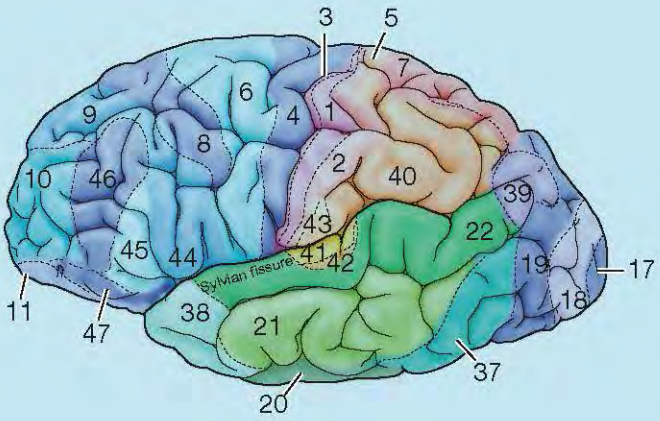
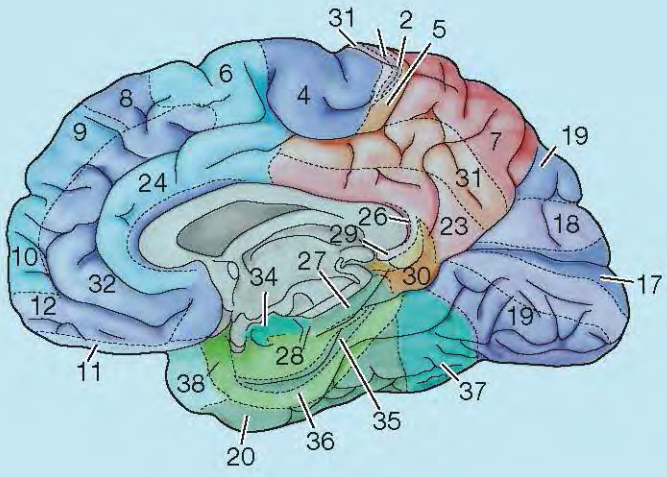
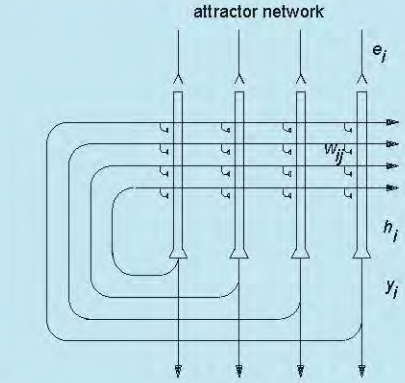
Brain Computations

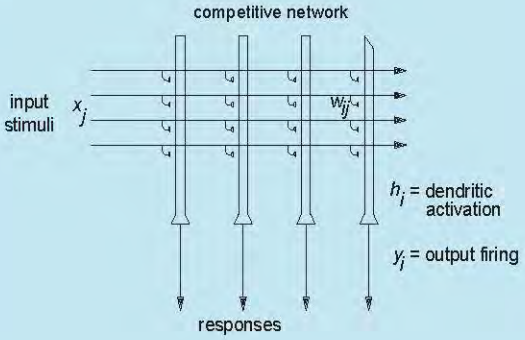
What and How

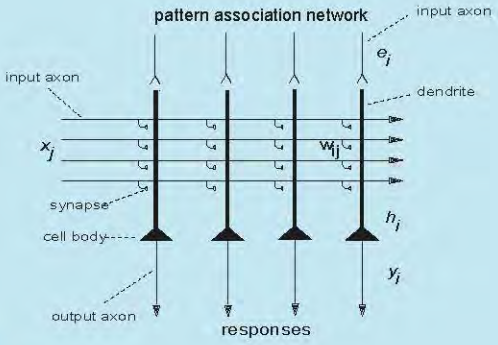
Edmund T. Rolls











OXFORD

Brain Computations

What and How

Edmund T. Rolls
Oxford Centre for Computational Neuroscience
Oxford
England

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

First Edition published in 2021
Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2020945700

ISBN 978–0–19–887110–1

DOI: 10.1093/oso/9780198871101.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Preface

Many scientists, and many others, are interested in how the brain works. In order to understand this, we need to know **what** computations are performed by different brain systems; and **how** they are computed by each of these systems.

The aim of this book is to elucidate what is computed in different brain systems; and to describe current computational approaches and models of how each of these brain systems computes.

To understand how our brains work, it is essential to know **what** is computed in each part of the brain. That can be addressed by utilising evidence relevant to computation from many areas of neuroscience. Knowledge of the connections between different brain areas is important, for this shows that the brain is organised as systems, with whole series of brain areas devoted for example to visual processing. That provides a foundation for examining the computation performed by each brain area, by comparing what is represented in a brain area with what is represented in the preceding and following brain area, using techniques of for example neurophysiology and functional neuroimaging. Neurophysiology at the single neuron level is needed because this is the level at which information is transmitted between the computing elements of the brain, the neurons. 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. Functional 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. So for each brain system, evidence on what is computed at each stage, and what the system as a whole computes, is essential.

To understand how our brains work, it is also essential to know **how** each part of the brain computes. That requires a knowledge of what is computed by each part of the brain, but it also requires knowledge of the network properties of each brain region. This involves knowledge of the connectivity between the neurons in each part of the brain, and knowledge of the synaptic and biophysical properties of the neurons. It also requires knowledge of the theory of what can be computed by networks with defined connectivity.

There are at least three key goals of the approaches described here. One is to understand ourselves better, and how we work and think. A second is to be better able to treat the system when it has problems, for example in mental illnesses. Medical applications are a very important aim of the type of research described here. A third, is to be able to emulate the operation of parts of our brains, which some in the field of artificial intelligence (AI) would like to do to produce useful machines. All of these goals require, and cannot get off the ground, without a firm foundation in what is computed by brain systems, and theories and models of how it is computed. To understand the operation of the whole brain, it is necessary to show how the different brain systems operate together: but a necessary foundation for this is to know what is computed in each brain system.

Part of the enterprise here is to stimulate new theories and models of how parts of the brain work. The evidence on what is computed in different brain systems had advanced rapidly in the last 50 years, and provides a reasonable foundation for the enterprise, though there is much that remains to be learned. Theories of how the computation is performed are less advanced, but progress is being made, and current models are described in this book for many brain systems, in the expectation that before further advances are made, knowledge of

the considerable current evidence on how the brain computes provides a useful stating point, especially as current theories do take into account the limitations that are likely to be imposed by the neural architectures present in our brains.

The simplest way to define **brain computation** is to examine what information is represented at each stage of processing, and how this is different from stage to stage. For example in the primary visual cortex (V1), neurons respond to simple stimuli such as bars or edges or gratings and have small receptive fields. Little can be read off from the firing rates about for example whose face is represented from a small number of neurons in V1. On the other hand, after four or five stages of processing, in the inferior temporal cortex, information can be read from the firing rates of neurons about whose face is being viewed, and indeed there is remarkable invariance with respect to the position, size, contrast and even in some cases view of the face. That is a major computation, and indicates what can be achieved by neural computation.

These approaches can only be taken to understand brain function because there is considerable localization of function in the brain, quite unlike a digital computer. One fundamental reason for localization of function in the brain is that this minimizes the total length of the connections between neurons, and thus brain size. Another is that it simplifies the genetic information that has to be provided in order to build the brain, because the connectivity instructions can refer considerably to local connections. These points are developed in my book *Cerebral Cortex: Principles of Operation* (Rolls, 2016b).

That brings me to what is different about the present book and *Cerebral Cortex: Principles of Operation* (Rolls, 2016b). The previous book took on the enormous task of making progress with understanding how the major part of our brains, the cerebral cortex, works, by understanding its principles of operation. The present book builds on that approach, and uses it as background, but has the different aim of taking each of our brain systems, and describing *what* they compute, and then what is known about *how* each system computes. The issue of how they compute relies for many brain systems on how the cortex operates, so *Cerebral Cortex: Principles of Operation* provides an important complement to the present book.

With its focus on what and how each brain system computes, a field that includes computational neuroscience, 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 brain actually performs some of its complex and necessarily computational functions in memory, perception, attention, decision-making, cognitive functions, and actions.

Indeed, as one of the key aims of this book is to describe **what computations** are performed by different brain systems, I have chosen to include in this book some of the key discoveries in neuroscience that I believe help to define what computations are performed in different brain systems.

That makes this book very different from many of the textbooks of neuroscience (such as *Principles of Neuroscience* (Kandel et al., 2013)), and some of the textbooks of theoretical neuroscience that describe principles of operation of neurons or of networks of neurons (Dayan and Abbott, 2001; Hertz et al., 1991; Gerstner et al., 2014), but not in general what is computed in different brain systems, and how it is computed. Further, there are likely to be great developments in our understanding of *how* the brain computes, and this book is intended to set out a framework for new developments, by providing an analysis of what is computed by different brain systems, and providing some current approaches to how these computations may be performed.

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 performed by the brain, and whether the simulation has similar properties to the real brain. 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 brain 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 brain 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 brain 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, and the major differences in their systems-level computational organization from that of rodents, as set out in Section 19.10.

The overall aims of the book are developed further, and the plan of the book is described, in Chapter 1. Appendix B describes the fundamental operation of key networks of the type that are likely to be the building blocks of brain function. Appendix C describes quantitative, information theoretic, approaches to how information is represented in the brain, which is an essential framework for understanding what is computed in a brain system, and how it is computed. 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; to show how the information represented by neurons can be measured; and to provide a tutorial version of the VisNet program for invariant visual object recognition described in Chapter 2. The neural networks programs are also provided in Python. The programs are available at https://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 Alessandro Treves, Gustavo Deco, and Simon M. Stringer, who have contributed greatly in always interesting and fruitful research collaborations 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. Charl Ning (University of Warwick) is thanked for help with translating the Matlab neural network programs described in Appendix D into Python. Professor Lorraine Tyler (University of Cambridge) is thanked for guidance to some of the literature on language in the brain. Dr Patrick Mills is thanked for very helpful comments on an earlier version of this book. 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. The book was typeset by the author using LaTeXand WinEdt.

