Cerebral Cortex
Principles of Operation

Edmund T. Rolls
Oxford Centre for Computational Neuroscience, Oxford, UK
Preface

The overall aim of this book is to provide insight into the principles of operation of the cerebral cortex. These are key to understanding how we, as humans, function.

There have been few previous attempts to set out some of the important principles of operation of the cortex, and this book is pioneering. I have asked some of the leading investigators in neuroscience about their views on this, and most have not had many well formulated answers or hypotheses. As clear hypotheses are needed in this most important area of 21st century science, how our brains work, I have formulated a set of hypotheses to guide thinking and future research. I present evidence for many of the hypotheses, but at the same time we must all recognise that hypotheses and theory in science are there to be tested, and hopefully refined rather than rejected. Nevertheless, such theories and hypotheses are essential to progress, and it is in this frame of reference that I present the theories, hypotheses, and ideas that I have produced and collected together.

This book focusses on the principles of operation of the cerebral cortex, because at this time it is possible to propose and describe many principles, and many are likely to stand the test of time, and provide a foundation I believe for further developments, even if some need to be changed. In this context, I have not attempted to produce an overall theory of operation of the cerebral cortex, because at this stage of our understanding, such a theory would be incorrect or incomplete. I believe though that many of the principles will be important, and that many will provide the foundations for more complete theories of the operation of the cerebral cortex.

Given that many different principles of operation of the cortex are proposed in this book, with often several principles in each Chapter, the reader may find it convenient to take one Chapter at a time, and think about the issues raised in each Chapter, as the overall enterprise is large. The Highlights sections provided at the end of each Chapter may be useful in helping the reader to appreciate the different principles being considered in each Chapter.

To understand how the cortex works, including how it functions in perception, memory, attention, decision-making, and cognitive functions, it is necessary to combine different approaches, including neural computation. Neurophysiology at the single neuron level is needed because this is the level at which information is exchanged between the computing elements of the brain. Evidence from the effects of brain damage, including that available from neuropsychology, is needed to help understand what different parts of the system do, and indeed what each part is necessary for. Neuroimaging is useful to indicate where in the human brain different processes take place, and to show which functions can be dissociated from each other. Knowledge of the biophysical and synaptic properties of neurons is essential to understand how the computing elements of the brain work, and therefore what the building blocks of biologically realistic computational models should be. Knowledge of the anatomical and functional architecture of the cortex is needed to show what types of neuronal network actually perform the computation. And finally the approach of neural computation is needed, as this is required to link together all the empirical evidence to produce an understanding of how the system actually works. This book utilizes evidence from all these disciplines to develop an understanding of how different types of memory, perception, attention, and decision-making are implemented by processing in the cerebral cortex.
I emphasize that to understand how memory, perception, attention, decision-making, cognitive functions, and actions are produced in the cortex, we are dealing with large-scale computational systems with interactions between the parts, and that this understanding requires analysis at the computational and global level of the operation of many neurons to perform together a useful function. Understanding at the molecular level is important for helping to understand how these large-scale computational processes are implemented in the brain, but will not by itself give any account of what computations are performed to implement these cognitive functions. Instead, understanding cognitive functions such as object recognition, memory recall, attention, and decision-making requires single neuron data to be closely linked to computational models of how the interactions between large numbers of neurons and many networks of neurons allow these cognitive problems to be solved. The single neuron level is important in this approach, for the single neurons can be thought of as the computational units of the system, and is the level at which the information is exchanged by the spiking activity between the computational elements of the brain. The single neuron level is therefore, because it is the level at which information is communicated between the computing elements of the brain, the fundamental level of information processing, and the level at which the information can be read out (by recording the spiking activity) in order to understand what information is being represented and processed in each brain area.

With its focus on how the brain and especially how the cortex works at the computational neuroscience level, this book is distinct from the many excellent books on neuroscience that describe much evidence about brain structure and function, but do not aim to provide an understanding of how the brain works at the computational level. This book aims to forge an understanding of how some key brain systems may operate at the computational level, so that we can understand how the cortex actually performs some of its complex and necessarily computational functions in memory, perception, attention, decision-making, cognitive functions, and actions.

A test of whether one’s understanding is correct is to simulate the processing on a computer, and to show whether the simulation can perform the tasks of cortical systems, and whether the simulation has similar properties to the real cortex. The approach of neural computation leads to a precise definition of how the computation is performed, and to precise and quantitative tests of the theories produced. How memory systems in the cortex work is a paradigm example of this approach, because memory-like operations which involve altered functionality as a result of synaptic modification are at the heart of how many computations in the cortex are performed. It happens that attention and decision-making can be understood in terms of interactions between and fundamental operations in memory systems in the cortex, and therefore it is natural to treat these areas of cognitive neuroscience in this book. The same fundamental concepts based on the operation of neuronal circuitry can be applied to all these functions, as is shown in this book.

One of the distinctive properties of this book is that it links the neural computation approach not only firmly to neuronal neurophysiology, which provides much of the primary data about how the cortex operates, but also to psychophysical studies (for example of attention); to neuropsychological studies of patients with brain damage; and to functional magnetic resonance imaging (fMRI) (and other neuroimaging) approaches. The empirical evidence that is brought to bear is largely from non-human primates and from humans, because of the considerable similarity of their cortical systems.

In this book, I have not attempted to produce a single computational theory of how the cortex operates. Instead, I have highlighted many different principles of cortical function, most of which are likely to be building blocks of how our cortex operates. The reason for this approach is that many of the principles may well be correct, and useful in understanding how the cortex operates, but some might turn out not to be useful or correct. The aim of this
book is therefore to propose some of the fundamental principles of operation of the cerebral cortex, many or most of which will provide a foundation for understanding the operation of the cortex, rather than to produce a single theory of operation of the cortex, which might be disproved if any one of its elements was found to be weak.

The overall aims of the book are developed further, and the plan of the book is described, in Chapter 1, Section 1.1. Some of the main Principles of Operation of the Cerebral Cortex that I describe can be found in the titles of Chapters 2–22; but in practice, most Chapters include several Principles of Operation, which will appear in the Highlights to each Chapter. Section 26.5 may be useful in addition to the Highlights, for Section 26.5 draws together in a synthesis some of the Principles of Operation of the Cerebral Cortex that are described in the book. Further evidence on how these principles are relevant to the operation of different cortical areas and systems and operate together is provided in Chapters 24–25. In these Chapters, the operation of two major cortical systems, those involved in memory and in visual object recognition, are considered to illustrate how the principles are combined to implement two different key cortical functions. The Appendices provide some of the more formal and quantitative properties of the operation of neuronal systems, and are provided because they provide a route to a deeper understanding on the principles, and to enable the presentation in earlier Chapters to be at a readily approachable level. The Appendices describe many of the building blocks of the neurocomputational approach, and are designed to be useful for teaching. Appendix D describes Matlab software that has been made available with this book to provide simple demonstrations of the operation of some key neuronal networks related to cortical function. The programs are available at http://www.oxcns.org.

Part of the material described in the book reflects work performed in collaboration with many colleagues, whose tremendous contributions are warmly appreciated. The contributions of many will be evident from the references cited in the text. Especial appreciation is due to Gustavo Deco, Simon M. Stringer, and Alessandro Treves who have contributed greatly in an always interesting and fruitful research collaboration on computational aspects of brain function, and to many neurophysiology and functional neuroimaging colleagues who have contributed to the empirical discoveries that provide the foundation to which the computational neuroscience must always be closely linked, and whose names are cited throughout the text. Much of the work described would not have been possible without financial support from a number of sources, particularly the Medical Research Council of the UK, the Human Frontier Science Program, the Wellcome Trust, and the James S. McDonnell Foundation. I am also grateful to many colleagues who I have consulted while writing this book, including Joel Price (Washington University School of Medicine), and Donald Wilson (New York University). Dr Patrick Mills is warmly thanked for his comments on the text. Section 24.3.12 on ars memoriae is warmly dedicated to my colleagues at Corpus Christi College, Oxford. The book was typeset by the author using \LaTeX and WinEdit.

The cover includes part of the picture *Pandora* painted in 1896 by J. W. Waterhouse. The metaphor is to look inside the system of the mind and the brain, in order to understand how the brain functions, and thereby better to understand and treat its disorders. The cover also includes an image of the dendritic morphology of excitatory neurons in S1 whisker barrel cortex (Fig. 1.14) (adapted from Marcel Oberlaender, Christiaan P.J. de Kock, Randy M. Bruno, Alejandro Ramirez, Hanno S. Meyer, Vincent J. Dercksen, Moritz Helmaesstetter and Bert Sakmann, Cell type-specific three-dimensional structure of thalamocortical circuits in a column of rat vibrissal cortex, Cerebral Cortex, 2012, Vol. 22, issue 10, pp. 2375–2391, by permission of Oxford University Press). The cover also includes a diagram of the computational circuitry of the hippocampus by the author (Fig. 24.1). The aim of these second two images is to highlight the importance of moving from the anatomy of the cortex using all the approaches available including neuronal network models that address and
incorporate neurophysiological discoveries to lead to an understanding of how the cortex operates computationally.


I dedicate this work to the overlapping group: my family, friends, and colleagues – in salutem praesentium, in memoriam absentium.
## Contents

1 Introduction 1
   1.1 Principles of operation of the cerebral cortex: introduction and plan 1
   1.2 Neurons 4
   1.3 Neurons in a network 6
   1.4 Synaptic modification 8
   1.5 Long-term potentiation and long-term depression 9
   1.6 Distributed representations 14
      1.6.1 Definitions 14
      1.6.2 Advantages of different types of coding 15
   1.7 Neuronal network approaches versus connectionism 16
   1.8 Introduction to three neuronal network architectures 17
   1.9 Systems-level analysis of brain function 18
      1.9.1 Ventral cortical visual stream 19
      1.9.2 Dorsal cortical visual stream 21
      1.9.3 Hippocampal memory system 23
      1.9.4 Frontal lobe systems 23
      1.9.5 Brodmann areas 24
   1.10 The fine structure of the cerebral neocortex 27
      1.10.1 The fine structure and connectivity of the neocortex 27
      1.10.2 Excitatory cells and connections 27
      1.10.3 Inhibitory cells and connections 29
      1.10.4 Quantitative aspects of cortical architecture 32
      1.10.5 Functional pathways through the cortical layers 34
      1.10.6 The scale of lateral excitatory and inhibitory effects, and modules 38
   1.11 Highlights 39

2 Hierarchical organization 40
   2.1 Introduction 40
   2.2 Hierarchical organization in sensory systems 41
      2.2.1 Hierarchical organization in the ventral visual system 41
      2.2.2 Hierarchical organization in the dorsal visual system 46
      2.2.3 Hierarchical organization of taste processing 48
      2.2.4 Hierarchical organization of olfactory processing 57
      2.2.5 Hierarchical multimodal convergence of taste, olfaction, and vision 59
      2.2.6 Hierarchical organization of auditory processing 64
   2.3 Hierarchical organization of reward value processing 67
   2.4 Hierarchical organization of connections to the frontal lobe for short-term memory 68
   2.5 Highlights 69

3 Localization of function 72
   3.1 Hierarchical processing 72
3.2 Short-range neocortical recurrent collaterals 72
3.3 Topographic maps 72
3.4 Modularity 72
3.5 Lateralization of function 73
3.6 Ventral and dorsal cortical areas 73
3.7 Highlights 74

4 Recurrent collateral connections and attractor networks 75
4.1 Introduction 75
4.2 Attractor networks implemented by the recurrent collaterals 75
4.3 Evidence for attractor networks implemented by recurrent collateral connections 76
4.3.1 Short-term Memory 77
4.3.2 Long-term Memory 80
4.3.3 Decision-Making 80
4.4 The storage capacity of attractor networks 80
4.5 A global attractor network in hippocampal CA3, but local in neocortex 81
4.6 The speed of operation of cortical attractor networks 83
4.7 Dilution of recurrent collateral cortical connectivity 83
4.8 Self-organizing topographic maps in the neocortex 85
4.9 Attractors formed by forward and backward connections between cortical areas? 85
4.10 Interacting attractor networks 86
4.11 Highlights 90

5 The noisy cortex: stochastic dynamics, decisions, and memory 91
5.1 Reasons why the brain is inherently noisy and stochastic 91
5.2 Attractor networks, energy landscapes, and stochastic neurodynamics 95
5.3 A multistable system with noise 98
5.4 Stochastic dynamics and the stability of short-term memory 101
5.4.1 Analysis of the stability of short-term memory 103
5.4.2 Stability and noise in a model of short-term memory 104
5.5 Long-term memory recall 106
5.6 Stochastic dynamics and probabilistic decision-making in an attractor network 106
5.6.1 Decision-making in an attractor network 107
5.6.2 Theoretical framework: a probabilistic attractor network 107
5.6.3 Stationary multistability analysis: mean-field 110
5.6.4 Integrate-and-fire simulations of decision-making: spiking dynamics 112
5.6.5 Reaction times of the neuronal responses 116
5.6.6 Percentage correct 117
5.6.7 Finite-size noise effects 117
5.6.8 Comparison with neuronal data during decision-making 119
5.6.9 Testing the model of decision-making with human functional neuroimaging 122
5.6.10 Decisions based on confidence in one's decisions: self-monitoring 129
5.6.11 Decision-making with multiple alternatives 131
5.6.12 The matching law 132
5.6.13 Comparison with other models of decision-making 132
5.7 Perceptual decision-making and rivalry 134
5.8 Symmetry-breaking 135
5.9 The evolutionary utility of probabilistic choice 135
5.10 Selection between conscious vs unconscious decision-making, and free will 136
5.11 Creative thought 137
5.12 Unpredictable behaviour 138
5.13 Predicting a decision before the evidence is applied 138
5.14 Highlights 140

6 Attention, short-term memory, and biased competition 141
6.1 Bottom-up attention 141
6.2 Top-down attention – biased competition 143
   6.2.1 The biased competition hypothesis 143
   6.2.2 Biased competition – single neuron studies 145
   6.2.3 Non-spatial attention 147
   6.2.4 Biased competition – fMRI 149
   6.2.5 A basic computational module for biased competition 149
   6.2.6 Architecture of a model of attention 150
   6.2.7 Simulations of basic experimental findings 154
   6.2.8 Object recognition and spatial search 158
   6.2.9 The neuronal and biophysical mechanisms of attention 163
   6.2.10 ‘Serial’ vs ‘parallel’ attentional processing 167
6.3 Top-down attention – biased activation 171
   6.3.1 Selective attention can selectively activate different cortical areas 171
   6.3.2 Sources of the top-down modulation of attention 173
   6.3.3 Granger causality used to investigate the source of the top-down biasing 174
   6.3.4 Top-down cognitive modulation 175
   6.3.5 A top-down biased activation model of attention 178
6.4 Conclusions 181
6.5 Highlights 184

7 Diluted connectivity 186
7.1 Introduction 186
7.2 Diluted connectivity and the storage capacity of attractor networks 187
   7.2.1 The autoassociative or attractor network architecture being studied 187
   7.2.2 The storage capacity of attractor networks with diluted connectivity 188
   7.2.3 The network simulated 190
   7.2.4 The effects of diluted connectivity on the capacity of attractor networks 192
   7.2.5 Synthesis of the effects of diluted connectivity in attractor networks 197
7.3 The effects of dilution on the capacity of pattern association networks 198
7.4 The effects of dilution on the performance of competitive networks 201
   7.4.1 Competitive Networks 201
   7.4.2 Competitive networks without learning but with diluted connectivity 202
   7.4.3 Competitive networks with learning and with diluted connectivity 203
   7.4.4 Competitive networks with learning and with full (undiluted) connectivity 205
   7.4.5 Overview and implications of diluted connectivity in competitive networks 206
7.5 The effects of dilution on the noise in attractor networks 207
7.6 Highlights 207

8 Coding principles 209
8.1 Types of encoding 209
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>8.2 Place coding with sparse distributed firing rate representations</strong></td>
<td></td>
</tr>
<tr>
<td>8.2.1 Reading the code used by single neurons</td>
<td>210</td>
</tr>
<tr>
<td>8.2.2 Understanding the code provided by populations of neurons</td>
<td>214</td>
</tr>
<tr>
<td><strong>8.3 Synchrony, coherence, and binding</strong></td>
<td>221</td>
</tr>
<tr>
<td><strong>8.4 Principles by which the representations are formed</strong></td>
<td></td>
</tr>
<tr>
<td><strong>8.5 Information encoding in the human cortex</strong></td>
<td>223</td>
</tr>
<tr>
<td><strong>8.6 Highlights</strong></td>
<td>226</td>
</tr>
<tr>
<td><strong>9 Synaptic modification for learning</strong></td>
<td>227</td>
</tr>
<tr>
<td>9.1 Introduction</td>
<td>227</td>
</tr>
<tr>
<td>9.2 Associative synaptic modification implemented by long-term potentiation</td>
<td>227</td>
</tr>
<tr>
<td>9.3 Forgetting in associative neural networks, and memory reconsolidation</td>
<td>228</td>
</tr>
<tr>
<td>9.3.1 Forgetting</td>
<td>228</td>
</tr>
<tr>
<td>9.3.2 Factors that influence synaptic modification</td>
<td>230</td>
</tr>
<tr>
<td>9.3.3 Reconsolidation</td>
<td>232</td>
</tr>
<tr>
<td>9.4 Spike-timing dependent plasticity</td>
<td>233</td>
</tr>
<tr>
<td>9.5 Long-term synaptic depression in the cerebellar cortex</td>
<td>233</td>
</tr>
<tr>
<td>9.6 Reward prediction error learning</td>
<td>234</td>
</tr>
<tr>
<td>9.6.1 Blocking and delta-rule learning</td>
<td>234</td>
</tr>
<tr>
<td>9.6.2 Dopamine neuron firing and reward prediction error learning</td>
<td>234</td>
</tr>
<tr>
<td>9.7 Highlights</td>
<td>240</td>
</tr>
<tr>
<td><strong>10 Synaptic and neuronal adaptation and facilitation</strong></td>
<td></td>
</tr>
<tr>
<td>10.1 Mechanisms for neuronal adaptation and synaptic depression and facilitation</td>
<td>241</td>
</tr>
<tr>
<td>10.1.1 Sodium inactivation leading to neuronal spike-frequency adaptation</td>
<td>241</td>
</tr>
<tr>
<td>10.1.2 Calcium activated hyper-polarizing potassium current</td>
<td>242</td>
</tr>
<tr>
<td>10.1.3 Short-term synaptic depression and facilitation</td>
<td>243</td>
</tr>
<tr>
<td>10.2 Short-term depression of thalamic input to the cortex</td>
<td>244</td>
</tr>
<tr>
<td>10.3 Relatively little adaptation in primate cortex when it is operating normally</td>
<td>244</td>
</tr>
<tr>
<td>10.4 Acetylcholine, noradrenaline, and other modulators of adaptation and facilitation</td>
<td>247</td>
</tr>
<tr>
<td>10.4.1 Acetylcholine</td>
<td>247</td>
</tr>
<tr>
<td>10.4.2 Noradrenergic neurons</td>
<td>248</td>
</tr>
<tr>
<td>10.5 Synaptic depression and sensory-specific satiety</td>
<td>249</td>
</tr>
<tr>
<td>10.6 Neuronal and synaptic adaptation, and the memory for sequential order</td>
<td>250</td>
</tr>
<tr>
<td>10.7 Destabilization of short-term memory by adaptation or synaptic depression</td>
<td>250</td>
</tr>
<tr>
<td>10.8 Non-reward computation in the orbitofrontal cortex using synaptic depression</td>
<td>251</td>
</tr>
<tr>
<td>10.9 Synaptic facilitation and a multiple-item short-term memory</td>
<td>253</td>
</tr>
<tr>
<td>10.10 Synaptic facilitation in decision-making</td>
<td>253</td>
</tr>
<tr>
<td>10.11 Highlights</td>
<td>254</td>
</tr>
<tr>
<td><strong>11 Backprojections in the neocortex</strong></td>
<td>255</td>
</tr>
<tr>
<td>11.1 Architecture</td>
<td>255</td>
</tr>
<tr>
<td>11.2 Learning</td>
<td>257</td>
</tr>
<tr>
<td>11.3 Recall</td>
<td>258</td>
</tr>
<tr>
<td>11.4 Semantic priming</td>
<td>259</td>
</tr>
<tr>
<td>11.5 Top-down Attention</td>
<td>259</td>
</tr>
<tr>
<td>11.6 Autoassociative storage, and constraint satisfaction</td>
<td>261</td>
</tr>
</tbody>
</table>
12 Memory and the hippocampus

12.1 Introduction 262
12.2 Hippocampal circuitry and connections 262
12.3 The hippocampus and episodic memory 262
12.4 Autoassociation in the CA3 network for episodic memory 263
12.5 The dentate gyrus as a pattern separation mechanism, and neurogenesis 265
12.6 Rodent place cells vs primate spatial view cells 265
12.7 Backprojections, and the recall of information from the hippocampus to neocortex 266
12.8 Subcortical structures connected to the hippocampo-cortical memory system 267
12.9 Highlights 267

13 Limited neurogenesis in the adult cortex

13.1 No neurogenesis in the adult neocortex 269
13.2 Limited neurogenesis in the adult hippocampal dentate gyrus 269
13.3 Neurogenesis in the chemosensing receptor systems 270
13.4 Highlights 271

14 Invariance learning and vision

14.1 Hierarchical cortical organization with convergence 272
14.2 Feature combinations 272
14.3 Sparse distributed representations 273
14.4 Self-organization by feedforward processing without a teacher 273
14.5 Learning guided by the statistics of the visual inputs 274
14.6 Bottom up saliency 275
14.7 Lateral interactions shape receptive fields 276
14.8 Top-down selective attention vs feedforward processing 277
14.9 Topological maps to simplify connectivity 278
14.10 Biologically decodable output representations 279
14.11 Highlights 279

15 Emotion, motivation, reward value, pleasure, and their mechanisms

15.1 Emotion, reward value, and their evolutionary adaptive utility 281
15.2 Motivation and reward value 283
15.3 Principles of cortical design for emotion and motivation 283
15.4 Objects are first represented independently of reward value 284
15.5 Specialized systems for face identity and expression processing in primates 286
15.6 Unimodal processing to the object level before multimodal convergence 287
15.7 A common scale for reward value 287
15.8 Sensory-specific satiety 287
15.9 Economic value is represented in the orbitofrontal cortex 288
15.10 Neuroeconomics vs classical microeconomics 288
15.11 Output systems influenced by orbitofrontal cortex reward value representations 289
15.12 Decision-making about rewards in the anterior orbitofrontal cortex 291
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.13</td>
<td>Probabilistic emotion-related decision-making</td>
<td>292</td>
</tr>
<tr>
<td>15.14</td>
<td>Non-reward, error, neurons in the orbitofrontal cortex</td>
<td>292</td>
</tr>
<tr>
<td>15.15</td>
<td>Reward reversal learning in the orbitofrontal cortex</td>
<td>296</td>
</tr>
<tr>
<td>15.16</td>
<td>Dopamine neurons and emotion</td>
<td>301</td>
</tr>
<tr>
<td>15.17</td>
<td>The explicit reasoning system vs the emotional system</td>
<td>301</td>
</tr>
<tr>
<td>15.18</td>
<td>Pleasure</td>
<td>302</td>
</tr>
<tr>
<td>15.19</td>
<td>Personality relates to differences in sensitivity to rewards and punishers</td>
<td>302</td>
</tr>
<tr>
<td>15.20</td>
<td>Highlights</td>
<td>303</td>
</tr>
<tr>
<td>16</td>
<td>Noise in the cortex, stability, psychiatric disease, and aging</td>
<td>305</td>
</tr>
<tr>
<td>16.1</td>
<td>Stochastic noise, attractor dynamics, and schizophrenia</td>
<td>305</td>
</tr>
<tr>
<td>16.1.1</td>
<td>Introduction</td>
<td>305</td>
</tr>
<tr>
<td>16.1.2</td>
<td>A dynamical systems hypothesis of the symptoms of schizophrenia</td>
<td>307</td>
</tr>
<tr>
<td>16.1.3</td>
<td>The depth of the basins of attraction: mean-field flow analysis</td>
<td>308</td>
</tr>
<tr>
<td>16.1.4</td>
<td>Decreased stability produced by reduced NMDA conductances</td>
<td>309</td>
</tr>
<tr>
<td>16.1.5</td>
<td>Increased distractibility produced by reduced NMDA conductances</td>
<td>311</td>
</tr>
<tr>
<td>16.1.6</td>
<td>Synthesis: network instability and schizophrenia</td>
<td>312</td>
</tr>
<tr>
<td>16.2</td>
<td>Stochastic noise, attractor dynamics, and obsessive-compulsive disorder</td>
<td>316</td>
</tr>
<tr>
<td>16.2.1</td>
<td>Introduction</td>
<td>316</td>
</tr>
<tr>
<td>16.2.2</td>
<td>A hypothesis about obsessive-compulsive disorder</td>
<td>317</td>
</tr>
<tr>
<td>16.2.3</td>
<td>Glutamate and increased depth of the basins of attraction</td>
<td>319</td>
</tr>
<tr>
<td>16.2.4</td>
<td>Synthesis on obsessive-compulsive disorder</td>
<td>322</td>
</tr>
<tr>
<td>16.3</td>
<td>Stochastic noise, attractor dynamics, and depression</td>
<td>325</td>
</tr>
<tr>
<td>16.3.1</td>
<td>Introduction</td>
<td>325</td>
</tr>
<tr>
<td>16.3.2</td>
<td>A non-reward attractor theory of depression</td>
<td>328</td>
</tr>
<tr>
<td>16.3.3</td>
<td>Evidence consistent with the theory</td>
<td>329</td>
</tr>
<tr>
<td>16.3.4</td>
<td>Relation to other brain systems implicated in depression</td>
<td>331</td>
</tr>
<tr>
<td>16.3.5</td>
<td>Implications for treatments</td>
<td>332</td>
</tr>
<tr>
<td>16.3.6</td>
<td>Mania and bipolar disorder</td>
<td>333</td>
</tr>
<tr>
<td>16.4</td>
<td>Stochastic noise, attractor dynamics, and aging</td>
<td>335</td>
</tr>
<tr>
<td>16.4.1</td>
<td>NMDA receptor hypofunction</td>
<td>335</td>
</tr>
<tr>
<td>16.4.2</td>
<td>Dopamine</td>
<td>338</td>
</tr>
<tr>
<td>16.4.3</td>
<td>Impaired synaptic modification</td>
<td>338</td>
</tr>
<tr>
<td>16.4.4</td>
<td>Cholinergic function and memory</td>
<td>339</td>
</tr>
<tr>
<td>16.5</td>
<td>Highlights</td>
<td>343</td>
</tr>
<tr>
<td>17</td>
<td>Syntax and Language</td>
<td>345</td>
</tr>
<tr>
<td>17.1</td>
<td>Neurodynamical hypotheses about language and syntax</td>
<td>345</td>
</tr>
<tr>
<td>17.1.1</td>
<td>Binding by synchrony?</td>
<td>345</td>
</tr>
<tr>
<td>17.1.2</td>
<td>Syntax using a place code</td>
<td>346</td>
</tr>
<tr>
<td>17.1.3</td>
<td>Temporal trajectories through a state space of attractors</td>
<td>347</td>
</tr>
<tr>
<td>17.1.4</td>
<td>Hypotheses about the implementation of language in the cerebral cortex</td>
<td>347</td>
</tr>
<tr>
<td>17.2</td>
<td>Tests of the hypotheses – a model</td>
<td>351</td>
</tr>
<tr>
<td>17.2.1</td>
<td>Attractor networks with stronger forward than backward connections</td>
<td>351</td>
</tr>
<tr>
<td>17.2.2</td>
<td>The operation of a single attractor network module</td>
<td>353</td>
</tr>
<tr>
<td>17.2.3</td>
<td>Spike frequency adaptation mechanism</td>
<td>355</td>
</tr>
<tr>
<td>17.3</td>
<td>Tests of the hypotheses – findings with the model</td>
<td>355</td>
</tr>
<tr>
<td>17.3.1</td>
<td>A production system</td>
<td>355</td>
</tr>
<tr>
<td>17.3.2</td>
<td>A decoding system</td>
<td>356</td>
</tr>
<tr>
<td>17.4</td>
<td>Evaluation of the hypotheses</td>
<td>359</td>
</tr>
</tbody>
</table>
18 Evolutionary trends in cortical design and principles of operation 364
18.1 Introduction 364
18.2 Different types of cerebral neocortex: towards a computational understanding 364
  18.2.1 Neocortex or isocortex 365
  18.2.2 Olfactory (pyriform) cortex 371
  18.2.3 Hippocampal cortex 374
18.3 Addition of areas in the neocortical hierarchy 376
18.4 Evolution of the orbitofrontal cortex 378
18.5 Evolution of the taste and flavour system 379
  18.5.1 Principles 379
  18.5.2 Taste processing in rodents 380
18.6 Evolution of the temporal lobe cortex 381
18.7 Evolution of the frontal lobe cortex 382
18.8 Highlights 382

19 Genetics and self-organization build the cortex 385
19.1 Introduction 385
19.2 Hypotheses about the genes that build cortical neural networks 386
19.3 Genetic selection of neuronal network parameters 390
19.4 Simulation of the evolution of neural networks using a genetic algorithm 391
  19.4.1 The neural networks 391
  19.4.2 The specification of the genes 392
  19.4.3 The genetic algorithm, and general procedure 397
  19.4.4 Pattern association networks 398
  19.4.5 Autoassociative networks 400
  19.4.6 Competitive networks 400
19.5 Evaluation of the gene-based evolution of single-layer networks 401
19.6 The gene-based evolution of multi-layer cortical systems 403
19.7 Highlights 404

20 Cortex versus basal ganglia design for selection 406
20.1 Systems-level architecture of the basal ganglia 406
20.2 What computations are performed by the basal ganglia? 408
20.3 How do the basal ganglia perform their computations? 410
20.4 Comparison of selection in the basal ganglia and cerebral cortex 413
20.5 Highlights 415

21 Sleep and Dreaming 416
21.1 Is sleep necessary for cortical function? 416
21.2 Is sleep involved in memory consolidation? 417
21.3 Dreams 418
21.4 Highlights 419

22 Which cortical computations underlie consciousness? 420
22.1 Introduction 420
### 22.2 A Higher-Order Syntactic Thought (HOST) theory of consciousness

