Neural Coding in the Ascending Somatosensory Pathway: 
A Metrical Information Theory Approach

Presented on December, 3rd 2010 to the jury composed of:

M. Angelo ARLEO . . . . . . . . . . . . . . . . . . . . . . . . . Advisor
M. Stefano PANZERI . . . . . . . . . . . . . . . . . . . . . . . . . Reviewer
M. Jonathan D. VICTOR . . . . . . . . . . . . . . . . . . . . . . . . Reviewer
M. Vincent HAYWARD . . . . . . . . . . . . . . . . . . . . . . . . . Examiner
M. Roland S. JOHANSSON . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Examiner
M. Simon THORPE . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Examiner
You are wrong, my dear, the boat is going at a high pace. But the Zuyderzee is a dead sea, or almost. With its flat borders, lost in the fog, we do not know where it begins, where it ends. So, we go without any mark, we cannot estimate our speed. We move on, but nothing changes. This is not navigation, this is dream.

Albert Camus

The whole mind strives for a single goal: move the boat forward.

Eric Tabarly
UNIVERSITÉ PIERRE ET MARIE CURIE
Laboratoire de Neurobiologie des Processus Adaptatifs

THÈSE DE DOCTORAT DE
l’UNIVERSITÉ PIERRE ET MARIE CURIE (PARIS VI)

Spécialité : Neuroscience

présentée par

Romain BRASSELET

pour obtenir le titre de
Docteur de l’Université Pierre et Marie Curie (Paris VI)

Codage Neuronal dans le Système Somatosensoriel Ascendant :
Une Approche de Théorie de l’Information Métrique

Soutenue le 3 Dcembre 2010 devant le jury composé de :

M. Angelo ARLEO ........... Directeur
M. Stefano PANZERI .......... Rapporteur
M. Jonathan D. VICTOR .... Rapporteur
M. Vincent HAYWARD ...... Examinateur
M. Roland S. JOHANSSON . Examinateur
M. Simon THORPE ......... Examinateur
Vous vous trompez, cher, le bateau file à bonne allure. Mais le Zuyderzee est une mer morte, ou presque. Avec ses bords plats, perdus dans la brume, on ne sait où elle commence, où elle finit. Alors, nous marchons sans aucun repère, nous ne pouvons évaluer notre vitesse. Nous avançons, et rien ne change. Ce n’est pas de la navigation, mais du rêve.

Albert Camus

L’esprit tout entier est tendu vers un seul but : faire avancer le bateau.

Eric Tabarly
As opposed to Albert Jacquard, I never considered science a passion. The latter word derives from the Latin *patior* which suggests a state of passivity, something that one cannot fight. On the contrary, my approach to science is active, it is a commitment. And I just do not see what else I could do. The reason is the following:

We are often faced with people neverendlessly arguing about some facts in politics, philosophy, or about how one should live his life. Their argument can usually be reduced to the fact that their conceptions of the world do not match at the very beginning. Whether one should act directly for the sake of the whole humanity or one should in the first place act for its own good to maximize the satisfaction of the whole. Whether our goal is to maximize the happiness right now or to develop technologies that will help us doing so. During these three years in Paris, I got involved in politics. I thus had the chance to study a bit Adam Smith or Karl Marx writings. It may be argued that the difference in their views can be traced back to a few underlying assumptions about the so-called *human nature* or about the emergence of macroscopic organizations from interactions between microscopic individuals. This highlights the necessity of understanding the law of the Universe and determine what we should do. After all, what is the goal we pursue? Does maximizing happiness make sense? Shouldn’t we instead minimize our footprint on the environment, as it is often argued these last few years? Do we have to let humanity follows it natural path or should we try to preserve a balance between species? Do we really have the choice? Shouldn’t we care instead about something else? We live in days of ecological concerns due to global warming or species extinction. Many say we should try to *save the planet*. But should we really care about our impact on environment? Do we have to depart from our normal way of living? Do these questions even make sense?

In the end, every conception of the world and of life entails some degree of arbitrariness. Science is the only thing that can reduce this arbitrariness by understanding what is a fundamental law of nature (Schrödinger 1922) and what is just an epiphenomenon, that should not be taken as a starting concept. This is, in the end, what allows us to understand what we should do as individuals and as a population. What are the aims we should follow and how to reach them.

That is why I started studying theoretical physics and mathematics, to understand these laws of nature, that can help us get free from this toxic arbitrariness. I switched to neuroscience 3 years ago in order to understand also how the ability to think, to understand the world, arises. To understand the way we can think about the physical world. How does the organization of the universe affect the way we think. Can we even understand it? to quote Richard Dawkins, is the universe even queerer than we can suppose? (Dawkins 2005) Or even as John Barrow writes: “A universe simple enough to be understood is too simple to produce a mind capable of understanding it” (Barrow). These are open questions that could further the knowledge in science in general if solved at least partially.
I thus started my PhD in neuroscience after a full master in theoretical physics. This probably gives this work a bit of a particular flavor. A major drive in my work was to try to find basic definitions so as to make neuroscience a physical science. I thus try to define rigorously objects in theoretical neuroscience, so as to have a framework just like we have in physics. A whole set of elementary objects and their interactions. I particularly focused on the geometry of the space where signals live. This is to me a solid starting point to start understanding the dynamics of the system.

In neural coding, a spike can be seen as the basic object. However, the meaningful objects are spike trains. This is why my work focused on trying to understand the space in which spike trains live. In the first place, it was crucial to understand the topology of this space. When this is done, one can start clarifying its geometry. This is still an open problem. But it may be the last step before being able to define “interactions” on this space. Interaction refer to neurons. They are the locus of the interactions between spike trains.

I once said half-jokingly that my dream would be to do gauge theory with spike trains. Crazy though it may seem, I think this is a fair goal.

*This is not navigation, this is dream.*
Acknowledgements

My first thanks go to Angelo who offered me the opportunity to do a PhD in neuroscience although I had no experience at all in this field. I owe him a lot on this one. For these three years, he made all the efforts he could so that I could focus only on science without having to be concerned with extra-scientific problems.

I am mostly indebted to Roland Johansson, who provided us with a truly amazing set of microneurography data, that has been the basis for all the work done during my PhD. I also thank Henrik Jorntell for the data of the recordings he made in the cuneate nucleus. The integration of his observations in our model was a significant improvement.

I also thank my team-mates, Denis, Luca, François, Manu, JB, Xavier, Éléonore. I really encourage Jérémie, the last arrived in the team, to keep on being curious about the general laws governing physics and neuroscience and to keep his critical thinking. I am sure he will do a great job of pursuing similar lines of work (among others) at ANC.

I would also like to thank all the scientists who welcomed me in their labs during these three years, starting with Roland Johansson and his collaborators who made this visit in Umea highly profitable: Göran Westling and Michael Dimitriou. In Granada, Eduardo Ros and his team, Richard, Jesus and Niceto. I thank Guillaume Hennequin for his kind invitation to present my work at EPFL (Lausanne) at a group meeting.

I thank Jonathan Victor for his visit and highly relevant remarks on my work. Christoph Kayser and Stefano Panzeri for giving me the opportunity to start a new project with them in Tübingen next year.

I thank Laurent Bonnasse-Gahot and Jean-Pierre Nadal for interesting discussions about the definition of metrical information and for giving inputs about the perceptual magnet effect.

I also feel indebted to all the young scientists I have met in conferences and with whom I discussed late at night around a bottle of wine: Thomas Kreuz, Daniel Chicharro, Richard Naud, Henning Sprekeler, Loic Mathey, Hermann Cuntz, Rui Costa, Cian O’Donnell, Remus Osan, David Higgins, Ioannis Taxidis, Helmut Schmidt, Sébastien Crouzet, Geoffrey Herman, Jean-Philippe Michel and others.

Merci mes mentors. Enfin, tout scientifique matérialiste forcé a besoin de quelques gouttes d’humanité pour rendre la vie moins aride : merci à mes amis.

Merci Laetitia, tu sais pourquoi.
Abbreviations

$S$ : set of stimulations
$R$ : set of responses
$d(x, y)$ : distance between two events $x$ and $y$
$\phi(x, y)$ : similarity measure between two events $x$ and $y$
$H(X)$ : Shannon entropy of a set $X$
$I(X; Y), MI$ : Shannon Mutual Information between two sets $X$ and $Y$
$H^*(X)$ : Galway entropy of a set $X$
$I^*(X; Y)$ : Galway Mutual Information between two sets $X$ and $Y$
STDP : Spike-Timing Dependent Plasticity
PSTH : Peristimulus time histogram
MR : Mechanoreceptor
CN : Cuneate Nucleus
AP : Action Potential
SRM : Spike Response Model

Glossary

Afferent signals : incoming signals.
Efferent signals : outcoming signals.
Spike : action potential.
Poisson process : a continuous-time stochastic point process, where the probability of occurrence of an event is independent of the time since the last event.
Fano factor : ratio of the variance over the mean of a stochastic process.
Firing rate : number of spikes emitted in a given time window.
Figure 0.1.: Location of this PhD on a map of Ireland.
## Contents

Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Introduction</strong></td>
<td>23</td>
</tr>
<tr>
<td>1.1. Scientific context</td>
<td>24</td>
</tr>
<tr>
<td>1.2. Statement of the problem</td>
<td>27</td>
</tr>
<tr>
<td><strong>I. Cracking the code</strong></td>
<td>31</td>
</tr>
<tr>
<td>2. Neural Coding</td>
<td>33</td>
</tr>
<tr>
<td>2.1. Description of spike trains</td>
<td>35</td>
</tr>
<tr>
<td>2.2. Rate coding</td>
<td>36</td>
</tr>
<tr>
<td>2.3. Temporal coding</td>
<td>40</td>
</tr>
<tr>
<td>2.4. Spike time coding</td>
<td>43</td>
</tr>
<tr>
<td>2.4.1. First-spike latency</td>
<td>45</td>
</tr>
<tr>
<td>2.4.2. Rank-Order coding</td>
<td>47</td>
</tr>
<tr>
<td>2.4.3. Parameters affecting reliability</td>
<td>48</td>
</tr>
<tr>
<td>2.4.4. Measures of reliability</td>
<td>48</td>
</tr>
<tr>
<td>3. Structure of the Spike Trains Space</td>
<td>51</td>
</tr>
<tr>
<td>3.1. Topology of the space of spike trains</td>
<td>51</td>
</tr>
<tr>
<td>3.1.1. Embedding in a higher-dimensional manifold or binning</td>
<td>52</td>
</tr>
<tr>
<td>3.1.2. Processing of the spike trains by synapses</td>
<td>54</td>
</tr>
<tr>
<td>3.2. Geometry of the space of spike trains</td>
<td>54</td>
</tr>
<tr>
<td>3.2.1. Victor-Purpura distance (1996)</td>
<td>54</td>
</tr>
<tr>
<td>3.2.2. Van Rossum distance (2001)</td>
<td>57</td>
</tr>
<tr>
<td>3.2.3. Houghton distance (2009)</td>
<td>59</td>
</tr>
<tr>
<td>3.2.4. Schreiber correlation (2003)</td>
<td>60</td>
</tr>
<tr>
<td>3.2.5. Quian Quiroga’s Event Synchronization (2002)</td>
<td>61</td>
</tr>
<tr>
<td>3.2.7. Carnell-Richardson (2005)</td>
<td>63</td>
</tr>
<tr>
<td>3.2.8. Kreuz distance (2007)</td>
<td>64</td>
</tr>
<tr>
<td>3.2.9. Multi-unit distances</td>
<td>64</td>
</tr>
<tr>
<td>3.3. Discussion</td>
<td>66</td>
</tr>
<tr>
<td><strong>4. Shannon Information Theory</strong></td>
<td>69</td>
</tr>
<tr>
<td>4.1. A crash course on information theory</td>
<td>70</td>
</tr>
<tr>
<td>4.2. Shannon Information Theory in neuroscience</td>
<td>73</td>
</tr>
<tr>
<td>4.3. Fisher information</td>
<td>74</td>
</tr>
<tr>
<td>4.4. Event space discretization</td>
<td>75</td>
</tr>
<tr>
<td>4.5. RDT, Information Bottleneck, Optimal state space quantization</td>
<td>76</td>
</tr>
<tr>
<td>4.6. Relevant Component Analysis</td>
<td>77</td>
</tr>
<tr>
<td>4.7. Spike train metrical analysis</td>
<td>77</td>
</tr>
</tbody>
</table>
## Contents

4.8. The specific information .................................................. 78  
4.9. Biological relevance of Shannon Information Theory ............... 79  

5. Galway Information Theory 81  
5.1. Invitation to Galway information theory .......................... 81  
5.2. Metrical information: from equivalence classes to distances .... 83  
  5.2.1. Galway entropy ..................................................... 83  
  5.2.2. Metrical Entropy ................................................... 85  
  5.2.3. Local Version of the Metrical Entropy ......................... 88  
5.3. Discussion ............................................................... 90  
5.4. Conclusions ............................................................... 92  

II. Theoretical Applications of the Galway Information Theory 99  

6. An information-theoretic account of the Perceptual Magnet 101  
6.1. Introduction ............................................................. 101  
6.2. Methods ................................................................. 102  
  6.2.1. Local version of metrical information ....................... 102  
  6.2.2. Generalization with a metrics on the stimulus space ...... 103  
6.3. Results ................................................................. 104  
6.4. Discussion .............................................................. 109  
6.5. Conclusion .............................................................. 111  
6.6. Appendix ............................................................... 111  

7. Galway entropy as a measure of Complexity 113  
7.1. Introduction ............................................................. 113  
7.2. Correspondance with an axiomatic approach to complexity ...... 114  
7.3. Variation of the complexity under perturbation .................. 114  
7.4. Results ................................................................. 116  
  7.4.1. Regular Graphs ................................................... 116  
  7.4.2. Erdős-Renyi graphs .............................................. 117  
  7.4.3. Watts-Strogatz graphs .......................................... 117  
7.5. Discussion .............................................................. 118  

III. Results on Somatosensory data 121  

8. Anatomy of the Ascending Somatosensory Pathway 123  
8.1. General organization of the nervous system and the brain ...... 123  
8.2. Ascending somatosensory system ................................. 124  
  8.2.1. Cutaneous receptors ............................................. 124  
  8.2.2. Mechanoreceptors ............................................... 125
Chapter 1: Introduction

Where I give an overview of my own approach to neuroscience. This involves explaining how I consider that we should really embed the individual and its nervous system in the physical world. How we should understand what we mean by plasticity or thoughts in physical terms. And thereby, get rid of notions that may be reminiscent of some vitalism or even animism.

Where I also outline the structure of this manuscript, starting from the standard methods that constitute the basis for this work: neural coding, spike train metrics, information theory. I then move on to the methods I have developed that merge information theory and spike train geometry and the reasons why I did so. Some of the theoretical outcomes of this method are discussed. We then go back to biology and apply this method to somatosensory data as well to the outputs of a neural network I designed. The importance of the geometry of spike trains is discussed in a final chapter.

Chapter 2: Neural Coding

Where we deal with this fundamental question of neuroscience, that is to understand how external signals are processed and transmitted by the nervous system. In particular, how the populations of spike trains encode the information about their inputs, what features are meaningful. In a seemingly chronological order, we survey the works that helped to understand the neural code, that is, the dictionary between the stimulations and the responses of a neuron or a population of neurons. In particular, we will review two of the main potential coding strategies that have been studied: rates and spike times. Among fundamental concepts developed, we briefly recall the notions of receptive fields, tuning curves or Poisson processes. We stress that spike time coding, first spike latency coding and rank-order coding appear as powerful and reliable methods of encoding information.

Chapter 3: Structure of the spike trains space

Where, instead of considering the object spike train, we explore the relations between them. We first attempt to understand the topological structure of the space of spike trains. This topology may differ according to the view we have on spike trains: either a collection of low-dimensional manifolds or a single high-dimensional manifold. We then move on to a study of the geometry of the space of spike trains. We review some of the existing distances and correlation measures that exist. We try to show what are their advantages and drawbacks. We end this chapter by a short discussion on the possible interpretations of these distances.
Chapter 4: Shannon Information Theory

Where we recall the basics of Shannon information theory. In particular, we put forward the fact that Shannon entropy is a topological measure of the uncertainty about the events. Then we review existing methods to estimate the information transfer capacity of neural systems. In particular, we recall standard methods in neuroscience to quantify the information contained in spike trains, such as the binning procedure or the spike train metrical analysis. We also discuss methods based on rate-distortion theory to estimate the extent to which stimulations can be reconstructed for a given level of information transfer.

Chapter 5: Galway Information Theory

Where we derive a definition of entropy that does embed an explicit measure of the metrical relations between input-output events. Indeed, the two approaches studied in the two previous chapters mutually ignore each other. Shannon information theory does not take the structure of the spike train space into account. We thus modify information theory so that it naturally depends on the metric organization of spike trains. This new parametric information analysis relies on similarity measures induced by the metrical relations between neural responses as spikes flow in. Thus, in order to assess the conditional entropy and the overall information transfer, this method does not require any a priori decoding algorithm to partition the space into equivalence classes. It therefore permits to determine the optimal parameters of a class of distances with respect to information transmission. We discuss the approach also with respect to existing methods. The proposed information theoretical measure proves to be a suitable generalization of Shannon Mutual Information in order to consider the metrics of temporal codes explicitly, that could permit an assessment of neurotransmission reliability in the presence of large spike train spaces (e.g. neural population codes) with high temporal precision.

Chapter 6: An information-theoretic account of the Perceptual Magnet

Where we extend the Galway information analysis. We propose an input-output symmetric version of the Galway information by introducing a metrics on the input space the same way it was done on the outputs. In addition, in contrast to the original optimization method that required the equivocation (or conditional entropy) to be zero and, simultaneously, the metrical information to be maximum, here an objective function setting a trade-off between minimizing the equivocation and maximizing the information is proposed. In the spirit of Butts and Goldman (2006) (Butts and Goldman 2006), we investigate here how an optimal decoder should behave in the low and high-noise regimes. It is shown that the well-known perceptual magnet effect (Kuhl 1991) can be understood as a consequence of maximizing the objective function in a low-noise regime. In a simple case with narrow Gaussian categories, it implies that an optimal decoding system should perform very fine discrimination near the boundaries of the categories, while being less selective near the centers of their distribution.
Chapter 7: Galway entropy as a measure of Complexity

Where we show that the Galway entropy can be understood as a measure of entropy on graphs, whose edges are representations of the similarities between events. We illustrate it on a few simple graphs to exhibit its behavior. We then compare this measure with the graph entropy introduced by Körner and to other approaches to graph complexity. Finally, we compute the Galway entropy of a well-known graphs such as the Erdős-Rényi and Watts-Strogatz graphs.

Chapter 8: Anatomy of the Ascending Somatosensory Pathway

Where we give an introduction to the ascending somatosensory pathway. We provide a description of the tactile apparatus and its functional unit, the mechanoreceptor. We also recall how their characteristics can be unveiled using microneurography experiments. Finally, we give an overview of the projections of the mechanoreceptors to higher-order structures, focusing on the first relay to the central nervous system: the cuneate nucleus, one of the two dorsal column nuclei, located near the brain stem. Of particular interest to us are the electrophysiological properties of the cuneate neurons and the connectivity patterns with the mechanoreceptors.

Chapter 9: Metrical Organization of Mechanoreceptors Signals

Where we apply the proposed information theoretical approach, we study precise temporal decoding of human somatosensory signals recorded via microneurography experiments. For this analysis, we employ a similarity measure based on the Victor-Purpura spike train metrics. We show that, with appropriate parameters of this distance, the relative spike times of the mechanoreceptors responses convey enough information to perform optimal discrimination – defined as maximum metrical information and zero conditional entropy – of 81 distinct stimuli within 40 ms of the first afferent spike.

Chapter 10: Cuneate Nucleus Model

Where we design a model of the Cuneate Nucleus in order to assess its capacity to convey the metrical information of signals from the mechanoreceptors. This model is a simple feedforward layer of Spike-Response Models. The electroresponsiveness properties of the cuneate neurons as well as the connectivity patterns from the mechanoreceptors to the cuneate cells is fitted to experimental data. The synaptic efficacies are modulated according to a plasticity rule designed to maximize Mutual Information. After a learning phase during which the microneurography data were fed to the nucleus, the capacity of our model to transmit metrical information was probed. We found that the perfect discrimination condition at the output of the CN is reached only 5 ms later than at its input. This shows that a simple model is able to convey a high metrical information. In addition, the optimal distribution of synaptic weights
turned out to be highly bimodal, with only 3 to 6 non-silent synapses. This surprising result is in agreement with recent recordings performed in the Cat Cuneate Nucleus.

Chapter 11: Organization of the Signals

We have so far shown that the mechanoreceptors are organized so that, not only Shannon topological discriminability is guaranteed, but also geometric discriminability. We have designed a model of cuneate nucleus that is able to transmit this organization in a very short time. We now want to understand in more details how signals are organized and how they can be decoded/reencoded. In order to do this, we define a new distance on the set of spike trains that mimics the behavior of a well-known neuron model, the Spike-Response Model. This distance takes spike trains as inputs and computes the distance between the time-course of the probabilities of firing of this neuron, i.e. it evaluates the reencoding of the signals. We then apply this new distance to the mechanoreceptors signals and exhibit, for some values of the parameters of the distance, some logical organization of the signals. In particular, we show that the outputs of the cuneate nucleus can be, when we consider the whole spike trains, isometric to the stimulations. We conclude by arguing about the potential role of this isometry property.
## List of Figures

0.1. Location of this PhD on a map of Ireland. ................................................. 10

2.1. Two examples of place cells activity in a square environment ...................... 37
2.2. The response of a primary visual cortex neurons to moving oriented bars ... 38
2.3. Receptive fields of two cutaneous mechanoreceptors probed with Von Frey hairs 39
2.4. Distributions of the numbers of events for Poisson processes .................. 40
2.5. Distribution of interspike intervals ............................................................. 41
2.6. Firing rate time-course of Inferior Temporal neurons in response to visual stimuli 42
2.7. Entropy and information of a neuron from the cricket cercal system ........... 44
2.8. Example of a neuron acting as a coincidence detector ............................. 45
2.9. Spiking responses of a FA-I mechanoreceptor to stimulations in 5 directions presented each 5 times ................................................................. 46
2.10. Spiking responses of a FA-I, SA-I, SA-II mechanoreceptors to stimulations in 4 directions ........................................................... 47
2.11. Transduction of a spatial pattern in a volley of spikes across afferents ........... 48

3.1. Redundancy of the set of non-ordered numbers for spike trains ................. 52
3.2. Geometric loss of the embedding of time into a high-dimensional space ....... 53
3.3. Example of the computation of the Victor-Purpura distance between two spike trains ................................................................. 55
3.4. Two configurations of tables in a room ..................................................... 55
3.5. (a) Example of two stimuli $A$ and $B$ each eliciting two distinct responses $A_1, A_2$ and $B_1, B_2$, respectively. (b) Calculation of all the VP distances $D_{VP}$ between the response spike trains the VP cost parameter $C_{VP}$ varies. .... 57
3.6. Examples of four spike trains that cannot be embedded in a 2-dimensional euclidean manifold ......................................................... 58
3.7. Two spikes trains filtered with a rectified exponential and their instantaneous difference ................................................................. 59
3.8. Scaling of the distances with the number of spikes when patterns are repeated 59
3.9. Two spikes trains filtered with a rectified exponential and their instantaneous difference ................................................................. 61
3.10. Event synchronization between two spike trains ....................................... 62
3.11. Example of the Hunter-Milton measure .................................................... 63
3.12. Two spikes trains filtered with their instantaneous ISI ............................. 65

4.1. Basic scheme of a communication system ............................................... 69
4.2. Examples of PCA and RCA on responses to two stimuli ......................... 78

5.1. Binary word decoding, example of metrical entropy of a gaussian and evolution of the distances between spike trains ........................................... 94
5.2. Examples of results on metrical information .......................................... 95
5.3. Results on optimality of similarity measures on a cartoon example .......... 96
5.4. Link between metrical information theory and signal detection theory ...... 97

6.1. Optimal discrimination in a low-noise regime ........................................... 105
List of Figures

6.2. Optimal discrimination in high-noise regime ........................................ 107
6.3. Optimal discrimination in low- and high-noise regimes .......................... 108
6.4. Perceptual magnet phase diagram ..................................................... 110
6.5. Example of implementation of the map from stimulations to responses that may induce the Perceptual Magnet ..................................................... 110

7.1. Galway entropy for all the topologically different graphs with 4 vertices ...... 115
7.2. From left to right, examples of regular, Erds-Renyi and Watts-Strogatz graphs 116
7.3. Galway entropy on Erdös-Renyi and regular graphs ............................... 117
7.4. Galway entropy on Watts-Strogatz graphs ............................................ 118

8.1. The sensory homunculus ........................................................................ 125
8.2. Example of forces applied to the fingertips during a load task .................. 126
8.3. A slice of the fingertip glabrous skin .................................................... 126
8.4. Morphological and functional properties of the 4 types of mechanoreceptors 127
8.5. Highly simplified representation of the somatosensory pathways ............... 128
8.6. Contact parameter set ........................................................................... 129
8.7. Characteristics of the mechanoreceptor spiking activity .......................... 130

9.1. Sketch of the somatosensory pathway we concentrate on .......................... 133
9.2. Information theoretical analysis on a limited stimulus set .......................... 135
9.3. Information theoretical analysis on the entire stimulus set ......................... 137
9.4. Information theoretical analysis on the complete stimulus set .................... 138
9.5. Time course of the $D_{VP}$ distances ..................................................... 140

10.1. Responsiveness and firing properties of a SRM neuron ........................... 142
10.2. Distance matrix at the output of the Cuneate Nucleus ............................. 143
10.3. Information $I^*(R;S)$ and conditional entropy $H^*(R|S)$ over time ............ 144
10.4. Time necessary to discriminate the entire set of 81 stimuli perfectly .......... 145
10.5. Summed information transfer as a function of the connection probability between the mechanoreceptors and the cuneate cells ......................... 147
10.6. Simulation of the impact of synaptic efficacy distribution on information transfer 148

11.1. Comparison of the SRM, Victor-Purpura and Van Rossum distances between two Poisson processes ............................................................. 151
11.2. Comparison of the SRM, Victor-Purpura and Van Rossum distances between two Poisson processes ............................................................. 152
11.3. Distances between 4 stimuli as spikes are flowing in ................................ 153
11.4. Latencies of a single mechanoreceptor to many stimulations .................... 154
11.5. Distance matrix of the 81 shuffled inputs .............................................. 155
11.6. Multi-dimensional scaled representation of responses to 7 stimulations .... 156
11.7. Multi-dimensional scaled representation of responses to 12 stimulations .... 156
11.8. A donkey, a mule and a horse surprisingly appearing in this manuscript ...... 157
11.9. A zebra, a zorse and a horse less surprisingly appearing in this manuscript ... 157

14.1. Exemple de distance de Victor-Purpura ................................................ 168
14.2. Exemple de distance de Van Rossum ................................................... 169
List of Figures

14.3. Propriétés de l’information métrique ................................................. 178
14.4. Discrimination optimale dans un régime à faible bruit .......................... 179
14.5. Discrimination optimale dans un régime à bruit élevé ........................... 180
14.6. Entropie métrique de tous les graphes à 4 sommets ............................ 181
14.7. Analyse de théorie de l’information métrique sur un ensemble limité de stimulations ............................................................... 182
14.8. Analyse de théorie de l’information métrique sur l’ensemble complet de stimulations .............................................................................. 183
14.9. Comparaison des distances de Victor-Purpura, Van Rossum et SRM entre des processus de Poisson .......................................................... 184
14.10. Comparaison des distances de Victor-Purpura, Van Rossum et SRM entre des processus de Poisson ......................................................... 184
14.11. Latences d’un seul mécanorécepteur en réponse à plusieurs stimuli ...... 185
14.12. Multi-dimensional scaling des réponses à 7 stimulations. .................... 185
14.13. Multi-dimensional scaling des réponses à 12 stimulations. ................... 186
Where I give an overview of my own approach to neuroscience. This involves explaining how I consider that we should really embed the individual and its nervous system in the physical world. How we should understand what we mean by plasticity or thoughts in physical terms. And thereby, get rid of notions that may be reminiscent of some vitalism or even animism.

Where I also outline the structure of this manuscript, starting from the standard methods that constitute the basis for this work: neural coding, spike train metrics, information theory. I then move on to the methods I have developed that merge information theory and spike train geometry and the reasons why I did so. Some of the theoretical outcomes of this method are discussed. We then go back to biology and apply this method to somatosensory data as well to the outputs of a neural network I designed. The importance of the geometry of spike trains is discussed in a final chapter.

This work attempts at unveiling some properties of how sensory stimulations are encoded into spiking signals and potentially decoded by the nervous system. The modality that is studied here is the human somatosensory system. I have used spiking signals evoked by the tactile receptors of the fingertip, recorded in awake humans, in response to well-controlled stimulations. It has been shown recently in many modalities, and in particular in the somatosensory system, that a decoding scheme using the precise timing of spikes would allow a much larger amount of information to be transmitted in a much shorter time than if only the firing rate modulations were considered (Johansson and Birznieks 2004). I propose to further these studies and to focus not only the discriminability of signals, but also their organization (in particular
1. Introduction

Indeed, the tactile receptors can be seen as devices mapping a set of stimulations into a set of spike trains. While information theory assesses the non-ambiguity of this map, it seems of prime interest to study this map in more details and to unveil some of its geometric features, beyond its non-ambiguity. This is the focus of this manuscript.

But, before going further into the details of my work, I think it is relevant to say in a few words what my own approach to neuroscience is. Since this is not the place for a detailed line of argument, I will only briefly expose the hypotheses underlying this work.

1.1. Scientific context

Objects involved in neuroscience

What are the questions we wish to answer in neuroscience? In order to know what these question are, it is first necessary to describe the physical objects that are involved. The first one is the universe, which is described by rather well-known physical laws. The second relevant object is the individual, made of cells, and among them, neurons. This individual, as a part of the physical world, is governed by the laws of physics.

Isolation of individuals

Note already that the very fact that we separate the individuals from the rest of the universe is to some extent arbitrary. Why do we consider some molecules together, while keeping other molecules outside? Especially when you reckon that the material constituting an individual is constantly renewed. Indeed, an individual is not an isolated system, it exchanges work, heat and matter with the outside world. We probably isolate an individual because we consider it an autonomous system. While this may be a reasonable argument, the arbitrariness of the border between an individual and the outside world remains a strong argument in favor of the reduction of the nervous system to physical laws. There is indeed no reason to believe that this set of molecules that is arbitrarily separated from the rest of the world should be governed by other physical (or non-physical) laws.

Subject of neuroscience

Neuroscience is about studying some of the interactions between an individual and the external world, but moreover, studying how these interactions are processed internally by the individual and how, in the end, relevant actions are performed in response to stimulations. These interactions may be of various types: the bombardment of the retina by photons, the electromagnetic contact between the hand and an object, the arrival of pressure waves at the ear, etc... all these interactions are encoded the same way by the nervous systems: with action potentials that propagate along the axons of the neurons. These signals elicit observable reactions of the

---

1In the field of neuroscience, this kind of problem was, in the very specific case of synaptic plasticity, put forward in particular by Lisman (Lisman 1985).
individual, for example, a movement of the hand, a sound, a blink of the eyes... The transformation of stimulations into actions is done through a communication channel that is precisely the nervous system.

**Physical laws**

We can in general consider physical laws as black boxes that are fed with a kinematic state (a point in phase space) and that spit out a new kinematic state. A necessary criterion to do physics is the reproducibility of the laws in space and in time. The laws of physics are invariant under spatio-temporal translations (actually under even more general transformations called spatio-temporal diffeomorphisms, but in non-relativistic cases, we may consider only translations and rotations, i.e. the euclidean group). This is what allows us to do physics. We could say a word on this. There is no reason why the laws of physics should hold anywhere in space and time. We have no proof of this. We should probe all points in the universe to claim such a thing. So, when a law is discovered, there is a strong induction that this law holds true through space and time. For example, the fact that the speed of light is constant in space and time is a strong induction. Though we have hints that it is true as long as we do not consider too far apart regions of space, it may break down at some point in space...or in time.

**Reactions as physical laws**

Since we aim at understanding neuroscience in terms of physics, we may, in a naïve first approach, interpret the reaction of an individual to a stimulation as a physical law. This is indeed no more than the complex product of elementary laws. However, in opposition to the features of a physical law, what is usually considered a fundamental feature of an individual is its plasticity, the fact that it evolves with time and with its environment. Throw a ball twice at someone exactly the same way. He will surely not behave similarly on the first and on the second trial. He will change his reaction and thus modify its output state even though the input state is the same. Thus, we can no longer identify the behavior of an individual with a physical law that is required to be reproducible.

**Plasticity**

We can however easily counter this argument. There are actions that, if repeated enough times, will have effects on the next trials. For example, if you hit the ground with a hammer many times, it will appear undamaged many times until it suddenly breaks, though the last hit was similar to the previous ones. This shows that individuals seen as black boxes are not the only ones to be plastic. The reason is that some phenomena are not universal physical laws but more complex reactions of a system that, at each trial, is modified by its previous history. There are some hidden variables that, if not taken into account, may lead to the belief that the physics is not reproducible. Similarly, in individuals, after each trial, the kinematic state of the nervous system is modified. The modification may, for example, involve the synaptic efficacies. These are the hidden variables that make the reaction to a stimulation different from trial to trial. This is why, in addition to the observable reactions, we have to take into account the fact that input signals modify the properties of the individuals, by tweaking some properties of its nervous system.
1. Introduction

system, phenomenon called plasticity. But this plasticity is again nothing more than physics. I was surprised to hear many times that what separates animals from the rest of the universe is plasticity. I think this is wrong. When I bought my first car, the clutch pedal was too hard to press and rendered driving unpleasant. After a few thousand kilometers, it started being softer and much easier to use. My car had learned!... Or had it? What makes me use the word changed or learned? Probably nothing than bad habits. The same way, the nervous system changes, is modified by inputs. We say it learns, probably because it usually becomes better and better when trained. Though this is far from being obvious that the nervous system only improves with time.

The point of this discussion is again that the nervous system is nothing more than a physical system and we should not be mistaken by the use of different words in different contexts that may make us think that the nervous system is special.

Spike trains space

Now, neurons communicate with each other by sending electric impulses called action potentials, or spikes. The shape of all these events does not vary much and is thought to convey no information. The only important quantities are thus the time of occurrence of these events. Understanding the structure of the spike trains is crucial for many reasons. First, the nervous system is the communication channel that transform sensory signals into actions. Second, these spike trains are the only signals that the nervous system has access to so as to change its parameters, or learn the structure of the world.

The nervous system as a computing machine

To be more precise, let us denote $s_t$ the state of an individual at time $t$. When he receives a sensory signal $S$, he acts according to this signal and to its present state. But, in addition, its state is modulated because of this input. Thus each time a sensory signal is received, the individual performs an action, but its current state (that may be dynamical) is modified as well. Thus a further presentation of the sensory signal $S$ may trigger a different action. This view is highly reminiscent of the functioning of a Turing machine (Turing 1937). It is also in the line of Jacques Loeb’s “mechanistic conception of life” (Loeb 1912).

The mechanistic conception of the brain

I think it is the right place, at this very early level of this manuscript, to mention (very briefly) one of the problems that we are faced with when studying neuroscience: the problem of consciousness and free-will. As a physicist, I do not see any reason to believe in such a concept. In particular, because of the above-cited arbitrariness in the separation of an individual from the universe. And I guess it is worth quoting Hebb (Hebb 1949): “There is no separate soul or lifeforce to stick a finger into the brain now and then and make neural cells do what they would not otherwise”. A recent article (Cashmore 2010) reviewed most of the reasons why we should definitely get rid of this concept. We can thus consider the problem of biology as a pure physical problem. I was struck by the fact that this does not seem to be such a rooted
I had the chance to discuss with tried to argue that maybe in the probabilistic nature of quantum events would be the key to free-will. As pointed out by Cashmore, this is definitely not an argument in favor of free-will. The main arguments that were used are the three limitations to our knowledge of the laws of the universe that arose in the twentieth century: quantum mechanics randomness, chaos unpredictability and Gödel incompleteness theorem. None of them can really be used to positively assert the existence of such a thing as free-will...or to deny its non-existence.

This discussion may seem far-fetched and not much relevant to the problem we study in this manuscript. I guess it is much more than it seems. It means that my view on biological systems is a highly materialistic one. There is to me no more than physical processes going on. Thus, we should be able to understand the whole functioning of the nervous system from a rather bottom-up approach, as it is done in physics, where we try to recover macroscopic phenomena from the interactions between particles. In addition, it states that everything in the behavior of individuals can be reduced to physics and chemistry. In particular, actions and thoughts are reducible to the spiking activity of neurons. We can phrase it the following way: “A spike train is a thought”. The argument presented here will be used in chapter 11.

Now that I have presented in a few words some of the hypotheses underlying my work, I will move on to stating more precisely the problem I have tried to address during these 3 years.

1.2. Statement of the problem

The external world is somehow mapped into the neural code. Receptors - whether they are retinal cells, hair cells or mechanoreceptors - transduce an analogous stimulation into a sequence of impulses. This difference in nature of these two codes makes the study of this map a highly non-trivial problem. However, it seems that the features of this map are crucial in order to understand how the nervous system functions. Our main questions will thus be: what does this map look like? Is it continuous or discrete? Is it finite or not? Is it euclidean?

The point of this thesis is to study this map. However, we have to be precise when referring to it.

Let us forget for a few minutes the plasticity of the brain. In essence, the nervous system is a map from sensory stimulations to motor commands. This map depends on the state of the nervous system. As we have put forward, the map is modulated by the state and the properties of the nervous system at each instant. In particular, the signals sent by the mechanoreceptors flow directly to the Cuneate Nucleus, a group of neuron located in the brain stem that is a compulsory path to the central nervous system. This nucleus is subject to top-down control from the central nervous system (Kandel et al. 1991) and thus process the inputs accordingly. I guess it is worth quoting Von Neumann in the Computer and the Brain (Von Neumann 1958): “In some of the very last analog machines a further trick was introduced. These had electrical, ‘plugged’ connections. These plugged connections were actually controlled by electromechanical relays, and hence they could be changed by electrical stimulation of the magnets that closed or opened these relays. These electrical stimuli could be controlled by punched paper tapes, and these tapes could be started and stopped (and restarted and restopped, etc.) by electrical signals derived at suitable moments from the calculations.” Von Neumann suggests that a machine can be fed with various control inputs such that it behaves as different machines.
1. Introduction

He refers to this control inputs as short codes.

