_I + \Phi_I(\mathbf{M}_{IE} \cdot \mathbf{v}_E + \mathbf{M}_{II} \cdot \mathbf{v}_I + \mathbf{I}_I) \end{cases} \tag{3.18}$$

There are now four different weight matrices describing the four types of connections: $E \to E$ ($\mathbf{M}_{EE}$ whose elements are positive), $I \to E$ ($\mathbf{M}_{EI}$ whose elements are negative), $E \to I$ ($\mathbf{M}_{IE}$ whose elements are positive), $I \to I$ ($\mathbf{M}_{II}$ whose elements are negative). Also, due to the different nature of the neurons, the latter may have different time constants, transfer functions and feedforward inputs (Dayan and Laurence F. Abbott 2001b; Wilson and Cowan 1972).

This type of networks is particularly useful to analyze oscillatory phenomena (Dayan and Laurence F. Abbott 2001b).

3.4 Network of interacting populations

The concepts for a network of neurons can be easily transposed to a network of *populations*. Instead of using a vector of firing rates of neurons, one needs to use a vector of firing rates of *populations*, that is, a single scalar will describe the firing rate of an entire population made of a large number of neurons. As seen in subsection 3.3.1, the firing rate of the population will thus be an average over the neurons making up the population. In order for this method to be valid, the population must be homogeneous. This implies that

- Only neurons with similar properties should be grouped together, that is, they have similar parameters and are statistically indistinguishable (Gerstner et al. 2014 Ch 12).
- The number of neurons within a population should be large (and finite) (Gerstner et al. 2014 Ch 12).

- Each neuron within a population receives inputs from many other neurons, either from the same population, or from other populations, or both (Gerstner et al. 2014 Ch 12).
- All neurons within a population receive the same (external) input (Gerstner et al. 2014 Ch 12).

When the populations are homogeneous, then the rate model (3.15) and its variants (Eqs. (3.16, 3.17, 3.18)) can be transposed to populations by replacing $\mathbf{r}(t)$ by $\bar{\mathbf{r}}(t)$, the population activity or equivalently, the firing rates of the entire populations (Gerstner et al. 2014; Gjorgjieva et al. 2021a; Gjorgjieva et al. 2021b).

3.4.1 Wilson-Cowan models

When the rate model uses populations of excitatory neurons and populations of inhibitory neurons, the model is often referred to as a *Wilson-Cowan* (WC) rate model. The original Wilson-Cowan model uses two variables, each coding for a population:

- $E(t)$ = proportion of excitatory cells firing per unit time at the instant t.
- $I(t)$ = proportion of inhibitory cells firing per unit time at the instant t.

Their dynamics is then modeled as (Wilson and Cowan 1972)

$$\begin{cases} \tau_E \frac{d\bar{E}}{dt} = -\bar{E} + (1 - s\bar{E})\Phi_E(w_{EE}\bar{E} - w_{EI}\bar{I} + I_E) \\ \tau_I \frac{d\bar{I}}{dt} = -\bar{I} + (1 - s\bar{I})\Phi_I(w_{IE}\bar{E} - w_{II}\bar{I} + I_I) \end{cases} \tag{3.19}$$

with $\bar{E}$ and $\bar{I}$ the average firing rates, $\Phi_E(.)$ and $\Phi_I(.)$ two sigmoid functions and s the fraction of cells that are unavailable for firing at time t. All weights here have positive values. The premultiplicative factors $(1 - s\bar{E})$ and $(1 - s\bar{I})$ do not actually make too much difference in the analysis of (3.19). The fraction is therefore often set to zero ($s = 0$) (Ermentrout and Terman 2010). The WC model then reduces to

$$\begin{cases} \tau_E \frac{d\bar{E}}{dt} = -\bar{E} + \Phi_E(w_{EE}\bar{E} - w_{EI}\bar{I} + I_E) \\ \tau_I \frac{d\bar{I}}{dt} = -\bar{I} + \Phi_I(w_{IE}\bar{E} - w_{II}\bar{I} + I_I) \end{cases} \tag{3.20}$$

which corresponds to a 2-population case of (3.18), hence the reference to a Wilson-Cowan model when using excitatory and inhibitory populations.

Part II

Language, Memory and Semantic priming

Chapter 4

Language and Memory

This chapter aims at introducing the concept of memory in order to understand why memory is an important component of one's behavior. The link between language and memory is also discussed.

Memory is the general term accounting for the ability to *acquire, retain* and *make use* of skills and/or knowledge (Tulving 2000). More important than that, memory is crucial in human beings' life because it allows them to *adapt* their behavior to what they have experienced previously (Schacter 2000; Vandewalle 2020-2021). For this reason, the concept of memory is often linked to that of *learning*. The distinction between the two concepts is often blurry but learning should be thought of as the process of *acquiring slowly* a new skill or knowledge, whereas memory should be thought of as *acquiring "instantly", making use of, expressing* that new skill or knowledge that has been learned (Thompson 2000; Tulving 2000).

Describing memory is difficult because it can take many forms and kinds. The usual classification is then made using *memory systems* (Schacter 2000) and, in a more integrated fashion, *brain systems* (Thompson 2000).

4.1 Memory systems

Schacter 2000 uses the terms "memory system" as "a set of interrelated brain processes that allow one to store and retrieve a specific type of information", and as a system "that can be characterized using lists of properties that describe its mode of operation". As a result, (at least) five memory systems can be distinguished.

4.1.1 Episodic memory

Conscious (or aware) recollection of personal past events that occurred at a particular time and place is referred to as *episodic memory* (Schacter 2000). In other words, episodic memory contains all memories that are unique to one's life. It is a self-centered point of view.

4.1.2 Semantic memory

Unlike episodic memory, *semantic memory* refers to the knowledge that could be shared by anyone or at least by a significant number of individuals. Put differently, semantic memory refers to facts and (verbal) concepts that are not linked to a specific time and/or place of one's life (Kumar 2021; Schacter 2000). A typical example would be the language spoken by individuals belonging to a population. More simply, knowing that the Eiffel Tower is located in Paris is part of semantic memory. Semantic memory is therefore an allocentric point of view. Moreover, semantic memory is involved in the representation of associative and conceptual (or featural) information (Schacter 2000). In other words, semantic memory is involved in the identification of the *meaning* of words, objects, ...

4.1.3 Perceptual representation system

The *perceptual representation system* is involved in identifying words, objects, ... based on their form and structure. This identification occurs before any further processing within the semantic memory (Schacter 2000).

4.1.4 Procedural memory

Acquiring new skills and habits, that is, knowing *how* to do a task rather than knowing *what* is the task, is often referred to as *procedural memory*. Repetition in doing the task is the key to construct procedural memories (Schacter 2000).

4.1.5 Working memory

Finally, *working memory* focuses on the retention of a small amount of information over a period of a few seconds (Schacter 2000; Vandewalle 2020-2021). Since working memory can only serve short-term purposes, typical applications are basic cognitive activities such as reasoning, problem solving, ... (Schacter 2000). For instance, memorizing a phone number before writing it down is in the scope of working memory.

4.2 Brain systems

Similarly to memory systems, memory can be categorized based on the brain systems that are responsible for a particular type of memory, but also based on the purpose(s) served by a particular type of memory as being embodied into the brain.

4.2.1 Short-term VS Long-term memory

A first distinction appears according to the time scale of the retention of information.

Short-term memory, whose time scale goes from seconds to minutes, retains information briefly. For this reason, short-term memory is sometimes referred to as working memory as seen above. Moreover, short-term memory contains all the information that one is aware at any given moment in time. In addition, short-term memory also contains information that is *retrieved* from other types of memory. This retrieved information is then used directly after retrieval (Thompson 2000).

On the other hand, *long-term* or (relatively) *permanent* memory contains information that is retained over periods of days, weeks, months and years. Long-term memory can further be categorized as follows (Thompson 2000).

4.2.2 Declarative VS Non-declarative memory

Declarative and non-declarative memories are both long-term memories but the distinction appears when the access to these types of memories is made consciously (also known as explicit) or unconsciously (also referred to as implicit) (Thompson 2000).

Declarative memory is said to be an explicit memory and usually includes episodic and semantic memories (Schacter 2000; Thompson 2000; Vandewalle 2020-2021).

On the other hand, non-declarative memory is said to be an implicit memory and is assumed to contain everything that is neither in declarative memory nor in short-term/working memory (Thompson 2000).

4.2.3 Nonassociative memory

Nonassociative memory is a type of non-declarative long-term memory that refers to the phenomena of *habituation* and *sensitization* (Thompson 2000). Put differently, this type of memory accounts for the changes in already existing responses to stimuli. Habituation will decrease the response to a repeated stimulus whereas sensitization will increase the response (Thompson 2000).

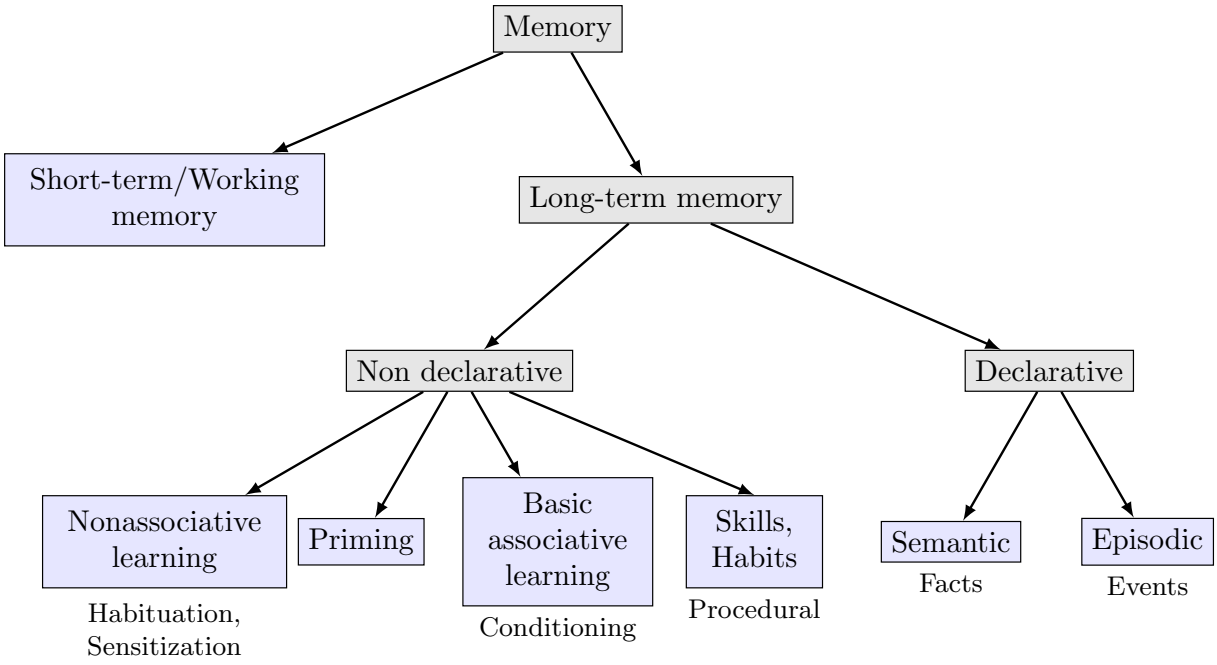


Figure 4.1: Memory classification. Adapted from Thompson 2000.

4.2.4 Basic associative memory

Basic associative memory contains associations between stimuli and/or associations between stimuli and responses. The most known process that makes up these associations is the *conditioning* (Thompson 2000). Conditioning is the process of associating a response A, initially due to a stimulus A, to a stimulus B, that was initially associated to a response B (usually a neutral response). For instance, listening to a tone (stimulus B) is associated to a neutral response (response B), but seeing a spider (stimulus A) is associated with fear (response A). Conditioning consists in repeatedly presenting together the stimuli "spider" and "tone" (that will give a "fear" response). Once conditioning is over, listening to the tone is then associated to fear (Vandewalle 2020-2021).

4.2.5 Priming

Priming is the phenomenon that appears when processing a concept/stimulus enhances the processing of a following related concept/stimulus (compared to when the following concept/stimulus is unrelated) (Heyman, De Deyne, and Storms 2013; McNamara 2005; Thompson 2000). Priming is thus important for human beings when interacting with their peers using language (see next).

4.3 Summary of memory classification

A summary of the classification of memory and brain systems can be found in Figure 4.1.

4.4 Link between language and memory

Language is not (absolutely) needed to have sophisticated learning and/or memory capabilities (Tulving 2000). For instance, even though bees do not talk, they remember and know how to go back to the hive once they fly in the outside world and pollinate flowers. However, when language comes into play as in human beings, memory and language certainly and greatly influence each other (Tulving 2000).

In order to interact with their peers, human beings use, among other things, language: either one listens or one speaks. When listening, one has first to *encode* the information received from the speaker and then *process* it. When speaking, one has to *retrieve* information (*i.e.* making use of vocabulary, grammar, ...) from memory, in particular *semantic* memory, in order to produce a coherent speech. Also, when being exposed to the same stimulus such as words (nouns, verbs, ...), sentences, ..., (semantic) memory becomes "updated" as well in the sense that new words with their meaning become part of the mental lexicon stored in memory.

In addition, this interaction happens in real-time in the sense that processing information does not take more than seconds (and usually take less). This efficiency is made possible through processes such as priming that speeds up the reaction to a stimulus.

Chapter 5

Semantic priming

This chapter aims at making an overview of the semantic priming paradigm, used as the context for the core computational study. Tasks, semantic relations and models are discussed.

As explained in subsection 4.1.2, semantic memory contains general knowledge (Hutchison 2003; Kumar 2021; Schacter 2000). Considering language, semantic memory would thus contain information about identity, spelling, pronunciation and especially meaning of a concept (Hutchison 2003; Kumar 2021); the most commonly used concepts being words and pictures (Sperber et al. 1979). However, it is still unclear *how* semantic memory is *structured*, that is, how this knowledge is *represented* in semantic memory exactly (Hutchison 2003; Kumar 2021).

The usual experimental procedure to investigate the semantic memory structure is called the *semantic priming paradigm* (Hutchison 2003). In this paradigm, participants are transiently presented a first concept (usually a word), called *prime*, that is followed after a controlled delay by another semantically (un)related concept, called *target* (Brunel and Lavigne 2009; Heyman, Bruninx, et al. 2018; Heyman, De Deyne, and Storms 2013; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011; McNamara 2005; Sperber et al. 1979). Participants are then asked to perform a task on the target concept usually (see section 5.1) (Brunel and Lavigne 2009; Heyman, De Deyne, and Storms 2013; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011).

The *semantic priming effect* comes then from the observation that participants process faster the target concept when it is related to the prime than when the target concept is unrelated to the prime. In other words, the reaction time of participants for related targets is smaller than that for unrelated targets (Brunel and Lavigne 2009; Heyman, Bruninx, et al. 2018; Heyman, De Deyne, and Storms 2013; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011; McNamara 2005; Sperber et al. 1979). The priming effects

are then usually computed by subtracting the average reaction time (RT) for the related targets from the average reaction time for the unrelated targets (*i.e.* $RT_U - RT_R$), although this computation does not give highly reliable figures (Heyman, Bruninx, et al. 2018). Thus, the larger the difference, the larger the priming effects and the faster the processing of related targets.

The "semantic" in "semantic priming" refers to *semantic* (*i.e.* “true relations of meaning” McNamara 2005) and/or *associative* relationships between the prime and the target (Heyman, Bruninx, et al. 2018; McNamara 2005). For instance, the concepts *dog* and *goat* are semantically related (because both belong to the same category *animal* and share features) whereas the concepts *light* and *dark* (antonyms) are associatively related (because when presenting the prime *light*, the primary concept that comes to mind is often *dark*). Actually, it is very rare to find *pure* semantic or associative relationships; these two act more like the extremes of a whole spectrum of relatedness (Heyman, Bruninx, et al. 2018; Hutchison 2003; Kumar 2021; McNamara 2005).

There are a few reasons why semantic priming is used to investigate the semantic memory structure and even used in general. First, when participants perform semantic priming tasks, semantic priming often occurs *automatically*, that is, the participants are not aware of this processing; they do not use it consciously. This is the fundamental reason why semantic priming is thought to *reflect* the semantic memory structure (Heyman, Bruninx, et al. 2018; Heyman, De Deyne, and Storms 2013; Hutchison 2003; Kumar 2021; McNamara 2005). Then, semantic priming occurs in wide variety of cognitive tasks such as lexical decision, pronunciation, ... (see section 5.1) (McNamara 2005). Finally, semantic priming can serve as a tool to explore other aspects of cognition and perception, namely word recognition, sentence comprehension, ... (McNamara 2005).

Unfortunately, priming effects are difficult to evaluate reliably (Heyman, Bruninx, et al. 2018) because they depend on the cognitive task, the timing of the experimental procedure, the pool of participants that is considered, the type and the (association) strength of the relationship between primes and targets, ... (Brunel and Lavigne 2009; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011). Models of priming attempt to take into account all or some of these factors to explain the dynamics of semantic memory (see section 5.3).

5.1 Experimental tasks

As already introduced above, a wide variety of tasks can be used to assess priming effects experimentally. The most frequently used task is called the *lexical decision task* (LDT) (Brunel and Lavigne 2009; Heyman, De Deyne, and Storms 2013; Hutchison 2003; Kumar 2021; McNamara 2005; Sperber et al. 1979). In this task, the stimuli are existing and correctly spelled words, and non existing and meaningless letter strings called non-words or pseudo-words. On each trial, participants first have to read (silently or aloud) the prime. Secondly, they have to decide whether the following target is a word or a non-word (Heyman, De Deyne, and Storms 2013; McNamara 2005). The finding is thus that responses are faster and more accurate when the target is semantically related to the prime, as predicted by the semantic priming effect. However, this task is prone to participants making use of strategies to increase their performance (Heyman, De Deyne, and Storms 2013; Hutchison 2003). Since these strategies cause an artefact in the results, variants of the LDT have been designed, notably a *continuous* LDT (cLDT) and a *letter decision task* (Heyman, De Deyne, and Storms 2013) to control for the possibility of strategies. A continuous LDT requires participants to decide, in addition to the target, whether the prime is a word or a non-word. The letter decision task requires participants to fill in a one-letter gap in the prime and the target words (e.g. *tom_ to-lett_ ce* would give *tom<u>a</u>to-lett<u>u</u>ce*) (Heyman, De Deyne, and Storms 2013).

The secondly most used task is the so-called *word naming* or *pronunciation* (Brunel and Lavigne 2009; Hutchison 2003; McNamara 2005). In this task, the stimuli are correctly spelled words and participants have to read aloud the target (and sometimes the prime as well) as fast and accurately as possible. The performance is again better when the target is semantically related to the prime (McNamara 2005).

5.2 Semantic relationships

Priming effects depend on the type of relationship that exists between a prime and a target. A non-exhaustive list of these relationships, categorized as "semantic" or "associative", is given in Table 5.1 with prime-target examples in parentheses.

More simply, the type of relationship between a prime and a target can be categorized as *direct* or *indirect*. Moreover, the number of steps (or, in a similar way, the number of associated concepts) separating the target from the prime further qualifies the relationship.

"Associative"	"Semantic"
Synonyms (*afraid-scared*)	Category members/coordinates (natural: *sheep-goat*; artificial: *table-chair*)
Antonyms (*light-dark*)	Supraordinate (*dog-animal*)
Property relations (perceptual: *canary-yellow*; functional: *broom-sweep*)	Subordinate (*animal-dog*)
Instrument relation (*broom-floor*)	
Script relation (*orchard-apple*)	
Phrasal associate (forward: *baby-boy*; backward: *boy-baby*)	

Table 5.1: Classification of prime-target relationships. Relationships are labeled as either "associative" or "semantic" but one should keep in mind that purely semantic or associative relations rarely exist. Inspired and adapted from Hutchison 2003.

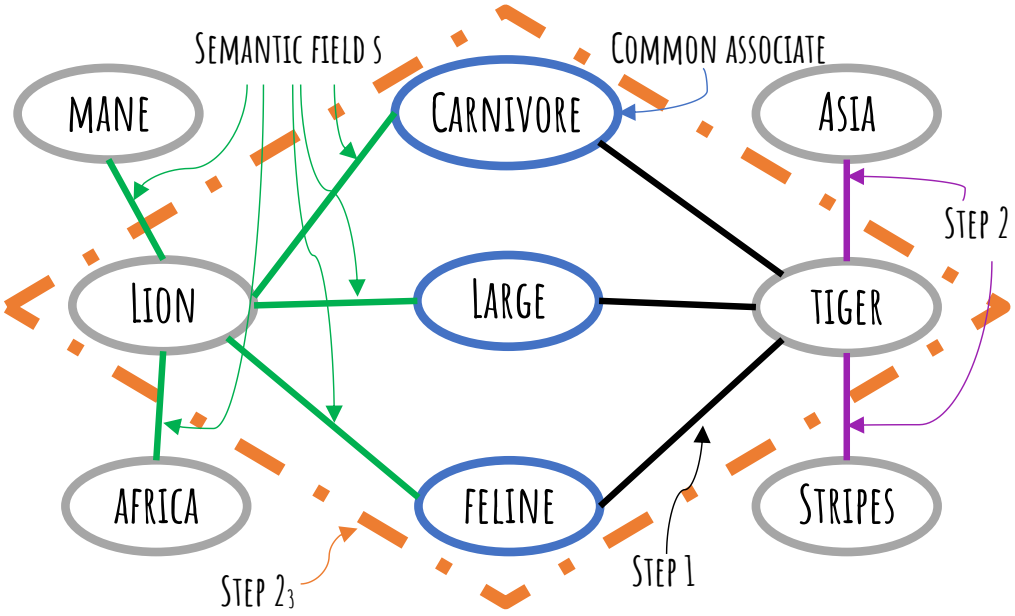


Figure 5.1: Types of prime-target relationships using an example of a semantic network. Concepts are represented by nodes. The links between the concepts represent the relations. Blue nodes account for the common associates to both the prime *lion* and the target *tiger*. Black lines illustrate direct (Step 1) relations between concepts. Purple lines illustrate indirect (Step 2) relations between concepts. Green lines account together for the semantic field *s* of the concept *lion*. Orange dash-dotted rhombus delimits a subnetwork accounting for a Step 2_3 priming for the prime-target pair *lion-tiger*. Example of semantic network adapted from McNamara 2005.

Step 1 priming (Figure 5.1 black lines) is a direct association/relation between a prime and a target (e.g. *carnivore-tiger*). The strength of this relation can be quantified using the *association strength*, defined as the percentage of people responding primarily with a given target word when given a prime cue (Hutchison 2003). Thus, the magnitude and the onset of Step 1 priming effects depend on the association strength a (Brunel and Lavigne 2009; Lavigne, Dumercy, and Darmon 2011).

Step 2 priming (Figure 5.1 purple lines), also known as *mediated priming* (Hutchison 2003; Kumar 2021), corresponds to an indirect association/relation through a *common associate* (or *mediator* (Hutchison 2003; Kumar 2021)) (e.g. *lion(-carnivore)-tiger*). Since the *semantic distance* separating the target from the prime is larger in two-step priming than in one-step priming, two-step priming effects are reported to be weaker than those of one-step priming but stronger than those of three-step priming (involving two intermediate associates) (Brunel and Lavigne 2009; Lavigne, Dumercy, and Darmon 2011).

Actually, two-step priming effects can also depend on the total number n of common associates to both the prime and the target, and is usually referred to as *Step* 2_n priming (Figure 5.1 orange dash-dotted rhombus including the nodes and edges within it). This priming is sometimes reported to be equivalent as, or even stronger than one-step priming (Brunel and Lavigne 2009; Lavigne, Dumercy, and Darmon 2011).

Generalization to N steps would give *Step* N_n^s with n the number of common associates to both the prime and the target, and s the *semantic field* (Figure 5.1 green lines) of a given word, that is the total number of associates that this given word has (Brunel and Lavigne 2009; Lavigne, Dumercy, and Darmon 2011).

5.3 Models of priming

All studies agree with the fact that semantic priming exists and should reflect the organization and the structure of semantic memory. However, a vivid debate about the *nature* (or *source*) of the semantic priming effects can be found in the semantic priming literature (Brunel and Lavigne 2009; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011). This debate is reflected in the different models of priming that are used to explain the representation of semantic memory, and to bridge the gap between the observed behaviors and the cellular levels responsible for these behaviors (Brunel and Lavigne 2009; Heyman, De Deyne, and Storms 2013).

According to association theories of semantic memory, the latter is assumed to be represented by a semantic network where concepts and features are encoded as whole

units, that is *nodes*. The network adopts more of a *localist* point of view (Brunel and Lavigne 2009; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011). A node makes a connection with another node if they are associated in some way. The link between these two nodes is then weighted with the association strength a characterizing the connection between the two nodes. Semantic priming effectively occurs in this network through a *spreading activation* mechanism. If the prime node is activated, then it leads to the activation of its direct neighboring nodes (if the weight of these connections is strong enough), and the network is traversed until the target node is reached and a response (*i.e.* activated or not) is made. The semantic distance separating the target from the prime reflects the reaction time needed to make a decision (Brunel and Lavigne 2009; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011).

On the other hand, *feature-based distributed* models represent concepts in semantic memory as a binary collection of features. Thus, nodes in a distributed network encode for single features that are either activated or not activated. Moreover, the network is fully connected. Activation or recall of a given concept in memory corresponds then to the convergence of the network to a particular pattern of activated features, that is an *attractor state* (Brunel and Lavigne 2009; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011). The distributed model is therefore similar to a Hopfield net (Gerstner et al. 2014 Ch 17). Semantically related concepts would then share common features, and the degree of overlap between two concepts' patterns would determine how similar the meaning of these two concepts is (Brunel and Lavigne 2009; Kumar 2021; Lavigne, Dumercy, and Darmon 2011). Semantic priming thus occurs through the facilitation of processing the target concept because the prime and the target concepts largely overlap. Two important remarks should be kept in mind as well: first, a concept is distributed over many different features but, in turn, a single feature can be contained in multiple concepts' representation. Second, priming occurs with direct associations between features rather than direct associations between concepts (Brunel and Lavigne 2009; Hutchison 2003; Kumar 2021; Lavigne, Dumercy, and Darmon 2011).

A drawback and limitation shared by both association-based and feature-based models is that they are unable to explain how *knowledge* of individual concepts and features was *learned* initially. For this reason, a third class of models, that is the *distributional semantic models* (DSM), attempts to overcome this issue.

"Distributional semantic models refer to a class of models that provide explicit mechanisms for how words or features for a concept may be learned from the natural language" (Kumar 2021). In other words, DSMs directly *build* representations of concepts and/or features by *extracting co-occurrence* of patterns and *inferring* associations between concepts and/or features, from texts with "natural language" (*i.e.* books, newspapers, arti-

cles, . . .) (Hutchison 2003; Kumar 2021). Two learning mechanisms, known as *error-free* and *error-driven* mechanisms, can be used by DSMs. An error-free learning mechanism is similar to a classic Hebbian learning mechanism ("Fire together, wire together") where learning of concepts/features occurs by identifying events that tend to co-occur in temporal proximity. On the other hand, an error-driven learning mechanism "posits that learning is accomplished by predicting events in response to a stimulus, and then applying an error-correction mechanism to learn associations" (Kumar 2021). Typically, words/features are represented by "environmental vectors" based on the texts and this representation is updated when co-occurrences are identified. Priming occurs via vector representation similarity (Kumar 2021).

Part III

Assessing model sensitivity and application to an experimental-like stimulus

Chapter 6

Model equivalence and parameter sensitivity assessment

6.1 Introduction

In section 4.4 and Chapter 5, it has been stressed that semantic priming is a cognitive process of crucial importance for the real-time and efficient interactions of human beings with their peers, and for the linguistic knowledge representation in semantic memory.

Brunel and Lavigne 2009 designed a cortical network model of priming that attempts to reproduce and explain behavioral findings on semantic priming experiments in humans. They investigated the magnitude of priming effects as a function of three key experimental parameters: the association strength between the concepts, the type of relationship between a prime and a target, and the time elapsed between the onset of the prime stimulus and the onset of the target stimulus, also referred to as the *stimulus onset asynchrony* (SOA). Their model also includes other (non-experimental) parameters but their values were kept constant.

In addition, in their model, Brunel and Lavigne 2009 used a transfer function that was computed analytically by N. Brunel and P.E. Latham[6], and that is thought to be qualitatively similar to that of cortical excitatory neurons. However, the shape of this transfer function significantly differs from the shape of a usual sigmoidal transfer function (as seen in subsection 3.3.2) and its use in numerical operations and manipulations is rather difficult. Moreover, it is unclear how and why this shape would be used physiologically.

For these reasons, this chapter focuses on assessing whether using a sigmoid transfer

[6]N. Brunel and P. E. Latham, "Firing Rate of the Noisy Quadratic Integrate-and-Fire Neuron," in Neural Computation, vol. 15, no. 10, pp. 2281-2306, 1 Oct. 2003, DOI: 10.1162/089976603322362365.

function (henceforth *Method* 2 with Φ_2) would give qualitatively the same dynamic behavior as the original model of Brunel and Lavigne 2009 (henceforth *Method* 1 with Φ_1). Moreover, this chapter will investigate the parameter sensitivity of the model of Brunel and Lavigne 2009 by varying the parameter values.