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.2.1 Multiple routes to action</td>
<td>421</td>
</tr>
<tr>
<td>22.2.2 A computational hypothesis of consciousness</td>
<td>423</td>
</tr>
<tr>
<td>22.2.3 Adaptive value of processing that is related to consciousness</td>
<td>425</td>
</tr>
<tr>
<td>22.2.4 Symbol grounding</td>
<td>426</td>
</tr>
<tr>
<td>22.2.5 Qualia</td>
<td>428</td>
</tr>
<tr>
<td>22.2.6 Pathways</td>
<td>429</td>
</tr>
<tr>
<td>22.2.7 Consciousness and causality</td>
<td>430</td>
</tr>
<tr>
<td>22.2.8 Consciousness and higher-order syntactic thoughts</td>
<td>431</td>
</tr>
<tr>
<td>22.3 Selection between conscious vs unconscious decision-making systems</td>
<td>432</td>
</tr>
<tr>
<td>22.3.1 Dual major routes to action: implicit and explicit</td>
<td>432</td>
</tr>
<tr>
<td>22.3.2 The Selfish Gene vs The Selfish Phenotype</td>
<td>439</td>
</tr>
<tr>
<td>22.3.3 Decision-making between the implicit and explicit systems</td>
<td>440</td>
</tr>
<tr>
<td>22.4 Determinism</td>
<td>441</td>
</tr>
<tr>
<td>22.5 Free will</td>
<td>442</td>
</tr>
<tr>
<td>22.6 Content and meaning in representations</td>
<td>443</td>
</tr>
<tr>
<td>22.7 The causal role of consciousness and the relation between the mind and the brain</td>
<td>445</td>
</tr>
<tr>
<td>22.8 Comparison with other theories of consciousness</td>
<td>447</td>
</tr>
<tr>
<td>22.8.1 Higher-order thought theories</td>
<td>447</td>
</tr>
<tr>
<td>22.8.2 Oscillations and temporal binding</td>
<td>449</td>
</tr>
<tr>
<td>22.8.3 A high neural threshold for information to reach consciousness</td>
<td>450</td>
</tr>
<tr>
<td>22.8.4 James–Lange theory and Damasio’s somatic marker hypothesis</td>
<td>451</td>
</tr>
<tr>
<td>22.8.5 LeDoux’s approach to emotion and consciousness</td>
<td>451</td>
</tr>
<tr>
<td>22.8.6 Panksepp’s approach to emotion and consciousness</td>
<td>452</td>
</tr>
<tr>
<td>22.8.7 Global workspace theories of consciousness</td>
<td>452</td>
</tr>
<tr>
<td>22.8.8 Monitoring and consciousness</td>
<td>452</td>
</tr>
<tr>
<td>22.9 Highlights</td>
<td>453</td>
</tr>
</tbody>
</table>

### 23 Cerebellar cortex

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.1 Introduction</td>
<td>455</td>
</tr>
<tr>
<td>23.2 Architecture of the cerebellum</td>
<td>456</td>
</tr>
<tr>
<td>23.2.1 The connections of the parallel fibres onto the Purkinje cells</td>
<td>456</td>
</tr>
<tr>
<td>23.2.2 The climbing fibre input to the Purkinje cell</td>
<td>457</td>
</tr>
<tr>
<td>23.2.3 The mossy fibre to granule cell connectivity</td>
<td>457</td>
</tr>
<tr>
<td>23.3 Modifiable synapses of parallel fibres onto Purkinje cell dendrites</td>
<td>460</td>
</tr>
<tr>
<td>23.4 The cerebellar cortex as a perceptron</td>
<td>460</td>
</tr>
<tr>
<td>23.5 Highlights: differences between cerebral and cerebellar cortex microcircuitry</td>
<td>461</td>
</tr>
</tbody>
</table>

### 24 The hippocampus and memory

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.1 Introduction</td>
<td>463</td>
</tr>
<tr>
<td>24.2 Systems-level functions of the hippocampus</td>
<td>464</td>
</tr>
<tr>
<td>24.2.1 Systems-level anatomy</td>
<td>465</td>
</tr>
<tr>
<td>24.2.2 Evidence from the effects of damage to the hippocampus</td>
<td>467</td>
</tr>
<tr>
<td>24.2.3 The necessity to recall information from the hippocampus</td>
<td>468</td>
</tr>
<tr>
<td>24.2.4 Systems-level neurophysiology of the primate hippocampus</td>
<td>470</td>
</tr>
<tr>
<td>24.2.5 Head direction cells in the presubiculum</td>
<td>478</td>
</tr>
<tr>
<td>24.2.6 Perirhinal cortex, recognition memory, and long-term familiarity memory</td>
<td>479</td>
</tr>
<tr>
<td>24.3 A theory of the operation of hippocampal circuitry as a memory system</td>
<td>486</td>
</tr>
<tr>
<td>24.3.1 Hippocampal circuitry</td>
<td>487</td>
</tr>
<tr>
<td>24.3.2 Entorhinal cortex</td>
<td>488</td>
</tr>
</tbody>
</table>
24.3.3 CA3 as an autoassociation memory 490
24.3.4 Dentate granule cells 509
24.3.5 CA1 cells 515
24.3.6 Recoding in CA1 to facilitate retrieval to the neocortex 515
24.3.7 Backprojections to the neocortex, memory recall, and consolidation 520
24.3.8 Backprojections to the neocortex – quantitative aspects 523
24.3.9 Simulations of hippocampal operation 526
24.3.10 The learning of spatial view and place cell representations 528
24.3.11 Linking the inferior temporal visual cortex to spatial view and place cells 529
24.3.12 A scientific theory of the art of memory: scientia artis memoriae 531
24.4 Tests of the theory of hippocampal cortex operation 531
24.4.1 Dentate gyrus (DG) subregion of the hippocampus 531
24.4.2 CA3 subregion of the hippocampus 535
24.4.3 CA1 subregion of the hippocampus 542
24.5 Evaluation of the theory of hippocampal cortex operation 546
24.5.1 Tests of the theory by hippocampal system subregion analyses 546
24.5.2 Comparison with other theories of hippocampal function 548
24.6 Highlights 552

25 Invariant visual object recognition learning 554
25.1 Introduction 554
25.2 Invariant representations of faces and objects in the inferior temporal visual cortex 555
25.2.1 Processing to the inferior temporal cortex in the primate visual system 555
25.2.2 Translation invariance and receptive field size 556
25.2.3 Reduced translation invariance in natural scenes 557
25.2.4 Size and spatial frequency invariance 560
25.2.5 Combinations of features in the correct spatial configuration 561
25.2.6 A view-invariant representation 562
25.2.7 Learning in the inferior temporal cortex 565
25.2.8 Distributed encoding 568
25.2.9 Face expression, gesture, and view 572
25.2.10 Specialized regions in the temporal cortical visual areas 572
25.3 Approaches to invariant object recognition 576
25.3.1 Feature spaces 577
25.3.2 Structural descriptions and syntactic pattern recognition 578
25.3.3 Template matching and the alignment approach 580
25.3.4 Invertible networks that can reconstruct their inputs 581
25.3.5 Feature hierarchies 582
25.4 Hypotheses about object recognition mechanisms 582
25.5 Computational issues in feature hierarchies 586
25.5.1 The architecture of VisNet 587
25.5.2 Initial experiments with VisNet 596
25.5.3 The optimal parameters for the temporal trace used in the learning rule 603
25.5.4 Different forms of the trace learning rule, and error correction 604
25.5.5 The issue of feature binding, and a solution 612
25.5.6 Operation in a cluttered environment 624
25.5.7 Learning 3D transforms 631
25.5.8 Capacity of the architecture, and an attractor implementation 636
25.5.9 Vision in natural scenes – effects of background versus attention 643
25.5.10 The representation of multiple objects in a scene 651
25.5.11 Learning invariant representations using spatial continuity 653
25.5.12 Lighting invariance 654
25.5.13 Invariant global motion in the dorsal visual system 656
25.5.14 Deformation-invariant object recognition 656
25.5.15 Learning invariant representations of scenes and places 657
25.5.16 Finding and recognising objects in natural scenes 659
25.6 Further approaches to invariant object recognition 663
25.6.1 Other types of slow learning 663
25.6.2 HMAX 663
25.6.3 Sigma-Pi synapses 668
25.6.4 Deep learning 668
25.7 Visuo-spatial scratchpad memory, and change blindness 669
25.8 Processes involved in object identification 670
25.9 Highlights 671
26 Synthesis 674
26.1 Principles of cortical operation, not a single theory 674
26.2 Levels of explanation, and the mind-brain problem 674
26.3 Brain computation compared to computation on a digital computer 676
26.4 Understanding how the brain works 681
26.5 Synthesis on principles of operation of the cerebral cortex 683
26.5.1 Hierarchical organization 683
26.5.2 Localization of function 684
26.5.3 Recurrent collaterals and attractor networks 684
26.5.4 The noisy cortex 685
26.5.5 Top-down attention 685
26.5.6 Diluted connectivity 685
26.5.7 Sparse distributed graded firing rate encoding 685
26.5.8 Synaptic modification 686
26.5.9 Adaptation and facilitation 686
26.5.10 Backprojections 686
26.5.11 Neurogenesis 687
26.5.12 Binding and syntax 687
26.5.13 Evolution of the cerebral cortex 687
26.5.14 Genetic specification of cortical design 687
26.5.15 The cortical systems for emotion 688
26.5.16 Memory systems 688
26.5.17 Visual cortical processing for invariant visual object recognition 689
26.5.18 Cortical lamination, operation, and evolution 689
26.6 Highlights 692
A Introduction to linear algebra for neural networks 694
A.1 Vectors 694
A.1.1 The inner or dot product of two vectors 694
A.1.2 The length of a vector 695
A.1.3 Normalizing the length of a vector 696
A.1.4 The angle between two vectors: the normalized dot product 696
A.1.5 The outer product of two vectors 697
A.1.6 Linear and non-linear systems 698
A.1.7 Linear combinations, linear independence, and linear separability 699
A.2 Application to understanding simple neural networks 700
A.2.1 Capability and limitations of single-layer networks 701
A.2.2 Non-linear networks: neurons with non-linear activation functions 703
B Neuronal network models

B.1 Introduction

B.2 Pattern association memory
B.2.1 Architecture and operation
B.2.2 A simple model
B.2.3 The vector interpretation
B.2.4 Properties
B.2.5 Prototype extraction, extraction of central tendency, and noise reduction
B.2.6 Speed
B.2.7 Local learning rule
B.2.8 Implications of different types of coding for storage in pattern associators

B.3 Autoassociation or attractor memory
B.3.1 Architecture and operation
B.3.2 Introduction to the analysis of the operation of autoassociation networks
B.3.3 Properties
B.3.4 Use of autoassociation networks in the brain

B.4 Competitive networks, including self-organizing maps
B.4.1 Function
B.4.2 Architecture and algorithm
B.4.3 Properties
B.4.4 Utility of competitive networks in information processing by the brain
B.4.5 Guidance of competitive learning
B.4.6 Topographic map formation
B.4.7 Invariance learning by competitive networks
B.4.8 Radial Basis Function networks
B.4.9 Further details of the algorithms used in competitive networks

B.5 Continuous attractor networks
B.5.1 Introduction
B.5.2 The generic model of a continuous attractor network
B.5.3 Learning the synaptic strengths in a continuous attractor network
B.5.4 The capacity of a continuous attractor network: multiple charts
B.5.5 Continuous attractor models: path integration
B.5.6 Stabilization of the activity packet within a continuous attractor network
B.5.7 Continuous attractor networks in two or more dimensions
B.5.8 Mixed continuous and discrete attractor networks

B.6 Network dynamics: the integrate-and-fire approach
B.6.1 From discrete to continuous time
B.6.2 Continuous dynamics with discontinuities
B.6.3 An integrate-and-fire implementation
B.6.4 The speed of processing of attractor networks
B.6.5 The speed of processing of a four-layer hierarchical network
B.6.6 Spike response model

B.7 Network dynamics: introduction to the mean-field approach

B.8 Mean-field based neurodynamics
B.8.1 Population activity
B.8.2 The mean-field approach used in a model of decision-making
B.8.3 The model parameters used in the mean-field analyses of decision-making
B.8.4 A basic computational module based on biased competition
B.8.5 Multimodular neurodynamical architectures

B.9 Sequence memory implemented by adaptation in an attractor network
<table>
<thead>
<tr>
<th>B.10</th>
<th>Error correction networks</th>
<th>792</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.10.1</td>
<td>Architecture and general description</td>
<td>792</td>
</tr>
<tr>
<td>B.10.2</td>
<td>Generic algorithm for a one-layer error correction network</td>
<td>793</td>
</tr>
<tr>
<td>B.10.3</td>
<td>Capability and limitations of single-layer error-correcting networks</td>
<td>793</td>
</tr>
<tr>
<td>B.10.4</td>
<td>Properties</td>
<td>797</td>
</tr>
<tr>
<td>B.11</td>
<td>Error backpropagation multilayer networks</td>
<td>799</td>
</tr>
<tr>
<td>B.11.1</td>
<td>Introduction</td>
<td>799</td>
</tr>
<tr>
<td>B.11.2</td>
<td>Architecture and algorithm</td>
<td>799</td>
</tr>
<tr>
<td>B.11.3</td>
<td>Properties of multilayer networks trained by error backpropagation</td>
<td>802</td>
</tr>
<tr>
<td>B.12</td>
<td>Biologically plausible networks vs backpropagation</td>
<td>803</td>
</tr>
<tr>
<td>B.13</td>
<td>Convolution networks</td>
<td>804</td>
</tr>
<tr>
<td>B.14</td>
<td>Contrastive Hebbian learning: the Boltzmann machine</td>
<td>806</td>
</tr>
<tr>
<td>B.15</td>
<td>Deep Belief Networks</td>
<td>807</td>
</tr>
<tr>
<td>B.16</td>
<td>Reinforcement learning</td>
<td>807</td>
</tr>
<tr>
<td>B.16.1</td>
<td>Associative reward–penalty algorithm of Barto and Sutton</td>
<td>808</td>
</tr>
<tr>
<td>B.16.2</td>
<td>Reward prediction error or delta rule learning, and classical conditioning</td>
<td>810</td>
</tr>
<tr>
<td>B.16.3</td>
<td>Temporal Difference (TD) learning</td>
<td>811</td>
</tr>
<tr>
<td>B.17</td>
<td>Highlights</td>
<td>814</td>
</tr>
<tr>
<td>C</td>
<td>Information theory, and neuronal encoding</td>
<td>815</td>
</tr>
<tr>
<td>C.1</td>
<td>Information theory</td>
<td>816</td>
</tr>
<tr>
<td>C.1.1</td>
<td>The information conveyed by definite statements</td>
<td>816</td>
</tr>
<tr>
<td>C.1.2</td>
<td>Information conveyed by probabilistic statements</td>
<td>817</td>
</tr>
<tr>
<td>C.1.3</td>
<td>Information sources, information channels, and information measures</td>
<td>818</td>
</tr>
<tr>
<td>C.1.4</td>
<td>The information carried by a neuronal response and its averages</td>
<td>819</td>
</tr>
<tr>
<td>C.1.5</td>
<td>The information conveyed by continuous variables</td>
<td>822</td>
</tr>
<tr>
<td>C.2</td>
<td>The information carried by neuronal responses</td>
<td>824</td>
</tr>
<tr>
<td>C.2.1</td>
<td>The limited sampling problem</td>
<td>824</td>
</tr>
<tr>
<td>C.2.2</td>
<td>Correction procedures for limited sampling</td>
<td>825</td>
</tr>
<tr>
<td>C.2.3</td>
<td>The information from multiple cells: decoding procedures</td>
<td>826</td>
</tr>
<tr>
<td>C.2.4</td>
<td>Information in the correlations between cells: a decoding approach</td>
<td>830</td>
</tr>
<tr>
<td>C.2.5</td>
<td>Information in the correlations between cells: second derivative approach</td>
<td>835</td>
</tr>
<tr>
<td>C.3</td>
<td>Information theory results</td>
<td>838</td>
</tr>
<tr>
<td>C.3.1</td>
<td>The sparseness of the distributed encoding used by the brain</td>
<td>839</td>
</tr>
<tr>
<td>C.3.2</td>
<td>The information from single neurons</td>
<td>850</td>
</tr>
<tr>
<td>C.3.3</td>
<td>The information from single neurons: temporal codes versus rate codes</td>
<td>852</td>
</tr>
<tr>
<td>C.3.4</td>
<td>The information from single neurons: the speed of information transfer</td>
<td>854</td>
</tr>
<tr>
<td>C.3.5</td>
<td>The information from multiple cells: independence versus redundancy</td>
<td>866</td>
</tr>
<tr>
<td>C.3.6</td>
<td>Should one neuron be as discriminative as the whole organism?</td>
<td>870</td>
</tr>
<tr>
<td>C.3.7</td>
<td>The information from multiple cells: the effects of cross-correlations</td>
<td>871</td>
</tr>
<tr>
<td>C.3.8</td>
<td>Conclusions on cortical neuronal encoding</td>
<td>875</td>
</tr>
<tr>
<td>C.4</td>
<td>Information theory terms – a short glossary</td>
<td>879</td>
</tr>
<tr>
<td>C.5</td>
<td>Highlights</td>
<td>880</td>
</tr>
<tr>
<td>D</td>
<td>Simulation software for neuronal network models</td>
<td>881</td>
</tr>
<tr>
<td>D.1</td>
<td>Introduction</td>
<td>881</td>
</tr>
<tr>
<td>D.2</td>
<td>Autoassociation or attractor networks</td>
<td>881</td>
</tr>
<tr>
<td>D.2.1</td>
<td>Running the simulation</td>
<td>881</td>
</tr>
<tr>
<td>D.2.2</td>
<td>Exercises</td>
<td>883</td>
</tr>
<tr>
<td>D.3</td>
<td>Pattern association networks</td>
<td>884</td>
</tr>
</tbody>
</table>
Appendix 3 Information theory, and neuronal encoding

In order to understand the operation of memory and perceptual systems in the brain, it is necessary to know how information is encoded by neurons and populations of neurons. The concepts and results found are essential for understanding cortical function, are introduced in Chapter 8, and are considered in this Appendix and by Rolls and Treves (2011).

We have seen that one parameter that influences the number of memories that can be stored in an associative memory is the sparseness of the representation, and it is therefore important to be able to quantify the sparseness of the representations.

We have also seen that the properties of an associative memory system depend on whether the representation is distributed or local (grandmother cell like), and it is important to be able to assess this quantitatively for neuronal representations.

It is also necessary to know how the information is encoded in order to understand how memory systems operate. Is the information that must be stored and retrieved present in the firing rates (the number of spikes in a fixed time), or is it present in synchronized firing of subsets of neurons? This has implications for how each stage of processing would need to operate. If the information is present in the firing rates, how much information is available from the spiking activity in a short period, of for example 20 or 50 ms? For each stage of cortical processing to operate quickly (in for example 20 ms), it is necessary for each stage to be able to read the code being provided by the previous cortical area within this order of time. Thus understanding the neural code is fundamental to understanding how each stage of processing works in the brain, and for understanding the speed of processing at each stage.

To treat all these questions quantitatively, we need quantitative ways of measuring sparseness, and also ways of measuring the information available from the spiking activity of single neurons and populations of neurons, and these are the topics addressed in this Appendix, together with some of the main results obtained, which provide answers to these questions.

Because single neurons are the computing elements of the brain and send the results of their processing by spiking activity to other neurons, we can understand brain processing by understanding what is encoded by the neuronal firing at each stage of the brain (e.g. each cortical area), and determining how what is encoded changes from stage to stage. Each neuron responds differently to a set of stimuli (with each neuron tuned differently to the members of the set of stimuli), and it is this that allows different stimuli to be represented. We can only address the richness of the representation therefore by understanding the differences in the responses of different neurons, and the impact that this has on the amount of information that is encoded. These issues can only be adequately and directly addressed at the level of the activity of single neurons and of populations of single neurons, and understanding at this neuronal level (rather than at the level of thousands or millions of neurons as revealed by functional neuroimaging) is essential for understanding brain computation.

Information theory provides the means for quantifying how much neurons communicate to other neurons, and thus provides a quantitative approach to fundamental questions about information processing in the brain. To investigate what in neuronal activity carries information, one must compare the amounts of information carried by different codes, that is different
descriptions of the same activity, to provide the answer. To investigate the speed of information transmission, one must define and measure information rates from neuronal responses. To investigate to what extent the information provided by different cells is redundant or instead independent, again one must measure amounts of information in order to provide quantitative evidence. To compare the information carried by the number of spikes, by the timing of the spikes within the response of a single neuron, and by the relative time of firing of different neurons reflecting for example stimulus-dependent neuronal synchronization, information theory again provides a quantitative and well-founded basis for the necessary comparisons. To compare the information carried by a single neuron or a group of neurons with that reflected in the behaviour of the human or animal, one must again use information theory, as it provides a single measure which can be applied to the measurement of the performance of all these different cases. In all these situations, there is no quantitative and well-founded alternative to information theory.

This Appendix briefly introduces the fundamental elements of information theory in Section C.1. A more complete treatment can be found in many books on the subject (e.g. Abramson (1963), Hamming (1990), and Cover and Thomas (1991)), including also Rieke, Warland, de Ruyter van Steveninck and Bialek (1997) which is specifically about information transmitted by neuronal firing. Section C.2 discusses the extraction of information measures from neuronal activity, in particular in experiments with mammals, in which the central issue is how to obtain accurate measures in conditions of limited sampling, that is where the numbers of trials of neuronal data that can be obtained are usually limited by the available recording time. Section C.3 summarizes some of the main results obtained so far on neuronal encoding. The essential terminology is summarized in a Glossary at the end of this Appendix in Section C.4. The approach taken in this Appendix is based on and updated from that provided by Rolls and Treves (1998), Rolls (2008d), and Rolls and Treves (2011).

**C.1 Information theory and its use in the analysis of formal models**

Although information theory was a surprisingly late starter as a mathematical discipline, having been developed and formalized by C. Shannon (1948), the intuitive notion of information is immediate to us. It is also very easy to understand why we use logarithms in order to quantify this intuitive notion, of how much we know about something, and why the resulting quantity is always defined in relative rather than absolute terms. An introduction to information theory is provided next, with a more formal summary given in Section C.1.3.

**C.1.1 The information conveyed by definite statements**

Suppose somebody, who did not know, is told that Reading is a town west of London. How much information is he given? Well, that depends. He may have known it was a town in England, but not whether it was east or west of London; in which case the new information amounts to the fact that of two a priori (i.e. initial) possibilities (E or W), one holds (W). It is also possible to interpret the statement in the more precise sense, that Reading is west of London, rather than east, north or south, i.e. one out of four possibilities; or else, west rather that north-west, north, etc. Clearly, the larger the number \( k \) of a priori possibilities, the more one is actually told, and a measure of information must take this into account. Moreover, we would like independent pieces of information to just add together. For example, our person may also be told that Cambridge is, out of \( l \) possible directions, north of London. Provided nothing was known on the mutual location of Reading and Cambridge, there are now overall
Information theory

$k \times l$ a priori (initial) possibilities, only one of which remains a posteriori (after receiving the information). Given that the number of possibilities for independent events are multiplicative, but that we would like the measure of information to be additive, we use logarithms when we measure information, as logarithms have this property. We thus define the amount $I$ of information gained when we are informed in which of $k$ possible locations Reading is located as

$$I(k) = \log_2 k.$$  \hfill (C.1)

Then when we combine independent information, for example producing $k \times l$ possibilities from independent events with $k$ and $l$ possibilities respectively, we obtain

$$I(k \times l) = \log_2(k \times l) = \log_2 k + \log_2 l = I(k) + I(l).$$ \hfill (C.2)

Thus in our example, the information about Cambridge adds up to that about Reading. We choose to take logarithms in base 2 as a mere convention, so that the answer to a yes/no question provides one unit, or bit, of information. Here it is just for the sake of clarity that we used different symbols for the number of possible directions with respect to which Reading and Cambridge are localized; if both locations are specified for example in terms of E, SE, S, SW, W, NW, N, NE, then obviously $k = l = 8, I(k) = I(l) = 3$ bits, and $I(k \times l) = 6$ bits. An important point to note is that the resolution with which the direction is specified determines the amount of information provided, and that in this example, as in many situations arising when analysing neuronal codings, the resolution could be made progressively finer, with a corresponding increase in information proportional to the log of the number of possibilities.

C.1.2 The information conveyed by probabilistic statements

The situation becomes slightly less trivial, and closer to what happens among neurons, if information is conveyed in less certain terms. Suppose for example that our friend is told, instead, that Reading has odds of 9 to 1 to be west, rather than east, of London (considering now just two a priori possibilities). He is certainly given some information, albeit less than in the previous case. We might put it this way: out of 18 equiprobable a priori possibilities (9 west + 9 east), 8 (east) are eliminated, and 10 remain, yielding

$$I = \log_2(18/10) = \log_2(9/5),$$ \hfill (C.3)

as the amount of information given. It is simpler to write this in terms of probabilities

$$I = \log_2 P_{\text{posterior}}(W)/P_{\text{prior}}(W) = \log_2(9/10)/(1/2) = \log_2(9/5).$$ \hfill (C.4)

This is of course equivalent to saying that the amount of information given by an uncertain statement is equal to the amount given by the absolute statement

$$I = -\log_2 P_{\text{prior}}(W)$$ \hfill (C.5)

minus the amount of uncertainty remaining after the statement, $I = -\log_2 P_{\text{posterior}}(W)$. A successive clarification that Reading is indeed west of London carries

$$I' = \log_2((1)/(9/10))$$ \hfill (C.6)

bits of information, because 9 out of 10 are now the a priori odds, while a posteriori there is certainty, $P_{\text{posterior}}(W) = 1$. In total we would seem to have

$$I_{\text{TOTAL}} = I + I' = \log_2(9/5) + \log_2(10/9) = 1 \text{ bit}$$ \hfill (C.7)

as if the whole information had been provided at one time. This is strange, given that the two pieces of information are clearly not independent, and only independent information
should be additive. In fact, we have cheated a little. Before the clarification, there was still one residual possibility (out of 10) that the answer was ‘east’, and this must be taken into account by writing

\[ I = \text{posterior}(W) \log_2\frac{\text{posterior}(W)}{\text{prior}(W)} + \text{posterior}(E) \log_2\frac{\text{posterior}(E)}{\text{prior}(E)} \]

as the information contained in the first message. This little detour should serve to emphasize two aspects that are easy to forget when reasoning intuitively about information, and that in this example cancel each other. In general, when uncertainty remains, that is there is more than one possible \textit{a posteriori} state, one has to average information values for each state with the corresponding \textit{a posteriori} probability measure. In the specific example, the sum \( I + I' \) totals slightly more than 1 bit, and the amount in excess is precisely the information ‘wasted’ by providing correlated messages.

### C.1.3 Information sources, information channels, and information measures

In summary, the expression quantifying the information provided by a definite statement that event \( s \), which had an \textit{a priori} probability \( P(s) \), has occurred is

\[ I(s) = \log_2(1/P(s)) = - \log_2 P(s), \]

whereas if the statement is probabilistic, that is several \textit{a posteriori} probabilities remain non-zero, the correct expression involves summing over all possibilities with the corresponding probabilities:

\[ I = \sum_s \left[ \text{posterior}(s) \log_2\frac{\text{posterior}(s)}{\text{prior}(s)} \right]. \]

When considering a discrete set of mutually exclusive events, it is convenient to use the metaphor of a set of symbols comprising an alphabet \( S \). The occurrence of each event is then referred to as the emission of the corresponding symbol by an information source. The entropy of the source, \( H \), is the average amount of information per source symbol, where the average is taken across the alphabet, with the corresponding probabilities

\[ H(S) = - \sum_{s \in S} P(s) \log_2 P(s). \]

An information channel receives symbols \( s \) from an alphabet \( S \) and emits symbols \( s' \) from alphabet \( S' \). If the joint probability of the channel receiving \( s \) and emitting \( s' \) is given by the product

\[ P(s, s') = P(s)P(s') \]

for any pair \( s, s' \), then the input and output symbols are independent of each other, and the channel transmits zero information. Instead of joint probabilities, this can be expressed with conditional probabilities: the conditional probability of \( s' \) given \( s \) is written \( P(s'|s) \), and if the two variables are independent, it is just equal to the unconditional probability \( P(s') \).
and in particular if the channel does transmit information, the variables are not independent, and one can express their joint probability in two ways in terms of conditional probabilities

$$P(s, s') = P(s'|s)P(s) = P(s'|s)P(s'),$$

(C.13)

from which it is clear that

$$P(s'|s) = P(s') \frac{P(s'|s)}{P(s)}.$$  

(C.14)

which is called Bayes’ theorem (although when expressed as here in terms of probabilities it is strictly speaking an identity rather than a theorem). The information transmitted by the channel conditional to its having emitted symbol $s'$ (or specific transinformation, $I(s')$) is given by equation C.10, once the unconditional probability $P(s)$ is inserted as the prior, and the conditional probability $P(s'|s)$ as the posterior:

$$I(s') = \sum_s P(s | s') \log_2 \frac{P(s | s')}{P(s)}.$$  

(C.15)

Symmetrically, one can define the transinformation conditional to the channel having received symbol $s$

$$I(s) = \sum_{s'} P(s' | s) \log_2 \frac{P(s' | s)}{P(s')}.$$  

(C.16)

Finally, the average transinformation, or **mutual information**, can be expressed in fully symmetrical form

$$I = \sum_s P(s) \sum_{s'} P(s' | s) \log_2 \frac{P(s' | s)}{P(s')}$$

$$= \sum_{s, s'} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')}.$$  

(C.17)

The **mutual information** can also be expressed as the entropy of the source using alphabet $S$ minus the equivocation of $S$ with respect to the new alphabet $S'$ used by the channel, written

$$I = H(S) - H(S|S') \equiv H(S) - \sum_{s'} P(s')H(S|s').$$  

(C.18)

A channel is characterized, once the alphabets are given, by the set of conditional probabilities for the output symbols, $P(s'|s)$, whereas the unconditional probabilities of the input symbols $P(s)$ depend of course on the source from which the channel receives. Then, the **capacity** of the channel can be defined as the maximal mutual information across all possible sets of input probabilities $P(s)$. Thus, the information transmitted by a channel can range from zero to the lower of two independent upper bounds: the entropy of the source, and the capacity of the channel.

**C.1.4 The information carried by a neuronal response and its averages**

Considering the processing of information in the brain, we are often interested in the amount of information the response $r$ of a neuron, or of a population of neurons, carries about an event happening in the outside world, for example a stimulus $s$ shown to the animal. Once the inputs and outputs are conceived of as sets of symbols from two alphabets, the neuron(s)
may be regarded as an information channel. We may denote with $P(s)$ the \textit{a priori} probability that the particular stimulus $s$ out of a given set was shown, while the conditional probability $P(s|r)$ is the \textit{a posteriori} probability, that is updated by the knowledge of the response $r$. The response-specific transinformation

$$I(r) = \sum_s P(s|r) \log_2 \frac{P(s|r)}{P(s)}$$  \hspace{1cm} (C.19)$$

takes the extreme values of $I(r) = -\log_2 P(s(r))$ if $r$ unequivocally determines $s(r)$ (that is, $P(s|r)$ equals 1 for that one stimulus and 0 for all others); and $I(r) = \sum_s P(s) \log_2 (P(s)/P(s)) = 0$ if there is no relation between $s$ and $r$, that is they are independent, so that the response tells us nothing new about the stimulus and thus $P(s|r) = P(s)$.