The cuneate nucleus can then be seen as a machine that receives short codes as inputs. Basically, this is a general machine that under the influence of cortical top-down control behaves as specific machine depending on the task to realize. We propose, in this manuscript, to provide the information about how the machine should function in various cases in order to maximize the information transmitted (although we do not tell how this emulation is implemented, what kind of top-down control would make it behave as required).

The point here is that the map from the inputs to the outputs depend on the state of the nervous system (the state itself being determined by its previous history). What does it mean? Simply that the geometry of the input space is determined by the state of the part of the nervous system that has to read it. The geometry of this space is a projection of the properties of the nervous system. The latter acting on the neurons that receive the inputs as a short-code generating machine.

We first review (chapter 3) some of the most important results concerning neural coding. As we want to understand some features of the map, we study the structure of the spike train space (chapter 4) (its topology and geometry). We state that, as long as we have not defined the observer (who is going to read the spike trains), there is no a priori geometry on the space. Only when an observer is defined, we have some notions of similarity or distance on it. These notions are projections of the properties of the reader onto the space of spike trains, that did not have a rigid structure beforehand.

We then review methods from Shannon information theory (chapter 5) that allows to assess quantitatively and rigorously the amount of information conveyed by spike trains. This is referred as discriminability or non-ambiguity. This information is an upper bound of what can be possibly extracted. We then tweak this information so as to embed the properties of the decoder in it (chapter 6). The properties of the decoder will be described, as suggested in the previous paragraph, by the geometry of the space they induce. The information obtained will be lower than Shannon information, but will have the advantage of encompassing the capacities of the reader to decode the information and will allow us to make predictions on the optimal values of the decoder. We will accordingly refer to this as the geometric discriminability. We then explore the properties of this new information measure (chapter 7 and 8). In particular, the fact that it allows to give an information-theoretic explanation of a psychoacoustic phenomenon called the Perceptual Magnet, whereas Shannon information theory cannot. We also show that this new information measure can be interpreted as a measure of graph complexity, where edges represent similarities between events.

We apply this method to microneurography data (chapter 10) of the somatosensory system with the Victor-Purpura distance and show that a set of 81 stimulations can be decoded in a time compatible with temporal constraints imposed by behavioral reactions. We then design a model of the Cuneate Nucleus as a feedforward layer of spiking neurons based on electrophysiological recordings (chapter 11). We connect it with the mechanoreceptors according to patterns observed in the cat. We then apply a synaptic plasticity rule that modifies the efficacy of synapses. After this learning procedure, we show that the output spike trains of the Cuneate Nucleus can also be perfectly discriminated. The metrical information reaches its maximum theoretical value only a few 5 ms after it has reached it with mechanoreceptors signals. This shows that, more than discriminability, the metrical organization of the mechanoreceptors signals can be transmitted quickly and reliably by a simple feedforward layer of neurons. In
1.2. Statement of the problem

addition to this, we show that the synaptic distribution that leads to the higher information transfer from the input signals to the output of the Cuneate Nucleus is strongly bimodal with as few as 5 non-silent synapses. Though it may appear as a surprising result, it is in agreement with observations in the cat CN that, out of 300 existing synapses at the input of a single cuneate cell, as few as 5 to 10 are non-silent.

A lot of efforts have been made to understand how the nervous system -seen as a huge bunch of connected neurons- works. Many algorithms have been developed to crack the so-called neural code. However, we have to pay attention not to introduce artificial observers/decoders. While these observers may extract a lot of information, they may not really exist and therefore, the information they extract may not be really extracted by a real observer. The only ones that have to read the code are neurons. This is the main point of this thesis: to study not how we can decode the signals, but to study how the signals are organized so as to provide a relevant and faithful representation of the world. The code used, just like the letters and words in a given language, is in itself irrelevant and arbitrary. The only thing that really matters is how the signals are organized. In a final part (chapter 12), we thus define a distance that mimics the behavior of a simple neuron model (the SRM). This is meant to provide a geometry to the space of spike trains that is biologically plausible. We show that by using correct parameters (or correct top-down control or short codes on the cuneate nucleus), the spike train space is isometric to the stimulus space. We end by discussing the potentialities of such a property.
Part I.

Cracking the code
Every spike is sacred,  
Every spike is great,  
If a spike is wasted,  
God gets quite irrate.

Spoof of the Monty Python’s song The Meaning of Life

Neural Coding

Where we deal with this fundamental question of neuroscience, that is to understand how external signals are processed and transmitted by the nervous system. In particular, how the populations of spike trains encode the information about their inputs, what features are meaningful. In a seemingly chronological order, we survey the works that helped to understand the neural code, that is, the dictionary between the stimulations and the responses of a neuron or a population of neurons. In particular, we will review two of the main potential coding strategies that have been studied: rates and spike times. Among fundamental concepts developed, we briefly recall the notions of receptive fields, tuning curves or Poisson processes. We stress that spike time coding, first spike latency coding and rank-order coding appear as powerful and reliable methods of encoding information.

Neural coding is about understanding how external (physical) stimulations are represented in the activity of the brain or how neurons map their inputs into output spike trains. It has been well-known, for almost a century, that neurons communicate with each other by sending electrical impulses along their fibers. These impulses are called action potentials or, more shortly, spikes. These impulses are stereotyped and are the only signals that propagate along the fibers of the neurons. The immediate consequence is that the only relevant quantities for coding are the times at which spikes are emitted or, on a coarse-grained scale, the frequency at which they are emitted.

Since there are many types of neurons, there may be many different answers to the question of the neural code. Some neurons are sensitive to external stimulations, such as light or pressure waves, while other only receive inputs from other neurons. The problem is highly non-trivial because of the complexity of the responses of neurons. Some express the problem as finding a dictionary (Rieke et al. 1997) between the stimulations and the responses of neurons. However, this dictionary is stochastic, it is not univocal: (i) two different stimuli may elicit the same response and (ii) a single stimulus presented many times may elicit many different responses. In addition, we would like to express caution when talking about a dictionary. Indeed, we have some a priori thought on what a dictionary is. Let us consider a usual French-English dictionary: for almost every word in English, there actually exists a well-defined word in French. Even if there are slight variations on the precise meaning of the words, there is a one
2. **Neural Coding**

Word to one word translation. The translation from stimulations to spike trains may be much more complex. There is language of some people in Oceania where there is no word for blue and green, but a word for greenish blue (or bluish green). The dictionary from English to this language may be much harder to write here because you may need some longer description for our actual blue and green. The neural dictionary is likely to be more similar to the latter case than to the usual English-French dictionary. It is thus far from being obvious what features of the responses are relevant in order to determine the nature of the stimulation.

As the mapping from the input to the output is (at least partially) stochastic, probability theory appears as a good candidate to formalize ideas. In general, we describe a set of stimulus $S = \{s\}$ and a set of responses $R = \{r\}$. Each of the event has a probability $p(s)$ or $p(r)$ to occur. On the one hand, the problem of *encoding*, i.e. determine how neurons translate inputs into outputs, as determining the complete set of conditional probabilities $p(r|s)$. On the other, the brain faces the opposite task, i.e. determining the stimulation according to the response. This is known as *decoding* and is formalized as determining the complete set of probabilities $p(s|r)$. These are two facets of the problem of neural coding. Note that the two questions are linked by Bayes’ rule $p(s|r)p(r) = p(r|s)p(s)$. Yet, they are not completely equivalent, since the relation requires knowledge of $p(s)$, i.e. the structure of the world. This is a euphemism to say that it virtually impossible to determine it completely. One of the main reason is that we constantly update our beliefs about the probabilities. For example, looking at a thermometer that tells whether the temperature is between $-100$ and $60$ celsius or not does not give much information. It is redundant to what everyone knows. The probability of havine temperatures outside of this range is indeed virtually null (However, having such a thermometer may be highly useful for a galactic explorer).

The problem when using information theory is thus to determine what prior probabilities we should use. Should we use some highly marginal probabilities? For example, the probability for an object to be at a given temperature should be based on a particular physical distribution. Or should we use *updated priors*? The probability of an object being between 0 and 100 is nearly one. Of course, the results of a study of the thermoceptive system highly depends on such choices. As for our study, we use a set of 81 equiprobable stimulations. We found no reason to believe that one of those stimuli would be more probable than another. But we keep in mind that we have only approximated the *true* distribution.

An issue arises from the fact that stimulations are dynamical. We have to determine the extent to which the impulses in a given window of time code for a determined stimulation or to the time-course of a stimulation. Indeed, in some modalities, the time-scale of the stimulation dynamics may be of the same order as that of the average interspike interval. How can we determine whether a spike codes for the instantaneous stimulation or should be considered with the previous spike as coding for previous values of the stimulation? Intuitively, this is the same as when you hear an unknown language, you never whether the sound you hear at a given instant $t$ is part of a word coding for the meaning at instant $t - \delta t$ or whether it actually codes for the meaning at instant $t$. Let us take an example, here is a sequence of Japanese words that describes the successive positions of a cat as it is walking in a room: *te e bu ru i su ne da i*. There may be a confusion in the way to assemble the syllables in order to give them meaning. For example, is the syllable *ru* coding for the instantaneous position of the cat or does it have to be considered with *bu* or even *te e bu* to describe the position of the cat at an instant slightly before. Actually, *te e bu ru, i su and ne da i* respectively means *table, chair and bed.*
2.1. Description of spike trains

In addition to these difficulties, stimulations are coded by populations of neurons, rendering the feature space of the responses highly dimensional and hardly tractable. Indeed, the dimension of the problem do not merely scale with the number of neurons, but as a power of this number since correlations between their activities should be taken into account (Thorpe et al. 2001).

In this part, we review some of the most important results on neural coding of the last century. Because of the difficulties cited above, this is still a subject of intense ongoing work.

2.1. Description of spike trains

Since spikes are short-timed, stereotyped and have an amplitude much larger than usual fluctuations of the membrane potential, the standard way to describe them is with Dirac functions at the time they are elicited. A spike train will thus be formalized as:

\[ s(t) = \sum_i \delta(t - \hat{t}_i) \]  
\[ (2.1) \]

where \( \{\hat{t}_i\}_{i \in \mathbb{N}} \) is the set of firing times and \( \delta(t - t_i) \) is the Dirac function that is null everywhere except at \( t = t_i \), where it is infinite such that its integral is equal to unity.

The spike count over a window of time \( T \) is obtained by integrating the spike train:

\[ N = \int_0^T s(t) dt \]  
\[ (2.2) \]

The time-course of the firing rate, based on a single trial, can be defined as:

\[ r_{\delta t}(t) = \int_t^{t+\delta t} s(t) dt \]  
\[ (2.3) \]

When \( \delta t \) goes to 0, the firing rate \( r(t) \) tends to the original spike train \( s(t) \). When \( \delta t \) becomes large, computing \( r(t) \) amounts to counting the spikes in large windows of time. These two extreme behaviors show that the distinction between spike time codes and rate codes may be more of a quantitative than qualitative nature. This can be made more precise. We can indeed define a fuzzy border between time and rate code: this would be the value of \( \delta t \) that is of the order of the interspike interval. If \( \delta t \) is smaller than this, we get a sparse distribution, with a few spikes here and there. If \( \delta t \) is larger, then we get many spikes in each bin, suggesting a time-dependent firing rate.

In many cases, the time-course of the firing rate is computed with many trials to the same stimulation. This is done by computing the post stimulus time histogram. Time is discretized into small bins, to which a probability of firing is attributed that is the ratio of the number of spikes elicited in this bin to the number of trials. One of the main drawback of such a method is that it is rather biologically unrealistic. The organism does not have access to such an averaged firing rate. It must make decisions based on a single trial.

For detailed discussions about the properties of this definition, we refer the reader to the textbook by Dayan and Abbott (Dayan and Abbott 2001).
2. Neural Coding

2.2. Rate coding

In 1926, Lord Adrian (Adrian 1926) observed that sensory neurons elicit all-or-none impulses, called action potentials, as we have explained above. He also observed that the rate of discharge of a neuron is positively correlated with the amplitude of the stimulation. This has led to a paradigm for neural coding for almost all the twentieth century: neurons encode the magnitude of a stimulus by the number of spikes they emit. In this section, we briefly review how the firing rate is determined with respect to the stimulation, how spikes are generated once the firing rate is known and finally, some of the limitations of this potential code.

Receptive fields and Tuning curves

Spike count has been the definition of neural activity for almost a century, this approach has been fruitful and led to a lot of highly interesting works and results. One of the concepts developed from this paradigm is the receptive field that is defined as the range of stimulation parameters that trigger a significant activity of the neuron. The activity is here taken as the spike count over a fixed time window. The center of the receptive field therefore corresponds to the values of the parameters that yield the highest firing rate. Receptive fields can be defined more generally for neurons that receive inputs from other neurons only (rather than responding to external stimulations). We can thus define the receptive field of a high-order neuron in the visual cortex as the set of visual scenes that make it fire above baseline.

In turn, it allowed to define tuning curves, i.e. functions mapping the stimulation parameter onto a firing rate. Rigorously, we should call them Rate Tuning Curves since the rate is taken as the activity of a neuron. However, the reader should keep in mind that, if neural activity was defined otherwise, the tuning curves would be different.

The two notions of receptive field and tuning curve are somehow overlapping as it can be seen in the example we provide below.

A demonstrative example is offered by place cells. These are neurons located in the brain that fire maximally when the individual is at a specific location in space. When moving away from this location, the firing rate decreases smoothly until it reaches its baseline activity, see Fig. 2.1. The receptive field is thus a portion of the space where the animal can evolve.

Another example is provided by the wonderful experiment by Hubel and Wiesel (Hubel and Wiesel 1962). They recorded a neuron in the primary visual cortex of the monkey and presented a bar in its visual field. The neuron turned out to fire differentially according to the orientation of the bar, see Fig. 2.2. Here, the receptive field is a range of angles with a “hot spot” at a given orientation.

Highly relevant to the subject of this thesis is the determination of the receptive fields of cutaneous mechanoreceptors (Johansson and Vallbo 1983). The standard method to do this is to record the receptor while touching the skin with Von Frey hair of different stiffness at different locations. This allows a complete determination of the receptive field. Though the approach here is slightly different. The receptive field is characterized at each location, not by its firing rate, but rather by the minimal stiffness of the hair that elicited a significant number of spikes. This kind of receptors will be studied in greater details in chapter 9.

Although the first two examples of tuning curves depicted above are bell-shaped, this is not compulsory. There are, for example, cosine receptive fields that are usually found when the
2.2. Rate coding

Figure 2.1.: Two examples of place cells activity in a square environment. Blue indicates a low baseline firing rate, while red is for high spiking activity. The receptive fields are roughly circular, with the center of the circle being the place of highest activity. From (O’Keefe and Burgess 1996).

The stimulation parameter is an angle. The receptive field is then:

\[ r(\theta) = r_0 + r_1 \cos(\omega(\theta - \theta_0)) \] (2.4)

We would also stress that this definition of neural activity led to the understanding of the specialization of brain regions. For example, when we watch some relevant visual scenes, some part or our brains undergo significant activity (large numbers of spikes elicited). This permitted to define zones that process visual signals. This has been done for much of the nervous system, and we know have a good understanding of the specialization of brain areas. There are of course a lot of caveats, but we do not want to delve into details here, our goal is rather to point out that the definition of the specialization of brain regions is done through the definition of neural activity as spike count.

Spike trains as Poisson processes

A common view is that once the rate of firing is determined, the neurons emit spikes according to a Poisson process. A Poisson process is a continuous-time stochastic point process, where the probability of occurrence of an event (a spike) is independent of the time since the last event (that is usually called the Markov property). This is called the independent spike hypothesis or, in probability theory, independent and identically distributed (i.i.d.) random variables. This means that the probability of obtaining two spikes at times \( \hat{t}_1 \) and \( \hat{t}_2 \) is the product of the two probabilities of having a spike at time \( \hat{t}_1 \) and that of having a spike at time \( \hat{t}_2 \).

If the expected number of events occurring in a window of time is \( N \), then the probability of observing \( k \) events is:

\[ p(k|N) = \frac{N^k e^{-N}}{k!} \] (2.5)

We refer to Fig. 2.3 for examples of the distributions \( p(k|N) \) of Poisson processes with mean...
2. Neural Coding

Figure 2.2.: The response of a primary visual cortex neurons to moving oriented bars. There is an direction of movement (D) that triggers a much larger activity. As the direction is changed, the spike count decreases until it is null (A and G). From (Hubel and Wiesel 1968).

numbers \( N = 5, 10, 20, 30 \).

A notable feature of a Poisson process is that, over a fixed time window, its variance is equal to its mean. A quantity to measure this is the Fano factor which is the ratio of the variance to the mean. A Poisson process thus has a Fano factor of 1. Thus, the larger the number of events, the larger the probability of having large absolute deviations from the mean. According to this picture, the spike times are highly irregular and carry absolutely no information. In addition, the observation of a single trial does not provide much information about the stimulation since the number of spikes is highly irregular, see Fig. 2.3.

Another interesting property of the Poisson process is that the waiting time (or holding time) between two spikes (or interspike interval) is distributed according to a decreasing exponential:

\[
p(t_{i+1} \mid t_i = t_0) \propto e^{-p(t-t_0)}
\]  

(2.6)

This means that short intervals are privileged.

Poisson process can be built in two ways:

- discretize time into short bins and assign a spike to each of this bin with a probability \( p \),
- start at the origin of the time axis and generate interspike interval according to the expo-
2.2. Rate coding

Figure 2.3.: Receptive fields of two cutaneous mechanoreceptors probed with Von Frey hairs. In the bottom, lines represent levels of same sensitivity. From (Johansson and Vallbo 1983)

The exponential distribution with parameter $p$.

These two constructive algorithms give the same results.

**Spike trains as Renewal processes**

However, the interspike distribution is clearly violated in real spike trains because of the refractoriness. Indeed, after a spike is emitted, a neuron undergoes a hyperpolarization period during which it cannot spike. Very short intervals are thus prevented. In order to take this fundamental feature of spike trains into account, it is necessary to move from Poisson process to renewal process, the latter being a generalization of the former. To build a renewal process, the standard method is to copy the second one to generate Poisson processes but with an arbitrary interspike interval distribution. It is observed in recorded spike trains that the distribution of interspike intervals can be roughly fitted with a Gamma function (Bair et al. 1994):

$$p(\hat{t}_{i+1} = t | \hat{t}_i = t_0) \propto (t - t_0)^\alpha e^{-p(t-t_0)} \tag{2.7}$$

where $\hat{t}_i$ denotes the time of the $i$-th spike.

A spike train can thus be built as a renewal process with a Gamma distribution of intervals. The renewal process view of spike trains somehow violates the independent spike hypothesis, since the timing of a spike depends on the previous spike, there is indeed a smallest
2. Neural Coding

Possible interval. However, the intervals are themselves still independent of each other. This is why the second method to build a point process is still relevant in the case of renewal process while the first does not work anymore.

Some neurons are selective to several features of the stimulation. In this case, many combination of the values of the features may lead to the same spike count. If a visual neuron is selective to orientation and contrast for example, then a stimulus at low contrast in the privileged orientation may lead to the same number of spikes as a stimulus at high contrast in another orientation. What may help to decode these two stimulations are the temporal modulations of the firing rate, even to a static stimulation.

2.3. Temporal coding

As we have explained above, temporal encoding is not an obvious phenomenon to exhibit. Stimulations are often dynamical, it is thus non-trivial to determine whether the fluctuations in the firing rate of a neuron are due to the time-course of the stimulations or are a manner of encoding a static stimulation with a little delay.

At the end of the eighties, a series of papers by Richmond and Optican (Richmond et al. 1987; Richmond and Optican 1987; Optican and Richmond 1987) showed that one could extract much more information about the signals by taking into account the instantaneous variations of the firing rate, i.e. temporal modulations of the firing rate Fig. 2.6. More precisely, the time-course of the activity of a neuron or of a population of neuron in response to a constant
2.3. Temporal coding

Figure 2.5.: Distribution of interspike intervals (left) recorded in the MT area of a behaving monkey and (right) obtained from a neuron model. They obviously depart from the distributions of waiting times in the Poisson process. Here, they can be fitted by a Gamma function, see text. From (Bair et al. 1994).

stimulus highly reduces the uncertainty about the stimulus that was presented (McClurkin et al. 1991; Eskandar et al. 1992). It is interesting to see that it does not suggest that spikes are well-timed, but rather that a constant stimulus elicit time-varying activities or firing rates, that, in turn, evoke spikes according to Poisson processes.

This shows that simple static presentations of stimulations may elicit complex temporal evolutions of the activity of a neuron, just like the cat in the room that walks on a ne da i. The signal that describes a single static fact is temporally complex.

This was made precise later by Theunissen and Miller (Theunissen and Miller 1995) who gave a rigorous definition of temporal encoding.

Limitations of rate codes

Let us briefly mention a few potential problems with rate coding:

- In 1989, Thorpe and Imbert (Thorpe and Imbert 1989) suggested that the responses observed in the primate brain 100 to 150 ms after the presentation of specific visual scenes rules out rate coding. In the same spirit, a fly is able to change its direction of motion as soon as 30 ms after a significant stimulation is presented visually (Rieke et al. 1997) ch. 2.2 and references therein.

- some neurons only fire once or twice in response to a stimulus. How can a rate be defined in these cases? Gautrais and Thorpe (Gautrais and Thorpe 1998) carried an exhaustive analysis of the coding capacity of rates. For example, they computed that, when only one spike is fired in a 10 ms window, there is 90% chance that the true firing rate lies in the range $[5, 474]$ Hz. In agreement with Fig. 2.4, this clearly shows that, in a short time, it is very hard to even guess what are the properties of a Poisson process. This partly rules out the possibility that neurons actually communicate with Poisson spike trains.

- there is also a computational resource problem. Lennie (Lennie 2003) pointed out that the cost to trigger a spike is high. He showed that a high spiking activity soon represents
2. Neural Coding

Figure 2.6.: Firing rate time-course of Inferior Temporal neurons in response to visual stimuli, consisting of black and white simple patterns. It is conspicuous here that the neuron encode static stimulations through temporal modulations of its firing rate. From (Eskandar et al. 1992).

• a metabolic burden. This suggests that using rate coding is energy-wise inefficient.

• in Fig. 2.6, the firing rate is modulated very quickly. It is virtually impossible to estimate these modulations based on a single spike train. Indeed, the time-scale of the modulations is shorter than the average interspike interval. This is the problem faced by organisms, they have to decode the signals on the fly.

• finally, the main neuron model is the Hodgkin-Huxley model, which is deterministic. There was thus a separation between the view on spike trains and the view on the biophysics of neurons.

• a pure rate code is ruled out for another important reason: the organism point of view or, otherwise stated, the biophysics of the neuron. Its nonlinear properties make it pretty clear that sending it many input spike trains with the same rate (first moment) but different higher-order moments will lead to completely different outputs. One may trigger spikes while the other do not. Neurons thus respond to higher-order statistics of the spike train than the average rate.

We can summarize the rate coding as slow, costly, rather inefficient and noisy.

However, it has been noticed by many (De Ruyter van Steveninck et al. 1997; Warzecha and Agelhaaf 1999; Berry et al. 1997) that the observed spike trains depart strongly from Poisson processes. They have Fano factors much below 1. This means that the number of spikes in real
2.4. Spike time coding

Spike trains is much more fixed than in a Poisson process. This first suggested that the number of spikes can carry more information than if spike trains were Poisson processes.

2.4. Spike time coding

Some studies went further and started to examine how information may be encoded in the exact spike timing. For example, it was noticed in many modalities that the spike times were highly reliable. This means that spikes are more precisely timed than suggested by the Poisson process view where the only relevant quantity is the time-varying firing rate.

In a pioneering work published in 1952, MacKay and MacCulloch showed that the capacity of neurons to transmit information would be much higher if spike timing (interspike interval) was taken into account compared to spike count only.¹

As a basis to much of the following work, in 1990, Thorpe (Thorpe 1990) showed that the brain was able to make a decision based on the timing of individual spikes. The argument is that we are able to process visual information in a very short time. Visually selective neurons in the inferotemporal cortex react to stimulations within 100 to 150 ms. There are about 10 synaptic relays from the retina to these neurons. It leaves 10 ms to each relay neurone to respond. Considering the average rate of these neurones, they only have time to fire once or twice. This means that information about the stimulation is contained in the first spikes of the neurones, and in particular in its latency (also called time-to-first-spike).

This suggested that, indeed, the timings of the spikes had to be precise. In 1994, a series of papers by Shadlen and Newsome (Shadlen and Newsome 1994; Shadlen and Newsome 1994) on the one hand and Softky (Softky 1995) on the other argued about the possibility that spikes were actually precisely timed. The relevance of spike timing became a serious candidate to stand for a neural code. It has in addition a computational power that is much higher than a simple rate code. In a nutshell, while the spike count in a given window of time is just a real number, the spike trains live in much higher dimensional spaces. A thorough examination of this fact is the central point of Chapter 4.

A seminal article by Mainen and Sejnowski (Mainen and Sejnowski 1995) corroborated this idea. They recorded neurons in rat neocortical slices, to which they sent many times two types of current: a constant and a frozen white-noise. The spikes elicited by the neurons were much more reliable in response to the white-noise than to the constant current. This was shown to be a general property of spiking neuron model by Brette and Guigon in 2003 (Brette and Guigon 2003).

In 1995, Hopfield (Hopfield 1995) proposed a computational model that could make explicit use of the precision of spike times. In particular, his model focused on the precision of spike times with respect to background oscillations.

In 1997, Berry et al. (Berry et al. 1997) recorded spikes from the retinal ganglion of the rabbit and of the salamander, they showed that spikes are timed to a precision of 1 ms. Thus the information transmitted by spikes is much larger than that transmitted by spike counts. This happened to be true for a wide variety of contrasts. In addition, they found that the major contribution to trial-to-trial variability was the insertion or deletion of spikes rather than the jitter on spike times.

¹Though some of the assumptions of the article are really questionable.
2. Neural Coding

A rigorous way to estimate the amount of information transmitted is to use Shannon information theory (Shannon 1948). In 1998, Strong et al. (Strong et al. 1998) proposed a rigorous way to estimate the entropy and information of spike trains by discretizing time into small bins. Of particular interest to our approach is the review by Borst and Theunissen (Borst and Theunissen 1999) that focuses on the application of information theory to the study of the neural code. Information theory and its applications to the study of the neural code will be the focus of chapter 3.

Most of these foundational works are reviewed in the book by Rieke, Warland, de Ruyter van Steveninck and Bialek (Rieke et al. 1997).

Many studies have shown that spike times can give information about the stimulations (deCharms deCharms and Mezernich 1996).

Synchrony between spikes (Abeles 1991) obviously requires that spikes are well-timed.

In 1999, Dimitrov and Miller (Dimitrov and Miller 2000) recorded from neurons of the cricket cercal system. As Strong et al. (Strong et al. 1998) proposed, they discretized time into small bins. They measured the entropy and information transmitted for different bin widths, see Fig. 2.7. The information obviously increases as the bin width is reduced. However, there is a value of the width at which the information saturates. Reducing the width does not allow more information to be transmitted. This is what they call a natural time scale for encoding. Departing from this natural time scale would either lead to a loss of information or would not provide much despite a finer decoding scheme requiring more computational resources.

![Figure 2.7: Entropy and information of a neuron from the cricket cercal system, with respect to the width of the bin used to discretize time. At a value of 0.1 ms, the information saturates. This is the natural time scale for encoding. From (Dimitrov and Miller 2000).](image)

A posteriori, it appears to be a surprise that it took so much time after the first insight into the functioning of a neuron (e.g. Lapicque’s integrate-and-fire (Brunel and van Rossum 2007)) to prove that the exact timing of spike is a highly relevant quantity. Indeed, it seems obvious
retrospectively, that two spike trains fed into a neuron can elicit very different outputs even if their rates are similar.

The idea that spike timing encodes information is in line with the fact that some neurons act as **coincidence detectors**. They only fire when they receive synchronous inputs Fig. 2.8.

**Figure 2.8.** Example of a neuron acting as a coincidence detector. When it receives two unsynchronized spike trains, it does not fire. However, if the two input spike trains are synchronized, it fires. Note that, even though the input rates are close, the outputs are highly different.

## 2.4.1. First-spike latency

A potential very fast way of encoding the stimulation is via the latency to the first spike. We can get an intuitive idea of the reason why it is relevant. In the rate coding view, the stronger a stimulus, the higher the rate and therefore the shorter the interspike intervals. The first spike latency can be seen as the first interval (though this is not an interspike interval). It may thus contain an information similar to the rate. However, it was shown by (Johansson and Birznieks 2004) that the first spike latency and the rates do not actually contain the same information.

In 1996, Gawne et al. (Gawne et al. 1996) showed that the latency of the first spike in response to a visual stimulation is strongly correlated to the contrast, whereas, as suggested by Richmond et al. (Richmond et al. 1987), the temporal modulations of the firing rate encode for other features such as the orientation of the visual stimulation.

In 1997, Heil (Heil 1997) showed that, in the auditory system, the first spike latency could be completely and reliably described by a few parameters such as derivatives of the peak pressure. In 2004, he (Heil 2004) showed that the first spike latency conveyed significant amount of information.

Meister and Berry presented results suggesting that the latency decreases when the magnitude of the stimulation increases. In addition, the variability is reduced when the strength of the stimulation is high. Wainrib et al. (Wainrib et al. 2010) performed a detailed mathematical analysis of the effect of stimulus strength on the first spike latency and on its variability. This analysis is carried on a biophysical neuron model: the Morris-Lecar model (Morris and Lecar 1981). They determined the precise function from the stimulus amplitude to the first-spike latency. They also exhibited that the variability decreases when either the amplitude of the stimulus or the number of ion channels increases.

Many studies have then followed concerning the capacity of first spikes to convey information about the stimulation. Among them, in 2001, Panzeri et al. (Panzeri et al. 2001) showed
2. Neural Coding

that in the rat whisker system, the timing of the first spike gives almost the same amount of information as the total spike train, suggesting that the subsequent spikes yield redundant information with respect to the first spike. This is, with respect to our work, a very important reference, since we will deal with this first spike / subsequent spike dichotomy later.

In 2004, Johansson and Birnieks (Johansson and Birznieks 2004) showed that the first spike emitted by a population so mechanoreceptors of the human tactile system could allow a perfect discrimination of 5 stimulations applied to the fingertip in 5 different directions. It was taken further by Saal et al. (Saal et al. 2009) who showed that this could allow for an accurate discrimination of probes with three different curvatures.

Figure 2.9.: Spiking responses of a FA-I mechanoreceptor to stimulations in 5 directions presented each 5 times. The variation in latency from a stimulation to another is much larger than the variation from trial to trial. This makes latency a potential robust and reliable code for information transmission. From (Johansson and Birznieks 2004).

Of particular interest is the discovery of latency tuning curves (LTC). While tuning curves were so far defined according to the number of spikes (or rate) emitted for a given value of the stimulation parameter, it was shown (Johansson and Birznieks 2004) that it is possible to find a tuning curve based on the latency of the first spike, see Fig. 2.10. Moreover, it was shown in the same article that the information conveyed by this latency is not redundant to that conveyed by the rate.

However, the first spike latency is defined with respect to the stimulus onset, to which the central nervous system does not have access. In 2007, Chase and Young (Chase and Young 2007) showed that the information may on average increase with respect to a population onset. They simulated an onset neuron that would fire and thereby signal the onset only when some definite number of cells have fired. Gollisch (Gollisch and Meister 2008) showed that the
2.4. Spike time coding

Figure 2.10.: Spiking responses of a FA-I, SA-I, SA-II mechanoreceptors to stimulations in 4 directions. The variation in latency from a stimulation to another describes a smooth curve. This suggests a privileged direction for spiking. From (Johansson and Birznieks 2004).

relative timing between the first spike of two neurons contain much information about the stimulus.

An open question is whether the jitter in spike latency is correlated among neurons. If it was, then the relative latency would be a very robust quantity. In the somatosensory system, it is suggested that the jitter is correlated among receptors. Indeed, the small variations in the spike timing may be due to factors such as the instantaneous blood pressure at the fingertip (Johansson).

2.4.2. Rank-Order coding

In 1990, Simon Thorpe proposed that the order of arrival of spikes may be a source of reliable information. Albeit potentially less powerful than spike timing, it would be more reliable. It would indeed be mostly insensitive to small jitter if the difference between times is large enough. In 2001, using evidence from the reliability of spike times in the rat whisker system, Panzeri et al. (Panzeri et al. 2001) suggested that Rank-Order Coding would indeed be reliable. Data from many modalities (see Van Rullen et al. (VanRullen et al. 2005) and references therein) suggest that spike timing is sufficiently reliable to make rank-order coding a suitable and efficient codes.

This information could be decoded by an inhibition network, as presented in (Thorpe et al. 2001).
2. Neural Coding

Figure 2.11.: Transduction of a spatial pattern in a volley of spikes across afferents. The central nervous system may profit from this asynchrony to reconstruct the stimulation pattern. From (VanRullen et al. 2005).

2.4.3. Parameters affecting reliability

Among the facts that make spike timing a potential neural code is the reliability of the timings. Neurons are more reliable when stimulated with natural dynamic stimuli than when presented artificial stimulations (De Ruyter van Steveninck et al. 1997), although (Warzecha and Agelhaaf 1999). The level of reliability can be seen as a consequence of the steepness of the membrane potential time-course when it crosses the threshold (Cecchi et al. 2000). There also seems to be a decrease in reliability from the periphery to the central nervous system (Kara et al. 2000).

In 2010, Wainrib (Wainrib et al. 2010) studied the influence of ion channel noise on the first-spike latency.

2.4.4. Measures of reliability

In order to quantify the reliability, Mainen and Sejnowski (Mainen and Sejnowski 1995) computed the peristimulus time histogram (PSTH). They convolved each spike with a filter (e.g. a gaussian) and then computed the sum of the functions obtained. This can be interpreted as the average firing rate. In case spike timing is reliable, large peaks appear in the PSTH. To detect these “events”, a threshold was set at three times the average firing rate. Reliability was then defined as the number of spikes elicited during this events with respect to the number of trials. Precision was defined as the width of this event.

Reich et al. (Reich et al. 1997) in addition to these measures, quantified variability as the variance of the single-cycle Fourier fundamental component. They showed that the variability is stimulus-independent whereas precision is stimulus-dependent.

The next chapter is devoted to the study of distances and correlation measures defined on
spike trains to assess their similarity or dissimilarity.
Structure of the Spike Trains Space

Where, instead of considering the object \textit{spike train}, we explore the relations between them. We first attempt to understand the topological structure of the space of spike trains. This topology may differ according to the view we have on spike trains: either a collection of low-dimensional manifolds or a single high-dimensional manifold. We then move on to a study of the geometry of the space of spike trains. We review some of the existing distances and correlation measures that exist. We try to show what are their advantages and drawbacks. We end this chapter by a short discussion on the possible interpretations of these distances.

Let us make a brief detour by the ideas that guided the development of physics. Let us in particular focus on the debate between Leibniz and Newton concerning the nature of space. Newton argued that there is an absolute space, implying that the position of a single object is relevant in itself. As for Leibniz, he was a proponent of what is now called \textit{relationism}: there is no absolute reference in space, the positions of objects are only relevant with respect to each other. Much of modern science has built on Leibniz’s view. We could, in this respect, refer to Mach’s bucket or to the development of Special and General Relativity. While this could be seen as less true in Quantum Mechanics, there is a trend toward interpreting it in a relationist way (Rovelli 1996). In traditional approaches to neural coding, the goal is to find the \textit{letters} or the \textit{words} with which neurons talk to each other. This may be retrospectively as trying to find absolute positions of the objects “spike trains”. However, guided by the principle of relationism, it may be more relevant to define relations between objects and consider them as the \textit{true} information. This may indeed allow to reconstruct the space from relative adjacency.

Let us consider spike trains in a window of time \([0, \infty]\). What is the topology of the space they live in? What is its geometry?

3.1. Topology of the space of spike trains

Each spike can be described by a real number: its timing. Thus, a spike train can be seen as a collection of real numbers \(\{t_i\}_{i \in \mathbb{N}}\). For example, \(\{34.2, 56.7, 89.9\}\) is a spike train with
3. Structure of the Spike Trains Space

three spikes at 34.2, 56.7 and 89.9 ms. A first possible confusion arises with this notation because spikes are not distinguishable: \{t_a, t_b\} \sim \{t_b, t_a\}. Thus, when representing the space of two-dimensional spike trains, some points are identified, the space is somehow folded in two (see fig. 3.1). This first difficulty can be easily overcome by requiring the collection of real numbers, describing spike times, to be ordered.

![Figure 3.1.: Redundancy of the set of non-ordered numbers for spike trains.](image)

As long as we deal with spike trains with a fixed number of spikes, things are not too bad. Spike trains with \(n\) spikes can be represented in a \(n\)-dimensional manifold (with the caveat of the folding explained in the paragraph above). This space can even be endowed with a euclidean metrics, see next chapter on geometry.

However, spike trains with different numbers of spikes live in different manifolds: where would a spike trains with a single spike live in the two-dimensional space of Fig. 3.1? The space of spike trains is thus a collection of folded manifolds of integer dimensions: the spike trains with \(n\) spikes live on a folded \(n\)-dimensional manifold. However, these manifolds are not even connected with each other. Indeed, it is only possible to have spike trains with an integer number of spikes. Thus, in order to go from a \(n\)-dimensional spike train to another \(m\)-dimensional spike train (with \(m \neq n\)), jumps must be made. The transition cannot be smooth. If we want to define a metrics on this collection of manifolds, some non-trivial operations need to be defined.

### 3.1.1. Embedding in a higher-dimensional manifold or binning

Of course, this aggregate-space can be embedded in a higher-dimensional space. The most common way to do this is to discretize time into small bins (e.g. 1 ms). Each bin is then assigned a value of 0 or 1 whether a spike occurred in it or not. The dimensionality of the space is then the number of bins. This space is not really convenient for many reasons:

- its dimensionality is highly dependent on the level of discretization,
- it is no longer continuous but discrete: a spike is an all-or-none event,
3.1. Topology of the space of spike trains

- all the dimensions are orthogonal and there is no distance between these dimensions.

In each dimension, only two points (0 and 1) are relevant. This is definitely not an economical description. It clearly does not take into account the fact that spikes are all-or-none events. Their amplitude is not described by a vector space and there is thus no need for a dimension to describe a single spike at a single time.

In addition, all pairs of bins are somehow considered identical. Let us consider three discretized spike trains $\mathbf{A} = \{100\}$, $\mathbf{B} = \{010\}$, $\mathbf{C} = \{001\}$ (see Fig. 3.2). These three objects are equidistant. However, it may be considered intuitive to say that $\mathbf{A}$ and $\mathbf{B}$ are closer than $\mathbf{A}$ and $\mathbf{C}$.