The structure for this chapter is as follows. As a first step, the full model of Brunel and Lavigne 2009 will be described, including its transfer function and its parameters. Then, the model will be simplified to a one-dimensional version and the exploration of its dynamics will be carried out. Finally, phase-portrait and bifurcation analyses will help determining the equivalence and the sensitivity of the model.

6.2 Brunel and Lavigne's full model description

Brunel and Lavigne 2009 (B&L) consider populations of cortical neurons that are *selective* to the same objects or concepts, that is, the population shows maintained or *persistent* activity following the presentation of those objects or concepts. They designed a simplified rate model: the dynamic behavior of a single population is described by the time evolution of its average firing rate (= dynamic variable) in a Wilson-Cowan type equation (Eq. (6.1a)). Moreover, the following assumptions are made.

Assumptions

1. The model consists of p *non-overlapping* populations of excitatory neurons coding for p distinct stimuli. In other words, neurons within a population are selective to a single object/concept, and each population codes for a different object/concept.

2. The global inhibitory current regulating the activity of all excitatory populations is proportional to the average activity of excitatory populations.

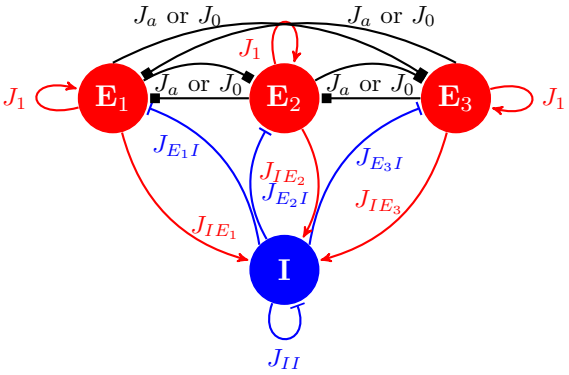
Assumption n° 2 actually implies that the transfer function for the global inhibitory population is linear. It also implies that the time scale of inhibitory population dynamics is much shorter than that of excitatory population dynamics. As a consequence, the inhibitory average firing rate can be set to its steady state value, that is the average activity of all excitatory populations.

All in all, the full model is mathematically expressed with Equations (6.1a), (6.1b), (6.1c) and Table 6.1, that describes the parameters[7].

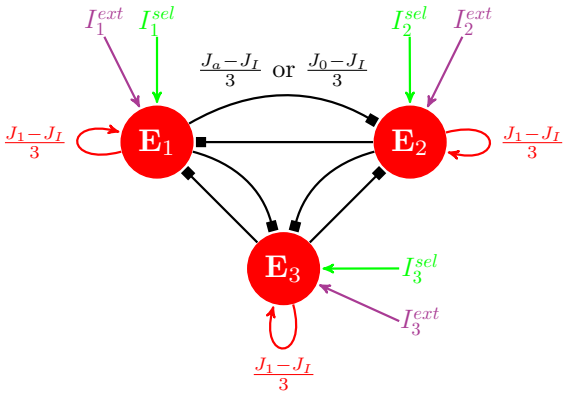
[7]The model from Brunel and Lavigne 2009 was reused by Lavigne, Dumercy, and Darmon 2011 to investigate the spike frequency adaptation phenomenon. This phenomenon is not of interest here but the same parameters had their value changed with respect to those in Brunel and Lavigne 2009. A comparison of values between these two studies seemed appropriate when considering parameter sensitivity.

Symb.	Description	Brunel and Lavigne 2009 value	Lavigne, *et al* (2011) value
p ($p_g p_i$)	Number of selective (excitatory) populations	100	99
p_g	Number of groups of selective (excitatory) populations	10	33
p_i	Number of selective (excitatory) populations within a single group	10	3
τ	Time constant of rate dynamics	10 [msec]	10 [msec]
a	Association strength between associated items	0.001-0.02 ($0 < a < 1$)	0.00675-0.00825 ($0 < a < 1$)
J_E	Average excitatory synaptic strength	3	3
J_S	Strength of synaptic potentiation (*i.e.* connection reinforcement)	3.65	3.65
J_1	Intrapopulation synaptic efficacy	$J_E + J_S = 6.65$	$J_E + J_S = 6.65$
J_0	Synaptic efficacy between nonassociated (or unrelated) populations (or items)	$J_E - J_S \frac{a(p_i-1)+1}{(1-a)(p_i-1)+p-p_i}$ ($< J_1$)	$J_E - J_S \frac{a(p_i-1)+1}{(1-a)(p_i-1)+p-p_i}$ ($< J_1$)
J_a	Synaptic efficacy between associated (or related) populations (or items)	$J_E + J_S \frac{a(p-p_i+1)-1}{(1-a)(p_i-1)+p-p_i} = J_0 + a(J_1 - J_0)$	$J_E + J_S \frac{a(p-p_i+1)-1}{(1-a)(p_i-1)+p-p_i} = J_0 + a(J_1 - J_0)$
J_I	Synaptic efficacy of global inhibition (non-selective)	$J_E = 3$	$J_E = 3$
r_{spont}	Spontaneous (or background) activity of each population	5 [Hz]	5 [Hz]
I_i^{ext}	Non-selective external input current (*i.e.* bias current) to population i	set to get $r_{spont} = 5$ [Hz] in the absence of selective external stimuli	set to get $r_{spont} = 5$ [Hz] in the absence of selective external stimuli
I_i^{sel}	Selective external input current to population i	150 [µA]	10-200 [µA]

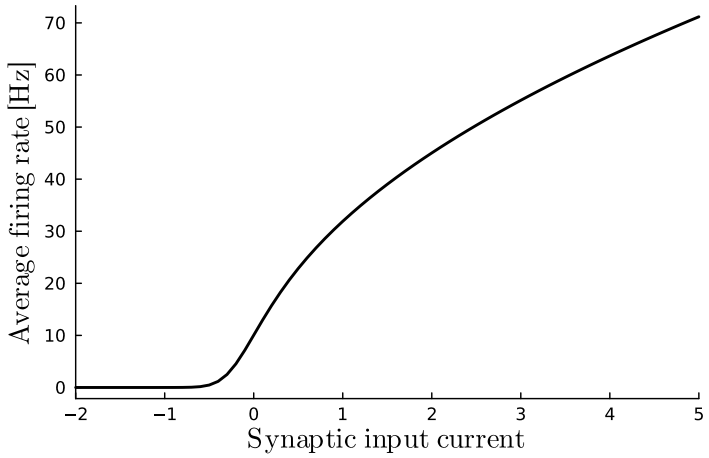
Table 6.1: Default parameters of the full model (Brunel and Lavigne 2009).



(a) Basic architecture.



(b) Model example with $p = 3$ and assumptions are made.



(c) Brunel and Lavigne 2009's transfer function

Figure 6.1: **(a)** Architecture of the excitatory-inhibitory network for $p = 3$. Inhibitory population (blue node I) is non-selective and applies global inhibition (flat blue arrows) to itself and to all excitatory populations (red nodes $E_i, i \in \{1, 2, 3\}$) selective to three distinct items. Black square arrows represent either excitatory (strength J_a) or inhibitory (strength J_0) connections depending on the relatedness between the corresponding items. Red arrows represent self excitatory feedback with strength J_1. **(b)** Same full model example as in **(a)** with assumptions made. Each population is characterized by its average firing rate r_i (Equation (6.1a)) and receives synaptic input from all populations, as well as from external sources. Purple arrows represent a non-selective external input current I_i^{ext}, that is a bias current, applied to obtain a spontaneous activity of $r_{spont} = 5$ Hz in each population. Green arrows stand for selective external input current I_i^{sel}. **(c)** Population $f - I$ curve (Equation (6.1c)). Adapted from Brunel and Lavigne 2009.

Eq. (6.1a) describes the time evolution of the average firing rate for each excitatory population $i = 1, \ldots, p$; Eq. (6.1b) mathematically expresses assumption n° 2, and Eq. (6.1c) describes the transfer function that Brunel and Lavigne 2009 used (Figure 6.1c). It specifies how the average firing rate of a population of excitatory neurons depends on the *total synaptic* inputs the population receives (as previously discussed in subsection 3.3.2). An example of the basic architecture of the excitatory-inhibitory network for $p = 3$ is shown in Figure 6.1a, and Figure 6.1b shows the corresponding network once the assumptions are made.

$$\tau \frac{dr_i}{dt} = -r_i + \Phi_1 \left(\frac{1}{p} \sum_{j=1}^{p} J_{ij} r_j + I_i^{ext} + I_i^{sel} - I_{inh} \right) \tag{6.1a}$$

$$I_{inh} = \frac{J_I}{p} \sum_{j=1}^{p} r_j \tag{6.1b}$$

$$\Phi_1(x) = \frac{1}{\tau_m \sqrt{\pi}} \left[\int_{-\infty}^{+\infty} \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48} \right) \, dz \right]^{-1} \tag{6.1c}$$

with $J_{ij} \in \{J_0, J_a, J_1\}$ the total synaptic strength from population j to population i, $\sigma = 0.5$ and $\tau_m = 10$ msec the membrane time constant. The other parameters can be found in Table 6.1. J_1, J_0 and J_I are chosen so that both the background state and the attractor states (*i.e.* single or multiple items activated) can be equilibria of the model.

6.3 Simplified one-dimensional version

To investigate the equivalence and the parameter sensitivity of the model from Brunel and Lavigne 2009, the corresponding one-dimensional model is derived. Thus, setting parameter p to one, the model from Brunel and Lavigne 2009 reduces to (Figure 6.2a)

$$\tau \frac{dr_T}{dt} = -r_T + \Phi(\underbrace{(J_1 - J_I)}_{w} \cdot r_T + \underbrace{I_{ext} + I_{sel}}_{I}) = -r_T + \Phi(w \cdot r_T + I) \tag{6.2}$$

where T stands for *target* population, τ is the time constant of rate dynamics (same value as in Table 6.1), w is the recurrent connection weight, I is the total external input current the population receives and Φ is a transfer function.

The model (6.2) actually amounts to study the behavior of a single population of excitatory neurons coding for a single (target) item.

The use of the 1D model is motivated by several reasons:

- No assumptions, in particular a minimum value, were made on parameter p. Thus,

varying p (and setting it to one) allows to investigate the model sensitivity to parameter p.

- A 1D model is the simplest model one can analyze without it to be trivial.
- This 1D model allows one to understand the dynamics and the behavior of a one-dimensional model but also of higher-order models since any model can usually be reduced to a 1D model (*i.e.* dominant eigendirection where all the dynamics happens).

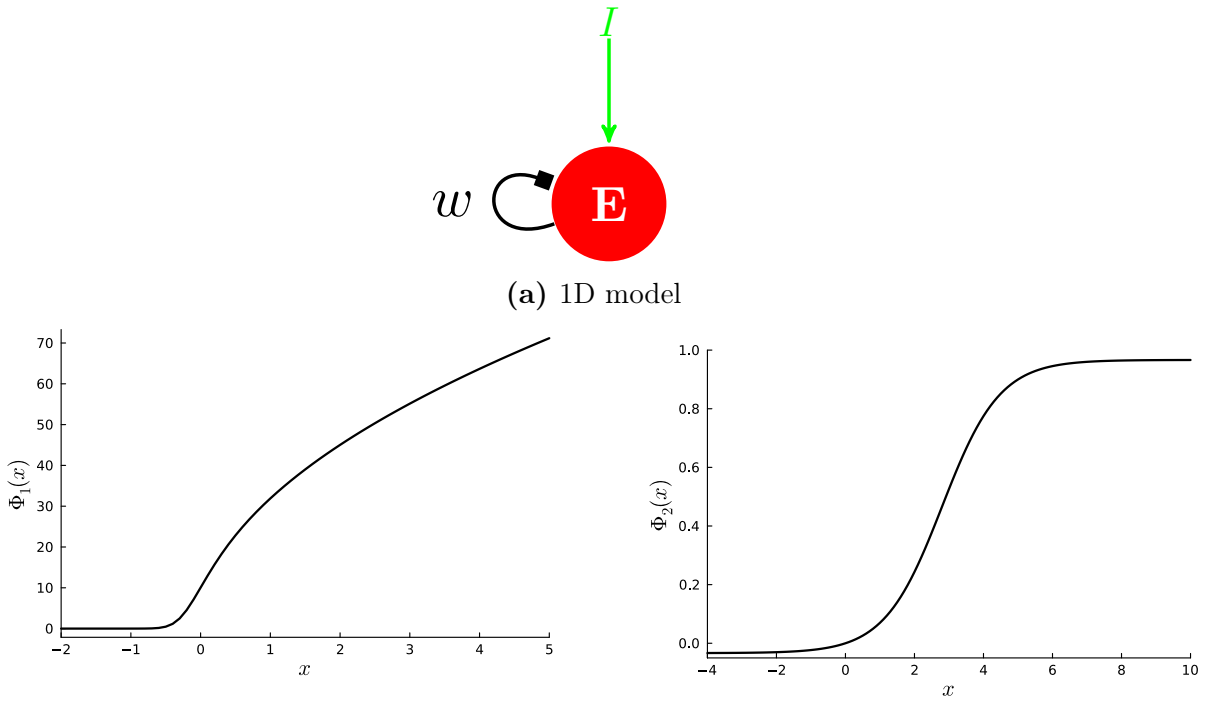
To assess model equivalence, two different transfer functions will be considered. The first transfer function will be that of Brunel and Lavigne 2009 ($\Phi_1(x)$; Figure 6.2b) while the second transfer function will be a more standard sigmoidal function ($\Phi_2(x)$; Figure 6.2c). The main difference between these two functions is the presence or absence of an upper saturation, and it will be assessed whether this property qualitatively has an impact over the behavior of the 1D model. The absence of upper saturation is not a problem as long as the range of input is restricted to a range giving a physiologically plausible firing rate output (L. F. Abbott 1991). Having an upper saturation in the transfer function is thus somehow a mark of safety, ensuring that the model cannot give unbounded values.

Another difference is the range of outputs: $\Phi_1(x)$ gives values in $[0;\ +\infty[$ while $\Phi_2(x)$ gives values in $\left[-\frac{1}{1+\exp(\alpha\theta)};\ 1-\frac{1}{1+\exp(\alpha\theta)}\right]$. However, a scaling factor can be applied to $\Phi_2(x)$ in order to get the desired range of firing rates.

The default parameter values used with $\Phi_2(x)$ are the same as in Gjorgjieva et al. 2021a and Gjorgjieva et al. 2021b, and can be found in Table 6.2.

Symbol	Description	Gjorgjieva et al. 2021a value
τ	Time constant of rate dynamics	1 [msec]
α	Gain/Slope of $f-I$ curve	1.2
θ	Threshold of $f-I$ curve	2.8
w	Recurrent connection strength	9
R	Spontaneous activity	0
I_{bias}	Bias current	0
I_{app}	External applied current	0

Table 6.2: Default parameters of the 1D alternative model from Gjorgjieva et al. 2021a and Gjorgjieva et al. 2021b.



(a) 1D model

(b) $\Phi_1(x) = \frac{1}{\tau_m\sqrt{\pi}}\left[\int_{-\infty}^{+\infty}\exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\,dz\right]^{-1}$

(c) $\Phi_2(x) = \frac{1}{1+\exp(-\alpha(x-\theta))} - \frac{1}{1+\exp(\alpha\theta)}$, $\alpha = 1.2$, $\theta = 2.8$

Figure 6.2: Setup for the model sensitivity assessment. **(a)** One-dimensional model. A single population of excitatory neurons coding for a single (target) item. The population makes a recurrent connection with itself with weight w (black arrow) that can either be excitatory ($w > 0$) or inhibitory ($w < 0$). The population also receives external synaptic input I (green arrow). **(b)** Transfer function from Brunel and Lavigne 2009 (see Appendix C.1 for an analysis of its behavior). **(c)** More standard sigmoidal transfer function from Gjorgjieva et al. 2021a and Gjorgjieva et al. 2021b. Parameters α and θ allow the modeler to tune the gain/slope and the midpoint of $\Phi_2(x)$ as desired (see Appendix C.2 for an analysis of its behavior).

6.4 Spontaneous activity

6.4.1 Method 1

In their paper, Brunel and Lavigne use the non-selective external currents I_i^{ext}, that actually act as bias currents, to get a spontaneous activity of $r_{spont} = 5$ [Hz] for each population[8] in the absence of any other external input. However, they do not specify these values. Using the 1D model (6.2), conditions on parameters w and I are investigated in order to reproduce this background activity.

In order to get a spontaneous activity of r_{spont} [Hz], the model must satisfy two conditions (as previously discussed in Chapter 2):

1. The spontaneous activity $r_T(t) = r_{spont}$ must be a fixed point (or equilibrium) of the model (6.2). That is

$$\dot{r}_T = 0 \quad \leftrightarrow \quad r_{spont} = \Phi_1(w \cdot r_{spont} + I_{ext} + 0) \tag{6.3}$$

2. This spontaneous activity state must be stable. In other words, the eigenvalue evaluated at $r_T = r_{spont}$ must be negative:

$$\left.\frac{d\dot{r}_T}{dr_T}\right|_{r_T = r_{spont}} < 0 \quad \leftrightarrow \quad \frac{1}{\tau}\left(-1 + w\Phi_1'(w \cdot r_{spont} + I_{ext})\right) < 0 \tag{6.4}$$

with $\Phi_1'(x)$ the first derivative of transfer function $\Phi_1(x)$. The factor $\frac{1}{\tau}$ can be ignored since $\tau > 0$.

Equations (6.3) and (6.4) form together a system of two non-linear equations with two unknowns (w and I_{ext}). It can be solved graphically and/or numerically with a non-linear solver. However, Eq. (6.4) requires first the computation of $\Phi_1'(x)$. One can find

$$\begin{aligned}
\Phi_1'(x) &= \frac{1}{\tau_m\sqrt{\pi}} \cdot (-1) \cdot \left[\int_{-\infty}^{+\infty} \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\, dz\right]^{-2} \cdot \frac{d}{dx}\left[\int_{-\infty}^{+\infty} \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\, dz\right] \\
&= -\frac{\tau_m\sqrt{\pi}}{(\tau_m\sqrt{\pi})^2} \cdot \left[\int_{-\infty}^{+\infty} \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\, dz\right]^{-2} \cdot \left[\int_{-\infty}^{+\infty} \frac{d}{dx}\left[\exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\right]\, dz\right] \\
&= -\tau_m\sqrt{\pi} \cdot (\Phi_1(x))^2 \cdot \left[\int_{-\infty}^{+\infty} -z^2 \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\, dz\right]
\end{aligned}$$

Finally, one has

$$\Phi_1'(x) = \tau_m\sqrt{\pi} \cdot (\Phi_1(x))^2 \cdot \left[\int_{-\infty}^{+\infty} z^2 \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\, dz\right] \tag{6.5}$$

[8] This background activity is very common in the brain since any cortical area has a resting state greater than 0 [Hz].

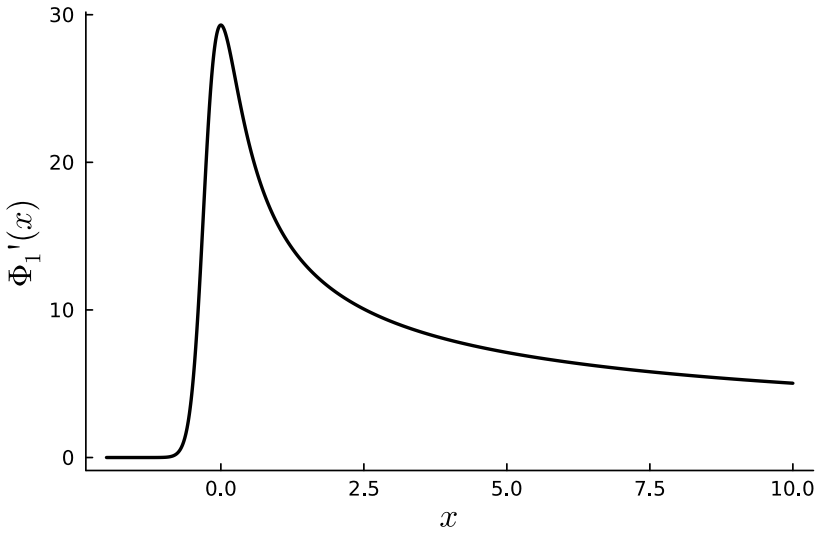


Figure 6.3: First derivative of the transfer function from Brunel and Lavigne 2009. Its expression is given by $\Phi_1'(x) = \tau_m\sqrt{\pi} \cdot (\Phi_1(x))^2 \cdot \left[\int_{-\infty}^{+\infty} z^2 \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\, dz\right]$

Its behavior is shown in Figure 6.3.

Considering again Eqs. (6.3) and (6.4), one can find the conditions on w and I:

$$\begin{cases} \Phi_1^{-1}(r_{spont}) = w \cdot r_{spont} + I_{ext} \\ w < \frac{1}{\Phi_1'(w \cdot r_{spont} + I_{ext})} \end{cases}$$

such that

$$I_{ext} = \Phi_1^{-1}(r_{spont}) - w \cdot r_{spont} \ (= I_{ext,w=0}) \tag{6.6a}$$

$$w < \frac{1}{\Phi_1'(\Phi_1^{-1}(r_{spont}))} \tag{6.6b}$$

with $\Phi_1^{-1}(x)$ the inverse transfer function of $\Phi_1(x)$. The inverse function is guaranteed to exist because $\Phi_1(x)$ is a monotonically increasing and continuous function. Thus, Eq. (6.6a) tells the bias current to apply in order to make the state $r_T(t) = r_{spont}$ a fixed point (whatever it is stable or not). Eq. (6.6b), in turn, makes sure that the fixed point $r_T(t) = r_{spont}$ is stable.

For $r_{spont} = 5$ [Hz] (default value from B&L), the numerical solution gives $w \lessapprox 0.04$ (see Appendix C.3.1 for a less accurate graphical approach). Thus, for default values of parameters J_1 and J_I, *i.e.* $w = J_1 - J_I = 3.65$, the spontaneous activity $r_{spont} = 5$ [Hz] would be unstable (*i.e.* $\frac{d\dot{r}_T}{dr_T} \approx 8.95 > 0$)! One needs to change the default value of w

(thus J_1 and J_I) to get back this property. For example, using $w = 0.02$, one would get a stable spontaneous activity with $\frac{d\dot{r}_T}{dr_T} \approx -0.05 < 0$ (see Appendix C.3.2 for graphical examples). Therefore, by varying p, other parameters such as w need to change as well to get back desired properties.

It is true that Eqs. (6.6a) and (6.6b) give the conditions for an *arbitrary* value of r_{spont} [Hz]. Moreover, these conditions simply indicate that the spontaneous activity is a stable equilibrium but they do not tell anything else with respect to *other stable* equilibria that could be associated to the pair (w, I). In other words, the stable spontaneous activity ensured by Eqs. (6.6a) and (6.6b) could potentially be part of a bistable system, that is, for the same w, there could exist another (stable) equilibrium r_T^* at $I = I_{ext}$. One could therefore wonder what are the conditions to have a single fixed point at r_{spont} [Hz] or even further, what are the conditions to have a single fixed point for *any* r_{spont} value.

The first requirement amounts to constrain I_{ext} from Eq. (6.6a) to be *outside* of the bistable region bounded by the saddle-nodes $[I_{SN,2};\ I_{SN,1}]$ associated with the chosen w. Thus, the conditions are

$$w < \frac{1}{\Phi_1'(\Phi_1^{-1}(r_{spont}))} \tag{6.7a}$$

$$I_{ext} = \Phi_1^{-1}(r_{spont}) - w \cdot r_{spont} < I_{SN,2}(w) \quad \underline{\text{or}} \quad I_{ext} > I_{SN,1}(w) \tag{6.7b}$$

with $I_{SN,1}$ and $I_{SN,2}$ obtained by solving (numerically) the conditions for a saddle-node point (see subsection 2.4.1):

$$\begin{cases} r_{SN} &= \Phi(w \cdot r_{SN} + I_{SN}) \\ 1 &= w\Phi'(w \cdot r_{SN} + I_{SN}) \end{cases}$$

for a fixed w and trying different initial conditions.

The second requirement amounts to determine the conditions to have a *monostable* system. A sufficient condition is that w should be less than the maximum value of $\Phi_1'(x)$, that is

$$w < \frac{1}{\Phi_{1,max}'} \tag{6.8}$$

For Method 1, the recurrent connection weight should satisfy $w \lessapprox 0.034$ to have a monostable system.

6.4.2 Method 2

Similarly to Method 1, the same methodology can be applied to Method 2 with $\Phi_2(x)$ in order to have a spontaneous activity R. Conditions (6.6a) and (6.6b) are still valid; one just needs to replace r_{spont} by R, $\Phi_1(x)$ by $\Phi_2(x)$, and $\Phi_1'(x)$ by $\Phi_2'(x)$ whose behavior is given in Figure 6.4 while its expression is given by

$$\Phi_2'(x) = \frac{\alpha \cdot \exp\left(-\alpha \cdot (x-\theta)\right)}{\left[1+\exp\left(-\alpha \cdot (x-\theta)\right)\right]^2} \tag{6.9}$$

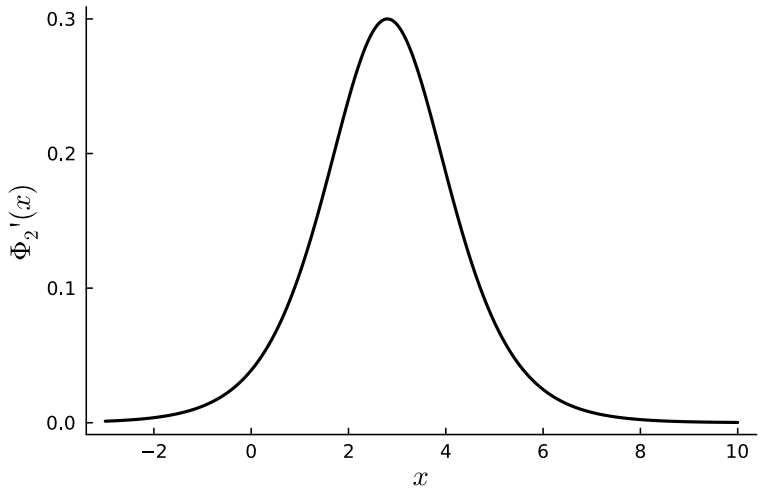


Figure 6.4: First derivative of the transfer function from Gjorgjieva et al. 2021a. Its expression is given by $\Phi_2'(x) = \frac{\alpha \cdot \exp(-\alpha \cdot (x-\theta))}{[1+\exp(-\alpha \cdot (x-\theta))]^2}$. Illustrated with default parameter values: $\alpha = 1.2$, $\theta = 2.8$.

Again, $\Phi_2^{-1}(x)$ is the inverse transfer function of $\Phi_2(x)$ and it exists since $\Phi_2(x)$ is a monotonically increasing and continuous function.

Equations (6.7a), (6.7b) and (6.8) can be easily transposed to Method 2 by changing all Φ_1 by Φ_2. For example, to have a stable resting state activity $R = 0$, w should satisfy $w \lessapprox 25.69$. Any R, in turn, will be stable if $w < \frac{1}{0.3} \approx 3.33$ (monostable system).

6.4.3 Comparison between both methods

The same methodology can be applied to both methods suggesting that the transfer function does not play a crucial role as long as it can be differentiable and possesses an inverse function. Also, a single constraint on r_{spont} or R is that the value should be in the range of outputs of Φ. In other words, r_{spont} or R should satisfy

$$\Phi(-\infty) \leq r_{spont} \text{ or } R \leq \Phi(+\infty)$$

All in all, it is possible to choose a desired stable spontaneous activity by applying a bias current and by choosing w and I_{bias} appropriately.

6.5 Linear filter dynamics

As a first step, in order to explore the temporal dynamics of Eq. (6.2), the recurrent connectivity w is set to zero. In other words, the target population only receives an external input current I, and the temporal behavior $r_T(t)$ in response to that current is investigated. If $w = 0$, then the model reduces to a *linear low-pass filter*, meaning that the target population simply integrates its input current linearly. Thus, the model becomes

$$\tau \frac{dr_T}{dt} = -r_T + \Phi(I) \tag{6.10}$$

Since this equation is linear, it can be solved analytically. The solution is therefore given by

$$r_T(t) = r_T(0) + (\Phi(I) - r_T(0)) \cdot \left[1 - \exp\left(-\frac{t}{\tau}\right)\right] \tag{6.11}$$

with $r_T(0)$ an initial condition (chosen by the user) for $r_T(t)$.