This is the information conveyed by each particular response. One is usually interested in further averaging this quantity over all possible responses $r$,

$$\langle I \rangle = \sum_r P(r) \left[ \sum_s P(s|r) \log_2 \frac{P(s|r)}{P(s)} \right].$$  \hspace{1cm} (C.20)$$

The angular brackets $\langle \rangle$ are used here to emphasize the averaging operation, in this case over responses. Denoting with $P(s,r)$ the \textit{joint probability} of the pair of events $s$ and $r$, and using Bayes’ theorem, this reduces to the symmetric form (equation C.18) for the \textbf{mutual information} $I(S;R)$

$$\langle I \rangle = \sum_{s,r} P(s,r) \log_2 \frac{P(s,r)}{P(s)P(r)}$$  \hspace{1cm} (C.21)$$

which emphasizes that responses tell us about stimuli just as much as stimuli tell us about responses. This is, of course, a general feature, independent of the two variables being in this instance stimuli and neuronal responses. In fact, what is of interest, besides the mutual information of equations C.20 and C.21, is often the information specifically conveyed about each stimulus,

$$I(s) = \sum_r P(r|s) \log_2 \frac{P(r|s)}{P(r)}$$  \hspace{1cm} (C.22)$$

which is a direct quantification of the variability in the responses elicited by that stimulus, compared to the overall variability. Since $P(r)$ is the probability distribution of responses averaged across stimuli, it is again evident that the stimulus-specific information measure of equation C.22 depends not only on the stimulus $s$, but also on all other stimuli used. Likewise, the mutual information measure, despite being of an average nature, is dependent on what set of stimuli has been used in the average. This emphasizes again the relative nature of all information measures. More specifically, it underscores the relevance of using, while measuring the information conveyed by a given neuronal population, stimuli that are either representative of real-life stimulus statistics, or of particular interest for the properties of the population being examined\textsuperscript{41}.

\textsuperscript{41}The quantity $I(s,R)$, which is what is shown in equation C.22 and where $R$ draws attention to the fact that this quantity is calculated across the full set of responses $R$, has also been called the stimulus-specific surprise, see DeWeese and Meister (1999). Its average across stimuli is the mutual information $I(S;R)$. 
To make these notions clearer, we can consider a specific example in which the response of a neuron to the presentation of, say, one of four visual stimuli (A, B, C, D) is recorded for 10 ms, during which the neuron emits either 0, 1, or 2 spikes, but no more. Imagine that the neuron tends to respond more vigorously to visual stimulus B, less to C, even less to A, and never to D, as described by the table of conditional probabilities $P(r|s)$ shown in Table C.1.

<table>
<thead>
<tr>
<th>s = A</th>
<th>r = 0</th>
<th>0.6</th>
<th>0.4</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>s = B</td>
<td>r = 0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.8</td>
</tr>
<tr>
<td>s = C</td>
<td>r = 0</td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>s = D</td>
<td>r = 0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Then, if different visual stimuli are presented with equal probability, the table of joint probabilities $P(s;r)$ will be as shown in Table C.2.

<table>
<thead>
<tr>
<th>s = A</th>
<th>r = 0</th>
<th>0.15</th>
<th>0.1</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>s = B</td>
<td>r = 0</td>
<td>0.0</td>
<td>0.05</td>
<td>0.2</td>
</tr>
<tr>
<td>s = C</td>
<td>r = 0</td>
<td>0.1</td>
<td>0.125</td>
<td>0.025</td>
</tr>
<tr>
<td>s = D</td>
<td>r = 0</td>
<td>0.25</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

From these two tables one can compute various information measures by directly applying the definitions above. Since visual stimuli are presented with equal probability, $P(s) = 1/4$, the entropy of the stimulus set, which corresponds to the maximum amount of information any transmission channel, no matter how efficient, could convey on the identity of the stimuli, is

$$H_s = -\sum_s [P(s) \log_2 P(s)] = -4[(1/4) \log_2(1/4)] = \log_2 4 = 2 \text{ bits}.$$ 

There is a more stringent upper bound on the mutual information that this cell’s responses convey on the stimuli, however, and this second bound is the channel capacity $T$ of the cell. Calculating this quantity involves maximizing the mutual information across prior visual stimulus probabilities, and it is a bit complicated to do, in general. In our particular case the maximum information is obtained when only stimuli B and D are presented, each with probability 0.5. The resulting capacity is $T = 1 \text{ bit}$. We can easily calculate, in general, the entropy of the responses. This is not an upper bound characterizing the source, like the entropy of the stimuli, nor an upper bound characterizing the channel, like the capacity, but simply a bound on the mutual information for this specific combination of source (with its related visual stimulus probabilities) and channel (with its conditional probabilities). Since only three response levels are possible within the short recording window, and they occur with uneven probability, their entropy is considerably lower than $H_s$, at

$$H_r = -\sum_r P(r) \log_2 P(r) = -[0.25 \log_2(0.25) + 0.25 \log_2(0.25) + 0.5 \log_2(0.5)] = 1.496 \text{ bits}.$$ 

The actual average information $I$ that the responses transmit about the stimuli, which is a measure of the correlation in the variability of stimuli and responses, does not exceed the absolute variability of either stimuli (as quantified by the first bound) or responses
(as quantified by the last bound), nor the capacity of the channel. An explicit calculation using
the joint probabilities of the second table in expression C.21 yields \( I = 0.733 \) bits. This is of
course only the average value, averaged both across stimuli and across responses.

The information conveyed by a particular response can be larger. For example, when the
cell emits two spikes it indicates with a relatively large probability stimulus B, and this is
reflected in the fact that it then transmits, according to expression C.19, \( I(r = 2) = 1.497 \)
bits, more than double the average value.

Similarly, the amount of information conveyed about each individual visual stimulus varies
with the stimulus, depending on the extent to which it tends to elicit a differential response.
Thus, expression C.22 yields that only \( I(s = C) = 0.185 \) bits are conveyed on average about
stimulus C, which tends to elicit responses with similar statistics to the average statistics
across stimuli, and are therefore not easily interpretable. On the other hand, exactly 1 bit of
information is conveyed about stimulus D, since this stimulus never elicits any response, and
when the neuron emits no spike there is a probability of 1/2 that the stimulus was stimulus D.

### C.1.5 The information conveyed by continuous variables

A general feature, relevant also to the case of neuronal information, is that if, among a
continuum of \textit{a priori} possibilities, only one, or a discrete number, remains \textit{a posteriori},
the information is strictly infinite. This would be the case if one were told, for example, that
Reading is exactly 10° west, 1° north of London. The \textit{a priori} probability of precisely this set of
coordinates among the continuum of possible ones is zero, and then the information diverges
to infinity. The problem is only theoretical, because in fact, with continuous distributions,
there are always one or several factors that limit the resolution in the \textit{a posteriori} knowledge,
rendering the information finite. Moreover, when considering the mutual information in the
conjoint probability of occurrence of two sets, e.g. stimuli and responses, it suffices that
at least one of the sets is discrete to make matters easy, that is, finite. Nevertheless, the
identification and appropriate consideration of these resolution-limiting factors in practical
cases may require careful analysis.

#### C.1.5.1 Example: the information retrieved from an autoassociative memory

One example is the evaluation of the information that can be retrieved from an autoassociative
memory. Such a memory stores a number of firing patterns, each one of which can be
considered, as in Appendix B, as a vector \( r^\mu \) with components the firing rates \( \{ r^\mu_i \} \), where
the subscript \( i \) indexes the neuron (and the superscript \( \mu \) indexes the pattern). In retrieving
pattern \( \mu \), the network in fact produces a distinct firing pattern, denoted for example simply as
\( r \). The quality of retrieval, or the similarity between \( r^\mu \) and \( r \), can be measured by the average
mutual information

\[
< I(r^\mu, r) > = \sum_{r^\mu, r} P(r^\mu, r) \log_2 \frac{P(r^\mu, r)}{P(r^\mu)P(r)}
\]

(C.23)

In this formula the ‘approximately equal’ sign \( \approx \) marks a simplification that is not necessarily
a reasonable approximation. If the simplification is valid, it means that in order to extract an
information measure, one need not compare whole vectors (the entire firing patterns) with
each other, and may instead compare the firing rates of individual cells at storage and retrieval,
and sum the resulting single-cell information values. The validity of the simplification is a
matter that will be discussed later and that has to be verified, in the end, experimentally, but
for the purposes of the present discussion we can focus on the single-cell terms. If either
r, or r^μ, has a continuous distribution of values, as it will if it represents not the number of spikes emitted in a fixed window, but more generally the firing rate of neuron i computed by convolving the firing train with a smoothing kernel, then one has to deal with probability densities, which we denote as p(r)dr, rather than the usual probabilities P(r). Substituting p(r)dr for P(r) and p(r^μ, r)dr^μdr for P(r^μ, r), one can write for each single-cell contribution (omitting the cell index i)

$$< I(r^μ, r) >_i = \int dr^μ dr \frac{p(r^μ, r) \log_2 \frac{p(r^μ, r)}{p(r^μ)p(r)}}{p(r^μ)p(r)} \tag{C.24}$$

and we see that the differentials dr^μdr cancel out between numerator and denominator inside the logarithm, rendering the quantity well defined and finite. If, however, r^μ were to exactly determine r, one would have

$$p(r^μ, r)dr^μdr = p(r^μ)\delta(r - r(r^μ))dr^μdr \tag{C.25}$$

and, by losing one differential on the way, the mutual information would become infinite. It is therefore important to consider what prevents r^μ from fully determining r in the case at hand – in other words, to consider the sources of noise in the system. In an autoassociative memory storing an extensive number of patterns (see Appendix A4 of Rolls and Treves (1998)), one source of noise always present is the interference effect due to the concurrent storage of all other patterns. Even neglecting other sources of noise, this produces a finite resolution width ρ, which allows one to write an expression of the type p(r|r^μ)dr = exp(-(r - r(r^μ))^2/2ρ^2)dr which ensures that the information is finite as long as the resolution ρ is larger than zero.

One further point that should be noted, in connection with estimating the information retrievable from an autoassociative memory, is that the mutual information between the current distribution of firing rates and that of the stored pattern does not coincide with the information gain provided by the memory device. Even when firing rates, or spike counts, are all that matter in terms of information carriers, as in the networks considered in this book, one more term should be taken into account in evaluating the information gain. This term, to be subtracted, is the information contained in the external input that elicits the retrieval. This may vary a lot from the retrieval of one particular memory to the next, but of course an efficient memory device is one that is able, when needed, to retrieve much more information than it requires to be present in the inputs, that is, a device that produces a large information gain.

Finally, one should appreciate the conceptual difference between the information a firing pattern carries about another one (that is, about the pattern stored), as considered above, and two different notions: (a) the information produced by the network in selecting the correct memory pattern and (b) the information a firing pattern carries about something in the outside world. Quantity (a), the information intrinsic to selecting the memory pattern, is ill defined when analysing a real system, but is a well-defined and particularly simple notion when considering a formal model. If p patterns are stored with equal strength, and the selection is errorless, this amounts to log_2 p bits of information, a quantity often, but not always, small compared with the information in the pattern itself. Quantity (b), the information conveyed about some outside correlate, is not defined when considering a formal model that does not include an explicit account of what the firing of each cell represents, but is well defined and measurable from the recorded activity of real cells. It is the quantity considered in the numerical example with the four visual stimuli, and it can be generalized to the information carried by the activity of several cells in a network, and specialized to the case that the network operates as an associative memory. One may note, in this case, that the capacity to retrieve memories with high fidelity, or high information content, is only useful to the extent that the
representation to be retrieved carries that amount of information about something relevant – or, in other words, that it is pointless to store and retrieve with great care largely meaningless messages. This type of argument has been used to discuss the role of the mossy fibres in the operation of the CA3 network in the hippocampus (Treves and Rolls 1992b, Rolls and Treves 1998).

C.2 Estimating the information carried by neuronal responses

C.2.1 The limited sampling problem

We now discuss in more detail the application of these general notions to the information transmitted by neurons. Suppose, to be concrete, that an animal has been presented with stimuli drawn from a discrete set, and that the responses of a set of $C$ cells have been recorded following the presentation of each stimulus. We may choose any quantity or set of quantities to characterize the responses; for example let us assume that we consider the firing rate of each cell, $r_i$, calculated by convolving the spike response with an appropriate smoothing kernel. The response space is then $C$ times the continuous set of all positive real numbers, $(\mathbb{R}/2)^C$. We want to evaluate the average information carried by such responses about which stimulus was shown. In principle, it is straightforward to apply the above formulas, e.g. in the form

$$<I(s; r) > = \sum_s P(s) \int \Pi_i dr_i p(r_i|s) \log_2 \frac{p(r_i|s)}{p(r)} \quad (C.26)$$

where it is important to note that $p(r)$ and $p(r_i|s)$ are now probability densities defined over the high-dimensional vector space of multi-cell responses. The product sign $\Pi$ signifies that this whole vector space has to be integrated over, along all its dimensions. $p(r)$ can be calculated as $\sum_s p(r_i|s)P(s)$, and therefore, in principle, all one has to do is to estimate, from the data, the conditional probability densities $p(r_i|s)$ – the distributions of responses following each stimulus. In practice, however, in contrast to what happens with formal models, in which there is usually no problem in calculating the exact probability densities, real data come in limited amounts, and thus sample only sparsely the vast response space. This limits the accuracy with which, from the experimental frequency of each possible response, we can estimate its probability, in turn seriously impairing our ability to estimate $<I>$ correctly. We refer to this as the limited sampling problem. This is a purely technical problem that arises, typically when recording from mammals, because of external constraints on the duration or number of repetitions of a given set of stimulus conditions. With computer simulation experiments, and also with recordings from, for example, insects, sufficient data can usually be obtained that straightforward estimates of information are accurate enough (Strong, Koberle, de Ruyter van Steveninck and Bialek 1998, Golomb, Kleinfeld, Reid, Shapley and Shraiman 1994). The problem is, however, so serious in connection with recordings from monkeys and rats in which limited numbers of trials are usually available for neuronal data, that it is worthwhile to discuss it, in order to appreciate the scope and limits of applying information theory to neuronal processing.

In particular, if the responses are continuous quantities, the probability of observing exactly the same response twice is infinitesimal. In the absence of further manipulation, this would imply that each stimulus generates its own set of unique responses, therefore any response that has actually occurred could be associated unequivocally with one stimulus, and the mutual information would always equal the entropy of the stimulus set. This absurdity
shows that in order to estimate probability densities from experimental frequencies, one has to resort to some regularizing manipulation, such as smoothing the point-like response values by convolution with suitable kernels, or binning them into a finite number of discrete bins.

C.2.1.1 Smoothing or binning neuronal response data

The issue is how to estimate the underlying probability distributions of neuronal responses to a set of stimuli from only a limited number of trials of data (e.g. 10–30) for each stimulus. Several strategies are possible. One is to discretize the response space into bins, and estimate the probability density as the histogram of the fraction of trials falling into each bin. If the bins are too narrow, almost every response is in a different bin, and the estimated information will be overestimated. Even if the bin width is increased to match the standard deviation of each underlying distribution, the information may still be overestimated. Alternatively, one may try to ‘smooth’ the data by convolving each response with a Gaussian with a width set to the standard deviation measured for each stimulus. Setting the standard deviation to this value may actually lead to an underestimation of the amount of information available, due to oversmoothing. Another possibility is to make a bold assumption as to what the general shape of the underlying densities should be, for example a Gaussian. This may produce closer estimates. Methods for regularizing the data are discussed further by Rolls and Treves (1998) in their Appendix A2, where a numerical example is given.

C.2.1.2 The effects of limited sampling

The crux of the problem is that, whatever procedure one adopts, limited sampling tends to produce distortions in the estimated probability densities. The resulting mutual information estimates are intrinsically biased. The bias, or average error of the estimate, is upward if the raw data have not been regularized much, and is downward if the regularization procedure chosen has been heavier. The bias can be, if the available trials are few, much larger than the true information values themselves. This is intuitive, as fluctuations due to the finite number of trials available would tend, on average, to either produce or emphasize differences among the distributions corresponding to different stimuli, differences that are preserved if the regularization is ‘light’, and that are interpreted in the calculation as carrying genuine information. This is illustrated with a quantitative example by Rolls and Treves (1998) in their Appendix A2.

Choosing the right amount of regularization, or the best regularizing procedure, is not possible a priori. Hertz, Kjaer, Eskander and Richmond (1992) have proposed the interesting procedure of using an artificial neural network to regularize the raw responses. The network can be trained on part of the data using backpropagation, and then used on the remaining part to produce what is in effect a clever data-driven regularization of the responses. This procedure is, however, rather computer intensive and not very safe, as shown by some self-evident inconsistency in the results (Heller, Hertz, Kjaer and Richmond 1995). Obviously, the best way to deal with the limited sampling problem is to try and use as many trials as possible. The improvement is slow, however, and generating as many trials as would be required for a reasonably unbiased estimate is often, in practice, impossible.

C.2.2 Correction procedures for limited sampling

The above point, that data drawn from a single distribution, when artificially paired, at random, to different stimulus labels, results in ‘spurious’ amounts of apparent information, suggests a simple way of checking the reliability of estimates produced from real data (Optican, Gawne, Richmond and Joseph 1991). One can disregard the true stimulus associated with each response, and generate a randomly reshuffled pairing of stimuli and responses, which should
therefore, being not linked by any underlying relationship, carry no mutual information about each other. Calculating, with some procedure of choice, the spurious information obtained in this way, and comparing with the information value estimated with the same procedure for the real pairing, one can get a feeling for how far the procedure goes into eliminating the apparent information due to limited sampling. Although this spurious information, \( I_s \), is only indicative of the amount of bias affecting the original estimate, a simple heuristic trick (called ‘bootstrap’\(^{42}\)) is to subtract the spurious from the original value, to obtain a somewhat ‘corrected’ estimate. This procedure can result in quite accurate estimates (see Rolls and Treves (1998), Tovee, Rolls, Treves and Bellis (1993))\(^{43}\).

A different correction procedure (called ‘jackknife’) is based on the assumption that the bias is proportional to \( 1/N \), where \( N \) is the number of responses (data points) used in the estimation. One computes, beside the original estimate \( < I_N > \), \( N \) auxiliary estimates \( < I_{N-1} >_k \), by taking out from the data set response \( k \), where \( k \) runs across the data set from 1 to \( N \). The corrected estimate

\[
< I > = N < I_N > - (1/N) \sum_k (N - 1) < I_{N-1} >_k
\]

is free from bias (to leading order in \( 1/N \)), if the proportionality factor is more or less the same in the original and auxiliary estimates. This procedure is very time-consuming, and it suffers from the same imprecision of any algorithm that tries to determine a quantity as the result of the subtraction of two large and nearly equal terms; in this case the terms have been made large on purpose, by multiplying them by \( N \) and \( N - 1 \).

A more fundamental approach (Miller 1955) is to derive an analytical expression for the bias (or, more precisely, for its leading terms in an expansion in \( 1/N \), the inverse of the sample size). This allows the estimation of the bias from the data itself, and its subsequent subtraction, as discussed in Treves and Panzeri (1995) and Panzeri and Treves (1996). Such a procedure produces satisfactory results, thereby lowering the size of the sample required for a given accuracy in the estimate by about an order of magnitude (Golomb, Hertz, Panzeri, Treves and Richmond 1997). However, it does not, in itself, make possible measures of the information contained in very complex responses with few trials. As a rule of thumb, the number of trials per stimulus required for a reasonable estimate of information, once the subtractive correction is applied, is of the order of the effectively independent (and utilized) bins in which the response space can be partitioned (Panzeri and Treves 1996). This correction procedure is the one that we use standardly (Rolls, Treves, Tovee and Panzeri 1997d, Rolls, Critchley and Treves 1996a, Rolls, Treves, Robertson, Georges-François and Panzeri 1998b, Booth and Rolls 1998, Rolls, Tovee and Panzeri 1999b, Rolls, Franco, Aggelopoulos and Jerez 2006b).

### C.2.3 The information from multiple cells: decoding procedures

The bias of information measures grows with the dimensionality of the response space, and for all practical purposes the limit on the number of dimensions that can lead to reasonably accurate direct measures, even when applying a correction procedure, is quite low, two to three. This implies, in particular, that it is not possible to apply equation C.26 to extract the information content in the responses of several cells (more than two to three) recorded

\(^{42}\)In technical usage bootstrap procedures utilize random pairings of responses with stimuli with replacement, while shuffling procedures utilize random pairings of responses with stimuli without replacement.

\(^{43}\)Subtracting the ‘square’ of the spurious fraction of information estimated by this bootstrap procedure as used by Optican, Gawne, Richmond and Joseph (1991) is unfounded and does not work correctly (see Rolls and Treves (1998) and Tovee, Rolls, Treves and Bellis (1993)).
The information carried by neuronal responses

simultaneously. One way to address the problem is then to apply some strong form of regularization to the multiple cell responses. Smoothing has already been mentioned as a form of regularization that can be tuned from very soft to very strong, and that preserves the structure of the response space. Binning is another form, which changes the nature of the responses from continuous to discrete, but otherwise preserves their general structure, and which can also be tuned from soft to strong. Other forms of regularization involve much more radical transformations, or changes of variables.

Of particular interest for information estimates is a change of variables that transforms the response space into the stimulus set, by applying an algorithm that derives a predicted stimulus from the response vector, i.e. the firing rates of all the cells, on each trial. Applying such an algorithm is called decoding. Of course, the predicted stimulus is not necessarily the same as the actual one. Therefore the term decoding should not be taken to imply that the algorithm works successfully, each time identifying the actual stimulus. The predicted stimulus is simply a function of the response, as determined by the algorithm considered. Just as with any regularizing transform, it is possible to compute the mutual information between actual stimuli $s$ and predicted stimuli $s'$, instead of the original one between stimuli $s$ and responses $r$. Since information about (real) stimuli can only be lost and not be created by the transform, the information measured in this way is bound to be lower in value than the real information in the responses. If the decoding algorithm is efficient, it manages to preserve nearly all the information contained in the raw responses, while if it is poor, it loses a large portion of it. If the responses themselves provided all the information about stimuli, and the decoding is optimal, then predicted stimuli coincide with the actual stimuli, and the information extracted equals the entropy of the stimulus set.

The procedure for extracting information values after applying a decoding algorithm is indicated in Fig. C.1 (in which $s'$ is $s'$). The underlying idea indicated in Fig. C.1 is that if we know the average firing rate of each cell in a population to each stimulus, then on any single trial we can guess (or decode) the stimulus that was present by taking into account the responses of all the cells. The decoded stimulus is $s'$, and the actual stimulus that was shown is $s$. What we wish to know is how the percentage correct, or better still the information, based on the evidence from any single trial about which stimulus was shown, increases as the number of cells in the population sampled increases. We can expect that the more cells there are in the sample, the more accurate the estimate of the stimulus is likely to be. If the encoding was local, the number of stimuli encoded by a population of neurons would be expected to rise approximately linearly with the number of neurons in the population. In contrast, with distributed encoding, provided that the neuronal responses are sufficiently independent, and are sufficiently reliable (not too noisy), information from the ensemble would be expected to rise linearly with the number of cells in the ensemble, and (as information is a log measure) the number of stimuli encodable by the population of neurons might be expected to rise exponentially as the number of neurons in the sample of the population was increased.

| Table C.3 Decoding. $s'$ is the decoded stimulus, i.e. that predicted from the neuronal responses $r$. |
| $s \Rightarrow r \Rightarrow s'$ |
| $I(s, r)$ ($I(s, s')$) |

The procedure is schematized in Table C.3 where the double arrow indicates the transformation from stimuli to responses operated by the nervous system, while the single arrow indicates the further transformation operated by the decoding procedure. $I(s, s')$ is the mutual
How well can one predict which stimulus was shown on a single trial from the mean responses of different neurons to each stimulus?

<table>
<thead>
<tr>
<th>Cell 1</th>
<th>Cell 2</th>
<th>Cell 3</th>
<th>Cell 4</th>
<th>Cell 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S?</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Fig. C.1** This diagram shows the average response for each of several cells (Cell 1, etc.) to each of several stimuli (S1, etc.). The change of firing rate from the spontaneous rate is indicated by the vertical line above or below the horizontal line, which represents the spontaneous rate. We can imagine guessing or predicting from such a table the predicted stimulus S? (i.e. $s'$) that was present on any one trial.

Information between the actual stimuli $s$ and the stimuli $s'$ that are predicted to have been shown based on the decoded responses.

A slightly more complex variant of this procedure is a decoding step that extracts from the response on each trial not a single predicted stimulus, but rather probabilities that each of the possible stimuli was the actual one. The joint probabilities of actual and posited stimuli can be averaged across trials, and information computed from the resulting probability matrix $(S \times S)$. Computing information in this way takes into account the relative uncertainty in assigning a predicted stimulus to each trial, an uncertainty that is instead not considered by the previous procedure based solely on the identification of the maximally likely stimulus (Treves 1997). Maximum likelihood information values $I_{ml}$ based on a single stimulus tend therefore to be higher than probability information values $I_p$ based on the whole set of stimuli, although in very specific situations the reverse could also be true.

The same correction procedures for limited sampling can be applied to information values computed after a decoding step. Values obtained from maximum likelihood decoding, $I_{ml}$, suffer from limited sampling more than those obtained from probability decoding, $I_p$, since each trial contributes a whole “brick” of weight $1/N$ ($N$ being the total number of trials), whereas with probabilities each brick is shared among several slots of the $(S \times S)$ probability
The information carried by neuronal responses

The neural network procedure devised by Hertz, Kjaer, Eskander and Richmond (1992) can in fact be thought of as a decoding procedure based on probabilities, which deals with limited sampling not by applying a correction but rather by strongly regularizing the original responses.

When decoding is used, the rule of thumb becomes that the minimal number of trials per stimulus required for accurate information measures is roughly equal to the size of the stimulus set, if the subtractive correction is applied (Panzeri and Treves 1996). This correction procedure is applied as standard in our multiple cell information analyses that use decoding (Rolls, Treves and Tovee 1997b, Booth and Rolls 1998, Rolls, Treves, Robertson, Georges-François and Panzeri 1998b, Franco, Rolls, Aggelopoulos and Treves 2004, Aggelopoulos, Franco and Rolls 2005, Rolls, Franco, Aggelopoulos and Jerez 2006b).

C.2.3.1 Decoding algorithms

Any transformation from the response space to the stimulus set could be used in decoding, but of particular interest are the transformations that either approach optimality, so as to minimize information loss and hence the effect of decoding, or else are implementable by mechanisms that could conceivably be operating in the real system, so as to extract information values that could be extracted by the system itself.

The optimal transformation is in theory well-defined: one should estimate from the data the conditional probabilities \( P(r|s) \), and use Bayes’ rule to convert them into the conditional probabilities \( P(s'|r) \). Having these for any value of \( r \), one could use them to estimate \( I_p \), and, after selecting for each particular real response the stimulus with the highest conditional probability, to estimate \( I_{ml} \). To avoid biasing the estimation of conditional probabilities, the responses used in estimating \( P(r|s) \) should not include the particular response for which \( P(s'|r) \) is going to be derived (jack-knife cross-validation). In practice, however, the estimation of \( P(r|s) \) in usable form involves the fitting of some simple function to the responses. This need for fitting, together with the approximations implied in the estimation of the various quantities, prevents us from defining the really optimal decoding, and leaves us with various algorithms, depending essentially on the fitting function used, which are hopefully close to optimal in some conditions. We have experimented extensively with two such algorithms, that both approximate Bayesian decoding (Rolls, Treves and Tovee 1997b). Both these algorithms fit the response vectors produced over several trials by the cells being recorded to a product of conditional probabilities for the response of each cell given the stimulus. In one case, the single cell conditional probability is assumed to be Gaussian (truncated at zero); in the other it is assumed to be Poisson (with an additional weight at zero). Details of these algorithms are given by Rolls, Treves and Tovee (1997b).

Biologically plausible decoding algorithms are those that limit the algebraic operations used to types that could be easily implemented by neurons, e.g. dot product summations, thresholding and other single-cell non-linearities, and competition and contrast enhancement among the outputs of nearby cells. There is then no need for ever fitting functions or other sophisticated approximations, but of course the degree of arbitrariness in selecting a particular algorithm remains substantial, and a comparison among different choices based on which yields the higher information values may favour one choice in a given situation and another choice with a different data set.

To summarize, the key idea in decoding, in our context of estimating information values, is that it allows substitution of a possibly very high-dimensional response space (which is difficult to sample and regularize) with a reduced object much easier to handle, that is with a discrete set equivalent to the stimulus set. The mutual information between the new set and the stimulus set is then easier to estimate even with limited data, and if the assumptions about population coding, underlying the particular decoding algorithm used, are justified, the
value obtained approximates the original target, the mutual information between stimuli and responses. For each response recorded, one can use all the responses except for that one to generate estimates of the average response vectors (the average response for each neuron in the population) to each stimulus. Then one considers how well the selected response vector matches the average response vectors, and uses the degree of matching to estimate, for all stimuli, the probability that they were the actual stimuli. The form of the matching embodies the general notions about population encoding, for example the ‘degree of matching’ might be simply the dot product between the current vector and the average vector \((r_{av})\), suitably normalized over all average vectors to generate probabilities

\[
P(s'|r(s)) = \frac{r(s) \cdot r_{av}(s')}{\sum_{s''} r(s) \cdot r_{av}(s'')}
\]

where \(s''\) is a dummy variable. (This is called dot product decoding in Fig. 25.13.) One ends up, then, with a table of conjoint probabilities \(P(s;s')\), and another table obtained by selecting for each trial the most likely (or predicted) single stimulus \(s^{p}\), \(P(s;s^{p})\). Both \(s'\) and \(s^{p}\) stand for all possible stimuli, and hence belong to the same set \(S\). These can be used to estimate mutual information values based on probability decoding \((I_{p})\) and on maximum likelihood decoding \((I_{ml})\):

\[
<I_{p}> = \sum_{s \in S} \sum_{s' \in S} P(s, s') \log_{2} \frac{P(s, s')}{P(s)P(s')}
\]

and

\[
<I_{ml}> = \sum_{s \in S} \sum_{s^{p} \in S} P(s, s^{p}) \log_{2} \frac{P(s, s^{p})}{P(s)P(s^{p})}.
\]

Examples of the use of these procedures are available (Rolls, Treves and Tovee 1997b, Booth and Rolls 1998, Rolls, Treves, Robertson, Georges-François and Panzeri 1998b, Rolls, Aggelopoulos, Franco and Treves 2004, Franco, Rolls, Aggelopoulos and Treves 2004, Rolls, Franco, Aggelopoulos and Jerez 2006b), and some of the results obtained are described in Section C.3.

**C.2.4 Information in the correlations between the spikes of different cells: a decoding approach**

Simultaneously recorded neurons sometimes shows cross-correlations in their firing, that is the firing of one is systematically related to the firing of the other cell. One example of this is neuronal response synchronization. The cross-correlation, to be defined below, shows the time difference between the cells at which the systematic relation appears. A significant peak or trough in the cross-correlation function could reveal a synaptic connection from one cell to the other, or a common input to each of the cells, or any of a considerable number of other possibilities. If the synchronization occurred for only some of the stimuli, then the presence of the significant cross-correlation for only those stimuli could provide additional evidence separate from any information in the firing rate of the neurons about which stimulus had been shown. Information theory in principle provides a way of quantitatively assessing the relative contributions from these two types of encoding, by expressing what can be learned from each type of encoding in the same units, bits of information.

Figure C.2 illustrates how synchronization occurring only for some of the stimuli could be used to encode information about which stimulus was presented. In the Figure the spike trains of three neurons are shown after the presentation of two different stimuli on one trial.
The information carried by neuronal responses

Fig. C.2 Illustration of the information that could be carried by spike trains. The responses of three cells to two different stimuli are shown on one trial. Cell 3 reflects which stimulus was shown in the number of spikes produced, and this can be measured as spike count or rate information. Cells 1 and 2 have no spike count or rate information, because the number of spikes is not different for the two stimuli. Cells 1 and 2 do show some synchronization, reflected in the cross-correlogram, that is stimulus dependent, as the synchronization is present only when stimulus 1 is shown. The contribution of this effect is measured as the stimulus-dependent synchronization information.