The cover shows on the left a human brain with a lateral view above and a medial view below. The numbers refer to Brodmann areas, and a summary of the functions of each area

is provided in Section 1.11 together with in Fig.1.9 a more fully labelled version of these images. These images refer to one of the aims of this book, to describe **what** computations are performed in each brain area and especially in brain systems that consist of a set of closely connected brain areas. It should be noted that there is not an exact correspondence between Brodmann areas and computationally relevant brain areas, and an aim of this book is to delineate more exactly the brain's computational systems, and the areas and subareas that are involved in each computational system. (The images are from Purves,D., Augustine,G.J., Fitzpatrick,D. et al., editors (2019) Neuroscience. International Edition. © Oxford University Press: Oxford.) The different neural networks on the right refer to some of the types of biologically plausible neuronal network architectures that are important in **how** the computations are performed in different cortical areas. These networks are introduced in Section 1.9, and described computationally in Appendix B.

Updates to and .pdfs of many of the publications cited in this book are available at https://www.oxcns.org. Updates and corrections to the text and notes are also available at https://www.oxcns.org.

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 What and how the brain computes: introduction 1

1.2 What and how the brain computes: plan of the book 3

1.3 Neurons 5

1.4 Neurons in a network 6

1.5 Synaptic modification 8

1.6 Long-term potentiation and long-term depression 10

1.7 Information encoding by neurons, and distributed representations 14
- 1.7.1 Definitions 16
- 1.7.2 Advantages of different types of coding 16

1.8 Neuronal network approaches versus connectionism 18

1.9 Introduction to three neuronal network architectures 18

1.10 Systems-level analysis of brain function 21

1.11 Brodmann areas 23

1.12 The fine structure of the cerebral neocortex 26
- 1.12.1 The fine structure and connectivity of the neocortex 27
- 1.12.2 Excitatory cells and connections 27
- 1.12.3 Inhibitory cells and connections 29
- 1.12.4 Quantitative aspects of cortical architecture 32
- 1.12.5 Functional pathways through the cortical layers 34
- 1.12.6 The scale of lateral excitatory and inhibitory effects, and modules 38

2 The ventral visual system **40**

2.1 Introduction and overview 40
- 2.1.1 Introduction 40
- 2.1.2 Overview of what is computed in the ventral visual system 40
- 2.1.3 Overview of how computations are performed in the ventral visual system 43
- 2.1.4 What is computed in the ventral visual system is unimodal, and is related to other 'what' systems after the inferior temporal visual cortex 45

2.2 What: V1 – primary visual cortex 47

2.3 What: V2 and V4 – intermediate processing areas in the ventral visual system 48

2.4 What: Invariant representations of faces and objects in the inferior temporal visual cortex 49
- 2.4.1 Reward value is not represented in the ventral visual system 49
- 2.4.2 Translation invariant representations 50
- 2.4.3 Reduced translation invariance in natural scenes 50
- 2.4.4 Size and spatial frequency invariance 53
- 2.4.5 Combinations of features in the correct spatial configuration 54
- 2.4.6 A view-invariant representation 54
- 2.4.7 Learning in the inferior temporal cortex 58
- 2.4.8 A sparse distributed representation is what is computed in the ventral visual system 60
- 2.4.9 Face expression, gesture, and view 66
- 2.4.10 Specialized regions in the temporal cortical visual areas 66

2.5 How the computations are performed: approaches to invariant object recognition 71
2.5.1 Feature spaces 72
2.5.2 Structural descriptions and syntactic pattern recognition 73
2.5.3 Template matching and the alignment approach 75
2.5.4 Invertible networks that can reconstruct their inputs 76
2.5.5 Deep learning 76
2.5.6 Feature hierarchies 77
2.6 Hypotheses about how the computations are performed in a feature hierarchy approach 82
2.7 VisNet: a model of how the computations are performed in the ventral visual system 86
2.7.1 The architecture of VisNet 87
2.7.2 Initial experiments with VisNet 96
2.7.3 The optimal parameters for the temporal trace used in the learning rule 102
2.7.4 Different forms of the trace learning rule, and error correction 104
2.7.5 The issue of feature binding, and a solution 112
2.7.6 Operation in a cluttered environment 125
2.7.7 Learning 3D transforms 132
2.7.8 Capacity of the architecture, and an attractor implementation 137
2.7.9 Vision in natural scenes – effects of background versus attention 140
2.7.10 The representation of multiple objects in a scene 148
2.7.11 Learning invariant representations using spatial continuity 150
2.7.12 Lighting invariance 151
2.7.13 Deformation-invariant object recognition 153
2.7.14 Learning invariant representations of scenes and places 154
2.7.15 Finding and recognising objects in natural scenes 156
2.7.16 Non-accidental properties, and transform invariant object recognition 159
2.8 Further approaches to invariant object recognition 161
2.8.1 Other types of slow learning 161
2.8.2 HMAX 161
2.8.3 Hierarchical convolutional deep neural networks 166
2.8.4 Sigma-Pi synapses 167
2.9 Visuo-spatial scratchpad memory, and change blindness 167
2.10 Processes involved in object identification 169
2.11 Top-down attentional modulation is implemented by biased competition 170
2.12 Highlights on how the computations are performed in the ventral visual system 173

3 The dorsal visual system 176

3.1 Introduction, and overview of the dorsal cortical visual stream 176
3.2 Global motion in the dorsal visual system 177
3.3 Invariant object-based motion in the dorsal visual system 179
3.4 What is computed in the dorsal visual system: visual coordinate transforms 181
3.4.1 The transform from retinal to head-based coordinates 182
3.4.2 The transform from head-based to allocentric bearing coordinates 183
3.4.3 A transform from allocentric bearing coordinates to allocentric spatial view coordinates 184
3.5 How visual coordinate transforms are computed in the dorsal visual system 185
3.5.1 Gain modulation 185
3.5.2 Mechanisms of gain modulation using a trace learning rule 186
3.5.3 Gain modulation by eye position to produce a head-centered representation in Layer 1 of VisNetCT 187
3.5.4 Gain modulation by head direction to produce an allocentric bearing to a landmark in Layer 2 of VisNetCT 188
3.5.5 Gain modulation by place to produce an allocentric spatial view representation in Layer 3 of VisNetCT 189
3.5.6 The utility of the coordinate transforms in the dorsal visual system 190

4 The taste and flavour system 192

4.1 Introduction and overview 192
- 4.1.1 Introduction 192
- 4.1.2 Overview of what is computed in the taste and flavour system 192
- 4.1.3 Overview of how computations are performed in the taste and flavour system 194

4.2 Taste and related pathways: what is computed 194
- 4.2.1 Hierarchically organised anatomical pathways 194
- 4.2.2 Taste neuronal tuning become more selective through the taste hierarchy 197
- 4.2.3 The primary, insular, taste cortex represents what taste is present and its intensity 198
- 4.2.4 The secondary, orbitofrontal, taste cortex, and its representation of the reward value and pleasantness of taste 199
- 4.2.5 Sensory-specific satiety is computed in the orbitofrontal cortex 201
- 4.2.6 Oral texture is represented in the primary and secondary taste cortex: viscosity and fat texture 204
- 4.2.7 Vision and olfaction converge using associative learning with taste to represent flavour in the secondary but not primary taste cortex 207
- 4.2.8 Top-down attention and cognition can modulate taste and flavour representations in the taste cortical areas 208
- 4.2.9 The tertiary taste cortex in the anterior cingulate cortex provides the rewards for action-reward learning 210

4.3 Taste and related pathways: how the computations are performed 212
- 4.3.1 Increased selectivity of taste and flavor neurons through the hierarchy by competitive learning and convergence 212
- 4.3.2 Pattern association learning of associations of visual and olfactory stimuli with taste 212
- 4.3.3 Rule-based reversal of visual to taste associations in the orbitofrontal cortex 212
- 4.3.4 Sensory-specific satiety is implemented by adaptation of synapses onto orbitofrontal cortex neurons 212
- 4.3.5 Top-down cognitive and attentional modulation is implemented by biased activation 213

5 The olfactory system 217

5.1 Introduction 217
- 5.1.1 Overview of what is computed in the olfactory system 217
- 5.1.2 Overview of how the computations are performed in the olfactory system 218

5.2 What is computed in the olfactory system 219
- 5.2.1 1000 gene-encoded olfactory receptor types, and 1000 corresponding glomerulus types in the olfactory bulb 219
- 5.2.2 The primary olfactory, pyriform, cortex: olfactory feature combinations are what is represented 221
- 5.2.3 Orbitofrontal cortex: olfactory neuronal response selectivity 221
- 5.2.4 Orbitofrontal cortex: olfactory to taste convergence 222
- 5.2.5 Orbitofrontal cortex: olfactory to taste association learning and reversal 223
- 5.2.6 Orbitofrontal cortex: olfactory reward value is represented 225

5.3 How computations are performed in the olfactory system 226
- 5.3.1 Olfactory receptors, and the olfactory bulb 226
- 5.3.2 Olfactory (pyriform) cortex 227
- 5.3.3 Orbitofrontal cortex 231

6 The somatosensory system 232

6.1 What is computed in the somatosensory system 232
- 6.1.1 The receptors and periphery 232
- 6.1.2 The anterior somatosensory cortex, areas 1, 2, 3a, and 3b, in the anterior parietal cortex 232
- 6.1.3 The ventral somatosensory stream: areas S2 and PV, in the lateral parietal cortex 233
- 6.1.4 The dorsal somatosensory stream to area 5 and then 7b, in the posterior parietal cortex 234

6.1.5 Somatosensory representations in the insula 236
6.1.6 Somatosensory and temperature inputs to the orbitofrontal cortex, affective value, pleasant touch, and pain 236
6.1.7 Decision-making in the somatosensory system 240
6.2 How computations are performed in the somatosensory system 242
6.2.1 Hierarchical computation in the somatosensory system 242
6.2.2 Computations for pleasant touch and pain 243
6.2.3 The mechanisms for somatosensory decision-making 243

7 The auditory system 244

7.1 Introduction, and overview of computations in the auditory system 244
7.2 Auditory Localization 245
7.3 Ventral and dorsal cortical auditory pathways 248
7.4 The ventral cortical auditory stream 249
7.5 The dorsal cortical auditory stream 251
7.6 How the computations are performed in the auditory system 251

8 The temporal cortex 253

8.1 Introduction and overview 253
8.2 Middle temporal gyrus and face expression and gesture 253
8.3 Semantic representations in the temporal lobe neocortex 255
8.3.1 Neurophysiology 255
8.3.2 Neuropsychology 256
8.3.3 Functional neuroimaging 256
8.3.4 Brain stimulation 257
8.4 The mechanisms for semantic learning in the human anterior temporal lobe 259