![Figure 3.2.](image)

Figure 3.2.: Geometric loss of the embedding of time into a high-dimensional space. The space of the spike trains discretized into three bins and the points representing $\mathbf{A} = \{100\}$, $\mathbf{B} = \{010\}$, $\mathbf{C} = \{001\}$. While intuitively, $\mathbf{A}$ and $\mathbf{C}$ more distant than $\mathbf{A}$ and $\mathbf{B}$ or $\mathbf{B}$ and $\mathbf{C}$, this is not the case here: $\mathbf{ABC}$ is an equilateral triangle.

As the discretization becomes arbitrarily fine, the dimension of the space goes to infinity. Even worse, the basis of this space is the set of real numbers which is uncountably infinite (this is the famous proof by Georg Cantor in 1874). At this point, it is similar to the instantaneous firing rate. We can thus make a connection with the space of functions, which is just an algebra over a field, where a natural euclidean distance can be defined: the norm-2 on the space of real functions.

However, an approach called kernel methods can be defined that may cure these issues. On the time-axis, we define a kernel, that is a real function. It can be proved that this amounts to a scalar product between an infinite number of dimensions. In Fig. 3.2, it amounts to saying that the scalar product between pairs of axes is not null. In the extreme case, it may be equal to 1, which means that these axes are similar. This is a way of forcing the proximity of some instants of time into the higher-dimensional space. Such a trick is the basis for the Van Rossum distance that is presented below.
3. Structure of the Spike Trains Space

3.1.2. Processing of the spike trains by synapses

We have seen that the space of spike trains is not a vector space, but interestingly, their processing by a further layer of neurons may be seen as a vector space. The membrane potential of the Spike-Response Model is:

\[ V(t) = \sum_i w_i \sum_j \epsilon(t - \hat{t}_j^i) \]  \hspace{1cm} (3.1)

where \( t_j^i \) is the \( j \)-th spike of the \( i \)-th input neuron. This is indeed a linear combination (with coefficients \( w_i \)) of spike trains. We may thus add two spike trains by adding their contributions to the membrane potential of a hypothetical neuron. See the Carnell-Richardson distance explained below.

3.2. Geometry of the space of spike trains

There has been many attempts to define distances on the space of spike trains. They may be classified according to a few features:

- they may be parametric or non-parametric
- they may be single or multi-unit
- they may rely on the “collection of manifolds” view or to the discretized view
- they may result of a filtering of spike trains

When computing distances, we will always consider two spike trains \( s^1 \) and \( s^2 \) with respectively \( N_1 \) and \( N_2 \) spikes. We will analyze the behavior of each distance under perturbations, i.e. when a spike is added or deleted and when a spike is shifted by a small \( \delta \). For a more detailed analysis of the properties and similarities of the distances, I refer to the exhaustive study by Kreuz et al. (Kreuz et al. 2007).


The distance we study most in this manuscript is the one by Victor-Purpura (Victor and Purpura 1996). The basic idea is that the distance between two spike trains is the minimum cost of transforming one into another by deleting, creating and shifting spikes in time. Deleting and creating spikes correspond to jumping from manifold to manifold and shifting spike amounts to moving inside a manifold.

This method finds its roots in a distance between sequences of genes (Sellers 1974). Let us consider two sequences of genes \( \{x_i\}_{i\in[1,m]} \) and \( \{y_i\}_{i\in[1,n]} \) where the possible symbols are the four nucleobases: Adenine \( A \), Guanine \( G \), Thymine \( T \) and Cytosine \( C \). A distance between the sequences can be defined as the minimal number of steps needed to transform one into another. A step may be:

- the deletion of a symbol from a sequence,
### 3.2. Geometry of the space of spike trains

- the transformation of a symbol of a sequence to make it match the other.

In the spike train version of this distance, one can delete or create spikes for a fixed cost of 1 and shift spikes in time for a cost of $c$ per unit time ($c$ being a free parameter of the distance).

![Figure 3.3](image1)

**Figure 3.3.**: Example of the computation of the Victor-Purpura distance between two spike trains.

I gave a few talks where I had to explain this distance to biologists not used to this kind of approach. I found a somewhat intuitive version: we may think of the configuration of tables in a room as our objects, see Fig. 3.4. The distance is the minimum energy it takes to switch from a configuration to another. In order to do this, one can shift tables in space for a cost of $c$ per unit space (rotation is free !) or call a stronger friend to take tables in or out of the room for a fixed cost of 1.

![Figure 3.4](image2)

**Figure 3.4.**: Two configurations of tables in a room.

Let us see what happens for extreme values of the parameter $c$:

- if $c = 0$, then shifting spikes in time is free. The distance is then merely the difference between the number of spikes: $d_{VP}(s_1, s_2) = |N_1 - N_2|$.

- if $c >= 2$, then it is always cheaper to delete a spike and create another one than to shift it in time. The distance is then the sum of the numbers of spikes minus twice the coincidences: $d_{VP}(s_1, s_2) = N_1 + N_2 - M_{12}$.

Let us note that there is some kind of natural time scale appearing. Indeed, as long as two spikes belonging to the two spike trains are separated by less than $\frac{2}{c \text{ms}}$, it is costless to make
them coincide by shifting them in time. However, if they are separated by more than $2/c\text{ms}$, it is cheaper to delete one and create the other. The critical value $2/c\text{ms}$ appears as the natural time scale for which two spikes are still somehow considered identical.

We can however be more precise than this. In order to clarify the relationship between the value of the optimal value of $c$ and the temporal precision of the spike train discrimination analysis, let us consider the cartoon example depicted on Fig. 3.5a, in which two distinct stimuli $A$ and $B$ elicit two distinct spike trains $A_1, A_2$ and $B_1, B_2$, respectively. We can compute the Victor-Purpura distances $d_{VP}$ between all response pairs as the value of $c$ varies, Fig. 3.5b. The cost parameter to be used to compute the temporal precision is the highest cost for which it is still better to shift a spike rather than to delete it and create another one elsewhere. In the example of Fig. 3.5, the intrastimulus distances $d_{VP}(A_1, A_2), d_{VP}(B_1, B_2)$ increase steadily until the cost reaches 2.0, after which they begin to saturate. This means that the temporal precision is $2/2\text{ms} \simeq 1\text{ms}$. On the other hand, the interstimulus distances (i.e., $d_{VP}(A_1, B_1), d_{VP}(A_1, B_2), d_{VP}(A_2, B_1)$, and $d_{VP}(A_2, B_2)$) tend to saturate at a cost of $0.3\text{ms}^{-1}$, i.e. the temporal precision is $2/0.3\text{ms} \simeq 7\text{ms}$. In this case the optimal cost $c$ for effective discrimination would be $0.3\text{ms}^{-1}$ (indeed, the interstimulus distances saturate at this cost, and increasing it would only increase the intrastimulus distances, which would disadvantage the discrimination process), and the temporal precision about $7\text{ms}$. This means that, for discrimination purpose, the best cost is the one that maximizes the ratio interstimulus distance to intrastimulus distance. This idea will be used in the design of our method based on information theory in chapter 7.

A prominent feature of this distance is that is non-euclidean (Aronov and Victor 2004). Indeed, let us consider 4 spike trains with at most 2 spikes. These spike trains thus live in a collection of a 0, 1 and 2 dimensional manifold. The Victor-Purpura distances between these 4 objects suggest a hyperbolic geometry.

If we start with two identical spike trains and add a spike to one of them, the distance becomes equal to 1, wherever the spike is added. If a spike is shifted by $\delta t$, the distance is equal to $c|\delta t|$, which is linear.

This distance was applied with success to the auditory (Huetz et al. 2006) and somatosensory (Saal et al. 2009).

One of the problem with this distance, is that adding or deleting a spike induces a jump between a $n$ and a $n+1$-dimensional manifold. When adding a spike, the point where you start is well-defined, but wherever you place the spike, the cost is the same. It means that somehow, all the points where you may end are equidistant from the point where you start.

The set of these points is thus a non-cyclic, infinite and equidistant to a single point. It is a euphemism to say that it is hard to get a picture of this space.

Another way of exhibiting this odd geometry is to consider two spike trains with one spike each $A = \{t_1\}$ and $B = \{t_2\}$. If we take the Victor-Purpura distance with parameter $c$. If $c|t_2 - t_1| < 2$, then $d_{VP} = c|t_2 - t_1|$. Otherwise, the distance is 2. This means that, once some threshold delay is reached, the distance remains the same. This leads to some odd geometry of the space. Indeed, it is a one-dimensional manifold whose points are all equidistant at large scales. Of course, this threshold delay depends on the parameter. Using a large $c$, the threshold (in terms of delay) is small.

If we were looking for non-parametric measures, we could obtain them from parametric ones by integrating over the parameter. For example, we could define a non-parametric Victor-
3.2. Geometry of the space of spike trains

(a)

Figure 3.5.: (a) Example of two stimuli $A$ and $B$ each eliciting two distinct responses $A_1, A_2$ and $B_1, B_2$, respectively. (b) Calculation of all the VP distances $D_{VP}$ between the response spike trains the VP cost parameter $C_{VP}$ varies.

Purpura distance:

$$d_{VP}^{free} = \int d_{VP}[D_c]$$

(3.2)

We however have to make sure that this integral is well-defined by choosing an appropriate measure. This distance would take into account the contributions of the distances with small and large $c$ and thus would not suffer from the hard-to-picture geometry suggested by the Victor-Purpura distance.

3.2.2. Van Rossum distance (2001)

Another distance was proposed for spike trains by Van Rossum (Van Rossum 2001). The spike trains, considered as distributions (or generalized functions):

$$s(t) = \sum_i \delta(t - \hat{t}_i)$$

(3.3)
3. Structure of the Spike Trains Space

![Diagram showing four spike trains with the numbers 1, 2, and 11.

Figure 3.6.: Examples of four spike trains that cannot be embedded in a 2-dimensional Euclidean manifold. Note that with an appropriate transformation ($x \rightarrow \sqrt{x}$) they could. However, it was proven in (Aronov and Victor 2004) that this is impossible in general.

are first convolved with an exponential kernel so as to obtain a function:

$$f(t) = \frac{1}{\tau} \sum_i H(t - \hat{t}_i)e^{-(t-\hat{t}_i)/\tau} \quad (3.4)$$

The spike trains are thus mapped onto real functions, which is a vector space (it is even an algebra over a field).

Note that this convolution can also be seen as the solution of the differential equation:

$$\dot{f} = -\frac{f}{\tau} + s(t) \quad (3.5)$$

The distance between two spike trains is then taken as the $2$-norm on the space of smooth functions of these two convolved spike trains, see Fig. 3.7:

$$d(s_1, s_2) = \sqrt{\int (f(t) - g(t))^2 dt} \quad (3.6)$$

This distance is null if and only if $s_1$ is exactly similar to $s_2$.

If we start with two identical spike trains and add a spike to one of them, the distance becomes equal to 1, wherever the spike is added. If a spike is shifted by $\delta t$, the distance is equal to $\exp(-\delta t/\tau)$, which is not linear, as opposed to the Victor-Purpura distance.

In opposition to the Victor-Purpura distance that sticks to the fundamental topology of the spike trains space, the Van Rossum distance maps it into a higher-dimensional space. Actually, it is even an uncountably infinite-dimensional space. The distance is thus Euclidean.

The filters clearly reminds of the processing of the input spikes by a Leaky Integrate-and-Fire (LIF) neuron. In this sense, this distance has some biological relevance. However, if there were a threshold for firing, then any two spike trains leading only to subthreshold potential should be treated as similar and thus have a distance equal to zero. This is not the case with this distance.
3.2. Geometry of the space of spike trains

Figure 3.7.: Two spikes trains filtered with a rectified exponential and their instantaneous difference. The distance is just the integral of the instantaneous difference over time.

Note that the two distances described so far scale with the number of spikes. The distance between two patterns is half the distance between these patterns repeated twice, see Fig. 3.8. In this sense, these distances are local, they do not take into account the global number of spikes or any global properties of the spike trains.

3.2.3. Houghton distance (2009)

In (Houghton 2009), Houghton carried further this analogy between the kernel used by Van Rossum and the processing of spikes by a neuron. He even added some short-term depression in the map from spike trains to real functions. Instead of convolving each spike with a kernel \( \frac{1}{\tau} H(t - t_i)e^{-\frac{(t-t_i)}{\tau}} \) with a fixed amplitude, he modulated the kernel by a synaptic function.
3. Structure of the Spike Trains Space

satisfying the differential equation:
\[ \dot{g} = -\frac{g}{\tau_g} \]  

but that, in addition undergoes a jump each time a spike is received:
\[ g(t^+) = g(t) + \delta g \frac{g_{\text{max}} - g}{g_{\text{max}}} \]  

The function of the spike train is then obtained by:
\[ f(t^+) = (1 - \mu) f(t) + 1 \]

Instead of obtaining a function reminiscent of the membrane potential time-course of a standard LIF, he obtains that of a LIF with short-term depression, which is much more biologically relevant. This has very interesting computational properties, that Houghton expressed very nicely:

*If a spike train is changed by adding a single spike, the distance between the new and the old spike train will be the same under the van Rossum metric, no matter where the spike is added. For the new metric, however, this distance will depend on the precise location of the extra spike. Note that the same remark holds for the Victor-Purpura distance.*

In the same line of thought, if a spike is slightly shifted, the distance depends on the surrounding spikes. In case, there are no spikes around, it yields the same distance as the Van Rossum one. However, if there are many spikes around, then shifting a spike influences the level of synaptic depression and thus the distance.

Just like the Van Rossum distance, however, it is subject to the criticism that is does not take into account a threshold for firing that is yet, a fundamental property of neurons. It is thus unable to deal with irrelevant spontaneous activity.

3.2.4. Schreiber correlation (2003)

In (Schreiber et al. 2003), a correlation-based measure of reliability was proposed. As with the Van Rossum metric, the first task consists in convolving the spike train with a filter. In the original article, the filter is taken as a gaussian function. A pairwise correlation is then computed between \( s_1 \) and \( s_2 \) as:
\[
C_{\text{Sch}}(s_1, s_2) = \frac{\tilde{s}_1 \cdot \tilde{s}_2}{|s_1||s_2|} = \frac{\int s_1(t)s_2(t)dt}{\sqrt{\int s_1^2(t)dt} \sqrt{\int s_2^2(t)dt}}
\]  

An example is depicted on Fig. 3.9. This quantity is equal to 1 if and only if \( s_1 \) and \( s_2 \) are exactly identical. It goes to zero as their difference increases. If the width \( \sigma \) of the gaussian kernel goes to zero, the correlation is the number of coincident spikes over the product of the spike counts. If \( \sigma \) goes to infinity, the correlation goes to 1.

Let us start with two identical spike trains and add a single spike far from the other ones to one of the spike trains. While the numerator is left unchanged, the denominator increases. The correlation is then decreased. Interestingly, it decreases less when there are already many spikes. For example, if there was a single spike beforehand, the addition of a spike decreases
3.2. Geometry of the space of spike trains

![Two spike trains filtered with a rectified exponential and their instantaneous difference. The correlation is just the integral of the instantaneous product over time.](image)

Figure 3.9.: Two spike trains filtered with a rectified exponential and their instantaneous difference. The correlation is just the integral of the instantaneous product over time.

the correlation to $\frac{1}{\sqrt{2}}$. However, if there were 10 spikes beforehand, the correlation after the insertion the spike is $\sqrt{\frac{10}{11}}$. As opposed to the Victor-Purpura or Van Rossum distance, this measure thus assesses perturbations with respect to the global number of spikes. It is thus linked with some signal-to-noise ratio.

If, instead of adding spikes, we shift one of them by $\delta t$, then its contribution to the correlation falls from 1 to $\frac{1}{\sqrt{2}}e^{-\frac{\delta t^2}{4\sigma^2}}$. Again, this contribution is all the smaller as the number of spikes is larger.

To estimate the average reliability of the response of a neuron to a single stimulus, a measure of reliability is defined as the average of the pairwise correlations:

$$R_{corr} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\vec{s}_i \cdot \vec{s}_j}{||\vec{s}_i|| \cdot ||\vec{s}_j||}$$  \hspace{1cm} (3.11)

3.2.5. Quian Quiroga’s Event Synchronization (2002)

Another measure was defined (Quian Quiroga et al. 2002) to quantify the level of synchrony between two spike trains by assessing the pairwise synchrony between individual spikes. The synchrony between spikes is assessed with respect to the surrounding interspike intervals. For any couple of spikes drawn from two spike trains, one first defines a temporal scale for synchrony $\tau_{ij}$:

$$\tau_{ij} = \min \{t_{i+1}^1 - t_i^1, t_i^1 - t_{i-1}^1, t_{j+1}^2 - t_j^2, t_j^2 - t_{j-1}^2\} / 2$$  \hspace{1cm} (3.12)

This is half the minimum interspike interval (ISI) before or after one of the spikes. A function estimating the number of times a spike appears in $s_1$ shortly after a spike in $s_2$ is then defined
3. Structure of the Spike Trains Space

as:

\[ c_{12} = \sum_i \sum_j J_{ij} \quad (3.13) \]

where

\[
J_{ij} = 1 \quad \text{if} \quad 0 < t_1^i - t_2^i < \tau_{ij} \quad (3.14)
\]

\[
J_{ij} = \frac{1}{2} \quad \text{if} \quad t_1^i = t_2^i \quad (3.15)
\]

\[
J_{ij} = 0 \quad \text{else} \quad (3.16)
\]

Note that for a given \( i \), only one \( J_{ij} \) can be non-null. Basically, for all couples of spikes, it amounts to determine whether they may be seen as the same spike slightly shifted in time.

The event synchronization is then defined as:

\[
Q = \frac{c_{12} + c_{21}}{\sqrt{N_1 N_2}} \quad (3.17)
\]

Figure 3.10.: Event synchronization between two spike trains. All the strictly positive terms are given, plus a few null terms. All spikes, except the third one of the first train, are grouped in pairs yielding a total of 1 (either \( 1 + 0 \) or \( 1/2 + 1/2 \)).

When two spike trains are exactly identical, their event synchronization is equal to 1 (all terms are equal to 1/2). When two spike trains are shifted by less than the typical interspike interval, the event synchronization is also equal to 1 (half the terms are equal to 1).

Let us consider two identical spike trains. If we shift one of the spikes of one spike train by less than half the minimum surrounding ISI, the event synchrony remains equal to 1. If it is shifted by more than half the minimum surrounding ISI, then the event synchrony decreases since some spikes are no longer grouped in pairs.

If, instead of shifting spikes in time, a spike is added, then the event synchrony decreases. Indeed, either this new spike does not find its pair, either it finds it by destroying an existing pair. The numerator remains the same, but the denominator increases. As Schreiber’s correlation, the level of the change depends on the total number of spikes, thus adding a spike to two identical spike trains with lots of spikes does not significantly decrease the event synchrony.

Note that this distance is non-parametric, it does not need a predefined time-constant. The reason is that it is self-adaptive, the relevant time constant is assessed by the average activity in a natural time-window (current or nearest interspike intervals). A short discussion on the relative merits of parametric and non-parametric measures is postponed at the end of this chapter.

A somewhat similar measure was proposed in (Hunter and Milton 2003), though it is parametric. For each spike in $s^1$, we look for the nearest spike in $s^2$ that we call $n_{12}^i$. We then define a weight that depends on the delay between them $r_{12}^i = e^{-|t_1^i - n_{12}^i|/\tau}$. The same is done for the spikes of $s^1$. The similarity is then defined as:

$$S(s^1, s^2) = \frac{<r_{12}^i> + <r_{21}^i>}{2}$$

(3.18)

where $\tau$ is a predefined parameter.

Just like Schreiber’s correlation and event synchrony, it is equal to 1 if $s^1$ and $s^2$ are strictly identical.

A problem with this distance is that the nearest spike in $s^2$ may be the same for two different spikes in $s^1$ while another spike in $s^2$ will never be the nearest spike. This makes the idea of synchrony a little hard to interpret here.

3.2.7. Carnell-Richardson (2005)

In 2005, a conceptually very different measure was proposed in (Carnell and Richardson 2005). In order to define this scalar product, let us first define the weighted spike trains of $s^1$ and $s^2$:

$$f^1(t) = \sum_i c_i \delta(t - t_1^i)$$

(3.19)

$$f^2(t) = \sum_j d_j \delta(t - t_2^j)$$

(3.20)

These functions live in an uncountably infinite-dimensional space. It is reminiscent of the binning procedure when the number of bins goes to infinity. However, here, not only the values 0 and 1 are relevant. Instead, all real values are. At each time, there is a 1-dimensional vector space describing the magnitude of the spike. In addition, spikes in a single train may have different weights.

We can then define a scalar product:

$$<f^1(t) | f^2(t)> = \sum_{i,j} c_i d_j e^{(|t_1^i - t_2^j|)/\tau}$$

(3.21)
3. Structure of the Spike Trains Space

There is again a time constant appearing in this definition.

When we consider two identical spike trains and shift a spike in time by \( \delta t \), the scalar product decreases by an amount \( c_i d_j (1 - e^{\delta t/\tau}) \). On the other hand, if we add a spike, the Carnell-Richardson measure always increases since a new positive term is added. This is one of the main problems is that adding spikes will always increase the scalar product. For example, let us take two copies of a spike train \( s = \{10, 20, 30\}\) ms, a third and fourth copy with more spikes \( s' = \{10, 15, 20, 25, 30\}\), \( s'' = \{10, 20, 30, 110, 120, 130\}\).

- if \( \tau \lesssim 20 \), we have \( <s|s> = <s|s''> \),
- if \( \tau \gtrsim 1 \), we have \( <s|s>> <s|s'> > \).

These two results are very counter-intuitive.


Kreuz et al. (Kreuz et al. 2007) proposed a non-parametric distance between spike trains they called ISI-distance. They assigned at each instant the value of the current Inter-Spike Interval (ISI).

\[
f(t) = \min_i \{ t_i; t_i > t \} - \max_i \{ t_i; t_i < t \}
\]  

(3.22)

This technique is applied to two spike trains to obtain two functions \( f \) and \( g \). Then we define a new function \( h \) as:

\[
h(t) = \frac{f(t)}{g(t)} - 1 \text{if } f(t) \leq g(t) \\
h(t) = 1 - \frac{g(t)}{f(t)} \text{if } f(t) > g(t)
\]  

(3.23) (3.24)

The distance is then 1-norm of the ratio function \( h \):

\[
D_K(s_1, s_2) = \int |h(t)| dt
\]  

(3.25)

The effects of the shifting of spikes in time or of the insertion of spikes are highly dependent on the statistics of the surrounding interspike intervals. A discussion of these effects would require a long and detailed analysis that we will not carry here. We will however mention the fact that this distance suffers from side-effects. Indeed, for short spike trains, the distance highly depends on the choice of origin and end of the time axis.

Just like event synchrony, this distance is self-adaptative, it does not need a predefined parameter.

3.2.9. Multi-unit distances

In general, when we have a distance between the spike train elicited by an axon, we compute the distance on the population by:

\[
D_{pop}(s^1, s^2) = (\sum_i d_i(s^1, s^2)^z)^{1/z}
\]  

(3.26)
3.2. Geometry of the space of spike trains

Figure 3.12.: Two spikes trains filtered with their instantaneous ISI. The distance is just the integral of the absolute value of ratio function over time.

However, the identity of the neuron that fired a spike may not be relevant. Thus, multi-unit distances have been defined to account for this.

**Aronov-Victor distance (2003)**

Aronov *et al.* (Aronov *et al.* 2003) proposed an extension of the Victor-Purpura for many units, as well as an algorithm to compute it quickly (Aronov 2003). In addition to the three operations authorized (creation, deletion, shift in time) in the single-unit distance, a shift in space of the neurons is allowed. The responses of neurons are treated as labeled lines, each neuron having its own label. Shifting a spike from a neuron to another (though keeping the same time) can be done at a cost $k$. This parameter represents the importance of knowing the label of the spike is introduced. If the parameter $k$ is equal to 0, then changing spikes from neuron to neuron is free. It amounts to consider that all the spikes are generated by a single axon. On the other extreme, if $k$ is large (larger than 2), then it will always be costless to delete a spike on one neuron and create it on another than to change its label. The distance induced would be the same as eq.3.26 with $z = 1$. This distance may be very helpful in determining whether we should consider a population code with all neurons equivalent or by taking into account their individuality (Schneidman *et al.* 2000). It may also give insights on the optimal connectivity of the neurons with a further layer.

**Multi-unit Van Rossum distance (2008)**

An extension of the Van Rossum distance to the case of multiple neurons has been proposed in (Houghton and Sen 2008). Let us sketch the idea for two neurons. The spike trains in
response to a first stimulation are filtered as in Van Rossum distance, let us call them \( f_1 \) and \( g_1 \). The filtered spike trains in response to a second stimulus are called \( f_2 \) and \( g_2 \). We then consider two unit vectors \( e_0 = (1, 0) \) and \( e_\theta = (\cos \theta, \sin \theta) \). With this, we define two vectors \( r_1 = f_1 \cdot e_0 + g_1 \cdot e_\theta \) and \( r_2 = f_2 \cdot e_0 + g_2 \cdot e_\theta \). The distance is taken as the 2-norm between these vectors:

\[
d = \| r_1 - r_2 \|_2
\] (3.27)

When \( \theta = 0 \), this boils down to the distance between the whole population of spikes, whatever the neurons that fired them. It amounts to grouping all the spikes of a population on a single axon and take the Van Rossum distance. At the opposite, when \( \theta = \frac{\pi}{2} \), this is the labeled line distance, i.e. where the identity of the neuron that fired a spike is taken into account and no interactions between these neurons is considered. In the case of the norm-2, this is

\[
D_{pop}(s^1, s^2) = \sqrt{d_1^2(s^1, s^2) + d_2^2(s^1, s^2)}
\] (see Eq. 3.26 with \( z = 2 \)), where the \( d_i \)'s are the distance between two stimuli for the \( i \)-th neuron.

Note that the angle \( \theta \) is akin to the parameter \( k \) in the Aronov distance.

### 3.3. Discussion

These distance are relying on various approaches to the spike train space (collection of manifolds, higher-dimensional spaces, etc...), they may focus on the relative timing or on the precise synchrony of spikes, they may be causal or not, they may be parametric or not.

I attended a workshop on spike-train metrics in Plymouth in 2010. During the discussion session, someone asks whether we ought to be happy with the distances we have so far. It really seems we have an embarrassment of riches. We have distances focusing on many different features of the spike trains. If we want to show that stimuli are well discriminated, we can try all these distances and find the more appropriate one.

A confusion of aims and a perfection of means seems to be our main problem.

Albert Einstein

However, one of the fundamental question we may ask is: do we want to remain as artificial observers of the nervous system? \(^1\) Is our goal finding the most artificial distance that allows a nice discrimination of the inputs without taking into account its possible implementation by neurons? This would show that there is irrelevant information, i.e. information that is here but cannot be read out because of the biological constraints on the neurons. We should define instead distances that are biologically plausible. The work by Conor Houghton where he implemented short-term depression on the Van Rossum distance clearly goes in this direction: trying to define distances that are interpretable in the language of neuron parameters.

Note that an additional problem with non-parametric measures is that you cannot make predictions on the parameters of the neurons. Indeed, if the task of the nervous system is represented by some objective function, then it is interesting to find the parameters of the distance that maximizes it. It then allows to make an testable prediction on the parameters of

\(^1\)in a first version of this manuscript, I wrote Copenhagen observers. Indeed, the idea to implement distances that have no straightforward interpretation in terms of neuronal features reminds me of the introduction of a classical observer in a quantum realm. Though it may help understanding what goes on, it cannot be taken as a satisfying picture of the phenomena at stake.
the neurons. However, a non-parametric distance cannot lead to such predictions, and even if this distance accounts for the outcomes of psychophysical experiments, we are left with wondering why neurons are tuned the way they are.
The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

Claude Shannon

Shannon Information Theory

Where we recall the basics of Shannon information theory. In particular, we put forward the fact that Shannon entropy is a topological measure of the uncertainty about the events. Then we review existing methods to estimate the information transfer capacity of neural systems. In particular, we recall standard methods in neuroscience to quantify the information contained in spike trains, such as the binning procedure or the spike train metrical analysis. We also discuss methods based on rate-distortion theory to estimate the extent to which stimulations can be reconstructed for a given level of information transfer.

In his seminal paper of 1948 (Shannon 1948) proposed a basic scheme of a communication system Fig. 4.1.

![Basic scheme of a communication system](image)

Figure 4.1.: Basic scheme of a communication system. Taken from Shannon 1948 (Shannon 1948).

On this scheme, he identifies an information source, a transmitter, a transmission channel, a receiver and a destination. If we can obviously find biological analogues of some elements (such as the information source, the transmitter or the receiver), it is not conspicuous to find one for the destination. Indeed, Shannon gives the receiver the role of “performing the inverse operation as that performed by the transmitter” (Shannon 1948). It is far from obvious that the
4. Shannon Information Theory

central nervous system performs such a perfect reconstruction of the stimulus. As stated in the introduction of this manuscript, the neurons that receive the signals introduce a bias: they only focus on the relevant features of the code. The relevance being a function of the state of the nervous system and of the task to be performed.

Shannon information theory provides a mathematical framework to characterize the input-output relationship of probabilistic communication systems (Shannon 1948; Cover and Thomas 1991). Neural information processing can be understood as a multi-stage transmission process in which neurons, as well as neural populations, are considered as stochastic communication channels. The reliability of neural encoding/decoding mechanisms can then be analyzed in terms of information content (MacKay and McCulloch 1952; Bialek et al. 1991; Deco and Obradovic 1997; Rieke et al. 1997; Borst and Theunissen 1999).

The information theoretical approach offers a suitable vantage point to compute a comprehensive estimate of the knowledge encoded in neural responses. In particular, Shannon Mutual Information (MI) provides an upper-bound of the knowledge that can be extracted from neural responses by an ideal observer, which produces a benchmark to assess the fraction of the overall information actually captured by decoding algorithms (Borst and Theunissen 1999; Quiang Quiroga and Panzeri 2009). It provides a non-parametric measure of the statistical dependence between stimulus and neural response, even in the presence of non-linear dynamics and complex input-output transformations (London et al. 2002; Fuhrmann et al. 2004; Chacron 2006). For instance, it enables to quantify to what extent specific properties and regularities of the stimuli, e.g. spatio-temporal correlations, are functionally relevant to neurotransmission.

It is a measure of the correlation between the variability of the stimulus and that of the responses. In a nutshell, if the variability of the response to a single stimulus is low with respect to its marginal variability, the response is informative.

4.1. A crash course on information theory

Let us consider two sets of events $X = \{x\}$ and $Y = \{y\}$, each event having a probability $p(x)$ or $p(y)$ to occur. Let us recall a few formula of few formula of probability theory. First, we require the events to be mutually exclusive and to form a partition of the universe, this can be formalized by $\sum_x p(x) = 1$. Second, the joint probability of an event $x$ and an event $y$ is $p(x, y) = p(x)p(x|y) = p(y)p(y|x)$. The last equality is equivalent to Bayes’ formula. Finally, we recall that for some arbitrary function $A(x)$ that does not depend on $y$, then $\sum_x p(x)A(x) = \sum_{x,y} p(x, y)A(x)$.

Let us now attribute the set $X$ an entropy $H(X)$ that measures the uncertainty on the event that will happen next:

$$H(X) = -\sum_x p(x) \log p(x) \quad (4.1)$$

For example, if we have a set of $N$ events (i.e. $Card(X) = N$), then the entropy is maximal when all the events are equiprobable, i.e. with probability $p(x) = \frac{1}{N}$. The entropy is then equal to $H(X) = \sum_n \frac{1}{N} \log \frac{1}{N} = \log N$. This is a monotonically increasing function of $N$. Thus the higher the number of equiprobable events, the higher the uncertainty and the higher the entropy.

Note already that the entropy does not depend on the states $x$, only on their probabilities.
Two distributions of different events, but whose probability distributions are equal have the same entropy.

When two sets are considered \( X = \{ x \} \) and \( Y = \{ y \} \), we define a joint entropy:

\[
H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y) \tag{4.2}
\]

where \( p(x, y) \) is the probability that \( x \) and \( y \) occur. Let us show that the joint entropy is always lower than the sum of the two individual entropies:

\[
H(X) = - \sum_x p(x) \log p(x) = - \sum_{x,y} p(x, y) \log p(x) \\
H(Y) = - \sum_y p(y) \log p(y) = - \sum_{x,y} p(x, y) \log p(y) \\
H(X) + H(Y) = - \sum_{x,y} p(x, y) \log p(x)p(y)
\]

Therefore:

\[
[H(X) + H(Y)] - H(X, Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{4.3}
\]

The right-hand side is in general non-zero. It is only when \( p(x, y) = p(x)p(y) \) or equivalently \( p(y|x) = p(y) \), i.e. when the events \( X \) and \( Y \) are independent (the knowledge of one of them does not help on the second), case in which the joint entropy is the sum of the individual entropies. The right-hand side gives the decrease in entropy due to some correlation between \( X \) and \( Y \).

We now define the conditional entropy (or noise entropy or equivocation). This is the expectation value of the entropy of \( Y \) for each value of \( x \) weighted by the probability of getting \( x \).

\[
H(Y|X) = - \sum_x p(x) \sum_y p(y|x) \log p(y|x) \tag{4.4}
\]

that can be written:

\[
H(Y|X) = - \sum_{x,y} p(y, x) \log p(y|x) \tag{4.5}
\]

We can now decompose the joint entropy in:
4. Shannon Information Theory

\[
H(X,Y) = -\sum_{x,y} p(x, y) \log p(x, y)
\]
\[
H(X,Y) = -\sum_{x,y} p(y)p(x|y) \log p(y)p(x|y)
\]
\[
H(X,Y) = -\sum_{y} p(y) \log p(y) \sum_{x} p(x|y) - \sum_{y} p(y) \sum_{x} p(x|y) \log p(x|y)
\]
\[
H(X,Y) = H(Y) + H(X|Y)
\]

We get back the previous result: the joint entropy of two sets is the sum of the individual entropies if and only if the events are independent, i.e. if \( H(X|Y) = H(X) \).

\[
H(X) - H(X|Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]

(4.6)

\[
H(Y) - H(Y|X) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]

(4.7)

We can now define the mutual information between \( X \) and \( Y \) as the reduction in uncertainty about \( x \) obtained from the knowledge of \( y \). This can then be formalized as:

- it is the difference between the uncertainty on \( X \) and that on \( X \) knowing \( Y \): \( I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \),

- it is the difference between \( H(X) + H(Y) \) and \( H(X,Y) \).

Let us take a simple example; Alice chooses a number in the set \( \{0,1\} \). If she thinks 0, she tells Bob the letter A, if she thinks 1, she tells him the letter B. Before Alice tells anything to Bob, he does not know what Alice thinks. 0 and 1 have a priori the same probability \( p = 0.5 \), the entropy is therefore \( H(\{0,1\}) = \log 2 \). As soon as Alice tells Bob a letter, he knows the number she was thinking, because \( H(\{0,1\}|A) = 0 \) and \( H(\{0,1\}|B) = 0 \). And then \( H(\{0,1\}|\{A,B\}) = 0 \). The signal Alice sends allows Bob to reconstruct the number Alice thinks unambiguously. The Mutual Information is thus equal to \( I(\{0,1\}; \{A,B\}) = \log 2 \).

Note that the same value of the Mutual Information may correspond to a lot of different cases. If, instead of telling Bob a letter, Alice tells him a random even number between 1 and 100 when she thinks 0, and a random odd number between 1 and 100 when she thinks 1. Then, although there is no longer a one-to-one map from the number Alice thinks and the one she tells Bob, the message still permits an unambiguous reconstruction of the number Alice thinks of. The Mutual Information is \( \log 2 \). The difference is in the entropy of the number she tells conditioned on the number she thinks, which is here equal to \( \log 50 \). In general, in Information Theory, not too much attention is paid to this conditional entropy.

The information is said mutual because it is symmetric \( I(X;Y) = I(Y;X) \). \( X \) brings as much information on \( Y \) as \( Y \) brings on \( X \). This measure is based on the fact that the knowledge
of $Y$ turns the marginal probability distribution $p(x)$ in a conditional probability distribution $p(x|y)$.

We can also write:

$$I(X;Y) = -\sum_x p(x) \log p(x) + \sum_y p(y) \sum_x p(x|y) \log p(x|y)$$

$$I(X;Y) = -\sum_{x,y} p(x|y)p(y) \log p(x) + \sum_{y} p(y) \sum_{x} p(x|y) \log p(x|y)$$

$$I(X;Y) = \sum_{x,y} p(x|y)p(y) \log \frac{p(x|y)}{p(x)}$$

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Note that the problem of decoding is unambiguous if $H(S|R) = 0$, i.e., if for a given set of stimulations, the information is at its highest value $I(R;S) = H(S)$, and this, whatever the value of $H(R|S)$.

It is in fact the Kullback-Leibler divergence or relative entropy (which is a pseudo-distance since it does not meet the symmetry requirement) between the marginal entropy $H(R)$ and the conditional entropy $H(R|S)$:

$$D(X,Y) = \sum_{x,y} p(x) \log \frac{p(x)}{p(y)}$$

One of the drawbacks of the mutual information is that it is an expectation value. It just measures on average how much information one gains by having access to the value of $X$ or $Y$. A few information measures were designed to know how much information is gained at each observation.

### 4.2. Shannon Information Theory in neuroscience

A large body of work has been done to assess the capacity of single neurons, as well as neural populations, to process afferent signals and transmit meaningful accounts of their inputs to downstream neural stages (seen (Quian Quiroga and Panzeri 2009) for a recent review). To a large extent, these studies have provided qualitative analyzes of the neural communication process based on Shannon information theory principles (Shannon 1948; Cover and Thomas 1991; Bialek et al. 1991; Deco and Obradovic 1997; Rieke et al. 1997; Borst and Theunissen 1999).

Information theoretical approaches have been used to characterize neurotransmission at different organization levels, from single synapses (Zador 1998; Manwani and Koch 2001; Tiesinga 2001; Fuhrmann et al. 2002; London et al. 2002; Manwani et al. 2002; Goldman 2004), to single cell responses (Theunissen and Miller 1991; De Ruyter van Steveninck et al. 1997; Borst and Theunissen 1999; Tiesinga et al. 2002; Butts and Goldman 2006; Sharpee et al. 2006; Arleo et al. 2010), to neural population activity (Brenner et al. 2000; Reich et al. 2001a; Lu and Wang 2004; Smith and Lewicki 2006; Saal et al. 2009; Quian Quiroga and...
4. Shannon Information Theory

Panzeri 2009). Also, the relevance of stimulus-specific information transmission measures (see e.g. (DeWeese and Meister 1999) for a review) has been demonstrated by numerous studies on sensory neural processing (Richmond and Optican 1990; Buracas et al. 1998; Panzeri et al. 2001; Paz and Vaadia 2004; Rainer et al. 2004; Einhauser et al. 2007). Finally, the relation between information theoretical analysis and more explicit decoding schemes to infer the most likely stimulus that elicited an observed response (e.g. Bayesian algorithms, k-nearest neighbor decoders, population vector approaches; see reviews by (Rieke et al. 1997; Brunel and Nadal 1998; Borst and Theunissen 1999; Pouget et al. 2000; Dayan and Abbott 2001)) has began to be elucidated in neuroscience (Thomson and Kristan 2005; Victor and Nirenberg 2008; Quian Quiroga and Panzeri 2009).