As shown by the cross-correlogram in the lower part of the figure, the responses of cell 1 and cell 2 are synchronized when stimulus 1 is presented, as whenever a spike from cell 1 is emitted, another spike from cell 2 is emitted after a short time lag. In contrast, when stimulus 2 is presented, synchronization effects do not appear. Thus, based on a measure of the synchrony between the responses of cells 1 and 2, it is possible to obtain some information about what stimulus has been presented. The contribution of this effect is measured as the stimulus-dependent synchronization information. Cells 1 and 2 have no information about what stimulus was presented from the number of spikes, as the same number is found for both stimuli. Cell 3 carries information in the spike count in the time window (which is also called the firing rate) about what stimulus was presented. (Cell 3 emits 6 spikes for stimulus 1 and 3 spikes for stimulus 2.)

The example shown in Fig. C.2 is for the neuronal responses on a single trial. Given that the neuronal responses are variable from trial to trial, we need a method to quantify the information that is gained from a single trial of spike data in the context of the measured
variability in the responses of all of the cells, including how the cells’ responses covary in a way which may be partly stimulus-dependent, and may include synchronization effects. The direct approach is to apply the Shannon mutual information measure (Shannon 1948, Cover and Thomas 1991)

\[ I(s, r) = \sum_{s \in S} \sum_{r} P(s, r) \log_2 \frac{P(s, r)}{P(s)P(r)}, \]  

where \( P(s, r) \) is a probability table embodying a relationship between the variable \( s \) (here, the stimulus) and \( r \) (a vector where each element is the firing rate of one neuron).

However, because the probability table of the relation between the neuronal responses and the stimuli, \( P(s, r) \), is so large (given that there may be many stimuli, and that the response space which has to include spike timing is very large), in practice it is difficult to obtain a sufficient number of trials for every stimulus to generate the probability table accurately, at least with data from mammals in which the experiment cannot usually be continued for many hours of recording from a whole population of cells. To circumvent this undersampling problem, Rolls, Treves and Tovee (1997b) developed a decoding procedure (described in Section C.2.3), in which an estimate (or guess) of which stimulus \( s' \) was shown on a given trial is made from a comparison of the neuronal responses on that trial with the responses made to the whole set of stimuli on other trials. One then obtains a conjoint probability table \( P(s';s) \), and then the mutual information based on probability estimation (PE) decoding \( I_p \) between the estimated stimuli \( s' \) and the actual stimuli \( s \) that were shown can be measured:

\[ < I_p > = \sum_{s \in S} \sum_{s' \in S} P(s, s') \log_2 \frac{P(s, s')}{P(s)P(s')} \]  

\[ = \sum_{s \in S} P(s) \sum_{s' \in S} P(s'|s) \log_2 \frac{P(s'|s)}{P(s')} . \]

These measurements are in the low dimensional space of the number of stimuli, and therefore the number of trials of data needed for each stimulus is of the order of the number of stimuli, which is feasible in experiments. In practice, it is found that for accurate information estimates with the decoding approach, the number of trials for each stimulus should be at least twice the number of stimuli (Franco, Rolls, Aggelopoulos and Treves 2004).

The nature of the decoding procedure is illustrated in Fig. C.3. The left part of the diagram shows the average firing rate (or equivalently spike count) responses of each of 3 cells (labelled as Rate Cell 1,2,3) to a set of 3 stimuli. The last row (labelled Response single trial) shows the data that might be obtained from a single trial and from which the stimulus that was shown (St. ?) must be estimated or decoded, using the average values across trials shown in the top part of the table, and the probability distribution of these values. The decoding step essentially compares the vector of responses on trial St.? with the average response vectors obtained previously to each stimulus. This decoding can be as simple as measuring the correlation, or dot (inner) product, between the test trial vector of responses and the response vectors to each of the stimuli. This procedure is very neuronally plausible, in that the dot product between an input vector of neuronal activity and the synaptic response vector on a single neuron (which might represent the average incoming activity previously to that stimulus) is the simplest operation that it is conceived that neurons might perform (Rolls and Treves 1998, Rolls and Deco 2002, Rolls 2008d). Other decoding procedures include a Bayesian procedure based on a Gaussian or Poisson assumption of the spike count distributions as described in detail by Rolls, Treves and Tovee (1997b). The Gaussian one is what we have used (Franco, Rolls, Aggelopoulos and Treves 2004, Aggelopoulos, Franco and Rolls 2005), and it is described below.
The information carried by neuronal responses

The new step taken by Franco, Rolls, Aggelopoulos and Treves (2004) is to introduce into the Table Data \( (s, r) \) shown in the upper part of Fig. C.3 new columns, shown on the right of the diagram, containing a measure of the cross-correlation (averaged across trials) for some pairs of cells (labelled as Corrln Cells 1−2 and 2−3). The decoding procedure can then take account of any cross-correlations between pairs of cells, and thus measure any contributions to the information from the population of cells that arise from cross-correlations between the neuronal responses. If these cross-correlations are stimulus-dependent, then their positive contribution to the information encoded can be measured. This is the new concept for information measurement from neuronal populations introduced by Franco, Rolls, Aggelopoulos and Treves (2004). We describe next how the cross-correlation information can be introduced into the Table, and then how the information analysis algorithm can be used to measure the contribution of different factors in the neuronal responses to the information that the population encodes.

To test different hypotheses, the decoding can be based on all the columns of the Table (to provide the total information available from both the firing rates and the stimulus-dependent synchronization), on only the columns with the firing rates (to provide the information available from the firing rates), and on only the columns with the cross-correlation values (to provide the information available from the stimulus-dependent cross-correlations). Any information from stimulus-dependent cross-correlations will not necessarily be orthogonal to the rate information, and the procedures allow this to be checked by comparing the total information to that from the sum of the two components. If cross-correlations are present but are
not stimulus-dependent, these will not contribute to the information available about which stimulus was shown.

The measure of the synchronization introduced into the Table Data \((s, r)\) on each trial is, for example, the value of the Pearson cross-correlation coefficient calculated for that trial at the appropriate lag for cell pairs that have significant cross-correlations (Franco, Rolls, Aggelopoulos and Treves 2004). This value of this Pearson cross-correlation coefficient for a single trial can be calculated from pairs of spike trains on a single trial by forming for each cell a vector of 0s and 1s, the 1s representing the time of occurrence of spikes with a temporal resolution of 1 ms. Resulting values within the range \(-1\) to \(1\) are shifted to obtain positive values. An advantage of basing the measure of synchronization on the Pearson cross-correlation coefficient is that it measures the amount of synchronization between a pair of neurons independently of the firing rate of the neurons. The lag at which the cross-correlation measure was computed for every single trial, and whether there was a significant cross-correlation between neuron pairs, can be identified from the location of the peak in the cross-correlogram taken across all trials. The cross-correlogram is calculated by, for every spike that occurred in one neuron, incrementing the bins of a histogram that correspond to the lag times of each of the spikes that occur for the other neuron. The raw cross-correlogram is corrected by subtracting the “shift predictor” cross-correlogram (which is produced by random re-pairings of the trials), to produce the corrected cross-correlogram.

Further details of the decoding procedures are as follows (see Rolls, Treves and Tovee (1997b) and Franco, Rolls, Aggelopoulos and Treves (2004)). The full probability table estimator (PE) algorithm uses a Bayesian approach to extract \(P(s'|r)\) for every single trial from an estimate of the probability \(P(r|s')\) of a stimulus–response pair made from all the other trials (as shown in Bayes’ rule shown in equation C.34) in a cross-validation procedure described by Rolls et al. (1997b).

\[
P(s'|r) = \frac{P(r|s')P(s')}{P(r)}. \tag{C.34}
\]

where \(P(r)\) (the probability of the vector containing the firing rate of each neuron, where each element of the vector is the firing rate of one neuron) is obtained as:

\[
P(r) = \sum_{s'} P(r|s')P(s'). \tag{C.35}
\]

This requires knowledge of the response probabilities \(P(r|s')\) which can be estimated for this purpose from \(P(r, s')\), which is equal to \(P(s') \prod_c P(r_c|s')\), where \(r_c\) is the firing rate of cell \(c\). We note that \(P(r_c|s')\) is derived from the responses of cell \(c\) from all of the trials except for the current trial for which the probability estimate is being made. The probabilities \(P(r_c|s')\) are fitted with a Gaussian (or Poisson) distribution whose amplitude at \(r_c\) gives \(P(r_c|s')\). By summing over different test trial responses to the same stimulus \(s\), we can extract the probability that by presenting stimulus \(s\) the neuronal response is interpreted as having been elicited by stimulus \(s'\),

\[
P(s'|s) = \sum_{r \in \text{test}} P(s'|r)P(r|s). \tag{C.36}
\]

After the decoding procedure, the estimated relative probabilities (normalized to 1) were averaged over all ‘test’ trials for all stimuli, to generate a (Regularized) table \(P_{RN}(s, s')\) describing the relative probability of each pair of actual stimulus \(s\) and posited stimulus \(s'\) (computed with \(N\) trials). From this probability table the mutual information measure \((I_p)\) was calculated as described above in equation C.33.
We also generate a second (Frequency) table \( P^F_N(s, s^p) \) from the fraction of times an actual stimulus \( s \) elicited a response that led to a predicted (single most likely) stimulus \( s^p \). From this probability Table the mutual information measure based on maximum likelihood decoding \( (I_{ml}) \) was calculated with equation C.37:

\[
< I_{ml} > = \sum_{s \in S} \sum_{s^p \in S} P(s, s^p) \log_2 \frac{P(s, s^p)}{P(s)P(s^p)} .
\]  

A detailed comparison of maximum likelihood and probability decoding is provided by Rolls, Treves and Tovee (1997b), but we note here that probability estimate decoding is more regularized (see below) and therefore may be safer to use when investigating the effect on the information of the number of cells. For this reason, the results described by Franco, Rolls, Aggelopoulos and Treves (2004) were obtained with probability estimation (PE) decoding. The maximum likelihood decoding does give an immediate measure of the percentage correct.

Another approach to decoding is the dot product (DP) algorithm which computes the normalized dot products between the current firing vector \( r \) on a “test” (i.e. the current) trial and each of the mean firing rate response vectors in the “training” trials for each stimulus \( s' \) in the cross-validation procedure. (The normalized dot product is the dot or inner product of two vectors divided by the product of the length of each vector. The length of each vector is the square root of the sum of the squares.) Thus, what is computed are the cosines of the angles of the test vector of cell rates with, in turn for each stimulus, the mean response vector to that stimulus. The highest dot product indicates the most likely stimulus that was presented, and this is taken as the predicted stimulus \( s^p \) for the probability table \( P(s, s^p) \). (It can also be used to provide percentage correct measures.)

We note that any decoding procedure can be used in conjunction with information estimates both from the full probability table (to produce \( I_p \)), and from the most likely estimated stimulus for each trial (to produce \( I_{ml} \)).

Because the probability tables from which the information is calculated may be unregularized with a small number of trials, a bias correction procedure to correct for the undersampling is applied, as described in detail by Rolls, Treves and Tovee (1997b) and Panzeri and Treves (1996). In practice, the bias correction that is needed with information estimates using the decoding procedures described by Franco, Rolls, Aggelopoulos and Treves (2004) and by Rolls et al. (1997b) is small, typically less than 10% of the uncorrected estimate of the information, provided that the number of trials for each stimulus is in the order of twice the number of stimuli. We also note that the distortion in the information estimate from the full probability table needs less bias correction than that from the predicted stimulus table (i.e. maximum likelihood) method, as the former is more regularized because every trial makes some contribution through much of the probability table (see Rolls et al. (1997b)). We further note that the bias correction term becomes very small when more than 10 cells are included in the analysis (Rolls et al. 1997b).

Examples of the use of these procedures are available (Franco, Rolls, Aggelopoulos and Treves 2004, Aggelopoulos, Franco and Rolls 2005), and some of the results obtained are described in Section C.3.

### C.2.5 Information in the correlations between the spikes of different cells: a second derivative approach

Another information theory-based approach to stimulus-dependent cross-correlation information has been developed as follows by Panzeri, Schultz, Treves and Rolls (1999a) and Rolls, Franco, Aggelopoulos and Reece (2003b). A problem that must be overcome is the fact that
Information theory, and neuronal encoding

with many simultaneously recorded neurons, each emitting perhaps many spikes at different times, the dimensionality of the response space becomes very large, the information tends to be overestimated, and even bias corrections cannot save the situation. The approach described in this Section (C.2.5) limits the problem by taking short time epochs for the information analysis, in which low numbers of spikes, in practice typically 0, 1, or 2, spikes are likely to occur from each neuron.

In a sufficiently short time window, at most two spikes are emitted from a population of neurons. Taking advantage of this, the response probabilities can be calculated in terms of pairwise correlations. These response probabilities are inserted into the Shannon information formula C.38 to obtain expressions quantifying the impact of the pairwise correlations on the information \( I(t) \) transmitted in a short time \( t \) by groups of spiking neurons:

\[
I(t) = \sum_{s \in S} \sum_{r} P(s, r) \log_2 \frac{P(s, r)}{P(s)P(r)}
\]

where \( r \) is the firing rate response vector comprised by the number of spikes emitted by each of the cells in the population in the short time \( t \), and \( P(s, r) \) refers to the joint probability distribution of stimuli with their respective neuronal response vectors.

The information depends upon the following two types of correlation:

**C.2.5.1 The correlations in the neuronal response variability from the average to each stimulus (sometimes called “noise” correlations) \( \gamma \):**

\( \gamma_{ij}(s) \) (for \( i \neq j \)) is the fraction of coincidences above (or below) that expected from uncorrelated responses, relative to the number of coincidences in the uncorrelated case (which is \( \bar{n}_i(s)\bar{n}_j(s) \), the bar denoting the average across trials belonging to stimulus \( s \), where \( n_i(s) \) is the number of spikes emitted by cell \( i \) to stimulus \( s \) on a given trial)

\[
\gamma_{ij}(s) = \frac{n_i(s)n_j(s)}{\bar{n}_i(s)\bar{n}_j(s)} - 1,
\]

and is named the ‘scaled cross-correlation density’. It can vary from \(-1\) to \(\infty\); negative \( \gamma_{ij}(s) \)’s indicate anticorrelation, whereas positive \( \gamma_{ij}(s) \)’s indicate correlation. \( \gamma_{ij}(s) \) can be thought of as the amount of trial by trial concurrent firing of the cells \( i \) and \( j \), compared to that expected in the uncorrelated case. \( \gamma_{ij}(s) \) (for \( i \neq j \)) is the ‘scaled cross-correlation density’ (Aertsen, Gerstein, Habib and Palm 1989, Panzeri, Schultz, Treves and Rolls 1999a), and is sometimes called the “noise” correlation (Gawne and Richmond 1993, Shadlen and Newsome 1995, Shadlen and Newsome 1998).

\( \gamma_{ij}(s) \) is an alternative, which produces a more compact information analysis, to the neuronal cross-correlation based on the Pearson correlation coefficient \( \rho_{ij}(s) \) (equation C.40), which normalizes the number of coincidences above independence to the standard deviation of the number of coincidences expected if the cells were independent.

The normalization used by the Pearson correlation coefficient has the advantage that it quantifies the strength of independence to the standard deviation of the number of coincidences expected if the cells were independent. For the information analysis, it is more convenient to use the scaled correlation density \( \gamma_{ij}(s) \) than the Pearson correlation coefficient, because of the compactness of the resulting formulation, and because of its scaling properties for small \( t \). \( \gamma_{ij}(s) \) remains finite as \( t \to 0 \), thus by using this measure we can keep the \( t \) expansion of the information explicit. Keeping the time-dependence of the resulting information components explicit greatly increases the amount of insight obtained from the series expansion.

In contrast, the Pearson noise-correlation measure applied to short timescales approaches zero at short time windows:

\[
\rho_{ij}(s) \equiv \frac{n_i(s)n_j(s) - \bar{n}_i(s)\bar{n}_j(s)}{\sigma_{n_i(s)}\sigma_{n_j(s)}} \simeq t \quad \gamma_{ij}(s) = \sqrt{\frac{\bar{n}_i(s)\bar{n}_j(s)}}{\sigma_{n_i(s)}\sigma_{n_j(s)}},
\]

where \( \sigma_{n_i(s)} \) is the standard deviation of the count of spikes emitted by cell \( i \) in response to stimulus \( s \).
C.2.5.2 The correlations in the mean responses of the neurons across the set of stimuli (sometimes called “signal” correlations) $\nu$:

$$\nu_{ij} = \frac{\langle \pi_i(s)\pi_j(s) \rangle}{\langle \pi_i(s) \rangle \langle \pi_j(s) \rangle} - 1 = \frac{\langle \tau_i(s)\tau_j(s) \rangle}{\langle \tau_i(s) \rangle \langle \tau_j(s) \rangle} - 1$$

(C.41)

where $\pi_i(s)$ is the mean rate of response of cell $i$ (among $C$ cells in total) to stimulus $s$ over all the trials in which that stimulus was present. $\nu_{ij}$ can be thought of as the degree of similarity in the mean response profiles (averaged across trials) of the cells $i$ and $j$ to different stimuli. $\nu_{ij}$ is sometimes called the “signal” correlation (Gawne and Richmond 1993, Shadlen and Newsome 1995, Shadlen and Newsome 1998).

C.2.5.3 Information in the cross-correlations in short time periods

In the short timescale limit, the first ($I_t$) and second ($I_{tt}$) information derivatives describe the information $I(t)$ available in the short time $t$

$$I(t) = t I_t + \frac{t^2}{2} I_{tt} .$$

(C.42)

(The zeroth order, time-independent term is zero, as no information can be transmitted by the neurons in a time window of zero length. Higher order terms are also excluded as they become negligible.)

The instantaneous information rate $I_t$ is

$$I_t = \sum_{i=1}^{C} \left\langle r_i(s) \log_2 \frac{\pi_i(s)}{\langle \pi_i(s') \rangle} \right\rangle .$$

(C.43)

This formula shows that this information rate (the first time derivative) should not be linked to a high signal to noise ratio, but only reflects the extent to which the mean responses of each cell are distributed across stimuli. It does not reflect anything of the variability of those responses, that is of their noisiness, nor anything of the correlations among the mean responses of different cells.

The effect of (pairwise) correlations between the cells begins to be expressed in the second time derivative of the information. The expression for the instantaneous information ‘acceleration’ $I_{tt}$ (the second time derivative of the information) breaks up into three terms:

$$I_{tt} = \frac{1}{\ln 2} \sum_{i=1}^{C} \sum_{j=1}^{C} \left\langle r_i(s)\gamma_{ij}(s) \right\rangle_s \left[ \nu_{ij} + (1 + \nu_{ij}) \ln \left( \frac{1}{1 + \nu_{ij}} \right) \right]$$

$$+ \sum_{i=1}^{C} \sum_{j=1}^{C} \left\langle \pi_i(s)\pi_j(s)\gamma_{ij}(s) \right\rangle_s \log_2 \left( \frac{1}{1 + \nu_{ij}} \right)$$

$$+ \sum_{i=1}^{C} \sum_{j=1}^{C} \left\langle r_i(s)\pi_j(s)(1 + \gamma_{ij}(s)) \log_2 \left[ \frac{1 + \gamma_{ij}(s)}{\langle \pi_i(s')\pi_j(s') \rangle_s(1 + \gamma_{ij}(s'))} \right] \right\rangle_s .$$

(C.44)

The first of these terms is all that survives if there is no noise correlation at all. Thus the rate component of the information is given by the sum of $I_t$ (which is always greater than or equal to zero) and of the first term of $I_{tt}$ (which is instead always less than or equal to zero).

\[\text{Note that } s' \text{ is used in equations C.43 and C.44 just as a dummy variable to stand for } s, \text{ as there are two summations performed over } s.\]
The second term is non-zero if there is some correlation in the variance to a given stimulus, even if it is independent of which stimulus is present; this term thus represents the contribution of stimulus-independent noise correlation to the information.

The third component of $I_{tt}$ represents the contribution of stimulus-modulated noise correlation, as it becomes non-zero only for stimulus-dependent noise correlations. These last two terms of $I_{tt}$ together are referred to as the correlational components of the information.

The application of this approach to measuring the information in the relative time of firing of simultaneously recorded cells, together with further details of the method, are described by Panzeri, Treves, Schultz and Rolls (1999b), Rolls, Franco, Aggelopoulos and Reece (2003b), and Rolls, Aggelopoulos, Franco and Treves (2004), and in Section C.3.7.

C.3 Neuronal encoding: results obtained from applying information-theoretic analyses

How is information encoded in cortical areas such as the inferior temporal visual cortex? Can we read the code being used by the cortex? What are the advantages of the encoding scheme used for the neuronal network computations being performed in different areas of the cortex? These are some of the key issues considered in this Section (C.3). Because information is exchanged between the computing elements of the cortex (the neurons) by their spiking activity, which is conveyed by their axon to synapses onto other neurons, the appropriate level of analysis is how single neurons, and populations of single neurons, encode information in their firing. More global measures that reflect the averaged activity of large numbers of neurons (for example, PET (positron emission tomography) and fMRI (functional magnetic resonance imaging), EEG (electroencephalographic recording), and ERPs (event-related potentials)) cannot reveal how the information is represented, or how the computation is being performed.

Although information theory provides the natural mathematical framework for analysing the performance of neuronal systems, its applications in neuroscience have been for many years rather sparse and episodic (e.g. MacKay and McCulloch (1952); Eckhorn and Popel (1974); Eckhorn and Popel (1975); Eckhorn, Grusser, Kroller, Pellnitz and Popel (1976)). One reason for this limited application of information theory has been the great effort that was apparently required, due essentially to the limited sampling problem, in order to obtain accurate results. Another reason has been the hesitation in analysing as a single complex ‘black-box’ large neuronal systems all the way from some external, easily controllable inputs, up to neuronal activity in some central cortical area of interest, for example including all visual stations from the periphery to the end of the ventral visual stream in the temporal lobe. In fact, two important bodies of work, that have greatly helped revive interest in applications of the theory in recent years, both sidestep these two problems. The problem with analysing a huge black-box is avoided by considering systems at the sensory periphery; the limited sampling problem is avoided either by working with insects, in which sampling can be extensive (Bialek, Rieke, de Ruyter van Steveninck and Warland 1991, de Ruyter van Steveninck and Laughlin 1996, Rieke, Warland, de Ruyter van Steveninck and Bialek 1997), or by utilizing a formal model instead of real data (Atick and Redlich 1990, Atick 1992). Both approaches have provided insightful quantitative analyses that are in the process of being extended to more central mammalian systems (see e.g. Atick, Griffin and Relich (1996)).

In the treatment provided here, we focus on applications to the mammalian brain, using examples from a whole series of investigations on information representation in visual cortical areas, the original papers on which refer to related publications.
C.3.1 The sparseness of the distributed encoding used by the brain

Some of the types of representation that might be found at the neuronal level are summarized next (cf. Section 8.2). A local representation is one in which all the information that a particular stimulus or event occurred is provided by the activity of one of the neurons. This is sometimes called a grandmother cell representation, because in a famous example, a single neuron might be active only if one’s grandmother was being seen (see Barlow (1995)). A fully distributed representation is one in which all the information that a particular stimulus or event occurred is provided by the activity of the full set of neurons. If the neurons are binary (for example, either active or not), the most distributed encoding is when half the neurons are active for any one stimulus or event. A sparse distributed representation is a distributed representation in which a small proportion of the neurons is active at any one time.

C.3.1.1 Single neuron sparseness \( a^s \)

Equation C.45 defines a measure of the single neuron sparseness, \( a^s \):

\[
a^s = \frac{\left( \sum_{s=1}^{S} y_s / S \right)^2}{\left( \sum_{s=1}^{S} y_s^2 / S \right)}
\]  

\[ \text{(C.45)} \]

where \( y_s \) is the mean firing rate of the neuron to stimulus \( s \) in the set of \( S \) stimuli (Rolls and Treves 1998). For a binary representation, \( a^s \) is 0.5 for a fully distributed representation, and \( 1 / S \) if a neuron responds to one of a set of \( S \) stimuli. Another measure of sparseness is the kurtosis of the distribution, which is the fourth moment of the distribution. It reflects the length of the tail of the distribution. (An actual distribution of the firing rates of a neuron to a set of 65 stimuli is shown in Fig. C.4. The sparseness \( a^s \) for this neuron was 0.69 (see Rolls, Treves, Tovee and Panzeri (1997d).)

It is important to understand and quantify the sparseness of representations in the brain, because many of the useful properties of neuronal networks such as generalization and completion only occur if the representations are not local (see Appendix B), and because the value of the sparseness is an important factor in how many memories can be stored in such neural networks. Relatively sparse representations (low values of \( a^s \)) might be expected in memory systems as this will increase the number of different memories that can be stored and retrieved. Less sparse representations might be expected in sensory systems, as this could allow more information to be represented (see Table B.2).

Barlow (1972) proposed a single neuron doctrine for perceptual psychology. He proposed that sensory systems are organized to achieve as complete a representation as possible with the minimum number of active neurons. He suggested that at progressively higher levels of sensory processing, fewer and fewer cells are active, and that each represents a more and more specific happening in the sensory environment. He suggested that 1,000 active neurons (which he called cardinal cells) might represent the whole of a visual scene. An important principle involved in forming such a representation was the reduction of redundancy. The implication of Barlow’s (1972) approach was that when an object is being recognized, there are, towards the end of the visual system, a small number of neurons (the cardinal cells) that are so specifically tuned that the activity of these neurons encodes the information that one particular object is being seen. (He thought that an active neuron conveys something of the order of complexity of a word.) The encoding of information in such a system is described as local, in that knowing the activity of just one neuron provides evidence that a particular stimulus (or, more exactly, a given ‘trigger feature’) is present. Barlow (1972) eschewed ‘combinatorial rules of usage of nerve cells’, and believed that the subtlety and sensitivity
of perception results from the mechanisms determining when a single cell becomes active. In contrast, with distributed or ensemble encoding, the activity of several or many neurons must be known in order to identify which stimulus is present, that is, to read the code. It is the relative firing of the different neurons in the ensemble that provides the information about which object is present.

At the time Barlow (1972) wrote, there was little actual evidence on the activity of neurons in the higher parts of the visual and other sensory systems. There is now considerable evidence, which is now described.

First, it has been shown that the representation of which particular object (face) is present is actually rather distributed. Baylis, Rolls and Leonard (1985) showed this with the responses of temporal cortical neurons that typically responded to several members of a set of five faces, with each neuron having a different profile of responses to each face (see examples in Fig. 25.12 on page 569). It would be difficult for most of these single cells to tell which of even five faces, let alone which of hundreds of faces, had been seen. (At the same time, the neurons discriminated between the faces reliably, as shown by the values of \(d'\), taken, in the case of the neurons, to be the number of standard deviations of the neuronal responses that separated the response to the best face in the set from that to the least effective face in the set. The values of \(d'\) were typically in the range 1–3.)

Second, the distributed nature of the representation can be further understood by the finding that the firing rate probability distribution of single neurons, when a wide range of natural visual stimuli are being viewed, is approximately exponential, with rather few stimuli producing high firing rates, and increasingly large numbers of stimuli producing lower and lower firing rates, as illustrated in Fig. C.5a (Rolls and Tovee 1995b, Baddeley, Abbott, Booth, Sengpiel, Freeman, Wakeman and Rolls 1997, Treves, Panzeri, Rolls, Booth and Wakeman 1999, Franco, Rolls, Aggelopoulos and Jerez 2007).

For example, the responses of a set of temporal cortical neurons to 23 faces and 42 non-face natural images were measured, and a distributed representation was found (Rolls and Tovee 1995b). The tuning was typically graded, with a range of different firing rates to the set of faces, and very little response to the non-face stimuli (see example in Fig. C.4). The spontaneous firing rate of the neuron in Fig. C.4 was 20 spikes/s, and the histogram bars indicate the change of firing rate from the spontaneous value produced by each stimulus. Stimuli that are faces are marked F, or P if they are in profile. B refers to images of scenes that included either a small face within the scene, sometimes as part of an image that included a whole person, or other body parts, such as hands (H) or legs. The non-face stimuli are unlabelled. The neuron responded best to three of the faces (profile views), had some response to some of the other faces, and had little or no response, and sometimes had a small decrease of firing rate below the spontaneous firing rate, to the non-face stimuli. The sparseness value \(a^*\) for this cell across all 68 stimuli was 0.69, and the response sparseness \(a^*_r\) (based on the evoked responses minus the spontaneous firing of the neuron) was 0.19. It was found that the sparseness of the representation of the 68 stimuli by each neuron had an average across all neurons of 0.65 (Rolls and Tovee 1995b). This indicates a rather distributed representation. (If neurons had a continuum of firing rates equally distributed between zero and maximum rate, \(a^*\) would be 0.75, while if the probability of each response decreased linearly, to reach zero at the maximum rate, \(a^*\) would be 0.67).

I comment that these values for \(a\) do not seem very sparse. But these values are calculated using the raw firing rates of the neurons, on the basis that these would be what a receiving neuron would receive as its input representation. However, neocortical neurons have a spontaneous firing rate of several spikes/s (with a lower value of 0.75 spikes/s for hippocampal pyramidal cells), and if this spontaneous value is subtracted from the firing rates to yield a ‘response sparseness’ \(a_r\), this value is considerably lower. For example, the sparseness \(a\) of
Fig. C.4 Firing rate distribution of a single neuron in the temporal visual cortex to a set of 23 face (F) and 45 non-face images of natural scenes. The firing rate to each of the 68 stimuli is shown. The neuron does not respond to just one of the 68 stimuli. Instead, it responds to a small proportion of stimuli with high rates, to more stimuli with intermediate rates, and to many stimuli with almost no change of firing. This is typical of the distributed representations found in temporal cortical visual areas. (After Rolls and Tovee 1995b.)

Inferior temporal cortex responses to a set of 68 stimuli had an average across all neurons that we analyzed in this study of 0.65 (Rolls and Tovee 1995b). If the spontaneous firing rate was subtracted from the firing rate of the neuron to each stimulus, so that the changes of firing rate, i.e., the responses of the neurons, were used in the sparseness calculation, then the ‘response sparseness’ had a lower value, with a mean of $a_r=0.33$ for the population of neurons, or 0.60 if calculated over the set of faces rather than over all the face and non-face stimuli. Further, the true sparseness of the representation is probably much less than this, for this is calculated only over the neurons that had responses to some of these stimuli. There were many more neurons that had no response to the stimuli. At least 10 times the number of inferior temporal cortex neurons had no responses to this set of 68 stimuli. So the true sparseness would be much lower than this value of 0.33. Further, it is important to remember the relative nature of sparseness measures, which (like the information measures to be discussed below) depend strongly on the stimulus set used. Thus we can reject a cardinal cell representation. As shown below, the readout of information from these cells is actually much better in any case than
would be obtained from a local representation, and this makes it unlikely that there is a further population of neurons with very specific tuning that use local encoding.