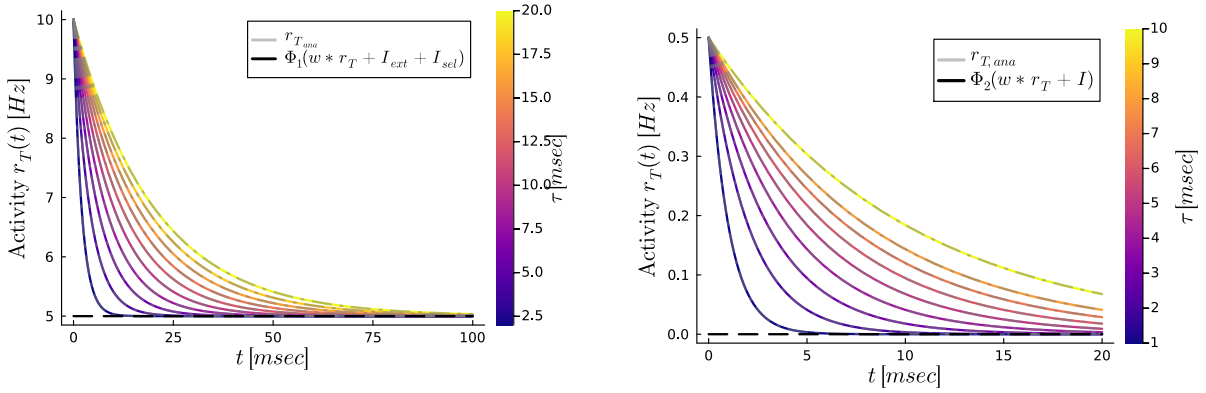
Thus the individual effects of rate dynamics τ and external input current I can be investigated numerically, and the numerical solution can then be compared with the analytical one.

6.5.1 Effect of the time constant τ of rate dynamics

It can be seen in Figure 6.5 that the effect of τ is similar for both methods. First, it can be observed that varying τ does not affect the final attractor (black dashed line), given by $\Phi(I)$, to which the firing rate converges. On the other hand, the *speed of convergence* is affected by changes in τ. Indeed, as τ increases, $r_T(t)$ needs more and more time to converge to the steady state. It can be explained by τ being at the denominator in the exponential function (Eq. (6.11)): if τ increases (decreases), then the argument of the exponential becomes slightly (largely) negative, suggesting therefore a slow (fast) decay towards $\Phi(I)$. Also, it can be seen that the analytical solution (gray dash-dotted curve) matches with the numerical solution.

6.5.2 Effect of external input current I

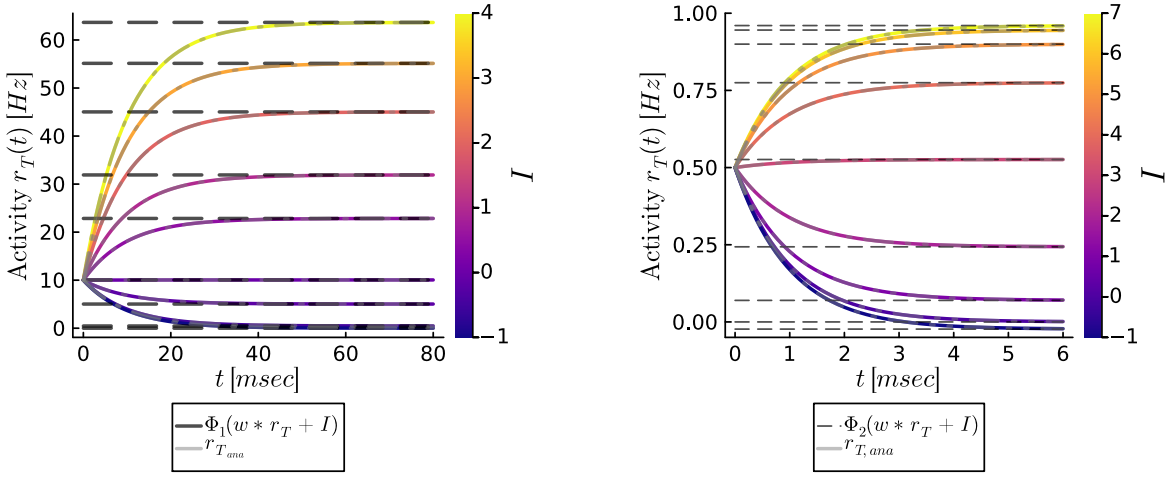
Similarly to the effect of τ, the effect of external input current I in a linear regime (*i.e.* $w = 0$) can be investigated (Figure 6.6). Parameter I tunes the final attractor to which the firing rate converges (black dashed lines), as anticipated with previous section. As I



(a) Method 1; I_{ext} set for $r_{spont} = 5$ [Hz]; $I_{sel} = 0$

(b) Method 2; $I = 0$

Figure 6.5: Effect of the time constant parameter τ on the temporal dynamics of the 1D model (6.2) when no recurrent connectivity exists ($w = 0$). The model reduces to a linear low-pass filter. Parameter τ has an impact on the speed of convergence but not on the final steady state (black dashed line), given by $\Phi(I)$, that is reached. Gray dash-dotted curves represent the analytical solution given by Eq. (6.11).



(a) Method 1

(b) Method 2

Figure 6.6: Effect of external input current parameter I on the temporal dynamics of the 1D model (6.2) in the linear regime ($w = 0$). Parameter I has an impact on the final steady state (black dashed lines), given by $\Phi(I)$, that is reached. If $\Phi(x)$ is bounded from above (below), then $\Phi(I)$ becomes independent of I (*i.e.* saturates) as I becomes largely positive (negative). Gray dash-dotted curves represent the analytical solutions given by Eq. (6.11).

increases, the steady state firing rate $\Phi(I)$ increases as well since $\Phi(x)$ is a monotonically increasing function in both methods. However, a difference between Method 1 and Method 2 can be observed. In Method 1 (Figure 6.6a), as I increases, $\Phi_1(I)$ increases and it can potentially increase indefinitely. On the other hand, in Method 2 (Figure 6.6b), as I increases, $\Phi_2(I)$ increases until it becomes independent of I, that is, $\Phi_2(I)$ always outputs the same value if I is large; it *saturates*. This observation is due to the presence (Method 2) or absence (Method 1) of and upper saturation in the transfer function. Since $\Phi_1(x)$ has no upper saturation, $\Phi_1(I)$ can potentially increase to infinity whereas $\Phi_2(x)$ will be bounded from above due to the presence of an upper saturation. In addition, both transfer functions are bounded from below, hence $\Phi(I)$ saturates as I decreases and becomes largely negative. Again, it can be observed that the analytical solution (gray dash-dotted curve) matches the numerical solution.

6.6 Phase portrait & Bifurcation analyses

Unfortunately, the simple linear behavior applies only with a strongly restricted set of values (a single one actually) for the recurrent connectivity w. However, the general case makes no assumption on the value of w. As a consequence, one should study the dynamics of the *non-linear* model (6.2) as a function of I (the total external input current) and w (the recurrent connection strength). The non-linear feature of

$$\tau \frac{dr_T}{dt} = -r_T + \Phi(w \cdot r_T + I) \tag{6.12}$$

does not make it possible to solve the system analytically. A dynamical analysis, such as phase portraits and bifurcations, is thus considered in order to have an idea of the qualitative behavior of the solution as explained in Chapter 2.

Geometric approach

The cleaner approach to explore the dynamics of (6.12) is to adopt a geometric point of view, that is, one should plot the curves $y = r_T$ and $y = \Phi(w \cdot r_T + I)$ in the same graph and study the effect of parameters w and I. The intersection(s) of these two curves correspond(s) then to the fixed point(s) (or *attractor(s)*) of the model. When the line $y = r_T$ is above (below) the curve $y = \Phi(w \cdot r_T + I)$, it implies that $\dot{r}_T$ is negative (positive) and the firing rate $r_T(t)$ will thus decrease (increase).

6.6.1 Geometric approach: Effect of external input current I

The effect of external input current $I = I_{bias} + I_{app}$ in the non-linear model is first investigated. As seen in section 1.3, this effect corresponds to a "time" shifting of the curve

$y = \Phi(w \cdot r_T + I)$ by I units. This shift will be to the right (left) if I is negative (positive). This effect is observed for both methods and for different fixed values of w (Figures 6.7 and 6.8).

Method 1

As expected, when the external input current I is positive (yellow shades), the curve $y = \Phi_1(w \cdot r_T + I)$ is shifted to the left with respect to the curve $y_0 = \Phi_1(w \cdot r_T + 0)$ (yellow-orange shade). On the contrary, the curve $y = \Phi_1(w \cdot r_T + I)$ is shifted to the right with respect to y_0 when I is negative. Moreover, one can observe different behaviors depending on the fixed value for w. For some values of w (*i.e.* $w < \frac{1}{\Phi'_{1,max}}$; Figure 6.7 top panels), the two curves $y = \Phi_1(w \cdot r_T + I)$ and $y = r_T$ (black line) intersect at exactly one point, suggesting therefore the existence of a single attractor in the model; the model is thus monostable. For the other values of w (Figure 6.7 bottom panels), the variation in the external input current I makes the curves intersect at either one point (at least yellow and dark blue shades) or at three points (curves in the gray shaded area bounded by the dash-dotted curves). Thus, for these values of w, there exist two particular values of I at which two equilibria are created/destroyed and where the two curves are *tangent* (*i.e.* their slopes are equal). These specific values of I correspond therefore to *saddle-nodes* (green dash-dotted curve $y = \Phi_1(w \cdot r_T + I_{SN,1})$ and light blue dash-dotted curve $y = \Phi_1(w \cdot r_T + I_{SN,2})$). In addition, the gray shaded area, indicating where the two curves intersect three times, becomes larger as w increases because the large positive feedback given by $w \cdot r_T$ is compensated by the strongly negative current I.

To determine graphically the stability of the attractor(s), one should look at the vector field given by the sign of $\dot{r}_T$. When $\Phi_1(w \cdot r_T + I) > r_T$ (*i.e.* the curve is above the line for a fixed r_T), the difference $\Phi_1(w \cdot r_T + I) - r_T > 0$ is positive giving $\dot{r}_T > 0$ positive as well. Since the vector field is positive in this region, the solution $r_T(t)$ will grow if an initial condition is put into that region. In other words, if the solution $r_T(t)$ starts initially below an equilibrium and in a region where $\dot{r}_T > 0$, then $r_T(t)$ will approach (and converge to) that equilibrium from below. Similarly, when $\Phi_1(w \cdot r_T + I) < r_T$ (*i.e.* the curve is below the line for a fixed r_T), the vector field is negative because the difference $\Phi_1(w \cdot r_T + I) - r_T < 0$ is negative giving $\dot{r}_T < 0$ negative as well. As a consequence, if the solution $r_T(t)$ starts initially above an equilibrium and in a region where $\dot{r}_T < 0$, then $r_T(t)$ will decay (and converge to) that equilibrium from above. Thus, when the system is monostable (top panels), the single equilibrium is a stable attractor. When the system admits three fixed points (bottom panels), the leftmost (or low) and rightmost (or high) intersections (*i.e.* attractors) are stable whereas the middle intersection is unstable. The model is thus *bistable*. The unstable equilibrium plays therefore the role of separator or *threshold* between the two stable attractors. Since $\Phi_1(w \cdot r_T + I)$ is shifted to the left as

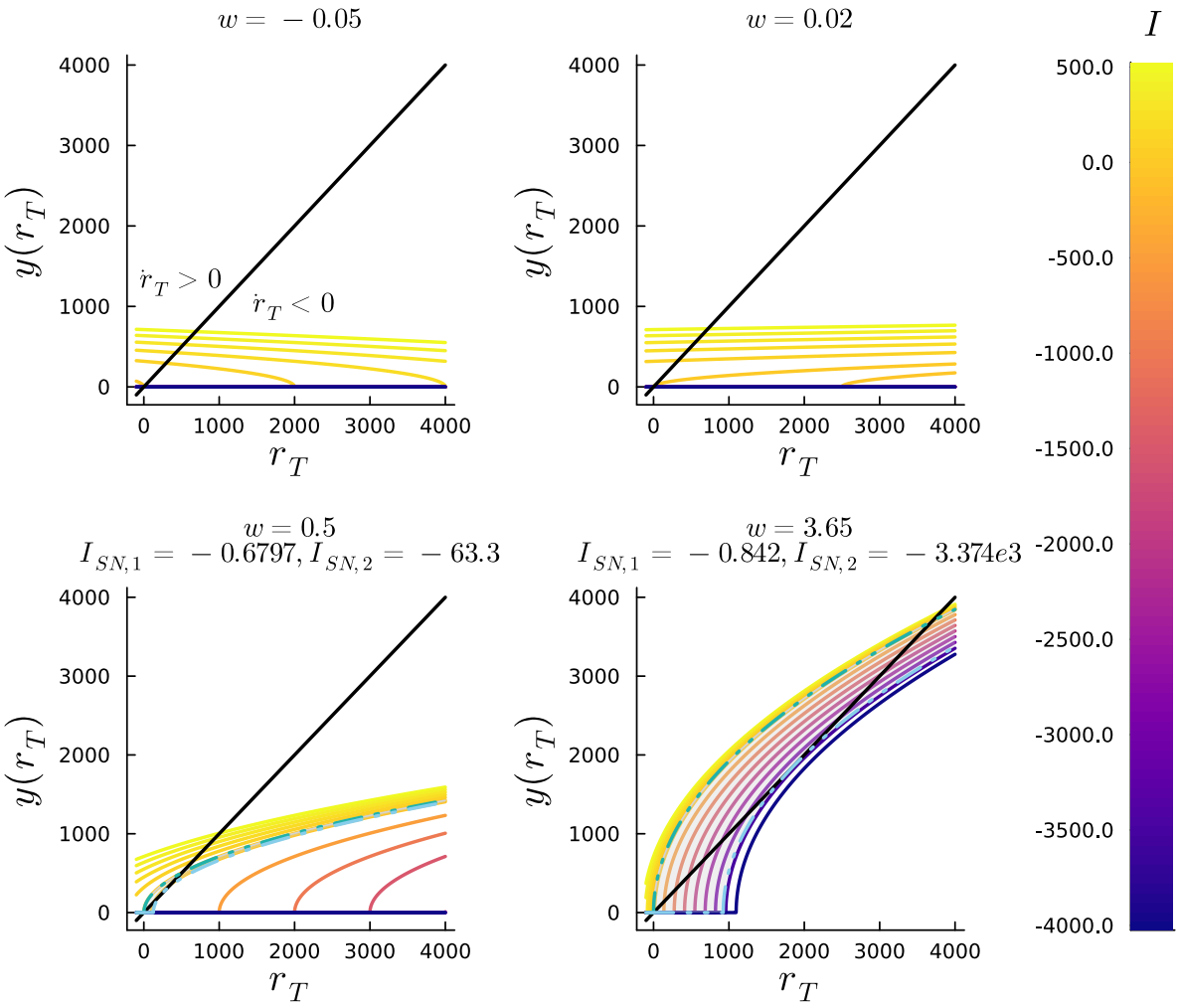


Figure 6.7: Effect of external input current I on the temporal dynamics of the model (6.12) for Method 1 with different fixed values of the recurrent connectivity w. Parameter I shifts the curve $y = \Phi_1(w \cdot r_T + I)$ to the left as I increases. For a fixed value of w, varying I makes up to three intersections with the line $y = r_T$ (black line). Green ($I_{SN,1}$) and light blue ($I_{SN,2}$) dash-dotted curves are tangent to $y = r_T$ therefore corresponding to saddle-node equilibria. Gray shaded area, bounded by saddle-node curves, corresponds to the bistable region where three fixed points coexist.

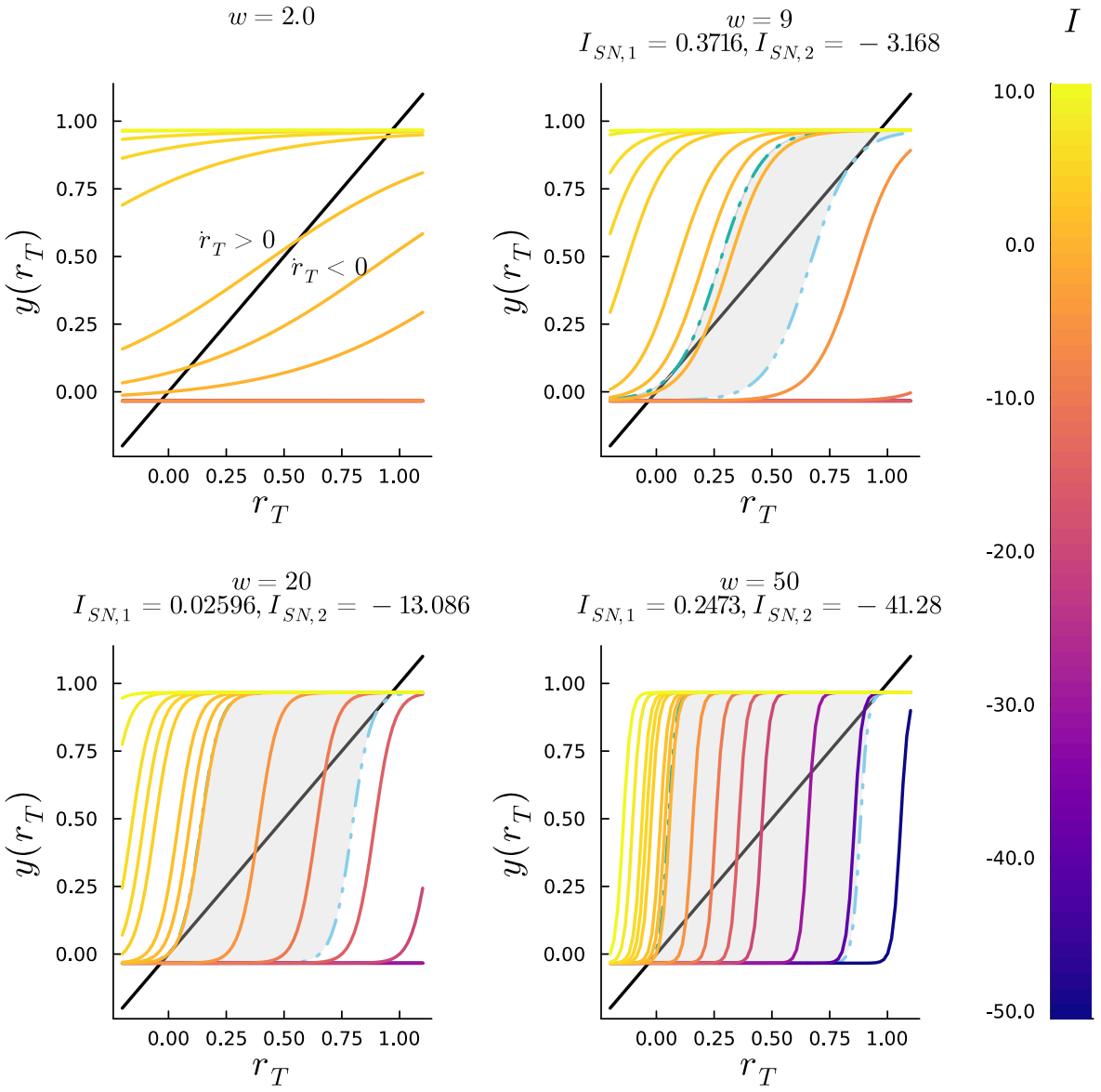


Figure 6.8: Effect of external input current I on the temporal dynamics of the model (6.12) for Method 2 with different fixed values of the recurrent connectivity w. Parameter I shifts the curve $y = \Phi_2(w \cdot r_T + I)$ to the left as I increases. For a fixed value of w, varying I makes up to three intersections with the line $y = r_T$ (black line). Green ($I_{SN,1}$) and light blue ($I_{SN,2}$) dash-dotted curves are tangent to $y = r_T$ therefore corresponding to saddle-node equilibria. Gray shaded area, bounded by saddle-node curves, corresponds to the bistable region where three fixed points coexist.

I increases, the middle intersection is getting closer and closer to the low stable attractor (*i.e.* leftmost intersection). As I increases, the threshold is thus lowered at the same time (see also Chapter 7).

Method 2

The exact same results as with Method 1 can be observed:

- For values of w such that $w < \frac{1}{\Phi'_{2,max}}$ (top-left panel), only one intersection between $y = \Phi_2(w \cdot r_T + I)$ and $y = r_T$. This equilibrium is a stable attractor based on the vector field.
- For other values of w, either one or three (in gray shaded area) intersections. There must therefore exist two values of I making the curves $y = \Phi_2(w \cdot r_T + I)$ and $y = r_T$ tangent to each other. These values correspond to saddle-nodes ($I_{SN,1} \rightarrow$ green curve, $I_{SN,2} \rightarrow$ light blue curve). When three intersections occur (bistable region), the leftmost and the rightmost intersections are stable attractors whereas the middle intersection is an unstable fixed point (similar to a threshold between both attractors).
- If the model is in a bistable regime, the threshold between both stable attractors is lowered as I increases.
- As w increases, the bistable region gets wider.

6.6.2 Geometric approach: Effect of recurrent synaptic weight w

Similarly to the effect of I, the effect of recurrent connectivity strength w can be investigated. This effect corresponds to a "time" scaling of the curve $y = \Phi(w \cdot r_T + I)$ (as seen in section 1.3). The curve will be contracted if $|w| > 1$ and expanded if $0 < |w| < 1$. Moreover, if $w < 0$, then a "time" folding occurs in addition to the "time" scaling, that is, the curve $y = \Phi(w \cdot r_T + I)$ undergoes an orthogonal symmetry (*i.e.* mirror effect) with respect to y-axis. The effect of w is observed for both methods and for different fixed values of I (Figures 6.9 and 6.10).

Method 1

As expected, the curve $\Phi_1(w \cdot r_T + I)$ is horizontally contracted ("accelerated" over the range of r_T) when $|w| > 1$ (light purple to yellow shades) and expanded (or dilated over the range of r_T) when $0 < |w| < 1$ (dark purples and blues). Moreover, when $w < 0$, the y-axis reflects the curves; they are folded. For a fixed value of I, the behavior of the model again depends on the value of the other parameter. For some values of I (Figure 6.9 top panels), the curve $y = \Phi_1(w \cdot r_T + I)$ and the line $y = r_T$ (black line) either intersect at

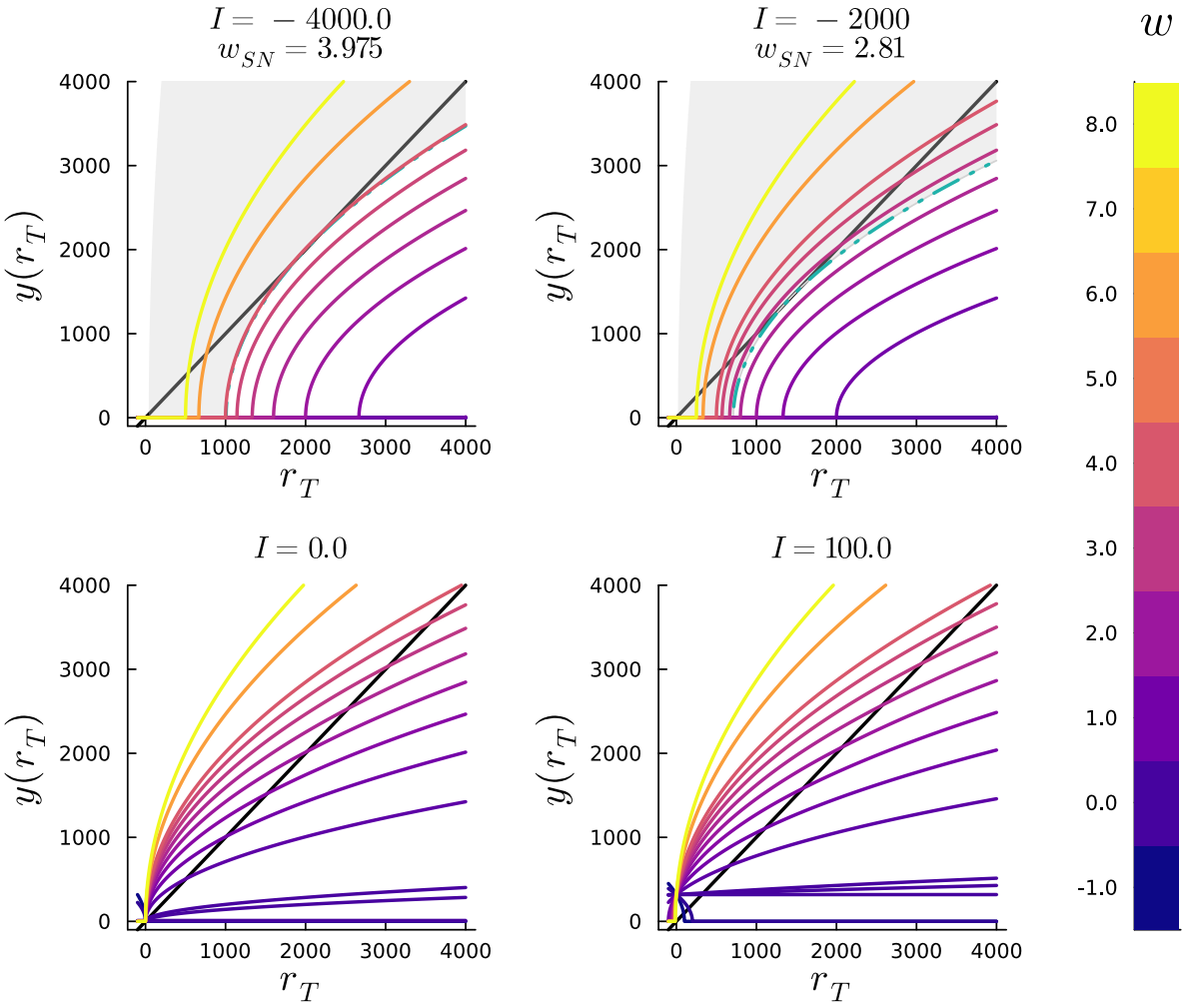


Figure 6.9: Effect of recurrent connection strength w on the temporal dynamics of the model (6.12) for Method 1 with different fixed values of the external current I. Parameter w contracts (expands) the curve $y = \Phi_1(w \cdot r_T + I)$ if $|w| > 1$ ($0 < |w| < 1$). If $w < 0$, then the curve is also folded. For a fixed value of I, varying w makes up to three intersections with the line $y = r_T$ (black line). Green (w_{SN}) dash-dotted curve is tangent to $y = r_T$ therefore corresponding to a saddle-node equilibrium. Gray shaded area corresponds to the bistable region where three fixed points coexist.

one point (dark blue and dark purple shades) or at three points (brighter shades). Again, it implies that a particular value of w makes the curve $y = \Phi_1(w_{SN} \cdot r_T + I)$ (green dash-dotted curve) and the line $y = r_T$ tangent to each other. The tangent point $(w_{SN}, r_{T,SN})$ is then a saddle-node. Unlike what has been observed with the effect of I, the three fixed point region (gray shaded area) is unbounded in the sense that as w tends towards infinity ($w \to +\infty$), the model still displays three fixed points. For other values of I (Figure 6.9 bottom panels), the model possesses a single equilibrium (*i.e.* single intersection between $y = \Phi_1(w \cdot r_T + I)$ and $y = r_T$).

Regarding the stability of the fixed points, using the vector field approach once again (*i.e.* $y = \Phi_1(w \cdot r_T + I)$ above (below) $y = r_T \Rightarrow \dot{r}_T > 0$ ($\dot{r}_T < 0$) $\Rightarrow r_T(t)$ grows (decays)), the single fixed point of the model in a monostable regime is always a stable attractor. When the model admits three fixed points, then the leftmost and the rightmost intersections are stable attractors whereas the middle intersection is an unstable fixed point, exactly as seen with the effect of I. This unstable fixed point plays once again the role of threshold between both stable attractors.

Method 2

The exact same results as seen in Method 1 can be found in Method 2 as well (Figure 6.10). However, two additional observations, that cannot be found in Method 1, should be made: first, in Method 2, a saddle-node strength w_{SN} *always* exists, that is, the model always possesses a bistable region, starting from $y = \Phi_2(w_{SN} \cdot r_T + I)$ (green dash-dotted curve). It is true though that this bistable region can be reduced to a single curve (*i.e.* binary switch or step function) as the considered value of I increases. Thus, this difference is only minor with respect to the model in Method 1. Second, for some values of I, the model admits *two* bistable regions (bottom-left panel, gray shaded areas bounded by green, light blue and light green dash-dotted curves). This observation can be explained by the choice of parameter θ in $\Phi_2(x)$. This parameter breaks the symmetry of the sigmoid with respect to the origin when $\theta > 0$. The appearance of another bistable region is the consequence of this asymmetry[9]. This second bistable region has not been found in Method 1. However, the shape of $\Phi_1(x)$ does not make it easy to manipulate it numerically (the non-linear solver has difficulties in solving systems involving $\Phi_1(x)$ and is extremely sensitive to initial conditions). It must be admitted that each individual value of current has not been tested, thus the existence of this second bistable region cannot be completely refuted.