In information theory, the average amount of information that can be extracted from the neural responses \( r \in R \) to the inputs \( s \in S \) (both considered as probabilistic variables) is estimated by the Mutual Information (MI) measure:

\[
I(R; S) = H(R) - H(R|S) \tag{4.9}
\]

where \( H(R) \) is the marginal entropy in the output events and quantifies the intrinsic variability of the response space:

\[
H(R) = - \sum_{r \in R} p(r) \log p(r) \tag{4.10}
\]

and \( H(R|S) \) is the conditional entropy in the response space given the input:

\[
H(R|S) = - \sum_{s \in S} p(s) \sum_{r \in R} p(r|s) \log p(r|s) \tag{4.11}
\]

Mutual information can then be rewritten as:

\[
I(R; S) = \sum_{r,s} p(r, s) \log \left( \frac{p(r, s)}{p(r)p(s)} \right) \tag{4.12}
\]

It estimates the difference between signal and noise entropy by measuring how much one can learn about the stimulus by observing the neural responses (or vice-versa). \( I(R; S) \) is zero if stimuli \( s \) and responses \( r \) are completely uncorrelated, otherwise \( I(R; S) > 0 \). For a given set of stimuli, Mutual Information is maximal if-and-only-if no response \( r \) is elicited by two different stimuli \( s \), i.e. if there is no ambiguity when reconstructing the input by observing the output (i.e. \( H(S|R) = 0 \)).

4.3. Fisher information

Fisher information (Blahut 1988; Clarke and Barron 1990; Cover and Thomas 1991; Rissanen 1996; Brunel and Nadal 1998) quantifies the precision of a stimulus reconstruction process in the presence of a continuous input state space. It is defined as:

\[
FI(s) = \sum_r p(r|s) \left[ \frac{\partial \ln p(r|s)}{\partial s} \right]^2 \tag{4.13}
\]
4.4. Event space discretization

When the event state space is considered discrete by the readout system, the probability distributions (i.e. input a priori probability \( p(s) \), output marginal probability \( p(r) \), conditional and joint probabilities \( p(r|s) \) and \( p(r,s) \), respectively) can be estimated empirically on the basis of natural equivalence classes (e.g. spike counts). In the case of continuous event spaces (e.g. spike time), the equivalence classes will depend on the precision of the discretization process used to partitioning the continuous state space. For instance, spike trains can be processed by means of a binning procedure that maps them into binary words (Fig. 7.1a) (Strong et al. 1998; Dimitrov and Miller 2000; London et al. 2002), which preserves the information about spike timing up to a certain precision (e.g. few milliseconds). Therefore, the binary word coding can be used to relax the temporal resolution constraint (see (Panzeri et al. 2007) for a review) and to allow equivalence classes (with arbitrary borders) to be defined by reducing the dimensionality of the input-output space. Once the equivalence classes have been defined, both Shannon entropy and MI can be computed (Eqs. 4.9-4.12).

When neural responses are processed according to spike count decoding, two spike trains \( r \) and \( r' \) are considered equal if-and-only-if they have the same number of spikes \( N \), i.e. by using the following similarity measure \( \phi(r, r') \):

\[
\phi(r, r') = \begin{cases} 
  1 & \text{if } N(r) = N(r') \\
  0 & \text{otherwise}
\end{cases}
\] (4.14)

When the neural decoder accounts for the timing \( t_i \) of the spikes, the similarity measure can be written as:

\[
\phi(r, r') = \begin{cases} 
  1 & \text{if } \forall i, \ t_i(r) \sim t_i(r') \\
  0 & \text{otherwise}
\end{cases}
\] (4.15)

This measure obviously depends on the temporal resolution taken to determine when \( t_i(r) \) and \( t_i(r') \) can be considered similar (i.e. \( t_i(r) \sim t_i(r') \)). The more severe the similarity measure (e.g. with a temporal resolution of 1 ms), the more precise the decoding process, and the smaller the information loss when processing the neural responses. Yet, precise temporal
4. Shannon Information Theory

decoding makes it computationally intractable to estimate reliably the Shannon entropy is large, or in the presence of large neural populations (i.e. curse of dimensionality, (Borst and Theunissen 1999; Quiroga and Panzeri 2009)) because of a dramatic under-sampling of the response space.

This discretization procedure was used on responses from the cricket cercal system to determine natural timescales for encoding (Dimitrov and Miller 2000) by finding the optimal cost for transmitting information. In essence, the information was computed for various bin sizes. For all bin sizes between a certain threshold (0.1 ms in the study), the information was constant, but for larger bins, the information started to decrease significantly.

Dimitrov et al. (Dimitrov et al. 2001) proposed to use a non-uniform discretization of time. They computed for various possible discretizations, the amount of information lost. This leads to large bins where the spikes are not precise and to small bins where the exact timing of spikes is highly relevant for information transmission. Thus, it allows to determine at each instant, the extent to which we should take into account the precision of spikes. It also allows to circumvent the problem of the computability of the information, by using coarse descriptions of periods of irrelevant activity.

It has to be noticed that binary word coding guarantees the transitivity of identity, i.e. \( r \sim r' \) and \( r' \sim r'' \) \( \Rightarrow \) \( r \sim r'' \). However, the binning procedure is prone to local anisotropy effects on the similarity measurements (Fig. 7.1a). For example, given a partitioning based on time bins of \( \Delta t = 5 \text{ ms} \), two spikes separated by 4 ms can be considered identical, whereas two spikes separated by 1 ms may be considered different.

4.5. RDT, Information Bottleneck, Optimal state space quantization

Rate distortion theory (RDT) (Cover and Thomas 1991) has been extensively used to study neural encoding/decoding principles. For a system that cannot transmit more than a given information \( I \) per unit of time (its rate), What is the inevitable level of distortion \( D(I) \). Or conversely, to determine the minimum amount of information \( I \) to transfer to maintain a small error on the reconstruction of the stimulus (the distortion) remains low \( I(D) \).

For example, let us consider a picture with \( N \) pixels that are either black or white. If Alice has to tell Bob everything about that picture, but can only communicate through a channel that allows a maximal rate of \( x \) bits, then if \( x < \log N \), there will be an error on Bob’s reconstruction of Alice’s picture. This error measure we call the distortion. Thus, for a given measure of distortion, how does its value depend on \( x \)? Conversely, in order to remain below some level of distortion, what is the minimum amount of information to transmit?

Rate distortion theory has many applications in compression techniques, for example video or audio formats. It is formalized the following way:

\[
\inf_{Q_{Y|X}(y|x)} I_Q(Y; X) \text{ with } D_Q \leq D^* \tag{4.16}
\]

In general, if the information versus the distortion is a convex function, we may be able to lower the information to be transmitted by allowing only slight errors on the message. This may be fundamental in neuroscience. If we do not mind if some distortion appears, it allows
4.6. Relevant Component Analysis

to decrease the amount of information transmitted: the lower the information, the shorter the message and thus the lowest the amount of computational resources needed.

When dealing with neural transmission, it is formulated slightly differently. The goal is to find a compression (discretization) of the responses such that the information about the stimuli is preserved. The compression keeps the features of the code that are relevant for information transmission and discard the irrelevant ones. More technically, it amounts to finding an optimal quantization (partitioning into equivalence classes, see Section 4.4) of the response space with respect to a distortion function.

• In (Tishby et al. 1999) and (Schneidman et al. 2002) the cost function to be minimized is \( \mathcal{L}(p(\tilde{r}|r)) = I(\tilde{R}; R) - \beta I(\tilde{R}; S) \), where \( \tilde{R} \) is a compressed version of \( R \). The goal is to find the quantization \( \tilde{R} \) capturing the most information about \( S \) while keeping its information with \( R \) the lowest possible. It amounts to discarding the unnecessary variability of the responses. The factor \( \beta \) determines the trade-off between the quality of the transmission and the compression level. If \( \beta = 0 \), all the \( r \) should be mapped to the same tilder while, if \( \beta \to \infty \), the quantization has to be arbitrarily fine.

• Another approach (Dimitrov et al. 2001; Samengo; Dimitrov et al. 2003) consists in finding the optimal state space discretization \( \tilde{R} \) or \( R \) with respect to a distortion function \( D_I(R, \tilde{R}) = I(S; R) - I(S; \tilde{R}) \) with the same notations as above. This function measures the discrepancy in information transmission with and without a quantization of the responses in \( N \) equivalence classes. The optimal quantization is the one that maximizes \( H(\tilde{R}|R) \) (it means that it makes no further assumption on the code) while keeping the distortion function below some fixed threshold \( D \).

4.6. Relevant Component Analysis

Another approach that I find highly relevant for the problem of neural coding is that by Shental et al. (Shental et al. 2002). They argue that Principal Component Analysis (PCA) may be flawed by the fact that directions of high variability may be irrelevant with respect to information transmission (see fig.4.2). The direction that is the most relevant is the one that, in the spirit of Fisher Linear Discriminant, maximizes the discriminability, i.e. the distance between the mean of the clusters over the variance of the clusters.

4.7. Spike train metrical analysis

A solution to avoid the discretization of the event space for the calculation of \( I(R; S) \) is to account for spike train metrics (Victor and Purpura 1996; Van Rossum 2001; Aronov et al. 2003; Schreiber et al. 2003; Schrauwen and Van Campenhout 2007; Houghton and Sen 2008). For instance, a well-established approach consists in estimating the Victor-Purpura (VP) distances between the neural responses (Victor and Purpura 1996) and classifying them according to a clustering method such as the \( k \)-nearest neighbor algorithm (Duda et al. 2001). That is, the response is assigned the class that constitutes the majority among its \( k \)-nearest neighbor responses (the positive-definite integer \( k \) being a free parameter of the method). This classification allows the so-called confusion matrix \( C_{ij} \) to be computed, whose terms are the probability...
4. Shannon Information Theory

![Figure 4.2.: Examples of PCA and RCA on responses to two stimuli. The principal component is the vertical one but it is clearly not relevant for information transmission. The one that is most relevant is the horizontal one, because it compresses the responses while keeping the discriminability between stimuli high.](image)

of classifying a response to the $i^{th}$ stimulus as a response to the $j^{th}$ stimulus. In the presence of perfect decoding, the main diagonal of the confusion matrix $C_{ij}$ should be equal to 1, whereas all the other entries should be zero. The use of VP spike-train metrics coupled to clustering methods (e.g. k-nearest neighbors) to generate confusion matrices has proved to be suitable to estimate lower bounds of Shannon mutual information $I(R; S)$ in sensory information processing, e.g. auditory (Huetz et al. 2006), and somatosensory (Saal et al. 2009).

Although this approach takes into account the metrical properties of the event space to compute Shannon mutual information, once equivalence classes have been defined, the metrics is forgotten: whether the clusters are close to each other or far apart no longer influences the estimation of $I(R; S)$. In other words, spike train metrics is not explicitly embedded into the calculation of mutual information, which does not depend on neither the shape of the clusters nor the distance between them.

4.8. The specific information

The mutual information is the average information brought by reading a response. It is the weighted sum of the specific information of each response:

$$I(R; S) = \sum_r p(r) i(r; S)$$  \hspace{1cm} (4.17)

where the specific information is

$$i(r; S) = \sum_s p(s|r) \log \frac{p(s|r)}{p(s)}$$  \hspace{1cm} (4.18)

It may alternatively be defined as:

$$i(r; S) = H(S) - H(S|r)$$  \hspace{1cm} (4.19)
4.9. Biological relevance of Shannon Information Theory

Shannon entropy and information are purely topological. They do not take into account the metrical organization of signals. This is in line with Shannon’s point of view: “as soon as two responses are different (no matter how much), they can be discriminated at a later stage of processing”. While this view can be easily defended in electronics, telecommunication, etc... it may not be relevant to our subject. There are many reasons for this. The goal of the nervous system is also to make different stimulations converge to the same output, while keeping some others different (though there may be some signals that are very hard to discriminate). This may not be an easy task since there are constraints on the biological systems. For example, neurons process signals by filtering them with an exponential-like excitatory postsynaptic potential, they then fire if the membrane potential reaches some threshold.

Given these observations, the values obtained with Shannon information theory may be overestimations of what can really be extracted by the nervous system. In the next chapter, we propose a modified version of Information Theory that may turn out to be more suitable to address these concerns.
Galway Information Theory

Where we derive a definition of entropy that does embed an explicit measure of the metrical relations between input-output events. Indeed, the two approaches studied in the two previous chapters mutually ignore each other. Shannon information theory does not take the structure of the spike train space into account. We thus modify information theory so that it naturally depends on the metric organization of spike trains. This new parametric information analysis relies on similarity measures induced by the metrical relations between neural responses as spikes flow in. Thus, in order to assess the conditional entropy and the overall information transfer, this method does not require any a priori decoding algorithm to partition the space into equivalence classes. It therefore permits to determine the optimal parameters of a class of distances with respect to information transmission. We discuss the approach also with respect to existing methods. The proposed information theoretical measure proves to be a suitable generalization of Shannon Mutual Information in order to consider the metrics of temporal codes explicitly, that could permit an assessment of neurotransmission reliability in the presence of large spike train spaces (e.g. neural population codes) with high temporal precision.

5.1. Invitation to Galway information theory

Alice has lived in Luxemburg, London, Brussels, Geneva, Amsterdam and Paris for one year each, and Bob has spent two years in Moscow, Tokyo and Buenos Aires each. Whose distribution has the larger entropy? According to Shannon, the first is \( \log(6) \) and the second \( \log(3) \). However, from a geographic (metric) point of view, it is conspicuous that there is a larger uncertainty on the location of Bob at a given time. Of course, the probability to get its exact position is higher, but the error (distance) is on average much larger. If we consider a very narrow similarity measure (e.g. two cities are considered identical if they are less than 10 km apart), then the entropy of Alice is larger, in agreement with Shannon. However, if we consider a more lenient similarity measure (e.g. two cities are considered identical if they are less than 500 km apart), then the entropy of the Alice is almost null, while that of Bob is still...
5. Galway Information Theory

$\log(3)$. Finally, if we consider a very broad similarity measure (e.g. after all living anywhere on Earth is pretty much the same since there are planes taking you from London to Auckland in 24 hours), then both entropies are null. Clearly, the result depends on the question asked via the similarity measure. We claim this works as well with neural networks. The similarity measure -the decoding scheme, temporal precision- to be used depends on the task (or in a more biological fashion, on the reward function).

In 1987, Linsker (Linsker 1988): The principle of “maximum information preservation” states that for a layer of cells L that is connected to and provides input to another layer M, the connections should develop so that the transformation of signals from L to M (in the presence of processing noise) has the property that the set of output values M conveys the maximum amount of information about the input values L, subject to various constraints on, e.g., the range of lateral connections and the processing power of each cell.

However, there may be two principles in neural coding: categorizing and preserving information.

- Preserving information consists in guaranteeing the possibility of the reconstruction of the stimulus by looking at the response. In general, it will require taking all the details into account, so as to be able to reconstruct any feature of the stimulation. Of course, this may require a high precision and thus large computational resources. A collateral may be the need for a large delay before reaching the maximal information.

- Categorizing is, on the contrary, about forgetting details. When one has to hold a glass of water, whether it is cold or warm does not matter. What matters is the weight of the glass or the friction level of the surface. One must therefore provide similar reactions to stimulations with different features.

When Linsker requires a “maximum information preservation”, he does not take categorization into account, he just requires that stimulation can be determined unambiguously by observing the response. However, if you convey all the information, you are bound to be overwhelmed with signals that may be irrelevant for your task. Indeed, since there is at least a bit of noise in the nervous system, one may get many different responses to a single stimulation. If the set of responses do not overlap for any two different stimulations, the information is maximal. But the task of determining the stimulation and in particular to cluster responses that were elicited by the same stimulation is far from being trivial. This is what we could call categorization. For example, when Alice tells Bob a non-prime number if she thinks A and a prime number if she thinks B, it is not only necessary for Bob to understand that 17 and 24 do not belong to the same cluster, it is also necessary to understand that 17 and 65537 (which turns out to be Fermat’s fourth number) belong to the same one.

One thus needs to compress the information and keeps only the relevant part of it. In computational neuroscience, it is always argued that one should maximize information. There are no statements on what the value of the conditional entropy must be. It is as though we were concerned with the divergence of signals: no two stimuli eliciting the same response. But not at all with the convergence: all relevantly identical stimuli must elicit the same response. Our method is a compromise between these two paradigms.
5.2. Metrical information: from equivalence classes to distances

- Ask someone to think of a number and to print it with characters of size 1. Ask him for the sheet, how much information did you get? Now perform the same experiment with a magnifying glass.

- Each second, I choose “prime” or “not prime” and I give you a number satisfying this requirement: 1337689, 67113443, 5678899,... What is the information rate? According to Shannon Information Theory, it is 1 bit per second. Is that what you got?

These Gedankenexperimente are designed to show that information is an observer-dependent property. It does not only depend on what is relevant for the observer and on its priors (Rieke et al. 1997), but also depends on what it is actually able to decode.

5.2.1. Galway entropy

Shannon entropy calls upon a similarity measure $\phi(r, r')$ that can be thought of as a Kronecker function of two responses $r, r' \in R$:

$$\phi(r, r') = 1 \iff r = r'$$

As a consequence, two events can only be either identical or different (i.e. two slightly different or two very different events are treated the same way). This makes Shannon entropy diffeomorphism-invariant (or topological) because whether the mapping between $S$ and $R$ is random or isometric does not influence Shannon mutual information.

We propose an extension of Shannon entropy on the response space that can be defined in the following way:

$$H^*(R) = -\sum_{r \in R} p(r) \log \left( \sum_{r' \in R} p(r') \phi(r, r') \right)$$

where $R$ is the set of events, and $\phi(r, r')$ is a yet-unspecified similarity measure between the events $r$ and $r'$. In the general framework, the similarity measure can be any real function with values in $[0, 1]$. The difference with Shannon entropy is the logarithm argument: instead of $p(r)$, we have $\sum_{r' \in R} p(r')\phi(r, r')$. It can thus be noted that:

- The logarithm argument is always higher than the probability of $r$, because

$$\sum_{r' \in R} p(r')\phi(r, r') = p(r) + \sum_{r' \in R, r' \neq r} p(r')\phi(r, r')$$

the second term being positive. Therefore, this entropy is always lower than Shannon entropy.

- If the similarity measure $\phi(r, r')$ is taken to be a Kronecker delta, then the logarithm argument of Eq. 5.2 is equal to $p(r)$, and thus $H^*(R)$ reduces to Shannon entropy.
The generalization of a concept is a process very often used in science. A method of generalization is not uniquely determined, for there are usually numerous ways of carrying it out. One requirement, however, must be rigorously satisfied: any generalized concept must reduce to the original one when the original conditions are fulfilled.


- Although the similarity measure is not specified yet, it can be seen that the broader \( \phi(r, r') \), the lower the entropy \( H^*(R) \).
- If \( \forall r, r' \in R, \phi(r, r') = 1 \), then the entropy is nil.

The Galway conditional entropy on the response given a specific stimulus \( s \in S \) can then be defined as:

\[
H^*(R|s) = - \sum_{r \in R} p(r|s) \log \left( \sum_{r' \in R} p(r'|s) \phi(r, r') \right)
\]

and its average over the entire set of stimuli \( S \) gives:

\[
H^*(R|S) = \sum_{s \in S} p(s) H(R|s) = - \sum_{s \in S} \sum_{r \in R} p(r, s) \log \left( \sum_{r' \in R} p(r'|s) \phi(r, r') \right)
\]

Finally, similar to Shannon MI, the Galway mutual information is taken as a difference between marginal and conditional Galway entropies:

\[
I^*(R; S) = H^*(R) - H^*(R|S)
\]

\[
= \sum_{s \in S} \sum_{r \in R} p(r, s) \log \left( \frac{\sum_{r' \in R} p(r'|s) \phi(r, r')}{\sum_{r' \in R} p(r') \phi(r, r')} \right)
\]

This is the general framework of the theory. It encompasses as a specific case the unbinned Shannon entropy but also the binning procedure, where all the events belonging to the same bin have similarity 1, while two events in different bins have similarity 0. The codebook method can also be restated in this framework.

**Additivity property of the extension of the entropy** Shannon required the entropy to satisfy the three following properties:

1. It is continuous on the probabilities \( p \).
2. If the equivalence classes are equiprobable, i.e. \( \forall p, \ p = \frac{1}{N} \) with \( N \) being the number of equivalence classes, then the entropy \( H(R) \) is a monotonic increasing function of \( N \).
3. If a probabilistic choice is further broken down into two successive choices, the overall entropy should be a weighted sum of the sub-choice entropies. In the example of Figure 5.2d (top row), \( H(a, bc, bd) = H(a, b) + bH(c, d) \), where \( a, b, c \) denote probabilities.
5.2. Metrical information: from equivalence classes to distances

The entropy $H^*$ satisfies the first two requirements. Does it also satisfy the third property, i.e. additivity? Consider the example of Figure 5.2d (bottom row). The joint entropy is:

$$H^*(a, bc, bd) = -a \log(a) - bc \log(bc + bdy) - bd \log(bd + bcy)$$ (5.7)

Since:

$$H^*(a, b) = -a \log(a) - b \log(b)$$ (5.8)

$$H^*(c, d) = -c \log(c + dy) - d \log(d + cy)$$ (5.9)

Then:

$$bH^*(c, d) + H^*(a, b) = -a \log(a) - b \log(b) - bc \log(c + dy) - bd \log(d + cy)$$ (5.10)

$$= -a \log(a) - bc \log(bc + bdy) - bd \log(bd + bcy)$$ (5.11)

$$= H^*(a, bc, bd)$$ (5.12)

The extension of the entropy we propose thus satisfies the three previous fundamental requirements.

5.2.2. Metrical Entropy

In order to account for the metrical relations between events, we here propose to define the similarity measure as a decreasing function of the distance between two events, following the intuitive idea that a distance provides a measure of dissimilarity. We thus call $H^*$ a metrical entropy.

Figure 5.1b shows an example of the metrical entropy (dashed curve) for a 1-dimensional Gaussian distribution of events (continuous curve) when defining the similarity measure as a Heaviside function of the distance between two events, i.e. $\phi(r, r') = H(D_{\text{critic}} - d(r, r'))$. Henceforth, we will focus on this form for $\phi$. When the cutoff $D_{\text{critic}}$ (or critical distance) of the Heaviside function is zero, we get back to Shannon entropy and when it goes to infinity, the metrical entropy becomes zero.

When applying the metrical information analysis to neurotransmission, we first need to select a distance function on the event space and define a class of similarity measures (Sec. 5.2.2), for example a Heaviside function of the distance. Then, we have to define an optimality condition for efficient neurotransmission (Sec. 5.2.2). Finally, we have to determine the optimal values of the parameters with respect to the defined optimality condition (Sec. 5.2.2).

Definition of the similarity measure

When the event space consists of spiking signals, a method to quantify the distance between neural responses is offered by the Victor-Purpura (VP) spike train metrics (Victor and Purpura 1996), though other metrics may be suitable as well. We define the similarity measure $\phi(r, r')$ as a function of the VP distance $D_{VP}(r, r')$ between two responses $r$ and $r'$. The distance $D_{VP}(r, r')$ depends on the VP cost parameter $C_{VP}$ (Victor and Purpura 1996), which determines the time scale of the temporal coding analysis (the cost parameter $C_{VP}$ regulates the influence of spike timing vs. spike count when calculating the VP distance between $r$ and $r'$).
5. Galway Information Theory

As stated above, we wish to define the similarity as a decreasing function of the distance. A simple way to define the similarity measure \( \phi(r, r') \) is to take a Heaviside step function of the distance \( D_{VP}(r, r') \):

\[
\phi(r, r') = H(D_{\text{critic}} - D_{VP}(r, r'))
\]  

where the critical distance \( D_{\text{critic}} \) is the cutoff parameter: as long as \( D_{VP}(r, r') < D_{\text{critic}} \) the responses \( r, r' \) are considered to be identical, otherwise they are classified as different. Note that if \( D_{\text{critic}} = 0 \) one recovers the Shannon entropy from Eq. 5.2.

It is worth mentioning that by taking a similarity measure different from the Kronecker delta we introduce a bias in the computation of the entropy \( H^*(R) \). This bias is supposed to depend on the characteristics of the readout system, e.g. the properties of a downstream neural population. A consequence of this bias is that the metrical information \( I^* \) is not diffeomorphism-invariant (in contrast to Shannon mutual information). Note also that since we take a similarity measure that depends on the \( D_{VP}(r, r') \) distance only, the metrical quantities are invariant under the isometry group and under isotropic homotheties (i.e. affine transformations preserving the ratios between event distances).

**Optimality condition for efficient neurotransmission**

In the zero-noise limit, optimal feed-forward information processing would require maximizing the metrical information \( I^*(R; S) \) according to an *infomax*-like principle (Linsker 1988; Bell and Sejnowski 1995; Nadal et al. 1998). However, since neural information processing is not noise-free and occurs through multiple encoding/decoding stages, the issue of minimizing the variability on the output representation at each stage constitutes a major requirement. Therefore, akin to the principle of redundancy reduction proposed for biological sensory processing by (Barlow 1961), the metrical conditional entropy \( H^*(R|S) \) constitutes a fundamental quantity for neurotransmission, and optimal neural information processing should both maximize the metrical information \( I^*(R; S) \) and (at the same time) minimize the conditional entropy \( H^*(R|S) \). This perfect discrimination condition is met when all the responses elicited by the same stimulus are strictly identical \( (\phi = 1) \), whereas all the pairs of outputs elicited by two distinct stimuli are strictly different \( (\phi = 0) \). In the sequel, we will refer to this situation as the **optimal discrimination**.

According to this optimality principle, a neural encoder should provide both that any response is not elicited by two different stimuli and that the set of responses elicited by any stimulus remains as small as possible. If a neural encoder were not behaving this way, then a single stimulus would possibly elicit millions of different responses. How could this plethora be effectively decoded by a downstream neural population? In other words, neural information processing requires encoding mechanisms capable of producing as few responses as possible to a given stimulus while keeping these responses different between stimuli (Földiák 1990; Nadal and Toulouse 1998; Olshausen and Field 1996; Willmore and Tolhurst 2001, i.e. sparse coding; see, for instance.).

**Determining the parameters of the similarity measures**

The similarity measure \( \phi(r, r') \) is the lever on which we can act to set the compromise between the overall information \( I^*(R; S) \) and the mean variability of the response to a stimulus
5.2. Metrical information: from equivalence classes to distances

There are two ways of understanding the similarity measure. On the one hand, provided that a comprehensive knowledge of the properties of the readout system (e.g. a neural decoder) is available, the selectivity level (i.e. the cutoff distance $D_{\text{critic}}$) of the similarity measure should account for the specific properties of the reader (e.g. time constant of synaptic integration). On the other hand, if no knowledge on the decoding system is available, the similarity measure should be set to allow optimal information transmission (in terms of both $I^*(R;S)$ and $H^*(R|S)$) to be achieved in order to predict the parameters that maximize the information transmission.

What appears to be meaningful in order to set the optimal parameters for $\phi(r,r')$ is to consider two sets of $D_{\text{VP}}(r,r')$ distances:

- the distances between the responses elicited by the same stimulus (henceforth referred to as *intra-stimulus distances*);
- the distances between the responses elicited by different stimuli (named *inter-stimulus distances*).

In order to determine the optimal value for $D_{\text{critic}}$, we compute the minimum and maximum *intra*-stimulus distances as well as the minimum and maximum *inter*-stimulus distances. The optimal coding condition, corresponding to maximum $I^*(R;S)$ and zero $H^*(R|S)$, occurs when the maximum intra-stimulus distance (which provides the size of the largest cluster of responses) becomes smaller than the minimum inter-stimulus distance (which estimates the smallest distance between clusters of responses). We can thus set the cut-off distance in-between the maximum intra-stimulus distance and the minimum inter-stimuli distance. See Fig. 5.2a for a toy example.

In the case of neurotransmission, the relationship between intra- and inter-stimulus distance distributions tends to evolve over time, as the input spike waves across multiple afferents flow in the readout system. Figure 5.2a shows an example of intra- and inter-stimulus distance distributions evolving over time. The two distributions separate from each other after about 110 ms. The critical parameter $D_{\text{critic}}$ can then be taken as the distance at which the maximum intra-stimulus distance becomes smaller than the minimum inter-stimulus distance (dashed line in Figure 5.2b). The time at which the critical distance $D_{\text{critic}}$ can be determined indicates when the perfect discrimination can be achieved. In other words, optimal discrimination occurs when the distributions of intra- and inter-stimulus distances stop overlapping, which means that: (i) the conditional entropy $H^*(R|S)$ is nil, because all the responses elicited by any stimulus are identical; (ii) the information $I^*(R;S)$ is maximum, because any two responses elicited by two different stimuli are always correctly discriminated. Fig. 5.2b illustrates an example of well separated intra- and inter-stimulus distance distributions (bottom). It also shows the values for $I^2(R;S)$ and $H^2(R|S)$ as a function of the critical distance parameter (top). It appears clearly that taking $6.5 < D_{\text{critic}} < 8$ guarantees perfect discrimination (because within that range $I^2(R;S)$ is maximum and $H^2(R|S)$ is nil).

As aforementioned, the critical distance $D_{\text{critic}}$ is interdependent on the VP cost parameter $C_{\text{VP}}$ (Victor and Purpura 1996). We define the optimum VP cost $C^*_{\text{VP}}$ as the one that leads to earliest perfect discrimination (in the example of Figure 5.2b, a cost $C_{\text{VP}} = 0.15$ led to perfect discrimination after 110 ms).

The table depicted on Fig. 5.2c presents a toy example comparing the estimates provided by Shannon mutual information (Sec. 4.2), Shannon mutual information based on metrical space
5. Galway Information Theory

analysis (Sec. 4.7), and metrical information analysis. It can be seen that the $I^*(R; S)$ captures
the metrical property of the mapping from input to output spaces. For instance, if a stimulus $s$
yields 10 different responses $r$ that are randomly distributed on the output space, then $I^*(R; S)$
is nil, because the critical distance will be very large.

Decoding scheme

When the perfect discrimination is reached, a very simple decoding scheme can be performed.
It is sufficient to compare the response with an example (a prototype) of a response to all of the
potential stimuli. Only one of the distance will be lower than the critical distance. This gives
the identity of the stimulus. In the same spirit, one does not even need the critical distance,
a 1-nearest neighbor algorithm would be enough. Interestingly, one does not even care about
the prototype of the response that is chosen. Any would fit.

This appears to us as a decoding scheme that could potentially be used by the nervous
system. Indeed, this allows a fast identification of the stimulus based on a single trial. Our
approach estimates the time at which this decoding scheme would be fully accurate.

5.2.3. Local Version of the Metrical Entropy

A critical property of the metrical information $I^*(R; S)$ is the global nature of the similarity
measure $\phi(r, r')$, i.e. the fact that the critical distance does not depend on the position in the
event space. In the cartoon example of Fig. 5.2e, $I^*(R; S)$ can not be maximized while keeping
$H^*(R|S) = 0$ because the size of one of the clusters is larger than the distance between two
others. This is an undesired outcome of the analysis because all the responses within a cluster
are closer to each other than the closest responses from other clusters.

Otherwise stated, the global version of the metrical information assumes a similarity mea-
ure that does not depend on the region of the event space. In order to circumvent this limi-
tation, the discrimination selectivity function can be made locally specific. A local version of
the metrical entropy can be derived by taking a similarity measure measure that varies over the
event space and thus where $\phi(r, r')$ is not necessarily equal to $\phi(r', r)$. This can be formalized
this way: $\phi(r, r') = \mathcal{H}(d(r, r') - D_{critic}(r))$ (Note that we still constrain the similarity measure
be isotropic). The critical distance for $\phi(r, r')$ becomes the size of the cluster of responses
(to a given stimulus) which $r$ belongs to. As a consequence, instead of the two parameters
$C_{VP}$ and $D_{critic}$, the local version of metrical information measure will depend on $C_{VP}$ and a
family of critical distances $D = \{D_{critic}\}$ Fig. 5.2e.

When considering a local similarity measure, the optimal discrimination condition, i.e. max-
mum $I^*(R; S)$ and zero $H^*(R|S)$, will occur when the size of each cluster in the event space
becomes smaller than the distance with its closest clusters.

Link with Kolmogorov-Chaitin Complexity and Graph Complexity

Let us suppose a distribution of events in a given space. There is an entropy associated to this
distribution. In fact, this entropy can be interpreted as the Kolmogorov-Chaitin complexity, i.e.
the shortest binary word that describe each of the events. When a binning is performed on that
space, some of the events are grouped. What does it mean ? It means that events all the events
in a bin are strictly identical. They are assigned similarity measures of 1. All the similarity
5.2. Metrical information: from equivalence classes to distances

Measures between events in two different bins are 0. Transitivity is guaranteed. There Shannon entropy is well-defined. However, if we consider more general “binnings”, Shannon entropy is no longer defined. Let us give examples to settle ideas. Let us consider 4 equiprobable events on an axis: \{1, 2, 4, 5\}. The entropy is 2 bits. If we use bins of width 3 starting at 0, then 1 and 2 are identical and 4 and 5 as well. The entropy is then 1 bit. Now, if we have a finer similarity measure (not transitive) with \(\phi(1, 2) = \phi(2, 4) = \phi(4, 5) = 1\) but \(\phi(1, 5) = 0\) (threshold for discrimination), then Shannon entropy is not defined. However, the metrical entropy is. More interesting, in the first case, it gives an entropy of 2 bits. In the second case, it yields \(-4 \frac{1}{4} \log \frac{1}{4} + \frac{1}{4} = 1\) bit. In the third case, the entropy is \(-2 \frac{1}{4} \log \frac{1}{4} + \frac{1}{4} = 1\) bit. It even allows similarity values different from 0 and 1. This is then a generalization of entropy for any generalized binning.

Transitivity property of the metrical information. Under the perfect discrimination condition, i.e. maximum \(I^*(R; S)\) and zero \(H^*(R|S)\), all the responses elicited by the same stimulus are strictly identical (\(\phi = 1\)), whereas all the pairs of outputs elicited by two distinct stimuli are strictly different (\(\phi = 0\)). Therefore, the optimality condition guarantees a grouping of responses to the same stimulus, case in which we recover transitivity, but not as an ad hoc property.

Note on the data-processing inequality theorem. In Shannon information theory, the data-processing inequality states that any processing of the signal, \(f(r)\), can neither increase nor create information, i.e. \(I(f(R); S) \leq I(R; S)\) (Cover and Thomas 1991). The metrical information is a parametric measure depending on the similarity measure, which is meant to account for the properties of the observer (readout system). Therefore, changing the similarity function after a processing of the signals (e.g. downstream a neural population) might result in an increase of \(I^*(f(R); S)\) compared to the upstream information estimate. For instance, if a unitary flat similarity measure were taken to compute the metrical information before processing, then \(I^*(R; S) = 0\). If after the processing stage a narrower similarity function were taken, it would eventually lead to \(I^*(f(R); S) > 0\). Notice, however, that this violation of the data-processing inequality would rather result from a misinterpretation of the metrical information theory, and it would merely be due to an inappropriate choice of the similarity measure before the processing stage.

Expansion of the similarity measure

In order to understand a bit more what the insertion of the similarity measure \(\phi\) changes, we expand it around a Dirac similarity measure:

\[
H^*(R) = -\sum_r p(r) \log(\sum_{r'} p(r') \phi(r, r'))
\]

Let’s write \(\phi(r, r')\) as:

\[
\phi(r, r') = \delta(r, r') + \epsilon(r, r')
\]

Then:

\[
H^*(R) = -\sum_r p(r) \log(p(r) + \sum_{r' \neq r} p(r') \epsilon(r, r'))
\]
5. Galway Information Theory

\[ H^*(R) = - \sum_r p(r) \log(p(r)) - \sum_r p(r) \log(1 + \sum_{r' \neq r} \frac{p(r')}{p(r)} \epsilon(r, r')) \] (5.17)

If we assume slow variations of the probabilities with \( r \) and that \( \epsilon(r, r') \ll 1 \), we can expand the logarithm:

\[ H^*(R) = H(R) - \sum_r p(r) \sum_{r' \neq r} \frac{p(r')}{p(r)} \epsilon(r, r') \] (5.18)

\[ H^*(R) = H(R) - \sum_{r, r' \neq r} p(r') \epsilon(r, r') \] (5.19)

Optimal values of the critical distance

It would be interesting to know what are the best values for the critical distance is simple cases. Let us, for example, consider two stimuli eliciting two different responses. Let us further consider that the 4 response are at the vertices of a tetrahedron Fig. 5.3.

Let us consider a few possibilities:

- In order to make the conditional entropy null, the similarities are all equal to 1. The information is then null as well.

- In order to make the information maximal (i.e. \( \log(2) \)), the similarities are all equal to 0. The conditional entropy is then non-null.

- Now, if we consider a trade-off between the information and the conditional entropy.

\[ Q(R, S) = H^*(R) - (\alpha + 1)H^*(R|S) = I^*(R, S) - \alpha H(R|S) \] (5.20)

Then according to the value of \( \alpha \), what is the best similarities? It is surprising to find that the best values for the similarities are either 0 or 1. On Figure 5.3, we plot \( Q \) as a function of the similarity value between pairs of events for several values of \( \alpha \).

It appears that the function \( Q \) reaches it higher value either for similarities equal to 0 or 1. A general proof of this property still remains to be given.