These data provide a clear answer to whether these neurons are grandmother cells: they are not, in the sense that each neuron has a graded set of responses to the different members of a set of stimuli, with the prototypical distribution similar to that of the neuron illustrated in Fig. C.4. On the other hand, each neuron does respond very much more to some stimuli than to many others, and in this sense is tuned to some stimuli.

Figure C.5 shows data of the type shown in Fig. C.4 as firing rate probability density functions, that is as the probability that the neuron will be firing with particular rates. These data were from inferior temporal cortex neurons, and show when tested with a set of 20 face and non-face stimuli how fast the neuron will be firing in a period 100–300 ms after the visual stimulus appears (Franco, Rolls, Aggelopoulos and Jerez 2007). Figure C.5a shows an example of a neuron where the data fit an exponential firing rate probability distribution, with many occasions on which the neuron was firing with a very low firing rate, and decreasingly few occasions on which it fired at higher rates. This shows that the neuron can have high firing rates, but only to a few stimuli. Figure C.5b shows an example of a neuron where the data do not fit an exponential firing rate probability distribution, with insufficiently few very low rates. Of the 41 responsive neurons in this data set, 15 had a good fit to an exponential firing rate probability distribution; the other 26 neurons did not fit an exponential but did fit a gamma distribution in the way illustrated in Fig. C.5b. For the neurons with an exponential distribution, the mean firing rate across the stimulus set was 5.7 spikes/s, and for the neurons with a gamma distribution was 21.1 spikes/s (t=4.5, df=25, p < 0.001). It may be that neurons with high mean rates to a stimulus set tend to have few low rates ever, and this accounts for their poor fit to an exponential firing rate probability distribution, which fits when there are many low firing rate values in the distribution as in Fig. C.5a.

The large set of 68 stimuli used by Rolls and Tovee (1995b) was chosen to produce an approximation to a set of stimuli that might be found to natural stimuli in a natural environment, and thus to provide evidence about the firing rate distribution of neurons to natural stimuli. Another approach to the same fundamental question was taken by Baddeley,
Abbott, Booth, Sengpiel, Freeman, Wakeman, and Rolls (1997) who measured the firing rates over short periods of individual inferior temporal cortex neurons while monkeys watched continuous videos of natural scenes. They found that the firing rates of the neurons were again approximately exponentially distributed (see Fig. C.6), providing further evidence that this type of representation is characteristic of inferior temporal cortex (and indeed also V1) neurons.

The actual distribution of the firing rates to a wide set of natural stimuli is of interest, because it has a rather stereotypical shape, typically following a graded unimodal distribution with a long tail extending to high rates (see for example Figs. C.5a and C.6). The mode of the distribution is close to the spontaneous firing rate, and sometimes it is at zero firing. If the number of spikes recorded in a fixed time window is taken to be constrained by a fixed maximum rate, one can try to interpret the distribution observed in terms of optimal information transmission (Shannon 1948), by making the additional assumption that the coding is noiseless. An exponential distribution, which maximizes entropy (and hence information transmission for noiseless codes) is the most efficient in terms of energy consumption if its mean takes an optimal value that is a decreasing function of the relative metabolic cost of emitting a spike (Levy and Baxter 1996). This argument would favour sparser coding schemes the more energy expensive neuronal firing is (relative to rest). Although the tail of actual firing rate distributions is often approximately exponential (see for example Figs. C.5a and C.6; Baddeley, Abbott, Booth, Sengpiel, Freeman, Wakeman and Rolls (1997); Rolls, Treves, Tovee and Panzeri (1997d); and Franco, Rolls, Aggelopoulos and Jerez (2007)), the maximum entropy argument cannot apply as such, because noise is present and the noise level varies as a function of the rate, which makes entropy maximization different from information maximization. Moreover, a mode at low but non-zero rate, which is often observed (see e.g. Fig. C.5b), is inconsistent with the energy efficiency hypothesis.

A simpler explanation for the characteristic firing rate distribution arises by appreciating
that the value of the activation of a neuron across stimuli, reflecting a multitude of contributing factors, will typically have a Gaussian distribution; and by considering a physiological input–output transform (i.e. activation function), and realistic noise levels. In fact, an input–output transform that is supralinear in a range above threshold results from a fundamentally linear transform and fluctuations in the activation, and produces a variance in the output rate, across repeated trials, that increases with the rate itself, consistent with common observations. At the same time, such a supralinear transform tends to convert the Gaussian tail of the activation distribution into an approximately exponential tail, without implying a fully exponential distribution with the mode at zero. Such basic assumptions yield excellent fits with observed distributions (Treves, Panzeri, Rolls, Booth and Wakeman 1999), which often differ from exponential in that there are too few very low rates observed, and too many low rates (Rolls, Treves, Tovee and Panzeri 1997d, Franco, Rolls, Aggelopoulos and Jerez 2007).

This peak at low but non-zero rates may be related to the low firing rate spontaneous activity that is typical of many cortical neurons. Keeping the neurons close to threshold in this way may maximize the speed with which a network can respond to new inputs (because time is not required to bring the neurons from a strongly hyperpolarized state up to threshold). The advantage of having low spontaneous firing rates may be a further reason why a curve such as an exponential cannot sometimes be exactly fitted to the experimental data.

A conclusion of this analysis was that the firing rate distribution may arise from the threshold non-linearity of neurons combined with short-term variability in the responses of neurons (Treves, Panzeri, Rolls, Booth and Wakeman 1999).

However, given that the firing rate distribution for some neurons is approximately exponential, some properties of this type of representation are worth elucidation. The sparseness of such an exponential distribution of firing rates is 0.5. This has interesting implications, for to the extent that the firing rates are exponentially distributed, this fixes an important parameter of cortical neuronal encoding to be close to 0.5. Indeed, only one parameter specifies the

---

*Fig. C.7* The set of 20 stimuli used to investigate the tuning of inferior temporal cortex neurons by Franco, Rolls, Aggelopoulos and Jerez 2007. These objects and faces are typical of those encoded in the ways described here by inferior temporal cortex neurons. The code can be read off simply from the firing rates of the neurons about which object or face was shown, and many of the neurons have invariant responses.
Fig. C.8 An exponential firing rate probability distribution obtained by pooling the firing rates of a population of 41 inferior temporal cortex neurons tested to a set of 20 face and non-face stimuli. The firing rate probability distribution for the 100–300 ms interval following stimulus onset was formed by adding the spike counts from all 41 neurons, and across all stimuli. The fit to the exponential distribution (dashed line) was high. (After Franco, Rolls, Aggelopoulos and Jerez 2007.)

The single neuron selectivity reflects response distributions of individual neurons across time to different stimuli. As we have seen, part of the interest of measuring the firing rate probability distributions of individual neurons is that one form of the probability distribution, the exponential, maximizes the entropy of the neuronal responses for a given mean firing rate, which could be used to maximize information transmission consistent with keeping the firing rate on average low, in order to minimize metabolic expenditure (Levy and Baxter 1996, Baddeley, Abbott, Booth, Sengpiel, Freeman, Wakeman and Rolls 1997). Franco, Rolls, Aggelopoulos and Jerez (2007) showed that while the firing rates of some single inferior temporal cortex neurons (tested in a visual fixation task to a set of 20 face and non-face stimuli illustrated in Fig. C.7) do fit an exponential distribution, and others with higher spontaneous firing rates do not, as described above, it turns out that there is a very close fit
to an exponential distribution of firing rates if all spikes from all the neurons are considered together. This interesting result is shown in Fig. C.8.

One implication of the result shown in Fig. C.8 is that a neuron with inputs from the inferior temporal visual cortex will receive an exponential distribution of firing rates on its afferents, and this is therefore the type of input that needs to be considered in theoretical models of neuronal network function in the brain (see Appendix B). The second implication is that at the level of single neurons, an exponential probability density function is consistent with minimizing energy utilization, and maximizing information transmission, for a given mean firing rate (Levy and Baxter 1996, Baddeley, Abbott, Booth, Sengpiel, Freeman, Wakeman and Rolls 1997).

C.3.1.2 Population sparseness $a^p$

If instead we consider the responses of a population of neurons taken at any one time (to one stimulus), we might also expect a sparse graded distribution, with few neurons firing fast to a particular stimulus. It is important to measure the population sparseness, for this is a key parameter that influences the number of different stimuli that can be stored and retrieved in networks such as those found in the cortex with recurrent collateral connections between the excitatory neurons, which can form autoassociation or attractor networks if the synapses are associatively modifiable (Hopfield 1982, Treves and Rolls 1991, Rolls and Treves 1998, Rolls and Deco 2002, Rolls 2008d) (see Appendix B). Further, in physics, if one can predict the distribution of the responses of the system at any one time (the population level) from the distribution of the responses of a component of the system across time, the system is described as ergodic, and a necessary condition for this is that the components are uncorrelated (Lehky, Sejnowski and Desimone 2005). Considering this in neuronal terms, the average sparseness of a population of neurons over multiple stimulus inputs must equal the average selectivity to the stimuli of the single neurons within the population provided that the responses of the neurons are uncorrelated (Földiák 2003).

The sparseness $a^p$ of the population code may be quantified (for any one stimulus) as

$$a^p = \frac{\left( \sum_{n=1}^{N} \frac{y_n}{N} \right)^2}{\sum_{n=1}^{N} \frac{y_n^2}{N}}$$

where $y_n$ is the mean firing rate of neuron $n$ in the set of $N$ neurons.

This measure, $a^p$, of the sparseness of the representation of a stimulus by a population of neurons has a number of advantages. One is that it is the same measure of sparseness that has proved to be useful and tractable in formal analyses of the capacity of associative neural networks and the interference between stimuli that use an approach derived from theoretical physics (Rolls and Treves 1990, Treves 1990, Treves and Rolls 1991, Rolls and Treves 1998) (see Appendix B). We note that high values of $a^p$ indicate broad tuning of the population, and that low values of $a^p$ indicate sparse population encoding.

Franco, Rolls, Aggelopoulos and Jerez (2007) measured the population sparseness of a set of 29 inferior temporal cortex neurons to a set of 20 stimuli that included faces and objects (see Fig. C.7). Figure C.9a shows, for any one stimulus picked at random, the normalized firing rates of the population of neurons. The rates are ranked with the neuron with the highest rate on the left. For different stimuli, the shape of this distribution is on average the same, though with the neurons in a different order. (The rates of each neuron were normalized to a mean of 10 spikes/s before this graph was made, so that the neurons can be combined in the same graph, and so that the population sparseness has a defined value, as described by Franco,
Population sparseness. (a) The firing rates of a population of inferior temporal cortex neurons to any one stimulus from a set of 20 face and non-face stimuli. The rates of each neuron were normalized to the same average value of 10 spikes/s, then for each stimulus, the cell firing rates were placed in rank order, and then the mean firing rates of the first ranked cell, second ranked cell, etc. were taken. The graph thus shows how, for any one stimulus picked at random, the expected normalized firing rates of the population of neurons. (b) The population normalized firing rate probability distributions for any one stimulus. This was computed effectively by taking the probability density function of the data shown in (a). (After Franco, Rolls, Aggelopoulos and Jerez 2007.)

Figure C.9a shows the probability distribution of the normalized firing rates of the population of (29) neurons to any stimulus from the set. This was calculated by taking the probability distribution of the data shown in Fig. C.9a. This distribution is not exponential because of the normalization of the firing rates of each neuron, but becomes exponential as shown in Fig. C.8 without the normalization step.

A very interesting finding of Franco, Rolls, Aggelopoulos and Jerez (2007) was that when the single cell sparseness $a_s$ and the population sparseness $a_p$ were measured from the same set of neurons in the same experiment, the values were very close, in this case 0.77. (This was found for a range of measurement intervals after stimulus onset, and also for a larger population of 41 neurons.)

The single cell sparseness $a_s$ and the population sparseness $a_p$ can take the same value if the response profiles of the neurons are uncorrelated, that is each neuron is independently tuned to the set of stimuli (Lehky et al. 2005). Franco, Rolls, Aggelopoulos and Jerez (2007) tested whether the response profiles of the neurons to the set of stimuli were uncorrelated in two ways. In a first test, they found that the mean (Pearson) correlation between the response profiles computed over the 406 neuron pairs was low, $0.049 \pm 0.013$ (sem). In a second test, they computed how the multiple cell information available from these neurons about which stimulus was shown increased as the number of neurons in the sample was increased, and showed that the information increased approximately linearly with the number of neurons in the ensemble. The implication is that the neurons convey independent (non-redundant) information, and this would be expected to occur if the response profiles of the neurons to the stimuli are uncorrelated.

We now consider the concept of ergodicity. The single neuron selectivity, $a_s^*$, reflects response distributions of individual neurons across time and therefore stimuli in the world...
(and has sometimes been termed “lifetime sparseness”). The population sparseness $a_p$ reflects response distributions across all neurons in a population measured simultaneously (to for example one stimulus). The similarity of the average values of $a_s$ and $a_p$ (both 0.77 for inferior temporal cortex neurons (Franco, Rolls, Aggelopoulos and Jerez 2007)) indicates, we believe for the first time experimentally, that the representation (at least in the inferior temporal cortex) is ergodic. The representation is ergodic in the sense of statistical physics, where the average of a single component (in this context a single neuron) across time is compared with the average of an ensemble of components at one time (cf. Masuda and Aihara (2003) and Lehky et al. (2005)). This is described further next.

In comparing the neuronal selectivities $a_s$ and population sparsenesses $a_p$, we formed a table in which the columns represent different neurons, and the stimuli different rows (Földiák 2003). We are interested in the probability distribution functions (and not just their summary values $a_s$ and $a_p$), of the columns (which represent the individual neuron selectivities) and the rows (which represent the population tuning to any one stimulus). We could call the system strongly ergodic (cf. Lehky et al. (2005)) if the selectivity (probability density or distribution function) of each individual neuron is the same as the average population sparseness (probability density function). (Each neuron would be tuned to different stimuli, but have the same shape of the probability density function.) We have seen that this is not the case, in that the firing rate probability distribution functions of different neurons are different, with some fitting an exponential function, and some a gamma function (see Fig. C.5). We can call the system weakly ergodic if individual neurons have different selectivities (i.e. different response probability density functions), but the average selectivity (measured in our case by $<a_s>$) is the same as the average population sparseness (measured by $<a_p>$), where $<a>$ indicates the ensemble average. We have seen that for inferior temporal cortex neurons the neuron selectivity probability density functions are different (see Fig. C.5), but that their average $<a_s>$ is the same as the average (across stimuli) $<a_p>$ of the population sparseness, 0.77, and thus conclude that the representation in the inferior temporal visual cortex of objects and faces is weakly ergodic (Franco, Rolls, Aggelopoulos and Jerez 2007).

We note that weak ergodicity necessarily occurs if $<a_s>$ and $<a_p>$ are the same and the neurons are uncorrelated, that is each neuron is independently tuned to the set of stimuli (Lehky et al. 2005). The fact that both hold for the inferior temporal cortex neurons studied by Franco, Rolls, Aggelopoulos and Jerez (2007) thus indicates that their responses are uncorrelated, and this is potentially an important conclusion about the encoding of stimuli by these neurons. This conclusion is confirmed by the linear increase in the information with the number of neurons which is the case not only for this set of neurons (Franco, Rolls, Aggelopoulos and Jerez 2007), but also in other data sets for the inferior temporal visual cortex (Rolls, Treves and Tovee 1997b, Booth and Rolls 1998). Both types of evidence thus indicate that the encoding provided by at least small subsets (up to e.g. 20 neurons) of inferior temporal cortex neurons is approximately independent (non-redundant), which is an important principle of cortical encoding.

C.3.1.3 Comparisons of sparseness between areas: the hippocampus, insula, orbitofrontal cortex, and amygdala

In the study of Franco, Rolls, Aggelopoulos and Jerez (2007) on inferior temporal visual cortex neurons, the selectivity of individual cells for the set of stimuli, or single cell sparseness $a_s$, had a mean value of 0.77. This is close to a previously measured estimate, 0.65, which was obtained with a larger stimulus set of 68 stimuli (Rolls and Tovee 1995b). Thus the single neuron probability density functions in these areas do not produce very sparse representations. Therefore the goal of the computations in the inferior temporal visual cortex may not be to produce sparse representations (as has been proposed for V1 (Field 1994, Olshausen and
Field 1997, Vinje and Gallant 2000, Olshausen and Field 2004)). Instead one of the goals of the computations in the inferior temporal visual cortex may be to compute invariant representations of objects and faces (Rolls 2000a, Rolls and Deco 2002, Rolls and Stringer 2006) (see Chapter 25), and to produce not very sparse distributed representations in order to maximize the information represented (see Table B.2 on page 722). In this context, it is very interesting that the representations of different stimuli provided by a population of inferior temporal cortex neurons are decorrelated, as shown by the finding that the mean (Pearson) correlation between the response profiles to a set of 20 stimuli computed over 406 neuron pairs was low, $0.049 \pm 0.013$ (sem) (Franco, Rolls, Aggelopoulos and Jerez 2007). The implication is that decorrelation is being achieved in the inferior temporal visual cortex, but not by forming a sparse code. It will be interesting to investigate the mechanisms for this.

In contrast, the representation in some memory systems may be more sparse. For example, in the hippocampus in which spatial view cells are found in macaques, further analysis of data described by Rolls, Treves, Robertson, Georges-François and Panzeri (1998b) shows that for the representation of 64 locations around the walls of the room, the mean single cell sparseness $<\alpha>$ was $0.34 \pm 0.13$ (sd), and the mean population sparseness $\alpha^p$ was $0.33 \pm 0.11$. The more sparse representation is consistent with the view that the hippocampus is involved in storing memories, and that for this, more sparse representations than in perceptual areas are relevant. These sparseness values are for spatial view neurons, but it is possible that when neurons respond to combinations of spatial view and object (Rolls, Xiang and Franco 2005c), or of spatial view and reward (Rolls and Xiang 2005), the representations are more sparse. It is of interest that the mean firing rate of these spatial view neurons across all spatial views was 1.77 spikes/s (Rolls, Treves, Robertson, Georges-François and Panzeri 1998b). (The mean spontaneous firing rate of the neurons was 0.1 spikes/s, and the average across neurons of the firing rate for the most effective spatial view was 13.2 spikes/s.) It is also notable that weak ergodicity is implied for this brain region too (given the similar values of $<\alpha>$ and $<\alpha^p>$), and the underlying basis for this is that the response profiles of the different hippocampal neurons to the spatial views are uncorrelated. Further support for these conclusions is that the information about spatial view increases linearly with the number of hippocampal spatial view neurons (Rolls, Treves, Robertson, Georges-François and Panzeri 1998b), again providing evidence that the response profiles of the different neurons are uncorrelated. The representations in the hippocampus may be more sparse than this, in line with the observation that in rodents, hippocampal neurons may have place fields in only one of several environments.

Further evidence is now available on ergodicity in three further brain areas, the macaque insular primary taste cortex, the orbitofrontal cortex, and the amygdala (Rolls, Critchley, Verhagen and Kadohisa 2010a). In all these brain areas sets of neurons were tested with an identical set of 24 oral taste, temperature, and texture stimuli. (The stimuli were: Taste - 0.1 M NaCl (salt), 1 M glucose (sweet), 0.01 M HCl (sour), 0.001 M quinine HCl (bitter), 0.1 M monosodium glutamate (umami), and water; Temperature - $10^\circ$C, $37^\circ$C and $42^\circ$C; flavour - blackcurrant juice; viscosity - carboxymethyl-cellulose 10 cPoiße, 100 cPoiße, 1000 cPoiße and 10000 cPoiße; fatty / oily - single cream, vegetable oil, mineral oil, silicone oil (100 cPoiße), coconut oil, and safflower oil; fatty acids - linoleic acid and lauric acid; capsaicin; and gritty texture.) Further analysis of data described by Verhagen, Kadohisa and Rolls (2004) showed that in the primary taste cortex the mean value of $\alpha^p$ across 58 neurons was 0.745 and of $\alpha^p$ (normalized) was 0.708. Further analysis of data described by Rolls, Verhagen and Kadohisa (2003c), Verhagen, Rolls and Kadohisa (2003), Kadohisa, Rolls and Verhagen (2004) and Kadohisa, Rolls and Verhagen (2005a) showed that in the orbitofrontal cortex the mean value of $\alpha^p$ across 30 neurons was 0.625 and of $\alpha^p$ was 0.611. Further analysis of data described by Kadohisa, Rolls and Verhagen (2005b) showed that in the amygdala the mean value of $\alpha^p$ across 38 neurons was 0.811 and of $\alpha^p$ was 0.813. Thus in all these cases, the mean
value of $a^*$ is close to that of $a^p$, and weak ergodicity is implied. The values of $a^*$ and $a^p$ are also relatively high, implying the importance of representing large amounts of information in these brain areas about this set of stimuli by using a very distributed code, and also perhaps about the stimulus set, some members of which may be rather similar to each other.

### C.3.2 The information from single neurons

Examples of the responses of single neurons (in this case in the inferior temporal visual cortex) to sets of objects and/or faces (of the type illustrated in Fig. C.7) are shown in Figs. 16.15, 25.12 and C.4. We now consider how much information these types of neuronal response convey about the set of stimuli $S$, and about each stimulus $s$ in the set. The mutual information $I(S,R)$ that the set of responses $R$ encode about the set of stimuli $S$ is calculated with equation C.21 and corrected for the limited sampling using the analytic bias correction procedure described by Panzeri and Treves (1996) as described in detail by Rolls, Treves, Tovee and Panzeri (1997d). The information $I(s,R)$ about each single stimulus $s$ in the set $S$, termed the stimulus-specific information (Rolls, Treves, Tovee and Panzeri 1997d) or stimulus-specific surprise (DeWeese and Meister 1999), obtained from the set of the responses $R$ of the single neuron is calculated with equation C.22 and corrected for the limited sampling using the analytic bias correction procedure described by Panzeri and Treves (1996) as described in detail by Rolls, Treves, Tovee and Panzeri (1997d). (The average of $I(s,R)$ across stimuli is the mutual information $I(S,R)$.)

Figure C.10 shows the stimulus-specific information $I(s,R)$ available in the neuronal response about each of 20 face stimuli calculated for the neuron (am242) whose firing rate response profile to the set of 65 stimuli is shown in Fig. C.4. Unless otherwise stated, the information measures given are for the information available on a single trial from the firing rate of the neuron in a 500 ms period starting 100 ms after the onset of the stimuli. It is shown in Fig. C.10 that 2.2, 2.0, and 1.5 bits of information were present about the three face stimuli to which the neuron had the highest firing rate responses. The neuron conveyed some but smaller amounts of information about the remaining face stimuli. The average information $I(S,R)$ about this set ($S$) of 20 faces for this neuron was 0.55 bits. The average firing rate of this neuron to these 20 face stimuli was 54 spikes/s. It is clear from Fig. C.10 that little information was available from the responses of the neuron to a particular face stimulus if that response was close to the average response of the neuron across all stimuli. At the same time, it is clear from Fig. C.10 that information was present depending on how far the firing rate to a particular stimulus was from the average response of the neuron to the stimuli. Of particular interest, it is evident that information is present from the neuronal response about which face was shown if that neuronal response was below the average response, as well as when the response was greater than the average response.

One intuitive way to understand the data shown in Fig. C.10 is to appreciate that low probability firing rate responses, whether they are greater than or less than the mean response rate, convey much information about which stimulus was seen. This is of course close to the definition of information. Given that the firing rates of neurons are always positive, and follow an asymmetric distribution about their mean, it is clear that deviations above the mean have a different probability to occur than deviations by the same amount below the mean. One may attempt to capture the relative likelihood of different firing rates above and below the mean by computing a $z$ score obtained by dividing the difference between the mean response to each stimulus and the overall mean response by the standard deviation of the response to that stimulus. The greater the number of standard deviations (i.e. the greater the $z$ score) from the mean response value, the greater the information might be expected to be. We therefore show in Fig. C.11 the relation between the $z$ score and $I(s,R)$. (The $z$ score was calculated by
The stimulus-specific information $I(s; R)$ available in the response of the same single neuron as in Fig. C.4 about each of the stimuli in the set of 20 face stimuli (abscissa), with the firing rate of the neuron to the corresponding stimulus plotted as a function of this on the ordinate. The horizontal line shows the mean firing rate across all stimuli. (Reproduced from Journal of Computational Neuroscience, 4: 309–333. Information in the neuronal representation of individual stimuli in the primate temporal visual cortex, Rolls, E. T., Treves, A., Tovee, M. and Panzeri, S. Copyright ©1997, Kluwer Academic Publishers, with permission of Springer.)

Obtaining the mean and standard deviation of the response of a neuron to a particular stimulus $s$, and dividing the difference of this response from the mean response to all stimuli by the calculated standard deviation for that stimulus.) This results in a C-shaped curve in Figs. C.10 and C.11, with more information being provided by the cell the further its response to a stimulus is in spikes per second or in $z$ scores either above or below the mean response to all stimuli (which was 54 spikes/s). The specific C-shape is discussed further in Section C.3.4.

The information $I(s; R)$ about each stimulus in the set of 65 stimuli is shown in Fig. C.12 for the same neuron, am242. The 23 face stimuli in the set are indicated by a diamond, and the 42 non-face stimuli by a cross. Using this much larger and more varied stimulus set, which is more representative of stimuli in the real world, a C-shaped function again describes the relation between the information conveyed by the cell about a stimulus and its firing rate to that stimulus. In particular, this neuron reflected information about most, but not all, of the faces in the set, that is those faces that produced a higher firing rate than the overall mean firing rate to all the 65 stimuli, which was 31 spikes/s. In addition, it conveyed information about the majority of the 42 non-face stimuli by responding at a rate below the overall mean response of the neuron to the 65 stimuli. This analysis usefully makes the point that the information available in the neuronal responses about which stimulus was shown is relative to (dependent upon) the nature and range of stimuli in the test set of stimuli.

This evidence makes it clear that a single cortical visual neuron tuned to faces conveys information not just about one face, but about a whole set of faces, with the information
conveyed on a single trial related to the difference in the firing rate response to a particular stimulus compared to the average response to all stimuli.

The analyses just described for neurons with visual responses are general, in that they apply in a very similar way to olfactory neurons recorded in the macaque orbitofrontal cortex (Rolls, Critchley and Treves 1996a, Rolls, Critchley, Verhagen and Kadohisa 2010a).

The neurons in this sample reflected in their firing rates for the post-stimulus period 100 to 600 ms on average 0.36 bits of mutual information about which of 20 face stimuli was presented (Rolls, Treves, Tovee and Panzeri 1997d). Similar values have been found in other experiments (Tovee, Rolls, Treves and Bellis 1993, Tovee and Rolls 1995, Rolls, Tovee and Panzeri 1999b, Rolls, Franco, Aggelopoulos and Jerez 2006b). The information in short temporal epochs of the neuronal responses is described in Section C.3.4.

C.3.3 The information from single neurons: temporal codes versus rate codes within the spike train of a single neuron

In the third of a series of papers that analyze the response of single neurons in the primate inferior temporal cortex to a set of static visual stimuli, Optican and Richmond (1987) applied information theory in a particularly direct and useful way. To ascertain the relevance of stimulus-locked temporal modulations in the firing of those neurons, they compared the amount of information about the stimuli that could be extracted from just the firing rate, computed over a relatively long interval of 384 ms, with the amount of information that could be extracted from a more complete description of the firing, that included temporal
modulation. To derive this latter description (the temporal code within the spike train of a single neuron) they applied principal component analysis (PCA) to the temporal response vectors recorded for each neuron on each trial. The PCA helped to reduce the dimensionality of the neuronal response measurements. A temporal response vector was defined as a vector with as components the firing rates in each of 64 successive 6 ms time bins. The \((64 \times 64)\) covariance matrix was calculated across all trials of a particular neuron, and diagonalized. The first few eigenvectors of the matrix, those with the largest eigenvalues, are the principal components of the response, and the weights of each response vector on these four to five components can be used as a reduced description of the response, which still preserves, unlike the single value giving the mean firing rate along the entire interval, the main features of the temporal modulation within the interval. Thus a four- to five-dimensional temporal code could be contrasted with a one-dimensional rate code, and the comparison made quantitative by measuring the respective values for the mutual information with the stimuli.

Although the initial claim (Optican, Gawne, Richmond and Joseph 1991, Eskandar, Richmond and Optican 1992), that the temporal code carried nearly three times as much information as the rate code, was later found to be an artefact of limited sampling, and more recent analyses tend to minimize the additional information in the temporal description (Tovee, Rolls, Treves and Bellis 1993, Heller, Hertz, Kjaer and Richmond 1995), this type of application has immediately appeared straightforward and important, and it has led to many developments.
By concentrating on the code expressed in the output rather than on the characterization of the neuronal channel itself, this approach is not affected much by the potential complexities of the preceding black box. Limited sampling, on the other hand, is a problem, particularly because it affects much more codes with a larger number of components, for example the four to five components of the PCA temporal description, than the one-dimensional firing rate code. This is made evident in the paper by Heller, Hertz, Kjaer and Richmond (1995), in which the comparison is extended to several more detailed temporal descriptions, including a binary vector description in which the presence or not of a spike in each 1 ms bin of the response constitutes a component of a 320-dimensional vector. Obviously, this binary vector must contain at least all the information present in the reduced descriptions, whereas in the results of Heller, Hertz, Kjaer and Richmond (1995), despite the use of a sophisticated neural network procedure to control limited sampling biases, the binary vector appears to be the code that carries the least information of all. In practice, with the data samples available in the experiments that have been done, and even when using analytic procedures to control limited sampling (Panzeri and Treves 1996), reliable comparison can be made only with up to two- to three-dimensional codes.

Tovee, Rolls, Treves and Bellis (1993) and Tovee and Rolls (1995) obtained further evidence that little information was encoded in the temporal aspects of firing within the spike train of a single neuron in the inferior temporal cortex by taking short epochs of the firing of neurons, lasting 20 ms or 50 ms, in which the opportunity for temporal encoding would be limited (because there were few spikes in these short time intervals). They found that a considerable proportion (30%) of the information available in a long time period of 400 ms utilizing temporal encoding within the spike train was available in time periods as short as 20 ms when only the number of spikes was taken into account.

Overall, the main result of these analyses applied to the responses to static stimuli in the temporal visual cortex of primates is that not much more information (perhaps only up to 10% more) can be extracted from temporal codes than from the firing rate measured over a judiciously chosen interval (Tovee, Rolls, Treves and Bellis 1993, Heller, Hertz, Kjaer and Richmond 1995). Indeed, it turns out that even this small amount of ‘temporal information’ is related primarily to the onset latency of the neuronal responses to different stimuli, rather than to anything more subtle (Tovee, Rolls, Treves and Bellis 1993). Consistent with this point, in earlier visual areas the additional ‘temporally encoded’ fraction of information can be larger, due especially to the increased relevance, earlier on, of precisely locked transient responses (Kjaer, Hertz and Richmond 1994, Golomb, Kleinfeld, Reid, Shapley and Shraiman 1994, Heller, Hertz, Kjaer and Richmond 1995). This is because if the responses to some stimuli are more transient and to others more sustained, this will result in more information if the temporal modulation of the response of the neuron is taken into account. However, the relevance of more substantial temporal codes for static visual stimuli remains to be demonstrated. For non-static visual stimuli and for other cortical systems, similar analyses have largely yet to be carried out, although clearly one expects to find much more prominent temporal effects e.g. in the auditory system (Nelken, Prut, Vaadia and Abeles 1994, deCharms and Merzenich 1996), for reasons similar to those just annunciated.