9 The hippocampus, memory, and spatial function 260

9.1 Introduction and overview 260
9.1.1 Overview of what is computed by the hippocampal system 260
9.1.2 Overview of how the computations are performed by the hippocampal system 262
9.2 What is computed in the hippocampus 263
9.2.1 Systems-level anatomy 263
9.2.2 Evidence from the effects of damage to the hippocampus 265
9.2.3 Episodic memories need to be recalled from the hippocampus, and can be used to help build neocortical semantic memories 267
9.2.4 Systems-level neurophysiology of the primate hippocampus 270
9.2.5 Head direction cells in the presubiculum 285
9.2.6 Perirhinal cortex, recognition memory, and long-term familiarity memory 286
9.3 How computations are performed in the hippocampal system 293
9.3.1 Historical development of the theory of the hippocampus 293
9.3.2 Hippocampal circuitry 296
9.3.3 Medial entorhinal cortex, spatial processing streams, and grid cells 297
9.3.4 Lateral entorhinal cortex, object processing streams, and the generation of time cells in the hippocampus 301
9.3.5 CA3 as an autoassociation memory 307
9.3.6 Dentate granule cells 325
9.3.7 CA1 cells 328
9.3.8 Backprojections to the neocortex, memory recall, and consolidation 333
9.3.9 Backprojections to the neocortex – quantitative aspects 336
9.3.10 Simulations of hippocampal operation 339
9.3.11 The learning of spatial view and place cell representations 341
9.3.12 Linking the inferior temporal visual cortex to spatial view and place cells 342
9.3.13 A scientific theory of the art of memory: scientia artis memoriae 344

9.4 Tests of the theory of hippocampal cortex operation 344
9.4.1 Dentate gyrus (DG) subregion of the hippocampus 345
9.4.2 CA3 subregion of the hippocampus 348
9.4.3 CA1 subregion of the hippocampus 355
9.5 Comparison with other theories of hippocampal function 358

10 The parietal cortex, spatial functions, and navigation 363

10.1 Introduction and overview 363
10.1.1 Introduction 363
10.1.2 Overview of what is computed in the parietal cortex 363
10.1.3 Overview of how the computations are performed in the parietal cortex 364
10.2 Precuneus and medial area 7 365
10.3 Navigation: What computations are performed in the parietal and related cortex 366
10.4 How navigation is performed 367
10.4.1 Navigation using a hippocampal allocentric Euclidean cognitive map 367
10.4.2 Navigation using an entorhinal cortex goal vector system 367
10.4.3 Transforms between allocentric and egocentric representations 368
10.4.4 Navigational computations using neuron types found in primates 371

11 The orbitofrontal cortex, amygdala, reward value, and emotion 379

11.1 Introduction and overview 379
11.1.1 Introduction 379
11.1.2 Overview of what is computed in the orbitofrontal cortex 379
11.1.3 Overview of how the computations are performed by the orbitofrontal cortex 381
11.2 The topology and connections of the orbitofrontal cortex 383
11.2.1 Inputs to the orbitofrontal cortex 383
11.2.2 Outputs of the orbitofrontal cortex 386
11.3 What is computed in the orbitofrontal cortex 387
11.3.1 The orbitofrontal cortex represents reward value 387
11.3.2 Neuroeconomic value is represented in the orbitofrontal cortex 392
11.3.3 A representation of face and voice expression and other socially relevant stimuli in the orbitofrontal cortex 396
11.3.4 Negative reward prediction error neurons in the orbitofrontal cortex 399
11.3.5 The human medial orbitofrontal cortex represents rewards, and the lateral orbitofrontal cortex non-reward and punishers 403
11.3.6 Decision-making in the orbitofrontal / ventromedial prefrontal cortex 406
11.3.7 The ventromedial prefrontal cortex and memory 408
11.3.8 The orbitofrontal cortex and emotion 409
11.3.9 Emotional orbitofrontal vs rational routes to action 411
11.3.10 Comparison between the functions of the orbitofrontal cortex and amygdala in emotion 419
11.4 How the computations are performed in the orbitofrontal cortex 428
11.4.1 Decision-making in attractor networks in the brain 429
11.4.2 Analyses of reward-related decision-making mechanisms in the orbitofrontal cortex 433
11.4.3 A model for reversal learning in the orbitofrontal cortex 438
11.4.4 A theory and model of non-reward neural mechanisms in the orbitofrontal cortex 443
11.5 Highlights: the special computational roles of the orbitofrontal cortex 444

12 The cingulate cortex 447

12.1 Introduction to and overview of the cingulate cortex 447
12.1.1 Introduction 447
12.1.2 Overview of what is computed in the cingulate cortex 447
12.1.3 Overview of how the computations are performed by the cingulate cortex 449
12.2 Anterior Cingulate Cortex 450

12.2.1 Anterior cingulate cortex anatomy and connections 450
12.2.2 Anterior cingulate cortex: A framework 451
12.2.3 Pregenual anterior cingulate representations of reward value, and supracallosal anterior cingulate representations of punishers and non-reward 453
12.2.4 Anterior cingulate cortex and action-outcome representations 455
12.2.5 Anterior cingulate cortex lesion effects 455
12.2.6 Subgenual cingulate cortex 456
12.3 Mid-cingulate cortex, the cingulate motor area, and action–outcome learning 457
12.4 The posterior cingulate cortex 458
12.5 How the computations are performed by the cingulate cortex 459
12.5.1 The anterior cingulate cortex and emotion 459
12.5.2 Action-outcome learning 459
12.5.3 Connectivity of the posterior cingulate cortex with the hippocampal memory system 460
12.6 Synthesis and conclusions 461

13 The motor cortical areas 464

13.1 Introduction and overview 464
13.2 What is computed in different cortical motor-related areas 464
13.2.1 Ventral parietal and ventral premotor cortex F4 464
13.2.2 Superior parietal areas with activity related to reaching 466
13.2.3 Inferior parietal areas with activity related to grasping, and ventral premotor cortex F5 466
13.3 The mirror neuron system 466
13.4 How the computations are performed in motor cortical and related areas 467

14 The basal ganglia 468

14.1 Introduction and overview 468
14.2 Systems-level architecture of the basal ganglia 469
14.3 What computations are performed by the basal ganglia? 471
14.3.1 Effects of striatal lesions 471
14.3.2 Neuronal activity in different parts of the striatum 473
14.4 How do the basal ganglia perform their computations? 485
14.4.1 Interaction between neurons and selection of output 485
14.4.2 Convergence within the basal ganglia, useful for stimulus-response habit learning 489
14.4.3 Dopamine as a reward prediction error signal for reinforcement learning in the striatum 491
14.5 Comparison of computations for selection in the basal ganglia and cerebral cortex 494

15 Cerebellar cortex 497

15.1 Introduction 497
15.2 Architecture of the cerebellum 498
15.2.1 The connections of the parallel fibres onto the Purkinje cells 498
15.2.2 The climbing fibre input to the Purkinje cell 499
15.2.3 The mossy fibre to granule cell connectivity 500
15.3 Modifiable synapses of parallel fibres onto Purkinje cell dendrites 502
15.4 The cerebellar cortex as a perceptron 502
15.5 Highlights: differences between cerebral and cerebellar cortex microcircuitry 503

16 The prefrontal cortex 505

16.1 Introduction and overview 505

16.2 Divisions of the lateral prefrontal cortex 508
16.2.1 The dorsolateral prefrontal cortex 509
16.2.2 The caudal prefrontal cortex 510
16.2.3 The ventrolateral prefrontal cortex 512

16.3 The lateral prefrontal cortex and top-down attention 512

16.4 How the computations are performed in the prefrontal cortex 515
16.4.1 Cortical short-term memory systems and attractor networks 515
16.4.2 Prefrontal cortex short-term memory networks, and their relation to perceptual networks 517
16.4.3 Mapping from one representation to another in short-term memory 522
16.4.4 The mechanisms of top-down attention 524
16.4.5 Computational necessity for a separate, prefrontal cortex, short-term memory system 525
16.4.6 Synaptic modification is needed to set up but not to reuse short-term memory systems 525
16.4.7 Sequence memory 525
16.4.8 Working memory, and planning 526

17 Language and syntax in the brain 527

17.1 Introduction and overview 527
17.1.1 Introduction 527
17.1.2 Overview 527

17.2 What is computed in different brain systems to implement language 529
17.2.1 The Wernicke-Lichtheim-Geschwind hypothesis 529
17.2.2 The dual-stream hypothesis of speech comprehension 529
17.2.3 Reading requires different brain systems to hearing speech 530
17.2.4 Semantic representations 531
17.2.5 Syntactic processing 533
17.2.6 The parietal cortex: supramarginal and angular gyri 533

17.3 Hypotheses about how semantic representations are computed 534

17.4 A neurodynamical hypothesis about how syntax is computed 535
17.4.1 Binding by synchrony? 535
17.4.2 Syntax using a place code 536
17.4.3 Temporal trajectories through a state space of attractors 536
17.4.4 Hypotheses about the implementation of language in the cerebral cortex 537
17.4.5 Tests of the hypotheses – a model 540
17.4.6 Tests of the hypotheses – findings with the model 545
17.4.7 Evaluation of the hypotheses 548
17.4.8 Further approaches 552

18 Cortical attractor dynamics and connectivity, stochasticity, psychiatric disorders, and aging 554

18.1 Introduction and overview 554
18.1.1 Introduction 554
18.1.2 Overview 554

18.2 The noisy cortex 555
18.2.1 Reasons why the brain is inherently noisy and stochastic 556
18.2.2 Attractor networks, energy landscapes, and stochastic neurodynamics 559
18.2.3 A multistable system with noise 564
18.2.4 Stochastic dynamics and the stability of short-term memory 566
18.2.5 Stochastic dynamics in decision-making, and the evolutionary utility of probabilistic choice 570
18.2.6 Selection between conscious vs unconscious decision-making, and free will 571
18.2.7 Stochastic dynamics and creative thought 572
18.2.8 Stochastic dynamics and unpredictable behaviour 573

18.3 Attractor dynamics and schizophrenia 573
18.3.1 Introduction 573

18.3.2 A dynamical systems hypothesis of the symptoms of schizophrenia 574
18.3.3 Reduced functional connectivity of some brain regions in schizophrenia 577
18.3.4 Beyond the disconnectivity hypothesis of schizophrenia: reduced forward but not backward connectivity 578
18.4 Attractor dynamics and obsessive-compulsive disorder 582
18.4.1 Introduction 582
18.4.2 A hypothesis about obsessive-compulsive disorder 582
18.4.3 Glutamate and increased depth of the basins of attraction 585
18.5 Depression and attractor dynamics 586
18.5.1 Introduction 586
18.5.2 A non-reward attractor theory of depression 587
18.5.3 The orbitofrontal cortex, and the theory of depression 588
18.5.4 Altered connectivity of the orbitofrontal cortex in depression 590
18.5.5 Activations of the orbitofrontal cortex related to depression 595
18.5.6 Implications, and possible treatments, and subtypes of depression 595
18.5.7 Mania and bipolar disorder 598
18.6 Attractor stochastic dynamics, aging, and memory 600
18.6.1 NMDA receptor hypofunction 600
18.6.2 Dopamine and norepinephrine 602
18.6.3 Impaired synaptic modification 602
18.6.4 Cholinergic function and memory 603
18.7 High blood pressure, reduced hippocampal functional connectivity, and impaired memory 607
18.8 Brain development, and structural differences in the brain 608