5.3. Discussion

The presented metrical information analysis proposes a complementary approach, compared to Shannon information theory, to study stochastic communication systems. Shannon mutual information provides a fundamental upper bound on the quality of the coding, no matter the decoding system and the conditional entropy. Metrical information analysis incorporates the notion of decoding system (through the distance-based similarity function) and the importance of conditional entropy for the optimization of information transmission. This complementarity relation leads to a shift from an approach that tells how much information is there but not how to read it, to a constructive-like approach that considers the parameters of the decoding system to actually extract the knowledge contained in the neural code.
This work attempts to contribute to existing methods, such as Fisher information analysis 
(Blahut 1988; Clarke and Barron 1990; Rissanen 1996; Brunel and Nadal 1998) and rate 
distortion theory (Bialek et al. 1991; Gabbiani and Koch 1996; Dimitrov et al. 2001; Dimitrov 
et al. 2003), that bridge the gap between pure decoding approaches (Rieke et al. 1997; Borst 
and Theunissen 1999; Pouget et al. 2000; Dayan and Abbott 2001, e.g. Bayesian algorithms, 
k-nearest neighbor decoders, population vector decoding;) and pure information theoretical 
analyzes (London et al. 2002; Sharpee et al. 2006; Butts and Goldman 2006, e.g.). It also 
relates to recent works by (Thomson and Kristan 2005), (Victor and Nirenberg 2008), and 
(Quian Quiroga and Panzeri 2009) that have studied the complementary properties of decoding 
and information theoretical algorithms to decipher the neural code.

Our work is related to the approach by (Treves 1997) that consists in estimating the struc-
ture of the perceptual space by comparing results from information theory and maximum-
likelihood estimations. Theoretical constraints may be derived between Shannon mutual in-
formation and the fraction of correct decodings. For a given fraction of correct decodings, it is 
possible to infer lower and upper-bounds to the MI depending on the structure of the perceptual 
space. When analyzing experimental data, knowing whether the results are close to the lower 
or upper-bound allows the dimension and structure of the response space to be determined (do 
they lie on a line or at the vertices of an $n$-simplex ?).

Both rate distortion theory (RDT) and metrical information analysis look for a compression 
of the responses such that the information about the stimuli is preserved. The compression is 
to be realized by a the neural decoder impose some constraints on the possible compressions. 
RDT methods aim at identifying the features of the code that are most relevant to information 
transmission while discarding the irrelevant ones so as to compress the code. Metrical infor-
mation analysis tries to find the optimal way in which neurons can look at the code (so as to 
minimize the noise entropy and maximize the information) given the constraints. These con-
straints are captured by the estimation of distances between spike trains. Indeed, the similarity 
measure is subordinated to distance estimation. It is then possible to infer what are the optimal 
parameters of the neurons (e.g. the membrane time constant) to transfer information while 
compressing the signals. In this framework, setting the critical distance so as to minimize the 
conditional entropy amounts to stress the importance of the compression of the code besides 
the importance of input discriminability. The relaxation of this optimality constraint may lead 
to biologically plausible trade-offs between compression and information transfer (currently 
under examination, see example in Supporting Figure S2).

From a machine learning perspective, the proposed method relates to statistical linear dis-
crimination (e.g. for dimensionality reduction and classification) such as Fisher discriminant 
analysis (Fisher 1936; Friedman 1989; Mika et al. 1999; McLachlan 2004). Indeed, there 
extists a qualitative link to the index of discriminability $d'$, which quantifies the separability of 
two populations of events that are supposed to have identical Gaussian conditional probability 
distributions. The $d'$ measure is defined as the distance between the means of the two distri-
butions, normalized by their standard deviation. The linear combination of parameters that 
maximizes the ratio $d'$ leads to optimal discrimination (Fisher 1936). For instance, $d'$ can be 
used to assess the stimulus discriminability based on the spike count of neural responses, both 
for single cell and population coding (Petersen et al. 2002a). As for the computation of $d'$, esti-
mating the metrical information $I^\ast(R; S)$ requires a comparison of the distances between the 
clusters of events with the distances within a cluster. Notice that instead of providing a con-
tinuous measure of discriminability without any upper bound (the ideal case would be reached when the two distributions are infinitely far apart), the measure presented here is bounded: two distributions are considered as fully separated as long as their size is smaller than their distance. Once this condition is met, no matter how far they are, their discriminability is estimated to be the same.

From a more general perspective, a link can be drawn between the proposed metrical information analysis and signal detection theory (SDT) (Dayan and Abbott 2001). SDT offers a set of statistical tools to assess the capacity of neural systems to encode and transmit information (Tolhurst et al. 1983; Britten et al. 1992; Guido et al. 1995; Cheng and Wasserman 1996; Petersen et al. 2002a). In contrast to classical information theoretical principles, SDT makes an extensive use (either implicit or explicit) of metrics. As suggested in Fig. 5.4, the metrical conditional entropy \( H^*(R|S) \) can be related to the SDT “false positive rate” concept, in the sense that when \( H^*(R|S) = 0 \) two identical stimuli are always correctly identified (0 false alarms). Likewise, \( I^*(R; S) \) is akin to the “true positive rate”, in the sense that when it is maximum the hit rate is maximum (i.e. two different inputs are always correctly discriminated). Under the optimal discrimination condition (i.e. maximum \( I^*(R; S) \) and nil \( H^*(R|S) \)), the Receiver-Operating Characteristic (ROC) curve (Barlow and Levick 1969; Cohn et al. 1975; Geisler et al. 1991) as a function of the critical distance follows the 0-miss and 0-false alarm axes. Therefore, the integral of the ROC curve is 1, and so it is the probability of discriminating correctly in a two-alternative forced-choice test (which holds here for any two stimuli) (Dayan and Abbott 2001). Notice that in metrical information analysis the emphasis is shifted from the \textit{values} to the \textit{distances between the values}. In SDT, the threshold is a certain \textit{value} of the output parameter, whereas in metrical information analysis the threshold is a certain \textit{distance} between two outputs.

5.4. Conclusions

This chapter highlights the importance of endowing an information theoretical analysis with the capability of taking into full account the metrical relations (e.g. distances) between spike trains in order to assess neurotransmission reliability. In contrast to Shannon Mutual Information, the proposed measure does not require any a priori partitioning of the event space into equivalence classes (e.g. temporal binning procedures or clustering of neural responses based on confusion matrices). Rather, the discretization of the state space evolves over time and is shaped according to the metrics (similarity measure) employed. It is shown how the distances between events (computed, for instance, based on a Victor-Purpura spike train metrics (Victor and Purpura 1996)) can be explicitly incorporated in the definition of the metrical entropy. The resultant information quantity, in contrast to Shannon information, is both a parametric and diffeomorphism-variant estimate of the statistical dependence between two variables (e.g. stimulus and neural response).

The parametric nature of the proposed method is due to its dependence on a similarity measure \( \phi(r, r') \) to assess, as spikes flow in, the metrical relations between neural responses \( r, r' \in R \). We put forth the hypothesis that the selectivity level represented by the similarity measure may be tuned to reflect some known properties of the readout system (e.g. actual temporal precision of a neural decoder). By contrast, Shannon information can be understood in...
5.4. Conclusions

terms of a Kronecker-like similarity measure, i.e. the selectivity capacity of an ideal observer. As a consequence, in certain situations, Shannon information measurements of the capacity of a neural code might happen to be not biologically relevant due to nonexistent neural decoders able to exploit the whole amount of estimated information (Quian Quiroga and Panzeri 2009). Since the similarity measure is a projection of the properties of the reader onto the input space, metrical information analysis can help solving this problem. Indeed, the metrical information $I^*(R; S)$ and conditional entropy $H^*(R|S)$ can be seen as quantitative answers to the question: is it possible for a readout system that implements a decoding scheme based on the similarity measure $\phi(r, r')$ to perform perfect decoding, i.e. maximize $I^*(R; S)$ and minimize $H^*(R|S)$?
5. Galway Information Theory

Figure 5.1.: Binary word decoding, example of metrical entropy of a gaussian and evolution of the distances between spike trains. (a) Binary word decoding of neural responses. Time is partitioned according to a fixed temporal resolution $\Delta t$. If a spike occurs within a given time bin, the latter is assigned 1, otherwise 0. Besides potential disregard of action potentials (i.e. multiple spikes might occur within the same temporal bin), the binning procedure is prone to anisotropy effects: spike trains $r_1$ and $r_2$ are considered as identical, whereas $r_2$ and $r_3$ are considered as different (although the spike times of $r_2$ and $r_3$ are closer than those of $r_1$ and $r_2$). (b) Example of evolution of the metrical entropy (dashed curve) of a Gaussian distribution of events (continuous curve) as a function of the cutoff when taking a Heaviside similarity measure (i.e. the similarity function is 1 for distances below the cutoff value and 0 otherwise). (c) In this toy example, the maximum intrastimulus distance $max D_{intra}$ is smaller than the minimal interstimulus distance $min D_{inter}$. Thus, we are in the perfect discrimination condition, i.e. $I^*(R; S)$ is maximal and $H^*(R|S)$ is nil.
5.4. Conclusions

Figure 5.2.: Examples of results on metrical information. (a) Example of intra- and inter-stimulus distance distributions (red and blue curves, respectively) over time for a VP cost parameter $C_{VP} = 0.15$. The optimal discrimination condition is met after about 110 ms, when the two distributions stop overlapping. (b) Example of how $I^* (R; S)$ and $H^* (R|S)$ varies as a function of the critical distance (top), given the distributions of intra- and inter-stimulus distances (bottom). (c) Simple comparison between Shannon MI, metrical space analysis, and metrical information analysis on a two-stimuli classification example. (d) Simple example showing the additivity property of Shannon entropy (top row) and metrical entropy (bottom row) with probabilities denoted a, b, c and similarities x, y, z. (e) Local vs. global metrical information measures. Since the size of the orange cluster is larger than the distance between the purple and green clusters, the global similarity measure would make it impossible to maximize $I^* (R; S)$ while maintaining $H^* (R|S) = 0$. By contrast, a local version of the metrical entropy would consider different similarity measures for the clusters. Thus, the $D_{critic}$ for the green cluster would be taken smaller than the one for the orange cluster.
5. Galway Information Theory

Figure 5.3.: (Left) Distribution of the 4 responses to 2 stimulations on a tetrahedron. (Right) The measure $Q$ as a function of the similarity between pairs of responses on the tetrahedron. The different lines correspond to different values of the trade-off parameter $\alpha$. The function is convex, so that the optimal value of the similarity is always either 0 or 1.
5.4. Conclusions

Figure 5.4.: Link between metrical information theory and signal detection theory. Signal detection theory (SDT) offers a set of statistical tools to evaluate the discrimination efficacy of decision-making processes (e.g. diagnostic tests). One of these tools is the Receiver-Operating Characteristic (ROC) (Barlow and Levick 1969; Cohn et al. 1975; Geisler et al. 1991), which can be used to determine the appropriate selectivity (or cutoff) threshold of a classification procedure. (a) Let \( R_1 \) and \( R_2 \) denote two populations of events characterized by overlapping Gaussian distributions. The exploration of all possible cutoff values to discriminate \( R_1 \) from \( R_2 \) events can be quantified via a ROC analysis, which plots the true positive (“hit”) rate (e.g. the fraction of correct classification of \( R_2 \) events) against the false positive (“false alarm”) rate (e.g. the fraction of \( R_1 \) events incorrectly classified as \( R_2 \) events) while varying the selectivity threshold. As a consequence, ROC curves permit to trade-off the relative importance of true positive and false positive rates to enhance the accuracy of a discrimination test. Different points on the curve denote different choices of the discriminatory threshold. The diagonal line corresponds to random guess (less accurate classification). The closer the curve to the top-left quadrant, the more accurate the discrimination. TP: true positive; FP: false positive; TN: true negative; FN: false negative. (b) The metrical information with respect to the conditional metrical entropy as a function of the critical distance for two Gaussians of variance 5 separated by \( d \). The behavior highly resembles that of a classical ROC, as the metrical information (resp. conditional metrical entropy) roughly gives the true positive rate (resp. false positive rate) of a discrimination task performed by an observer acting according to the similarity measure used.
Part II.

Theoretical Applications of the Galway Information Theory
An information-theoretic account of the Perceptual Magnet

Where we extend the Galway information analysis. We propose an input-output symmetric version of the Galway information by introducing a metrics on the input space the same way it was done on the outputs. In addition, in contrast to the original optimization method that required the equivocation (or conditional entropy) to be zero and, simultaneously, the metrical information to be maximum, here an objective function setting a trade-off between minimizing the equivocation and maximizing the information is proposed. In the spirit of Butts and Goldman (2006) (Butts and Goldman 2006), we investigate here how an optimal decoder should behave in the low and high-noise regimes. It is shown that the well-known perceptual magnet effect (Kuhl 1991) can be understood as a consequence of maximizing the objective function in a low-noise regime. In a simple case with narrow Gaussian categories, it implies that an optimal decoding system should perform very fine discrimination near the boundaries of the categories, while being less selective near the centers of their distribution.

6.1. Introduction

Shannon information theory (Shannon 1948) provides a mathematical framework to estimate the amount of information conveyed by a noisy transmission system. It tells how much a decoder can possibly know about the source by reading the message transmitted. Information theory has been extensively used in the field of neuroscience (Bialek et al. 1991; Borst and Theunissen 1999; Quiroga Quiroga and Panzeri 2009). However, readout neurons have properties that make them differ from ideal observers. In addition, there may be details of the code that are not relevant to convey. In a recent proposal (Brasselet et al. 2009) -akin to the bottleneck information (Tishby et al. 1999) and to relevant component analysis (Shental et al. 2002)- a metrics-based information theoretical measure was proposed to account for the properties of downstream neurons (i.e. the decoder system) that make them transmit the most
6. An information-theoretic account of the Perceptual Magnet

relevant information.

An entropy measure embedding the metrical relations between events was defined as:

\[ H^*(X) = - \sum_x p(x) \log \sum_y p(y) \phi(x, y) \]  \hspace{1cm} (6.1)

with \( \phi(x, y) \) denoting a similarity measure between the events \( x \) and \( y \) that depends on the distance between them. The distance and the similarity measure can be thought of as representations of the properties of the downstream decoder. It was then proposed that an optimal decoder should, at the same time, minimize the equivocation (metrical conditional entropy) and maximize the metrical information so as to be able to reconstruct the stimulus unambiguously. The theory was validated on a tactile discrimination task using a data set of human microneurography recordings. In that application, the similarity measure was taken as an all-or-none function of the distance with a cut-off value called the critical distance.

This work extends the previously introduced metrical information theory along two axes. First, the previous measure was \textit{global} in the sense that the similarity measure did not depend on the region in the event space. That is, the discriminability was invariant over all the space. Here, a \textit{local} version of the metrical information theory is developed, which allows the similarity function to be location dependent. Second, the previous measure was \textit{asymmetric} in the sense that a similarity measure was introduced on the response space only, neglecting the stimulus space. Here, a \textit{metric} is also considered for the input space, making the new metrical information measure \textit{symmetric}.

In the following sections we define the local and symmetric version of the metrical entropy and we show that the introduction of a similarity measure on the input space can be used to describe the discrimination task to be solved by the downstream reader. We also propose that the attempt to maximize the metrical information and minimize the metrical conditional entropy can be formalized as maximizing a single objective function that takes a trade-off parameter as input. When applying this optimization to Gaussian distributed events, we find results very similar to the physiological phenomenon of perceptual magnet (Kuhl 1991).

6.2. Methods

6.2.1. Local version of metrical information

A local version of the metrical entropy (Eq. 6.1) can be derived by making the similarity function locally specific:

\[ H^*(R) = - \sum_{r \in R} p(r) \log \left( \sum_{r' \in R} p(r') \phi(r, r') \right) \]  \hspace{1cm} (6.2)

The local metrical information can then be taken as:

\[ I^*(R; S) = H^*(R) - H^*(R|S) = \sum_{r, s} p(r, s) \log \frac{\sum_{r'} p(r'|s) \phi(r, r')}{\sum_{r'} p(r') \phi(r, r')} \]  \hspace{1cm} (6.3)
6.2. Methods

where \( r, r' \in R \) and \( s \in S \) denote neural responses and input stimuli, respectively, and where the similarity measure \( \varphi(r, r') \) is not necessarily equal to \( \varphi(r', r) \) and varies over the event state space.

Given this metrical information definition, if the similarity measures are taken as functions of the distance between \( r \) and \( r' \), the information (resp. equivocation) is maximum (resp. minimum) if the outputs of a category are all closer than they are from the outputs of the other categories. Indeed, in that case, a similarity measure at a given point includes all the responses to the same category while excluding all other responses. Note that here, the similarity measure can be thought of as the processing of the inputs by the nervous system. It is symbolic in the sense that we do not claim it should be implemented at some precise location, but could be the result of a multi-stage processing. The crucial variables in these measures are the similarity measures at each point in the response space.

Instead of trying to maximize the information \( I^*(R; S) \), we will look for conditions maximizing an objective function defined as:

\[
Q(R, S) = H^*(R) - (\alpha + 1)H^*(R|S) = I^*(R, S) - \alpha H(R|S)
\]  

where the factor \( \alpha \) determines the trade-off between maximizing the metrical information and minimizing the metrical equivocation. If \( \alpha = 0 \), the objective function reduces to the metrical information. Thus, we do not want an observer that acts as a the perfect Shannon observer by differentiating as much as it can by taking into account every slight differences between responses. The problem is more that of finding the characteristics of an observer who would really reconstruct the categories, without taking into account differences in the responses to the same category.

Since we now consider a local version of the metrical information, we can define the specific metrical information associated to a single response \( r \):

\[
I^*(r; S) = \sum_s p(r|s) \log \frac{\sum_{r'} p(r'|s) \varphi(r, r')} {\sum_{r'} p(r') \varphi(r, r')}
\]  

Similarly, the specific equivocation can be defined as:

\[
H^*(r|S) = -\sum_s p(r|s) \log \sum_{r'} p(r'|s) \varphi(r, r')
\]  

and a specific objective function can be taken:

\[
Q(r, S) = I^*(r; S) - \alpha H^*(r|S)
\]  

which will be maximized for each \( r \) by tuning the similarity functions \( \varphi \).

6.2.2. Generalization with a metrics on the stimulus space

The previous version of the metrical entropy suffers from one main drawback. How does it behave when the stimuli become closer and closer? Indeed, the clusters of responses will necessarily overlap at some point. If the stimuli are continuous, how can one define the responses
elicited by a single stimulus? This is why it is necessary to introduce a metrics on the stimulus space to define what the task is and what stimuli must be actually discriminated.

A symmetric version of the metrical equivocation can be defined as:

$$H^{**}(R|S) = \sum_{r,s} p(r, s) \text{log} \frac{\sum_{r',s'} p(r', s') \varphi(\{r, s\}, \{r', s'\})}{\sum_{s'} p(s') \varphi(s, s')}$$

(6.8)

where the input-output similarity measure \(\varphi(\{r, s\}, \{r', s'\})\) is taken as the product of the two input and output similarity measures \(\varphi(r, r')\) and \(\varphi(s, s')\). Note that if we assume no correlation on the stimulus space, i.e. \(\forall \{s, s'\} \in S \times S \; s \neq s' \Rightarrow \varphi(s, s') = 0\), we return to the original metrical information theory.

The generalized metrical information is then equal to:

$$I^{**}(R; S) = H^{**}(R) - H^{**}(R|S) = \sum_{r,s} p(r, s) \text{log} \frac{\sum_{r',s'} p(r', s') \varphi(\{r, s\}, \{r', s'\})}{(\sum_{s'} p(s') \varphi(s, s'))(\sum_{r'} p(r') \varphi(r, r'))}$$

(6.9)

which is symmetric in \(r\) and \(s\).

The metrics on the response space was interpreted as the properties of the decoder. How can we interpret the metrics on the stimulus space? The answer is that it formalizes the discrimination goal of the reader. In case two stimuli are to be distinguished, their similarity measure should be equal to 0. Otherwise, it should be set to 1. Let us assume there are only two stimuli. If their similarity is 1, then the entropy on the stimulus space, defined as:

$$H^{**}(S) = \sum_{s} p(s) \text{log} \sum_{s'} p(s') \varphi(s, s')$$

(6.10)

is zero, and no information can be conveyed since there is no attempt to discriminate whatsoever. The symmetric metrical entropy formulation can be thought of a tool to define the best decoding scheme (response metrics) according to a given discrimination task (stimulus metrics).

Note that the similarity measure between stimuli \(\varphi(s, s')\) may take intermediate values as well, depending on the features of the stimuli that should be decoded in priority. In some cases, it may be impossible to discriminate fully different categories. It could then be appropriate to put the emphasis on some pairs by using a zero similarity measure and less emphasis on others with a \([0, 1]\) similarity measure.

### 6.3. Results

Let us consider a simple case with a few categories and responses lying in the set of real numbers \(\mathbb{R}\). For simplicity, let us assume a Gaussian distribution for each object \(i\), with mean \(\mu_i\) and standard deviation \(\sigma_i\). Let us simplify even more by considering the variances to be equal, \(\sigma_i = \sigma, \forall i\). Henceforth, optimal discrimination is studied by considering the objective function defined in Eq. 6.4 and by varying the trade-off parameter \(\alpha\). Recall that a large \(\alpha\) implies maximizing the metrical information and minimizing the metrical equivocation simultaneously, whereas small \(\alpha\) values relax the minimality constraint on the metrical equivocation.
6.3. Results

We will consider Gaussian similarity kernel functions \( \varphi(x, y) = e^{-(x-y)^2/2\beta^2} \), which are more realistic than Heaviside-like all-or-none functions. Different standard deviations \( \beta \) will be employed in order to modulate the selectivity of the similarity function (the lower \( \beta \), the more selective \( \varphi \)).

Figure 6.1.: Optimal discrimination in a low-noise regime. Five categories (blue lines) must be discriminated. The results were obtained for 4 different values of the trade-off parameter \( \alpha = 0, 10^{-7}, 10^{-5}, 1 \). For low values of \( \alpha \), the width \( \beta \) of the similarity measure at each point (black curve) is larger at the center of the categories and smaller between them (indicating higher and lower selectivity, respectively). As the trade-off parameter \( \alpha \) increases, the width becomes highest in-between categories. At \( \alpha = 10^{-5} \), the information (resp. equivocation) is still high (resp. already low) at the center of the categories.

Figure 6.1 presents the results obtained with four different values of the trade-off parameter, \( \alpha = 0, 10^{-7}, 10^{-5}, 1 \), in the presence of low-noise category distributions (blue curves). The optimal width \( \beta \) of the local similarity kernel is shown at each value of the response space (black curves). The specific information \( I^*(r; S) \), Eq. 6.5, and equivocation \( H^*(r|S) \), Eq. 6.6, are also plotted (solid and dashed red curves, respectively) to show the contribution of all the responses to the expectation values over the entire space. For \( \alpha = 0 \) (Fig. 6.1a), the goal is to maximize the information without taking into account the equivocation. The best solution is to have very narrow similarity functions so as to discriminate every pairs of responses: the width \( \beta \) is thus always zero. For low (positive) values of \( \alpha \), the width of the similarity function
can be large at the center of a category distribution since it decreases the equivocation without impinging on the information. This is apparent in Fig. 6.1b, corresponding to $\alpha = 10^{-7}$, where the values of the equivocation drop at the center of the categories compared to the $\alpha = 0$ case. On the other hand, to keep the information maximum, the kernel function $\varphi$ must be narrow in-between categories (increasing the selectivity of the decoder). Indeed, a higher selectivity in-between categories guarantees a high information, at the price of a large equivocation, since objects at the center of a category will not be considered similar to those at the edge. For $\alpha = 10^{-5}$, Fig. 6.1c, the equivocation is already almost nil at the center of the categories (with a full information), though it is far from being zero in-between them. However, the contributions to the equivocation of these events are small because their probability is low. The equivocation further decreases only when $\alpha$ is drastically increased (Fig. 6.1d, $\alpha = 1$). The price to pay is a loss of information in-between categories since now, objects belonging to some category are considered similar to objects of other categories. We here see the effect of the trade-off parameter at each point: in general, a response near the edge of a category contributes largely to the equivocation as long as its similarity with the center of the category (which has a high probability) is low. However, increasing its width impinge on the information. At high values of the trade-off parameter (e.g. $\alpha = 1$), the similarity measure is wider in-between categories than at their centers. The reason is that, for the specific equivocation to be zero (i.e. all objects from the same category considered identical), the similarity measure at the boundaries of a category needs to be twice as wide as that at the center of a category. If we arbitrarily give Shannon information and equivocation values of 1, the metrical information and equivocation are respectively:

| $\alpha$   | $I^*(R, S)$ | $H^*(R|S)$ |
|------------|-------------|-------------|
| 0          | 1.00        | 1.00        |
| $10^{-7}$  | 1.00        | 0.63        |
| $10^{-5}$  | 1.00        | 0.41        |
| 1          | 0.96        | 0.05        |

which shows that, in this case, it is feasible to decrease the equivocation without impinging on the information.
6.3. Results

Figure 6.2.: Optimal discrimination in high-noise regime. Five categories (blue) are to be discriminated. The results were obtained for 4 different values of the trade-off parameter $\alpha = 0, 10^{-7}, 10^{-5}, 10^{-1}$. For all values of $\alpha$, the width is largest in-between categories.
6. An information-theoretic account of the Perceptual Magnet

Figure 6.3.: Optimal discrimination in low- and high-noise regimes. When two categories (centered at 375 and 450) must not be distinguished. Top row: low-noise categories. Bottom row: high-noise categories. Left column: trade-off parameter \( \alpha = 0 \). Right column: \( \alpha = 10^{-5} \). Compared to the case where all categories had to be discriminated, Shannon information is lower. When the trade-off parameter is non-nil, and in the low-noise regime, the width of the similarity measure exhibits a dramatic increase between the two categories that must not be discriminated, whereas this increase is weaker in the high-noise regime.

Figure 6.2 shows that in the high-noise regime (large variance of the distributions) the optimal cut-off distances behave differently than in the low-noise regime. The optimal widths \( \beta \) of the similarity kernels tend to be large at the interface between classes and rather small near the centers. The situation in which the similarity measures are wide at the center of categories never occurs in the high-noise regime. As a consequence, it is almost impossible to decrease the equivocation without diminishing the information. If the Shannon information and equivocation are attributed values of 1, the information and equivocation are respectively:

| \( \alpha \) | \( I^*(R, S) \) | \( H^*(R|S) \) |
|----------|----------------|----------------|
| 0        | 1.00           | 1.00           |
| \( 10^{-5} \) | 1.00          | 0.94           |
| \( 10^{-5} \) | 1.00          | 0.69           |
| \( 10^{-1} \) | 0.98          | 0.19           |

The results of Figs. 6.1, 6.2 suggest that for low values of both noise level and trade-off parameter \( \alpha \), a phenomenon of high selectivity in-between categories and low selectivity at
6.4. Discussion

When attempting to maximize the objective function defined with respect to the metrical entropy on a set of Gaussian categories, two types of behavior were observed: the selectivity of the similarity measure can be larger either at the center of the distribution of each category or between categories. The first scenario can be linked to the perceptual magnet effect (Kuhl 1991), which has been first introduced in the field of psychoacoustics in the early 90s (Kuhl 1991). Each category of sound (for example each type of vowel) has a prototype, an element that best represents the vowel. The other elements of the category are perceived closer to the prototype than their physical distance would suggest, as if the prototype was pulling the elements of the category toward itself, hence the term ‘perceptual magnet’. The perceptual space is thus warped with between-category expansion and within-category compression.

The results presented here in the low-noise regime for low values of the trade-off parameter $\alpha$ can then be related to the perceptual magnet. Indeed, the width of the similarity measure at the center of a category is large, meaning that objects around are seen as very similar, while the width is small in-between categories, making the close objects dissimilar. In addition, the large values of the standard deviation of the similarity measure between the two categories that are not to be discriminated is also closely related to the perceptual magnet effect. In the case where many close categories are treated as the same one, the similarity measure is high in the bulk of the responses to them. The discriminability will be much lower in the center of this aggregate of categories than at their borders.

It is interesting to highlight the fact that the perceptual magnet may not be a highly generic phenomenon, but rather appear in limited ranges of parameters only. The above results suggest that it seems to be a low-noise regime effect in circumstances where the emphasis is on discrimination of different stimuli rather than identification of similar ones. We studied in more details the predominance of the two types of behavior (i.e. highest width inside categories or in-between them). We defined the amplitude of the perceptual magnet as the logarithm of the ratio between the width of the similarity measure at the exact middle between categories and at the exact center of categories. A highly negative value of this amplitude would correspond to a very pronounced perceptual magnet effect. The phase diagram of Figure 6.4 displays the amplitude of the perceptual magnet effect with respect to the noise level and to the trade-off parameters.
6. An information-theoretic account of the Perceptual Magnet

parameter $\alpha$.

Figure 6.4.: Perceptual magnet phase diagram. The perceptual magnet amplitude is defined as the logarithm of the ratio of the width of the similarity measure between categories and at the center of categories. This amplitude is plotted with respect to the noise (category width) and to the trade-off parameter. The perceptual magnet appears to occupy only a small portion of the phase space: in the low-noise regime and small trade-off parameter.

Figure 6.5.: Example of implementation of the map from stimulations to responses that may induce the Perceptual Magnet with 2 categories (red curves) and a population of tuning curves (blue lines). In the response space of the population of neurons the distance between $x = 2$ and $x = 4$ is equal to $d_1 \sim 2.8$, while that between $x = 4$ and $x = 6$ is $d_2 \sim 3.8$. Therefore, the categories are much smaller compared to their mutual distances.

The varying width of the similarity measure can be understood in two different ways. The first one is that the input space is isometrically represented in the nervous patterns and some neural decoder can perform the task of the similarity measure. Yet, it could also be the result
of a remapping through an ensemble of neurons with a non-uniform density (Bonnasse-Gahot and Nadal 2008) followed by a decoder using a uniform similarity measure. Bonnasse-Gahot and Nadal (2008) (Bonnasse-Gahot and Nadal 2008) presented a study based on Fisher information analysis and showed that a way to implement this mechanisms may be to have a higher receptive field density between the categories than within categories. In the light of metrical information analysis, this possible neural coding principle can interpreted the following way. The input space is mapped to the space of discharge rate of the neurons (i.e. tuning curves). In this space, categories are much more separated with large no man’s land between them. Then using a uniform similarity measure on this space yields the maximum objective function. Here, the optimal similarity measure is obtained in two steps: first an optimal neural coding that relatively contracts the categories with respect to the distances between them and then a decoding scheme (whose implementation is not given) that maximizes the quality factor. The simple example of Figure 6.5 suggest that the two approaches (Fisher and metrical information) yield similar results.

6.5. Conclusion

We have extended the metrical information to endow a metrics on the stimulus space and showed that it may be seen as a formalization of the task that is to be fulfilled. We have shown that the metrical information yields results that can be related to the perceptual magnet effect. It theoretically allows us to give estimates on the presence of this phenomenon in a two-dimensional phase space consisting of the noise level and the goal of the communication system (maximizing true positives or minimizing false positives). In the low noise regime, approaches based on Fisher information (Bonnasse-Gahot and Nadal 2008) yielded similar results if we make a straightforward link between the firing rates of a population of neurons and the distances they implement. It is indeed known that Fisher information is somewhat dependent on the metrics of the event space. This observation explains in part why both approaches yield similar results. But, more than this, we think that future work may help clarifying further the relations between metrical information and Fisher information and thus bring light on the relations between classical information theory and Fisher information (Brunel and Nadal 1998).

6.6. Appendix

In this appendix, we show that when the similarity measure on the stimulus space is a Dirac, then the full-metrical information reduces to the metrical information measure.

The stimulus entropy just reduces to the Shannon entropy:

$$H^{	ext{**}}(S) = \sum_s p(s) \log p(s) = H(S)$$

(6.11)
6. An information-theoretic account of the Perceptual Magnet

The joint entropy is then:

\[ H^{**}(R, S) = \sum_{r,s} p(r, s) \log \left( \sum_{r'} p(r', s) < r | r' > \right) = H^*(R, S) \]  \hspace{1cm} (6.12)

The conditional entropy:

\[ H^{**}(R|S) = \sum_{r,s} p(r, s) \log \left( \sum_{r'} p(r', s) < r | r' > \right) \quad \text{and} \quad H^{**}(R|S) = \sum_{r,s} p(r, s) \log \left( \sum_{r'} p(r'|s) < r | r' > \right) = H^*(R|S) \]  \hspace{1cm} (6.13, 6.14)

The information is then:

\[ I^{**}(R; S) = \sum_{r,s} p(r, s) \log \left( \frac{\sum_{r'} p(r', s) < r | r' >}{\sum_{r'} p(r') < r | r' >} \right) \quad \text{and} \quad I^{**}(R; S) = \sum_{r,s} p(r, s) \log \left( \frac{\sum_{r'} p(r'|s) < r | r' >}{\sum_{r'} p(r'|s) < r | r' >} \right) \]  \hspace{1cm} (6.15)
7

Galway entropy as a measure of Complexity

Where we show that the Galway entropy can be understood as a measure of entropy on graphs, whose edges are representations of the similarities between events. We illustrate it on a few simple graphs to exhibit its behavior. We then compare this measure with the graph entropy introduced by Körner and to other approaches to graph complexity. Finally, we compute the Galway entropy of a well-known graphs such as the Erdös-Renyi and Watts-Strogatz graphs.

7.1. Introduction

We have defined, in the previous chapters, a measure of the entropy of a set of events incorporating their relations. It is defined on a set of events with various probabilities that may be considered similar to one another. This similarity can take arbitrary values between 0 and 1, depending on the level of difference between the events. These elements can be seen as a graph whose vertices are the events and whose weighted edges are the similarities between events. Whereas the Galway entropy deals with graphs induced by a metric on the space of events, without loss of generality, we here consider graphs as starting objects and show that the Galway entropy can be naturally interpreted as an entropy on these graphs.

Let us consider a graph $G = \{V, E, W\}$ where $V$ is a set of vertices (or events), $E$ is a set of edges whose weights are given by $W$. The metric entropy is defined as:

$$H^*(G) = -\sum_i \sum_j v^i \log v^j w_{ij}$$

(7.1)

using balanced notations for summation.
7. Galway entropy as a measure of Complexity

7.2. Correspondance with an axiomatic approach to complexity

In 1999, Carter Butts proposed a set of axioms to define the complexity of a graph (with non-weighted bonds). He first proposed 4 initial axioms:

Axiom 1 (Cardinality): The complexity of a graph, $C(G)$, must take the set of labeled graphs into the real numbers.

Axiom 2 (Determinism): $\forall G, \exists a \in (R): p(C(G) = a) = 1$.

Axiom 3 (Existence): $C(G)$ exists for all finite graphs $G$.

Axiom 4 (Finiteness): For all finite graphs $G$, $\exists a \in (R): C(G) \leq a$.

Galway entropy verifies such basic axioms.

Butts went on and proposed 4 reasonable axioms that requires the definition of the smallest graph, the single point with no loop, that is called $K1$.

Axiom 5 (Floor value): $\forall GC(G) \geq C(K1)$

Axiom 6 (Labeling Insensitivity): In labeled graphs, the complexity should be independent of a particular labeling.

Axiom 7 (Monotonicity): $\forall H \subset GC(G) \geq C(H)$

Axiom 8 (Complementarity): $\forall GC(G) = C(\bar{G})$

Galway entropy verifies axiom 5, 6 and 7. However, it violates axiom 8. This should come as no surprise: Butts showed that measures verifying all these axioms are very simple (close to trivial). For example, the number of vertices verifies all 8 axioms.

7.3. Variation of the complexity under perturbation

Imagine I have 4 equiprobable events $A, B, C, D$ with $A = B$, $B = C$, $C = D$ and $D = A$. The entropy is then 1 bit. It means that to completely specify the member of the set, I need to give a single binary number.

Adding a single edge with weight $\alpha$ to a graph between two vertices of probabilities $p$ and $q$ that are already linked to other edges such that their confusions are $P$ and $Q$ induces a change $\Delta H$.

\[
\Delta H = H_2 - H_1
\]

\[
\Delta H = H + p \log P + q \log Q - H - p \log(P + \alpha q) - q \log(Q + \alpha p)
\]

\[
\Delta H = -p \log \frac{P + \alpha q}{P} - q \log \frac{Q + \alpha p}{Q}
\]

Obviously, the larger $\alpha$, the larger the decrease in entropy.

The Galway entropy can be understood as the average number of modules in the network.

Natural encompassing of binning It naturally encompasses the classical binnings but can deal with more complex binnings where there are no well-defined clusters and where similarities may take arbitrary values between 0 and 1.

Link with Graph Entropy Graph Entropy (Körner 1973; Simonyi 1995) is defined on a topo-
7.3. Variation of the complexity under perturbation

Figure 7.1.: Galway entropy for all the topologically different graphs with 4 vertices. In case, Shannon entropy could be defined as well, we added the label Sh.

logical set of vertices (events) that may be linked by an edge if they are distinguishable (Note that the graph defined here is complementary to the one described above). When such a graph is obtained, the graph entropy is defined as the minimum information between the vertices and the vertex packing polytope (i.e. the maximally independent sets of distinguishable vertices). The minimization process sometimes yields an entropy exactly similar to the Galway entropy presented here, though this is not always the case. In any case, the qualitative behavior of both entropies is the same, though there may be some small differences.
Our method has the advantage that, contrary to graph entropy, it can be expanded to similarities that take real values between 0 and 1.

Link with EAP In EAP (Symmetry based Structure Entropy of Complex Networks), the network is partitioned into sets of vertices having the same role in the network. For example, in a square \( \{A, B, C, D\} \) with links \( \{AB, BC, CD\} \), B and C have the same role, as well as A and D. We thus have two sets with the same number of elements in them, the entropy is \( \log 2 \).
If we had links \( \{AB, AC, BC, CD, BD\} \), we would have the same sets and thus the same entropy. The fact is that, once the partition is done, there is no dependence on the roles of each set.
7. Galway entropy as a measure of Complexity

Figure 7.2.: From left to right, examples of regular, Erdős-Renyi and Watts-Strogatz graphs.

In Galway entropy, each vertex contributes independently to the entropy, but two vertices with the same role will make the same contribution, this contribution depending on the role of this vertex.

**Interpretation of the Galway entropy.** The Shannon entropy of a set $X$ is usually interpreted as the average description length for a member of this set. This still holds with Galway entropy, though in a more subtle way. Let us consider 4 equiprobable events on an axis: $\{1, 2, 4, 5\}$. The entropy is 2 bits. If 1 and 2 are identical and 4 and 5 as well, the entropy is 1 bit. Now, if $\phi(1, 2) = \phi(2, 4) = \phi(4, 5) = 1$ but $\phi(1, 5) = 0$ (threshold for discrimination), then Shannon entropy is not defined. However, Galway entropy is. More interesting, in the first case, it gives an entropy of 2 bits. In the second case, it yields $-4 \frac{1}{4} \log \frac{1}{4} + \frac{1}{4} = 1$ bit. In the third case, the entropy is $-2 \frac{1}{3} \log \frac{1}{3} + \frac{1}{3} - 2 \frac{1}{3} \log \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 0.71$ bit. It seems highly relevant that by adding similarities between events, the entropy decreases. The problem of interpreting these values as Kolmogorov complexities remains. The fact is that the interpretation in terms of the average description length for a member of the set only holds for well-defined disconnected graphs (with so-called cliques or modules).