C.3.4 The information from single neurons: the speed of information transfer

It is intuitive that if short periods of firing of single cells are considered, there is less time for temporal modulation effects. The information conveyed about stimuli by the firing rate and that conveyed by more detailed temporal codes become similar in value. When the firing periods analyzed become shorter than roughly the mean interspike interval, even the statistics
of firing rate values on individual trials cease to be relevant, and the information content of the firing depends solely on the mean firing rates across all trials with each stimulus. This is expressed mathematically by considering the amount of information provided as a function of the length \(t\) of the time window over which firing is analyzed, and taking the limit for \(t \to 0\) (Skaggs, McNaughton, Gothard and Markus 1993, Panzeri, Biella, Rolls, Skaggs and Treves 1996). To first order in \(t\), only two responses can occur in a short window of length \(t\): either the emission of an action potential, with probability \(tr_s\), where \(r_s\) is the mean firing rate calculated over many trials using the same window and stimulus; or no action potential, with probability \(1 - tr_s\). Inserting these conditional probabilities into equation C.22, taking the limit and dividing by \(t\), one obtains for the derivative of the stimulus-specific transinformation

\[
dI(s)/dt = r_s \log_2(r_s/\langle r \rangle) + (\langle r \rangle - r_s)/\ln 2,
\]

where \(\langle r \rangle\) is the grand mean rate across stimuli. This formula thus gives the rate, in bits/s, at which information about a stimulus begins to accumulate when the firing of a cell is recorded. Such an information rate depends only on the mean firing rate to that stimulus and on the grand mean rate across stimuli. As a function of \(r_s\), it follows the U-shaped curve in Fig. C.13. The curve is universal, in the sense that it applies irrespective of the detailed firing statistics of the cell, and it expresses the fact that the emission or not of a spike in a short window conveys information in as much as the mean response to a given stimulus is above or below the overall mean rate. No information is conveyed about those stimuli the mean response to which is the same as the overall mean. In practice, although the curve describes only the universal behaviour of the initial slope of the specific information as a function of time, it approximates well the full stimulus-specific information \(I(s;R)\) computed even over rather long periods (Rolls, Critchley and Treves 1996a, Rolls, Treves, Tovee and Panzeri 1997d).

Averaging equation C.47 across stimuli one obtains the time derivative of the mutual information. Further dividing by the overall mean rate yields the adimensional quantity

\[
\chi = \sum_s P(s) (r_s/\langle r \rangle) \log_2(r_s/\langle r \rangle)
\]

which measures, in bits, the mutual information per spike provided by the cell (Bialek, Rieke, de Ruyter van Steveninck and Warland 1991, Skaggs, McNaughton, Gothard and Markus 1993). One can prove that this quantity can range from 0 to \(\log_2(1/a)\)

\[
0 < \chi < \log_2(1/a),
\]

where \(a\) is the single neuron sparseness \(a^*\) defined in Section C.3.1.1. For mean rates \(r_s\) distributed in a nearly binary fashion, \(\chi\) is close to its upper limit \(\log_2(1/a)\), whereas for
mean rates that are nearly uniform, or at least unimodally distributed, \( \chi \) is relatively close to zero (Panzeri, Biella, Rolls, Skaggs and Treves 1996). In practice, whenever a large number of more or less ‘ecological’ stimuli are considered, mean rates are not distributed in arbitrary ways, but rather tend to follow stereotyped distributions (which for some neurons approximate an exponential distribution of firing rates – see Section C.3.1 (Treves, Panzeri, Rolls, Booth and Wakeman 1999, Baddeley, Abbott, Booth, Sengpiel, Freeman, Wakeman and Rolls 1997, Rolls and Treves 1998, Rolls and Deco 2002, Franco, Rolls, Aggelopoulos and Jerez 2007, Rolls 2008d, Rolls and Treves 2011)), and as a consequence \( \chi \) and \( a \) (or, equivalently, its logarithm) tend to covary (rather than to be independent variables (Skaggs and McNaughton 1992)). Therefore, measuring sparseness is in practice nearly equivalent to measuring information per spike, and the rate of rise in mutual information, \( \chi < r > \), is largely determined by the sparseness \( a \) and the overall mean firing rate \( < r > \).

The important point to note about the single-cell information rate \( \chi < r > \) is that, to the extent that different cells express non-redundant codes, as discussed below, the instantaneous information flow across a population of \( C \) cells can be taken to be simply \( C \chi < r > \), and this quantity can easily be measured directly without major limited sampling biases, or else inferred indirectly through measurements of the sparseness \( a \). Values for the information rate \( \chi < r > \) that have been published range from 2–3 bits/s for rat hippocampal cells (Skaggs, McNaughton, Gothard and Markus 1993), to 10–30 bits/s for primate temporal cortex visual cells (Rolls, Treves and Tovee 1997b), and could be compared with analogous measurements in the sensory systems of frogs and crickets, in the 100–300 bits/s range (Rieke, Warland and Bialek 1993).

If the first time-derivative of the mutual information measures information flow, successive derivatives characterize, at the single-cell level, different firing modes. This is because whereas the first derivative is universal and depends only on the mean firing rates to each stimulus, the next derivatives depend also on the variability of the firing rate around its mean value, across trials, and take different forms in different firing regimes. Thus they can serve as a measure of discrimination among firing regimes with limited variability, for which, for example, the second derivative is large and positive, and firing regimes with large variability, for which the second derivative is large and negative. Poisson firing, in which in every short period of time there is a fixed probability of emitting a spike irrespective of previous firing, is an example of large variability, and the second derivative of the mutual information can be calculated to be

\[
\frac{d^2 I}{dt^2} = \ln a + (1 - a) \frac{< r >^2}{a \ln 2},
\]

where \( a \) is the single neuron sparseness \( a^s \) defined in Section C.3.1.1. This quantity is always negative. Strictly periodic firing is an example of zero variability, and in fact the second time-derivative of the mutual information becomes infinitely large in this case (although actual information values measured in a short time interval remain of course finite even for exactly periodic firing, because there is still some variability, \( \pm 1 \), in the number of spikes recorded in the interval). Measures of mutual information from short intervals of firing of temporal cortex visual cells have revealed a degree of variability intermediate between that of periodic and of Poisson regimes (Rolls, Treves, Tovee and Panzeri 1997d). Similar measures can also be used to contrast the effect of the graded nature of neuronal responses, once they are analyzed over a finite period of time, with the information content that would characterize neuronal activity if it reduced to a binary variable (Panzeri, Biella, Rolls, Skaggs and Treves 1996). A binary variable with the same degree of variability would convey information at the same instantaneous rate (the first derivative being universal), but in for example 20–30% reduced amounts when analyzed over times of the order of the interspike interval or longer.
Utilizing these approaches, Tovee, Rolls, Treves and Bellis (1993) and Tovee and Rolls (1995) measured the information available in short epochs of the firing of single neurons, and found that a considerable proportion of the information available in a long time period of 400 ms was available in time periods as short as 20 ms and 50 ms. For example, in periods of 20 ms, 30% of the information present in 400 ms using temporal encoding with the first three principal components was available. Moreover, the exact time when the epoch was taken was not crucial, with the main effect being that rather more information was available if information was measured near the start of the spike train, when the firing rate of the neuron tended to be highest (see Figs. C.14 and C.15). The conclusion was that much information was available when temporal encoding could not be used easily, that is in very short time epochs of 20 or 50 ms.

It is also useful to note from Figs. C.14, C.15 and 16.15 the typical time course of the responses of many temporal cortex visual neurons in the awake behaving primate. Although the firing rate and availability of information is highest in the first 50–100 ms of the neuronal response, the firing is overall well sustained in the 500 ms stimulus presentation period. Cortical neurons in the primate temporal lobe visual system, in the taste cortex (Rolls, Yaxley and Sienkiewicz 1990), and in the olfactory cortex (Rolls, Critchley and Treves 1996a), do not in general have rapidly adapting neuronal responses to sensory stimuli. This may be important for associative learning: the outputs of these sensory systems can be maintained for sufficiently long while the stimuli are present for synaptic modification to occur. Although rapid synaptic adaptation within a spike train is seen in some experiments in brain slices (Markram and Tsodyks 1996, Abbott, Varela, Sen and Nelson 1997), it is not a very marked effect in at least some brain systems in vivo, when they operate in normal physiological conditions with normal levels of acetylcholine, etc.
To pursue this issue of the speed of processing and information availability even further, Rolls, Tovee, Purcell, Stewart and Azzopardi (1994b) and Rolls and Tovee (1994) limited the period for which visual cortical neurons could respond by using backward masking. In this paradigm, a short (16 ms) presentation of the test stimulus (a face) was followed after a delay of 0, 20, 40, 60, etc. ms by a masking stimulus (which was a high contrast set of letters) (see Fig. C.16). They showed that the mask did actually interrupt the neuronal response, and that at the shortest interval between the stimulus and the mask (a delay of 0 ms, or a ‘Stimulus Onset Asynchrony’ of 20 ms), the neurons in the temporal cortical areas fired for approximately 30 ms (see Fig. C.17). Under these conditions, the subjects could identify which of five faces had been shown much better than chance. Interestingly, under these conditions, when the inferior temporal cortex neurons were firing for 30 ms, the subjects felt that they were guessing, and conscious perception was minimal (Rolls, Tovee, Purcell, Stewart and Azzopardi 1994b), the neurons conveyed on average 0.10 bits of information (Rolls, Tovee and Panzeri 1999b). With a stimulus onset asynchrony of 40 ms, when the inferior temporal cortex neurons were firing for 50 ms, not only did the subjects’ performance improve, but the stimuli were now perceived clearly, consciously, and the neurons conveyed on average 0.16 bits of information. This has contributed to the view that consciousness has a higher threshold of activity in a given pathway, in this case a pathway for face analysis, than does unconscious processing and performance using the same pathway (Rolls 2003, Rolls 2006a).

The issue of how rapidly information can be read from neurons is crucial and fundamental to understanding how rapidly memory systems in the brain could operate in terms of reading
Information theory results

Fig. C.16 Backward masking paradigm. The visual stimulus appeared at time 0 for 16 ms. The time between the start of the visual stimulus and the masking image is the Stimulus Onset Asynchrony (SOA). A visual fixation task was being performed to ensure correct fixation of the stimulus. In the fixation task, the fixation spot appeared in the middle of the screen at time +500 ms, was switched off 100 ms before the test stimulus was shown, and was switched on again at the end of the mask stimulus. Then when the fixation spot dimmed after a random time, fruit juice could be obtained by licking. No eye movements could be performed after the onset of the fixation spot. (After Rolls and Tovee 1994.)

In this context, Delorme and Thorpe (2001) have suggested that just one spike from each neuron is sufficient, and indeed it has been suggested that the order of the first spike in different neurons may be part of the code (Delorme and Thorpe 2001, Thorpe, Delorme and Van Rullen 2001, VanRullen, Guyonneau and Thorpe 2005). (Implicit in the spike order hypothesis is that the first spike is particularly important, for it would be difficult to measure the
Fig. C.17 Firing of a temporal cortex cell to a 20 ms presentation of a face stimulus when the face was followed with different stimulus onset asynchronies (SOAs) by a masking visual stimulus. At an SOA of 20 ms, when the mask immediately followed the face, the neuron fired for only approximately 30 ms, yet identification above change (by ‘guessing’) of the face at this SOA by human observers was possible. (After Rolls and Tovee 1994; and Rolls, Tovee, Purcell et al. 1994.)
order for anything other than the first spike.) An alternative view is that the number of spikes in a fixed time window over which a postsynaptic neuron could integrate information is more realistic, and this time might be in the order of 20 ms for a single receiving neuron, or much longer if the receiving neurons are connected by recurrent collateral associative synapses and so can integrate information over time (Deco and Rolls 2006, Rolls and Deco 2002, Panzeri, Rolls, Battaglia and Lavis 2001, Rolls 2008d). Although the number of spikes in a short time window of e.g. 20 ms is likely to be 0, 1, or 2, the information available may be more than that from the first spike alone, and Rolls, Franco, Aggelopoulos and Jerez (2006b) examined this by measuring neuronal activity in the inferior temporal visual cortex, and then applying quantitative information theoretic methods to measure the information transmitted by single spikes, and within short time windows.

The cumulative single cell information about which of the twenty stimuli (Fig. C.7) was shown from all spikes and from the first spike starting at 100 ms after stimulus onset is shown in Fig. C.18. A period of 100 ms is just longer than the shortest response latency of the neurons from which recordings were made, so starting the measure at this time provides the best chance for the single spike measurement to catch a spike that is related to the stimulus. The means and standard errors across the 21 different neurons are shown. The cumulated information from the total number of spikes is larger than that from the first spike, and this is evident and significant within 50 ms of the start of the time epoch. In calculating the information from the first spike, just the first spike in the analysis window starting in this case at 100 ms after stimulus onset was used.

Because any one neuron receiving information from the population being analyzed has multiple inputs, we show in Fig. C.19 the cumulative information that would be available from multiple cells (21) about which of the 20 stimuli was shown, taking both the first spike after the time of stimulus onset (0 ms), and the total number of spikes after 0 ms from each neuron. The cumulative information even from multiple cells is much greater when all the spikes rather than just the first spike are used.

An attractor network might be able to integrate the information arriving over a long time period of several hundred milliseconds (see Chapter 5.6), and might produce the advantage shown in Fig. C.19 for the whole spike train compared to the first spike only. However a
single layer pattern association network might only be able to integrate the information over
the time constants of its synapses and cell membrane, which might be in the order of 15–30
ms (Panzeri, Rolls, Battaglia and Lavis 2001, Rolls and Deco 2002) (see Section B.2). In a
hierarchical processing system such as the visual cortical areas, there may only be a short time
during which each stage may decode the information from the preceding stage, and then pass
on information sufficient to support recognition to the next stage (Rolls and Deco 2002) (see
Chapter 25). We therefore analyzed the information that would be available in short epochs
from multiple inputs to a neuron, and show the multiple cell information for the population
of 21 neurons in Fig. C.20 (for 20 ms and 50 epochs). We see in this case that the first spike
information, because it is being made available from many different neurons (in this case 21
selective neurons discriminating between the stimuli each with p<0.001 in an ANOVA), fares
better relative to the information from all the spikes in these short epochs, but is still less than
the information from all the spikes (particularly in the 50 ms epoch). In particular, for the
epoch starting 100 ms after stimulus onset in Fig. C.21 the information in the 20 ms epoch is
0.37 bits, and from the first spike is 0.24 bits. Correspondingly, for a 50 ms epoch, the values
in the epoch starting at 100 ms post stimulus were 0.66 bits for the 50 ms epoch, and 0.40
bits for the first spike. Thus with a population of neurons, having just one spike from each
can allow considerable information to be read if only a limited period (of e.g. 20 or 50 ms)
is available for the readout, though even in these cases, more information was available if all
the spikes in the short window are considered (Fig. C.20).

To show how the information increases with the number of neurons in the ensemble in
these short epochs, we show in Fig. C.21 the information from different numbers of neurons
for a 20 ms epoch starting at time = 100 ms with respect to stimulus onset, for both the first
spike condition and the condition with all the spikes in the 20 ms window. The linear increase
in the information in both cases indicates that the neurons provide independent information,
which could be because there is no redundancy or synergy, or because these cancel (Rolls,
is also clear from Fig. C.21 that even with the population of neurons, and with just a short time epoch of 20 ms, more information is available from the population if all the spikes in 20 ms are considered, and not just the first spike. The 20 ms epoch analyzed for Fig. C.21 is for the post-stimulus time period of 100–120 ms.

To assess whether there is information that is specifically related to the order in which the spikes arrive from the different neurons, Rolls, Franco, Aggelopoulos and Jerez (2006b) computed for every trial the order across the different simultaneously recorded neurons in which the first spike arrived to each stimulus, and used this in the information theoretic analysis. The control condition was to randomly allocate the order values for each trial between the neurons that had any spikes on that trial, thus shuffling or scrambling the order of the spike arrival times in the time window. In both cases, just the first spike in the time window was used in the information analysis. (In both the order and the shuffled control conditions, on some trials some neurons had no spikes, and this itself, in comparison with the fact that some neurons had spiked on that trial, provided some information about which stimulus had been shown. However, by explicitly shuffling in the control condition the order of the spikes for the neurons that had spiked on that trial, comparison of the control with the unshuffled order condition provides a clear measure of whether the order of spike arrival from the different neurons itself carries useful information about which stimulus was shown.)

The data set was 36 cells with significantly different (p<0.05) responses to the stimulus set where it was possible to record simultaneously from groups of 3 and 4 cells (so that the order on each trial could be measured) in 11 experiments. Taking a 75 ms time window starting 100 ms after stimulus onset, the information with the order of arrival times of the spikes was $0.142 \pm 0.02$ bits, and in the control (shuffled order) condition was $0.138 \pm 0.02$ bits (mean across the 11 experiments ± sem). Thus the information increase by taking into account the order of spike arrival times relative to the control condition was only $(0.142 - 0.138) = 0.004$ bits per experiment (which was not significant). For comparison, the information calculated for the first spike using the same dot product decoding as described above was $0.136 \pm 0.03$ bits per experiment. Analogous results were obtained for different time windows. Thus taking the spike order into account compared to a control condition in which the spike order was
scrambled made essentially no difference to the amount of information that was available from the populations of neurons about which stimulus was shown.

The results show that although considerable information is present in the first spike, more information is available under the more biologically realistic assumption that neurons integrate spikes over a short time window (depending on their time constants) of for example 20 ms. The results shown in Fig. C.21 are of considerable interest, for they show that even when one increases the number of neurons in the population, the information available from the number of spikes in a 20 ms time window is larger than the information available from just the first spike. Thus although intuitively one might think that one can compensate by taking a population of neurons rather than just a single neuron when using just the first spike instead of the number of spikes available in a fixed time window, this compensation by increasing neuron numbers is insufficient to make the first spike code as efficient as taking the number of spikes.

Further, in this first empirical test of the hypothesis that there is information that is specifically related to the order in which the spikes arrive from the different neurons, which has been proposed by Thorpe et al (Delorme and Thorpe 2001, Thorpe, Delorme and Van Rullen 2001, VanRullen, Guyonneau and Thorpe 2005), we found that in the inferior temporal visual cortex there was no significant evidence that the order of the spike arrival times from different simultaneously recorded neurons is important. Indeed, the evidence found in the experiments was that the number of spikes in the time window is the important property that is related to the amount of information encoded by the spike trains of simultaneously recorded neurons.
The fact that there was also more information in the number of spikes in a fixed time window than from the first spike only is also evidence that is not consistent with the spike order hypothesis, for the order between neurons can only be easily read from the first spike, and just using information from the first spike would discard extra information available from further spikes even in short time windows.

The encoding of information that uses the number of spikes in a short time window that is supported by the analyses described by Rolls, Franco, Aggelopoulos and Jerez (2006b) deserves further elaboration. It could be thought of as a rate code, in that the number of spikes in a short time window is relevant, but is not a rate code in the rather artificial sense considered by Thorpe et al. (Delorme and Thorpe 2001, Thorpe et al. 2001, VanRullen et al. 2005) in which a rate is estimated from the interspike interval. This is not just artificial, but also begs the question of how, once the rate is calculated from the interspike interval, this decoded rate is passed on to the receiving neurons, or how, if the receiving neurons calculate the interspike interval at every synapse, they utilize it. In contrast, the spike count code in a short time window that is considered here is very biologically plausible, in that each spike would inject current into the post-synaptic neuron, and the neuron would integrate all such currents in a dendrite over a time period set by the synaptic and membrane time constants, which will result in an integration time constant in the order of 15–20 ms. Explicit models of exactly this dynamical processing at the integrate-and-fire neuronal level have been described to define precisely these operations (Deco and Rolls 2003, Deco and Rolls 2005d, Deco, Rolls and Horwitz 2004, Deco and Rolls 2005b, Rolls and Deco 2002, Rolls 2008d). Even though the number of spikes in a short time window of e.g. 20 ms is likely to be 0, 1, or 2, it can be 3 or more for effective stimuli (Rolls, Franco, Aggelopoulos and Jerez 2006b), and this is more efficient than using the first spike.

To add some detail here, a neuron receiving information from a population of inferior temporal cortex neurons of the type described here would have a membrane potential that varied continuously in time reflecting with a time constant in the order of 15–20 ms (resulting from a time constant of order 10 ms for AMPA synapses, 100 ms for NMDA synapses, and 20 ms for the cell membrane) a dot (inner) product over all synapses of each spike count and the synaptic strength. This continuously time varying membrane potential would lead to spikes whenever the results of this integration process produced a depolarization that exceeded the firing threshold. The result is that the spike train of the neuron would reflect continuously with a time constant in the order of 15–20 ms the likelihood that the input spikes it was receiving matched its set of synaptic weights. The spike train would thus indicate in continuous time how closely the stimulus or input matched its most effective stimulus (for a dot product is essentially a correlation). In this sense, no particular starting time is needed for the analysis, and in this respect it is a much better component of a dynamical system than is a decoding that utilizes an order in which the order of the spike arrival times is important and a start time for the analysis must be assumed.

I note that an autoassociation or attractor network implemented by recurrent collateral connections between the neurons can, using its short-term memory, integrate its inputs over much longer periods, for example over 500 ms in a model of how decisions are made (Deco and Rolls 2006) (see Chapter 5.6), and thus if there is time, the extra information available in more than the first spike or even the first few spikes that is evident in Figs. C.18 and C.19 could be used by the brain.

The conclusions from the single cell information analyses are thus that most of the information is encoded in the spike count; that large parts of this information are available in short temporal epochs of e.g. 20 ms or 50 ms; and that any additional information which appears to be temporally encoded is related to the latency of the neuronal response, and reflects sudden changes in the visual stimuli. Therefore a neuron in the next cortical area would obtain
considerable information within 20–50 ms by measuring the firing rate of a single neuron. Moreover, if it took a short sample of the firing rate of many neurons in the preceding area, then very much information is made available in a short time, as shown above and in Section C.3.5.

C.3.5 The information from multiple cells: independent information versus redundancy across cells

The rate at which a single cell provides information translates into an instantaneous information flow across a population (with a simple multiplication by the number of cells) only to the extent that different cells provide different (independent) information. To verify whether this condition holds, one cannot extend to multiple cells the simplified formula for the first time-derivative, because it is made simple precisely by the assumption of independence between spikes, and one cannot even measure directly the full information provided by multiple (more than two to three) cells, because of the limited sampling problem discussed above. Therefore one has to analyze the degree of independence (or conversely of redundancy) either directly among pairs – at most triplets – of cells, or indirectly by using decoding procedures to transform population responses. Obviously, the results of the analysis will vary a great deal with the particular neural system considered and the particular set of stimuli, or in general of neuronal correlates, used. For many systems, before undertaking to quantify the analysis in terms of information measures, it takes only a simple qualitative description of the responses to realize that there is a lot of redundancy and very little diversity in the responses. For example, if one selects pain-responsive cells in the somatosensory system and uses painful electrical stimulation of different intensities, most of the recorded cells are likely to convey pretty much the same information, signalling the intensity of the stimulation with the intensity of their single-cell response. Therefore, an analysis of redundancy makes sense only for a neuronal system that functions to represent, and enable discriminations between, a large variety of stimuli, and only when using a set of stimuli representative, in some sense, of that large variety.

Rolls, Treves and Tovee (1997b) measured the information available from a population of inferior temporal cortex neurons using the decoding method described in Section C.2.3, and found that the information increased approximately linearly, as shown in Fig. 25.13 on page 570, and in Fig. C.22 for a 50 ms interval as well as for a 500 ms measuring period. (It is shown below that the increase is limited only by the information ceiling of 4.32 bits necessary to encode the 20 stimuli. If it were not for this approach to the ceiling, the increase would be approximately linear (Rolls, Treves and Tovee 1997b).) To the extent that the information increases linearly with the number of neurons, the neurons convey independent information, and there is no redundancy, at least with numbers of neurons in this range. Although these and some of the other results described in this Appendix are for face-selective neurons in the inferior temporal visual cortex, similar results were obtained for neurons responding to objects in the inferior temporal visual cortex (Booth and Rolls 1998), and for neurons responding to spatial view in the hippocampus (Rolls, Treves, Robertson, Georges-François and Panzeri 1998b).

Although those neurons were not simultaneously recorded, a similar approximately linear increase in the information from simultaneously recorded cells as the number of neurons in the sample increased also occurs (Rolls, Franco, Aggelopoulos and Reece 2003b, Rolls, Aggelopoulos, Franco and Treves 2004, Franco, Rolls, Aggelopoulos and Treves 2004, Aggelopoulos, Franco and Rolls 2005, Rolls, Franco, Aggelopoulos and Jerez 2006b). These findings imply little redundancy, and that the number of stimuli that can be encoded increases
approximate exponentially with the number of neurons in the population, as illustrated in Figs. 25.14 and C.22.

The issue of redundancy is considered in more detail now. Redundancy can be defined with reference to a multiple channel of capacity $T(C)$ which can be decomposed into $C$ separate channels of capacities $T_i, i = 1, ..., C$:

$$R = 1 - T(C)/\sum_i T_i$$  \hspace{1cm} (C.51)

so that when the $C$ channels are multiplexed with maximal efficiency, $T(C) = \sum_i T_i$ and $R = 0$. What is measured more easily, in practice, is the redundancy defined with reference to a specific source (the set of stimuli with their probabilities). Then in terms of mutual information

$$R' = 1 - I(C)/\sum_i I_i.$$  \hspace{1cm} (C.52)

Gawne and Richmond (1993) measured the redundancy $R'$ among pairs of nearby primate inferior temporal cortex visual neurons, in their response to a set of 32 Walsh patterns. They found values with a mean $< R' > = 0.1$ (and a mean single-cell transinformation of 0.23 bits). Since to discriminate 32 different patterns takes 5 bits of information, in principle one would need at least 22 cells each providing 0.23 bits of strictly orthogonal information to represent the full entropy of the stimulus set. Gawne and Richmond reasoned, however, that, because of the overlap, $y$, in the information they provided, more cells would be needed than if the redundancy had been zero. They constructed a simple model based on the notion that the overlap, $y$, in the information provided by any two cells in the population always corresponds to the average redundancy measured for nearby pairs. A redundancy $R' = 0.1$
corresponds to an overlap $y = 0.2$ in the information provided by the two neurons, since, counting the overlapping information only once, two cells would yield 1.8 times the amount transmitted by one cell alone. If a fraction of $1 - y = 0.8$ of the information provided by a cell is novel with respect to that provided by another cell, a fraction $(1 - y)^2$ of the information provided by a third cell will be novel with respect to what was known from the first pair, and so on, yielding an estimate of $I(C) = I(1) \sum_{i=0}^{C-1} (1 - y)^i$ for the total information conveyed by $C$ cells. However such a sum saturates, in the limit of an infinite number of cells, at the level $I(\infty) = I(1)/y$, implying in their case that even with very many cells, no more than 0.23/0.2 = 1.15 bits could be read off their activity, or less than a quarter of what was available as entropy in the stimulus set! Gawne and Richmond (1993) concluded, therefore, that the average overlap among non-nearby cells must be considerably lower than that measured for cells close to each other.

The model above is simple and attractive, but experimental verification of the actual scaling of redundancy with the number of cells entails collecting the responses of several cells interspersed in a population of interest. Gochin, Colombo, Dorfman, Gerstein and Gross (1994) recorded from up to 58 cells in the primate temporal visual cortex, using sets of two to five visual stimuli, and applied decoding procedures to measure the information content in the population response. The recordings were not simultaneous, but comparison with simultaneous recordings from a smaller number of cells indicated that the effect of recording the individual responses on separate trials was minor. The results were expressed in terms of the novelty $N$ in the information provided by $C$ cells, which being defined as the ratio of such information to $C$ times the average single-cell information, can be expressed as

$$N = 1 - R'$$

and is thus the complement of the redundancy. An analysis of two different data sets, which included three information measures per data set, indicated a behaviour $N(C) \approx 1/\sqrt{C}$, reminiscent of the improvement in the overall noise-to-signal ratio characterizing $C$ independent processes contributing to the same signal. The analysis neglected however to consider limited sampling effects, and more seriously it neglected to consider saturation effects due to the information content approaching its ceiling, given by the entropy of the stimulus set. Since this ceiling was quite low, for 5 stimuli at $\log_2 5 = 2.32$ bits, relative to the mutual information values measured from the population (an average of 0.26 bits, or 1/9 of the ceiling, was provided by single cells), it is conceivable that the novelty would have taken much larger values if larger stimulus sets had been used.

A simple formula describing the approach to the ceiling, and thus the saturation of information values as they come close to the entropy of the stimulus set, can be derived from a natural extension of the Gawne and Richmond (1993) model. In this extension, the information provided by single cells, measured as a fraction of the ceiling, is taken to coincide with the average overlap among pairs of randomly selected, not necessarily nearby, cells from the population. The actual value measured by Gawne and Richmond would have been, again, $1/22 = 0.045$, below the overlap among nearby cells, $y = 0.2$. The assumption that $y$, measured across any pair of cells, would have been as low as the fraction of information provided by single cells is equivalent to conceiving of single cells as ‘covering’ a random portion $y$ of information space, and thus of randomly selected pairs of cells as overlapping in a fraction $(y)^2$ of that space, and so on, as postulated by the Gawne and Richmond (1993) model, for higher numbers of cells. The approach to the ceiling is then described by the formula

$$I(C) \approx H\{1 - \exp[C \ln(1 - y)]\}$$
that is, a simple exponential saturation to the ceiling. This simple law indeed describes remarkably well the trend in the data analyzed by Rolls, Treves and Tovee (1997b). Although the model has no reason to be exact, and therefore its agreement with the data should not be expected to be accurate, the crucial point it embodies is that deviations from a purely linear increase in information with the number of cells analyzed are due solely to the ceiling effect. Aside from the ceiling, due to the sampling of an information space of finite entropy, the information contents of different cells’ responses are independent of each other. Thus, in the model, the observed redundancy (or indeed the overlap) is purely a consequence of the finite size of the stimulus set. If the population were probed with larger and larger sets of stimuli, or more precisely with sets of increasing entropy, and the amount of information conveyed by single cells were to remain approximately the same, then the fraction of space ‘covered’ by each cell, again $y$, would get smaller and smaller, tending to eliminate redundancy for very large stimulus entropies (and a fixed number of cells). The actual data were obtained with limited numbers of stimuli, and therefore cannot probe directly the conditions in which redundancy might reduce to zero. The data are consistent, however, with the hypothesis embodied in the simple model, as shown also by the near exponential approach to lower ceilings found for information values calculated with reduced subsets of the original set of stimuli (Rolls, Treves and Tovee 1997b).