[9] If one uses $\theta = 0$, then the sigmoid is symmetric again and graphs show that this second bistable region disappears. However, using $\theta = 0$ implies that the output range of values of $\Phi_2(x)$ is $[-0.5, 0.5]$. The user/modeler should therefore choose his/her parameter values carefully and according to his/her goals.

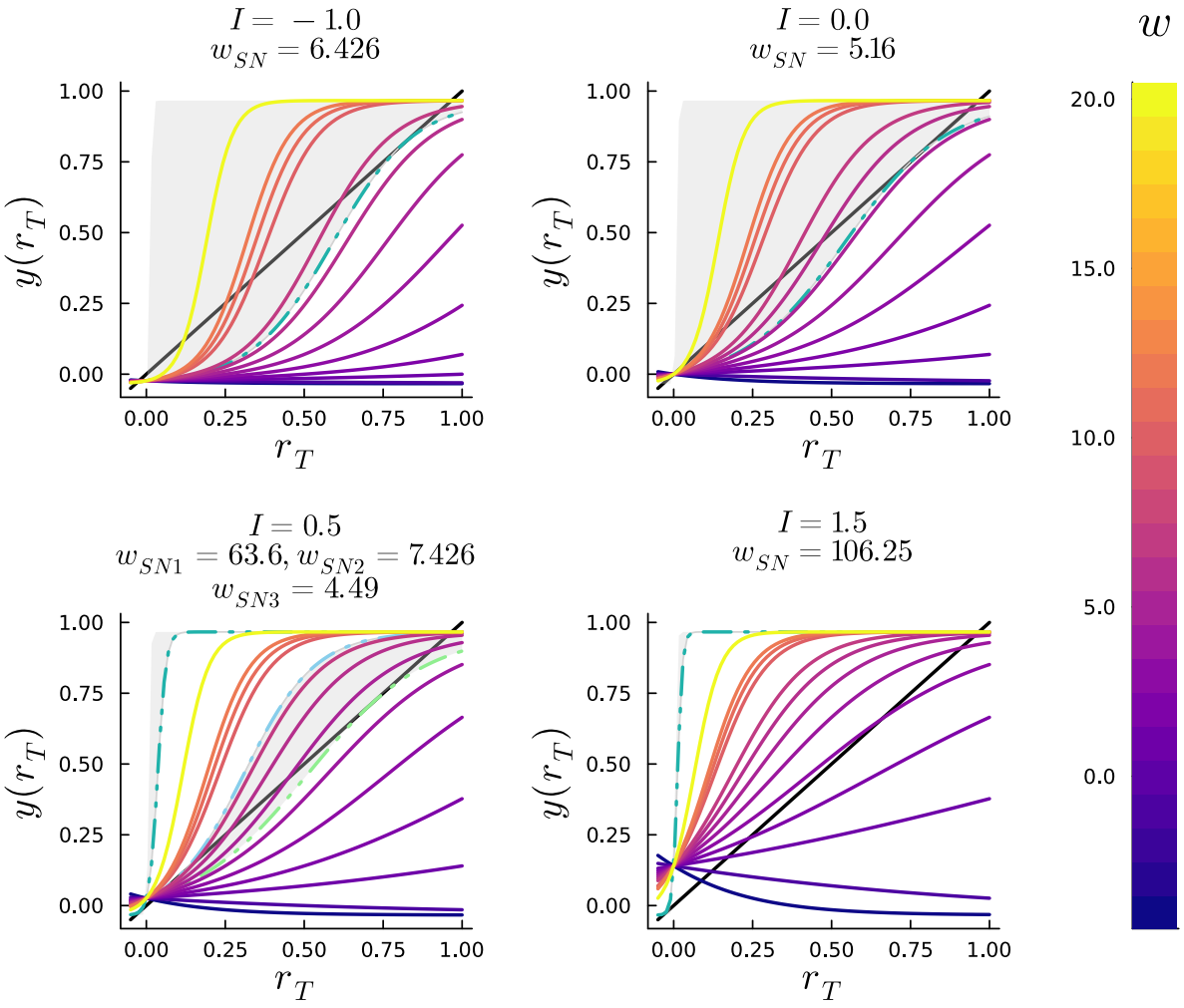


Figure 6.10: Effect of recurrent connection strength w on the temporal dynamics of the model (6.12) for Method 2 with different fixed values of the external current I. Parameter w contracts (expands) the curve $y = \Phi_2(w \cdot r_T + I)$ if $|w| > 1$ ($|w| < 1$). If $w < 0$, then the curve is also folded. For a fixed value of I, varying w makes up to three intersections with the line $y = r_T$ (black line). Green ($w_{SN,1}$), light blue ($w_{SN,2}$) and light green ($w_{SN,3}$) dash-dotted curves are tangent to $y = r_T$ therefore corresponding to saddle-node equilibria. Gray shaded area(s) correspond(s) to the bistable region(s) where three fixed points coexist.

All in all, both models behave globally similarly for the effects of w and I with minor differences between them.

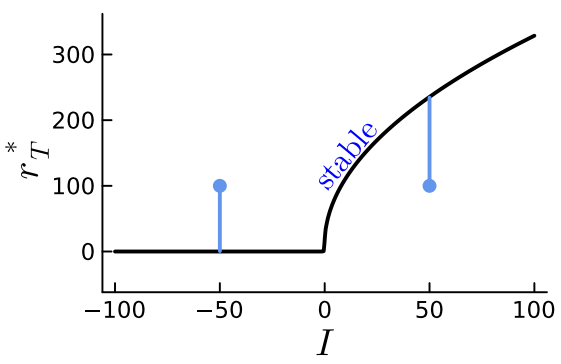
Bifurcation diagrams

The geometric approach is a very useful tool to understand the global behavior of the vector field, and to determine the number of intersections (thus equilibria of the model). This approach is also useful to investigate the individual effects of each parameter considering the others as frozen. Based on this approach, it has been seen that the model, in both methods, changes completely its behavior (monostable $\leftrightarrow$ bistable) for a range of parameter values. A drawback of this geometric point of view though is that it does not give clear information about the exact values of the equilibria; having access to the values at which both the curve and the straight line intersect is rather inaccurate based only on the geometric approach graphs. This the reason why *bifurcation diagrams* are drawn. Bifurcation diagrams show the equilibria values as a function of one parameter, the others being considered as frozen again. In other words, bifurcation diagrams show the curve(s) $\dot{r}_T = 0$ as a function of a parameter. These diagrams give thus *complementary* information to the geometric approach/phase portrait graphs.

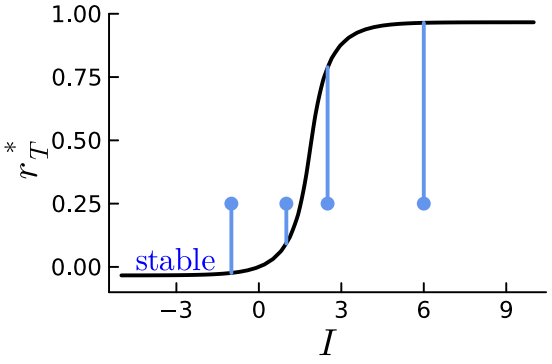
6.6.3 $\{I, r_T^*\}-$Bifurcation diagrams

The bifurcation diagrams are once again drawn for both methods and for different fixed values of w (Figure 6.11). The considered values of w are actually the same as in Figures 6.7 and 6.8.

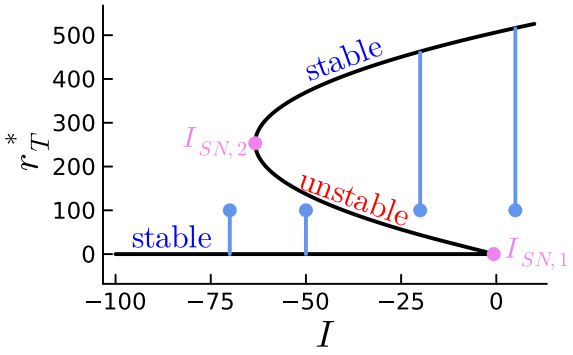
Bifurcation diagrams (black lines) show consistent information with respect to that of the geometric approach. Indeed, when $w < \frac{1}{\Phi'_{max}}$ (Figure 6.11 **(a)** and **(b)**), each single value of current I corresponds to exactly one value of r_T^* thus exactly one equilibrium. Moreover, in these cases, one can see that the value of the equilibrium increases as I increases, and the relationship between r_T^* and I looks like a "time" scaled version of the corresponding transfer functions. For the other values of w (middle and bottom panels), the bifurcation diagrams show a transition from one equilibrium ($I < I_{SN,2}$) to three equilibria ($I_{SN,2} < I < I_{SN,1}$) to one equilibrium again ($I_{SN,1} < I$). This transition is once again coherent with what has been found in the geometric approach (Figures 6.7 and 6.8). In addition, the saddle-node values (violet dots) correspond as well to the values found in the geometric approach, and therefore define the bistable region. This bistable region becomes wider as w increases. The shape of the bifurcation diagrams for values of w allowing a bistable regime, confirms a saddle-node bifurcation (or fold) as seen in subsection 2.4.1. Thus, the behavior of the model is similar with both methods giving



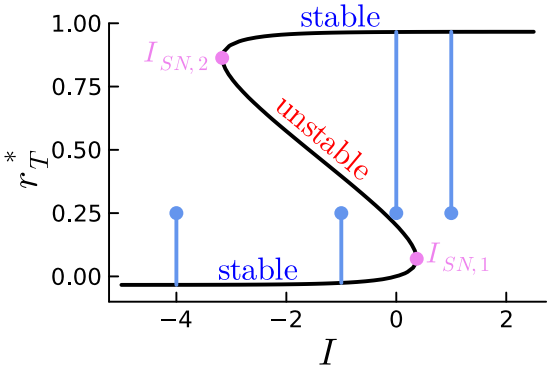
(a) Method 1; $w < \frac{1}{\Phi'_{1,max}}$ (*i.e.* $w = 0.02$)



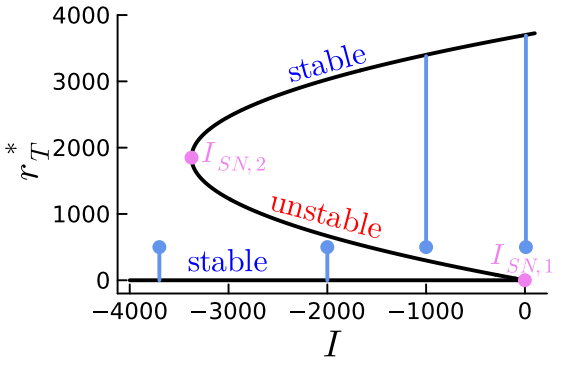
(b) Method 2; $w < \frac{1}{\Phi'_{2,max}}$ (*i.e.* $w = 2$)



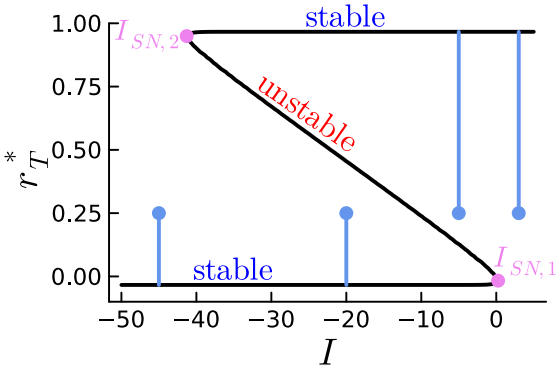
(c) Method 1; $w > \frac{1}{\Phi'_{1,max}}$ (*i.e.* $w = 0.5$)



(d) Method 2; Default value for w (*i.e.* $w = 9$)



(e) Method 1; Default value for w (*i.e.* $w = 3.65$)



(f) Method 2; $w > \frac{1}{\Phi'_{2,max}}$ (*i.e.* $w = 50$)

Figure 6.11: $\{(I, r_T^*)\}$−Bifurcation diagrams (black line) for Method 1 (**(a)**, **(c)** and **(e)**) and Method 2 (**(b)**, **(d)** and **(f)**) with different fixed values of w (increasing from top to bottom panels). Stability of branches, indicated by nearby colored letter strings, is determined by one-shot simulations (light blue lines) starting all at the same initial condition within a graph (light blue dots). Saddle-nodes (violet dots) define the bistable region and the values correspond to those found in Figures 6.7 and 6.8. The information from bifurcation diagrams is thus complementary to that of the geometric approach.

further evidence that the two methods are equivalent.

Link between stability of branches and initial conditions

A drawback of the bifurcation diagrams is that the stability of the different branches is not obvious right away. However, one can easily determine this stability by simulating the model with values of I that lie in different regimes (for convenience, the same initial condition (light blue dot) is used for all simulations). From these simulations, it appears that when the model has only one equilibrium (*i.e.* monostable system or bistable system with $I < I_{SN,2}$ or $I_{SN,1} < I$), this equilibrium is stable. The model is thus in a monostable regime. When I lies in the range giving a bistable regime, then depending on the value of I, the solution converges to the lower branch (e.g. **(e)** $I = -2000$) or the upper branch (e.g. **(e)** $I = -1000$)[10]. The lower and upper branches are thus stable while the middle branch is unstable. These results are in agreement with those of the geometric approach. Thus, the unstable branch of the bistable region acts as a separator or threshold between the two other stable attractors. More than that, it is also a boundary between the *basins of attraction* of stable attractors. In other words, any initial cue lying above (below) the unstable branch will give an output corresponding to the high (low) stable attractor.

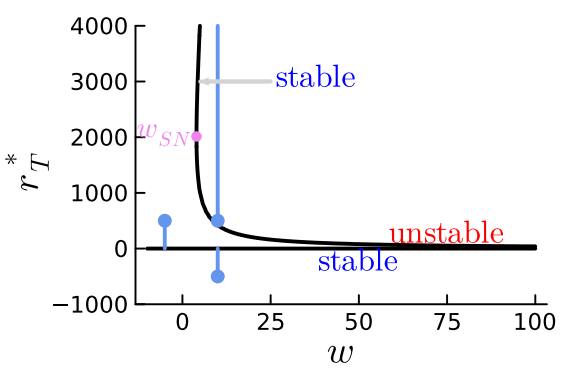
All in all, both methods give similar results, and the bifurcation diagrams give complementary information to that of the geometric approach in the sense that with bifurcation diagrams, one has immediate access to the values of the equilibria and to the behavior of these equilibria as a function of I.

6.6.4 $\{w, r_T^*\}-$Bifurcation diagrams

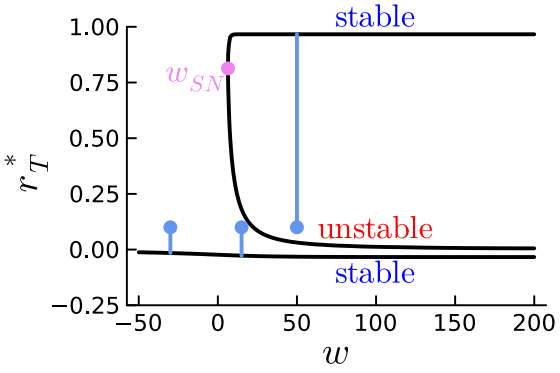
Similarly to the $\{I, r_T^*\}-$bifurcation diagrams, the $\{w, r_T^*\}-$bifurcation diagrams can also be drawn for both methods and for different fixed values of I (Figure 6.12). The considered values of I are actually the same as in Figures 6.9 and 6.10 except for one value (Figure 6.12c).

When $I < 0$ (Figure 6.12 **(a)**, **(b)** and **(c)**), both methods behave similarly, that is, bifurcation diagrams (black curves) show a transition from one equilibrium ($w < w_{SN}$) to three equilibria ($w_{SN} < w$). The region with three equilibria is also unbounded, that is, the model still displays three equilibria as w tends towards infinity. These results are consistent with the results from the geometric approach. Regarding the stability of the different branches, one-shot simulations suggest the same results as seen before: when the model displays a single equilibrium, this equilibrium is stable, and the upper and lower

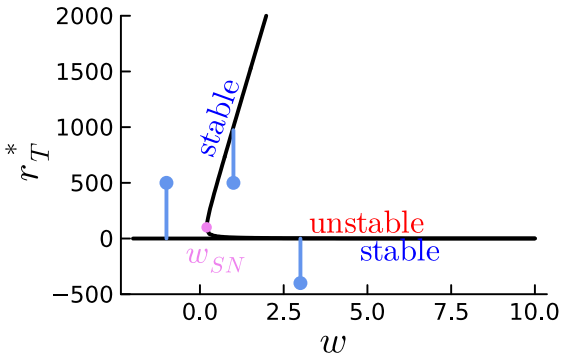
[10]The same result could also have been obtained by considering the same value of I (e.g. $I = -2000$) but with different initial conditions (e.g. $r_T(0) = 500$ and $r_T(0) = 1000$).



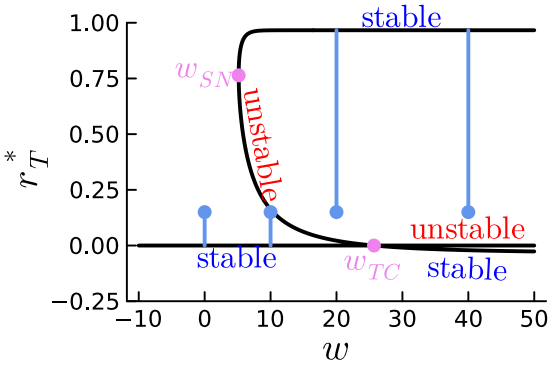
(a) Method 1; $I \ll 0$ (*i.e.* $I = -4000$)



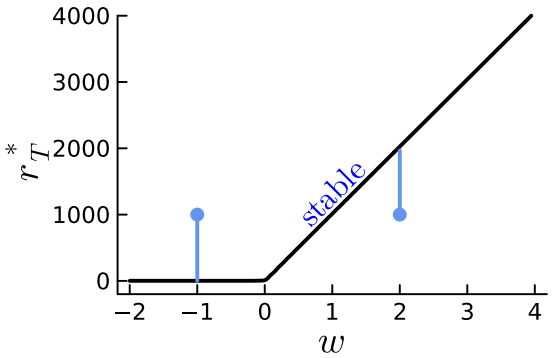
(b) Method 2; $I < 0$ (*i.e.* $I = -1$)



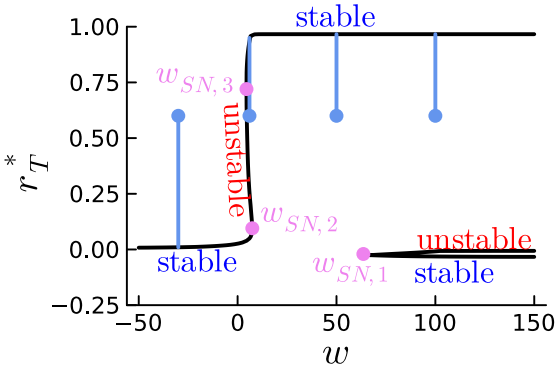
(c) Method 1; $I < 0$ (*i.e.* $I = -10$)



(d) Method 2; $I = 0$



(e) Method 1; $I \geq 0$ (*i.e.* $I = 0$)



(f) Method 2; $I > 0$ (*i.e.* $I = 0.5$)

Figure 6.12: $\{(w, r_T^*)\}$−Bifurcation diagrams (black line) for Method 1 (**(a)**, **(c)** and **(e)**) and Method 2 (**(b)**, **(d)** and **(f)**) with different fixed values for I (increasing from top panels to bottom panels). Stability of branches, denoted by colored letter strings, is determined by one-shot simulations (light blue lines) starting at different initial conditions (light blue dots). Saddle-nodes (violet dots) define the bistable region and the values correspond to those found in Figures 6.9 and 6.10. The information from bifurcation diagrams is thus complementary to that of the geometric approach.

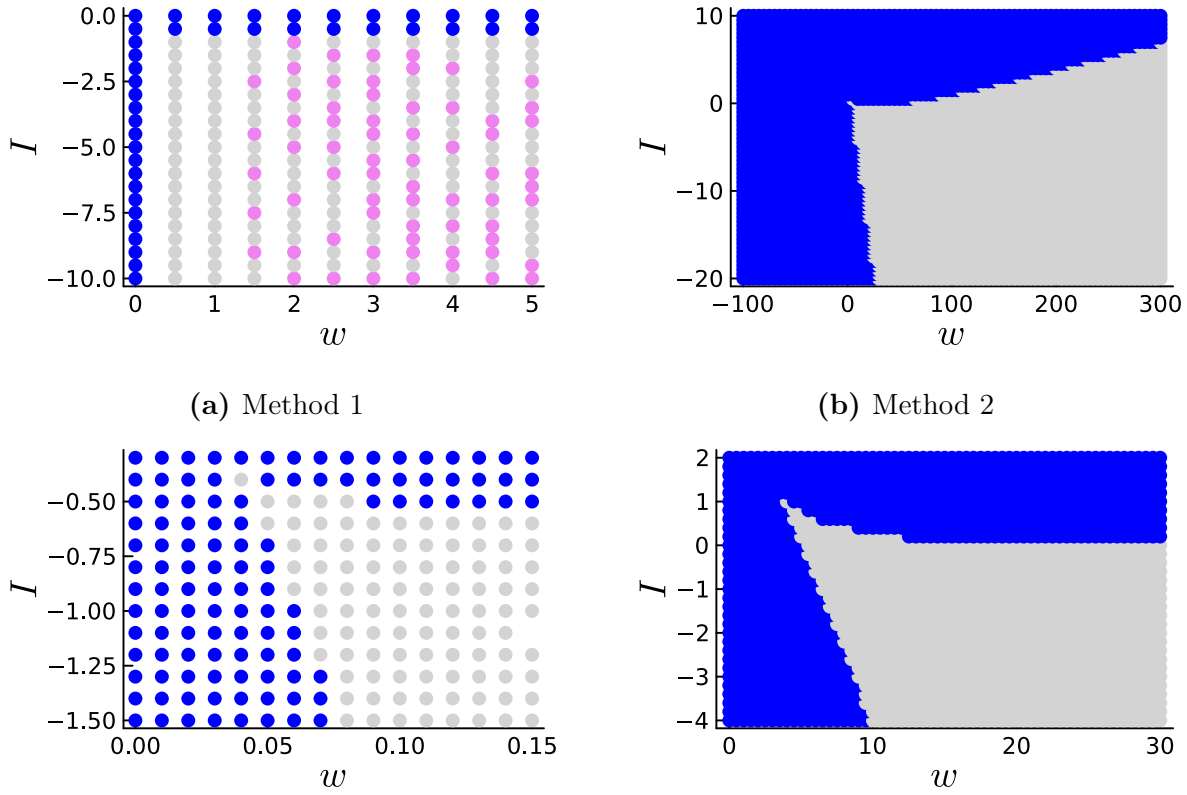
branches of the bistable region are stable as well. The middle branch in the bistable region is therefore unstable and acts at the same time as a threshold between both stable attractors, and as a frontier between basins of attraction of these attractors. Moreover, considering $I < 0$ implies that the model is biased towards the low stable attractor. For example, in Figure 6.12b, if $w = -50$ at first, the corresponding equilibrium is the low stable attractor $r_T(t) = 0$. As w increases, the equilibrium smoothly goes from 0 to -0.034, the low stable attractor of the bistable region. It is true though that this bias is not huge and is somehow counterbalanced by the wide basin of attraction of the high stable attractor in the bistable region.

When $I = 0$, Method 1 and Method 2 seem to display different behaviors. In particular, when $I = 0$, Method 1 (Figure 6.12e) displays a monostable model for any w whereas Method 2 (Figure 6.12d) displays a monostable model for some values of w and a bistable model for other values of w. Method 2 even displays a *transcritical* bifurcation point (w_{TC}). This transcritical bifurcation point is due to the shape of the transfer function $\Phi_2(x)$. For Method 2, $\Phi_2(0) = 0$ thus zero is always an equilibrium when $I = 0$ ($\dot{r}_T = 0 = -0 + \Phi_2(0 + 0)$) but changes stability at $w = \frac{1}{\Phi_2'(\Phi_2^{-1}(0))} \approx 26$. Theoretically, Method 1 should also display this transcritical point with $I \approx -1.5$ rather than $I = 0$ because $\dot{r}_T = 0 = -0 + \Phi_1(0 - 1.5)$. Moreover, this bifurcation should appear at $w \to +\infty$. The difference in behavior regarding the presence/absence of transcritical bifurcation is therefore due to the choice of parameters for $\Phi_2(x)$. Theoretically, there should exist a pair (α, θ) giving the same behavior as $\Phi_1(x)$ regarding the transcritical bifurcation. Both methods can thus display similar behaviors but with different values of I (Method $1 \to I \approx -1.5$; Method $2 \to I = 0$).

When $I > 0$ (or $I > -1.5$ for Method 1 strictly speaking), both methods seem to display again different behaviors (Figures 6.12e and 6.12f). Method 1 shows a monostable system whereas Method 2 shows monostable and bistable systems depending on the value of w. It should be noted that as I increases ($I > 1$), the second bistable region identified earlier in Method 2, defined by $w_{SN,2}$ and $w_{SN,3}$, disappears and only $w_{SN,1}$ remains. As a consequence, both methods show similar behaviors (*i.e.* monostable system) as long as $w < w_{SN,1}$ for such values of I in Method 2. Since the value of $w_{SN,1}$ is rather large and further increases as I increases, both methods are equivalent most of the time.

6.6.5 Stability diagrams

A drawback common to both the geometric approach and the bifurcation analysis is that one of the two parameters must be frozen in order to assess the effects of the other parameter. In addition, it has been seen that both parameters influence the number and



(a) Method 1

(b) Method 2

(c) Method 1; zoom in on bistable region's start

(d) Method 2; zoom in on bistable region's start

Figure 6.13: Stability diagrams for the model $\dot{r}_T = -r_T + \Phi(w \cdot r_T + I)$. Blue dots region illustrates a monostable behavior whereas light gray dots region illustrates a bistable behavior. Violet dots should have accounted for bifurcation curves (*i.e.* 2-fixed points curves) but it appears that the violet dots seen here are actually artefacts. **(c)** and **(d)** zoom in on the start of the bistable region.

the values of equilibria of the model. One could therefore wonder how does the number of fixed points behave as *both* parameters vary? The answer to this question lies in the *stability diagram* (Strogatz 1994) that shows the number of equilibria as a function of the two independent parameters (w and I).

In order to determine this stability diagram, one needs to compute numerically the fixed point(s) for a particular pair (w, I) and repeat for a large number of different pairs.

The stability diagrams for both methods are shown in Figure 6.13.

For Method 1 (Figures 6.13a and 6.13c), the stability diagrams indeed show a transition from a monostable regime (1 FP; blue dots) to a bistable regime (3 FP; light gray dots). Unfortunately, due to the complex shape of the transfer function $\Phi_1(x)$ and the high sensitivity of the solver to initial conditions, artefacts (violet dots) are present in Figure

6.13a. Violet dots were initially meant to illustrate the frontier between the monostable regime and the bistable regime, that is, showing the *bifurcation curves* where the model possesses two equilibria and where one of the two equilibria is a saddle-node. Due to the considered discretization of the (w, I) grid, these curves did not show. However, violet dots, thus indicating the presence of two FPs, appeared randomly in the gray-dotted region that, in turn, indicates clearly the bistable region. These violet dots are therefore actually artefacts of the solver that did not manage to find the third fixed point (usually the unstable one) of the model.

The gray uppermost point on the left in Figure 6.13c ($(w, I) \approx (0.04, -0.4)$) that starts the bistable region is close to a point called the *cusp point*, that is, the point at which both bifurcation curves meet tangentially and create a *co-dimension* 2 *bifurcation* (which is different from a saddle-node or transcritical bifurcation!).

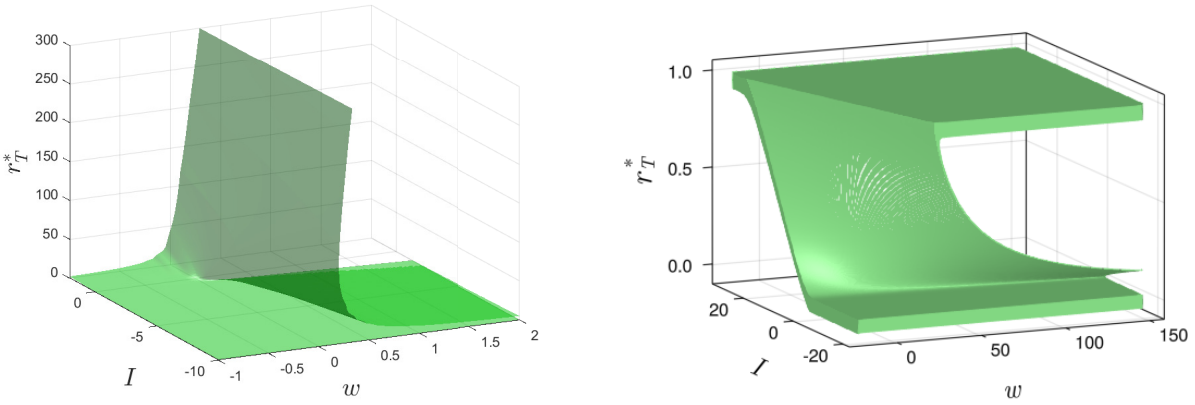
For Method 2, similar results can be observed. Stability diagrams also show a transition from a monostable regime to a bistable regime. The frontiers between these two regimes are the bifurcation curves along which the model displays saddle-nodes. Again, due to the discretization of the (w, I) grid, these curves did not show in the stability diagrams. The gray uppermost point on the left in Figure 6.13d is again close to the cusp point where bifurcation curves meet tangentially.