19 Computations by different types of brain, and by artificial neural systems 609

19.1 Introduction and overview 609
19.2 Computations that combine different computational systems in the brain to produce behaviour 610
19.3 Brain computation compared to computation on a digital computer 610
19.4 Brain computation compared with artificial deep learning networks 616
19.5 Reinforcement Learning 618
19.6 Levels of explanation, and the mind-brain problem 620
19.7 Levels of explanation, and levels of investigation 622
19.8 Brain-Inspired Intelligence 623
19.9 Brain-Inspired Medicine 624
19.9.1 Computational psychiatry and neurology 624
19.9.2 Reward systems in the brain, and their application to understanding food intake control and obesity 625
19.9.3 Multiple Routes to Action 628
19.10 Primates including humans have different brain organisation than rodents 628
19.10.1 The visual system 628
19.10.2 The taste system 629
19.10.3 The olfactory system 629
19.10.4 The somatosensory system 630
19.10.5 The auditory system 630
19.10.6 The hippocampal system and memory 630
19.10.7 The orbitofrontal cortex and amygdala 631
19.10.8 The cingulate cortex 632
19.10.9 The motor system 632
19.10.10 Language 633

A Introduction to linear algebra for neural networks 634

A.1 Vectors 634
A.1.1 The inner or dot product of two vectors 634
A.1.2 The length of a vector 636
A.1.3 Normalizing the length of a vector 636
A.1.4 The angle between two vectors: the normalized dot product 636
A.1.5 The outer product of two vectors 637
A.1.6 Linear and non-linear systems 638
A.1.7 Linear combinations, linear independence, and linear separability 639
A.2 Application to understanding simple neural networks 640
A.2.1 Capability and limitations of single-layer networks 641
A.2.2 Non-linear networks: neurons with non-linear activation functions 643
A.2.3 Non-linear networks: neurons with non-linear activations 644

B Neuronal network models 646

B.1 Introduction 646
B.2 Pattern association memory 646
B.2.1 Architecture and operation 647
B.2.2 A simple model 649
B.2.3 The vector interpretation 652
B.2.4 Properties 653
B.2.5 Prototype extraction, extraction of central tendency, and noise reduction 656
B.2.6 Speed 656
B.2.7 Local learning rule 657
B.2.8 Implications of different types of coding for storage in pattern associators 662
B.3 Autoassociation or attractor memory 663
B.3.1 Architecture and operation 663
B.3.2 Introduction to the analysis of the operation of autoassociation networks 665
B.3.3 Properties 667
B.3.4 Diluted connectivity and the storage capacity of attractor networks 674
B.3.5 Use of autoassociation networks in the brain 685
B.4 Competitive networks, including self-organizing maps 686
B.4.1 Function 686
B.4.2 Architecture and algorithm 687
B.4.3 Properties 688
B.4.4 Utility of competitive networks in information processing by the brain 693
B.4.5 Guidance of competitive learning 694
B.4.6 Topographic map formation 696
B.4.7 Invariance learning by competitive networks 700
B.4.8 Radial Basis Function networks 702
B.4.9 Further details of the algorithms used in competitive networks 703
B.5 Continuous attractor networks 707
B.5.1 Introduction 707
B.5.2 The generic model of a continuous attractor network 709
B.5.3 Learning the synaptic strengths in a continuous attractor network 709
B.5.4 The capacity of a continuous attractor network: multiple charts 711
B.5.5 Continuous attractor models: path integration 712
B.5.6 Stabilization of the activity packet within a continuous attractor network 715
B.5.7 Continuous attractor networks in two or more dimensions 717
B.5.8 Mixed continuous and discrete attractor networks 717
B.6 Network dynamics: the integrate-and-fire approach 718
B.6.1 From discrete to continuous time 718
B.6.2 Continuous dynamics with discontinuities 720
B.6.3 An integrate-and-fire implementation 723
B.6.4 The speed of processing of attractor networks 725
B.6.5 The speed of processing of a four-layer hierarchical network 727
B.6.6 Spike response model 731
B.7 Network dynamics: introduction to the mean-field approach 732
B.8 Mean-field based neurodynamics 733

B.8.1 Population activity 733
B.8.2 The mean-field approach used in a model of decision-making 735
B.8.3 The model parameters used in the mean-field analyses of decision-making 737
B.8.4 A basic computational module based on biased competition 737
B.8.5 Multimodular neurodynamical architectures 739

B.9 Interacting attractor networks 742

B.10 Sequence memory implemented by adaptation in an attractor network 745

B.11 Error correction networks 745
B.11.1 Architecture and general description 746
B.11.2 Generic algorithm for a one-layer error correction network 746
B.11.3 Capability and limitations of single-layer error-correcting networks 747
B.11.4 Properties 751

B.12 Error backpropagation multilayer networks 753
B.12.1 Introduction 753
B.12.2 Architecture and algorithm 753
B.12.3 Properties of multilayer networks trained by error backpropagation 756

B.13 Convolution networks 757

B.14 Contrastive Hebbian learning: the Boltzmann machine 758

B.15 Deep Belief Networks 760

B.16 Reinforcement learning 760
B.16.1 Associative reward–penalty algorithm of Barto and Sutton 761
B.16.2 Reward prediction error or delta rule learning, and classical conditioning 763
B.16.3 Temporal Difference (TD) learning 764

B.17 Learning in the neocortex 767

B.18 Forgetting in cortical associative neural networks, and memory reconsolidation 769

B.19 Highlights 773

C Information theory, and neuronal encoding 774

C.1 Information theory 775
C.1.1 The information conveyed by definite statements 775
C.1.2 Information conveyed by probabilistic statements 776
C.1.3 Information sources, information channels, and information measures 777
C.1.4 The information carried by a neuronal response and its averages 778
C.1.5 The information conveyed by continuous variables 781

C.2 The information carried by neuronal responses 783
C.2.1 The limited sampling problem 783
C.2.2 Correction procedures for limited sampling 784
C.2.3 The information from multiple cells: decoding procedures 785
C.2.4 Information in the correlations between cells: a decoding approach 789
C.2.5 Information in the correlations between cells: second derivative approach 794

C.3 Information theory results 797
C.3.1 The sparseness of the distributed encoding used by the brain 798
C.3.2 The information from single neurons 809
C.3.3 The information from single neurons: temporal codes versus rate codes 812
C.3.4 The information from single neurons: the speed of information transfer 814
C.3.5 The information from multiple cells: independence versus redundancy 825
C.3.6 Should one neuron be as discriminative as the whole organism? 829
C.3.7 The information from multiple cells: the effects of cross-correlations 830
C.3.8 Conclusions on cortical neuronal encoding 834

C.4 Information theory terms – a short glossary 838

C.5 Highlights 839

D Simulation software for neuronal networks, and information analysis of neuronal encoding 840

D.1 Introduction 840
D.2 Autoassociation or attractor networks 841
D.2.1 Running the simulation 841
D.2.2 Exercises 843
D.3 Pattern association networks 843
D.3.1 Running the simulation 843
D.3.2 Exercises 845
D.4 Competitive networks and Self-Organizing Maps 846
D.4.1 Running the simulation 846
D.4.2 Exercises 847
D.5 Further developments 848
D.6 Matlab code for a tutorial version of VisNet 848
D.7 Matlab code for information analysis of neuronal encoding 849
D.8 Matlab code to illustrate the use of spatial view cells in navigation 849
D.9 Highlights 849

References 850

Index 924

Appendix 3 Information theory, and neuronal encoding

In order to understand what is computed in the brain, and how it is computed, it is essential to understand how information is represented in the brain. Is it by the firing rates of neurons, by the latency of neuronal responses, by the order in which action potentials arrive in different neurons, by any stimulus-dependent cross-correlations between the firing of different neurons, etc? It is essential to understand how information is encoded by single neurons, and by populations of neurons. Neuronal encoding, and the use of information theory to analyze it, are described 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 (2016b), 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 being 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 $k \times l$ *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

$$I(k) = \log_2 k. \tag{C.1}$$

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

$$I(k \times l) = \log_2(k \times l) = \log_2 k + \log_2 l = I(k) + I(l). \tag{C.2}$$

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

$$I = \log_2(18/10) = \log_2(9/5) \tag{C.3}$$

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

$$I = \log_2 \mathrm{P}^{\mathrm{posterior}}(\mathrm{W})/\mathrm{P}^{\mathrm{prior}}(\mathrm{W}) = \log_2(9/10)/(1/2) = \log_2(9/5). \tag{C.4}$$

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

$$I = -\log_2 \mathrm{P}^{\mathrm{prior}}(\mathrm{W}) \tag{C.5}$$

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

$$I' = \log_2((1)/(9/10)) \tag{C.6}$$

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

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

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

$$I = \mathrm{P}^{\text{posterior}}(\mathrm{W}) \log_2 \frac{\mathrm{P}^{\text{posterior}}(\mathrm{W})}{\mathrm{P}^{\text{prior}}(\mathrm{W})} + \mathrm{P}^{\text{posterior}}(\mathrm{E}) \log_2 \frac{\mathrm{P}^{\text{posterior}}(\mathrm{E}}{\mathrm{P}^{\text{prior}}(\mathrm{E})} \tag{C.8}$$

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 *a posteriori* state, one has to average information values for each state with the corresponding *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 *a priori* probability $\mathrm{P}(s)$, has occurred is

$$I(s) = \log_2(1/\mathrm{P}(s)) = -\log_2 \mathrm{P}(s), \tag{C.9}$$

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

$$I = \sum_s \left[\mathrm{P}^{\text{posterior}}(s) \log_2 \frac{\mathrm{P}^{\text{posterior}}(s)}{\mathrm{P}^{\text{prior}}(s)} \right]. \tag{C.10}$$