The implication of this set of analyses, some performed towards the end of the ventral visual stream of the monkey, is that the representation of at least some classes of objects in those areas is achieved with minimal redundancy by cells that are allocated each to analyse a different aspect of the visual stimulus. This minimal redundancy is what would be expected of a self-organizing system in which different cells acquired their response selectivities through a random process, with or without local competition among nearby cells (see Section B.4). At the same time, such low redundancy could also very well result in a system that is organized under some strong teaching input, so that the emerging picture is compatible with a simple random process, but could be produced in other ways. The finding that, at least with small numbers of neurons, redundancy may be effectively minimized, is consistent not only with the concept of efficient encoding, but also with the general idea that one of the functions of the early visual system is to progressively minimize redundancy in the representation of visual stimuli (Attneave 1954, Barlow 1961). However, the ventral visual system does much more than produce a non-redundant representation of an image, for it transforms the representation from an image to an invariant representation of objects, as described in Chapter 25. Moreover, what is shown in this section is that the information about objects can be read off from just the spike count of a population of neurons, using decoding as simple as the simplest that could be performed by a receiving neuron, dot product decoding. In this sense, the information about objects is made explicit in the firing rate of the neurons in the inferior temporal cortex, in that it can be read off in this way.

We consider in Section C.3.7 whether there is more to it than this. Does the synchronization of neurons (and it would have to be stimulus-dependent synchronization) add significantly to the information that could be encoded by the number of spikes, as has been suggested by some?

Before this, we consider why encoding by a population of neurons is more powerful than the encoding than is possible by single neurons, adding to previous arguments that a distributed representation is much more computationally useful than a local representation, by allowing properties such as generalization, completion, and graceful degradation in associative neuronal networks (see Appendix B).
C.3.6 Should one neuron be as discriminative as the whole organism, in object encoding systems?

In the analysis of random dot motion with a given level of correlation among the moving dots, single neurons in area MT in the dorsal visual system of the primate can be approximately as sensitive or discriminative as the psychophysical performance of the whole animal (Zohary, Shadlen and Newsome 1994). The arguments and evidence presented here (e.g. in Section C.3.5) suggest that this is not the case for the ventral visual system, concerned with object identification. Why should there be this difference?

Rolls and Treves (1998) suggest that the dimensionality of what is being computed may account for the difference. In the case of visual motion (at least in the study referred to), the problem was effectively one-dimensional, in that the direction of motion of the stimulus along a line in 2D space was extracted from the activity of the neurons. In this low-dimensional stimulus space, the neurons may each perform one of a few similar computations on a particular (local) portion of 2D space, with the side effect that, by averaging over a larger receptive field than in V1, one can extract a signal of a more global nature. Indeed, in the case of more global motion, it is the average of the neuronal activity that can be computed by the larger receptive fields of MT neurons that specifies the average or global direction of motion.

In contrast, in the higher dimensional space of objects, in which there are very many different objects to represent as being different from each other, and in a system that is not concerned with location in visual space but on the contrary tends to be relatively invariant with respect to location, the goal of the representation is to reflect the many aspects of the input information in a way that enables many different objects to be represented, in what is effectively a very high dimensional space. This is achieved by allocating cells, each with an intrinsically limited discriminative power, to sample as thoroughly as possible the many dimensions of the space. Thus the system is geared to use efficiently the parallel computations of all its neurons precisely for tasks such as that of face discrimination, which was used as an experimental probe. Moreover, object representation must be kept higher dimensional, in that it may have to be decoded by dot product decoders in associative memories, in which the input patterns must be in a space that is as high-dimensional as possible (i.e. the activity on different input axons should not be too highly correlated). In this situation, each neuron should act somewhat independently of its neighbours, so that each provides its own separate contribution that adds together with that of the other neurons (in a linear manner, see above and Figs. 25.13, C.22 and 25.14) to provide in toto sufficient information to specify which out of perhaps several thousand visual stimuli was seen. The computation involves in this case not an average of neuronal activity (which would be useful for e.g. head direction (Robertson, Rolls, Georges-François and Panzeri 1999)), but instead comparing the dot product of the activity of the population of neurons with a previously learned vector, stored in, for example, associative memories as the weight vector on a receiving neuron or neurons.

Zohary, Shadlen and Newsome (1994) put forward another argument which suggested to them that the brain could hardly benefit from taking into account the activity of more than a very limited number of neurons. The argument was based on their measurement of a small (0.12) correlation between the activity of simultaneously recorded neurons in area MT. They suggested that there would because of this be decreasing signal-to-noise ratio advantages as more neurons were included in the population, and that this would limit the number of neurons that it would be useful to decode to approximately 100. However, a measure of correlations in the activity of different neurons depends entirely on the way the space of neuronal activity is sampled, that is on the task chosen to probe the system. Among face cells in the temporal cortex, for example, much higher correlations would be observed when the task is a simple two-way discrimination between a face and a non-face, than when the task involves finer
identification of several different faces. (It is also entirely possible that some face cells could be found that perform as well in a given particular face / non-face discrimination as the whole animal.) Moreover, their argument depends on the type of decoding of the activity of the population that is envisaged (see further Robertson, Rolls, Georges-François and Panzeri (1999)). It implies that the average of the neuronal activity must be estimated accurately. If a set of neurons uses dot product decoding, and then the activity of the decoding population is scaled or normalized by some negative feedback through inhibitory interneurons, then the effect of such correlated firing in the sending population is reduced, for the decoding effectively measures the relative firing of the different neurons in the population to be decoded. This is equivalent to measuring the angle between the current vector formed by the population of neurons firing, and a previously learned vector, stored in synaptic weights. Thus, with for example this biologically plausible decoding, it is not clear whether the correlation Zohary, Shadlen and Newsome (1994) describe would place a severe limit on the ability of the brain to utilize the information available in a population of neurons.

The main conclusion from this and the preceding Section is that the information available from a set or ensemble of temporal cortex visual neurons increases approximately linearly as more neurons are added to the sample. This is powerful evidence that distributed encoding is used by the brain; and the code can be read just by knowing the firing rates in a short time of the population of neurons. The fact that the code can be read off from the firing rates, and by a principle as simple and neuron-like as dot product decoding, provides strong support for the general approach taken in this book to brain function.

It is possible that more information would be available in the relative time of occurrence of the spikes, either within the spike train of a single neuron, or between the spike trains of different neurons, and it is to this that we now turn.

### C.3.7 The information from multiple cells: the effects of cross-correlations between cells

Using the second derivative methods described in Section C.2.5 (see Rolls, Franco, Aggelopoulos and Reece (2003b)), the information available from the number of spikes vs that from the cross-correlations between simultaneously recorded cells has been analyzed for a population of neurons in the inferior temporal visual cortex (Rolls, Aggelopoulos, Franco and Treves 2004). The stimuli were a set of 20 objects, faces, and scenes presented while the monkey performed a visual discrimination task. If synchronization was being used to bind the parts of each object into the correct spatial relationship to other parts, this might be expected to be revealed by stimulus-dependent cross-correlations in the firing of simultaneously recorded groups of 2–4 cells using multiple single-neuron microelectrodes.

A typical result from the information analysis described in Section C.2.5 on a set of three simultaneously recorded cells from this experiment is shown in Fig. C.23. This shows that most of the information available in a 100 ms time period was available in the rates, and that there was little contribution to the information from stimulus-dependent (‘noise’) correlations (which would have shown as positive values if for example there was stimulus-dependent synchronization of the neuronal responses); or from stimulus-independent ‘noise’ correlation effects, which might if present have reflected common input to the different neurons so that their responses tended to be correlated independently of which stimulus was shown.

The results for the 20 experiments with groups of 2–4 simultaneously recorded inferior temporal cortex neurons are shown in Table C.4. (The total information is the total from equations C.43 and C.44 in a 100 ms time window, and is not expected to be the sum of the contributions shown in Table C.4 because only the information from the cross terms (for $i \neq j$) is shown in the table for the contributions related to the stimulus-dependent contributions and
Fig. C.23 A typical result from the information analysis described in Section C.2.5 on a set of 3 simultaneously recorded inferior temporal cortex neurons in an experiment in which 20 complex stimuli effective for IT neurons (objects, faces and scenes) were shown. The graphs show the contributions to the information from the different terms in equations C.43 and C.44 on page 837, as a function of the length of the time window, which started 100 ms after stimulus onset, which is when IT neurons start to respond. The rate information is the sum of the term in equation C.43 and the first term of equation C.44. The contribution of the stimulus-independent noise correlation to the information is the second term of equation C.44, and is separated into components arising from the correlations between cells (the cross component, for \( i \neq j \)) and from the autocorrelation within a cell (the auto component, for \( i = j \)). This term is non-zero if there is some correlation in the variance to a given stimulus, even if it is independent of which stimulus is present. The contribution of the stimulus-dependent noise correlation to the information is the third term of equation C.44, and only the cross term is shown (for \( i \neq j \)), as this is the term of interest. (After Rolls, Aggelopoulos, Franco and Treves 2004.)

Table C.4 The average contributions (in bits) of different components of equations C.43 and C.44 to the information available in a 100 ms time window from 13 sets of simultaneously recorded inferior temporal cortex neurons when shown 20 stimuli effective for the cells.

<table>
<thead>
<tr>
<th>Component</th>
<th>Average Contribution (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rate</td>
<td>0.26</td>
</tr>
<tr>
<td>stimulus-dependent “noise” correlation-related, cross term</td>
<td>0.04</td>
</tr>
<tr>
<td>stimulus-independent “noise” correlation-related, cross term</td>
<td>-0.05</td>
</tr>
<tr>
<td>total information</td>
<td>0.31</td>
</tr>
</tbody>
</table>

The stimulus-independent contributions arising from the ‘noise’ correlations.) The results show that the greatest contribution to the information is that from the rates, that is from the numbers of spikes from each neuron in the time window of 100 ms. The average value of \(-0.05\) for the cross term of the stimulus independent ‘noise’ correlation-related contribution is consistent with on average a small amount of common input to neurons in the inferior temporal visual cortex. A positive value for the cross term of the stimulus-dependent ‘noise’ correlation related contribution would be consistent with on average a small amount of stimulus-dependent synchronization, but the actual value found, 0.04 bits, is so small that for 17 of the 20 experiments it is less than that which can arise by chance statistical fluctuations of the time of arrival of the spikes, as shown by MonteCarlo control rearrangements of the same data. Thus on average there was no significant contribution to the information from stimulus-dependent synchronization effects (Rolls, Aggelopoulos, Franco and Treves 2004).

Thus, this data set provides evidence for considerable information available from the
number of spikes that each cell produces to different stimuli, and evidence for little impact of common input, or of synchronization, on the amount of information provided by sets of simultaneously recorded inferior temporal cortex neurons. Further supporting data for the inferior temporal visual cortex are provided by Rolls, Franco, Aggelopoulos and Reece (2003b). In that parts as well as whole objects are represented in the inferior temporal cortex (Perrett, Rolls and Caan 1982), and in that the parts must be bound together in the correct spatial configuration for the inferior temporal cortex neurons to respond (Rolls, Tovee, Purcell, Stewart and Azzopardi 1994b), we might have expected temporal synchrony, if used to implement feature binding, to have been evident in these experiments.

We have also explored neuronal encoding under natural scene conditions in a task in which top-down attention must be used, a visual search task. We applied the decoding information theoretic method of Section C.2.4 to the responses of neurons in the inferior temporal visual cortex recorded under conditions in which feature binding is likely to be needed, that is when the monkey had to choose to touch one of two simultaneously presented objects, with the stimuli presented in a complex natural background (Aggelopoulos, Franco and Rolls 2005). The investigation is thus directly relevant to whether stimulus-dependent synchrony contributes to encoding under natural conditions, and when an attentional task was being performed. In the attentional task, the monkey had to find one of two objects and to touch it to obtain reward. This is thus an object-based attentional visual search task, where the top-down bias is for the object that has to be found in the scene (Aggelopoulos, Franco and Rolls 2005). The objects could be presented against a complex natural scene background. Neurons in the inferior temporal visual cortex respond in some cases to object features or parts, and in other cases to whole objects provided that the parts are in the correct spatial configuration (Perrett, Rolls and Caan 1982, Desimone, Albright, Gross and Bruce 1984, Rolls, Tovee, Purcell, Stewart and Azzopardi 1994b, Tanaka 1996), and so it is very appropriate to measure whether stimulus-dependent synchrony contributes to information encoding in the inferior temporal visual cortex when two objects are present in the visual field, and when they must be segmented from the background in a natural visual scene, which are the conditions in which it has been postulated that stimulus-dependent synchrony would be useful (Singer 1999, Singer 2000).

Aggelopoulos, Franco and Rolls (2005) found that between 99% and 94% of the information was present in the firing rates of inferior temporal cortex neurons, and less that 5% in any stimulus-dependent synchrony that was present, as illustrated in Fig. C.24. The implication of these results is that any stimulus-dependent synchrony that is present is not quantitatively important as measured by information theoretic analyses under natural scene conditions. This has been found for the inferior temporal visual cortex, a brain region where features are put together to form representations of objects (Rolls 2008d) (Chapter 25), where attention has strong effects, at least in scenes with blank backgrounds (Rolls, Aggelopoulos and Zheng 2003a), and in an object-based attentional search task.

The finding as assessed by information theoretic methods of the importance of firing rates and not stimulus-dependent synchrony is consistent with previous information theoretic approaches (Rolls, Franco, Aggelopoulos and Reece 2003b, Rolls, Aggelopoulos, Franco and Treves 2004, Franco, Rolls, Aggelopoulos and Treves 2004). It would of course also be of interest to test the same hypothesis in earlier visual areas, such as V4, with quantitative, information theoretic, techniques. In connection with rate codes, it should be noted that the findings indicate that the number of spikes that arrive in a given time is what is important for very useful amounts of information to be made available from a population of neurons; and that this time can be very short, as little as 20–50 ms (Tovee and Rolls 1995, Rolls and Tovee 1994, Rolls, Tovee and Panzeri 1999b, Rolls and Deco 2002, Rolls, Tovee, Purcell, Stewart and Azzopardi 1994b, Rolls 2003, Rolls, Franco, Aggelopoulos and Jerez 2006b).
Fig. C.24 Left: the objects against the plain background, and in a natural scene. Right: the information available from the firing rates (Rate Inf) or from stimulus-dependent synchrony (Cross-Corr Inf) from populations of simultaneously recorded inferior temporal cortex neurons about which stimulus had been presented in a complex natural scene. The total information (Total Inf) is that available from both the rate and the stimulus-dependent synchrony, which do not necessarily contribute independently. Bottom: eye position recordings and spiking activity from two neurons on a single trial of the task. (Neuron 31 tended to fire more when the macaque looked at one of the stimuli, S-, and neuron 21 tended to fire more when the macaque looked at the other stimulus, S+. Both stimuli were within the receptive field of the neuron.) (After Aggelopoulos, Franco and Rolls 2005.)

Further, it was shown that there was little redundancy (less than 6%) between the information
Information theory results

provided by the spike counts of the simultaneously recorded neurons, making spike counts an efficient population code with a high encoding capacity.

The findings (Aggelopoulos, Franco and Rolls 2005) are consistent with the hypothesis that feature binding is implemented by neurons that respond to features in the correct relative spatial locations (Rolls and Deco 2002, Elliffe, Rolls and Stringer 2002, Rolls 2008d, Rolls 2012c) (Chapter 25), and not by temporal synchrony and attention (Malsburg 1990a, Singer, Gray, Engel, Konig, Artola and Brocher 1990, Abeles 1991, Hummel and Biederman 1992, Singer and Gray 1995, Singer 1999, Singer 2000) (Section 8.3). In any case, the computational point made in Section 25.5.5.1 is that even if stimulus-dependent synchrony was useful for grouping, it would not without much extra machinery be useful for binding the relative spatial positions of features within an object, or for that matter of the positions of objects in a scene which appears to be encoded in a different way (Aggelopoulos and Rolls 2005) (see Section 25.5.10).

So far, we know of no analyses that have shown with information theoretic methods that considerable amounts of information are available about the stimulus from the stimulus-dependent correlations between the responses of neurons in the primate ventral visual system. The use of such methods is needed to test quantitatively the hypothesis that stimulus-dependent synchronization contributes substantially to the encoding of information by neurons.

C.3.8 Conclusions on cortical neuronal encoding

The conclusions emerging from this set of information theoretic analyses, many in cortical areas towards the end of the ventral visual stream of the monkey, and others in the hippocampus for spatial view cells (Rolls, Treves, Robertson, Georges-François and Panzeri 1998b), in the presubiculum for head direction cells (Robertson, Rolls, Georges-François and Panzeri 1999), and in the orbitofrontal cortex and related areas for olfactory and taste cells (Rolls, Critchley and Treves 1996a, Rolls, Critchley, Verhagen and Kadohisa 2010a) for which subsequent analyses have shown a linear increase in information with the number of cells in the population, are as follows (see also Rolls and Treves (2011)).

The representation of at least some classes of objects in those areas is achieved with minimal redundancy by cells that are allocated each to analyze a different aspect of the visual stimulus (Abbott, Rolls and Tovee 1996, Rolls, Treves and Tovee 1997b) (as shown in Sections C.3.5 and C.3.7). This minimal redundancy is what would be expected of a self-organizing system in which different cells acquired their response selectivities through processes that include some randomness in the initial connectivity, and local competition among nearby cells (see Appendix B). Towards the end of the ventral visual stream redundancy may thus be effectively minimized, a finding consistent with the general idea that one of the functions of the early visual system is indeed that of progressively minimizing redundancy in the representation of visual stimuli (Attneave 1954, Barlow 1961). Indeed, the evidence described in Sections C.3.5, C.3.7 and C.3.4 shows that the exponential rise in the number of stimuli that can be decoded when the firing rates of different numbers of neurons are analyzed indicates that the encoding of information using firing rates (in practice the number of spikes emitted by each of a large population of neurons in a short time period) is a very powerful coding scheme used by the cerebral cortex, and that the information carried by different neurons is close to independent provided that the number of stimuli being considered is sufficiently large.

Quantitatively, the encoding of information using firing rates (in practice the number of spikes emitted by each of a large population of neurons in a short time period) is likely to be far more important than temporal encoding, in terms of the number of stimuli that can be encoded. Moreover, the information available from an ensemble of cortical neurons when only the firing rates are read, that is with no temporal encoding within or between neurons, is
made available very rapidly (see Figs. C.14 and C.15 and Section C.3.4). Further, the neuronal responses in most ventral or ‘what’ processing streams of behaving monkeys show sustained firing rate differences to different stimuli (see for example Fig. 16.15 for visual representations, for the olfactory pathways Rolls, Critchley and Treves (1996a), for spatial view cells in the hippocampus Rolls, Treves, Robertson, Georges-François and Panzeri (1998b), and for head direction cells in the presubiculum Robertson, Rolls, Georges-François and Panzeri (1999)), so that it may not usually be necessary to invoke temporal encoding for the information about the stimulus. Further, as indicated in Section C.3.7, information theoretic approaches have enabled the information that is available from the firing rate and from the relative time of firing (synchronization) of inferior temporal cortex neurons to be directly compared with the same metric, and most of the information appears to be encoded in the numbers of spikes emitted by a population of cells in a short time period, rather than by the temporal synchronization of the responses of different neurons when certain stimuli appear (see Section C.3.7 and Aggelopoulos, Franco and Rolls (2005)).

Information theoretic approaches have also enabled different types of readout or decoding that could be performed by the brain of the information available in the responses of cell populations to be compared (Rolls, Treves and Tovee 1997b, Robertson, Rolls, Georges-François and Panzeri 1999). It has been shown for example that the multiple cell representation of information used by the brain in the inferior temporal visual cortex (Rolls, Treves and Tovee 1997b, Aggelopoulos, Franco and Rolls 2005), olfactory cortex (Rolls, Critchley and Treves 1996a), hippocampus (Rolls, Treves, Robertson, Georges-François and Panzeri 1998b), and presubiculum (Robertson, Rolls, Georges-François and Panzeri 1999) can be read fairly efficiently by the neuronally plausible dot product decoding, and that the representation has all the desirable properties of generalization and graceful degradation, as well as exponential coding capacity (see Sections C.3.5 and C.3.7).

Information theoretic approaches have also enabled the information available about different aspects of stimuli to be directly compared. For example, it has been shown that inferior temporal cortex neurons make explicit much more information about what stimulus has been shown rather than where the stimulus is in the visual field (Tovee, Rolls and Azzopardi 1994), and this is part of the evidence that inferior temporal cortex neurons provide translation invariant representations. In a similar way, information theoretic analysis has provided clear evidence that view invariant representations of objects and faces are present in the inferior temporal visual cortex, in that for example much information is available about what object has been shown from any single trial on which any view of any object is presented (Booth and Rolls 1998).

Information theory has also helped to elucidate the way in which the inferior temporal visual cortex provides a representation of objects and faces, in which information about which object or face is shown is made explicit in the firing of the neurons in such a way that the information can be read off very simply by memory systems such as the orbitofrontal cortex, amygdala, and perirhinal cortex / hippocampal systems. The information can be read off using dot product decoding, that is by using a synaptically weighted sum of inputs from inferior temporal cortex neurons (see further Section 24.2.6 and Chapter 25). Moreover, information theory has helped to show that for many neurons considerable invariance in the representations of objects and faces are shown by inferior temporal cortex neurons (e.g. Booth and Rolls (1998)). Examples of some of the types of objects and faces that are encoded in this way are shown in Fig. C.7. Information theory has also helped to show that inferior temporal cortex neurons maintain their object selectivity even when the objects are presented in complex natural backgrounds (Aggelopoulos, Franco and Rolls 2005) (see further Chapter 25 and Section 24.2.6).

Information theory has also enabled the information available in neuronal representations
Information theory results

Finally, information theory also provides a metric for directly comparing the information available from neurons in the brain (see Chapter 25 and this Appendix) with that available from single neurons and populations of neurons in simulations of visual information processing (see Chapter 25).

In summary, the evidence from the application of information theoretic and related approaches to how information is encoded in the visual, hippocampal, and olfactory cortical systems described during behaviour leads to the following working hypotheses:

1. Much information is available about the stimulus presented in the number of spikes emitted by single neurons in a fixed time period, the firing rate.

2. Much of this firing rate information is available in short periods, with a considerable proportion available in as little as 20 ms. This rapid availability of information enables the next stage of processing to read the information quickly, and thus for multistage processing to operate rapidly. This time is the order of time over which a receiving neuron might be able to utilize the information, given its synaptic and membrane time constants. In this time, a sending neuron is most likely to emit 0, 1, or 2 spikes.

3. This rapid availability of information is confirmed by population analyses, which indicate that across a population on neurons, much information is available in short time periods.

4. More information is available using this rate code in a short period (of e.g. 20 ms) than from just the first spike.

5. Little information is available by time variations within the spike train of individual neurons for static visual stimuli (in periods of several hundred milliseconds), apart from a small amount of information from the onset latency of the neuronal response. A static stimulus encompasses what might be seen in a single visual fixation, what might be tasted with a stimulus in the mouth, what might be smelled in a single sniff, etc. For a time-varying stimulus, clearly the firing rate will vary as a function of time.

6. Across a population of neurons, the firing rate information provided by each neuron tends to be independent; that is, the information increases approximately linearly with the number of neurons. This applies of course only when there is a large amount of information to be encoded, that is with a large number of stimuli. The outcome is that the number of stimuli that can be encoded rises exponentially in the number of neurons in the ensemble. (For a small stimulus set, the information saturates gradually as the amount of information available from the neuronal population approaches that required to code for the stimulus set.) This applies up to the number of neurons tested and the stimulus set sizes used, but as the number of neurons becomes very large, this is likely to hold less well. An implication of the independence is that the response profiles to a set of stimuli of different neurons are uncorrelated.

7. The information in the firing rate across a population of neurons can be read moderately efficiently by a decoding procedure as simple as a dot product. This is the simplest type of processing that might be performed by a neuron, as it involves taking a dot product of the incoming firing rates with the receiving synaptic weights to obtain the activation (e.g. depolarization) of the neuron. This type of information encoding ensures that the simple emergent properties of associative neuronal networks such as generalization, completion, and graceful
degradation (see Appendix B) can be realized very naturally and simply.

8. There is little additional information to the great deal available in the firing rates from any stimulus-dependent cross-correlations or synchronization that may be present. Stimulus-dependent synchronization might in any case only be useful for grouping different neuronal populations, and would not easily provide a solution to the binding problem in vision. Instead, the binding problem in vision may be solved by the presence of neurons that respond to combinations of features in a given spatial position with respect to each other.

9. There is little information available in the order of the spike arrival times of different neurons for different stimuli that is separate or additional to that provided by a rate code. The presence of spontaneous activity in cortical neurons facilitates rapid neuronal responses, because some neurons are close to threshold at any given time, but this also would make a spike order code difficult to implement.

10. Analysis of the responses of single neurons to measure the sparseness of the representation indicates that the representation is distributed, and not grandmother cell like (or local). Moreover, the nature of the distributed representation, that it can be read by dot product decoding, allows simple emergent properties of associative neuronal networks such as generalization, completion, and graceful degradation (see Appendix B) to be realized very naturally and simply. The evidence becoming available from humans is consistent with this summary (Fried, Rutishauser, Cerf and Kreiman 2014, Rolls 2015d, Rolls 2017a).

11. The representation is not very sparse in the perceptual systems studied (as shown for example by the values of the single cell sparseness \( a_s \)), and this may allow much information to be represented. At the same time, the responses of different neurons to a set of stimuli are decorrelated, in the sense that the correlations between the response profiles of different neurons to a set of stimuli are low. Consistent with this, the neurons convey independent information, at least up to reasonable numbers of neurons. The representation may be more sparse in memory systems such as the hippocampus, and this may help to maximize the number of memories that can be stored in associative networks.

12. The nature of the distributed representation can be understood further by the firing rate probability distribution, which has a long tail with low probabilities of high firing rates. The firing rate probability distributions for some neurons fit an exponential distribution, and for others there are too few very low rates for a good fit to the exponential distribution. An implication of an exponential distribution is that this maximizes the entropy of the neuronal responses for a given mean firing rate under some conditions. It is of interest that in the inferior temporal visual cortex, the firing rate probability distribution is very close to exponential if a large number of neurons are included without scaling of the firing rates of each neuron. An implication is that a receiving neuron would see an exponential firing rate probability distribution.

13. The population sparseness \( a_p \), that is the sparseness of the firing of a population of neurons to a given stimulus (or at one time), is the important measure for setting the capacity of associative neuronal networks. In populations of neurons studied in the inferior temporal cortex, hippocampus, and orbitofrontal cortex, it takes the same value as the single cell sparseness \( a_s \), and this is a situation of weak ergodicity that occurs if the response profiles of the different neurons to a set of stimuli are uncorrelated.
Understanding the neuronal code, the subject of this Appendix, is fundamental for understanding how memory and related perceptual systems in the brain operate, as follows:

Understanding the neuronal code helps to clarify what neuronal operations would be useful in memory and in fact in most mammalian brain systems (e.g. dot product decoding, that is taking a sum in a short time of the incoming firing rates weighted by the synaptic weights).

It clarifies how rapidly memory and perceptual systems in the brain could operate, in terms of how long it takes a receiving neuron to read the code.

It helps to confirm how the properties of those memory systems in terms of generalization, completion, and graceful degradation occur, in that the representation is in the correct form for these properties to be realized.

Understanding the neuronal code also provides evidence essential for understanding the storage capacity of memory systems, and the representational capacity of perceptual systems.

Understanding the neuronal code is also important for interpreting functional neuroimaging, for it shows that functional imaging that reflects incoming firing rates and thus currents injected into neurons, and probably not stimulus-dependent synchronization, is likely to lead to useful interpretations of the underlying neuronal activity and processing. Of course, functional neuroimaging cannot address the details of the representation of information in the brain in the way that is essential for understanding how neuronal networks in the brain could operate, for this level of understanding (in terms of all the properties and working hypotheses described above) comes only from an understanding of how single neurons and populations of neurons encode information.

C.4 Information theory terms – a short glossary

1. The amount of information, or surprise, in the occurrence of an event (or symbol) \( s_i \) of probability \( P(s_i) \) is

\[
I(s_i) = \log_2(1/P(s_i)) = -\log_2 P(s_i).
\]

(C.55)

(The measure is in bits if logs to the base 2 are used.) This is also the amount of uncertainty removed by the occurrence of the event.

2. The average amount of information per source symbol over the whole alphabet \( S \) of symbols \( s_i \) is the entropy,

\[
H(S) = -\sum_i P(s_i) \log_2 P(s_i)
\]

(C.56)

(or a priori entropy).

3. The probability of the pair of symbols \( s \) and \( s' \) is denoted \( P(s, s') \), and is \( P(s)P(s') \) only when the two symbols are independent.

4. Bayes theorem (given the output \( s' \), what was the input \( s \) ?) states that

\[
P(s|s') = \frac{P(s'|s)P(s)}{P(s')}
\]

(C.57)

where \( P(s'|s) \) is the forward conditional probability (given the input \( s \), what will be the output \( s' \) ?), and \( P(s|s') \) is the backward (or posterior) conditional probability (given the
output $s'$, what was the input $s$?). The prior probability is $P(s)$.

5. **Mutual information.** Prior to reception of $s'$, the probability of the input symbol $s$ was $P(s)$. This is the *a priori* probability of $s$. After reception of $s'$, the probability that the input symbol was $s$ becomes $P(s|s')$, the conditional probability that $s$ was sent given that $s'$ was received. This is the *a posteriori* probability of $s$. The difference between the *a priori* and *a posteriori* uncertainties measures the gain of information due to the reception of $s'$. Once averaged across the values of both symbols $s$ and $s'$, this is the **mutual information**, or transinformation

$$I(S, S') = \sum_{s, s'} P(s, s') \left\{ \log_2 \left[ \frac{1}{P(s)} \right] - \log_2 \left[ \frac{1}{P(s|s')} \right] \right\}$$

(C.58)

$$= \sum_{s, s'} P(s, s') \log_2 \left[ \frac{P(s|s')}{P(s)} \right].$$

Alternatively,

$$I(S, S') = H(S) - H(S|S').$$

(C.59)

$H(S|S')$ is sometimes called the **equivocation** (of $S$ with respect to $S'$).

**C.5 Highlights**

1. The encoding of information by neuronal responses is described using Shannon information theory.
2. Information is encoded by sparse distributed place coded representations, with almost independent information conveyed by each neuron, up to reasonable numbers of neurons.
3. Place cell encoding conveys much more information than stimulus-dependent synchronicity (‘oscillations’, coherence) of firing of groups of neurons in the awake behaving animal.
References


References


References


References


Fernhouteau, J., Holsinger, R. M. and McDonald, R. J. (1999). Lesions of the medial or lateral perforant path have different effects on hippocampal contributions to place learning and on fear conditioning to context, *Behavioral Brain Research* **101**: 65–84.
Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells,
References


Gabbott, P. and Rolls, E. T. (2013). Increased neuronal firing in resting and sleep in areas of the macaque medial prefrontal cortex (mPFC) that are part of the default mode network, European Journal of Neuroscience 37: 1737–1746.


References

with different rewards, *Cerebral Cortex* **10**: 263–271.


References


References


Markov, N. T., Ercsey-Ravasz, M., Lamy, C., Ribeiro Gomes, A. R., Magrou, L., Misery, P., Giroud, P., Barone, P.,


References


Naber, P. A., Lopes da Silva, F. H. and Witter, M. P. (2001). Reciprocal connections between the entorhinal cortex and hippocampal fields CA1 and the subiculum are in register with the projections from CA1 to the subiculum, Hippocampus 11: 99–104.
Nakazawa, K., McHugh, T. J., Wilson, M. A. and Tonegawa, S. (2004). NMDA receptors, place cells and hippocampal


Passingham, R. E. P. and Wise, S. P. (2012). *The Neurobiology of the Prefrontal Cortex*, Oxford University Press,
References

Oxford.