A difference between both models still persists though. When $I \geq 0$, Method 1 always displays a monostable regime whereas Method 2 always displays at least one bistable region. As a consequence, when $I \geq 0$, both models can be considered equivalent only when w lies in a restricted range, that is $w < w_{SN,1}$.

A note on stability of fixed points

Thus far, the stability of fixed points in the analyses using a geometric approach and in the bifurcation analyses was assessed graphically only. However, as explained in subsection 2.3.1, having a quantitative measure of *how stable* a fixed point is, is very common and often desired. Since the fixed points can be computed numerically using a non-linear solver (as done in this section to determine the number of equilibria), their stability can then be assessed mathematically by computing the eigenvalues of the jacobian evaluated at a fixed point's value. In other words, for a 1D model, one needs to compute

$$\lambda = \left.\frac{d\dot{r}_T}{dr_T}\right|_{r_T = r_T^*} = \frac{1}{\tau}\left(-1 + w\Phi'(w \cdot r_T^* + I)\right)$$



(a) Method 1 **(b)** Method 2

Figure 6.14: Cusp catastrophe surface of the model $\dot{r}_T = -r_T + \Phi(w \cdot r_T + I)$ for both methods.

and look at the sign of the real part to determine whether the fixed point is stable ($\mathcal{R}e\{\lambda\} < 0$) or unstable ($\mathcal{R}e\{\lambda\} > 0$). If $\mathcal{R}e\{\lambda\} = 0$, then the system is said to be marginally stable (Drion 2019-2020; Strogatz 1994) in the sense that there is no evolution of $r_T(t)$; it simply remains constant. This mathematical computation, performed on the fixed points found numerically for Figure 6.13, confirms the monostable and the bistable regimes.

6.6.6 Cusp catastrophe surface

Stability diagrams are useful to investigate the effects of both parameters w and I together but unfortunately, they loss the information about the *values* of the corresponding *equilibria*. A last way to plot all the effects and results together is then to draw the *cusp catastrophe surface*, that is, a three-dimensional surface illustrating the values of the equilibria as a function of w and I.

The surfaces are drawn for both methods in Figure 6.14. Exceptionally, these surfaces were drawn using the package[11] *GLMakie.jl* (Danisch and Krumbiegel 2021) for Method 2, or drawn using the *MATLAB* software[12] (MathWorks 2023; https://nl.mathworks.com/) for Method 1.

As expected, both methods show qualitatively similar results in the sense that each surface folds over on itself in certain places. These folds show the transition from a

[11] Traditional plotting tools did not allow to draw such complex 3D surfaces. Other more sophisticated visualization tools were thus needed.

[12] For an unknown reason, *GLMakie* did not produce any output for Method 1, although the same procedure as for Method 2 has been used. Hopefully, *MATLAB* managed to produce an output graph but at the cost of a well larger amount of computing time: ≈ 3 h.

monostable to a bistable system. With this single surface, one can actually obtain the bifurcation and the stability diagrams. The projection of the folds of the surface onto the (w, I) plane yields the stability diagram with the bifurcation curves. A slice (or cross section) at fixed w yields the $\{I, r_T^*\}-$bifurcation diagram whereas a slice at fixed I yields the $\{w, r_T^*\}-$bifurcation diagram (Strogatz 1994). The term *catastrophe* illustrates the fact that the state r_T of the model can be carried over the edge of the upper surface as the parameters are varied, and can then abruptly (or discontinuously) jump to the lower surface (Strogatz 1994).

The major difference between both surfaces is the presence/absence of upper saturation, as already discussed. This difference is thus explained by the shape of each individual transfer function. Method 1 uses a transfer function unbounded from above, allowing thus the equilibria to grow (potentially) to infinity. On the other hand, Method 2 uses a sigmoid-shaped transfer function that saturates as its argument becomes too large.

6.7 Conclusion on model equivalence and parameter sensitivity

The purpose of this chapter was to assess whether using a standard sigmoidal transfer function would give qualitatively the same dynamic behavior as the original model of Brunel and Lavigne 2009. Moreover, this chapter investigated the parameter sensitivity of the model of Brunel and Lavigne 2009 by varying the parameter values. To simplify the analyses and have a good understanding of the underlying dynamics, a one-dimensional model coding for a single target word was derived and used.

All the analyses of this chapter suggest that using a sigmoidal transfer function is equivalent, from a dynamic behavior perspective, to using the original transfer function. Both transfer functions allow the model to display monostable and/or bistable regimes. It is true, however, that this equivalence is only qualitative in the sense that the cusp point (*i.e.* start of bistable regime) occurs at different (w, I) pairs between the two methods. The monostable and bistable regimes occur therefore for different (w, I) pairs in general. Since these transfer functions are equivalent, it could be recommended to use a more standard transfer function that is easy to manipulate numerically and that gives a greater numerical stability.

In addition, the analyses suggest that the model from Brunel and Lavigne 2009 is sensitive to the parameter values of $p, J_1, J_I(\rightarrow w)$ and I. Indeed, by varying parameter p (the total number of populations or words), other parameters must change as well in

order to get back some properties (e.g. spontaneous activity). Moreover, the model is "sensitive" to parameters J_1, J_I and I because depending on their values, the model settles in one regime or another (monostable $\leftrightarrow$ bistable).

Chapter 7

Application to an experimental-like stimulus

Thus far, the dynamics of the 1D model was investigated using a *constant* input current. However, Chapter 5 discussed about different experimental tasks used to assess semantic priming. These tasks usually consist in the transient presentation of a prime word followed after a controlled delay by the transient presentation of a semantically (un)related target word. The presentation of a word acts therefore as an external excitatory stimulus/input for the population of neurons coding for that word. This experimental protocol can then be modeled by two current *pulses* of width *width* and amplitude *ampli*. Moreover, the two pulses are separated by a time delay *delay* (Figure 7.1). The stimulus onset asynchrony (SOA) is then computed as $SOA = width_P + delay$. The parameters *ampli, width* and *delay* can be tuned for both the prime (P) and the target (T) words ($\rightarrow$ 5 parameters in total). For a 1D model, a single pulse remains since only one word is coded.

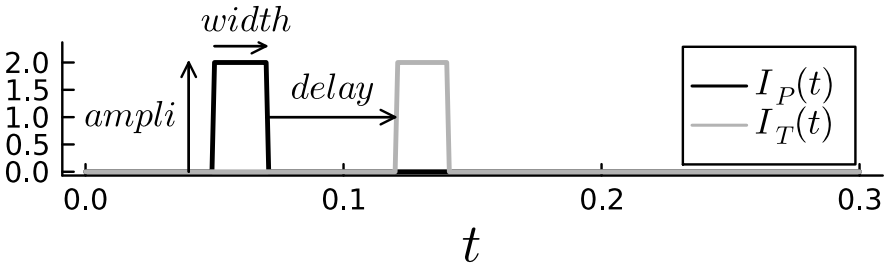


Figure 7.1: Traditional experimental protocol for semantic priming tasks. A prime word (P) is transiently presented with a strength *ampli* and a duration *width*. The transient presentation of a target word (T) follows that of the prime after a controlled delay *delay*. Parameters *ampli, width* and *delay* can be tuned according to the modeler/user's needs.

This chapter aims at investigating the response of the model to an input current pulse rather than a constant input current. The goal is to understand what changes with respect to simulations with a constant input. Moreover, since Chapter 6 suggested that the

model is equivalent when using $\Phi_1(x)$ or $\Phi_2(x)$, the investigation will be performed using $\Phi_2(x)$ due to its easier numerical manipulation. Using $\Phi_1(x)$ would still give the same results of course (see Appendix D).

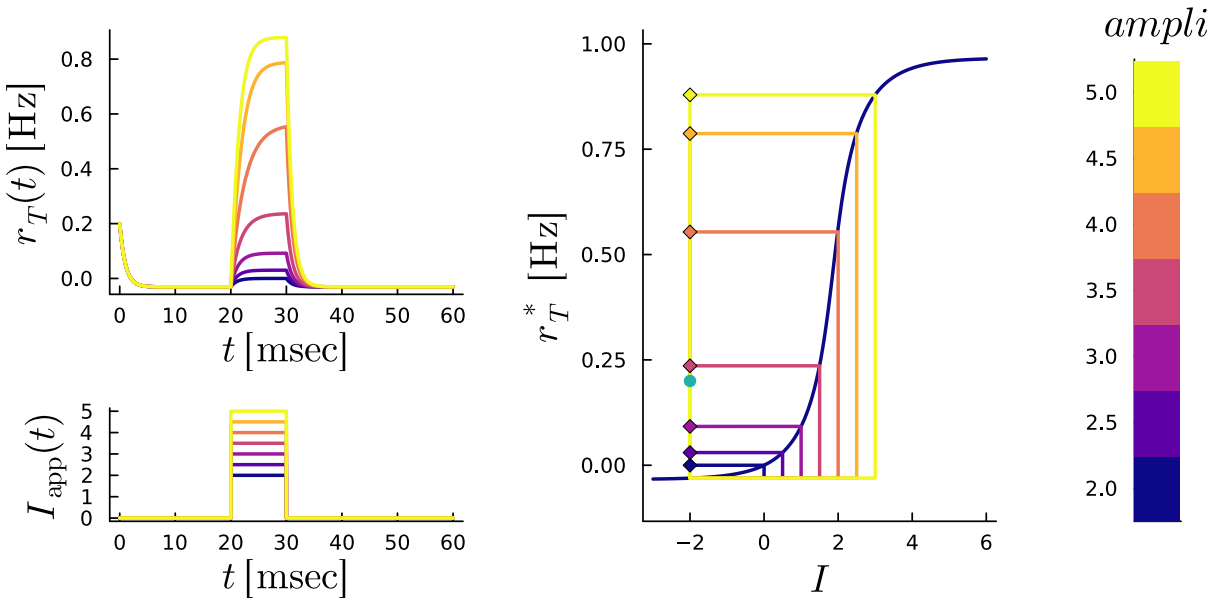
This chapter is structured as follows. First, the effects of the current pulse are assessed as a function of its amplitude. The effects are explored in the monostable regime as well as in the bistable regime. Then, the effects of the duration alone are examined again in both regimes. Furthermore, the combined effects of the amplitude and the duration are illustrated using a graph similar to a stability diagram. Finally, the evolution of response time as a function of the amplitude but for a fixed duration is shown as well.

7.1 Effect of current pulse amplitude

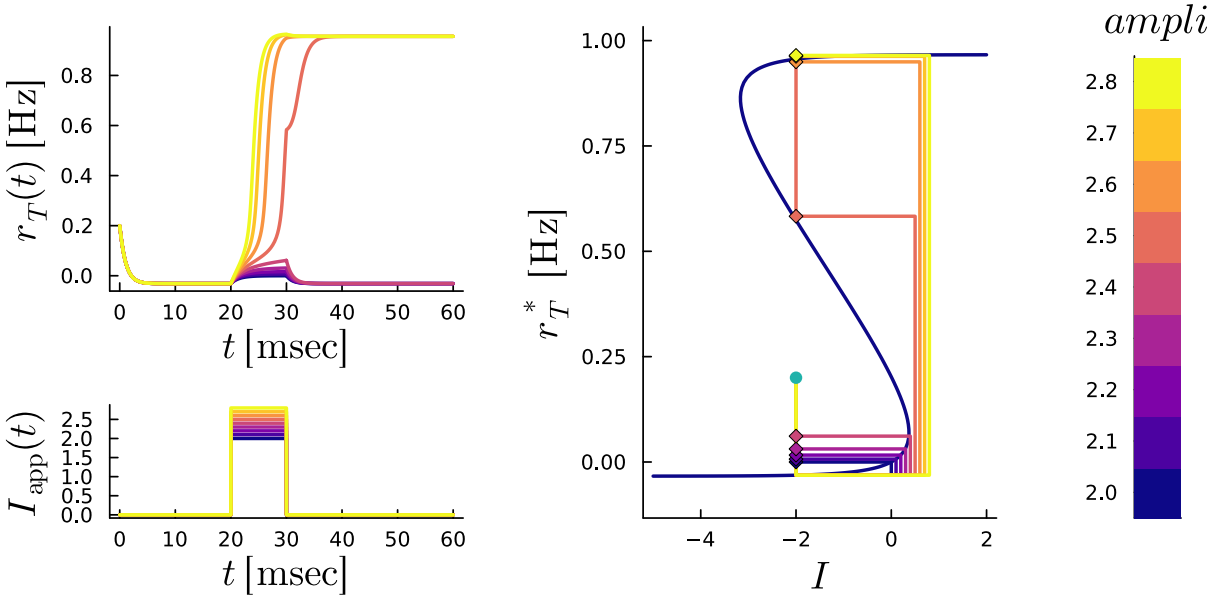
The effect of the amplitude of the pulse is shown in Figure 7.2a for a monostable system, and in Figure 7.2b for a bistable system.

For the monostable system, when the pulse is OFF at first (*i.e.* $I = I_{bias} + 0$), the population activity converges to the stable equilibrium associated to the bias current value. As long as the pulse is OFF, the system remains in this steady state. When the pulse becomes ON (*i.e.* $I = I_{bias} + ampli$), the system responds as an integrator, that is, the population activity evolves almost linearly and increases in order to reach the stable equilibrium associated to the new value of I. If the pulse is ON long enough, then the system reaches that stable equilibrium and remains constant as long as the pulse is ON. When the pulse ends, the system converges back to the stable equilibrium associated to the bias current. Since the system is monostable, the stable equilibrium that is reached at the end of the pulse is the same as that before the pulse became ON. This result can indeed be observed in the time evolution graph and the corresponding bifurcation diagram. Also, as the amplitude *ampli* increases, the steady state associated to I when the pulse is ON also increases according to the bifurcation diagram.

When the bias current is chosen so that the system lies in its bistable regime, the response of the system to the current pulse changes. When the pulse is OFF at first, the behavior is similar to that of a monostable system. When the pulse becomes ON, then depending on the amplitude of the applied current pulse, the system displays either the behavior of a monostable system (**(b)** purple and dark blue shades in bifurcation diagram and time evolution of the population activity) or it displays *persistent* (or *maintained, retrospective*) *activity* when the pulse becomes OFF again (brighter shades). In the semantic priming paradigm, this persistent activity would mean that the word is activated and retrieved from semantic memory. Two conclusions can therefore be made when the



(a) Monostable ($w = 2$)



(b) Bistable ($w = 9$)

Figure 7.2: Effect of input current pulse amplitude $I_{\text{app}}(t)$ for **(a)** a monostable system and **(b)** a bistable system. The input current pulse $I_{\text{app}}(t)$ (bottom left of a subfigure) is applied for a fixed duration of 10 [msec] from $t = 20$ to $t = 30$. The amplitude of the pulse is color-coded. The time evolution of the population activity for the different amplitudes (top left of a subfigure) shows a difference in behaviors between monostable and bistable regimes. The associated $\{I, r_T^*\}-$bifurcation diagram (dark-blue/black curve; right of a subfigure) allows one to make the link between variations in $I = I_{bias} + I_{\text{app}}$ and the activity to which the population converges. Colored straight lines in the bifurcation diagram illustrate the $r_T(t)$ VS $I(t)$ trajectory. For visualization purposes, a constant negative bias current ($I_{bias} = -2$) has been applied. Green dot stands for the initial condition of all trajectories. Diamond markers spot the end of the applied current pulse (*i.e.* $t = 30$ [msec]).

system is in a bistable regime:

1. The high stable attractor can only be reached if the amplitude of the applied current pulse allows the system to go beyond the *saddle-node* value of the bistable region. In other words, the amplitude of the applied current must satisfy $I_{\text{app}} > I_{thresh} = I_{SN,1} - I_{bias}$.

2. Semantic priming can only occur if the semantic memory system is in a bistable regime for the prime and the target at least. Otherwise, word representation could not display persistent activity after the presentation of the word.

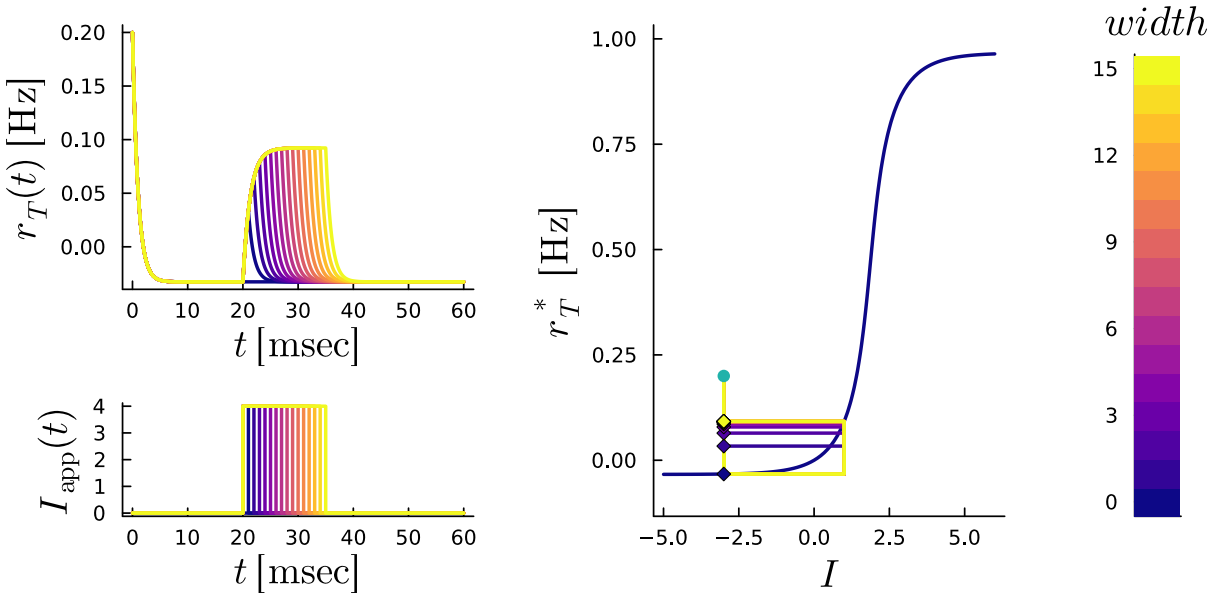
The bifurcation diagram in Figure 7.2b suggests that when the pulse ends, the system converges to the stable equilibrium associated to I_{bias}, but in addition it converges to the stable equilibrium whose basin of attraction includes the value of the population activity at the end of the pulse. Put differently, $r_T(t)$ converges to the high stable equilibrium if $r_T(t = 30) > r^*_{T,unstable}$ and to the low stable equilibrium otherwise. Moreover, this result suggests an influence of the pulse duration. This influence is thus investigated hereafter.

7.2 Effect of pulse duration

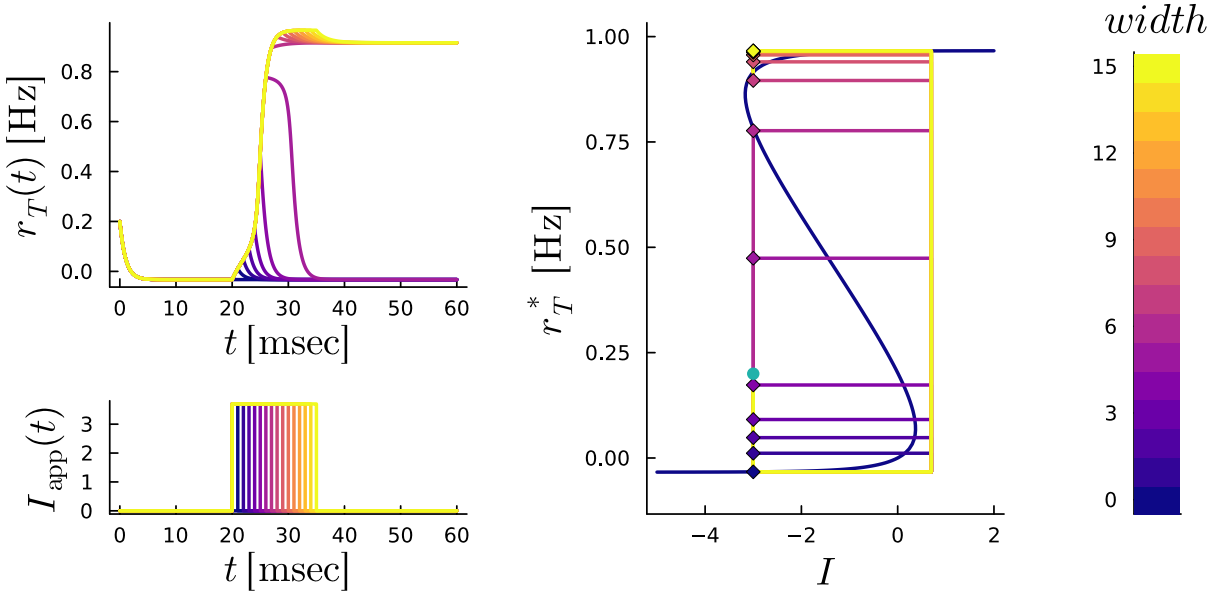
Similarly to the effect of amplitude, the effect of pulse duration is shown in Figure 7.3a for a monostable system and in Figure 7.3b for a bistable system.

For a monostable system, the duration only affects whether the system will reach the stable equilibrium associated to $I = I_{bias} + I_{app}$ when the pulse is ON. When the pulse is too short (dark blue shades), the system does not have the time to reach that stable equilibrium and thus converges back, from where it managed to go, to the state associated to I_{bias} when the pulse ends. On the other hand, if the pulse lasts long enough (purple and brighter shades), then the system manages to reach the stable steady state associated to $I = I_{bias} + I_{app}$ and remains in that state as long as the pulse is ON. When the pulse ends, the system converges back to its original steady state as usual.

For the bistable system, the duration has an impact on whether the high stable attractor is reached. For visualization purposes, a bias current ($I_{bias} = -3$) has been applied and the amplitude of the applied input current pulse is chosen to satisfy $I_{app} > I_{thresh}$. The pulse can thus potentially make the system reach the high stable attractor. When looking at the time evolution of the population activity (Figure 7.3b top left), one can see that only the pulse width affects the final steady state. When looking at the associated bifurcation diagram, one can then understand that the duration of the pulse should make the population activity $r_T(t)$ greater than the value of the unstable fixed



(a) Monostable ($w = 2$)



(b) Bistable ($w = 9$)

Figure 7.3: Effect of duration of the input current pulse $I_{\text{app}}(t)$ for **(a)** a monostable system and **(b)** a bistable system. The input current pulse $I_{\text{app}}(t)$ (bottom left of a subfigure) has a fixed amplitude and is applied for a variable duration. The fixed amplitude is chosen to be greater than the threshold value $I_{SN,1} - I_{bias}$. The duration of the pulse is color-coded. The time evolution of the population activity for the variable durations (top left of a subfigure) shows a difference in behaviors between monostable and bistable regimes. The associated $\{I, r_T^*\}-$bifurcation diagram (dark-blue/black curve; right of a subfigure) allows one to make the link between variations in duration and the activity to which the population converges. Colored straight lines in the bifurcation diagram illustrate the $r_T(t)$ VS $I(t)$ trajectory. For visualization purposes, a constant negative bias current ($I_{bias} = -3$) has been applied. Green dot stands for the initial condition of all trajectories. Diamond markers spot the end of the applied current pulse (*i.e.* $t = 20 + width$ [msec]).

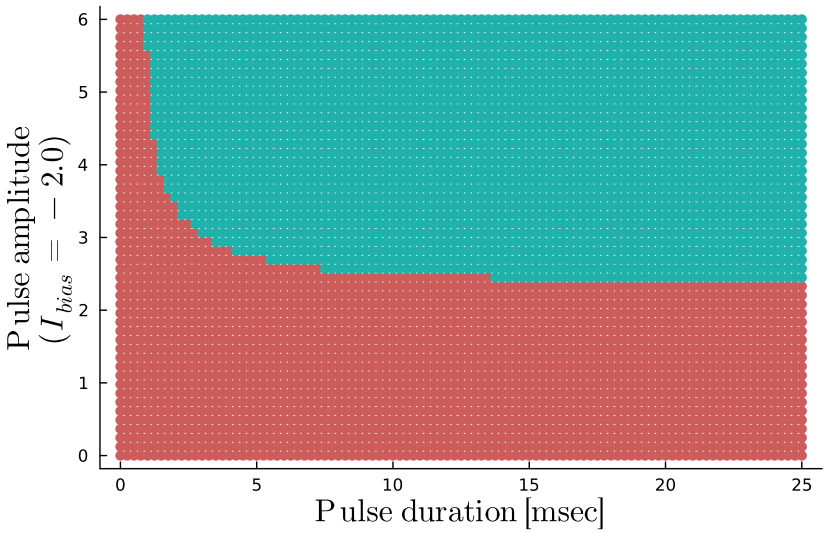


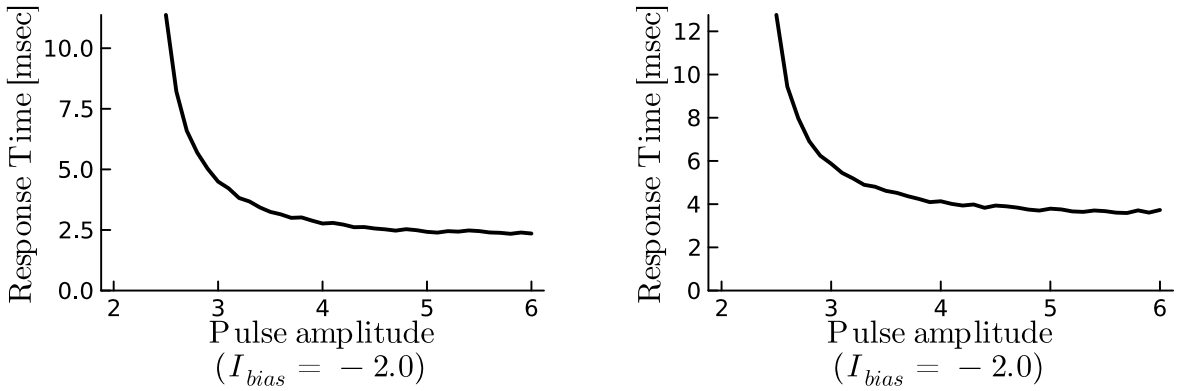
Figure 7.4: Combined effects of pulse duration and pulse amplitude for a bistable system ($w = 9$) with a bias current ($I_{bias} = -2$) in the bistable region. Red area indicates that the system did not jump to the high stable state associated to I_{bias} whereas green area indicates that the system did jump to that high state.

point at the end of the pulse. The perfect illustration is for $width = 7$ (light purple) with $r_T(t = 27)$ being a tiny below the unstable fixed point. This study of the pulse duration allows one to conclude that there exist two thresholds actually: one for the pulse amplitude ($I_{thresh} = I_{SN,1} - I_{bias}$) and the other for the pulse duration ($r_T(t = \text{end pulse}) > r^*_{T,unstable}$ associated to I_{bias}).

7.3 Combined effects of pulse duration and amplitude

For a monostable system, the combined effects of pulse duration and amplitude only make the system remain more or less long in the stable equilibrium associated to $I = I_{bias} + I_{app}$. When the pulse ends, the population activity inevitably converges back to the steady state associated to I_{bias}.

For a bistable system, in turn, it has been seen that both the duration and the amplitude affect whether the system jumps to the high stable attractor or not. Figure 7.4 shows both effects together (similarly to a stability diagram). With this Figure, one can indeed see the threshold on the current amplitude because the system jumps to the high stable state (green area) associated to $I_{bias} = -2$ only from $ampli \gtrapprox 2.4 = I_{SN,1} - I_{bias}$. In addition, for amplitudes greater than this threshold, one can observe that for a fixed value of pulse amplitude, a minimum duration for the pulse is required in order for the system to reach the high steady state. Put differently, a minimum duration for the pulse



(a) Response criterion: $r_{SN,2}$ associated to $I_{SN,2}$

(b) Response criterion: 98% of the high stable state associated to I_{bias}

Figure 7.5: Evolution of response times of a bistable system ($w = 9$) as a function of pulse amplitude. The pulse duration is fixed to 15 [msec]. A fixed bias current $I_{bias} = -2$ is applied to study the bistable regime. **(a)** The response criterion corresponds to the value of the saddle-node activity associated to the low saddle-node current $I_{SN,2}$. **(b)** The response criterion is 98% of the final value of the population activity when the system jumps.

is required to cross the unstable fixed point associated to I_{bias}. Moreover, the relationship between the pulse amplitude and the minimum pulse duration seems to be inverse, that is, as the pulse amplitude increases, the minimum pulse duration that is required decreases. This result suggests that the pulse amplitude has an impact on the speed of convergence to the high stable state. Indeed, if the amplitude increases, then $\Phi(w \cdot r_T + I)$ increases making then $\dot{r}_T$ larger. Crossing the unstable fixed point occurs then earlier thus making the minimum duration shorter.