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

$$H(S) = -\sum_{s \in S} \mathrm{P}(s) \log_2 \mathrm{P}(s). \tag{C.11}$$

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

$$\mathrm{P}(s, s') = \mathrm{P}(s)\mathrm{P}(s') \tag{C.12}$$

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 $\mathrm{P}(s'|s)$, and if the two variables are independent, it is just equal to the unconditional probability $\mathrm{P}(s')$. In general, 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

$$\mathrm{P}(s, s') = \mathrm{P}(s'|s)\mathrm{P}(s) = \mathrm{P}(s|s')\mathrm{P}(s'), \tag{C.13}$$

from which it is clear that

$$\mathrm{P}(s'|s) = \mathrm{P}(s|s')\frac{\mathrm{P}(s')}{\mathrm{P}(s)}, \tag{C.14}$$

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 $\mathrm{P}(s)$ is inserted as the prior, and the conditional probability $\mathrm{P}(s|s')$ as the posterior:

$$I(s') = \sum_s \mathrm{P}(s|s') \log_2 \frac{\mathrm{P}(s|s')}{\mathrm{P}(s)}. \tag{C.15}$$

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

$$I(s) = \sum_{s'} \mathrm{P}(s'|s) \log_2 \frac{\mathrm{P}(s'|s)}{\mathrm{P}(s')}. \tag{C.16}$$

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

$$\begin{aligned} I &= \sum_s \mathrm{P}(s) \sum_{s'} \mathrm{P}(s'|s) \log_2 \frac{\mathrm{P}(s'|s)}{\mathrm{P}(s')} \\ &= \sum_{s,s'} \mathrm{P}(s, s') \log_2 \frac{\mathrm{P}(s, s')}{\mathrm{P}(s)\mathrm{P}(s')}. \end{aligned} \tag{C.17}$$

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

$$I = H(S) - H(S|S') \equiv H(S) - \sum_{s'} \mathrm{P}(s')H(S|s'). \tag{C.18}$$

A channel is characterized, once the alphabets are given, by the set of conditional probabilities for the output symbols, $\mathrm{P}(s'|s)$, whereas the unconditional probabilities of the input symbols $\mathrm{P}(s)$ 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 $\mathrm{P}(s)$. 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 $\mathrm{P}(s)$ the *a priori* probability that the particular stimulus s out of a given set was shown, while the conditional probability $\mathrm{P}(s|r)$ is the *a posteriori* probability, that is updated by the knowledge of the response r. The response-specific transinformation

$$I(r) = \sum_s \mathrm{P}(s|r) \log_2 \frac{\mathrm{P}(s|r)}{\mathrm{P}(s)} \tag{C.19}$$

takes the extreme values of $I(r) = -\log_2 \mathrm{P}(s(r))$ if r unequivocally determines $s(r)$ (that is, $\mathrm{P}(s|r)$ equals 1 for that one stimulus and 0 for all others);
and $I(r) = \sum_s \mathrm{P}(s) \log_2(\mathrm{P}(s)/\mathrm{P}(s)) = 0$ 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 $\mathrm{P}(s|r) = \mathrm{P}(s)$.

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

$$< I >= \sum_r \mathrm{P}(r) \left[\sum_s \mathrm{P}(s|r) \log_2 \frac{\mathrm{P}(s|r)}{\mathrm{P}(s)} \right]. \tag{C.20}$$

The angular brackets $<>$ are used here to emphasize the averaging operation, in this case over responses. Denoting with $\mathrm{P}(s, r)$ the *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 **mutual information** $I(S, R)$

$$< I >= \sum_{s,r} \mathrm{P}(s, r) \log_2 \frac{\mathrm{P}(s, r)}{\mathrm{P}(s)\mathrm{P}(r)} \tag{C.21}$$

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,

$$I(s) = \sum_r \mathrm{P}(r|s) \log_2 \frac{\mathrm{P}(r|s)}{\mathrm{P}(r)} \tag{C.22}$$

which is a direct quantification of the variability in the responses elicited by that stimulus, compared to the overall variability. Since $\mathrm{P}(r)$ 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[27].

[27] 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 $I(S, R)$.

C.1.4.1 A numerical example

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 $\mathrm{P}(r|s)$ shown in Table C.1. Then, if different visual stimuli are presented with equal probability, the table of joint

Table C.1 The conditional probabilities $\mathrm{P}(r|s)$ that different neuronal responses (r=0, 1, or 2 spikes) will be produced by each of four stimuli (A–D).

	r=0	r=1	r=2
s=A	0.6	0.4	0.0
s=B	0.0	0.2	0.8
s=C	0.4	0.5	0.1
s=D	1.0	0.0	0.0

probabilities $\mathrm{P}(s, r)$ will be as shown in Table C.2. From these two tables one can compute

Table C.2 Joint probabilities $\mathrm{P}(s, r)$ that different neuronal responses (r=0, 1, or 2 spikes) will be produced by each of four equiprobable stimuli (A–D).

	r=0	r=1	r=2
s=A	0.15	0.1	0.0
s=B	0.0	0.05	0.2
s=C	0.1	0.125	0.025
s=D	0.25	0.0	0.0

various information measures by directly applying the definitions above. Since visual stimuli are presented with equal probability, $\mathrm{P}(s) = 1/4$, 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 $H_s = -\sum_s [\mathrm{P}(s) \log_2 \mathrm{P}(s)] = -4[(1/4) \log_2 (1/4)] = \log_2 4 = 2$ 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 $T = 1$ 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 $H_r = -\sum_r \mathrm{P}(r) \log_2 \mathrm{P}(r) = -\mathrm{P}(0) \log_2 \mathrm{P}(0) - \mathrm{P}(1) \log_2 \mathrm{P}(1) - \mathrm{P}(2) \log_2 \mathrm{P}(2) = -0.5 \log_2 0.5 - 0.275 \log_2 0.275 - 0.225 \log_2 0.225 = 1.496$ 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 *a priori* possibilities, only one, or a discrete number, remains *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 *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 *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 $\mathbf{r}^\mu$ with components the firing rates $\{r_i^\mu\}$, where the subscript i indexes the neuron (and the superscript μ indexes the pattern). In retrieving pattern μ, the network in fact produces a distinct firing pattern, denoted for example simply as $\mathbf{r}$. The quality of retrieval, or the similarity between $\mathbf{r}^\mu$ and $\mathbf{r}$, can be measured by the average mutual information

$$< I(\mathbf{r}^\mu, \mathbf{r}) > \; = \sum_{\mathbf{r}^\mu, \mathbf{r}} \mathrm{P}(\mathbf{r}^\mu, \mathbf{r}) \log_2 \frac{\mathrm{P}(\mathbf{r}^\mu, \mathbf{r})}{\mathrm{P}(\mathbf{r}^\mu)\mathrm{P}(\mathbf{r})} \tag{C.23}$$

$$\approx \sum_i \sum_{r_i^\mu, r_i} \mathrm{P}(r_i^\mu, r_i) \log_2 \frac{\mathrm{P}(r_i^\mu, r_i)}{\mathrm{P}(r_i^\mu)\mathrm{P}(r_i)}.$$

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_i or r_i^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 $p(r)\mathrm{d}r$, rather than the usual probabilities $\mathrm{P}(r)$. Substituting $p(r)\mathrm{d}r$ for $\mathrm{P}(r)$ and $p(r^\mu, r)\mathrm{d}r\mathrm{d}r^\mu$ for $\mathrm{P}(r^\mu, r)$, one can write for each single-cell contribution (omitting the cell index i)

$$< I(r^\mu, r) >_i = \int \mathrm{d}r^\mu \mathrm{d}r \; p(r^\mu, r) \log_2 \frac{p(r^\mu, r)}{p(r^\mu)p(r)} \tag{C.24}$$

and we see that the differentials $\mathrm{d}r^\mu \mathrm{d}r$ 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

$$p(r^\mu, r)\mathrm{d}r^\mu \mathrm{d}r = p(r^\mu)\delta(r - r(r^\mu))\mathrm{d}r^\mu \mathrm{d}r = p(r^\mu)\mathrm{d}r^\mu \tag{C.25}$$

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 $p(r|r^\mu)\mathrm{d}r = \exp -(r - r(r^\mu))^2/2\rho^2 \mathrm{d}r$ 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, 1992; 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, $(\mathbf{R}/2)^C$. 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

$$< I(s,\mathbf{r}) > = \sum_s \mathrm{P}(s) \int \Pi_i \mathrm{d}r_i \; p(\mathbf{r}|s) \log_2 \frac{p(\mathbf{r}|s)}{p(\mathbf{r})} \tag{C.26}$$

where it is important to note that $p(\mathbf{r})$ and $p(\mathbf{r}|s)$ are now probability densities defined over the high-dimensional vector space of multi-cell responses. The product sign Π signifies that this whole vector space has to be integrated over, along all its dimensions. $p(\mathbf{r})$ can be calculated as $\sum_s p(\mathbf{r}|s)\mathrm{P}(s)$, and therefore, in principle, all one has to do is to estimate, from the data, the conditional probability densities $p(\mathbf{r}|s)$ – 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'[28]) 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))[29].

A different correction procedure (called 'jack-knife') 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

$$< I > = N < I_N > -(1/N) \sum_k (N-1) < I_{N-1} >_k \tag{C.27}$$

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

[28] 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.