7.4. Results

7.4.1. Regular Graphs

A regular graph is a set of $N$ equiprobable vertices that are connected each to $K$ other vertices with edge weight 1. On these graphs, the Galway entropy is $H^* = \log \frac{N}{K+1}$.

- If $K = 0$, i.e. there are no edges, then $H^* = \log N$.

- If $K = N - 1$, i.e. the graph is fully connected, then $H^* = 0$.

The value of the Galway entropy on such graphs with varying numbers of edges (or probability of connection) is given on Fig. 7.3 (red line).
7.4. Results

Figure 7.3.: Galway entropy on Erdős-Renyi and regular graphs with 100 nodes and with varying probability of connection $p$. As expected, the entropy decreases with $p$.

7.4.2. Erdős-Renyi graphs

An Erdős-Renyi graph is a set of vertices with edges connecting them with a probability of presence of $p$. We computed the Galway entropy on these graphs for values of the probability between 0 and 1 on Fig. 7.3 (blue points). As $p$ increases, the Galway entropy decreases.

Note that the Galway entropy yields similar results for regular and Erdős-Renyi graphs.

7.4.3. Watts-Strogatz graphs

A Watts-Strogatz graph is a set of vertices along a circle that are at first connected to their $2k$ nearest neighbours. Then, each vertex is visited and its links are rewired the following way: remove one of the links with one of its nearest neighbor in the clockwise direction with probability $\beta$ and rewire it to another randomly chosen vertex (avoiding loops and duplications). Long-range connections are thereby introduced at a rate $\beta$. This was one the first methods to build so-called small-world networks.

We computed the Galway entropy on these graphs for values of $k = 4, 40, 80$ with varying value of $\beta$, see Fig. 7.4.

The Galway entropy distinguishes the various parameters of the small-world networks. In this sense, it allows for a characterization of these graphs.

The effect of rewiring can be computed analytically (at least partially). Starting from a regular graph with $N$ vertices connected to their $K$ nearest neighbors, $H^* = \log \frac{N}{K}$. Erasing a link $v_{ij}$ and creating another $v_{ik}$ induces a change

$$\Delta H^* = -\frac{1}{N} \log \frac{K}{K+1} - \frac{1}{N} \log \frac{K+2}{K+1}$$

(7.5)

i.e.:

$$\Delta H^* = -\frac{1}{N} \log \frac{K(K+2)}{(K+1)^2}$$

(7.6)

Note that the term in the logarithm is always smaller than 1, thus the change in the entropy
7. Galway entropy as a measure of Complexity

Figure 7.4.: Galway entropy on Watts-Strogatz graphs with 100 nodes as $\beta$ is varying. From top to bottom, $k = 4$, $k = 40$ and $k = 80$. As expected, the entropy decreases when $k$ increases. In addition, we see that the more we rewire the graph, the higher the entropy.

7.5. Discussion

We have here shown that the Galway entropy can be understood as a measure of complexity on graphs. We have in particular shown how it is linked with existing measures such as Shannon entropy (that is defined on a restricted set of graphs), graph entropy or EAP. This new entropy was computed on a few classes of well-known networks such as the Erdős-Renyi or Watts-Strogatz graphs. However, the Galway entropy is much more general and may be applied to graphs with edges between vertices that have intermediate values. It would be interesting, for example, to try to apply them to Penrose’s spin networks. These are sets of vertices that are
linked by edges labeled by elements of a spin group, for example, usually $SU(2)$. 
Part III.

Results on Somatosensory data
The hands? Why not? The eyes where nothing more than sense organs and the brain nothing more than a large standard in a box made of bones, isolated from the active surface of the body. These were the hands, the active surface of the body, the hands that were touching and manipulating the universe. Men were thinking with their hands. The hands responded to his curiosity, they touched and pinched and turned and lifted and weighted. There were animals with brains of significant size, but they were devoid of hands, and that was the difference.

Isaac Asimov

Anatomy of the Ascending Somatosensory Pathway

Where we give an introduction to the ascending somatosensory pathway. We provide a description of the tactile apparatus and its functional unit, the mechanoreceptor. We also recall how their characteristics can be unveiled using microneurography experiments. Finally, we give an overview of the projections of the mechanoreceptors to higher-order structures, focusing on the first relay to the central nervous system: the cuneate nucleus, one of the two dorsal column nuclei, located near the brain stem. Of particular interest to us are the electrophysiological properties of the cuneate neurons and the connectivity patterns with the mechanoreceptors.

8.1. General organization of the nervous system and the brain

The nervous system can be considered a rather large system with respect to its functional unit: the neuron. The brain of Sapiens is about 1400 cm$^3$ and it comprises around $10^{11}$ basic units. This large discrepancy between the size of a single microscopic unit and that of the macroscopic whole system is the basis for lots of mesoscopic levels of organization. Unlike a large number of particles in a box, the nervous system can indeed be decomposed in many modalities and anatomic areas (a reason for this may also be the fact that neurons are not all identical). This thesis is not the right place to list exhaustively these areas and their relationships. In particular, since we will mainly focus on the peripheral part of the nervous system. However, it is highly important to have a rough idea of the way things work up there to understand the

---

1I did not find an English version of the book, therefore this quote was translated back from French to English by myself.
behavior of the periphery. The brain is not a mere bunch of neurons connected together in a random fashion. Though they have large similarities, the behaviors of individual neurons are highly evolved. Though displaying a part of seemingly randomness, the connections are logically organized. These are two facts that should be taken into account seriously when studying the brain. What is the part of genomic and acquired information is unclear. However, the information needed to specify the whole brain is many order of magnitude larger than the amount of information contained in our genes.

This can probably linked with what Jacques Monod wrote in “Change and Necessity”: while you would need 2000 bits to specify the order and positions of all the atoms of a protein, DNA only gives a tenth of this. Physics and chemistry then makes the chain of atoms converge to a single conformation.

In this work, we study the somatosensory system (though we may sometimes go further and deal with the sensorimotor system). This system may be in a first approach described as ensembles of afferent neurons carrying information from the surface of the body to the central nervous system. They collect information that is carried and processed in higher-order structures. This allows the brain to provide relevant reactions with respect to the state and the changes of the environment. It is in reality a little bit more complex than this: some of the neurons are sensitive to proprioception, the perception of the position of the body. Some other respond to the state or changes of the visceral system.

8.2. Ascending somatosensory system

8.2.1. Cutaneous receptors

The epidermis is innervated with thousands of cutaneous sensory cells that respond to external stimulations. Most of these cells have nerve endings in direct contact with the tissue that are thus called free nerve endings. In contrast, some of them possess structures around the endings and are thus called encapsulated. All of these cells may be divided into three groups depending on the type of stimulations they are sensitive to:

- the mechanoreceptors respond to the mechanical indentations and deformations of the skin.

- the cutaneous thermoceptors are free nerve endings that signal the temperature level of the skin. There are cold detectors, sensitive to temperatures between 0 and 40 degrees Celsius; and warm detectors sensitive to temperatures between 30 and 50 degrees Celsius.

- the cutaneous nociceptors are also free nerve endings signal pain. They are high threshold receptors: they detect (too) high levels of stimulations.

Some of the sensory cells are called thermonociceptors as they are sensitive to very low or very high temperatures.
8.2. Ascending somatosensory system

8.2.2. Mechanoreceptors

Our efforts are focused on understanding the signals of the mechanoreceptors. Let us detail their characteristics. Unlike thermoceptors and nociceptors, mechanoreceptors are encapsulated. This means that their dendrites are not in direct contact with the tissue, but rather isolated in some structure. In the glabrous skin, there are four types of structure, leading to four types of mechanoreceptors (Fig. 8.3).

- Meissner’s corpuscles are responsible for light touch sensing and are selective to low frequency stimulation (below $50 \text{Hz}$).
- Pacinian corpuscles signal high pressure levels at high frequency.
- Merkel disks provide information about pressure and texture.
- Ruffini endings are responsible for signaling stretching of the skin.

Mechanoreceptors can also be classified according to functional properties such as the adaptation rate (fast-adapting, FA, and slow-adapting, SA), and the shape of the receptive field (small with sharp borders, type I, and large with smooth borders, type II). There is a one-to-one correspondence between the morphological and functional classifications (Fig. 8.4). Mechanoreceptors mostly carry information about fine touch (also called discriminative touch) as opposed to other types of cells in the dermis (called free-nerve endings because they are not encapsulated in any structure like the mechanoreceptors) which carry information about crude touch, temperature (thermoceptors), and pain (nociceptors). The information provided by mechanoreceptors and free-nerve endings is transmitted along two separate sensory pathways. Here, we will concentrate on the neural pathway processing the mechanoreceptor signals (Fig. 8.5).

In terms of cellular morphology, mechanoreceptors are pseudo-unipolar cells whose somas are located in the dorsal root ganglion near the spinal cord.
The behavioral responses to tactile stimuli (e.g., a stronger grasp when a cup is slipping from the hand) can be classified according to their time latencies (Deuschl and Lücking 1990; Kurtzer et al. 2008), which depend on the anatomical pathway subserving the processing of somatosensory signals and the generation of the motor commands. Very short-latency responses (≈ 25 – 50 ms) are likely to correspond to purely spinal reflexes mediated by the mechanoreceptors → spinal marginal nucleus → somatic motoneurons pathway. More complex responses (with latencies of ≈ 100 ms) are likely to involve the processing of tactile information at CNS level, and therefore the ascending somatosensory pathway through the cuneate nucleus relay.

Fig. 9.1 presents a sketch of the somatosensory pathway we studied. At the periphery, stimuli are encoded by mechanoreceptors that send spikes along axons that synapse (according to a connectivity scheme) cuneate nucleus (CN) cells. These synapses are plastic and may therefore undergo changes in their strength. The readout of CN discharges is also studied to take into account the information transmitted from the CN to downstream structures, such as the cerebellum.
8.2.3. Human microneurography data

A large set of microneurography recordings was employed to validate the developed encoding/decoding system. The microneurography experiments were conducted by R. Johansson’s group at UMEA University. In a nutshell, electrodes were percutaneously introduced near the elbow of an awake subject whose arm was preliminarily immobilized and nails glued to a fixed object in order to avoid proprioceptive effects on the discharge of mechanoreceptors. The tip of the electrode was introduced in the nerve and in particular very close to a single axon so as to be able to record spikes from a single afferent. The human tactile apparatus has been studied in detail since the advent of microneurography forty years ago (Vallbo and Hagbarth 1968). This technique allows one to record the response of a single afferent to various tactile stimulations. In particular, it allows to define precisely the receptive field of each mechanoreceptor (Johansson and Vallbo 1983), as well as many of its characteristics (fast or slow adapting, low or high threshold,...). It also allowed to count the number, density, overlap of the mechanoreceptors at each fingertip.

In the experiments conducted by R. Johansson, a highly precise device then applied tactile stimulations to the fingertip. The stimulation probes had various curvatures, they were applied with various (thoroughly controlled) forces, directions, angles, and speed profiles. A preprocessing study of the recorded spike train patterns permitted to classify the mechanoreceptor responses in one of the four aforementioned categories (FA-I, FA-II, SA-I or SA-II). The reader is referred to (Vallbo and Hagbarth 1968) and (Johansson and Birznieks 2004) for more details on microneurography recording methods. The original data set made available by UMEA included the recordings of 145 mechanoreceptors (of different types) in the presence of 111 distinct stimuli. Notice, however, that only a subset of all mechanoreceptor responses could be recorded for the entire set of stimuli. Indeed, it is nearly impossible to keep the electrode
Figure 8.5.: Highly simplified representation of the somatosensory pathways. In red is the one we focus on: the dorsal column medial lemniscus pathway. DRG: dorsal root ganglion, VPL: ventral posteriolarateral nucleus, MD: medial dorsal nucleus, VMP: ventral medial posterior nucleus.

really close to a specific afferent for long recording periods. For instance, it could happen that, after a while, the electrode slightly moved, thus breaking the sequence of recordings for a given mechanoreceptor.

In order to focus on rapid haptic discrimination mechanisms, we decided to concentrate on the responses of FA-I mechanoreceptors only. The stimulus state space was defined by considering a set of 4 primary input features (Fig. 8.6a):

- the curvature of the probe (n=3, i.e. \(0 m^{-1}, 100 m^{-1}, \text{and } 200 m^{-1}\)),
- the magnitude of applied force (n=3, i.e. \(1 N, 2 N, \text{and } 4 N\)),
- the direction of the force (n=5, i.e. ulnar, radial, distal, proximal, normal),
- the angle of the force relative to the normal direction (n=3, i.e. \(5^\circ, 10^\circ \text{or } 20^\circ\)).

In total, we were able to extract the responses of 42 FA-I mechanoreceptors to 81 distinct stimuli. To study how the information about all the contact parameters could be encoded &
8.2. Ascending somatosensory system

decoded very fast, we put the emphasis on the relative spike time correlations across multiple peripheral afferents. It is worth noting, that only the first spike of each recorded mechanoreceptor was taken into account in our study. Fig. 8.7 suggests that the variability in the first-spike latency domain with respect to the stimulus feature (e.g., the direction of the force) was much higher than the variability within repetitions of the same stimulus. Because we considered only the first spike of each mechanoreceptor response, each stimulus led to a single volley of spikes across all the mechanoreceptors, forming a response pattern defined by the first-spike latencies across the 42 afferents. The underlying working hypothesis was that rapid information transfer can prominently occur in the presence of spatio-temporal codes: spatial because the information is encoded in the neural population activity, and temporal because spike times are taken into account rather than firing rates.

8.2.4. Cuneate Nucleus

Mechanoreceptors send direct projections to both the marginal nucleus of the spinal cord and the cuneate nucleus, which is located behind the brain stem and is known to receive somatosensory afferents from the upper body (arms and trunk) – as opposed to the gracile nucleus, which is located next to the cuneate nucleus and receives somatosensory information from the lower body (mainly the legs). The marginal nucleus then sends axons to the cuneate nucleus and also
directly to the thalamus. The cuneate nucleus thus receives inputs from the mechanoreceptors both directly and indirectly via the spinal cord (Brown 1981; Kandel et al. 1991; Palecek 2004). Therefore, the cuneate nucleus constitutes an important synaptic relay of the ascending somatosensory pathway, and it projects to several areas of the central nervous system (CNS) including: (i) the thalamic ventrolateral posterior nucleus, which in turn projects to the primary somatosensory cortex, and (ii) the cerebellum, which in turn may elicit a motor command via the cerebello-spinal tract.

In the sequel, we will only consider the direct pathway from the mechanoreceptors to the cuneate nucleus, which is called the *dorsal column-medial lemniscus pathway*. This is the pathway that conveys information about fine (discriminative) touch.

Of prime interest to our work is the connectivity pattern. There are about 2000 mechanoreceptors at each fingertip. Data from the cat (Jones 2000) give a divergence of 1700 and a convergence of 300. This leads to a number of 11000 cuneate nucleus cells. The connection probability between a mechanoreceptor and a cuneate neuron is around 0.18.

As for the electrophysiological properties of the individual cuneate neuron, our collaborators in Lund, Fredrik Bengtsson and Henrik Jörntell have performed recordings in the awake cat. They found that the resting potential of these neurons is $V_r = -53\, mV$ and the threshold potential is $V_{th} = -50\, mV$. The time constant of the membrane is roughly $6\, ms$. 

---

Figure 8.7.: Characteristics of the mechanoreceptor spiking activity. (a) Responses of a FA-I mechanoreceptor to 5 trials of 5 stimuli with 5 different orientations of the force applied (P=proximal, U=ulnar, D=distal, R=radial, N=normal). (b) Probability density of a given jitter for a population of mechanoreceptors. Taken from (Johansson and Birznieks 2004).
Where we apply the proposed information theoretical approach, we study precise temporal decoding of human somatosensory signals recorded via microneurography experiments. For this analysis, we employ a similarity measure based on the Victor-Purpura spike train metrics. We show that, with appropriate parameters of this distance, the relative spike times of the mechanoreceptors responses convey enough information to perform optimal discrimination—defined as maximum metrical information and zero conditional entropy—of 81 distinct stimuli within 40 ms of the first afferent spike.

9.1. Haptic encoding/decoding based on relative spike timing

Recent microneurography studies in humans (Johansson and Birznieks 2004) suggest that the relative timing of impulses from ensembles of mechanoreceptor afferents can convey information about important contact parameters faster than the fastest possible rate code, and fast enough to account for the use of tactile signals in natural manipulation. Specifically, the relative timing of the first impulses elicited in individual members of ensembles of responsive afferents reliably conveys information about the direction of fingertip force and the shape of the surface contacting the fingertip (Johansson and Birznieks 2004). Changes in either of these contact parameters differentially influence the first-spike latency of individual afferents rather than having systematic effects on the latencies within an afferent population. For example, if the fingertip contacts a certain surface shape, the responsive afferents will be recruited in a particular order. With another shape, some afferents are recruited earlier, and others later because the fingertip deforms in another manner, leading to a different recruitment order. Likewise, with a single stimulus delivered to the fingertip in a given direction, the responsive afferents will be recruited in a particular order and with another force direction, the order will be different because some afferents are recruited earlier, and others later. The influence of changes in contact parameters on the first spike latencies is on average 7 to 15 ms across the mem-
bers of the afferent populations. Because these effects are large compared to the variability in first-spike latencies across repetitions of the same stimulus (s.d. = 1 ms), changes in contact parameters influence reliably the sequence in which different members of populations of tactile afferents discharge in response to discrete fingertip events. However, besides low noise in tactile afferents (Edin et al. 1995; Wheat et al. 1995; Vega-Bermudez and Johnson 1999; Johansson and Birznieks 2004) the effectiveness of a code based on relative spike timing for fast encoding of complex spatial features of fingertip events requires that the afferents are recruited at an adequately high speed. This is ensured by the combination of high density of afferents in the fingertips (Johansson and Vallbo 1979b), and variation in threshold sensitivity from afferent to afferent (Knibestol 1973; Knibestol 1975; Johansson and Vallbo 1979a; Johansson et al. 1980). Estimates of how early the sequence of recruitment of afferents contains information useful for discrimination of contact parameters (see rank-order code, (Thorpe et al. 2001)), indicate that the FA-I population provides correct discrimination faster than the SA-I afferents of both surface curvature and force direction. Reliable discrimination can take place after as few as some five FA-I afferents are recruited, which can occur a few milliseconds after the appearance of the first impulse in the population response. A correspondingly reliable discrimination by the SA-I population, requires recruitments of about twice as many (but still very few) because of a larger variability in first-spike latencies. Even under the most favorable conditions, discrimination based on firing rates (first-interspike rate) takes on average 15 to 20 ms longer than discrimination based on first spike latency. Populations of SA-II are much poorer in fast discrimination of force direction or object shape based on both first spike latencies and firing rate, in part because of the background discharge in many of these afferents. The relative timing of impulses presumably also contains early information about other contact parameters critical for control of manipulation. For example, the fast adjustment of force coordination to the friction between the skin and an object soon after it is contacted presumably occurs because the contact responses in different FA-I afferents are differentially influenced by surface friction (Johansson and Westling 1987). Information contained in relative spike timing in ensembles of tactile afferents is also likely to rapidly encode various spatial features of contact events that occur between held objects (e.g., tools) and other objects during manipulation tasks.

Encoding and decoding of sensory information based on the timing of neural discharges, rather than (or in addition to) their rate, has received increasing attention in the past decade and it has become clear that the timing, with precision in the order of milliseconds or even less, of individual spikes can carry far more information than average rates alone (Heil 2004; VanRullen et al. 2005). In particular, the high information content in timing of the first spikes in ensembles of central neurons is emphasized in several sensory modalities, including the auditory (Furukawa et al. 2000; Nelken et al. 2005), visual (Gawne et al. 1996; Reich et al. 2001b), and somatosensory (Panzeri et al. 2001; Petersen et al. 2001; Petersen et al. 2002b; Petersen et al. 2002a) systems. If relative spike timing is fundamental for rapid encoding and transfer of tactile events in manipulation, then how do neurons read out information carried by a temporal code? Various decoding schemes have been proposed to discriminate between different spatiotemporal sequences of incoming spike patterns. Many of these include central neural mechanisms that quantify first-spike latency information under the assumption that the brain has an independent and precise reference for stimulus onset from which to extract latency (Gerstner and Kistler 2002). Although object manipulation is built on proactively controlled and well-demarcated sequential action phases, it is unlikely that the brain can predict contact
9.2. Results: temporal decoding of human microneurography haptic signals

In order to validate the metrical information analysis presented above, we study the temporal decoding of the responses of fingertip mechanoreceptors to tactile stimuli. Mechanoreceptors innervate the epidermis and discharge as a function of the mechanical indentations and deformations of the skin. Recent microneurography studies in humans (Johansson and Birznieks 2004) suggest that the relative spike timing of mechanoreceptor responses can convey information about contact parameters faster than the fastest possible rate code, and fast enough to account for the use of tactile signals in natural manipulation (Johansson and Flanagan 2009).

Human microneurography dataset

To investigate fast encoding/decoding of tactile signals, we concentrate on the responses of fast adapting (FA-I) mechanoreceptors only (Johansson and Birznieks 2004). The stimulus state space is defined according to a set of four primary contact parameters:

- the curvature of the probe \( C = \{0, 100, 200\} \text{ m}^{-1}, |C| = 3 \),
9. Metrical Organization of Mechanoreceptors Signals

- the magnitude of the applied force \( F = \{1, 2, 4\} \text{ N}, |F| = 3 \),
- the direction of the force \( O = \{\text{Ulnar, Radial, Distal, Proximal, Normal}\}, |D| = 5 \),
- the angle of the force relative to the normal direction \( A = \{5, 10, 20\}^\circ, |A| = 3 \).

The overall dataset includes the responses of 42 FA-I mechanoreceptors to 81 distinct stimuli.

**Metrical information analysis on a limited input-output space**

First, we apply the metrical information analysis to decode a limited set of microneurography data: only the 5 force directions (ulnar, radial, distal, proximal, normal) are considered as variable primary features (Johansson and Birznieks 2004; Saal et al. 2009, similar to). Each of the 5 stimuli is presented 100 times, and the VP distances \( D_{VP} \) are computed across the population of 42 mechanoreceptor afferents.

Here, we focus on the first spike of each FA-I mechanoreceptor, according to the hypothesis that the variability in the first-spike latency domain with respect to stimulus feature (e.g. the direction of the force) is larger than the variability within repetitions of the same stimulus (Johansson and Birznieks 2004). Thus, each tactile stimulus consists of a single volley of spikes forming a spatiotemporal response pattern defined by the first-spike latencies across the afferent population (see Figure 9.2a for three sample recordings).

Figure 9.2b shows how the intra- and inter-stimulus distance distributions evolve over time. Within 25 ms of the the first mechanoreceptor discharge (and within 55 ms of stimulus onset), the critical cutoff \( D_{\text{critic}} = 6.2 \) of the Heaviside function can be set, which ensures that the perfect discrimination condition (i.e. maximum \( I^*(R; S) \) and zero \( H^*(R|S) \)) is met.

The red curves of Figure 9.2c confirm this result, by showing that the \( I^*(R; S) \) (continuous red curve), computed based on the global cutoff \( D_{\text{critic}} = 6.2 \), saturates at \( \log_2(5) \) within 25 ms of the first spike arrival, while the condition entropy \( H^*(R|S) \) (dashed red line) is nil. The green curves of Figure 9.2c show the time course of the local metrical information. Convergence to the perfect discrimination condition does not occur earlier than in the case of the global \( I^*(R; S) \) (red curve). The reason is that the clusters happen to have similar sizes, therefore all the local critical distances \( D = \{D_{\text{critic}}\} = \{5.3, 4.8, 4.8, 4.7, 6.2\} \) happen to be close to the global critical distance \( D_{\text{critic}} = 6.2 \). The blue curves of Figure 9.2c demonstrate that Shannon mutual information (blue continuous line), computed by considering a temporal binning of 1 ms resolution, increases faster than the metrical information. However, the conditional entropy (dashed continuous line) diverges.

Figure 9.2d displays two samples of distance matrices indicating how the input spike waves across the 42 mechanoreceptor afferents are clustered by the decoding system over time. Before the occurrence of the perfect discrimination condition (at 40 ms, left matrix) different stimuli can have relatively small distances, i.e some interferences can impair the decoding process. After 60 ms (right matrix), all the initially overlapping contexts become pulled apart, which removes all interferences across inputs and leads to a 100% accuracy in the discrimination process.
9.2. Results: temporal decoding of human microneurography haptic signals

Figure 9.2.: Information theoretical analysis on a limited stimulus set. The first spike latency patterns of \(|S| = 5\) distinct stimuli, each of which presented 100 times, are considered by this analysis. (a) Three examples of human microneurography FA-I recordings showing the first spike waves across 42 afferents evoked by 3 distinct tactile stimuli (while varying the orientation parameter only). (b) Evolution of the intra- and inter-stimulus distances as spikes flow in. The perfect discrimination condition is met about 40 ms after the stimulus onset, when the critical cutoff \(D_{\text{critic}}\) can be determined. (c) Shannon MI (blue curves), global metrical information (red curves) and local metrical information (green curves) over time. (d) Distance matrices before (left) and after (right) the occurrence of perfect discrimination. Only 10 presentations per stimulus were considered to generate these matrices. Whereas at 40 ms some interferences exist in the input-output mapping, all the events are well separated at 60 ms.
Metrical information analysis on an extended region of the input-output space

First spike wave analysis

We then scale up the analysis to the entire set of microneurography recordings (81 distinct stimuli encoded by 42 mechanoreceptor spike trains). Figure 9.3a compares Shannon mutual information against both the global and local versions of metrical information. In this case, only first spike latencies are considered by the analysis. Again, Shannon MI (blue continuous line) reaches its maximum faster than the metrical information, but the drawback is the large conditional entropy (blue dashed line). Global metrical information (red line) shows that perfect discrimination \( I^*(R; S) = H^*(R) \) is reached within about 40 ms of the first afferent spike. The local version of metrical information (green line) converges faster than the global measure. This can be explained by the distribution of the local critical distances (i.e. the sizes of the event clusters) shown on Figure 9.3b. The global version of the \( I^* \) uses a critical distance \( D_{\text{critic}} = 6.8 \), which is much too large for some stimuli. By contrast, the local version uses the whole distribution of critical distances \( D = \{1.7, ..., 6.8\} \) and it is therefore more adapted to the configuration of the input-output relationship.

It is worth noting that the time needed to achieve perfect discrimination is longer compared to the 5 stimulus case (40 vs. 25 ms) not only because of the larger input set but also because some stimuli are rather difficult to separate in the \( D_{VP} \) distance space (e.g. two skin indentations with the same curvature, force, direction parameters and with only a 5° difference on the angle). Figure 9.3c illustrates two examples of distance matrices largely (left) and shortly (right) before convergence.
9.2. Results: temporal decoding of human microneurography haptic signals

Figure 9.3.: Information theoretical analysis on the entire stimulus set. The first spike latency patterns of \(|S| = 81\) distinct stimuli, each of which presented 100 times, are considered by this analysis. (a) Shannon MI (blue lines), global (red lines) and (green lines) local metrical information measures as spikes flow in. (b) Distribution of critical distances \(D\). (c) Distance matrices at 40 ms (left) and at 65 ms (right), i.e. slightly before the occurrence of the perfect discrimination condition (~ 80 ms) (for sake of clarity, these matrices were obtained on the basis of 10 presentations per stimulus). Bottom: sequence of stimulus presentations according to the combinations of all contact parameters: curvature of the probe (0 m\(^{-1}\), 100 m\(^{-1}\), and 200 m\(^{-1}\)); angle of the force relative to the normal direction (0°, 5°, 10°, 20°); magnitude of the force (1 N, 2 N, 4 N); direction of the force (Normal, Radial, Distal, Ulnar, Proximal).
9. Metrical Organization of Mechanoreceptors Signals

**Full spike train analysis**

In a subsequent analysis, the complete spike train responses of the 42 recorded mechanoreceptors are taken into account, rather than the first spike waves only. Figure 9.4a shows an example of population spike latency pattern across 42 mechanoreceptors. When comparing again Shannon MI vs. global and local metrical information (Figure 9.4b), we find that the time course of Shannon MI (blue line) does not change significantly compared to the 1st spike latency case.

Interestingly, the global metrical information measure (red line) is impaired when taking into account the entire spike train. A plausible reason may be that the $D_{VP}$ distance scales up with the number of spikes: a stimulus with a large force amplitude (i.e. 4 N) elicits many spikes per mechanoreceptor, making the size of the cluster corresponding to this stimulus large. Meanwhile, the distances between clusters corresponding to stimuli with low forces (1 N) remain small. As suggested by Figure 9.4b, it is more efficient to use the local version of $I^{*}$ (green line), which behaves similarly to the 1st spike wave case.

![Figure 9.4: Information theoretical analysis on the complete stimulus set. The spike latencies of the entire spike trains, i.e. $|S| = 81$ distinct stimuli each presented 100 times, are considered by this analysis. (a) An example of human microneurography FA-I recordings showing the entire spike trains of 42 afferents. (b) Evolution of Shannon MI (blue lines), global (red lines) and (green lines) local metrical information measures over time.](image-url)
9.2. Results: temporal decoding of human microneurography haptic signals

Detecting regularities and isometric mapping

Since the time necessary to reach the optimal discrimination is the same when dealing with the first spike wave or the whole spike trains, we may wonder what are the contributions of the second and following spike waves. The results shown in Figure 9.5 highlight a meaningful property of the whole spike trains. If we employ the Victor-Purpura distance by taking $C_{VP} = 0$, and we measure the distances between the mechanoreceptor responses to stimuli with force amplitudes 1 and 2 N, i.e. $D_{VP}(1N, 2N)$, then to stimuli with force 2 and 4 N, i.e. $D_{VP}(2N, 4N)$, and finally to stimuli with force 1 and 4 N, i.e. $D_{VP}(1N, 4N)$, the following relation is verified for any set of the other free contact parameters: $D_{VP}(1N, 2N) + D_{VP}(2N, 4N) = D_{VP}(1N, 4N)$. This result holds when both only the first spike waves are considered (Figure 9.5a) and the entire spike trains are analyzed (Figure 9.5b). This means that the one-dimensional stimulus space is mapped onto a non-curved one-dimensional response space. This “alignment” property can be suitable to dissociate the problem of decoding the force of the stimulus from that of determining other features of the stimulus (i.e. pruning of the search state space).

Figure 9.5b also shows that when considering the entire spike trains, the distances between the outputs tend to reflect the distances between the stimuli. Indeed, $D_{VP}(2N, 4N)$ is approximately two times $D_{VP}(1N, 2N)$, which suggests an isometry-like property of the decoding measure. This result can be interpreted as follows: temporal decoding based on first spike latencies allows complete stimulus discrimination to be achieved very rapidly. Yet, taking into account the spike timing of the entire responses may add to perfect discrimination the possibility of capturing some regularities of the input space (e.g. isometric mapping). Not only are the responses organized so that a quick discrimination is possible, but in addition, there are logically organized so that the mapping from the stimulus space to the response space is a simple transformation. This may be the basis for the ability to extrapolate or to generalize the reconstruction of the stimulation when previously never-encountered stimulations are presented.
9. Metrical Organization of Mechanoreceptors Signals

Figure 9.5.: Time course of the $D_{VP}$ distances. The distance between the mechanoreceptor responses to stimuli with force amplitudes 1 and 2 N, i.e. $D_{VP}(1N, 2N)$, then to stimuli with force 2 and 4 N, i.e. $D_{VP}(2N, 4N)$, and finally to stimuli with force 1 and 4 N, i.e. $D_{VP}(1N, 4N)$. For each combination, all the other contact parameters are varied. (a) Results obtained when only the first spike of each mechanoreceptor is considered. (b) Results when the complete spike train is taken into account. In both cases, the equality $D_{VP}(1N, 2N) + D_{VP}(2N, 4N) = D_{VP}(1N, 4N)$ holds through time.
Where we design a model of the Cuneate Nucleus in order to assess its capacity to convey the metrical information of signals from the mechanoreceptors. This model is a simple feedforward layer of Spike-Response Models. The electroresponsiveness properties of the cuneate neurons as well as the connectivity patterns from the mechanoreceptors to the cuneate cells is fitted to experimental data. The synaptic efficacies are modulated according to a plasticity rule designed to maximize Mutual Information. After a learning phase during which the microneurography data were fed to the nucleus, the capacity of our model to transmit metrical information was probed. We found that the perfect discrimination condition at the output of the CN is reached only 5 ms later than at its input. This shows that a simple model is able to convey a high metrical information. In addition, the optimal distribution of synaptic weights turned out to be highly bimodal, with only 3 to 6 non-silent synapses. This surprising result is in agreement with recent recordings performed in the Cat Cuneate Nucleus.

10.1. Spike-Response Model

Single unit discharges at the level of the cuneate nucleus (CN) were modeled based on the spike-response model (SRM) (Gerstner and Kistler 2002). The SRM formalism provides a linear probabilistic neuronal model, as opposed to the more classical integrate-and-fire model which is non-linear and deterministic. Compared to the Hodgkin-Huxley formalism, the SRM permits a higher transparency and controllability of all free parameters (e.g. synaptic integration time constant, amplitude and shape of excitatory post-synaptic potentials, and so on).

In order to describe it, let us consider a neuron that receives inputs from N mechanoreceptors (labeled $j$) that fire at times $t^i_j$. We assume that each input spike induces a stereotyped excitatory postsynaptic potential of time course $\epsilon(t)$ whose amplitude is modulated by a synaptic weight $w_j$. The membrane potential of the neuron under study is then:
10. Cuneate Nucleus Model

\[ V(t) = V_{\text{rest}} + \sum_{j=1}^{N} w_j \sum_{i=1}^{M} \epsilon(t - t_j^i) \]  

(10.1)

where \( V_0 \) is the quasi-threshold, i.e. the potential at which the probability of firing is 1/2.

We can then define the probability of firing as a function of both the instantaneous firing rate \( g(t) \) and the refractoriness function \( R(t) \).

\[ \rho(t) = 1 - \exp[-g(V(t))R(t)] \]  

(10.2)

The refractoriness is a function with values between 0 and 1 that represents the absolute or relative impossibility to fire again after a first discharge:

\[ R(t) = \frac{(t - \hat{t} - \tau_{\text{abs}})^2}{\tau_{\text{rel}}^2 + (t - \hat{t} - \tau_{\text{abs}})^2} \Theta(t - \hat{t} - \tau_{\text{abs}}) \]  

(10.3)

This is discretized in: the time step is \( \Delta t \) and we note \( t^k = k \Delta t \), where \( k \in \mathbb{N} \)

\[ V(t^k) = V_{\text{rest}} + \sum_{n=1}^{N} w_n \sum_{i=1}^{M} \epsilon(t^k - t^i)x_n^i \]  

(10.4)

\[ \rho^k = 1 - \exp[-g(u(t^k))R(t^k)\Delta t] \]  

(10.5)

Figure 10.1.: Responsiveness and firing properties of a SRM neuron. Input current, membrane potential, probability of firing, refractoriness and spiking activity.

The CN network is modeled as a population of SRM units. The connectivity layout of
the mechanoreceptor-to-CN projections is based on neuroanatomical data reported by (Jones 2000), which suggests an average divergence/convergence ratio of 1700/300. This asymmetry coupling is in favor of a fast feedforward encoding/decoding process occurring at the CN network level. In order to test the hypothesis of a purely feedforward information transfer at the CN level, no collateral projections between CN neurons are considered in the current version of the model. Based on this divergence/convergence data, and given that there are around 2000 mechanoreceptors at each fingertip (and that the CN is somatotopically organized at least to the precision of the finger), there must exist around 12000 CN neurons coding for the tactile information coming from each fingertip. This data suggests a probability of connection between a mechanoreceptor and a CN cell of 0.18. The weights of these synapses are initialized uniformly and randomly between 0 and 1.

The efficacy of CN synapses was modulated according to a spike-timing-dependent plasticity (STDP, (Bi and Poo 1999; Markram et al. 1997)) rule specifically developed for the SRM (Toyoizumi et al. 2005). This learning rule optimizes the information transmission property of a single SRM neuron, it accounts for the coincidence detection across multiple afferents, and provides a biologically-plausible principle that generalizes the Bienenstock-Cooper-Munro (BCM) rule (Bienenstock et al. 1982) to spiking neurons. We adapted the learning rule developed by (Toyoizumi et al. 2005) to very short transient stimuli – to focus on the first spike latencies of the mechanoreceptor signals –, and we applied it to maximize the information transfer at the level of a neural population (rather than a single SRM neuron). The training phase consisted of 200 presentations of the sequence of 81 stimuli with a linearly decreasing learning factor.

10.2. results

Figure 10.2.: Distance matrix at the output of the Cuneate Nucleus.
10. Cuneate Nucleus Model

Figure 10.3.: Information $I^*(R; S)$ and conditional entropy $H^*(R|S)$ over time. The CN population consists of 50 cells. The 81 tactile stimuli are presented 100 times each. (A) Optimal discrimination is reached 35 ms after the first afferent spikes. (B) If the perfect discrimination constraint is relaxed by reducing the critical distance, then the system can perform partial discrimination (80% of maximum $I^*(R; S)$ and non null $H^*(R|S)$) already within 15 ms after the first spike time. (C-D) Distributions of intra- and interstimulus distances (computed 100 ms after stimulus onset) before and after training, respectively. (E) Distribution of synaptic weights of 10000 cuneate neurons (the total number of null synapses, absent and silent, is shown in gray) after learning.

Second, the entire set of microneurography recordings (81 stimuli) was employed to analyze the information transmission properties of a network of 50 CN neurons in the presence of synaptic adaptation. To compute $I^*(R; S)$, the VP distances $D_{VP}(r, r')$ between any two CN population responses $r, r'$ were considered. Again, the distance $D_{critic}$ was used to identify the perfect discrimination condition, and the VP cost parameter $C_{VP}^* = 0.1$ yielding the fastest perfect discrimination was selected. Fig. 10.3A shows that optimal haptic context separation is performed by the CN population after 35 ms from the arrival of the first presynaptic spikes. Selecting the optimal value of the critical distance, as done for Fig. 10.3A, corresponds to the situation in which a readout system downstream the CN would need a complete separation of haptic percepts (e.g. for highly precise feature recognition). Relaxing this optimality requirement (e.g. to the extent of very rapid though less precise reactions) can further speed up the discrimination process. For instance, Fig. 10.3B indicates that setting $D_{critic}$ to a suboptimal value would lead to the partial discrimination condition in which 80% of the maximum $I^*(R; S)$
10.2. results

Figure 10.4.: Time necessary to discriminate the entire set of 81 stimuli perfectly as a function of the size of the CN population. Each stimulus is presented 100 times. The average numbers of spikes emitted by each CN neuron when optimal discrimination occurs are also indicated in the diagram.