Quinlan, P. T. and Humphreys, G. W. (1987). Visual search for targets defined by combinations of color, shape, and
size: An examination of the task constraints on feature and conjunction searches, *Perception and Psychophysics* 41: 455–472.


Progress in Neurobiology 86: 216–244.
Rolls, E. T., Burton, M. J. and Mora, F. (1976). Hypothalamic neuronal responses associated with the sight of food,
References


across the macaque visual system, Journal of Neurophysiology 79: 3272–3277.
across the macaque visual system, Journal of Neurophysiology 79: 3272–3277.
Schulte, S. and Singer, J. (1962). Cognitive, social and physiological determinants of emotional state, Psychological
Scheuerecker, J., Ufer, S., Zipse, M., Frodl, T., Koutsouleris, N., Zetsche, T., Wiesmann, M., Albrecht, J., Bruckmann,
Scheuerecker, J., Ufer, S., Zipse, M., Frodl, T., Koutsouleris, N., Zetsche, T., Wiesmann, M., Albrecht, J., Bruckmann,
Scheuerecker, J., Ufer, S., Zipse, M., Frodl, T., Koutsouleris, N., Zetsche, T., Wiesmann, M., Albrecht, J., Bruckmann,
References


References


References


Index

ΔF, 119–131
3D model, 578–580
3D object recognition, 631–636
acumulation of evidence, 98
accumulator model of decision-making, 132–134
acetylcholine, 230, 247–248, 339–344, 481, 508, 857
acetylcholine and memory, 542
action, 281–304, 432–439
  multiple routes to, 420–454
activation, 6
activation function, 6, 188, 725
active vision, 581
adaptation, 7, 229–241–254, 339, 341, 355, 481, 502,
  686, 771–772, 791, 857
adaptation model, 242
addiction, 286
ADHD, 338
advantages of cortical architecture, 413
aesthetics, 692
ageing, 230, 727
aging, 335–344
alignment approach to object recognition, 580–581
Alzheimer’s disease, 339–344, 481
amnesia, 479–486, 493, 550
retrograde, 470, 493
amygdala, 289, 848
and reversal, 300
AND, 700
ars memoriae, 531
art, 692
art of memory, 531
associative learning, 653, 706–734
associative reward–penalty algorithm, 808–810
asymmetric receptive fields, 651
attention, 141–185, 311, 642–652, 685
biased activation theory of, 171–181
biased competition, 145–147, 149–185
binding, 143
bottom-up, 141
computational models, 145–185
feature integration theory, 143
inferior temporal cortex, 147
neuroimaging, 149, 183
neurophysiology, 145–146
neuropsychology, 171–184
non-spatial, 147–149
object, 158–163
saliency, 141
temporal visual areas, 167–171
spatial, 158–163
V4, V2 and V1, 146–147
attention deficit hyperactivity disorder, ADHD, 338
attentional bias, 147
attractor basins
flow analysis, 308
attractor dynamics, 684
attractor network dynamics, 91–140, 773–780
attractor network simulations, 881–884
attractor networks, 684
diluted connectivity, 187–198, 207
discrete and continuous, 767
obsessive-compulsive disorder, 317–325
storage capacity, 80
auditory processing, 64–66
autism, 286
autoassociation network, 636–642, 723–734
autoassociation network simulations, 881–884
autobiographical memory, 442
autocorrelation memory, 726
autonomic responses, 420, 438, 451
backprojections, 85, 255–261, 297, 370, 421, 518,
  520–526, 586, 686, 734, 743–745
quantitative analysis, 523–526
backprojections and attention, 259–260
backprojections and constraint satisfaction, 261
backprojections and learning, 257–258
backprojections and memory recall, 258–259
backprojections and recall, 266
backpropagation of error network, 16, 799–804
backward masking, 450
basal forebrain, 339–344
basal ganglia, 406–415
basal ganglia computation, 408–415
basal magnocellular cholinergic neurons, 340–344
basal magnocellular forebrain nuclei of Meynert, 339
Bayes’ theorem, 879
biased activation hypothesis of attention, 171–185
biased activation theory of attention, 171–181
biased competition, 141–185, 642–652, 788–789
biased competition hypothesis of attention, 143–171
binding, 143, 221, 345–363, 448, 449, 559, 579–580,
  584, 612–624, 687, 875
binocular rivalry, 134
biologically plausible models of object recognition,
  663–688
biologically plausible networks, 16, 803–804
bipolar disorder, 333–335
blindsight, 421, 434
body motion cells, 477
BOLD signal, 122–131
Boltzmann machine, 806–807
brain vs digital computers, 676–680
Brodman areas, 24–26
CA1 neurons, 515–520
CA3 neurons, 490–515
capacity, 700
autoassociate, 730–731
competitive network, 738
pattern associator, 718–720, 722–723
catastrophic changes in the qualitative shape descrip-
tors, 631
categorization, 734, 737
caudate nucleus
head, 407
causality, 674
central tendency, 716, 728
cerebellar cortex, 455–462
cerebellar Purkinje cell, 460
cerebral cortex
principles of operation, 674–693
change blindness, 669–670
cholinergic neurons, 339–344
cingulate cortex, 55, 289
clamped inputs, 727
classical conditioning, 810
classification, 734
climbing fibres in the cerebellum, 457
cluster analysis, 741
cluttered environment, 624–631
course coding, 14
coding, 15, 209–226, 568–572, 677, 685, 715, 722,
814–880
distributed, 214–221
labelled line, 211
place, 211
cognition and attention, 176–178
Cognitron, 582
coherence, 221
columnar organization, 575
combinatorial explosion, 586, 624
common scale of reward, 287
communication through coherence, 221
competition, 269
competitive network, 509, 583, 734–756
competitive network simulations, 886–889
competitive networks
diluted connectivity, 201–207
completion, 537, 542, 725, 727
computers, 676–693
conditional expected value neurons, 63
conditional reward neurons, 63, 298
conditioning, 706–723
conductances, 769–773
confabulation, 137, 421–441
confidence, 115, 126–131
decisions about, 129–131
configural learning, 513
conflict, 437
conjunction search, 171
connectionism, 16
connectome, 255
consciousness, 136, 420–454, 857
access, 420
causal role, 445
phenomenal, 420
process, 420
threshold, 450
consolidation, 417, 520, 545
constraint satisfaction, 261
content, 443–445
content addressability, 676, 723
cortex, 731
continuous attractor, 131
continuous attractor networks, 479, 495–499, 530,
540, 659, 669, 756–767
continuous attractors, 495
continuous spatial transformation learning, 653–654
contrastive Hebbian learning, 581, 806–807
convergence, 583
convolution networks, 668, 804–806
correlation, 696
cortex, 107
decision-making, 305–344
noisy operation, 305–344
object processing before reward processing, 284–
286
principles of operation, 674–693
rodents, 380–381
stability, 305–344
stochastic dynamics, 91–140
cortical
backprojections, 520–526
cortical architecture
advantages, 413
disadvantages, 413
quantitative aspects, 32–39
cortical attractor networks, 75–90
cortical circuits, 34–39, 364–384, 689–693
cortical computations, 455–462
693
cortical dynamics, 684
cortical evolution, 364–384
cortical processing
Tiers 1–3, 284–286
cortical structure, 27–39, 364–379, 689–693
cost–benefit analysis, 433–436
coupled attractor networks, 86–90
creative thought, 137
credit assignment, 426, 447, 802
critical band masking, 561
critical period, 230, 269
cross-correlation, 871–875
cross-correlations
between neuronal spike trains, 830–838
cytoarchitecture, 377

Damasio’s theory of emotion, 451
Darwin, 445
deception, 436, 437
decision confidence
decisions about, 129–131
decision prediction, 138
decision time, 98–101
decision-making
accumulator or race models, 132–134
probabilistic, 106–131
synaptic facilitation, 253–254
decision-making between rewards, 291–292
decisions, 106–131
  about decision confidence, 129–131
decoding, 568, 826–830
Deep Belief Networks, 807
deep learning, 668, 804–806
default mode network, 416
deformation-invariant object recognition, 656–657
delayed match to sample, 300, 339
delta rule, 793
delta rule learning, 810–814
dentate granule cells, 265, 269, 506–515, 531
depression, 325–335, 428
  ketamine, 329
  theory of, 325–335
determinism, 136, 432, 441
deterministic behaviour, 136
difference of Gaussian (DOG) filtering, 592–594
difficult vs easy decisions, 112–131
diffusion process, 132–134, 291
digital computers vs brain computation, 676–680
diluted connectivity, 83, 142, 186–208, 277, 502–505, 526, 685
discoveries
cortical systems for emotion, 281–304
  face neurons, 211
  hippocampal spatial view neurons, 213
  neural encoding in the primate brain, 210–221
  properties of inferior temporal visual cortex neurons, 553–673
  reward systems in the orbitofrontal cortex, 212
  secondary taste cortex in the orbitofrontal cortex, 211
  sensory-specific satiety, 212
theories of schizophrenia, obsessive-compulsive disorder, depression, and normal aging, 305–344
theory of hippocampal operation, 463–553
theory of invariant visual object recognition, 553–673
  distractibility, 311
  distributed computation, 678
  distributed encoding, 586
    advantages, 15–16, 209–226
  divisive inhibition, 30
  dopamine, 301, 338
  and reward prediction error, 813
  dopamine and schizophrenia, 313–316
dopamine neurons
  error signal, 234–240
  reward, 234–240
dorsal visual pathway, 18–23
dorsal visual system, 46–48, 73, 275, 656, 659, 749
dot product, 576, 694–697, 703
  neuronal operation, 712
dot product decoding, 209–226, 829–830
dot-product decoding, 279
  dreaming, 416–419
  dreams, 418
drift-diffusion model of decision-making, 132–134
dual origin concept of cortex, 73
dual routes to action, 432–439
Duns Scotus, 135
dynamical systems properties, 680
dynamics, 767–791
economic value, 288–289
economics, 692
emotion, 281–304, 688
  action vs affect, 438
  and memory, 339–344
  definition, 281
  feelings, 420–454
  emotional feelings, 420–454
  encoding, 209–221, 568, 814–880
  encoding in humans, 223–226
energy
  in an autoassociator, 726
entorhinal cortex, 477, 488, 497–498, 510–513, 517
epilepsy, 680
episodic memory, 262, 463–553, 733, 767
ergodicity, 682, 783, 846–848
error correction, 791–804
error correction learning, 604–612, 810–814
error learning, 809–814
error neurons, 234–240, 292–297, 326, 328, 813
error signal, 809
  escaping time, 98–101
ethics, 692
evolution, 107, 364–405, 425, 447, 687
evolution of cortex, 73, 364–405, 689–693
evolution of emotion, 440
  evolutionary utility, 135
excitatory cortical neurons, 27–29
executive function, 435
expansion recording, 457, 721, 742
expected reward value, 810–814
expected utility, 234–240
expected value, 61, 62, 810
explicit processing, 136
explicit systems, 421–454
exponential firing rate distribution, 205, 842–852
expression, 572–576
  face, 572
  extinction, 171–184, 810
  face expression, 286, 434, 572–576
  face identity, 286, 555–576
  face processing
    neuromaging, 575
    facilitation, 686
false binding errors, 623–624
Fano factor, 92
fat, 52
fault tolerance, 676, 714, 727
fear, 451
  feature
    binding, 584
    combinations, 584
  feature analysis, 736
  feature analyzer, 741
feature binding, 612–624, 875
feature combinations, 272, 561, 582, 597, 612–624
feature hierarchies, 582–673
introduction, 582
feature hierarchy, 582, 586–673
feature integration theory, 143
feature spaces, 577–578
feature subsets and supersets, 613–617
features, 582
feed-forward cortical processing, 277
feedback inhibition, 32
feedback processing, 773–780
feedforward processing, 777–780
feelings
emotional, 420–454
finite size noise, 93
finite-size effects, 117–119
firing rate distribution, 209–226, 731, 839–850
exponential, 205, 839–850
firing rate encoding, 504
fixation, 275
flow, 95–101
fMRI, 122–131, 218
food texture, 52
foraging, 136
forgetting, 228–233, 493
forgetting as a principle of operation of the cortex,
264
free will, 136, 432, 441–443
frequency encoding, 209
Freud, 418
frontal lobes, 23, 382
functional magnetic resonance neuroimaging, 122–131

GABA, 410
GABA, gamma-aminobutyric acid, 29–32, 725
Gabor filters, 592, 593
gamma oscillations, 449–450
gene-defined goal, 281–283, 444
generalization, 621–624, 713, 727
generic view, 631
genes, 385
number, 40
genetic algorithm, 390, 749
genetic algorithms, 385
genetic specification
of brain connectivity, 385–405, 679
genetic specification of cortical connectivity, 687
global attractor network, 81
global motion, 46–48, 656
global workspace, 431, 452
Go/NoGo task, 298
goal for action, 281–283, 285, 444, 445
graceful degradation, 676, 714, 727
gradient of retrograde amnesia, 494
grammar, 345–363
grandmother cell, 14, 214, 568, 583
Granger causality, 174, 675
granule cells, 456–462, 487–515, 689–692
granule cells in the cerebellum, 457
grid cells, 477, 497–498, 510–513

habit learning, 428
head direction cells, 478–479, 497, 549, 758–764
Hebb rule, 8–13, 227–240, 707, 708, 717, 724, 729
heteroclinic channel, 684
Hick’s law, 131
hidden units, 803
hierarchical feature analysis networks for object recognition, 582–673
hierarchical organisation in the visual system, 272
hierarchical organization, 683
hierarchical processing, 40–71, 680, 745, 777–780
hierarchy, 68, 440, 583
feature, 586–673
higher order thought theories of consciousness, 447–449
higher-order thoughts, 420–454
hippocampal CA3 as a global attractor network, 263–264
hippocampal cortex
functional architecture, 374–376, 691–692
hippocampal simulation, 526
hippocampal theory
evaluation, 546–553
hippocampus, 23, 81, 269, 340, 421, 422, 461, 463–553, 742, 848
anatomy, 465–467
backprojections, 518, 520–526
CA1, 542–546
CA3, 535–542
CA3 neurons, 490–509
completion, 537
lesions, 467–468
mossy fibres, 206
neurophysiology, 470–477
object–spatial view neurons, 473
recall, 518
reward–spatial view neurons, 475
Rolls’ theory of, 463–553
speed cells, 477
subcortical connections, 267
tests of the theory, 531–546
theory, 486–528
hippocampus and memory, 262–268, 463–553
HMAX, 663–668
Hodgkin-Huxley equations, 769
Hopfield, 726
human cortical coding, 878
human neuroimaging, 218
hypothalamus
lateral, 339–344
hypotheses, invariant object recognition, 582–586
idiiothetic (self-motion) inputs, 761–764
imaging, 575
implicit processing, 136
implicit responses, 432–439
incentive motivation, 432
inferior temporal cortex
attention, 147
models, 553–673
subdivisions, 572–576
topology, 278, 746
inferior temporal visual cortex, 443, 450
learning, 565
information
continuous variables, 822–824
in the cross-correlation, 871–875
in the rate, 871–875
limited sampling problem, 824–830
multiple cell, 826–838, 866–879
mutual, 819, 880
neuronal, 824–879
neuronal synchronization, 830–838, 871–875
single neuron, 819–822
speed of transfer, 854–866
temporal encoding within a spike train, 838–854
information encoding, 209, 814–879
information from multiple neurons, 866–875
information theoretic analysis of neuronal responses, 838–879
information theory, 568, 814–880
inhibitory cortical neurons, 29–32
inhibitory neurons, 297, 413, 725
inner product, 694, 712
instrumental reinforcers, 296, 444
insula, 48–51
insular taste cortex, 48–51, 848
integrate-and-fire model, 149
integrate-and-fire neuronal networks, 93–106, 767–780
interacting attractor networks, 86–90
interests, 439
gene-defined, 439
of the individual defined by the reasoning system, 439
interference, 720
invariance, 272–280, 554–673
dendritic computation, 668
rotation, 584
scale, 584
size, 584
spatial frequency, 584
translation, 584, 621–624
view, 585, 631–636
invariance learning, 749–751
invariant object recognition, 680
approaches, 576–582
hypotheses, 582–586
models, 553–673
invariant representations, 555–565, 576
invariant visual object recognition, 689
invertebrate neural systems, 401
invertible networks, 581–582
James–Lange theory, 438, 451
just noticeable difference, 119
ketamine and depression, 329
Kohonen map, 745–749
labelled line, 211
labelled line encoding, 209
lamination, 689
language, 345–363, 421–454, 679
lateral hypothalamus, 339–344
lateral inhibition, 29, 30, 735, 745
lateral intraparietal cortex, LIP, 141, 275, 659
lateralization of function, 72, 467
learning, 227–240, 338
associative, 296–301, 706–723
in a cluttered environment, 629
instrumental, 296
learning in IT, 565
learning rule
local, 717, 729
learning set, 296–301
least mean squares rule, 793
LeDoux’s theory of emotion, 451
levels of explanation, 674
lighting invariance, 654–655
linear algebra, 694–705
linear separability, 698–704, 739
linearity, 698
linearly independent vectors, 698–704
local attractor network, 81
local field potential (LFP), 218
local learning rule, 9, 16, 717, 729
local representation, 14, 214, 722
localization of function, 72–74, 684
logical operations, 679
long-term depression, 460
long-term depression (LTD), 9–14, 228, 296, 300
long-term memory, 80, 106, 479–486
long-term potentiation (LTP), 9–14, 227–240, 708
LTD, 9–14, 227–240, 460
LTP, 9–14, 227–240
Machiavellian intelligence, 437
machine learning, 804
magnetoencephalography, 218
mania, 333–335
map, 575, 745–749
masking, 857–858
backward, 434, 450
matching law, 132
McGurk effect, 87–89, 745
mean field neurodynamics, 781–791
mean-field analysis of decision-making, 110–112
for sequences, 500
hippocampus, 463–553
object recognition, 553–673
perirhinal cortex, 479–486
recognition, 479–486
sequential order, 250
memory and sleep, 231, 416–419
memory capacity, 576
memory consolidation, 417
memory recall, 106
memory reconsolidation, 230–233
memory systems, 688
mind–body problem, 445–447, 674
mind–brain problem, 674
mind–brain relationship, 445–447, 674–676
mixture states, 732
model of adaptation, 242
model of reversal learning, 296–301
modular organization, 747
modularity, 32–39, 72–74, 680, 681
monitoring, 129–131, 352, 447–449, 452
mossy fibres, 506–508
motivation, 281–304
motor control, 406–415, 455–462, 766
MT, 749
multicompartiment neurons, 777
multilayer perceptron, 799–804
multimodal representations, 744
multimodal dynamics, 789–791
multiple cell information, 866–879
multiple decision-making systems, 432–441
multiple memory systems, 560
multiple objects, 623, 642–652
multiple objects in a scene, 651
multiple-choice decisions, 131
multiplicative interactions, 790
multistability, 98–101
mutual information, 880
myeloarchitecture, 377
natural scene, 624–631
natural scene object recognition, 659–662
natural scenes, 557–560, 614, 642–652
Necker cube, 134
negative reward prediction error, 292
neglect, 171
object-based, 171
Neocognitron, 582
neocortex, 18–39, 365–370, 461, 689–692
evolution, 376
net value, 288–289
neural encoding, 209–221, 568–572, 715
neuroeconomics, 288–289
neurogenesis, 206, 269–271, 514, 533, 681, 687
neuroimaging, 122–131, 149, 183, 567, 575
task difficulty, 126–129
neuronal adaptation, 241–254
neuronal encoding, 814–880
neuronal encoding in humans, 223–226
neuronal networks, 17–18, 706–814
neuronal recordings in animals, 218
neurons, 4–16, 27–39
neuropsychology of attention, 171–184
nightmares, 418
NMDA receptors, 10–14, 164, 296, 300, 335, 720, 753, 765
noise, 98–119, 678
definition, 93
finite size, 93
measurement, 93
noise and diluted connectivity, 504
noise in the brain
source, 91–93
noise reduction, 716, 728
noisy cortex, 91–140, 685
non-linear networks, 703–705
non-linearity, 698
in the learning rule, 753
non-reward, 282–301
non-reward and depression, 328–333
non-reward computation, 251–252
non-reward neurons, 292–297, 325–326
noradrenaline, 248, 338
noradrenergic neurons, 248
norepinephrine, 338
normalization
of neuronal activity, 30
of synaptic weights, 752
object identification, 575, 670
object motion, 46–48, 656
object recognition, 272–280, 553–673
biologically plausible models, 663–668
Rolls’ theory of, 553–673
object recognition in natural scenes, 659–662
object representation, 689
object representations, 484, 571, 838–879
object-based encoding, 562–565
obsessive-compulsive disorder, 316–325
occlusion, 564, 629–631
Oja learning rule, 753
olfactory processing, 57–64
olfactory receptor genes, 270
orbitofrontal cortex, 51–64, 251–252, 377–381, 408, 433, 848
evolution, 378–381
rodent, 377–384
orthogonalization, 509, 737, 742
oscillations, 221, 449
outcome, 810
output systems for emotion, 422–424, 432–439
palimpsest, 228
Panksepp’s theory of emotion, 452
parallel computation, 678
parallel distributed computation, 677, 678
parallel fibres in the cerebellum, 456
parallel processing, 167–171
Parkinson’s disease, 338
path integration, 479, 497–498, 541, 548, 761–766
pattern association, 526
pattern association memory, 706–723
pattern association network simulations, 884–886
pattern association networks
diluted connectivity, 198–201
pattern rivalry, 134
pattern separation, 265, 461, 505–515, 532–534, 739
Pavlovian conditioning, 810
percent correct, 117
perceptron, 460, 700, 791–804
perceptual learning, 565
perforant path, 508
perirhinal cortex, 479–486
personality, 302
place cells, 265, 470–478, 529
place cells vs spatial view cells, 528–529
place code, 211, 346
place coding, 209
planning, 422–443
plasticity, 269
pleasure, 281–304, 420–430
Poisson firing, 678
Poisson process, 92
population encoding, 214–218
population sparseness, 846–850
positive reinforcement, 429
predicted reward value, 810–814
predicting a decision, 138
prediction error, 292
prediction error hypothesis, 234–240, 813
preference relative, 62
prefrontal cortex, 23, 68, 150–163, 377
primary reinforcers, 282, 444, 445
priming, 259
principal component analysis (PCA), 741
probabilistic emotion-related decision-making, 292
prospective memory, 414
prototype extraction, 716, 728
psychiatric states, 305–344
qualia, 420, 428, 429, 451
radial basis function (RBF) networks, 751–752
rate coding, 209
ratio coding, 435–437
rationality, 423
raw feels, 420, 428
reaction time, 98–101, 167–171
reaction times, 116
reasoning, 423
recall, 258, 266, 425, 442, 447, 468, 520–526, 707, 709, 725
in autoassociation memories, 725
semantic, 520
recall of semantic representation, 520
receptive field size, 276, 583
receptive fields asymmetric, 651
recognition of objects, 553–673
recognition memory, 479–486
reconsolidation, 230–233
recurrent collateral connections, 107
recurrent collaterals, 75–90, 684
redundancy, 737, 741, 866–871
reinforcement learning, 804, 807–814
reinforcers, 282–304, 436, 444, 445
relative preference, 62
religion, 443
representation semantic, 520
representations, 443–445
of objects, 484, 571, 838–879
resting state, 416
retrograde amnesia, 468–470, 493, 494
retrospective memory, 414
reversal learning, 326
reversal learning set, 296–301
reversal learning, model of, 296–301
reward, 234–240
common scale, 287
reward devaluation, 287
reward magnitude, 810
reward outcome, 810
reward outcome value, 52
reward predicting neuron, 62
reward prediction error, 234–240, 297, 810–811
reward processing, 67–68
reward value, 52, 234–240, 288–289
Riesenhuber and Poggio model of invariant recognition, 663–668
rivalry, 134
Rolls’ biased activation theory of attention, 171–181
conjectures on free will, 432–443
conjectures on symbol grounding, 426–428, 443–445
discoveries on neuronal encoding, 209–226, 814–880
discovery of conditional reward neurons, 297–301
discovery of error neurons, 292–301
discovery of face cells, 42, 211, 225, 286, 556
discovery of fat texture neurons, 52–58
discovery of hippocampal spatial view cells, 470–478
discovery of olfactory reward neurons, 57–62
discovery of sensory-specific satiety, 52–58
discovery of taste reward neurons, 48–64
discovery of visual reward neurons, 62–64
hypotheses about the implementation of syntax in the cortex, 345–363
hypotheses of how the cortex is specified and evolves, 850–854
theory of consciousness, 420–454
theory of depression, 325–335
theory of dreams, 418–419
theory of emotion, 281–304
theory of hippocampal function, 463–553
theory of invariant global motion recognition in the dorsal visual system, 46–48, 656
theory of normal aging, 335–343
theory of obsessive-compulsive disorder, 316–325
theory of schizophrenia, 305–316
theory of visual object recognition, 553–673
Rolls’ discoveries on neuronal encoding, 210–221
Rolls’ discovery of face neurons, 211
Rolls’ discovery of hippocampal spatial view neurons, 213
Rolls’ discovery of sensory-specific satiety, 212
Rolls’ model of invariant recognition, 582–673
Rolls’ theory of invariant object recognition, 680
saccade, 275
saliency, 240
saliency attention, 141
satiety, 52–58, 379–381
scene representation, 657–659
schizophrenia, 305–316
search conjunction, 167
feature, 167
neuropsychology, 171–184
segregation of processing streams, 749
segregation of visual pathways
principles, 749
selection of action, 410
self-identity, 442
self-monitoring, 129–131
self-organization, 736
self-organization of cortical processing without a teacher,
273
self-organizing map, 575
selfish gene, 439–440
selfish individual, 439–440
selfish phenotype, 439–440
semantic memory, 470
semantic representation, 520
sensory-specific satiety, 52–58, 249, 287
sequence memory, 500, 516, 517, 732, 791
serial vs parallel processing, 167–171
shift invariance, 584
short-term memory, 68, 77–80, 86, 143–185, 250,
253, 297, 339, 414, 567, 706, 723, 733
for visual stimuli, 567
stability, 101–106
visual, 669
shunting inhibition, 30
Sigma–Pi neurons, 9, 615, 668, 704–705
sigmoid activation function, 205
signal-detection theory, 92
signal-to-noise ratio, 93
simulations: attractor networks, 881–884
simulations: autoassociation networks, 881–884
simulations: competitive networks, 886–889
simulations: pattern association networks, 884–886
simplification, 669–670
size constancy, 561
size invariance, 560
sleep, 416–419, 469
sleep and memory, 231, 416–419
slow learning, 663
soft competition, 753–754
soft max, 754
sparse distributed coding, 685
sparse distributed representation, 214
sparse distributed representations, 273
sparse representations, 209
sparseness, 14, 214, 568, 715, 720, 722, 738, 839–850
sparsification, 742
spatial continuity, 846–850
spatial frequency, 560
spatial view cells, 265, 470–478, 529
speed cells, 477
speed of operation, 83
speed of operation of memory systems, 499
speed of processing, 32, 716, 728, 767–780
spike response model, 780–781
spike timing-dependent plasticity (STDP), 13, 233
spin glass, 726
spontaneous firing
principles, 91–94, 729
spontaneous firing of neurons, 91
spontaneous firing rate, 32
stability
short-term memory, 101–106
stability of attractor networks in the cortex, 305–344
statistical fluctuations, 117–119
statistical mechanics, 726
STDP, 13, 233
stimulus–reinforcer association learning, 706–723
stimulus–reinforcer reversal learning, model of, 296–301
stochastic dynamics, 678, 685
striatum, 406–415
structural description, 578
structural shape descriptions, 578–580
subgenual cingulate cortex, 289
subliminal processing, 450
superficial vs deep cortical layers, 689
symbol grounding, 426, 443, 445
symbols, 421–454
symmetric synaptic weights, 724
symmetry-breaking, 135
synaptic adaptation, 241–254
synaptic facilitation, 241–254
synaptic learning, 227–240
synaptic modification, 8–14, 296–297, 300, 338, 686
synaptic weight vector, 707–723
synchronization, 613–615, 871–875
synchrony, 221, 345
syntactic binding, 613–615
syntactic pattern recognition, 578–580
syntax, 345–363, 421–454, 679, 684, 687
synthesis, 674–693
taste processing, 48–64
taste reward, 48–64, 379–381
template matching, 580–581
temporal difference (TD) error, 811–814
temporal difference (TD) learning, 604–612, 811–814
temporal difference learning, 607–612, 811–814
temporal encoding within a spike train, 838–854
temporal lobes, 381
temporal synchronization, 613, 871–875
texture, 52
thalamo-cortical adaptation, 244
thalamus, 374
functions of, 366, 374
theories of hippocampal function, 548–553
theory of emotion
Damasio, 451
James–Lange, 451
LeDoux, 451
Panksepp, 452
Rolls, 281–304
theory of mind, 423
time cells, 517–518
top-down processing, 255–261
topographic map, 745–749, 886–889
topographic maps, 85
topographical map, 42
topological maps, 278
trace learning rule, 584, 587–589, 604–612
trace rule value $\eta$, 600, 603–604
trajectory through a state space, 684
translation invariance, 584, 621–624

umami taste, 48–52
uncertainty, 234–240, 813
unclamped inputs, 727
unconditioned stimulus, 706–715
unconscious processing, 136
unifying principles, 48, 158, 159, 163, 182, 529, 552, 657, 671, 674–693
unpredictable behaviour, 136, 138
unsupervised learning, 586
unsupervised networks, 734–756
V1, 146
V2, 146
V4, 145–147, 749
value, 288–289
vector, 694, 707–723
angle, 696
correlation, 696
dot product, 694
length, 695
linear combination, 698–704
linear independence, 698–704
linear separability, 698–704
normalization, 696
normalized dot product, 696
outer product, 697
vector quantization, 737
vector similarity computation, 676–680
ventral cortical stream, 73
ventral striatum, 409, 572
ventral visual pathway, 18–23, 553–673
principles, 749
ventral visual stream, 749
ventral visual system, 41–46, 272–280, 285, 659
vestibular inputs, 477
view invariance, 585, 601–603, 631–636
view-based object recognition, 582–673
view-invariant representation, 362–565
viscosity, 52
vision, 553–673
VisNet, 529, 586–673, 680
3D transforms, 631–636
and spatial view cells, 529, 657
architecture, 587–594
attention, 642–652
attraction, 636–642
capacity, 636–642
cluttered environment, 624–631
continuous spatial transformation learning, 653
feature binding, 612–624
generalization, 621–624
lighting invariance, 654
motion, 46, 656
multiple objects, 642–652
natural scenes, 642–652
object motion, 46, 656
occlusion, 629–631
performance measures, 594–596
receptive field size, 642–652
trace learning rule, 587–589
trace rule, 604–612
trace rule value $\eta$, 600, 603–604
translation invariance, 601
view invariance, 601–603
visual object recognition, 272–280
visual system learning, 274
visuo-spatial scratchpad, 669–670
wake-sleep algorithm, 806
Weber’s law, 106–122
weight normalization, 752
what visual stream, 18–23, 553–673
where visual stream, 18–23
Widrow–Hoff rule, 793
winner-take-all, 583, 736
wiring length, 575, 747
word, 520
XOR, 700