[29] 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 content in the responses of several cells (more than two to three) recorded 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 mut-

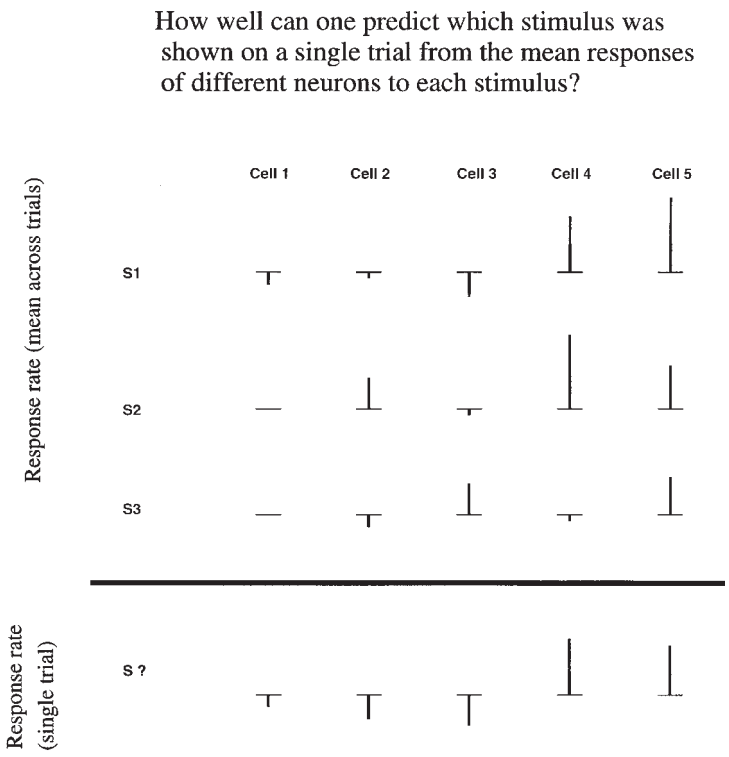


Fig. C.1 Decoding which stimulus (S1–S3) was present with a set of neuronal responses on a single trial to stimulus 'S ?'. The information available from multiple cells can then be calculated by comparing the decoded stimuli to the real stimuli that were presented. The figure shows the average response 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. (After Rolls, E. T., Treves,A. and Tovee,M.J. (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. Experimental Brain Research 114: 149–162. © Springer Nature.)

ual 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 matrix. 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 $\mathrm{P}(r|s)$, and use Bayes' rule to convert them into the conditional probabilities $\mathrm{P}(s'|r)$. 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 $\mathrm{P}(r|s)$ should not include the particular response for which $\mathrm{P}(s'|r)$ is going to be derived (jack-knife cross-validation). In practice, however, the estimation of $\mathrm{P}(r|s)$ 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 ($\mathbf{r}^{\text{av}}$), suitably normalized over all average vectors to generate probabilities

$$\mathrm{P}(s'|\mathbf{r}(s)) = \frac{\mathbf{r}(s) \cdot \mathbf{r}^{\text{av}}(s')}{\sum_{s''} \mathbf{r}(s) \cdot \mathbf{r}^{\text{av}}(s'')} \tag{C.28}$$

where s'' is a dummy variable. (This is called dot product decoding in Fig. 2.18.) One ends up, then, with a table of conjoint probabilities $\mathrm{P}(s, s')$, and another table obtained by selecting for each trial the most likely (or predicted) single stimulus s^{p}, $\mathrm{P}(s, s^{\text{p}})$. 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}):

$$< I_{\text{p}} > = \sum_{s \in S} \sum_{s' \in S} \mathrm{P}(s, s') \log_2 \frac{\mathrm{P}(s, s')}{\mathrm{P}(s)\mathrm{P}(s')} \tag{C.29}$$

and

$$< I_{\text{ml}} > = \sum_{s \in S} \sum_{s^{\text{p}} \in S} \mathrm{P}(s, s^{\text{p}}) \log_2 \frac{\mathrm{P}(s, s^{\text{p}})}{\mathrm{P}(s)\mathrm{P}(s^{\text{p}})} \, . \tag{C.30}$$

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

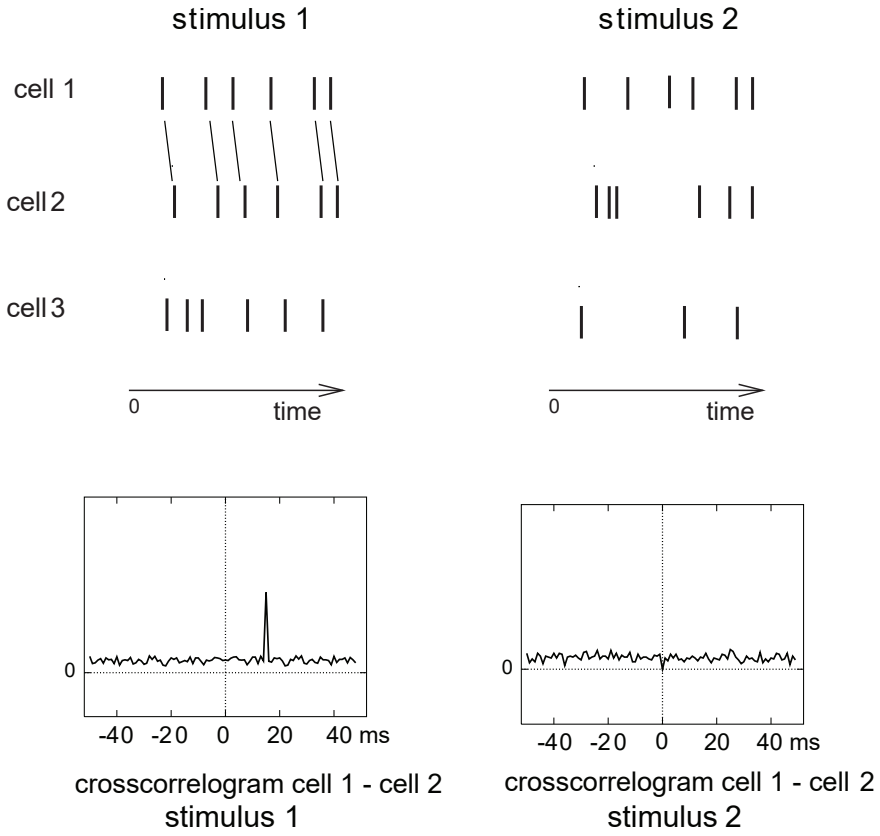


Fig. C.2 Stimulus-dependent cross-correlations between the spike trains of different neurons might encode information. The responses of three cells to two different stimuli are shown on one trial. Cell 3 reflects which stimulus was shown in the number of spikes produced, and this can be measured as spike 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. (From Rolls, E. T. and Treves.A. (2011) The neuronal encoding of information in the brain. Progress in Neurobiology 95: 448–490. © Elsevier Ltd.)

trains of three neurons are shown after the presentation of two different stimuli on one trial. 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)

$$I(s, \mathbf{r}) = \sum_{s \in S} \sum_{\mathbf{r}} \mathrm{P}(s, \mathbf{r}) \log_2 \frac{\mathrm{P}(s, \mathbf{r})}{\mathrm{P}(s)\mathrm{P}(\mathbf{r})}, \tag{C.31}$$

where $\mathrm{P}(s, \mathbf{r})$ is a probability table embodying a relationship between the variable s (here, the stimulus) and $\mathbf{r}$ (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, $\mathrm{P}(s, \mathbf{r})$, 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 (called 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 $\mathrm{P}(s, s')$, 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:

$$< I_{\mathrm{p}} > = \sum_{s \in S} \sum_{s' \in S} \mathrm{P}(s, s') \log_2 \frac{\mathrm{P}(s, s')}{\mathrm{P}(s)\mathrm{P}(s')} \tag{C.32}$$

$$= \sum_{s \in S} \mathrm{P}(s) \sum_{s' \in S} \mathrm{P}(s'|s) \log_2 \frac{\mathrm{P}(s'|s)}{\mathrm{P}(s')}. \tag{C.33}$$

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, 2016b). 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.

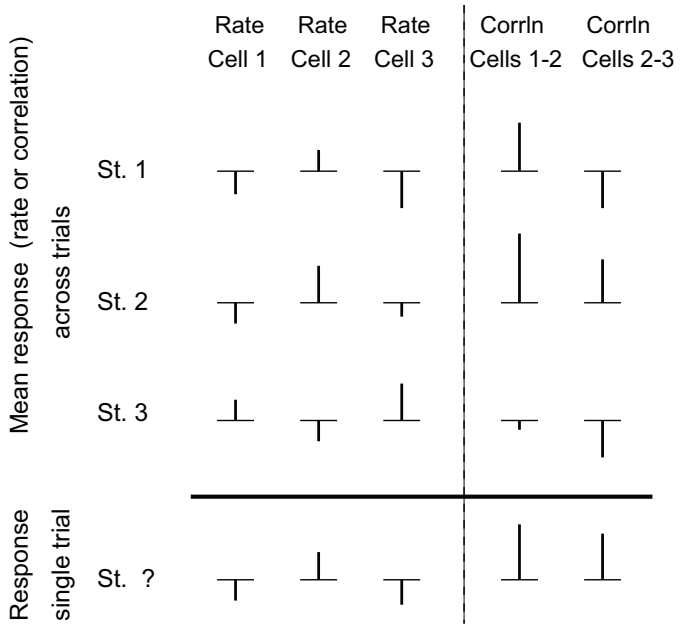


Fig. C.3 Decoding stimulus-dependent cross-correlations between the spike trains of different neurons that might encode information. The left part of the diagram shows the average firing rate (or equivalently spike count) responses of each of 3 cells (labelled as Rate Cell 1,2,3) to a set of 3 stimuli. The right two columns show a measure of the cross-correlation (averaged across trials) for some pairs of cells (labelled as Corrln Cells 1–2 and 2–3). 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. ? or s') must be estimated or decoded, using the average values across trials shown in the top part of the table. From the responses on the single trial, the most probable decoded stimulus is stimulus 2, based on the values of both the rates and the cross-correlations. (After Franco,L., Rolls, E. T., Aggelopoulos,N.C. and Treves,A. (2004) The use of decoding to analyze the contribution to the information of the correlations between the firing of simultaneously recorded neurons. Experimental Brain Research 155: 370–384. © Springer Nature.)

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 in the upper part of the table) 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 only on 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 $\mathrm{P}(s'|\mathbf{r})$ for every single trial from an estimate of the probability $\mathrm{P}(\mathbf{r}|s')$ 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).

$$\mathrm{P}(s'|\mathbf{r}) = \frac{\mathrm{P}(\mathbf{r}|s')\mathrm{P}(s')}{\mathrm{P}(\mathbf{r})}. \tag{C.34}$$

where $\mathrm{P}(\mathbf{r})$ (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:

$$\mathrm{P}(\mathbf{r}) = \sum_{s'} \mathrm{P}(\mathbf{r}|s')\mathrm{P}(s'). \tag{C.35}$$

This requires knowledge of the response probabilities $\mathrm{P}(\mathbf{r}|s')$ which can be estimated for this purpose from $\mathrm{P}(\mathbf{r}, s')$, which is equal to $\mathrm{P}(s') \prod_c \mathrm{P}(r_c|s')$, where r_c is the firing rate of cell c. I note that $\mathrm{P}(r_c|s')$ 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 $\mathrm{P}(r_c|s')$ are fitted with a Gaussian (or Poisson) distribution whose amplitude at r_c gives $\mathrm{P}(r_c|s')$. 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',

$$\mathrm{P}(s'|s) = \sum_{\mathbf{r} \in \mathrm{test}} \mathrm{P}(s'|\mathbf{r})\mathrm{P}(\mathbf{r}|s). \tag{C.36}$$

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 ${\mathrm{P}^{\mathrm{R}}}_N(s, s')$ 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 ${\mathrm{P^F}}_N(s, s^{\mathrm{p}})$ 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:

$$< I_{\mathrm{ml}} > = \sum_{s \in S} \sum_{s^{\mathrm{p}} \in S} \mathrm{P}(s, s^{\mathrm{p}}) \log_2 \frac{\mathrm{P}(s, s^{\mathrm{p}})}{\mathrm{P}(s)\mathrm{P}(s^{\mathrm{p}})} \ . \tag{C.37}$$

A detailed comparison of maximum likelihood and probability decoding is provided by Rolls, Treves and Tovee (1997b), but I 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 $\mathbf{r}$ 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 $\mathrm{P}(s, s^{\mathrm{p}})$. (It can also be used to provide percentage correct measures.)