(with non-null $H^+(R|S)$) can be achieved within 15 ms after the first presynaptic spike arrival.

Figs. 10.3C-D illustrate the distributions of intra- and interstimulus distances 100 ms after stimulus onset before and after learning. It is shown that while the distributions are well-separated after learning, they are still largely overlapping before training (implying the impossibility of a perfect discrimination). It is also interesting to note that after (resp. before) learning the CN fired on average $n=217$ (resp. 39) spikes, and that the maximum intrastimulus distance was about $D_{\text{VP}}^{\text{max}}=14$ (resp. 45). The average uncertainty on the timing of a single spike can be expressed by $\Delta t = D_{\text{VP}}^{\text{max}} / C_{\text{VP}} n$. Since $C_{\text{VP}} = 0.1$, $\Delta t = 0.6$ ms after learning and $\sim 12$ ms before. This shows that the plasticity rule helped reducing the jitter on CN spikes, thus reducing the metrical conditional entropy compared to the pre-learning condition.

Fig. 10.3E suggests that the plasticity rule leads to stable weight distributions that are invariant with respect to initial random conditions (uniform distribution between $[0,1]$). After learning, the synaptic efficacies of the mechanoreceptor-to-CN projections converge toward a bimodal distribution with one peak close to zero and the other peak close to the maximum weight.

A third analysis was performed to study how the size of the CN network may influence the discrimination capacity of the system. This analysis reveals that a lower bound on the number of CN neurons exists in order to perform optimal encoding of the 81 microneurography spike trains. As shown in Fig. 10.4, the perfect discrimination condition cannot be met with a population of less than 50 CN neurons. This result corroborates the hypothesis that a spatiotemporal population code is a necessary condition to perform optimal (i.e. both fast and reliable) context separation of complex spiking signals (Furukawa et al. 2000; Gollisch and Meister 2008). By increasing the number of neurons, the discrimination becomes faster and it saturates at 72 ms (which corresponds to the time at which the first spike from the slowest volley of pulses arrives at the CN). It is also shown that the average number of spikes emitted...
by CN cells under the optimal discrimination condition decreases from 2.1 to 1.3 with the size of the CN population, supporting the idea that one spike per neuron is enough to convey a significant amount of information.

10.3. Discussion

This study shows how a population of model 2nd-order somatosensory neurons can optimally encode incoming spike trains (obtained via microneurography recordings in humans) by separating them in a metrical abstract space. A large body of work has focused on measuring information transfer of spiking signals based on Shannon’s mutual information (MI), either directly (which is often impracticable due to the size of the response space and to the jitter in the spike times) or with classification algorithms such as a k-nearest neighbor coupled with confusion matrix decoding schemes (Victor and Purpura 1996). Here, we put forth a novel entropy definition to assess the efficiency of spiking neurotransmission in the presence of a large output space and with a 1 ms temporal precision. This measure allows us to corroborate the hypothesis that the CN has the ability to transmit to the central nervous system an optimal account of the peripheral haptic stimulations.

10.4. A few synapses are optimal for information processing

Fredrik Bengtsson and Henrik Jörntell in Lund performed electrophysiological experiments during which they recorded from neurons in the Cuneate Nucleus of the cat. They made several measurements that may turn out to be very relevant with respect to the information processing of this nucleus. One of the main observation is that there are only 5 to 10 synapses that have actually significant weights. This is a very surprising result when compared to the observation by Jones (Jones 2000) that each cuneate cell synapses with around 300 receptors. This means that 95% of all the synapses are actually near-silent. In terms of information processing, this does not appear straightforwardly like the optimal case.

However, with the model we have described in this chapter, we conducted a few experiments. We simulated a single cuneate neuron receiving inputs from the 42 mechanoreceptors recorded by Roland Johansson. The synaptic weights were initially drawn from a random flat distribution between 0 and 1. Their efficacies were modulated according to the MI-based learning rule as above. At the end of the simulation, as few as 3 synapses turned out to be non-silent. In Fig. 10.6a, we show the evolution of the number of effective synapses as the simulation takes place, as well as the mutual information. The latter increases significantly as the number of synapses decrease.

As suggested by Jones, the number of existing synapses is close to 300. However, we only had the responses from 42 mechanoreceptors to 81 stimuli or from 131 mechanoreceptors to 27 stimuli. We thus simulated cells receiving inputs from an increasing number of receptors. The number of non-silent synapses at the end of the simulation naturally increased, but only slightly (see Fig. 10.6b). We extrapolated to 300 mechanoreceptors and obtained a value close to 8, in agreement with Bengtsson and Jörntell’s observations.

Another analysis was carried out to study the effect of the structural connectivity on the results. Let us consider a population of cuneate neurons that receive signals from a population
10.4. A few synapses are optimal for information processing

![Graph showing the relationship between the probability of connection and mean cumulative information.](image)

Figure 10.5.: Summed information transfer as a function of the connection probability between the mechanoreceptors and the cuneate cells.

of mechanoreceptors. If the probability of connection is 0, then no information can be transmitted. If the probability of connection is 1, then all the cuneate neurons tend to focus on the same mechanoreceptors and therefore become highly redundant. Provided that it is improbable that a single cuneate neuron can transmit unambiguous information about 81 stimuli, this population will not transmit a full information (i.e. equal to the entropy of the stimulus set). There must therefore be an optimal value of the probability of connection that is strictly comprised between 0 and 1. At this value, each cuneate cell would receive signals from enough mechanoreceptors to choose those that convey information and from not too much neurons so as not to be redundant.

We launched many simulations with the whole spike trains elicited by mechanoreceptors. In order to estimate the information transfer with a single number, we summed the information over time. This makes neurons that transmit late information not very efficient. The plasticity rule was this time standard STDP. We plotted this summed information as a function of the probability of connection on Fig. 10.5. The results suggest that there is an optimal value for the probability of connection in the range [0.1, 0.4] with a peak around 0.2. It is of prime interest to compare this range with the observations of connectivity patterns in the cat CN by Jones (Jones 2000). He observed a divergence of 1700 and a convergence of 300, which leads to a probability of connection between a mechanoreceptor and a Cuneate cell of 0.18.
10. Cuneate Nucleus Model

Figure 10.6.: Simulation of the impact of synaptic efficacy distribution on information transfer. (a) Relationship between the number of effective synapses (mediating 100% unitary EPSP-to-spike coupling) and the information transfer in the primary afferent-to-cuneate neuron synaptic relay. The number of effective synapses gradually decreased as the result of a simulated process shaping the weight distribution of primary afferent synapses on the cuneate neuron from random to bimodal (Hartigan test, dip=0.0275, p<0.001). The mutual information between cuneate responses and primary inputs increased with the number of silent synapses (i.e. synapses evoking no unitary EPSPs). Under the bimodal distribution regime, near-silent synapses also existed that evoked small post-synaptic potentials with 0% unitary EPSP-to-spike coupling. (b) Effective primary afferent synapses evoked unitary EPSPs with maximum amplitude of 4 mV and with all-or-nothing behavior. (c) Relationship between the total number of primary afferent synapses and the number of effective synapses required to achieve maximal information transfer. (d) Relationship between the maximal EPSP amplitude, the EPSP time constant and the optimal number of effective synapses to obtain maximal information transfer. White lines indicate sample values for a cuneate neuron recorded in the present study.
We have so far shown that the mechanoreceptors are organized so that, not only Shannon topological discriminability is guaranteed, but also geometric discriminability. We have designed a model of cuneate nucleus that is able to transmit this organization in a very short time. We now want to understand in more details how signals are organized and how they can be decoded/reencoded. In order to do this, we define a new distance on the set of spike trains that mimics the behavior of a well-known neuron model, the Spike-Response Model. This distance takes spike trains as inputs and computes the distance between the time-course of the probabilities of firing of this neuron, i.e. it evaluates the reencoding of the signals. We then apply this new distance to the mechanoreceptors signals and exhibit, for some values of the parameters of the distance, some logical organization of the signals. In particular, we show that the outputs of the cuneate nucleus can be, when we consider the whole spike trains, isometric to the stimulations. We conclude by arguing about the potential role of this isometry property.

The question arises: Can we find some logical organization in the precise spike times? Can we link some parameter in the spike times to some parameter in the stimulus space? More than this, how do neurons tune to read this logical organization to transmit efficiently?

11.1. The SRM distance

As we have stated above, the distance we use in the information-theoretic quantities we have defined should represent the properties of a possible decoder (should be a projection of the properties of the neurons onto the topological spike train space). In this section, we attempt to define the distance that is implemented by a Spike-Response-Model neuron.

When a SRM receives an input spike train $s_i$, it first computes the membrane potential $V_i$ and then the probability of firing $p_i$. We think of this probability of firing as the fundamental
output of the SRM. It does not depend on a particular trial, but represents the way the signal is processed (on average). The advantage of the SRM is that probabilities can be computed in a straightforward way from the membrane potential time-course.

In the first place, the membrane potential time-course is computed with:

$$V(t) = \frac{1}{\tau} \sum_i H(t - \hat{t}_i) \sqrt{t} e^{-(t - \hat{t}_i)/\tau}$$

(11.1)

where $H$ is the Heaviside function.

The probability is a sigmoid of the membrane potential:

$$p(t) = \frac{1}{1 + e^{(V_0 - V(t))/\Delta V}}$$

(11.2)

We thus define the SRM distance as the difference between the probabilities for two input spike trains:

$$d(s_1, s_2) = \left( \int_t (p_1 - p_2)^z dt \right)^{1/z}$$

(11.3)

We will choose $z = 2$ in this study.

**Preliminary Discussion**

This distance may seem highly similar to the Van Rossum distance for example. Indeed, we begin by filtering the spike train with an exponential function. However, we then take into account a nonlinearity, that is highly relevant, since it represents the threshold behavior of a neuron. In the present case, it is modeled with escape noise. Interpreting the Van Rossum distance retrospectively, it would be as though the probabilities of spiking would be linearly dependent on the membrane potential. One of the main advantage of this distance is that all the spikes are not considered identically as before. It may really implement synchrony detection: indeed two spike trains whose interspike intervals are larger than 20 ms may have a null distance if $p(V < 0.8 V_{th})$ is very low. This may be very relevant in the case of spontaneous input activity that should not be taken into account. None of the distances above can do this. Indeed, the usual distances are likely to be high between two spontaneous spike trains.

It, of course, requires more parameters (membrane time constant, threshold and slope of the nonlinearity), but these parameters are neuronal! So this may help determining the parameters of the neurons that lead to the highest information or discrimination level.

The fundamental idea is that the space of spike trains is at first topological, it has no metric a priori. The metric is a projection of the reader properties onto this space. Thus the precise shape of the space is only relevant with respect to the reader. The idea is to determine the reader that has the best properties with respect to the task that is to be performed. Depending on the task one wants to perform, there will be a top-down control on the nuclei (such as the cuneate nucleus) that will change its properties and thus the way it will look at its inputs. This multi-parametric distance allows to determine how the parameters affect the shape of the input space, and thus how they should be tuned in order to perform various tasks.

Note that this distance suggests that the independents-spike and independent-neuron hy-
potheses do not hold. This distance indeed takes into account the correlations between spikes and between neurons.

This distance was implemented in response to a question that appears in a lot of articles dealing with the efficiency of spike timing codes. Here, I excerpt one from (Berry et al. 1997): “Does the brain use all of the information in (retinal) spike trains?”.

11.2. Results

In order to validate the efficiency of this new SRM distance, we have simulated stochastic input spike trains according to a Poisson process with refractory period. At each time $t$, the effective probability of spiking was:

$$p_{\text{eff}}(t) = \frac{p(t)}{1 + e^{-t + \hat{t} + \tau_{\text{ref}}}}$$

(11.4)

where $\hat{t}$ is the time of the latest spike. The refractory period was set at 5 ms each.

In a first protocol, the baseline probability was chosen $p_0 = 0.01$. We simulated two stimuli: the first increased the probability of spiking to 0.2 from 300 to 600 ms, while the second increased it to 0.2 from 400 to 700 ms. We computed 400 Victor-Purpura, Van Rossum and SRM distances between pairs of spike trains from identical stimuli, and 400 between pairs of spike trains from different stimuli see fig.11.1.

![Figure 11.1.](image)

Figure 11.1.: Comparison of the SRM, Victor-Purpura and Van Rossum distances between two Poisson processes. (top) Victor-Purpura, (middle) Van Rossum and (bottom) SRM distances computed for the first protocol. In blue are the intrastimulus distance and in red the interstimulus distances.

In a second protocol, the baseline probability was chosen $p_0 = 0.01$. We simulated two stimuli: the first increased the probability of spiking to 0.2 from 300 to 600 ms, while the
second increased it to 0.4 from 300 to 600 ms. We computed 400 Victor-Purpura, Van Rossum and SRM distances between pairs of spike trains from identical stimuli, and 400 between pairs of spike trains from different stimuli see fig.11.2.

Figure 11.2.: Comparison of the SRM, Victor-Purpura and Van Rossum distances between two Poisson processes.(top) Victor-Purpura, (middle) Van Rossum and (bottom) SRM distances computed for the first protocol. In blue are the intrastimulus distance and in red the interstimulus distances.

In the two protocols, two observations should be made:

- First, the SRM distance provides a much better discrimination than the Victor-Purpura or Van Rossum distances, i.e. the overlap between intra- and interstimulus distance is much smaller.

- Second, the intrastimulus distances are much smaller relative to the interstimulus ones, suggesting a very high signal-to-noise ratio. This may be a very helpful feature to keep the metrical conditional entropy low.

### 11.3. Idea

Let us consider mechanoreceptors encoding a direction parameter $\theta$ through a first spike latency that is a sinusoidal function of this parameter. Let us, in addition, assume a uniform distribution of the privileged directions. Let us first consider a single mechanoreceptor with privileged direction $\theta_1 = 0$. Then the Victor-Purpura distance between the 4 orthogonal directions is

\[
\begin{pmatrix}
0 & 1 & 2 & 1 \\
1 & 0 & 1 & 0 \\
2 & 1 & 0 & 1 \\
1 & 0 & 1 & 0
\end{pmatrix}
\]

Two observations should be made: in the first place, two different
stimulations evoke the same response, the mutual information is thus non-maximum. Secondly, the representation space is not at all isometric to the stimulus space. However, if we add a second mechanoreceptor whose privileged direction is orthogonal $\theta_2 = \frac{\pi}{2}$, the population VP distance is now 

$$
\begin{pmatrix}
0 & 2 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 2 & 0 & 2 \\
2 & 2 & 2 & 0 \\
\end{pmatrix}.
$$

The mutual information is now maximum and the representation space is now a tetrahedron. Note that, in order to mirror the stimulus space, one would expect a square instead of a tetrahedron.

\[\begin{array}{ccc}
C=0 & 0<C<\infty & C=\infty \\
\begin{pmatrix}
t_1 & 0 & 2 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 2 & 0 & 2 \\
\end{pmatrix} & \begin{pmatrix}
t_2 & 0 & 2 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 2 & 0 & 2 \\
\end{pmatrix} & \begin{pmatrix}
t_3 & 0 & 2 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 2 & 0 & 2 \\
\end{pmatrix} \\
\begin{pmatrix}
t_3 & 0 & 2 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 2 & 0 & 2 \\
\end{pmatrix} & \begin{pmatrix}
t_3 & 0 & 2 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 2 & 0 & 2 \\
\end{pmatrix} & \begin{pmatrix}
t_3 & 0 & 2 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 0 & 2 & 2 \\
2 & 2 & 0 & 2 \\
\end{pmatrix} \\
\end{array}\]

Figure 11.3.: Distances between 4 stimuli as spikes are flowing in. The best VP cost with respect to the fidelity of the metrical organization varies with time.

### 11.4. Multidimensional latency tuning curve

In order to corroborate this idea, we plotted the first spike latency with respect to the direction for the stimulation for all the forces, angles and curvatures on Fig. 11.4. For many mechanoreceptors, there is a nice organization of the latencies. In particular:

- the latencies with respect to the direction are always sinusoidal,
- the privileged direction is the same, whatever other the parameters of the stimulation,
- the latencies increase when the force decreases,
- the amplitude of the sinusoid increases with the angle to the normal direction.

These observations show the high relevance of latency as a coding scheme. But in addition, it shows that there is a deep organization of the spiking activity of the mechanoreceptors. Indeed, we can describe the trends in a few words only.
11. **Organization of the Signals**

![Figure 11.4: Latencies of a single mechanoreceptor to many stimulations. Solid, dash and dash-dot lines stand respectively for forces of 4, 2, 1N. Blue, red and green lines correspond to angles of 20, 10 and 5 degrees. Dot, circle and star markers correspond to the three different curvatures of the probe. On the abscissa are the directions of the force.](image)

11.5. **There is more to spiking signals than discrimination**

In order to understand more precisely how the stimuli are encoded, we shuffled the responses from the mechanoreceptors. We took all the responses to the 81 stimuli of a single mechanoreceptor and shuffled them. The resulting 81 artificial stimuli were composed of the responses of the 42 mechanoreceptors to any natural stimulus.

We performed the same analysis on this data. Surprising at first, the discrimination was much faster: it took about 60 ms to reach an optimal discrimination. The reason for this is that, for weak stimulations, it took longer for all mechanoreceptors to fire. These stimuli therefore took longer to discriminate. However, after shuffling, each artificial stimulus contained the response of a mechanoreceptor to a strong stimulation.

Despite this advantage in discrimination speed, a major drawback appeared. The distance matrix was very simple: a diagonal with low values and uniform off-diagonal values. This simply means that there was no logic at all anymore in the responses of the population of mechanoreceptors. The mapping from the stimulation to the output is random. This may show that something is lost in that case.

11.6. **Organization with the SRM distance: a potential role for subsequent spikes**

We can even go further by applying the SRM distance to the responses of populations of mechanoreceptors to these stimuli. If we keep biologically plausible values, we cannot obtain
very well-organized clusters of responses if we focus on the first spike only. This is, on the contrary, possible if we consider the whole spike trains (Fig. 11.6). It shows that stimulations in the normal direction with 1, 2 and 4 form a line that passes orthogonally right in the middle of a square formed by the responses to 4 orthogonal stimulations.

On Fig. 11.7, 12 stimulations were used: 5, 10 and 20 degrees in 4 orthogonal directions. On the left part, we see that the responses are obviously distributed along 4 axes. On the right part, the pyramidal shape of the set of responses is shown. These results are currently under more thorough examination.

We want to stress that it is possible to find such an organization only by taking into account the whole spike trains. This suggests the following view on spike trains: first spikes provide a non-ambiguous reconstruction of the stimulation, subsequent spikes shape the space of spike trains so as to allow high-order processes to take place, such as generalization. Thus, if the goal is to discriminate quickly, first spikes are enough. But if a more thorough representation is needed, one should wait for subsequent spikes.

11.7. Discussion

This organization of spiking signals depends on the characteristics of the observer, i.e. the neural network that receives these signals as inputs. We suggest that, if watched by a well-
11. Organization of the Signals

Figure 11.6.: Multi-dimensional scaled representation of responses to 7 stimulations.

Figure 11.7.: Multi-dimensional scaled representation of responses to 12 stimulations: 5, 10 and 20 degrees in 4 orthogonal directions. (left) from the top, (right) from the front.

tuned system, spiking signals may be highly organized, much more than just guaranteeing discrimination. We here show that by using neurons with the right properties, it is possible to have signals that mirror the stimulus space. This means that the relative distances between the spiking signals are similar to those between the stimuli.

This property may be of considerable help for the central nervous system. Indeed, if the spiking signals are organized similarly to the stimulus space, then, it is easy for the central nervous to determine the identity of a stimulation it has never experienced before by the location of the response in the spiking space.

In order to give an intuitive idea of this, let us take a macroscopic example: while you can make a picture of a horse and a donkey, it is not obvious you exactly know what is a mule. Because the word mule does not represent the fact that the animal it describes lies at the middle between a horse and a donkey, see Fig. 11.8.

However, if you know what is a zebra and a horse, you probably can guess what a zorse is. The property that allows you to do so is that the words zebra, zorse and horse are organized as the animals they describe. Indeed, the word zorse lies between the words horse and zebra, see Fig. 11.9 (though it is not exactly in the middle).
11.7. Discussion

Donkey  Mule  Horse

Figure 11.8.: A donkey, a mule and a horse surprisingly appearing in this manuscript. While the mule is the offspring of a donkey and a horse, its name does not represent this kinship. It is virtually impossible for someone who knows what a horse and a donkey is, to guess what a mule is if she has never seen one.

Zebra  Zorse  Horse

Figure 11.9.: A zebra, a zorse and a horse less surprisingly appearing in this manuscript. The zorse if the offspring of a zebra and a horse. Its name fully respect this kinship. It is thus easy for someone to imagine what a zorse is, even if never seen before.

Here, we make an analogy between the animals and the stimulus on the one hand, and between the words and the spiking signals on the other. We argue that, if spiking signals are organized logically with respect to the stimulation, then the central nervous system can perform the same kind of recognition as that described above.

If this is the case, then the job of the brain is much easier than Shannon information theory suggest. By watching its inputs the appropriate way, it may be able to decode the stimulation (even if never encountered in the past) quickly and faithfully.

This brings us to the main point of this manuscript. In essence, the nervous system can be seen as a map from sensory signals to motor outputs. This map depends on the state of the nervous system. This map transforms a set of relations into another set of relations. The actual code does not matter. What matters is the organization (the relations). We do not care about transformations of the spike trains that let the relations invariant (this is our gauge group). Take the set of stimulations in 4 directions. They are mapped onto 4 populations of spike trains $s_0 \rightarrow r_0$, $s_{90} \rightarrow r_{90}$, $s_{180} \rightarrow r_{180}$, $s_{270} \rightarrow r_{270}$. We can now use find a (biologically
plausible) distance that sends these responses at the 4 vertices of a square (or equivalently at 4 points on a circle). We can now permute them so that $s_0 \rightarrow r_{90}$, $s_{90} \rightarrow r_{180}$, $s_{180} \rightarrow r_{270}$, $s_{270} \rightarrow r_{0}$, the distances between them are still the same. The spike trains still do a good job of representing the stimulus space. In this sense, whether $s_0$ is represented by $r_0$ or $r_{90}$ does not matter as long as the other responses are consistent with it. In this sense, the precise neural code does not matter. The only one that matters is the relative one.
In this manuscript, we aimed at understanding in more details how tactile receptors in the dermis of the fingertip encode information with spikes about the external stimulations and how the cuneate nucleus, the first relay to the central nervous system, may decode them. We have used a set of spike trains elicited by 42 mechanoreceptors, recorded via microneurography experiments, when stimulated with 81 different stimulations with various parameters such as the magnitude of the force, the direction, the angle and the curvature of the probe used. We have started using two hypotheses, that are well-supported in neural coding: spike timing is important and information is encoded by populations of cells.

We have put forward the fact that Shannon information theory, in spite of its great successes in neuroscience, may fail to account for some highly relevant coding principles. The main reason being that it is purely topological and thus does not take into account the geometry of the space of spike trains. We have therefore proposed a modified version of Shannon entropy that may embed the distances between spiking responses. It is not a completely new entropy with respect to Shannon. Indeed, our entropy reduces to Shannon in some well-defined cases. Whereas Shannon assesses the information transfer of a system irrespective of the properties of the reader, our information takes them into account under the form of a distance-based similarity measure between spiking inputs. Instead of focusing only on information as in Shannon’s theory usual applications, we also put forth the idea that in biological systems, special attention should be taken of the conditional entropy. Indeed, for an individual to provide a relevant systematic reaction to a given stimulus, the number of responses elicited by the nervous system should be kept low or at least highly organized.

We then showed that this new information measure may account for a psychoacoustics phenomenon: the perceptual magnet. It states that the central element of a category pulls other elements towards it and thus make them appear more similar than they really are. While Shannon information cannot explain this, our new information measure, conjugated with the necessity to maintain conditional entropy low straightforwardly account for this phenomenon.

As a starting point to analyze the microneurography data, we have used a common distance, the Victor-Purpura distance that measures, for some predefined time constant, the extent to which spike trains are different. We have shown that, in order to reach a fast complete discrimination (with maximal information and zero conditional entropy), the best time constant is 7 ms. With this parameter, the complete discrimination occurs after only 70 ms or equivalently 35 ms after the first spike arrival time.

We then devised a model of the Cuneate Nucleus to estimate how fast and reliably it can transmit this information. It was modeled as a feedforward layer of Spike-Response Model. The properties of individual neurons were fitted according to experimental data performed
by our collaborators, Fredrik Bengtsson and Henrik Jörntell from Lund, Sweden. The connectivity properties were adapted from observations in the cat Cuneate Nucleus, suggesting a divergence/convergence of 1700/300, i.e. a probability of connection of 0.18. The synaptic weights were initially drawn randomly from a uniform distribution. They are then modulated according to a plasticity rule specifically designed for the SRM that maximized Shannon Mutual Information between the inputs to the neuron and its outputs. We then fed a layer of 50 neurons with the data from the microneurography experiments. After a phase of training, we fixed the weights and measured the metrical information transfer between the stimulations and the outputs of the CN. We showed that by using the Victor-Purpura distance, the perfect discrimination can be reached only some 5ms after it was reached at the arrival at the CN. This means that the organization of mechanoreceptors signals can be transmitted quickly and reliably by a feedforward layer of neurons.

Another interesting result obtained with this simulation is the surprising distribution of synaptic weights. Indeed, of the initial 42 synapses, only 3 to 6 remain non-silent after learning. We showed that this final number of non-silent synapses hold when there are inputs from more mechanoreceptors (around 130). The same numbers were observed by our collaborators in Lund. Of the 300 potential effective synapses, only 5 to 10 are actually non-silent. This suggests that the cuneate nucleus may be tuned so as to maximize information transmission.

In a final part, as we wanted to formalize the properties of potential decoders, we have defined a new distance based on the properties of the Spike-Response Model. This distance may be more efficient in terms of discrimination for inputs with irrelevant (non-informative) spontaneous activity. We have applied this distance to the microneurography data and showed that, with well-tuned parameters, the space of responses is isometric to that of the stimuli. This property may be of high interest for the central nervous system. Indeed, it is very easy then to generalize to unknown stimuli, since the latter is located with respect to other stimuli exactly as the response it elicits is located to other responses. In terms of neural coding, what appears meaningful is that, while first spikes are sufficient for geometric discriminability, this high organization requires to take into account the following spikes. This may appear as a potential role for subsequent spikes.

Let me draw the picture of the nervous system functioning that arises during these three years. The physical world has some structure. By interacting with this world, we probe this structure. There is thus a first map from the physical world to the spike trains by peripheral sensory organs. These spike trains are then processed by more central neurons until it reaches outputs, such as motor neurons. Our point is that the first map must preserve the structure of the world, so that central neurons have access to the structure of the physical world. Then central neurons, according to their inner state and to the goal to perform, transform this map into a more relevant one: they expand zones of interest, compress others. They behave as short-code machines (Von Neumann 1958). The perceptual space is warped by processing in higher-order structures, but the faithful image of the world is present at the periphery.

In line with this, the final conclusion of this manuscript may be a negative one: we have not succeeded in presenting a gauge theory of the neural code. However, to some extent, we have laid down some basic ideas for it. In particular, the idea that we have come up with is that the neural code is invariant under transformation that keep the relative geometry of spike trains unchanged. This is our gauge group.
La structure du système nerveux est aujourd’hui relativement bien connue : de nombreux modèles existent, aussi bien au niveau du neurone, que des assemblées de neurones. Différents types de neurones ont été analysés et leurs propriétés déterminées. De nombreuses structures et sous-structures du système nerveux central ont été également identifiées. Nonobstant cette compréhension détaillée de la structure du système nerveux, ses modes de fonctionnement restent en revanche largement non élucidés. La phénoménologie est riche, mais les principes fondamentaux se font rares.

En particulier, un problème qui se trouve être le thème majeur de cette thèse est celui du *code neuronal*. On peut définir ce dernier de façon simple comme le langage des neurones. Comme il a été mentionné au paragraphe précédent, si les connexions entre les neurones sont assez bien connues, le langage qu’ils utilisent restent non-déchiffré. On sait que les neurones communiquent entre eux par des successions d’impulsions brèves qu’on appelle *potentiel d’action* ou *spike* pour faire court. Ce sont les seuls signaux qui se propagent le long des axones des neurones et ce sont donc les seuls signaux auxquels les autres neurones ont accès. Ces impulsions sont de plus stéréotypées : les spikes émis par un neurone donné sont tous identiques. Il est également généralement supposé que tous les spikes émis par une population de neurones sont identiques. La conclusion directe de ces observations est que les seules caractéristiques variables d’une séquence de spikes sont le nombre d’impulsion et leurs instants d’émission. Ainsi, un neurone émet un “train de spikes” qu’on peut représenter par une succession de nombres réels correspondant aux temps d’émission des spikes : par exemple, \{23, 67, 98\} ms pour un train de 3 spikes émis aux temps 23, 67 et 98 millisecondes.

Le problème du code neuronal est en fait de déterminer quelles sont les caractéristiques des trains de spikes qui sont intéressantes pour transmettre de l’information. Si, à deux stimulations différents, un neurone émet les trains suivants \( A = \{23, 67, 98\} \) et \( B = \{23, 37, 48, 67, 72, 98\} \), alors le nombre de spikes émis est un bon indicateur de l’identité de la stimulation. En revanche, le minutage exact des spikes n’est que peu pertinent. De même pour le temps auquel est émis le premier spike. Si, au contraire, un neurone stimulé de deux façons différentes, émet les trains \( A = \{23, 67, 98\} \) et \( C = \{37, 48, 72\} \), alors le nombre de spikes émis n’a que peu d’importance. C’est le minutage exact des impulsions qui va permettre d’identifier la stimulation. La problématique de cette thèse est d’étudier quels sont les caractéristiques pertinentes des trains de spikes.

Pendant la majeure partie du XXe siècle, le paradigme en code neuronal a été que les neurones codent la stimulation en émettant des spikes à une fréquence qui croît avec l’amplitude de la stimulation, en accord avec les travaux pionniers d’Adrian (Adrian 1926). Selon cette vision, pour une stimulation donnée, le neurone émet des spikes à une certaine fréquence,
mais les instants précis d’émissions restent indéterminés. Néanmoins, au début des années 90, il est apparu que le minuitage des spikes étaient beaucoup plus précis que suggéré par cette vision. Pour une stimulation donnée, non seulement la fréquence d’émission est déterminée, mais il a également été observé que certains spikes réapparaissaient exactement au même instant lorsque cette stimulation était répétée (Mainen and Sejnowski 1995; De Ruyter van Steveninck et al. 1997). Ceci suggérant qu’il est possible de déterminer l’identité de la stimulation en étudiant le minuitage des impulsions. Ces observations ont une grande influence sur la suite des travaux sur le code neuronal. En effet, la capacité à transmettre de l’information rapidement est beaucoup plus grande si l’on utilise le minuitage des spikes plutôt que leur fréquence. Si un individu doit déterminer en un temps très court l’identité d’une stimulation.


Une approche classique pour estimer de façon quantitative et rigoureuse la possibilité de déterminer l’identité de la stimulation grâce à l’observation des trains de spikes en réponse à celle-ci est d’utiliser la théorie de l’information de Shannon. Cette théorie permet de déterminer la réduction de l’incertitude conséquente à l’observation des réponses des neurones. Par exemple, si Alice pense à un nombre qui est soit 0 soit 1. Si elle ne dit rien à Benoît, ce dernier ne sait pas à quel nombre elle pense : il y a une probabilité 1/2 pour qu’elle pense à l’un de ces deux nombres. Si, maintenant, Alice dit à Benoît que le nombre auquel elle pense est impair, alors Benoît sait qu’Alice pense 1 avec une probabilité égale à 1, c’est-à-dire qu’il est certain qu’Alice pensait 1. Le message d’Alice a donc réduit l’incertitude de Benoît sur le nombre auquel elle pensait. Le but de la théorie de l’information de Shannon est de formaliser mathématiquement ces idées.

Cette théorie peut aisément être appliquée au problème du codage neuronal. Si un neurone peut être stimulé de 2 façons différentes et qu’il émet les trains $A = \{23, 67, 98\}$ et $C = \{37, 48, 72\}$ en réponse à ces deux stimulations. Tant que je ne connais pas la réponse du neurone, il y a une probabilité 1/2 pour qu’une des deux stimulations ait été présentée. Si, maintenant, j’ai accès au nombre de spikes émis par le neurone, i.e. 3, je ne peux pas déterminer quelles stimulation a été présentée, car les deux stimulations évoquent 3 spikes. L’information transmise est donc nulle. En revanche, si j’ai accès au minutage exact des spikes et que je lis $\{23, 67, 98\}$, je sais quelle stimulation a été appliquée. L’information transmise est donc strictement positive. On voit dans cet exemple que le minuitage exact des spikes peut être hautement informatif par rapport au simple nombre de spikes.
Nous avons jusqu’à présent omis un élément crucial : le bruit. En effet, lorsque l’on applique plusieurs fois la même stimulation, on a en général des réponses différentes. Par exemple, dans le cas des mécanorécepteurs, lorsqu’on applique une stimulation donnée, on observe une certaine variation du minutage des spikes. Le premier spike est émis en moyenne après 42 ms, mais peut apparaître après 41 ou 43 ms également. Ce fait peut potentiellement diminuer la possibilité de déterminer l’identité de la stimulation. En revanche, ce qui est intéressant est d’observer que, pour des stimulations dans quatre directions différentes, la variation du minutage lorsque l’on varie la direction est plus grande que lorsque l’on présente plusieurs fois la même direction. Ce qui fait qu’aucun train de spikes n’est émis en réponse à deux stimulations différentes. Il est donc possible, selon la théorie de l’information de Shannon, de déterminer quelle stimulation a été appliquée en observant les trains de spikes émis par les mécanorécepteurs.

En revanche, la théorie de l’information de Shannon ne prend pas en compte certaines caractéristiques des réponses des neurones. Reprenons notre exemple avec Alice et Benoît. Alice pense a un nombre, soit 0, soit 1. Si elle pense 1, elle donne à Benoît un nombre premier, sinon, un nombre non-premier. Du point de vue de la théorie de l’information de Shannon, il n’y a aucune ambiguïté dans les messages d’Alice. Si Benoît parvient à déterminer si le nombre est premier ou non, il peut alors savoir si Alice pense 0 ou 1 avec une probabilité égale à 1. Il peut donc réduire son incertitude. Il apparaît pourtant intuitivement que si Alice donne de très grands nombres, Benoît ne pourra pas déterminer si le nombre était premier ou pas, il ne pourra donc pas réduire l’incertitude sur le nombre auquel Alice pense. Il n’y a donc pas d’information transmise par le message d’Alice. La raison est que le message d’Alice a une structure qui n’est que peu pertinente pour Benoît. Celui-ci est inapte à décoder le message. Si Alice lui fournissait un nombre pair ou impair suivant qu’elle pense 0 ou 1, Benoît pourrait décoder le message en écoutant seulement le dernier chiffre. Mais la théorie de l’information de Shannon ne prend pas ceci en compte. En un sens, elle considère que la personne qui va lire le message est un lecteur idéal capable de faire des opérations complexes en un temps très court.

Or, un neurone (ou une population de neurones) n’est pas un décodeur idéal. Il a des contraintes qui sont liées à sa structure et à son mode de fonctionnement. Un des travaux principaux de cette thèse a été de modifier la théorie de l’information de Shannon de façon à ce que les résultats sur la réduction de l’incertitude dépende de l’organisation des messages fournis. Ceci a été fait en introduisant des notions de similarité/distance entre les messages, c’est-à-dire, dans notre cas d’étude, entre les trains de spikes des mécanorécepteurs. Il existe de nombreuses notions de distance entre les trains de spikes. L’une des plus connues est celle définie par Victor et Purpura (Victor 2005). Prenons deux trains de spikes $A = \{23, 32, 67, 98\}$ et $B = \{21, 70, 95\}$. Quelle est la distance entre eux ? Victor et Purpura la définissent comme le coût minimal pour transformer l’un en l’autre. Afin de calculer cette distance, nous définissons un paramètre $c$ qui s’exprime comme l’inverse d’un temps. Partons de $A$ :

- déplaçons le spike de 23 à 21 pour un coût de $(23 - 21)c$,
- déplaçons le spike de 67 à 70 pour un coût de $(70 - 67)c$,
- déplaçons le spike de 98 à 95 pour un coût de $(98 - 95)c$,
- il reste un spike supplémentaire dans $A$, éliminons-le pour un coût de 1.
La distance entre $A$ et $B$ est donc: $d(A, B) = 1 + 10c$. Remarquons ici que, si $c = 0$, alors cette distance est la différence entre les nombres de spikes. Si $c > 1$, alors cette distance est le nombre total de spikes dans les deux réponses diminué de deux fois les coïncidences exactes. Nous proposons dans cette thèse une version de l’information qui prend en compte ces distances entre les trains de spikes, contrairement à l’information de Shannon. Pour Shannon, deux réponses qui diffèrent ne serait-ce que très peu sont considérés comme différentes, de la même façon que deux réponses qui diffèrent beaucoup. Dans notre version, nous considérons les relations entre réponses de façon plus nuancée. Ainsi, tandis que l’information de Shannon est grande, que la transformation des stimulations en trains de spikes soit régulière ou anarchique, notre information est beaucoup plus grande dans le cas d’une transformation régulière. Bien sûr, la notion de régularité dépend de la distance que nous utilisons. Finalement, notre mesure d’information n’est égale à celle de Shannon qu’en cas de régularité très forte des réponses, c’est-à-dire lorsque les réponses à une stimulation sont toutes plus proches les unes des autres que des réponses à d’autres stimulations. De façon imagée, lorsque les réponses à différentes stimulations forment des archipels distincts les uns des autres.

Nous avons montré que si l’on applique cette nouvelle information aux données du système somatosensoriel, il est possible d’obtenir cette classification en archipel en un temps très court, suffisamment court pour être compatible avec des expériences ou on mesure les temps de réaction d’individus à des stimulations tactiles. De façon intéressante, une valeur du paramètre $c = 0.15ms^{-1}$ apparaît comme étant optimale pour obtenir une classification rapide. Ceci suggère une échelle temporelle optimale pour décoder les signaux qui est de l’ordre de 7ms. Un de nos collaborateurs, Henrik Jörntell, de l’université de Lund, en Suède a effectué des enregistrements de neurones du noyau cunéiforme, qui est la destination des signaux des mécanorécepteurs. Il a observé que les constantes de temps des neurones dans ce noyau sont de l’ordre de 7ms également. Ceci suggère que ces neurones sont calibrés de façon optimale pour décoder les signaux qu’ils reçoivent.