7.4 A look at response times

Response or reaction times are often used in psychology experiments to measure cognitive abilities of participants. In the semantic priming paradigm, this response time is measured from the target onset until the population activity of a item reaches a response criterion. Figure 7.5 shows the evolution of response times (RT) when the response criterion is the saddle-node activity $r_{SN,2}$ associated to the lowest saddle-node current $I_{SN,2}$ (left), or when the response criterion is set to 98% of the final high steady state value (right). Moreover, this evolution is observed as a function of pulse amplitude. The pulse duration is fixed to 15 [msec].

Using one or the other criterion gives actually the same trend: as the amplitude increases, the response time decreases. Put differently, the response criterion changes the *quantitative* value of response time but not its *qualitative* evolution. In addition, if the

pulse amplitude does not satisfy the threshold value, then the response time tends towards infinity because the system does not jump to the high stable state. Furthermore, response times seem to become independent of the pulse amplitude as the latter becomes very large. This result suggests that a minimum amount of time is necessary to process information.

Using $r_{SN,2}$ as response criterion has the advantage that the response time value is independent of the bias current value that is applied. In other words, the trend and the value of response times will be the same no matter what the x-axis is precisely (due to a change in I_{bias}). However, if the transfer function is not saturated from above, then the high stable state could potentially be far away from $r_{SN,2}$, suggesting therefore the potential need for an extra amount of time to process information. This is the reason why choosing a high percentage of the high steady state value as a response criterion can be used as well. The response time is then specific to the high steady state that is considered (and therefore to the bias current that is considered). A drawback of this second response criterion is that any percentage can "do the trick" *a priori* (*i.e.* the value of the percentage is totally arbitrary).

7.5 Conclusions on the application of the model to a pulse-shaped stimulus

This chapter aimed at investigating the response of the 1D model to a pulse-shaped stimulus. The pulse-shaped stimulus mimics the real experimental protocol (*i.e.* transient presentation of a word) that is followed during psychology experiments. The amplitude and the duration of the pulse are two parameters that can be tuned independently to the experimenter's desires.

Analyses revealed that both the amplitude and the duration determine together the final steady state to which the system converges, especially for a bistable system. When the system is biased in the bistable region, the amplitude must be greater than the threshold value $I_{thresh} = I_{SN,1} - I_{bias}$ in order for the system to potentially reach the high stable state. If the amplitude satisfies this condition, then the pulse should still be long enough to allow the system to cross the unstable fixed point associated to the bias current value. This unstable fixed point is thus a key player in the sense that it is a separator between both stable attractors and their basins of attraction in the bistable region, but it is also a threshold for the pulse duration. In addition, these analyses suggested that semantic priming can only occur if the semantic memory system is in a bistable regime in order for the word representations to show persistent (or retrospective) activity.

The last part of this chapter focused on the response time computation when the system is in a bistable regime. Although several and different response criteria can be used, the evolution of response times as a function of the pulse amplitude is always the same: as the amplitude increases, the response time decreases. When the amplitude does not satisfy the threshold value, the response time tends towards infinity because the system never jumps to the high stable state. On the other side, the response time becomes independent of the amplitude as the latter becomes very large. The strictly positive response times suggest that a minimum amount of time to process information is always required.

Conclusions and Perspectives

This master thesis revisited the network model of Brunel and Lavigne 2009 in order to assess whether the same model with the same assumptions but using a different transfer function would give a qualitatively similar dynamic behavior. In addition, this master thesis investigated the parameter sensitivity. These two components were motivated by the facts that Brunel and Lavigne used a transfer function that is mathematically good-looking but rather difficult to manipulate numerically, and by the highly specific definitions for some parameters. In order to assess this model equivalence and this parameter sensitivity, the one-dimensional version of Brunel and Lavigne's model has been derived and used, and a more standard sigmoid transfer function has been considered.

Part I and **Part II** reviewed the literature on mathematical modeling, 1D model analysis, rate models, memory and semantic priming. They gave all the tools, notably the phase portrait and the bifurcation tools, that were necessary to carry out the dynamical analysis of Brunel and Lavigne's model.

Phase portrait and bifurcation analyses suggest that both transfer functions are qualitatively equivalent from a dynamic behavior perspective. Both allow the model to display monostable and/or bistable regimes, the latter being induced by particular ranges of values for the recurrent connection strength w and the total external input I. Using the sigmoid function would therefore give the same qualitative result with greater numerical stability.

In addition, phase portrait and bifurcation analyses also suggest that the model from Brunel and Lavigne 2009 is rather sensitive to parameter values because by varying one parameter (e.g. p), others should vary as well in order to get back some properties (e.g. spontaneous activity).

To be complete, a pulse-shaped experimental-like stimulus has been applied to the one-dimensional model to observe its response. The results show that both the amplitude and the duration of the input current pulse determine together the final steady state to which the system converges especially for a bistable system. When the system is biased in

the bistable region, the amplitude and the duration must be greater than their respective threshold ($I_{thresh} = I_{SN,1} - I_{bias}$ and long enough to cross the unstable fixed point) in order for the system to reach the high stable attractor. The key role of the unstable fixed point has been stressed. Regarding response times, several response criteria can be used but the trend is always the same: the response time decreases as the pulse amplitude increases. Moreover the response time is bounded by infinity (*i.e.* amplitude below threshold) and by a strictly positive value (*i.e.* necessary amount of time to process information).

Perspectives

Several limitations and thus perspectives can be noticed in this work.

First, the 1D model is very useful to investigate the full dynamic behavior of the model but it does not allow to study semantic priming as such. Indeed, semantic priming requires at least a *pair* of words (prime and target) such that priming of the target word can effectively occur. By adding one variable (and thus one word) to the 1D version, the model would become a *two-dimensional* model where semantic priming can be studied as such. Also, this 2D model should have extended versions of the tools (phase portrait, bifurcations, ...) used in this work. An introduction of this 2D model can be found as a bonus chapter in Appendix A.

Then, why limit oneself to two words? A common associate to the prime and the target words could be added to form a three-dimensional model. This 3D model could then be studied considering the slow and fast dimensions, that is, the dimensions associated to slightly negative eigenvalues ($\rightarrow$ slow) and to strongly negative eigenvalues ($\rightarrow$ fast). The slow dimension(s) is (are) usually where all the dynamics happens since the other fast dimensions are exponentially contracting (Franci 2023b).

Also, this work used the same assumptions as those of Brunel and Lavigne: the global inhibitory current regulating the activity of all excitatory populations has a linear transfer function and the time scale of inhibitory dynamics is much shorter than that of excitatory populations. However, this master thesis did not check whether these assumptions are still valid when $p = 1$. It is actually common that inhibitory populations have a different transfer function and a different time scale from excitatory populations (Gjorgjieva et al. 2021a), but the validity of the assumption of the inhibitory firing rate being proportional to the mean of excitatory average firing rates could be further explored. In particular, determining the minimum value for p such that this assumption is valid could be useful.

Finally, Brunel and Lavigne used symmetric connectivity matrices (see Appendix A for the 2D matrices) suggesting that the association between words is symmetric. This symmetry also enables to determine the conditions to have the same spontaneous activity for each node. However, their model does not therefore take into account pairs of words that are strongly associated in a forward or in a backward direction. For example, *baby-boy* has a strong forward association whereas *boy-baby* has a weaker backward association. The need for such a model is therefore present.

Appendices

Appendix A

Bonus chapter: A glimpse at the two-dimensional model

A drawback of the 1D model is that it cannot be used as such to investigate semantic priming since the cognitive process requires at least a *pair* of words. This bonus chapter aims then at introducing the two-dimensional (2D) model and explaining how concepts and tools seen in 1D can be extended to higher-order dimensions.

Similarly to the 1D model, by setting $p = 2$, the Brunel and Lavigne's model reduces to

$$\begin{cases} \tau \frac{dr_P}{dt} &= -r_P + \Phi \left[\underbrace{\frac{J_1 - J_I}{2}}_{w_{PP}} \cdot r_P + \underbrace{\frac{J_{PT} - J_I}{2}}_{w_{PT}} \cdot r_T + I^P_{ext} + I^P_{sel} \right] \\ \tau \frac{dr_T}{dt} &= -r_T + \Phi \left[\underbrace{\frac{J_{TP} - J_I}{2}}_{w_{TP}} \cdot r_P + \underbrace{\frac{J_1 - J_I}{2}}_{w_{TT}} \cdot r_T + I^T_{ext} + I^T_{sel} \right] \end{cases} \tag{A.1}$$

where P and T stand for prime and target, respectively, and J_{PT} (and J_{TP}) $\in \{J_a, J_0\}$ are the connection weights from T (P) to P (T). The other parameters have the same meaning as for the 1D model and their values can be found in Table 6.1.

This 2D model can be written more compactly using matrix and vector notations (denoted by uppercase and bold lowercase letters, respectively):

$$\tau \dot{\mathbf{r}} = -\mathbf{r} + \Phi \left[\underbrace{\begin{pmatrix} w_{PP} & w_{PT} \\ w_{TP} & w_{TT} \end{pmatrix}}_{W} \cdot \mathbf{r} + \mathbf{I_{ext}} + \mathbf{I_{sel}} \right] \tag{A.2}$$

with $\mathbf{r} = (r_P, r_T)^T$ (^{T} stands for the transpose), $\mathbf{I_{ext}} = (I^P_{ext}, I^T_{ext})^T$ and $\mathbf{I_{sel}} = (I^P_{sel}, I^T_{sel})^T$.

The network model in Eq. (A.2) amounts to two non-overlapping populations of excitatory neurons being selective to a single word or item. Each population or word is

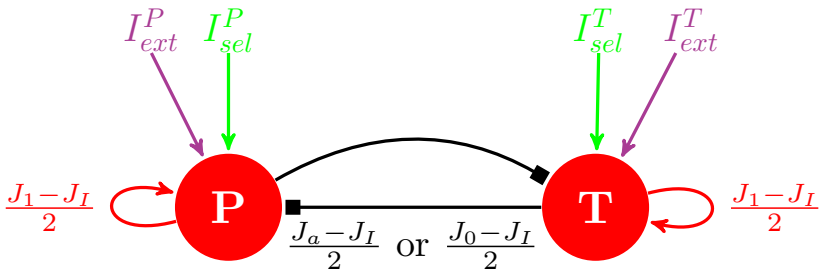


Figure A.1: Prime-Target network model. Each node receives synaptic inputs from itself and the other node (red and black arrows). Each node also receives external inputs that are either selective (green arrow) or non-selective (purple arrows) to the word encoded by each node.

then modeled as a node in a graph. The two nodes make connections to themselves (w_{PP}, and w_{TT}) and to each other (w_{PT}, and w_{TP}). In addition, the nodes receive non-selective external inputs (I^i_{ext}) that bias them into a specific spontaneous activity. They also receive selective external inputs (I^i_{sel}) that excite or activate the word to which they are selective. Figure A.1 illustrates the 2D model. Furthermore, the two words can be considered as belonging within the same semantic group (*i.e.* $p = p_g \cdot p_i$ with $p_g = 1, p_i = 2$) or into two different semantic groups (*i.e.* $p_g = 2, p_i = 1$).

A.1 Spontaneous activity

Similarly to the 1D model, the conditions to have spontaneous activities r^P_{spont} and r^T_{spont} in the absence of any other external input can be extended to a 2D model. In order to have these spontaneous activities, the model should satisfy several conditions:

1. The state $\mathbf{r}_{spont} = (r^P_{spont}, r^T_{spont})$ should be a fixed point (or equilibrium) of the model (A.2). That is

$$\begin{cases} \dot{r}_P = 0 & \leftrightarrow r_P = \Phi\left(w_{PP} \cdot r_P + w_{PT} \cdot r_T + I^P_{ext}\right) \\ \dot{r}_T = 0 & \leftrightarrow r_T = \Phi\left(w_{TP} \cdot r_P + w_{TT} \cdot r_T + I^T_{ext}\right) \\ \dot{r}_P = \dot{r}_T = 0 & \leftrightarrow w_{PP} \cdot r^P_{spont} + w_{PT} \cdot r^T_{spont} + I^P_{ext} - \Phi^{-1}(r^P_{spont}) \\ & \quad = w_{TP} \cdot r^P_{spont} + w_{TT} \cdot r^T_{spont} + I^T_{ext} - \Phi^{-1}(r^T_{spont}) \end{cases} \tag{A.3}$$

   When either $\dot{r}_P = 0$ or $\dot{r}_T = 0$ (first two equations), then the corresponding curves are called *nullclines* or *null isoclines*. The fixed points are therefore the intersections of nullclines. The last statement is actually valid for any higher-order dimension.

2. The spontaneous activities should be stable, that is, the so-called *Jacobian* matrix

$$\left.\frac{d\dot{\mathbf{r}}}{d\mathbf{r}}\right|_{\mathbf{r}=(r^P_{spont}, r^T_{spont})} = \left.\begin{pmatrix} \frac{d\dot{r}_P}{dr_P} & \frac{d\dot{r}_P}{dr_T} \\ \frac{d\dot{r}_T}{dr_P} & \frac{d\dot{r}_T}{dr_T} \end{pmatrix}\right|_{\mathbf{r}=(r^P_{spont}, r^T_{spont})}$$

   evaluated at the spontaneous activities should have eigenvalues with a negative real part to ensure the decay of small perturbations nearby the fixed point. This 2 by 2

matrix is the 2D extension of the slope $f'(x)$ in 1D (see section 2.3.1). The N by N square matrix would then be the N-D extension of that 1D slope $f'(x)$.

When $r^P_{spont} = r^T_{spont} = r_{spont}$, the following sufficient conditions ensure that this spontaneous activity for each node is a fixed point (but they do not say anything about the stability!)

$$\begin{cases} w_{PP} = w_{TT} = w_1 \\ w_{TP} = w_{PT} = w_2 \\ I^P_{ext} = I^T_{ext} = I_{ext} \\ I_{ext} = \Phi^{-1}(r_{spont}) - (w_1 + w_2) \cdot r_{spont} \end{cases} \tag{A.4}$$

with $\Phi^{-1}(x)$ the inverse transfer function that is guaranteed to exist since $\Phi(x)$ is a continuous and monotonically increasing function. The last expression in Eq (A.4) reminds that found for the 1D model (see Eq. (6.6a)) with $w = w_1 + w_2$.

Condition 2. amounts to computing the eigenvalues of

$$\left.\frac{d\dot{\mathbf{r}}}{d\mathbf{r}}\right|_{\mathbf{r}=(r^P_{spont}, r^T_{spont})} = -I_{2\times 2} + diag\left(\mathbf{\Phi'}\left[W \cdot \mathbf{r}_{spont} + \mathbf{I_{ext}}\right]\right) \cdot W \tag{A.5}$$

with $I_{2\times 2}$ the 2D identity matrix, $\mathbf{r}_{spont} = (r^P_{spont}, r^T_{spont})$ and $diag(.)$ a 2D diagonal matrix whose entries are

$$\Phi'(w_{PP} \cdot r^P_{spont} + w_{PT} \cdot r^T_{spont} + I^P_{ext}) \quad \text{and} \quad \Phi'(w_{TP} \cdot r^P_{spont} + w_{TT} \cdot r^T_{spont} + I^T_{ext})$$

The identity matrix simply shifts the eigenvalues λ of $diag\left(\mathbf{\Phi'}\left[W \cdot \mathbf{r}_{spont} + \mathbf{I_{ext}}\right]\right) \cdot W$ by one unit. Thus, computing the eigenvalues $\tilde{\lambda}$ of Eq. (A.5) amounts to computing the eigenvalues λ.

Assuming that $r^P_{spont} = r^T_{spont} = r_{spont}$, the problem reduces to finding

$$\tilde{\lambda}_{1,2} = -1 + f \cdot \lambda_{1,2}(W) \quad \text{with} \quad f = \Phi'(\Phi^{-1}(r_{spont}))$$

One finds[13]

$$\lambda_{1,2}(W) = \frac{(w_{PP} + w_{TT}) \pm \sqrt{(w_{PP} + w_{TT})^2 - 4(w_{PP}w_{TT} - w_{TP}w_{PT})}}{2}$$

Considering the sufficient conditions from Eq. (A.4), the eigenvalues simplify into

$$\lambda_{1,2}(W) = w_1 \pm w_2$$

As a result, the spontaneous activities $r^P_{spont} = r^T_{spont} = r_{spont}$ give a stable background

[13] As a reminder, the eigenvalues of a matrix A are found by solving the equation $\det(A - \lambda I) = 0$. For a 2D matrix, the eigenvalues are thus given by $\lambda_{1,2} = \frac{-\tau \pm \sqrt{\tau^2 - 4\Delta}}{2}$ with $\tau = a_{11} + a_{22}$ the trace of A and $\Delta = a_{11}a_{22} - a_{12}a_{21}$ the determinant of A.

	P-T Associated (A)	P-T Non associated (NA)
P-T within same group ($p_i = 2$)	$\begin{pmatrix} \frac{J_S}{2} & -\frac{J_S}{2} \\ -\frac{J_S}{2} & \frac{J_S}{2} \end{pmatrix}$	$\begin{pmatrix} \frac{J_S}{2} & -\frac{J_S}{2}\frac{(1+a)}{(1-a)} \\ -\frac{J_S}{2}\frac{(1+a)}{(1-a)} & \frac{J_S}{2} \end{pmatrix}$
P-T within different groups ($p_i = 1$)	$\begin{pmatrix} \frac{J_S}{2} & \frac{J_S(2a-1)}{2} \\ \frac{J_S(2a-1)}{2} & \frac{J_S}{2} \end{pmatrix}$	$\begin{pmatrix} \frac{J_S}{2} & -\frac{J_S}{2} \\ -\frac{J_S}{2} & \frac{J_S}{2} \end{pmatrix}$

Table A.1: Possible configurations for the connectivity matrix W in the 2D network model of Brunel and Lavigne 2009. J_S is the synaptic potentiation strength (> 0) and a is the association strength ($0 < a < 1$) between the prime (P) and the target (T) words (see Table 6.1).

state if all the following conditions are met

$$w_{PP} = w_{TT} = w_1 \tag{A.6a}$$
$$w_{TP} = w_{PT} = w_2 \tag{A.6b}$$
$$I^P_{ext} = I^T_{ext} = I_{ext} \tag{A.6c}$$
$$I_{ext} = \Phi^{-1}(r_{spont}) - \underbrace{(w_1 + w_2)}_{w} \cdot r_{spont} \tag{A.6d}$$
$$\lambda_{1,2}(W) = w_1 \pm w_2 < \frac{1}{f} \tag{A.6e}$$

To choose w_1 and w_2, one should first find

$$w_{max} = \frac{1}{f} = \frac{1}{\Phi'(I_{w=0})}$$

(similarly to the 1D case), and then choose w_1 satisfying

$$w_1 < \frac{1}{2}(w + \frac{1}{f})$$

The other connection weight w_2 is then automatiquely determined as $w_2 = w - w_1$. These conditions on w and w_1 indeed satisfy Eqs. (A.6d) and (A.6e), and can be checked numerically.

Thus far, the equations are actually valid for any 2D model but Brunel and Lavigne 2009 used very specific definitions for their connection strengths J_a and J_0 (see Table 6.1). Four possible configurations for W therefore arise (Table A.1).

From Table A.1, one can see that all configurations of W are symmetric suggesting that the sufficient conditions (A.4) to have $r^P_{spont} = r^T_{spont} = r_{spont}$ can be met. The stability of that background activity state depends then on the values of J_S (> 0; strength

	P-T Associated (A)	P-T Non associated (NA)
P-T within same group ($p_i = 2$)	$0 < J_S < \frac{1}{f}$ $0 < a < 1$	$0 < J_S < \frac{1}{f}$ $0 < a < 1 - fJ_S$
P-T within different groups ($p_i = 1$)	$0 < J_S < \frac{2}{f}$ $\max\{1 - \frac{1}{fJ_S}, 0\} < a$ $a < \min\{\frac{1}{fJ_S}, 1\}$	$0 < J_S < \frac{1}{f}$ $0 < a < 1$

Table A.2: Stability conditions on synaptic potentiation strength J_S and association strength a. These conditions ensure that the background activity state $r^P_{spont} = r^T_{spont} = r_{spont}$ is stable.

of synaptic potentiation) and a ($0 < a < 1$; association strength between prime P and target T). Based on Eq. (A.6e), one can then find the conditions on J_S and a for each W (Table A.2). These conditions can also be checked numerically.

The configurations with $P-T$ belonging within the same group are the most commonly encountered cases. Thus, similarly to the 1D case, it is possible to choose a desired stable background activity state, and to bias each node accordingly using Eq. (A.6d) and Table A.2. One will notice that the default value for J_S (*i.e.* $3.65 > \frac{1}{f} \approx 0.04$) still does not allow the $P - T$ network to have a stable background activity state at the default value for $r_{spont} = 5$ [Hz], suggesting again a parameter sensitivity.

A.2 Phase plane analysis

The vector (r_P, r_T) is called the *state* of the system because it is the minimal information that is necessary and sufficient to know in order to predict any future value of the solution of Eq. (A.2) (Strogatz 1994).

The vector $(\dot{r}_P, \dot{r}_T)$ is in turn called the *velocity vector* at point (r_P, r_T) on the (r_P, r_T) plane because it indicates *how fast* or *how slow* each variable evolves with time. The length of the vector is proportional to the magnitude of $\dot{r}_P$ and $\dot{r}_T$ (Strogatz 1994).

The solutions of Eq. (A.2) are thus the trajectories $(r_P(t), r_T(t))$ similarly to the one-dimensional case. Moreover, the trajectories can also be visualized on the (r_P, r_T) plane called the *phase plane*. Equation (A.2) therefore represents the *two-dimensional vector field* on the phase plane. The trajectories $(r_P(t), r_T(t))$ can thus be obtained by flowing along the vector field from any point in the plane. The last statement also implies that the entire phase plane is full of trajectories because each point in the plane can act as an initial condition (Strogatz 1994).

Similarly to the one-dimensional case, one can sketch the phase plane of a 2D model

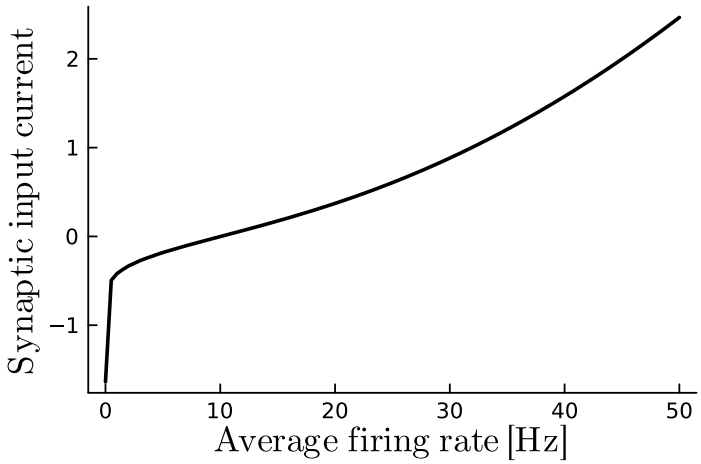


Figure A.2: Inverse transfer function of $\Phi_1(x)$. No analytical closed form can be obtained but the inverse function can be numerically approximated using interpolation methods.

in order to get an idea of the global pattern formed by the trajectories. It is often useful to first draw the *nullclines* of the system onto the phase plane. As explained previously, the nullclines indicate where $\mathbf{r}(t)$ is purely horizontal or vertical (Strogatz 1994). In other words, the nullclines are the curves (for 2D models) where exactly one component of the vector field is equal to zero. Considering the 2D model (A.2), the nullclines are given by the curves where either $\dot{r}_P = 0$ or $\dot{r}_T = 0$, that is

$$\dot{r_P} = 0 \leftrightarrow r_P = \Phi\left(w_1 \cdot r_P + w_2 \cdot r_T + I^P_{ext} + I^P_{sel}\right) \tag{A.7}$$

and

$$\dot{r_T} = 0 \leftrightarrow r_T = \Phi\left(w_2 \cdot r_P + w_1 \cdot r_T + I^T_{ext} + I^T_{sel}\right) \tag{A.8}$$

The P (T) nullcline, *i.e.* Eq. (A.7) (*i.e.* Eq. (A.8)) can also be expressed in this case as a function of one of the two variables only thanks to the inverse transfer function $\Phi^{-1}(x)$ (that exists since $\Phi(x)$ is a continuous and monotonically increasing function). One finds for the P nullcline and the T nullcline, respectively

$$r_T = \frac{1}{w_2}\left[\Phi^{-1}(r_P) - w_1 \cdot r_P - (I^P_{ext} + I^P_{sel})\right] \tag{A.9}$$

$$r_P = \frac{1}{w_2}\left[\Phi^{-1}(r_T) - w_1 \cdot r_T - (I^T_{ext} + I^T_{sel})\right] \tag{A.10}$$

It should be noted that if $w_2 = 0$, then the system is a 2D *decoupled* system which actually amounts to two 1D models. The variables then evolve independently of each other and the nullclines do not have a closed form. In addition, for Method 1 with $\Phi_1(x)$, the corresponding inverse transfer function $\Phi_1^{-1}(x)$ has no closed form either but it can still be approximated numerically using *interpolation* methods. As a result, the inverse transfer function of Method 1 looks like in Figure A.2.

The nullclines therefore partition the phase plane into regions (Table A.3) (Drion 2019-2020; Drion 2021-2022; Strogatz 1994).

$\dot{r}_T$ \ $\dot{r}_P$	$\dot{r}_P < 0 : \downarrow$	$\dot{r}_P = 0 : /$	$\dot{r}_P > 0 : \uparrow$
$\dot{r}_T < 0 : \leftarrow$	$\swarrow$	$\leftarrow$	$\nwarrow$
$\dot{r}_T = 0 : /$	$\downarrow$	FP	$\uparrow$
$\dot{r}_T > 0 : \rightarrow$	$\searrow$	$\rightarrow$	$\nearrow$

Table A.3: Vector field directions. On the nullclines ($\dot{r}_P = 0$ or $\dot{r}_T = 0$; blue row and column), the movement is purely vertical or horizontal. Besides nullclines, the global vector field is a combination of each individual vector field of each variable. Fixed points (FP) are obtained when both nullclines intersect.

Similarly to the phase portrait, the stability of a fixed point can be assessed graphically by looking at the vector field. If the vector field points towards the FP, then the latter is stable, otherwise the FP is unstable if the vector field points away from it. The FP can be a saddle node if a trajectory is attracted towards the FP whereas the other trajectories are repelled from that FP. A simple example of 2D phase plane analysis can be seen in Figure A.3 where the parameters have been chosen to have a monostable system ($J_S = 0.02 = a$) in the associated−one group case (top left of Table A.1).