I 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 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:

$$I(t) = \sum_{s \in S} \sum_{\mathbf{r}} \mathrm{P}(s, \mathbf{r}) \log_2 \frac{\mathrm{P}(s, \mathbf{r})}{\mathrm{P}(s) P(\mathbf{r})} \tag{C.38}$$

where $\mathbf{r}$ 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 $\mathrm{P}(s, \mathbf{r})$ 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) γ:

$\gamma_{ij}(s)$ (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 $\overline{n}_i(s)\overline{n}_j(s)$, 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)

$$\gamma_{ij}(s) = \frac{\overline{n_i(s) n_j(s)}}{(\overline{n}_i(s)\overline{n}_j(s))} - 1, \tag{C.39}$$

and is named the 'scaled cross-correlation density'. It can vary from -1 to ∞; negative $\gamma_{ij}(s)$'s indicate anticorrelation, whereas positive $\gamma_{ij}(s)$'s indicate correlation[30]. $\gamma_{ij}(s)$ 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, 1998).

[30] $\gamma_{ij}(s)$ 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 correlations between neurons in a rate-independent way. 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:

$$\rho_{ij}(s) \equiv \frac{\overline{n_i(s) n_j(s)} - \overline{n}_i(s)\overline{n}_j(s)}{\sigma_{n_i(s)} \sigma_{n_j(s)}} \simeq t \qquad \gamma_{ij}(s) = \sqrt{\overline{r}_i(s)\overline{r}_j(s)}, \tag{C.40}$$

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) ν:

$$\nu_{ij} = \frac{< \overline{n}_i(s)\overline{n}_j(s) >_s}{< \overline{n}_i(s) >_s < \overline{n}_j(s) >_s} - 1 = \frac{< \overline{r}_i(s)\overline{r}_j(s) >_s}{< \overline{r}_i(s) >_s < \overline{r}_j(s) >_s} - 1 \qquad \text{(C.41)}$$

where $\overline{r}_i(s)$ 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. ν_{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. ν_{ij} is sometimes called the "signal" correlation (Gawne and Richmond, 1993; Shadlen and Newsome, 1995, 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

$$I(t) = t\, I_t + \frac{t^2}{2}\, I_{tt} \quad . \qquad \text{(C.42)}$$

(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[31]

$$I_t = \sum_{i=1}^{C} \left\langle \overline{r}_i(s) \log_2 \frac{\overline{r}_i(s)}{\langle \overline{r}_i(s') \rangle_{s'}} \right\rangle_s \quad . \qquad \text{(C.43)}$$

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:

$$\begin{aligned} I_{tt} = {} & \frac{1}{\ln 2} \sum_{i=1}^{C} \sum_{j=1}^{C} \langle \overline{r}_i(s) \rangle_s \langle \overline{r}_j(s) \rangle_s \left[\nu_{ij} + (1+\nu_{ij}) \ln(\frac{1}{1+\nu_{ij}}) \right] \\ & + \sum_{i=1}^{C} \sum_{j=1}^{C} \left[\langle \overline{r}_i(s)\overline{r}_j(s)\gamma_{ij}(s) \rangle_s \right] \log_2(\frac{1}{1+\nu_{ij}}) \\ & + \sum_{i=1}^{C} \sum_{j=1}^{C} \left\langle \overline{r}_i(s)\overline{r}_j(s)(1+\gamma_{ij}(s)) \log_2 \left[\frac{(1+\gamma_{ij}(s)) \langle \overline{r}_i(s')\overline{r}_j(s') \rangle_{s'}}{\langle \overline{r}_i(s')\overline{r}_j(s')(1+\gamma_{ij}(s')) \rangle_{s'}} \right] \right\rangle_s . \end{aligned} \qquad \text{(C.44)}$$

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).

[31] Note that s' is used in equations C.43 and C.44 just as a dummy variable to stand for s, 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 analyzing 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 1.7). 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:

$$a^s = \frac{(\sum_{s=1}^{S} y_s/S)^2}{(\sum_{s=1}^{S} y_s^2)/S} \tag{C.45}$$

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. 2.17 on page 62). 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^{s} for this cell across all 68 stimuli was 0.69, and the response sparseness $a_{\mathrm{r}}^{\mathrm{s}}$ (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^{s} would be 0.75, while if the probability of each response decreased linearly, to reach zero at the maximum rate, a^{s} 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

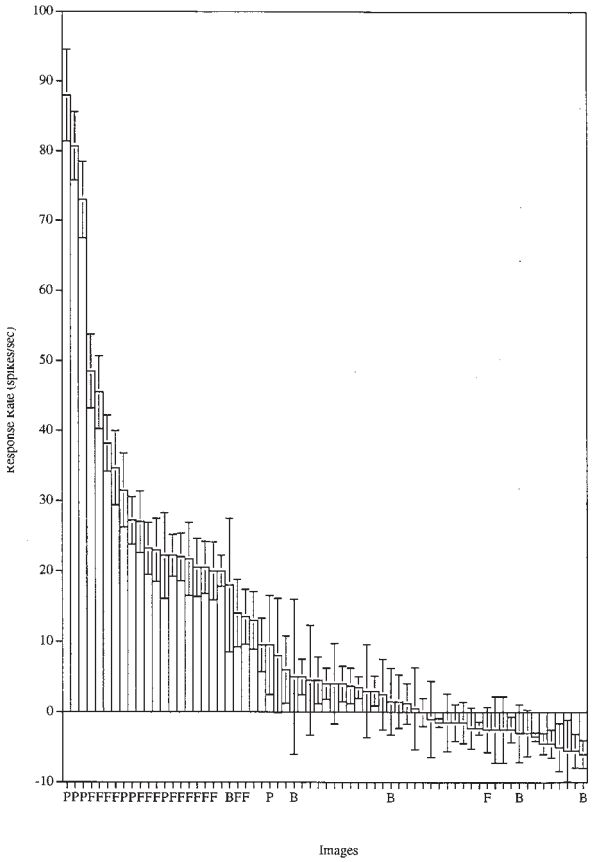


Fig. C.4 Single neuron sparseness. Firing rate distribution of a single neuron in the temporal visual cortex to a set of 23 face (F) and 45 non-face images of natural scenes. The firing rate to each of the 68 stimuli is shown. The neuron does not respond to just one of the 68 stimuli. Instead, it responds 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, E. T. and Tovee,M.J. (1995) Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. Journal of Neurophysiology 73: 713–726.)

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 that 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

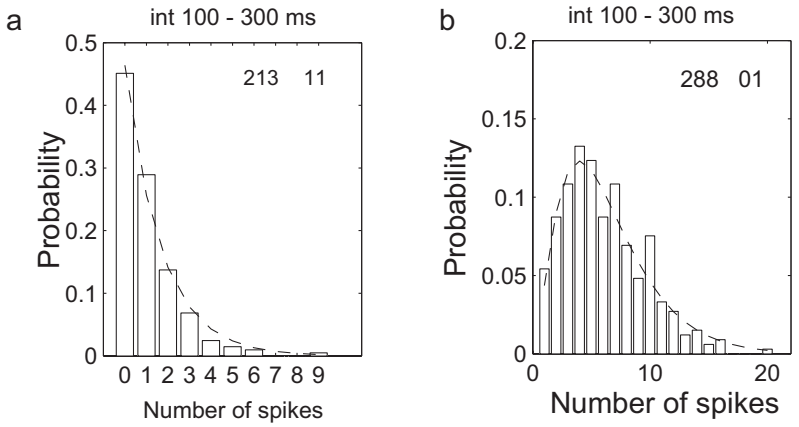


Fig. C.5 Firing rate probability distributions. These are shown for two neurons in the inferior temporal visual cortex tested with a set of 20 face and non-face stimuli. (a) A neuron with a good fit to an exponential probability distribution (dashed line). (b) A neuron that did not fit an exponential firing rate distribution (but which could be fitted by a gamma distribution, dashed line). The firing rates were measured in an interval 100–300 ms after the onset of the visual stimuli, and similar distributions are obtained in other intervals. (After Franco,L., Rolls, E. T., Aggelopoulos,N.C. and Jerez,J.M. (2007) Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. Biological Cybernetics 96: 547-560. © Springer Nature.)

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 approx-

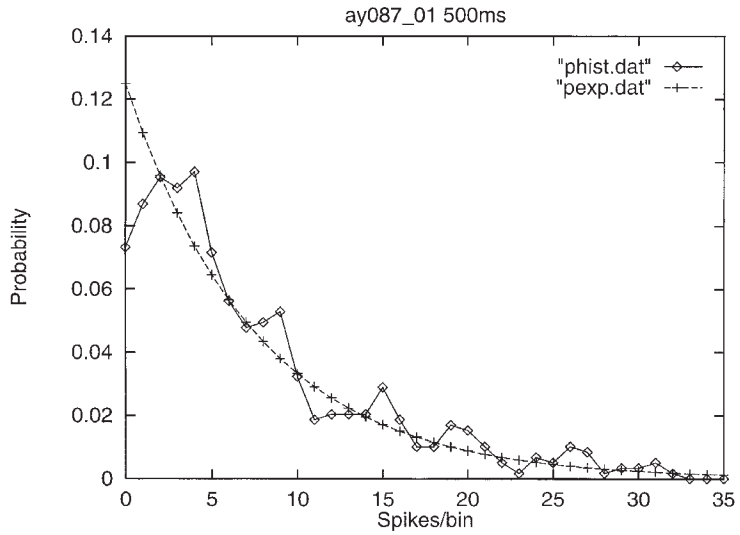


Fig. C.6 Firing rate probability distribution for a single neuron, using natural world image statistics. The probability of different firing rates measured in short (e.g. 100 ms or 500 ms) time windows of a temporal cortex neuron calculated over a 5 min period in which the macaque watched a video showing natural scenes, including faces. An exponential fit (+) to the data (diamonds) is shown. (After Baddeley,R.J., Abbott,L.F., Booth,M.J.A., Sengpiel,F., Freeman,T., Wakeman,E.A., and Rolls, E. T. (1997) Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proceedings of the Royal Society B 264: 1775-1783.)