Nous avons défini notre mesure de l’information dans le souci de tenir compte des propriétés de l’observateur. Or la distance de Victor et Purpura n’a pas d’interprétation directe en termes neuronaux. Nous avons donc défini une distance inspirée d’un modèle de neurone reconnu. L’idée est de reproduire la façon dont un neurone traite les signaux qu’il reçoit. Ceci nous a fourni une distance qui dépend de plusieurs paramètres qui ont une interprétation neuronale simple.


Alice pense à une lettre et transmet à Benoît un nombre en fonction de cette lettre. Si avant cela, Alice explique à Benoît qu’elle attribue 13 à A, 22 à B et 2 à Z, alors il est difficile pour Benoît de savoir à quelle lettre Alice pense lorsqu’elle lui dit, par exemple, 11. En revanche,
si Alice attribue 1 à A, 2 à B et 26 à Z, il devient évident pour Benoît que 13 correspond à M et de même pour les 26 lettres. Ainsi, si l’organisation des messages est en miroir de celle des stimulations, il devient facile de généraliser à des stimulations dont on ne connaît pas les messages.

Cette notion de ressemblance entre l’organisation des entrées et des sorties n’est pas ancrée dans notre langage quotidien. Pourtant, les mots inventés récemment vont dans ce sens. Ainsi, si vous devez deviner ce qu’est un *ligre* sachant que c’est un animal, vous aurez peu de mal à savoir qu’il s’agit du croisement d’un lion et d’une tigresse. De même, un *zorse* est le croisement d’un zèbre et d’une jument (*horse* en anglais).

Ainsi, si on définit un observateur pour qui les réponses apparaissent organisées de la même façon que le sont les stimulations, cet observateur déterminera aisément l’identité d’une stimulation - même s’il n’avait jamais eu à déterminer avant- juste en observant la réponse.
La structure du système nerveux est aujourd'hui relativement bien connue (Kandel, Schwartz, and Jessell 1991) et de nombreux modèles existent, aussi bien au niveau du neurone (Gerstner and Kistler 2002; Izhikevich 2007), que des assemblées de neurones. Différents types de neurones ont été analysés et leurs propriétés déterminées. De nombreuses structures et sous-structures du système nerveux central ont été également identifiées. Nonobstant cette compréhension détaillée de la structure du système nerveux, ses modes de fonctionnement restent en revanche largement non élucidés. La phénoménologie est riche, mais les principes fondamentaux se font rares.

En particulier, un problème qui se trouve être le thème majeur de cette thèse est celui du **code neuronal**. On peut définir ce dernier de façon simple comme le **langage** des neurones. Comme il a été mentionné au paragraphe précédent, si les connections entre les neurones sont assez bien connues, le langage qu’ils utilisent restent non-déchiffré. On sait que les neurones communiquent entre eux par des successions d’impulsions brèves qu’on appelle **potentiels d’action** (PA) ou **spikes** pour faire court. Ce sont les seuls signaux qui se propagent le long des axones des neurones et ce sont donc les seuls signaux auxquels les autres neurones ont accès. Ces impulsions sont de plus stéréotypés : les spikes émis par un neurone donné sont tous identiques. On fait de plus la supposition que tous les spikes émis par une population de neurones sont identiques. La conclusion directe de ces observations est que les seules caractéristiques variables d’une séquence de spikes sont le nombre d’impulsion et leurs instants d’émission. Ainsi, un neurone émet un **train de spikes** qu’on peut représenter par une succession de nombres réels correspondants aux temps d’émission des spikes : par exemple, \{23, 67, 98ms\} pour un train de 3 spikes émis aux temps 23, 67 et 98 millisecondes.

Le problème du code neuronal est en fait de déterminer quelles sont les caractéristiques des trains de spikes qui sont intéressantes pour transmettre de l’information. Si, à deux stimulations différentes, un neurone émet les trains suivants $A = \{23, 67, 98\}$ et $B = \{23, 37, 48, 67, 72, 98\}$, alors le nombre de spikes émis est un bon indicateur de l’identité de la stimulation. En revanche, le minutage exact des spikes n’est que peu pertinent. De même pour le temps auquel est émis le premier spike. Si, au contraire, un neurone stimulé de deux façons différentes, émet les trains $A = \{23, 67, 98\}$ et $C = \{37, 48, 72\}$, alors le nombre de spikes émis n’a que peu d’importance. C’est le minutage exact des impulsions qui va permettre d’identifier la stimulation. La problématique de cette thèse est d’étudier quelles sont les caractéristiques pertinentes...
14. Résumé scientifique

Figure 14.1.: Exemple de distance de Victor-Purpura. Etapes successives pour transformer la séquence de potentiels d’action (SPA) du haut en celle du bas.

des trains de spikes.

Notre modèle d’étude est le système somatosensoriel ascendant.

14.1. Géométrie des trains de spikes

Récemment, de nombreuses distances ont été définies sur les trains de spikes. Ces distances permettent notamment de mieux comprendre la géométrie de l’espace des trains de spikes. En effet, ceci permet d’estimer dans quelle mesure deux trains de spikes sont similaires ou dissimilaires.

Celle que nous utilisons dans ce manuscript est la distance de Victor-Purpura, voir figure 14.1. Cette distance entre deux SPA $A$ et $B$ se définit ainsi : c’est le coût minimal pour transformer l’une en l’autre. Pour ce faire, trois opérations sont disponibles :

- créer un PA pour un coût de 1.
- éliminer un PA pour un coût de 1.
- déplacer un PA dans le temps pour un coût de $c$ par unité de temps (par exemple, une milliseconde).

D’autres distances existent, telles celle de Van Rossum (Van Rossum 2001), voir figure 14.2. Cette dernière consiste à convoluer chaque SPA avec un filtre exponentiel $\exp(-t/\tau)$ de façon à obtenir deux fonction $f_A$ et $f_B$ et de prendre ensuite la norme-2 sur ces fonctions, c’est-à-dire $d_{VR}(A,B) = \int (f_A - f_B)^2 dt$. Notons que le paramètre $\tau$ de la distance de Van Rossum s’apparente à l’inverse du paramètre $c$ de la distance de Victor-Purpura.

14.2. Théorie de l’information de Shannon

Une des techniques les plus utilisées ces dernières années pour étudier la capacité de différentes stratégies de codage à transmettre de l’information est la théorie de l’information de Shannon (Shannon 1948). Lorsque l’on a un ensemble $R = \{r\}$ d’événements possibles, formant une partition de l’univers, ayant chacun une probabilité $a priori$ d’apparaître $p(r)$, on peut
14.2. Théorie de l’information de Shannon

Pour définir une *entropie* qui quantifie la variabilité ou l’incertitude sur l’événement qui a lieu. Cette entropie s’écrit:

\[ H(R) = - \sum_r p(r) \log p(r) \quad (14.1) \]

Si un des événements \( r \) est très probable, i.e. \( p(r) \sim 1 \), et tous les autres \( r' \) très improbables, i.e. \( p(r') \sim 0 \), alors l’entropie est très faible. En revanche, s’il existe plusieurs événements dont la probabilité est moyenne, alors l’entropie est élevée.

De la même façon, on définit une entropie conditionnelle ou equivocation qui mesure l’incertitude d’un événement étant donné un autre.

\[ H(R|S) = - \sum_s p(s) \sum_r p(r|s) \log p(r|s) \quad (14.2) \]

Cette entropie fait intervenir les probabilités mises à jour ou *a posteriori*.

On définit alors l’information mutuelle entre \( R \) et \( S \) comme la différence entre l’entropie marginale \( H(R) \) et l’entropie conditionnelle \( H(R|S) \):

\[ I(R; S) = \sum_{r,s} p(r,s) \log \frac{p(r,s)}{p(r)p(s)} \quad (14.3) \]

Une caractéristique fondamentale de la théorie de l’information de Shannon est qu’elle est purement topologique, c’est-à-dire qu’elle ne dépend pas de la géométrie de l’espace des réponses possibles. En effet, dans le cas de la transformation de stimuli \( s \) en réponses \( r \), une condition suffisante pour que l’information soit maximale est qu’aucune réponse \( r \) ne soit évoquée par deux éléments \( s \) et \( s' \). Quelque soit l’organisation géométrique des réponses, si cette condition est réalisée, l’information est maximale.
14. Résumé scientifique

14.3. Théorie de l’information métrique

Dans cette partie, nous présentons une extension de la théorie de l’information de Shannon qui prend explicitement en compte la géométrie de l’espace dans lequel se situent les événements. La façon dont elle est prise est générale et ne fait pas d’hypothèses quant à la topologie ou à la dimensionnalité de l’espace.

La géométrie de l’espace des réponses est considérée via une mesure de similarité $\phi(r, r')$ entre les événements $r$ et $r'$ :

$$H^*(R) = - \sum_r p(r) \log \sum_{r'} p(r') \phi(r, r') \quad (14.4)$$

Cette mesure de similarité peut être une fonction quelconque à valeurs entre 0 et 1. La différence avec l’entropie de Shannon réside donc dans l’argument du logarithme : à la place de $p(r)$, on a $\sum_{r'} p(r') \phi(r, r')$. On peut noter quelques propriétés intéressantes :

- Cette entropie est toujours plus faible que celle de Shannon puisque le terme dans le logarithme est toujours plus grand.
- Si on prend un delta de Kronecker pour la mesure de similarité, cette entropie se réduit à celle de Shannon.
- Plus la mesure de similarité est grande, plus l’entropie est faible.
- Dans le cas extrême, si la mesure de similarité vaut 1 pour tout couple de réponses, alors l’entropie est nulle.

L’entropie conditionnée à un stimulus $s$ peut être définie comme :

$$H^*(R|s) = - \sum_r p(r) \log \sum_{r'} p(r'|s) \phi(r, r') \quad (14.5)$$

L’entropie conditionnelle est donc :

$$H^*(R|S) = - \sum_r p(r, s) \log \sum_{r'} p(r'|s) \phi(r, r') \quad (14.6)$$

L’information est alors définie comme la différence entre l’entropie marginale et l’entropie conditionnelle :

$$I^*(R; S) = H^*(R) − H^*(R|S) \quad (14.7)$$

$$= \sum_r p(r, s) \log \frac{\sum_{r'} p(r'|s) \phi(r, r')}{\sum_{r'} p(r') \phi(r, r')} \quad (14.8)$$

Ceci pose les fondations de notre extension de la théorie de l’information de Shannon. De façon à prendre en compte la géométrie de l’espace, nous définissons la mesure de similarité entre deux événements comme une fonction décroissante de la distance entre ces deux événements, en suivant l’idée que la distance fournit une estimation naturelle de la dissimilarité. On appelle alors l’entropie définie précédemment *entropie métrique*.
Dans la suite, nous prenons une mesure de similarité qui est une fonction de Heaviside de la distance :

\[ \phi(r, r') = H(D_{\text{critic}} - D(r, r')) \]  

(14.9)

où la distance critique \( D_{\text{critic}} \) est la valeur de la distance à laquelle la similarité chute de 1 à 0.

En théorie de l’information de Shannon, l’objectif est souvent de maximiser l’information sans tenir compte de l’entropie conditionnelle. Pourtant, l’indication donnée par l’entropie conditionnelle est très importante : elle estime la dispersion des réponses à un seul stimulus. En ce sens, plus l’entropie conditionnelle est grande et plus la tâche d’identifier différentes réponses provenant du même stimulus est ardue. Ainsi, dans le cas de la neurotransmission, il semble primordial de prendre en compte la valeur de l’entropie conditionnelle. Nous choisissons donc la valeur de la distance critique qui minimise l’entropie conditionnelle métrique et simultanément maximise l’information métrique. Pour ce faire, nous calculons les distances entre toutes les réponses à tous les stimuli. Nous les regroupons en deux ensembles : les distances intra stimuli, c’est-à-dire les distances entre réponses provenant du même stimulus, et les distances inter stimuli, c’est-à-dire les distances entre réponses provenant de stimuli différents, voir figure 14.3a Si la distance maximale intra est inférieure à la distance minimale inter, alors on peut définir la distance critique comme étant entre les deux. On a alors des mesures de similarité qui valent toutes 1 entre réponses à un même stimulus et donc une entropie conditionnelle métrique qui est nulle. On a également des mesures de similarité qui valent 0 entre réponses provenant de deux stimuli différents et donc une information métrique maximale.

Jusqu’à présent, on a considéré une seule distance critique pour l’ensemble des paires de réponses. Considérons le cas de la figure 14.3e, les ensembles de réponses à un stimulus sont tous plus petits qu’ils ne sont distants. Mais la présence d’un ensemble de grande taille loin des autres petits ensembles nous oblige à choisir un distance critique élevée qui va mener à des similarités non-nulles entre petits ensembles. Ceci a pour conséquence que l’information métrique ne peut pas être maximale en même temps que l’entropie conditionnelle nulle. Pour remédier à cette pathologie, nous proposons une version locale de cette théorie. Désormais, la mesure de similarité s’écrira :

\[ \phi(r, r') = H(D_{\text{critic}}(r) - D(r, r')) \]  

(14.10)

avec \( D_{\text{critic}}(r) \) qui dépend de la région de l’espace des réponses. Ainsi, dans le cas décrit dans le paragraphe ci-dessus, on peut choisir une distance critique petite dans la région des petits ensembles et une distance critique élevée dans la région du grand ensemble. Ceci revient à demander une discrimination plus fine dans la région des petits ensembles.

14.4. Applications théoriques de l’information métrique

14.4.1. Une interprétation de l’aimant perceptuel par la théorie de l’information métrique

Nous avons analysé le phénomène d’aimant perceptuel (ou perceptual magnet en anglais) à l’aide de l’information métrique. Rappelons dans un premier temps à quoi correspond cet
aimant perceptuel. C’est un phénomène psychoacoustique qui se traduit par une perception déformée de certains sons. Par exemple, nous avons un prototype sonore pour le son correspondant à la voyelle \(i\). Lorsque nous entendons un son qui diffère peu en termes de contenu spectral de ce son, nous le percevons encore plus proche du prototype qu’il ne l’est réellement. De façon plus générale, à amplitude égale, nous percevons mieux les différences entre les catégories qu’à l’intérieur d’une catégorie. D’où l’appellation d’aimant perceptuel : le prototype agit comme un aimant sur les sons qui sont dans son voisinage.

Nous proposons d’étudier dans quelle mesure ce phénomène peut être vu comme une conséquence d’une maximisation de l’information métrique et d’une minimisation de l’entropie conditionnelle. Pour être plus précis, nous proposons en chaque point de l’espace de réponse de maximiser une fonction que l’on définit ainsi :

\[
Q(r, S) = I^*(r; S) - \alpha H^*(r|S)
\]  

(14.11)

où \(\alpha\) est un paramètre qui règle le compromis entre maximisation de l’information et minimisation de l’entropie conditionnelle. Lorsque \(\alpha\) est petit, l’accent est mis sur la maximisation de l’information et lorsque \(\alpha\) est grand, c’est la minimisation de l’entropie conditionnelle qui est privilégiée.

Considérons maintenant un ensemble de catégories de stimulations sonores distribuées selon un seul paramètre réel. Chaque catégorie est une distribution gaussienne centrée sur \(\mu_i\) et de variance \(\sigma\). Dorénavant, la discrimination optimale de ces stimulations est étudiée en considérant la fonction \(Q\) en variant le paramètre \(\alpha\).

Le noyau (ou mesure de similarité) considéré ici est une fonction gaussienne de la distance :

\[
\phi(x, y) = e^{(x-y)^2/2\beta^2}
\]  

(14.12)

La déviation standard \(\beta\) régule la finesse de la discrimination : en effet, si \(\beta\) est faible, la mesure de similarité n’est non-nulle que pour les stimulations \(x\) et \(y\) très proches. En revanche, si \(\beta\) est grande, la mesure de similarité décroît lentement et donc, des stimulations distantes sont considérées similaires. Différentes déviations standards \(\beta\) seront utilisées afin de déterminer, pour chaque valeur de \(\alpha\), laquelle est la meilleure.

Considérons dans un premier temps un cas à faible bruit, c’est-à-dire dans lequel les distributions des catégories sont très peu confondues (faible superposition).

Nous avons tracé la valeur optimale de la déviation standard \(\beta\) en chaque point de l’espace, voir figure 14.4. Pour de faibles valeurs de \(\alpha\), nous observons des mesures de similarité plus larges au centre des catégories, c’est-à-dire une selectivité moindre, en accord avec l’aimant perceptuel. En effet, celui-ci peut s’interpréter comme une faible discrimination des stimuli proches du prototype de la catégorie. En revanche, pour des valeurs de \(\alpha\) élevées, cet effet disparaît et les mesures de similarité sont plus larges entre les catégories. Ceci suggère que le phénomène d’aimant perceptuel est le fruit d’une tentative de maximiser l’information par rapport à une minimisation de l’entropie conditionnelle.

Nous avons effectué des calculs analogues pour le cas à haut bruit, c’est-à-dire avec des catégories qui se superposent assez largement, voir figure 14.5. Dans ce cas, quelque soit la valeur de \(\alpha\), la mesure de similarité est plus large entre les catégories. Ceci ne correspond donc pas à l’aimant perceptuel.
Nous avons donc montré que le phénomène d’aimant perceptuel, qui n’est observé que dans les situations à faible bruit, peut être retrouvé grâce à l’information métrique. Nous suggérons également que cet effet ne devrait pas être observé à haut bruit. De plus, notre étude suggère que ce phénomène est dû à un effort de discrimination (séparer des stimuli différents) plutôt que de catégorisation (regrouper des stimuli identiques).

14.4.2. L’entropie métrique comme mesure de complexité des graphes
Nous proposons ici d’utiliser l’entropie métrique comme mesure de la complexité de graphes. Il existe de nombreuses mesures de complexité de graphes, mais, à notre connaissance, aucune ne permet de mesurer la complexité de graphes généraux.

Considérons un ensemble de points \( V \) (ou événements) qui peuvent être reliés par des liens \( E \) pondérés par des poids \( W \) entre 0 et 1. Ceci est un graphe \( G = \{V, E, W\} \).

L’entropie métrique, définie précédemment, a comme arguments des événements avec des probabilités et des mesures de similarité entre ces événements. Nous faisons ici une correspondance entre les poids \( W \) des liens du graphe et les similarités entre événements. Si deux événements ont une similarité égale à 1, alors les deux points correspondants sont reliés par un lien de poids 1. L’ensemble des graphes à 4 sommets avec des liens qui valent soit 0, soit 1 sont listés sur la figure 14.6. L’entropie métrique associée est donnée, ainsi que celle de Shannon si elle est définie.

L’entropie métrique apparaît ainsi comme une mesure naturelle de la complexité des graphes. Ceci d’autant plus que Kolmogorov et Chaitin ont montré que l’entropie de Shannon est une mesure naturelle de la complexité. Le fait que notre mesure soit dérivée de la théorie de l’information de Shannon en fait un candidat sérieux comme mesure de complexité des graphes.

14.5. Applications de l’information métrique aux données de microneurographie du système somatosensoriel ascendant

14.5.1. Analyse des signaux des mécanorécepteurs
Nous avons appliqué la méthode de théorie de l’information métrique aux données de microneurographie. Puisqu’il a été récemment démontré (Johansson and Birznieks 2004) que le minutage précis des potentiels d’action (PA) permet un plus grand transfert d’information, nous voulons prendre en compte ces caractéristiques pour estimer les similarités entre séquences de potentiels d’action (SPA). Nous utilisons ici la distance de Victor-Purpura. Afin de calculer la distance entre deux populations \( D_{pop} \), nous additionnons les distances pour chaque neurone \( d_i \). Ainsi :

\[
D_{pop}(r, r') = \sum_i d_i(r, r')
\]

Nous calculons alors les distances entre les réponses neuronales provenant de la même stimulation et les distances entre les réponses neuronales provenant de deux stimulations...
différentes à chaque étape de temps. Nous traçons ensuite les bornes inférieures et supérieures de ces deux ensembles. Ceci nous donne donc la taille maximale d’un groupe de réponses à une stimulation et la distance minimale entre deux groupes de réponses provenant de deux stimulations.

Au départ, certains groupes sont plus grands que les distances entre les groupes. Mais, après quelques dizaines de millisecondes, ce n’est plus le cas, c’est-à-dire que les groupes sont tous plus petits que les distances entre eux. Nous pouvons donc définir la distance critique comme la taille maximale d’un groupe, ce qui garantit une entropie conditionnelle nulle et une information maximale (égale à $\log(|S|))$.

### 14.5.2. Modèle de Noyau Cunéiforme

Afin d’étudier s’il était possible de transmettre l’information présente dans les SPA des mécanorécepteurs en temps court et de façon fiable, nous avons défini un modèle de noyau cunéiforme, qui est le premier relai depuis les récepteurs cutanés jusqu’au système nerveux central.

Nous l’avons modélisé par un réseau de neurones sans récurrence ni interactions latérales. En effet, ce type de connections est sujet à des délais et donc sont peu aptes à transmettre les signaux rapidement. Chaque neurone est représenté par un modèle de neurone appelé Spike Response Model (SRM) (Gerstner and Kistler 2002). Ce sont des neurones non-linéaires et stochastiques. Leur dynamique est la suivante. L’état du neurone est décrit par un potentiel de membrane $V$ exprimé en millivolts. Ce potentiel est au repos à $V_r$. Lorsqu’il reçoit un PA, la membrane est dépolarisée puis de repolarise. On modélise ce phénomène transitoire par une fonction proportionnelle à $\sqrt(t) \exp(-t/\tau)$. En considérant l’ensemble des PA incidents, on peut donc calculer le potentiel de membrane au cours du temps. En fonction du potentiel de membrane, on détermine un taux de décharge instantané $g(t) = r_0 \log(1 + \exp(V(t)-V_{\text{seuil}}/\Delta V))$ où $r_0$, $V_{\text{seuil}}$ et $\Delta V$ sont des paramètres régulant respectivement le taux de décharge spontané, le quasi-seuil et la pente de la non-linéarité du taux de décharge instantané. On peut alors déterminer la probabilité de décharger comme $p(t) = 1 - \exp(-g(t))$. Un caractère réfractaire peut également être ajouté. On définit désormais la probabilité de décharger comme $p(t) = 1 - \exp(-g(t)R(t))$ avec $R$ une fonction qui vaut 1 sauf juste après la décharge d’un PA où elle chute à 0 pendant quelques millisecondes, induisant une probabilité de décharger nulle (période réfractaire absolue) puis relaxe vers 1 diminuant la probabilité de décharger (période réfractaire relative).

Pour ce qui est de la connectivité entre les mécanorécepteurs et les cellules du noyau cunéiforme, nous nous sommes inspirés de données enregistrées chez le chat (Jones 2000). Ce dernier a observé une divergence de 300 et une convergence de 1700. Ceci signifie que chaque axone de mécanorécepteur se ramifie pour envoyer des signaux à 300 cellules du noyau cunéiforme et que chacune de ces dernières reçoit des signaux de 1700 mécanorécepteurs.

Nous ne disposons de données de microneurographie que pour 42 mécanorécepteurs et ne pouvons donc modéliser le réseau en entier. Nous avons donc réduit ce réseau en gardant la probabilité de connection identique, c’est-à-dire 0.18.

Nous avons donc modélisé un noyau cunéiforme avec 50 cellules qui recevaient des signaux de 9 à 10 mécanorécepteurs en moyenne.
14.5.3. Définition d’une nouvelle distance sur les SPA

Nous avons interprété la géométrie des SPA comme représentant les propriétés d’un observateur potentiel. Nous avons dans les sections précédentes utilisé la distance de Victor-Purpura sur les SPA. Or il n’est pas évident de déterminer si un neurone ou un réseau de neurones peut effectivement observer ses entrées avec la géométrie induite par la distance VP. En particulier, nous savons que les neurones ont un seuil membranaire en dessous duquel ils n’émettent pas de PA. Ce seuil peut n’être atteint que lorsque plusieurs PA entrants arrivent de façon synchrone. Il y a donc toute une gamme de régimes de signaux pour lesquels la distance est nulle. La distance que nous avons définie prend cette propriété en compte.

Cette distance est appelée distance SRM car elle s’inspire du fonctionnement du Spike-Response Model décrit plus haut.

Dans un premier temps, on calcule le potentiel de membrane en fonction du temps :

$$V(t) = \frac{1}{\tau} \sum_i \mathcal{H}(t - \hat{t}_i) \sqrt{t} e^{-(t - \hat{t}_i)/\tau}$$  \hspace{1cm} (14.14)

où $\mathcal{H}$ est la fonction de Heaviside.

La probabilité d’émettre un potential d’action est une sigmoïde du potentiel de membrane :

$$p(t) = \frac{1}{1 + e^{(V_0 - V(t))/\Delta V}}$$  \hspace{1cm} (14.15)

On définit alors la distance SRM comme la différence entre les probabilités de décharger au cours du temps :

$$d(s_1, s_2) = \left( \int p_1 - p_2 \right)^{1/z}$$  \hspace{1cm} (14.16)

Nous choisissons ici $z = 2$.

A première vue, cette distance peut paraître très similaire à celle de Van Rossum, qui est la différence entre les convolutions des SPA avec une fonction exponentielle. En revanche, ici, nous appliquons un filtre non-linéaire au potentiel de membrane de façon à obtenir la probabilité de décharger. A posteriori, la distance de Van Rossum apparaît comme la différence entre les potentiels de membrane ou entre des probabilités de décharger qui seraient des fonctions linéaires du potentiel de membrane. Ainsi, la distance SRM n’est pas sensible à tous les PA entrants, contrairement à celle de Van Rossum. Ceci peut se légitimer par le fait qu’un neurone peut ne s’activer que lorsqu’il reçoit des PA synchrones.

Cette nouvelle distance requiert plus de paramètres (constante de temps $\tau$, seuil $V_0$, pente de la sigmoïde $\Delta V$) mais ceux-ci ont des interprétations neuronales directes.

14.6. Résultats

De façon à valider cette nouvelle distance SRM, nous avons généré des SPA selon des processus de Poisson avec période réfractaire. A chaque étape de temps $t$, la probabilité effective d’avoir un PA était :
14. Résumé scientifique

\[ p_{\text{eff}}(t) = \frac{p(t)}{1 + e^{-t+t+\tau_{\text{ref}}}} \]  
(14.17)

où \( \hat{t} \) est le temps du dernier PA. La période réfractaire \( \tau_{\text{ref}} \) était 5 ms.

Dans un premier protocole, la probabilité de base de décharger était \( p_0 = 0.01 \). Nous avons alors généré deux stimuli : le premier voyait sa probabilité de décharger augmenter à 0.2 entre 300 et 600 ms, tandis que le second la voyait augmenter à 0.2 entre 400 et 700 ms. Nous avons alors calculé 400 distances de Victor-Purpura, Van Rossum et SRM entre paires de SPA provenant de stimuli identiques ainsi que 400 distances entre paires de SPA provenant de stimuli différents, voir figure 14.9.

Dans un second protocole, la probabilité de base de décharger était \( p_0 = 0.01 \). Nous avons alors généré deux stimuli : le premier voyait sa probabilité de décharger augmenter à 0.2 entre 300 et 600 ms, tandis que le second la voyait augmenter à 0.4 entre 300 et 600 ms. Nous avons alors calculé 400 distances de Victor-Purpura, Van Rossum et SRM entre paires de SPA provenant de stimuli identiques ainsi que 400 distances entre paires de SPA provenant de stimuli différents, voir figure 14.10.

Dans les deux protocoles, deux observations peuvent être faites :

- la distance SRM fournit une meilleure discrimination des stimuli que celle de Victor-Purpura et de Van Rossum, c’est-à-dire que la superposition des distributions intra et inter est plus faible.
- avec la distance SRM, les distances intra sont beaucoup plus petites comparées aux distances inter, suggérant un rapport signal-sur-bruit plus élevé. Ceci peut être très favorables à une entropie conditionnelle métrique faible.

14.6.1. Organisation des réponses des mécanorécepteurs

Lorsque l’on observe les réponses des mécanorécepteurs, on observe une certaine logique, voir figure 14.11.

- Les latences sont toujours globalement sinusoidales en fonction de la direction,
- la direction privilégiée est toujours la même, quelques soient les autres paramètres,
- les latences augmentent lorsque force diminue,
- l’amplitude de la sinusoïde augmente lorsqu’on augmente l’angle.

Ces observations montrent la haute organisation des réponses des mécanorécepteurs.

14.6.2. Organisation avec la distance SRM : un rôle potentiel pour les PA suivants

Pour quantifier davantage ces observations, nous avons appliqué la distance SRM aux réponses de population de mécanorécepteurs. Nous avons ensuite appliqué une technique de multi-dimensional scaling à l’ensemble de distances obtenues. Lorsqu’on a un ensemble de points
14.6. Résultats

et les distances entre eux, on peut les représenter dans l’espace. Mais ce n’est pas toujours possible en 3 dimensions (exemple, on ne peut pas trouver 5 points équidistants dans un espace à 3 dimensions). On cherche alors la projection dans un espace à 3 dimensions qui conserve au mieux les distances.

Si on ne considère que le premier PA de chaque mécanorécepteur, en utilisant des valeurs des paramètres de la distance plausibles, il est impossible d’obtenir une représentation logique des réponses des mécanorécepteurs, c’est-à-dire que les points associés à un même stimulus forment des amas, mais ces amas sont distribués de façon quasi aléatoire. On a donc une information métrique maximale mais une faible organisation. En revanche, si on considère les PA suivants, en plus de l’information métrique maximale, on observe des amas distribués logiquement, voir figure 14.12). Par exemple, les réponses aux 4 stimulations dans des directions orthogonales apparaissent aux sommets d’un carré.

Sur la figure 14.13, 12 stimulations ont été utilisées : 5, 10 et 20 degrés dans 4 directions orthogonales.

De telles organisations ne s’obtiennent qu’en prenant en compte la totalité de la séquence de PA. En conjonction avec les résultats sur l’information métrique, ceci suggère que les premiers PA permettent une reconstruction non-ambigue des stimulations et que les PA suivants fournissent une représentation plus précise de l’espace des stimulations, permettant ainsi des processus de haut-niveau, telle que la généralisation.
Figure 14.3.: (a) Exemple de distances intra et inter au cours du temps. La condition de discrimination optimale est atteinte après 105 ms, lorsque les distributions cessent de se superposer. (b) Exemple d’information métrique et d’entropie conditionnelle métrique en fonction de la distance critique (haut) pour les distributions de distances intra et inter (bas). (c) Valeurs de l’information de Shannon, de l’information obtenue par algorithme des k plus proches voisins et information métrique pour quelques cas simplifiés. (d) Exemple simple montrant l’additivité de l’entropie de Shannon et celle métrique avec des probabilités a, b, c et des similarités x, y, z. (e) Versions locales et globales de l’information métrique. La taille de l’ensemble orange est plus grande que la distance entre les verts et violets, l’information métrique globale ne peut donc être maximale avec une entropie conditionnelle métrique nulle. En revanche, chaque ensemble a une taille inférieure à la distance avec le plus proche ensemble. L’information métrique locale peut donc être maximale avec une entropie conditionnelle nulle.
Figure 14.4.: Discrimination optimale dans un régime à faible bruit. Cinq catégories (lignes bleues) doivent être discriminées. Ces résultats ont été obtenus avec 4 valeurs différentes du paramètre de compromis $\alpha = 0, 10^{-7}, 10^{-5}, 1$. Pour de faibles valeurs de $\alpha$, la déviation standard $\beta$ de la mesure de similarité (ligne noire) est plus grande au centre des catégories qu’entre elles. Lorsque $\alpha$ augmente, la déviation standard devient plus grande entre les catégories.
14. Résumé scientifique

Figure 14.5.: Discrimination optimale dans un régime à bruit élevé. Cinq catégories (lignes bleues) doivent être discriminées. Ces résultats ont été obtenus avec 4 valeurs différentes du paramètre de compromis $\alpha = 0, 10^{-7}, 10^{-5}, 1$. Pour toutes les valeurs de $\alpha$, la déviation standard devient plus grande entre les catégories.
Figure 14.6.: Entropie métrique de tous les graphes à 4 sommets. Lorsque l’entropie de Shannon est également définie, nous l’indiquons par un Sh.
Figure 14.7.: Analyse de théorie de l’information métrique sur un ensemble limité de stimulations. Les latences des premiers PA de 42 mécanorécepteurs en réponse à 5 stimulations (dans 5 directions différentes) présentées 100 fois chacune sont utilisées ici. (a) Trois exemples de réponses de la population de 42 mécanorécepteurs pour trois paramètres de contact différents. (b) Evolution des distances intra (rouge) et inter (bleu) stimuli avec le temps. La condition de discrimination parfaite est atteinte après 52 ms. (c) Information de Shannon (bleu), information métrique globale (rouge) et information métrique locale (vert) au cours du temps. (d) Matrices de distances entre 10 présentations de chacune des 5 stimulations avant la discrimination parfaite (40 ms) et après (60 ms).
Figure 14.8.: Analyse de théorie de l’information métrique sur l’ensemble complet de stimulations. Les latences des premiers PA de 42 mécanorécepteurs en réponse à 81 stimulations présentées 100 fois chacune. (a) Information de Shannon (bleu), information métrique globale (rouge) et information métrique locale (vert) au cours du temps. (b) distribution des distances critiques pour les 81 stimulations. (c) matrices de distances entre 10 présentations de chacune des 5 stimulations avant la discrimination parfaite (40 ms) et peu avant (65 ms).
14. Résumé scientifique

Figure 14.9.: Comparaison des distances de Victor-Purpura, Van Rossum et SRM entre des processus de Poisson. (haut) Victor-Purpura, (milieu) Van Rossum et (bas) SRM. Les distances intra stimuli sont en bleu et les inter stimuli sont en rouge.

Figure 14.10.: Comparaison des distances de Victor-Purpura, Van Rossum et SRM entre des processus de Poisson. (haut) Victor-Purpura, (milieu) Van Rossum et (bas) SRM. Les distances intra stimuli sont en bleu et les inter stimuli sont en rouge.
14.6. Résultats

Figure 14.11.: Latences d’un seul mécanorécepteur en réponse à plusieurs stimuli. Les lignes pleines, discontinues et mixtes correspondent aux forces 4, 2, 1N. Les lignes bleues, rouges et vertes correspondent aux angles 20, 10 and 5 degrés. Les marqueurs point, cercle et étoile correspondent aux trois courbures de la sonde. En abscisse se trouve la direction de la force.

Figure 14.12.: Multi-dimensional scaling des réponses à 7 stimulations.
Figure 14.13.: Multi-dimensional scaling des réponses à 12 stimulations : 5, 10 et 20 degrés dans 4 directions orthogonales. (gauche) vue du dessu, (droite) vue de face.


Barrow, J. D. *The World Within the World* (Oxford University Press ed.).


BIBLIOGRAPHY


Dawkins, R. (2005). The universe is queerer than we can suppose. TED talk.


Johansson, R. S. personal communication.


Schrödinger, E. (1922). What is a law of nature? Inaugural Lecture at University of Zurich.


References about the authors of the quotes


Eric Tabarly: Nantes (France) 1931 - Irish Sea 1998. French sailor known for his series of boats all called Pen Duick (small black head in Breton)

Alber Jacquard: Lyon (France) 1925. French scientist and essayist.

Félix le Dantec: Plougastel-Daoulas (France) 1869 - Paris 1917. French biologist and philosopher of science.


Bernhard Riemann: Breselenz (Hannover state) 1826 - Selasca (Italy) 1866. German mathematician.

Ludwig Wittgenstein: Vienna 1889 - Cambridge (United Kingdom) 1951. Austrian-British philosopher best know for his Tractatus Logico-Philosophicus.

Roger Penrose: Colchester (United-Kingdom) 1931. British mathematician.


Programme: French experimental rock band.
Résumé

Pour étudier la neurotransmission, nous proposons une extension de l’information de Shannon, appelée information métrique, qui intègre explicitement les relations métriques entre les signaux via une distance sur les trains d’impulsions (TI). La métrique est interprétée comme une projection des propriétés du décodeur sur l’espace des TI. Ceci permet donc de déterminer quels sont les paramètres optimaux des neurones qui reçoivent ces signaux pour transmettre une quantité d’information maximale. Nous appliquons cette méthode à des données de microneurographie des mécanorécepteurs du doigt avec une distance de Victor-Purpura. Après quelques dizaines de millisecondes, l’information métrique est maximale. Nous proposons ensuite un modèle de Noyau Cunéiforme (NC), le premier relai des mécanorécepteurs au système nerveux central, sans récurrence que nous connectons aux mécanorécepteurs d’après des données de convergence et divergence et des données électrophysiologiques enregistrées chez le chat. Pour estimer la capacité d’une telle couche de neurones à propager l’information, nous implémentons un règle de plasticité biologiquement plausible dépendante du minutage des TI. Ce réseau parvient à transmettre l’information métrique et l’organisation précise des signaux des mécanorécepteurs en un temps très court (5-10 ms). Enfin, nous définissons une nouvelle distance inspirée du traitement des TI par un neurone réel. Avec des paramètres adéquats, l’espace de réponse de la population de mécanorécepteurs peut être isométrique à celui des stimulations. Cette organisation métrique peut être vue comme la possibilité pour le système nerveux central de généraliser.

Mots-clé : codage neuronal, métrique des séquences de potentiels d’action, théorie de l’information, système somatosensoriel, noyau cunéiforme, microneurographie.

Abstract

In order to study neurotransmission, we propose an extension of Shannon information, dubbed metrical information, that explicitly embeds the metrical relations between signals through a spike train metrics. The metrics is interpreted as a projection of the properties of the decoder onto the space of spike trains. Thus, it allows to predict what are the optimal parameters of the neurons that receive these signals to transmit a maximal amount of information. We apply this new method to microneurography recordings of fingertip mechanoreceptors with the Victor-Purpura spike train metrics. We show that after a few tens of milliseconds, it is possible to reach a maximal metrical information. We then design a feedforward model of the cuneate nucleus (CN) that we connect to the cutaneous receptors according to divergence and convergence recordings in the cat. The properties of the individual CN neurons are also based on electrophysiological recordings in the cat. In order to assess the capacity of such a layer of neurons to convey information, we implemented a biologically plausible spike-time dependent plasticity rule. We show that the layer can transmit the metrical information and the high organization of the mechanoreceptors signals with a very short delay (5-10 ms). Finally, we design a new spike train distance that is inspired by the processing of signals by real neurons. We show with this distance that, with appropriate parameters, the space of the population spike trains can be isometric to the stimulus space. This metrical organization can be seen as a basis for the ability of the central nervous system to generalize.

Keywords: neural coding, spike train metrics, information theory, somatosensory system, cuneate nucleus, microneurography.