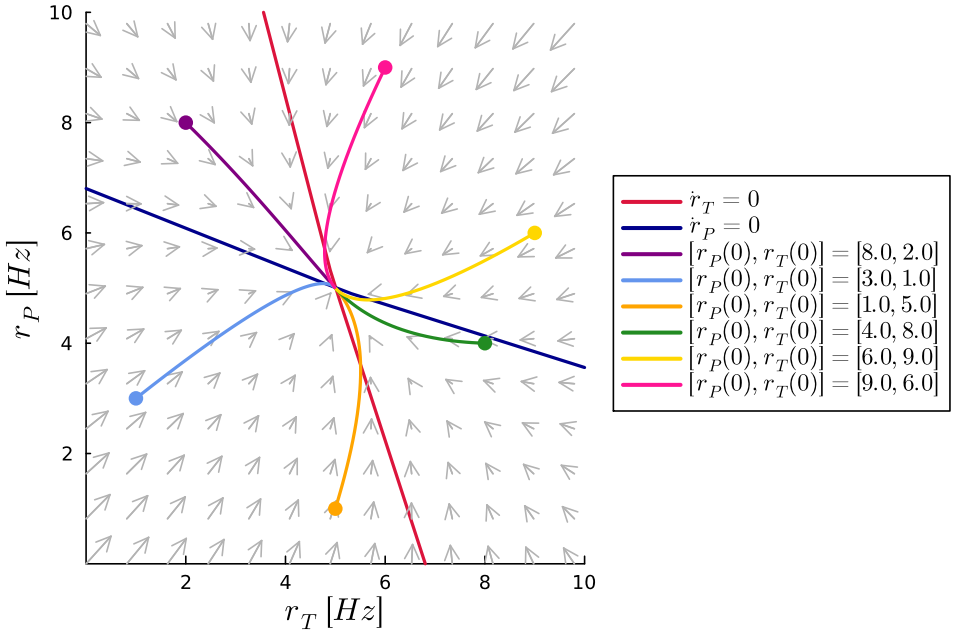


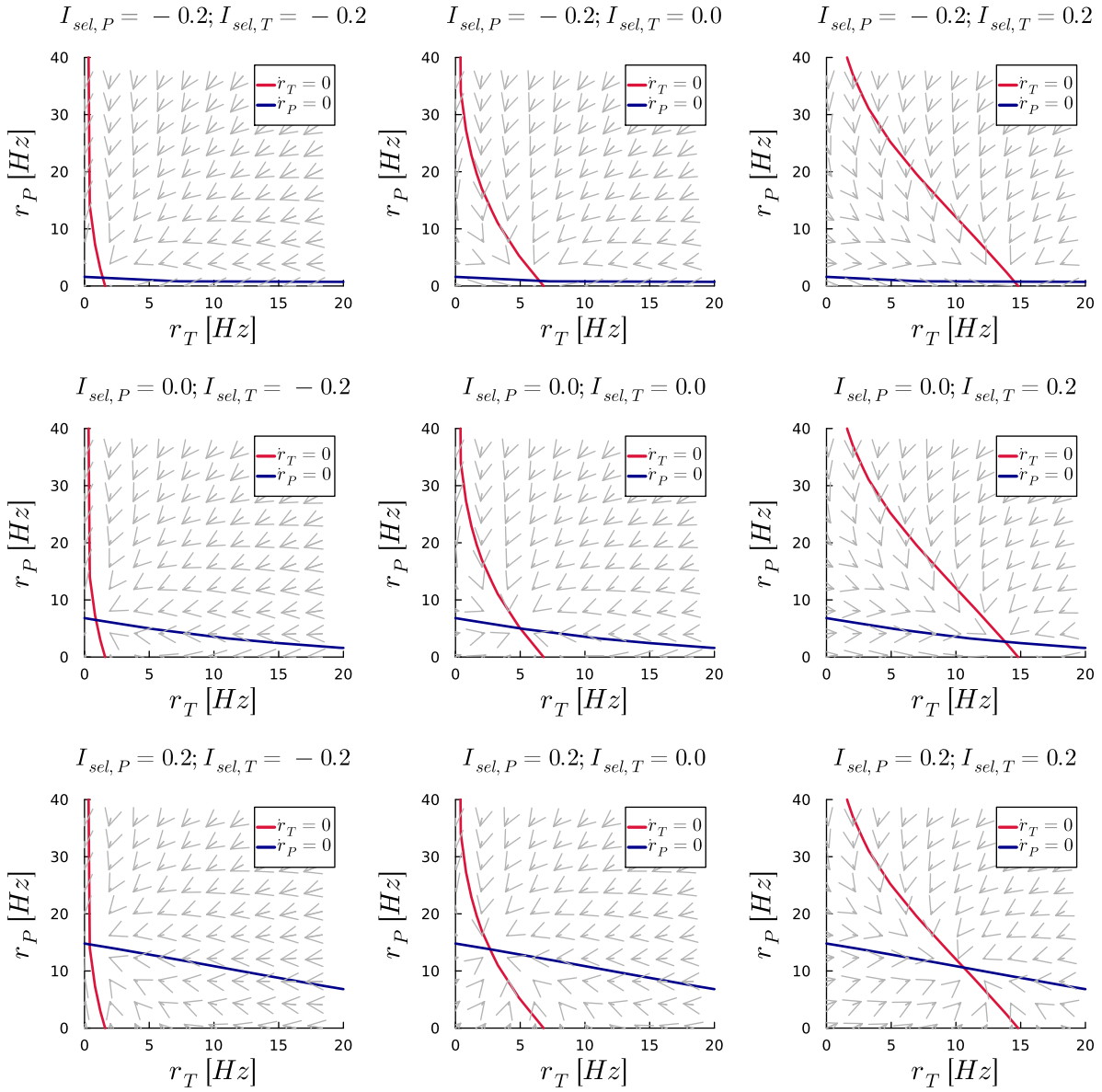
Figure A.3: Example of phase plane analysis for the 2D system (A.2) in a monostable regime. Prime and target words are associated and within the same group. Nullclines' and trajectories' color code is explicit in the legend. Vector field is represented by gray arrows.

From Figure A.3, it can be seen that trajectories, wherever they start (colored circles), follow the vector field and converge to the intersection of nullclines (dark blue and red curves).

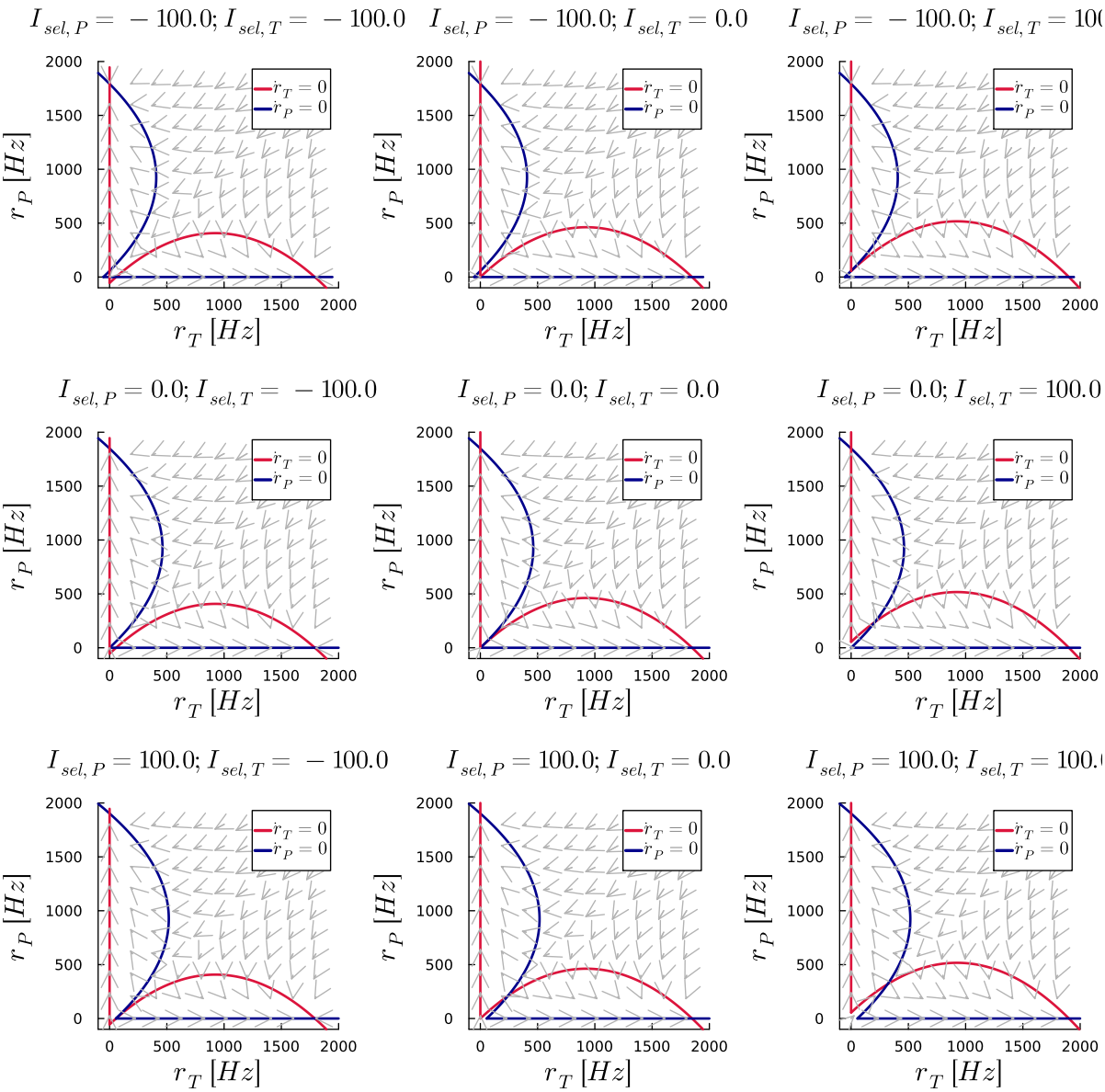
Since the prime and the target words are mostly encountered as belonging to the same group, different phase planes (A VS NA; $I = I_{ext} + I_{sel}$ varies) are shown hereafter.

A.2.1 Prime & Target associated

Figures A.4a and A.4b show the phase plane analysis for a monostable regime and a bistable regime (with W being the top left matrix of Table A.1). When J_S is chosen so that the system is in a monostable regime, the selective currents simply shift that stable equilibrium. It shifts up when $I_{sel,P}$ increases and it shifts to the right when $I_{sel,T}$ increases. When J_S is chosen so that the system is in a bistable regime, the system actually behaves as a "Winner-Takes-All" model: either the prime or the target word is activated. The equilibrium where both words are slightly activated is unstable (based on the vector field). The change in behavior when J_S varies suggests that a *bifurcation* occurs.



(a) Monostable



(b) Bistable

Figure A.4: Phase plane analysis for the Prime-Target network where the prime and the target are associated within a semantic group. Red curve is the T nullcline while the blue curve is the P nullcline. **(a)** Monostable regime ($J_S = 0.02$). **(b)** Bistable regime (Default $J_S = 3.65$).

A.2.2 Prime & Target non associated

The same behaviors as for the associated case can be observed when J_S and a vary, suggesting again that at least one bifurcation happens when varying J_S and a.

A.3 Conclusions on the 2D model

This bonus chapter introduced the Prime-Target network, that is the two-dimensional version of the Brunel and Lavigne's model. Although the analysis was not complete (bifurcation diagrams and responses to pulse-shaped stimuli were missing), this introduction still showed that all the concepts seen in the 1D model could be extended to the 2D model ... at the cost of longer and more complicated calculations, graphs, figures, ... In particular, the spontaneous activity for each node can still be chosen as the modeler desires but the degrees of freedom rapidly increase if assumptions are not made. When the spontaneous activity is the same for all nodes, sufficient conditions and ranges of values for appropriate parameters could be found in order to guarantee this stable background activity state.

Phase plane analysis allows one to graphically visualize the solutions' behavior, that is the trajectories, of the two-dimensional model. Phase plane analysis can actually be applied to any higher-order dimension. Important components of the phase plane are the nullclines where one variable does not evolve anymore. The intersections of nullclines are the fixed points (or equilibria). In addition, the nullclines partition the phase plane into different regions that display different evolving temporal behaviors. Phase plane analysis for the Prime-Target network, where the prime and the target belong to the same semantic group, suggests that a bifurcation is occurring when parameters J_S and a vary because the model goes from a monostable regime to a "Winner-Takes-All" bistable regime. The symmetry of the phase plane suggests that this bifurcation would be a three-dimensional pitchfork.

The next step would have been to investigate the semantic priming as such using pulse-shaped stimuli and see whether the target is effectively primed by the prime word, that is, the target would show prospective activity before the target word is presented. The effects of pulse amplitude, pulse duration and stimulus onset asynchrony could also have been explored.

Appendix B

Modeling background details

B.1 1D model computation details

Here below is the mathematical development used to find the analytical solution to the equation $\dot{x} = \cos x$.

The solution is found by separating the variables and then by integrating on both sides.

$$\begin{aligned}
\dot{x} = \frac{dx}{dt} &= \cos x \\
\leftrightarrow dt &= \frac{dx}{\cos x} \\
\leftrightarrow \int dt &= \int \frac{dx}{\cos x} \\
\leftrightarrow t + C &= \int \frac{dx}{\cos x} = \int \frac{\cos x}{(\cos x)^2} dx = \int \frac{\cos x}{1 - (\sin x)^2} dx
\end{aligned}$$

where C is an integration constant. By making a change of variable $u = \sin x$, one has

$$du = \cos x \; dx$$

and thus

$$\begin{aligned}
t + C &= \int \frac{1}{1 - u^2} \; du = \int \frac{2}{2(1-u)(1+u)} \; du = \int \frac{(1-u) + (1+u)}{2(1-u)(1+u)} \; du \\
\leftrightarrow t + C &= \frac{1}{2} \left(\int \frac{1}{1+u} \; du + \int \frac{1}{1-u} \; du \right) \\
\leftrightarrow t + C &= \frac{1}{2} \left(\ln |1+u| - \ln |1-u| \right) \\
\leftrightarrow t + C &= \frac{1}{2} \ln \frac{|1+\sin x|}{|1-\sin x|}
\end{aligned}$$

Assuming $x = x_0$ at $t = 0$, one can find the constant C as

$$C = \frac{1}{2} \ln \frac{|1+\sin x_0|}{|1-\sin x_0|}$$

The final solution is thus

$$\boxed{t = \frac{1}{2} \ln \left(\frac{|1 + \sin x|}{|1 - \sin x|} \cdot \frac{|1 - \sin x_0|}{|1 + \sin x_0|} \right)}$$

B.2 Taylor's expansion

The Taylor's expansion formula allows one to approximate a function $f(x)$ by a sum of function f and its derivatives, all evaluated at some point a in the neighborhood of x. This formula also gives an expression for the error associated to the approximation. Another advantage of Taylor's expansion is the possibility to improve the order of approximation of function f, that is, it gives a formula to approximate the function f by a relation more complicated than a simple line.

The formula here below is extracted from Delhez 2018-2019.

"If the real-valued function f is n times continuously differentiable on an interval $[a, x]$ (or $[x, a]$) and $n + 1$ times differentiable on the associated open interval $]a, x[$ (or $]x, a[$), then there exist at least one point $\xi \in]a, x[$ (or $]x, a[$) such that

$$f(x) = f(a) + \frac{(x-a)}{1!} f'(a) + \frac{(x-a)^2}{2!} f''(a) + \cdots + \frac{(x-a)^n}{n!} f^{(n)}(a) + \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(\xi) \tag{B.1}$$

"

The error associated to the approximation is thus given by

$$\epsilon = \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(\xi) \tag{B.2}$$

Appendix C

Transfer functions and spontaneous activity : Extra analyses

C.1 Understanding the transfer function of Brunel and Lavigne

Considering the transfer function from Brunel and Lavigne 2009 (Figure C.1), this small extra analysis attempts to understand the behavior of

$$\Phi_1(x) = \frac{1}{\tau_m\sqrt{\pi}}\left[\int_{-\infty}^{+\infty} \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\ dz\right]^{-1}$$

when varying x.

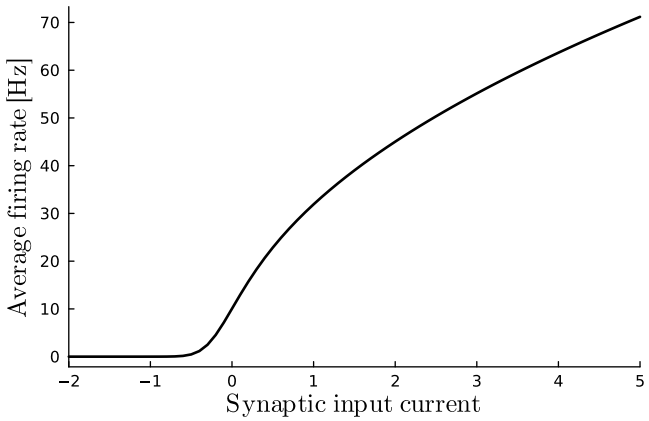
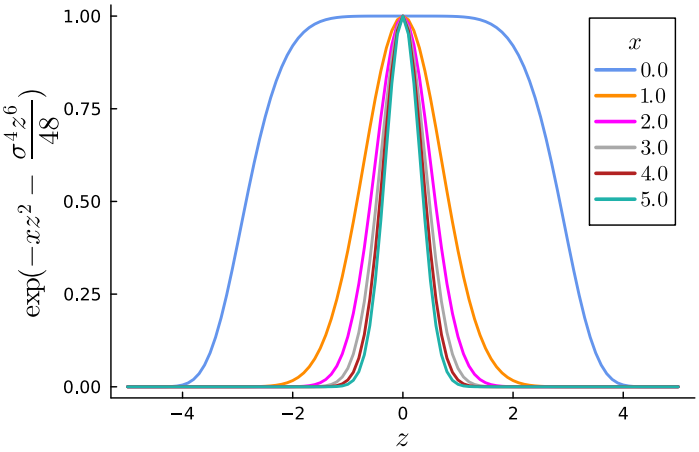


Figure C.1: $\Phi_1(x) = \frac{1}{\tau_m\sqrt{\pi}}\left[\int_{-\infty}^{+\infty} \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\ dz\right]^{-1}$. Transfer function from Brunel and Lavigne 2009.

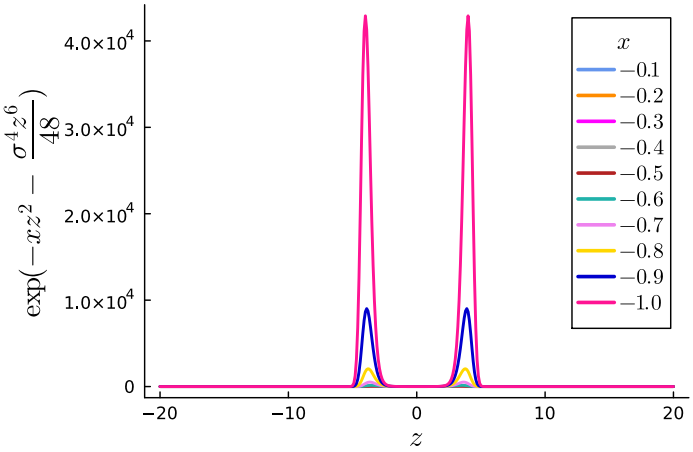
One should first notice that the integrand

$$\exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)$$

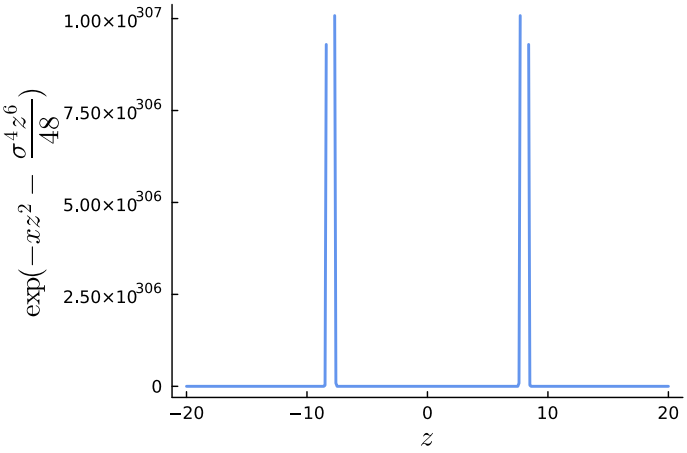
is symmetric with respect to the integrated variable z. In other words, the integrand is an even function (*i.e.* $f(z) = f(-z)\ \forall z$).



(a) $x \geq 0$



(b) $x < 0$



(c) $x = -16.5$

Figure C.2: Behavior of the integrand of the transfer function $\Phi_1(x) = \frac{1}{\tau_m\sqrt{\pi}}\left[\int_{-\infty}^{+\infty} \exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right)\, dz\right]^{-1}$ for different values of x.

When x is positive (Figure C.2a), the integrand is always positive but, above all, it is decaying. Thus, when $z = 0$, $\exp\left(-xz^2 - \frac{\sigma^4 z^6}{48}\right) = \exp 0 = 1$. As z increases/decreases, the integrand decays exponentially towards zero. Now, if one varies x, then when $x = 0$ (Figure C.2a light blue curve), the integrand reduces to $\exp\left(-\frac{\sigma^4 z^6}{48}\right)$ such that if $\sigma^4 z^6 < 48$, then the integrand has a value close to 1 whereas if $\sigma^4 z^6 > 48$, then it decays strongly. If $x > 0$, then the integrand monotonically and continuously decays towards zero and this decay is more pronounced as x increases. Thus, as x increases, the area under the curve decreases. As a result, the whole integral in the expression of $\Phi_1(x)$ decreases. However, since it is at the denominator of $\Phi_1(x)$, it implies that the value of $\Phi_1(x)$ increases as x is positive and increases.

When x is negative (Figure C.2b), the behavior is different. Since $x < 0$, the exponent of the integrand has a growing term $(-xz^2)$ and a decaying term $(-\frac{\sigma^4 z^6}{48})$. As a consequence, when z is small, the term $-xz^2$ is dominant over $-\frac{\sigma^4 z^6}{48}$ and the exponential is thus growing. On the other hand, when z is large, it is the opposite; the term $-\frac{\sigma^4 z^6}{48}$ dominates over $-xz^2$ and the exponential is decaying. The growth and the decay are even more pronounced as x becomes more and more negative. Thus, as x decreases in the negative values, the area under the curve increases. The whole integral in the expression of $\Phi_1(x)$ increases and the corresponding value $\Phi_1(x)$ converges towards zero rapidly.

An artificial trick must even be used when x becomes "too" negative. Indeed, when x becomes smaller than roughly -16 (Figure C.2c), the integrand is not representable numerically anymore because its maximum points are too high. However, since the value of $\Phi_1(x)$ theoretically exists and is equal to zero, $\Phi_1(x)$ can be *assigned* the value zero for such negative values of x.

C.2 Understanding the transfer function of Gjorgjieva *et al*

Considering the transfer function of Gjorgjieva et al. 2021a and Gjorgjieva et al. 2021b (Figure C.3), this small extra analysis attempts to understand the behavior of

$$\Phi_2(x) = \frac{1}{1 + \exp(-\alpha(x - \theta))} - \frac{1}{1 + \exp(\alpha\theta)}$$

when varying x, α and θ.

Similarly to the transfer function of Brunel and Lavigne 2009, the sigmoidal $f - I$ curve from Gjorgjieva et al. 2021a; Gjorgjieva et al. 2021b is a continuous and monotonically increasing function. Thus, as x increases, $\Phi_2(x)$ increases as well until saturation. For large x, $\Phi_2(x)$ becomes independent of the exact value of x and saturates (or remains constant) at the same value. The same phenomenon appears for largely negative x with the lower saturation.

Parameter α tunes the linear gain (or slope) of the transfer function (Figure C.4a). In other words, the larger α, the sharper the transition between the lower and the upper saturations. Thus, as α increases, the slope becomes more and more vertical and $\Phi_2(x)$

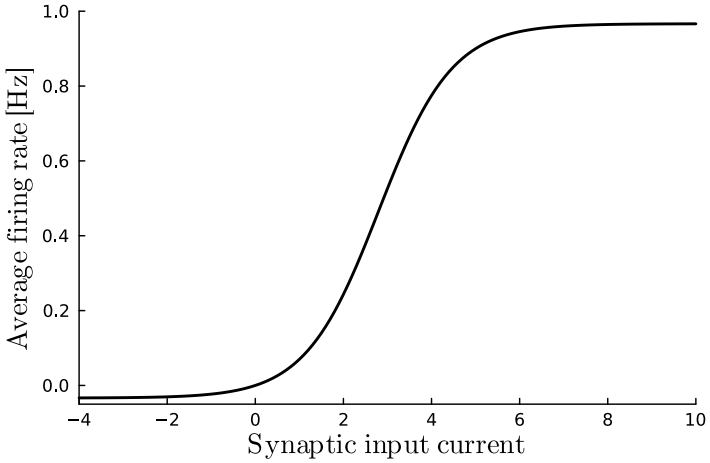
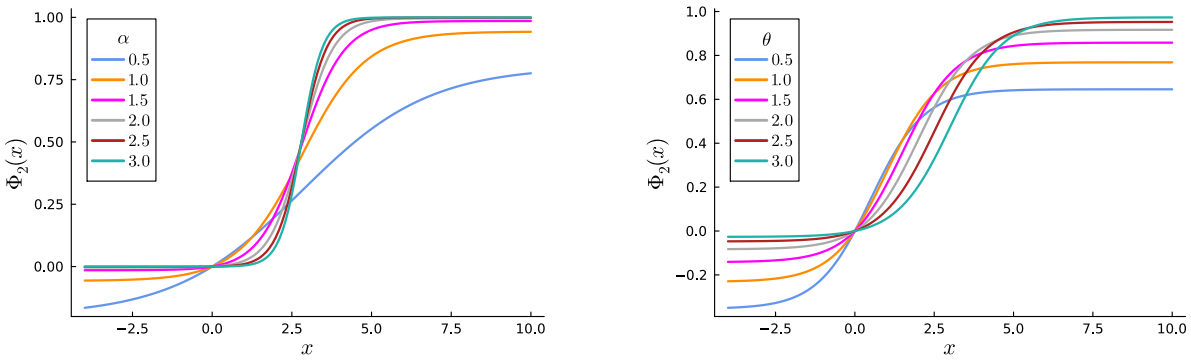


Figure C.3: $\Phi_2(x) = \frac{1}{1+\exp(-\alpha(x-\theta))} - \frac{1}{1+\exp(\alpha\theta)}$. Transfer function from Gjorgjieva et al. 2021a and Gjorgjieva et al. 2021b. The term $-\frac{1}{1+\exp(\alpha\theta)}$ allows one to get $\Phi_2(0) = 0$ for convenience.

takes a switch-like shape.

Parameter θ, in turn, tunes the threshold (or the input giving the midpoint) of the transfer function (Figure C.4b). In other words, the larger θ, the more the midpoint of the transfer function is shifted to the right. Because of the requirement that $\Phi_2(0) = 0$, $\Phi_2(x)$ is shifted up to the right as θ increases.

Parameters α and θ can be tuned independently according to the modeler/user's desires. In the current work, the same values as in Gjorgjieva et al. 2021a and Gjorgjieva et al. 2021b are used, *i.e.* $\alpha = 1.2$ and $\theta = 2.8$.



(a) Effect of α **(b)** Effect of θ

Figure C.4: Behavior of the transfer function $\Phi_2(x) = \frac{1}{1+\exp(-\alpha(x-\theta))} - \frac{1}{1+\exp(\alpha\theta)}$ as a function of parameters α and θ. **(a)** Parameter α tunes the slope (or gain) of the sigmoid. **(b)** Parameter θ tunes the input value at which $\Phi_2(x)$ is half its final value.

C.3 Spontaneous activity: supplementary figures

C.3.1 Graphical approach for estimating w_{max} in Method 1

In Method 1, the conditions on w and I_{ext} to have a stable spontaneous activity $r_T(t) = r_{spont}$ are

$$I_{ext} = \Phi_1^{-1}(r_{spont}) - w \cdot r_{spont} \ (= I_{ext,w=0})$$
$$w < \frac{1}{\Phi_1'(\Phi_1^{-1}(r_{spont}))}$$

with $\Phi_1'(x)$ the transfer function derivative and $\Phi_1^{-1}(x)$ the inverse transfer function of $\Phi_1(x)$.

A graphical approach to estimate the maximum w allowing a stable spontaneous activity is to plot the behavior of $\frac{d\dot{r}_T}{dr_T}\Big|_{r_T=r_{spont}}$ as a function of $w = J_1 - J_I$ and I_{ext} (r_{spont} is fixed and I_{sel} is assumed to be zero). For $r_{spont} = 5$ [Hz] (default value from Brunel and Lavigne 2009), this behavior looks like Figure C.5.

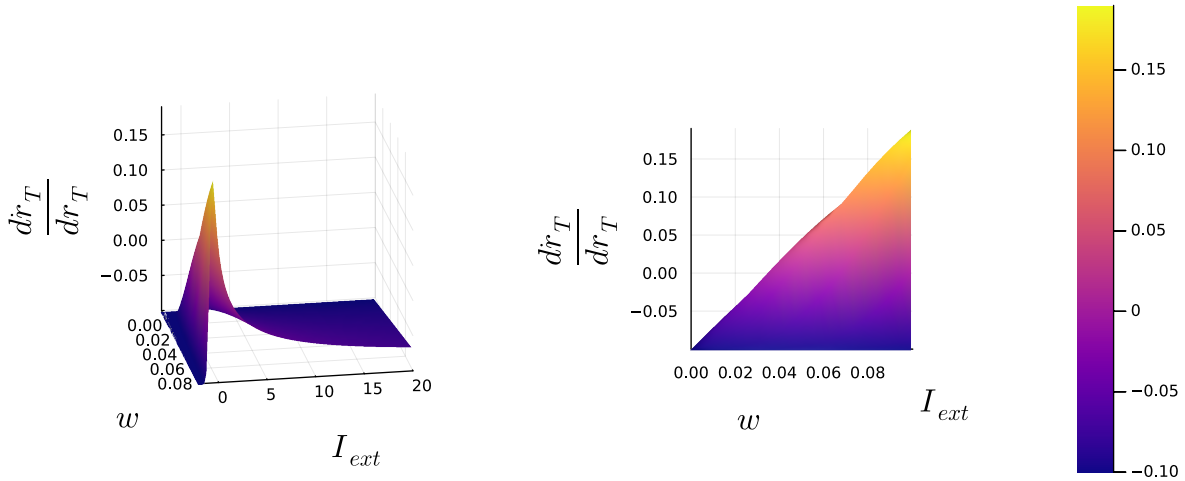
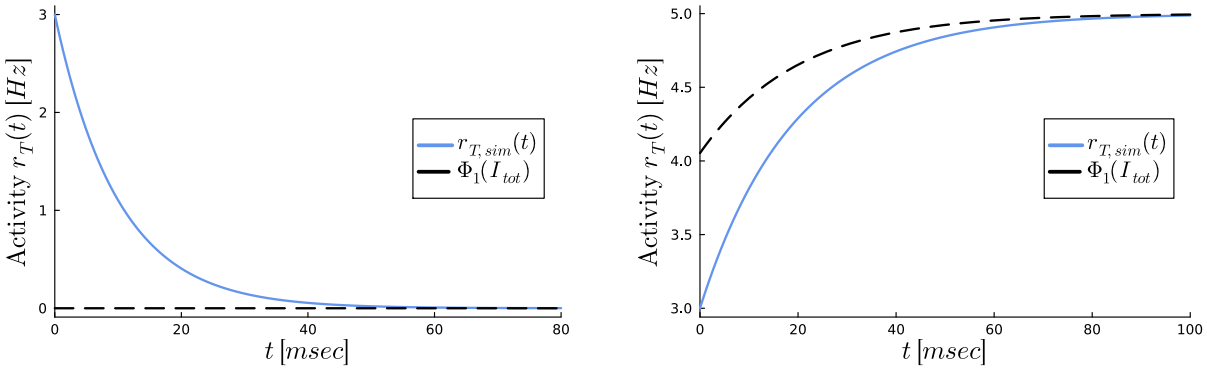


Figure C.5: Behavior of $\frac{d\dot{r}_T}{dr_T}\Big|_{r_T=r_{spont}}$ as a function of recurrent connection weight w and bias current I_{ext}. The spontaneous firing rate is set to default value from B&L ($r_{spont} = 5$ [Hz]) and I_{sel} is assumed to be zero.

When looking at the projection onto the $(\frac{d\dot{r}_T}{dr_T}, w)$ plane (Figure C.5 right), one could roughly estimate that w should be smaller than ≈ 0.035 so that $\frac{d\dot{r}_T}{dr_T} < 0$, giving the state $r_T = r_{spont}$ stable.

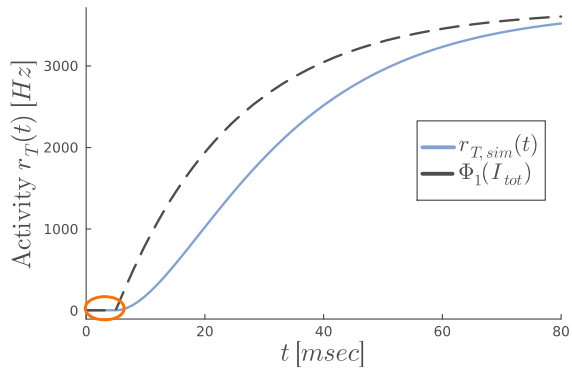
C.3.2 Examples of temporal dynamics for assessing spontaneous activity in Method 1

For a stable background activity $r_{spont} = 5$ [Hz], the recurrent connection weight should satisfy $w \lessapprox 0.04$ and the bias current should be chosen such that $I_{ext} = \Phi_1^{-1}(5) - w \cdot 5$.

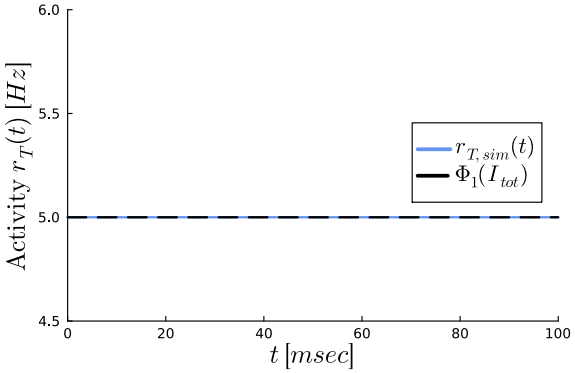


(a) $w = 3.65$; $r_T(0) = 3$

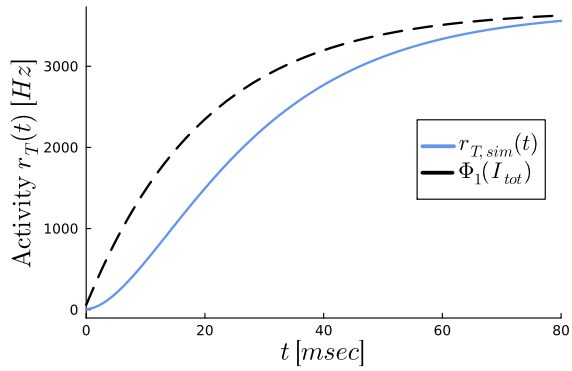
(b) $w = 0.02$; $r_T(0) = 3$



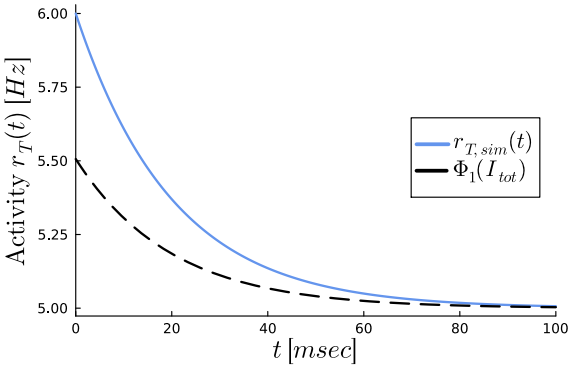
(c) $w = 3.65$; $r_T(0) = r_{spont}$



(d) $w = 0.02$; $r_T(0) = r_{spont}$



(e) $w = 3.65$; $r_T(0) = 6$



(f) $w = 0.02$; $r_T(0) = 6$

Figure C.6: Time evolution of $r_T(t)$ (blue curve) for the 1D model $\dot{r}_T = -r_T + \Phi_1(w \cdot r_T + I_{ext})$ with $w = 3.65$ (left) and $w = 0.02$ (right), with different initial conditions $r_T(0)$ (3: top; 5: middle; 6:bottom), and with I_{ext} set to have $r_{spont} = 5$ [Hz]. Black dashed curve shows the value that $r_T(t)$ tracks.

Default values of J_1 and J_I (*i.e.* $w = 3.65$) violate the condition on w. Therefore, "$r_{spont} = 5$ [Hz]" is an unstable equilibrium as can be seen in Figures C.6a, C.6c and C.6e. On the other hand, if $w = 0.02$, then $r_{spont} = 5$ [Hz] is indeed a stable equilibrium (Figures C.6b, C.6d and C.6f).

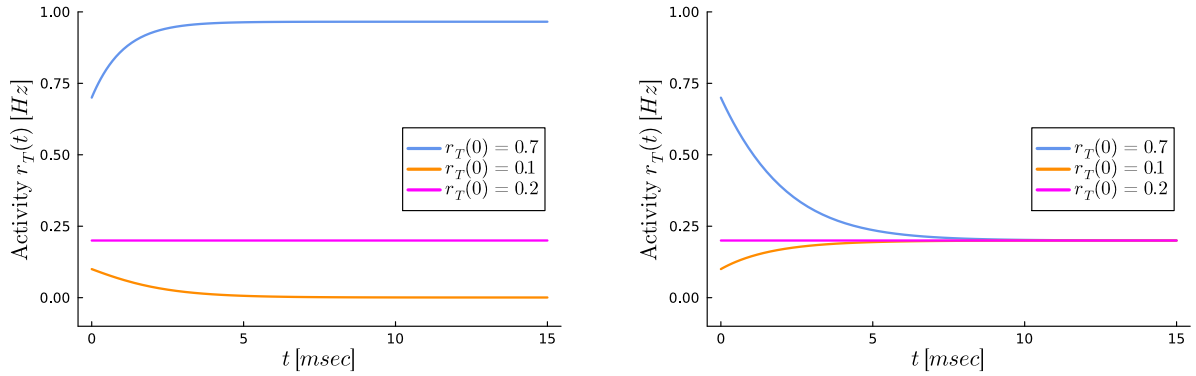
When w violates its condition, any initial condition (Figures C.6a and C.6e) will make r_T converge to a steady state other than $r_{spont} = 5$ [Hz]. When $r_T(0) = r_{spont}$ exactly (Figure C.6c), r_T remains in that equilibrium for a while before numerical error accumulation becomes too big and makes r_T converge to another stable steady state. Thus, when w violates its condition, $r_T = r_{spont}$ is an unstable FP.

Similarly, when w meets its condition, $r_T = r_{spont}$ is a stable FP and any initial condition $r_T(0)$ will make r_T converge to r_{spont}, assuming that $I_{sel} = 0$ of course (Figure C.6 right panels).

C.3.3 Examples of temporal dynamics for assessing spontaneous activity in Method 2

For example, if $R = 0.2$, then w should satisfy $w \lessapprox 4.66$ and the bias current should be chosen such that $I_{ext} = \Phi_2^{-1}(0.2) - w \cdot 0.2$.

Similarly to Method 1, default value of w (*i.e.* $w = 9$) violates the condition on w. Therefore, "$R = 0.2$ [Hz]" is an unstable equilibrium as can be seen in Figure C.7a. On the other hand, if $w = 2$, then $R = 0.2$ [Hz] is indeed a stable equilibrium (Figure C.7b).



(a) $w = 9$ **(b)** $w = 2$

Figure C.7: Time evolution of $r_T(t)$ for the 1D model $\dot{r}_T = -r_T + \Phi_2(w \cdot r_T + I_{bias})$ with **(a)** $w = 9$ and **(b)** $w = 2$, with different initial conditions (see legend in graphs). The bias current I_{bias} is set to have a spontaneous activity $R = 0.2$ [Hz].

The methodology and the results are thus similar for both methods.

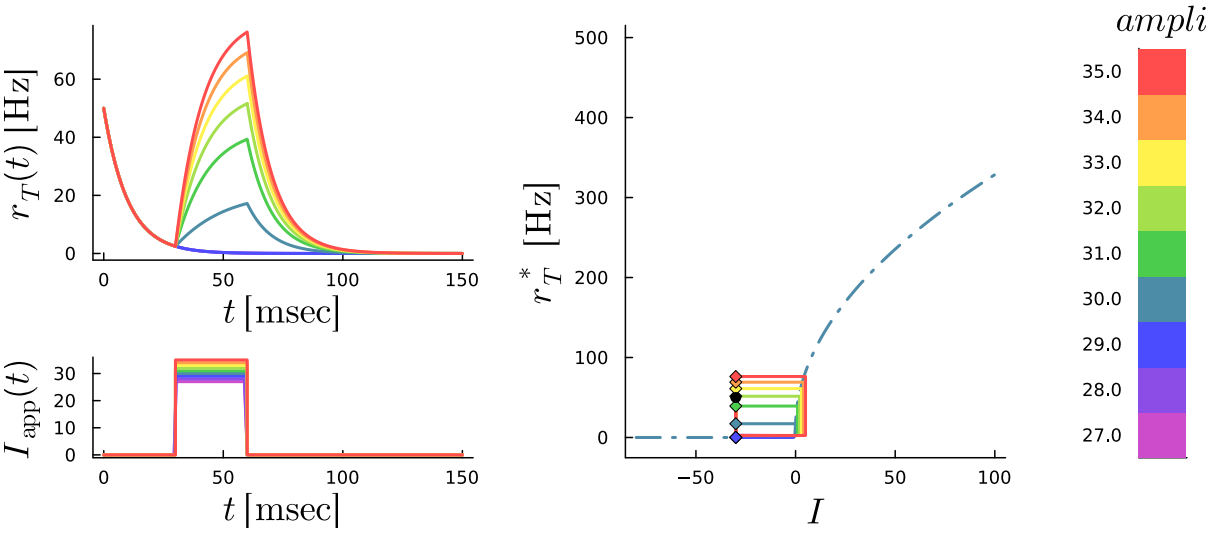
Appendix D

Pulse-shaped stimulus and Method 1

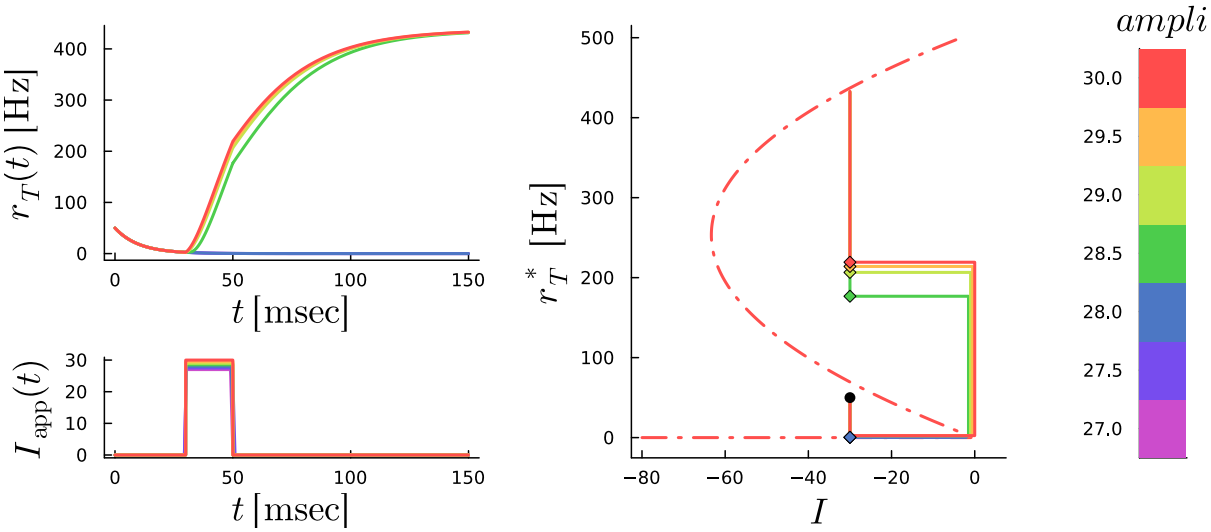
Similarly to Method 2, the effects of pulse amplitude and pulse duration can be investigated for Method 1. The same results will be found:

- For a bistable system, the amplitude of the pulse must satisfy $I_{app} > I_{thresh} = I_{SN,1} - I_{bias}$ to potentially reach the high steady state associated to I_{bias} (Figure D.1).
- The unstable fixed point associated to the bias current acts as a threshold on the minimum pulse duration in order to reach the high stable state (Figure D.2).
- Pulse duration and pulse amplitude determine together whether the bistable system reaches the high stable attractor and displays persistent activity (Figure D.3).
- For a monostable system, the final steady state is the same as before the pulse became ON, whatever the duration and/or the amplitude. Pulse amplitude and pulse duration only affect whether the monostable system reaches the new steady state associated to $I = I_{app} + I_{bias}$ when the pulse is ON, and how long the system remains in that state before converging back to its original state.

D.1 Effect of pulse amplitude



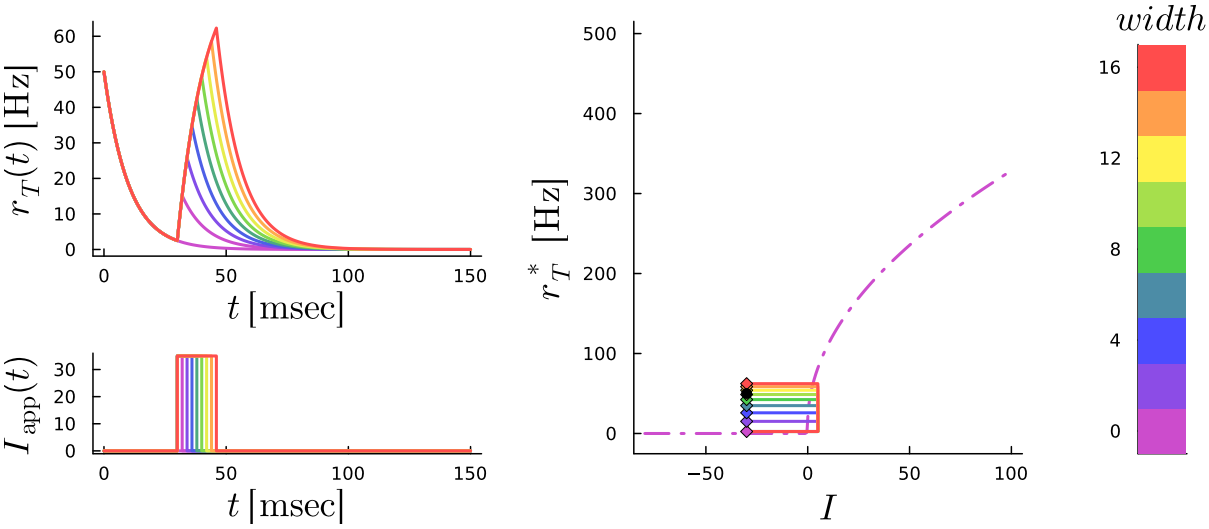
(a) Monostable ($w = 0.02$)



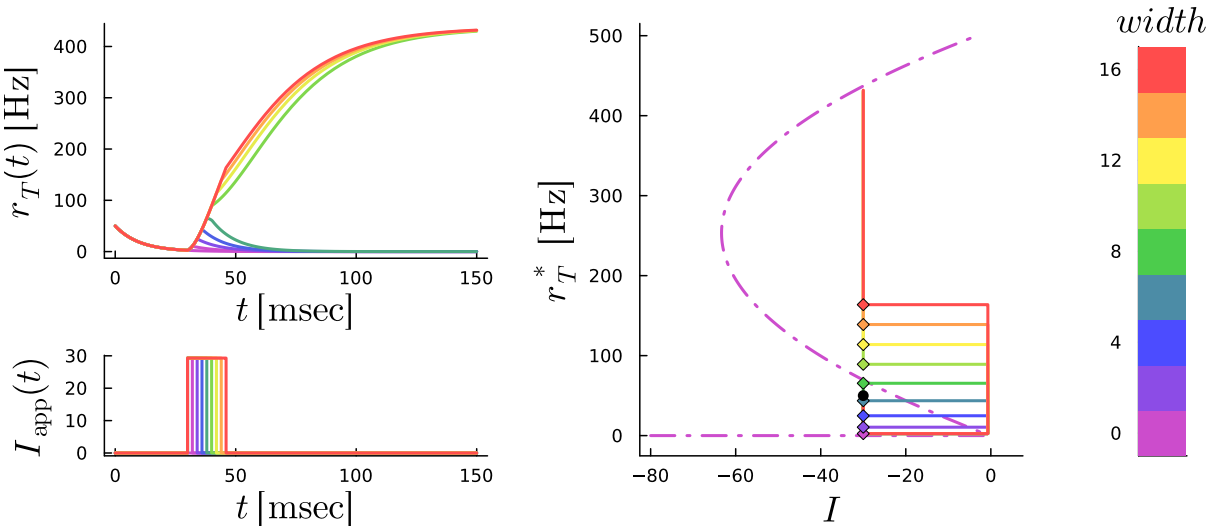
(b) Bistable ($w = 0.5$)

Figure D.1: Effect of pulse amplitude on the model in **(a)** a monostable regime or **b** a bistable regime with Method 1. The layout is the same as in Chapter 7 with Method 2. Markers have also the same meaning. The dash-dotted curve is the bifurcation diagram. Black dot in the bifurcation diagram is the initial condition for all trajectories.

D.2 Effect of pulse duration



(a) Monostable ($w = 0.02$)



(b) Bistable ($w = 0.5$)

Figure D.2: Effect of pulse duration on the model in **(a)** a monostable regime or (**b**) a bistable regime with Method 1. The layout is the same as in Chapter 7 with Method 2. Markers have also the same meaning. The dash-dotted curve is the bifurcation diagram. Black dot in the bifurcation diagram is the initial condition for all trajectories.

D.3 Combined effects

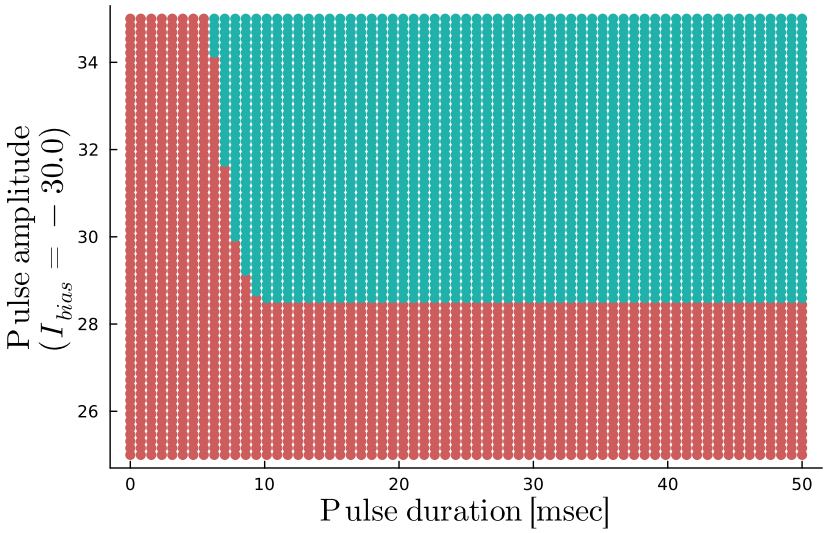


Figure D.3: Combined effects of pulse amplitude and pulse duration for Method 1 when the system is in a bistable regime. Both parameters determine together whether the system jumps to the high steady state (green area) or not (red area).

Bibliography

Abbott, L. F. (1991). "Firing-rate models for neural populations". In: *Neural networks: From biology to high-energy physics*, pp. 179–196.

Brunel, Nicolas (2021). *Dynamic Networks Intro, Week 2, Day 4*. URL: https://compneuro.neuromatch.io/tutorials/W2D4_DynamicNetworks/student/W2D4_Intro.html. (last accessed: 10.08.2023).

Brunel, Nicolas and Frédéric Lavigne (Dec. 2009). "Semantic Priming in a Cortical Network Model". In: *Journal of Cognitive Neuroscience* 21.12, pp. 2300–2319. ISSN: 0898-929X. DOI: 10.1162/jocn.2008.21156. eprint: https://direct.mit.edu/jocn/article-pdf/21/12/2300/1937667/jocn.2008.21156.pdf. URL: https://doi.org/10.1162/jocn.2008.21156.

Danisch, Simon and Julius Krumbiegel (2021). "Makie.jl: Flexible high-performance data visualization for Julia". In: *Journal of Open Source Software* 6.65, p. 3349. DOI: 10.21105/joss.03349. URL: https://doi.org/10.21105/joss.03349.

Dayan, Peter and Laurence F. Abbott (2001a). "Theoretical neuroscience : computational and mathematical modeling of neural systems". In: Computational neuroscience. The Massachusetts Institute of Technology (MIT) Press. Chap. 1, pp. 3–45. ISBN: 0-262-04199-5.

— (2001b). "Theoretical neuroscience : computational and mathematical modeling of neural systems". In: Computational neuroscience. The Massachusetts Institute of Technology (MIT) Press. Chap. 7, pp. 229–279. ISBN: 0-262-04199-5.

Delhez, Eric J.M. (2018-2019). "Analyse mathématique 1, Tome 1". In: Centrale des cours de la FSA ASBL. Chap. 1, p. 67.

Drion, Guillaume (2021-2022). *Advanced topics in systems and control*. Lecture notes.

— (2019-2020). *Introduction to signals and systems*. Lecture notes.

Ermentrout, G. Bard and David H. Terman (2010). "Mathematical foundations of neuroscience". In: Springer New York. Chap. 11. ISBN: 978-0-387-87707-5. DOI: https://doi.org/10.1007/978-0-387-87708-2.

Franci, Alessio (2023a). *Brain-Inpired Computing Class 2 : Flexible representations*. Lecture notes.

Franci, Alessio (2023b). *Brain-Inpired Computing Class 5 : From multi-option (event-based) representations to network (event-based) representations.* Lecture notes.

Gerstner, Wulfram et al. (2014). *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press. DOI: 10.1017/CBO9781107447615. Chapter 3, pp 58–80; Chapter 7, pp 168–201; Chapter 12, pp 291–324; Chapter 15, pp 395-416; Chapter 17, pp 442-466.

Gjorgjieva, Julijana et al. (2021a). *Tutorial 1: Neural Rate Models, Week 2, Day 4: Dynamic Networks*. URL: https://compneuro.neuromatch.io/tutorials/W2D4_DynamicNetworks/student/W2D4_Tutorial1.html. (last accessed: 10.08.2023).

— (2021b). *Tutorial 2: Wilson-Cowan Models, Week 2, Day 4: Dynamic Networks*. URL: https://compneuro.neuromatch.io/tutorials/W2D4_DynamicNetworks/student/W2D4_Tutorial2.html. (last accessed: 10.08.2023).

Heyman, Tom, Anke Bruninx, et al. (2018). "The (un)reliability of item-level semantic priming effects". In: *Behavior Research Methods* 50, pp. 2173–2183. DOI: https://doi.org/10.3758/s13428-018-1040-9.

Heyman, Tom, Simon De Deyne, and Gert Storms (2013). "Using the letter decision task to examine semantic priming". In: *Cooperative Minds: Social Interaction and Group Dynamics. Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Cognitive Science Society; Austin, TX, pp. 2542–2547.

Hutchison, Keith A (2003). "Is semantic priming due to association strength or feature overlap? A microanalytic review". In: *Psychonomic bulletin & review* 10, pp. 785–813. DOI: https://doi.org/10.3758/BF03196544.

Kumar, Abhilasha A (2021). "Semantic memory: A review of methods, models, and current challenges". In: *Psychonomic Bulletin & Review* 28, pp. 40–80. DOI: https://doi.org/10.3758/s13423-020-01792-x.

Lavigne, Frédéric, Laurent Dumercy, and Nelly Darmon (June 2011). "Determinants of Multiple Semantic Priming: A Meta-analysis and Spike Frequency Adaptive Model of a Cortical Network". In: *Journal of Cognitive Neuroscience* 23.6, pp. 1447–1474. ISSN: 0898-929X. DOI: 10.1162/jocn.2010.21504. eprint: https://direct.mit.edu/jocn/article-pdf/23/6/1447/1941639/jocn.2010.21504.pdf. URL: https://doi.org/10.1162/jocn.2010.21504.

Marting, Howard and Alan Ng (n.d.). *13.1: The motion of a spring-mass system*. URL: https://phys.libretexts.org/Bookshelves/University_Physics/Book%3A_Introductory_Physics_-_Building_Models_to_Describe_Our_World_(Martin_Neary_Rinaldo_and_Woodman)/13%3A_Simple_Harmonic_Motion/13.01%3A_The_motion_of_a_spring-mass_system. (accessed: 05.06.2023).

MathWorks, The MathWorks Incorporation (2023). *MATLAB version: 9.9.0.1495850 (R2020b) Update 1*. Natick, Massachusetts, United States. URL: https://www.mathworks.com.

McNamara, Timothy P (2005). "Semantic priming: Perspectives from memory and word recognition". In: Psychology Press. Chap. 1. ISBN: 9780415651677.

Palani, S. (2022a). "Signals and Systems". In: Springer Cham. Chap. 1, pp. 1–197. ISBN: 978-3-030-75744-1. DOI: https://doi.org/10.1007/978-3-030-75742-7.

— (2022b). "Signals and Systems". In: Springer Cham. Chap. 2, pp. 197–271. ISBN: 978-3-030-75744-1. DOI: https://doi.org/10.1007/978-3-030-75742-7.

Schacter, Daniel L. (2000). "Memory: memory systems". In: *Encyclopedia of psychology*. Vol. 5. Oxford University Press. Chap. 78, pp. 169–172. DOI: http://dx.doi.org/10.1037/10520-081.

Sperber, Richard D et al. (1979). "Semantic priming effects on picture and word processing". In: *Memory & Cognition* 7.5, pp. 339–345. DOI: https://doi.org/10.3758/BF03196937.

Strogatz, Steven Henry (1994). *Nonlinear Dynamics and Chaos: with applications to physics, biology, chemistry and engineering (studies in nonlinearity)*. Perseus Books Publishing. ISBN: 978-0-201-54344-5.

Thompson, Richard E. (2000). "Memory: brain systems". In: *Encyclopedia of psychology*. Vol. 5. Oxford University Press. Chap. 80, pp. 175–178. DOI: http://dx.doi.org/10.1037/10520-083.

Tulving, E. (2000). "Memory: an overview". In: *Encyclopedia of psychology*. Vol. 5. Oxford University Press. Chap. 75, pp. 161–162. DOI: http://dx.doi.org/10.1037/10520-078.

Vandewalle, Gilles (2020-2021). *Introduction to cognitive neuroscience*. Lecture notes.

Wilson, Hugh R. and Jack D. Cowan (1972). "Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons". In: *Biophysical Journal* 12.1, pp. 1–24. ISSN: 0006-3495. DOI: https://doi.org/10.1016/S0006-3495(72)86068-5.