imately 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

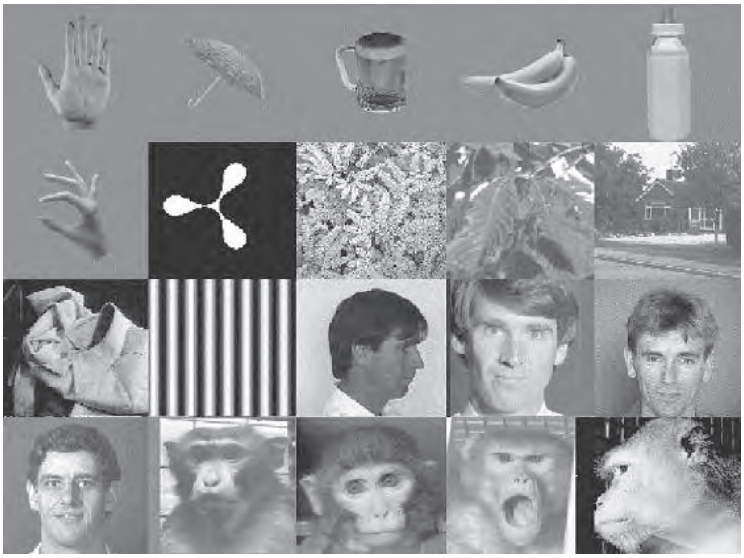


Fig. C.7 The set of 20 stimuli used to investigate the tuning of inferior temporal cortex neurons by Franco, Rolls, Aggelopoulos and Jerez 2007. These objects and faces are typical of those encoded in the ways described here by inferior temporal cortex neurons. The code can be read off simply from the firing rates of the neurons about which object or face was shown, and many of the neurons have invariant responses. (After Franco,L., Rolls, E. T., Aggelopoulos,N.C. and Jerez,J.M. (2007) Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. Biological Cybernetics 96: 547-560. © Springer Nature.)

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 shape of the exponential distribution, and the fact that the exponential distribution is at least a close approximation to the firing rate distribution of some real cortical neurons implies that the sparseness of the cortical representation of stimuli is kept under precise control. The utility of this may be to ensure that any neuron receiving from this representation can perform a dot product operation between its inputs and its synaptic weights that produces

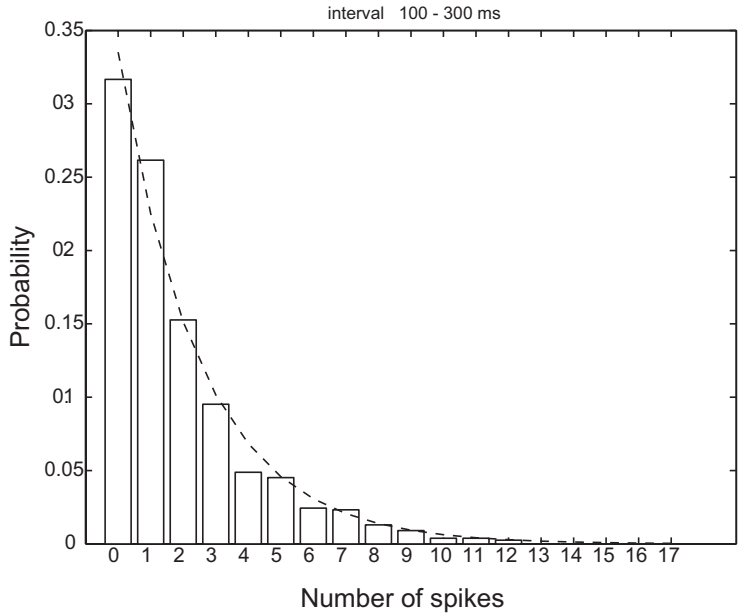


Fig. C.8 The firing rate probability distribution of single neurons is approximately exponential. An exponential firing rate probability distribution obtained by pooling the firing rates of a population of 41 inferior temporal cortex neurons tested to a set of 20 face and non-face stimuli. The firing 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,L., Rolls, E. T., Aggelopoulos,N.C. and Jerez,J.M. (2007) Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. Biological Cybernetics 96: 547–560. © Springer Nature.)

similarly distributed outputs; and that the information being represented by a population of cortical neurons is kept high. It is interesting to realize that the representation that is stored in an associative network (see Appendix B) may be more sparse than the 0.5 value for an exponential firing rate distribution, because the non-linearity of learning introduced by the voltage dependence of the NMDA receptors (see Appendix B) effectively means that synaptic modification in, for example, an autoassociative network will occur only for the neurons with relatively high firing rates, i.e. for those that are strongly depolarized.

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, 2016b) (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 et al., 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

$$a^{\mathrm{p}} = \frac{(\sum_{n=1}^{N} y_n/N)^2}{(\sum_{n=1}^{N} y_n^2)/N} \tag{C.46}$$

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). I 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, Rolls, Aggelopoulos and Jerez (2007).) The population sparseness a^{p} of this normalized (i.e. scaled) set of firing rates is 0.77.

Figure C.9b 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

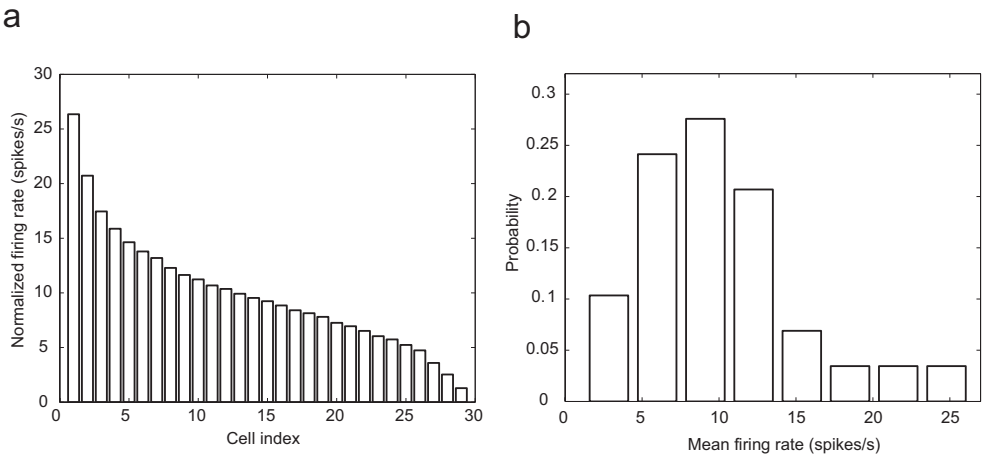


Fig. C.9 Population sparseness. (a) The firing rates of a population of inferior temporal cortex neurons to any one stimulus from a set of 20 face and non-face stimuli. The rates of each neuron were normalized to the same average value of 10 spikes/s, then for each stimulus, the cell firing rates were 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,L., Rolls, E. T., Aggelopoulos,N.C. and Jerez,J.M. (2007) Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. Biological Cybernetics 96: 547–560. © Springer Nature.)

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 ± 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 $< ... >$ 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).

I 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, 2007c; Rolls and Stringer, 2006) (see Chapter 2), and to produce not very sparse distributed representations in order to maximize the information represented (see Table B.2 on page 662). 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 ± 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 ¡a^s¿ was 0.34 ± 0.13 (sd), and the mean population sparseness a^p was 0.33 ± 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 ¡a^s¿ and ¡a^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 - 10ºC, 37ºC and 42ºC; flavour - blackcurrant juice; viscosity - carboxymethyl-cellulose 10 cPoise, 100 cPoise, 1000 cPoise and 10000 cPoise; fatty / oily - single cream, vegetable oil, mineral oil, silicone oil (100 cPoise), 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 a^s across 58 neurons was 0.745 and of a^p (normalized) was 0.708. Further analysis of data described by Rolls, Verhagen and Kadohisa (2003e), 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 a^s across 30 neurons was 0.625 and of a^p was 0.611. Further analysis of data described by Kadohisa, Rolls and Verhagen (2005b) showed that in the amygdala the mean value of a^s across 38 neurons was 0.811 and of a^p was 0.813. Thus in all these cases, the mean value of a^s is close to that of a^p, and weak ergodicity is implied. The values of a^s 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.

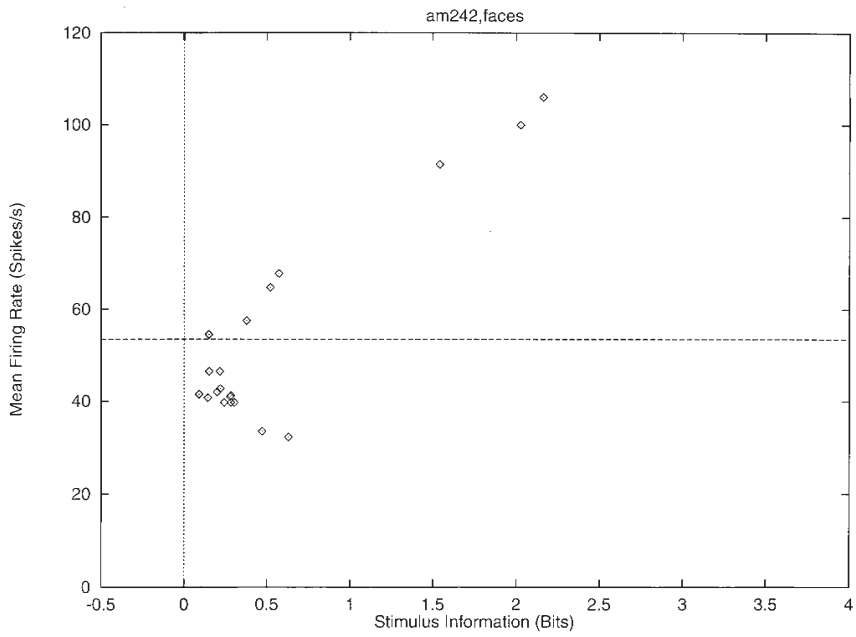


Fig. C.10 The stimulus-specific information $I(s, R)$ available in the response of the same single neuron as in Fig. C.4 about each of the stimuli in the set of 20 face stimuli (abscissa), with the firing rate of the neuron to the corresponding stimulus plotted as a function of this on the ordinate. The horizontal line shows the mean firing rate across all stimuli. (Reproduced from Rolls, E. T., Treves, A., Tovee, M. and Panzeri, S. (1997) Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. Journal of Computational Neuroscience 4: 309–333. © Springer Nature.)

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. 2.16, 2.17 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

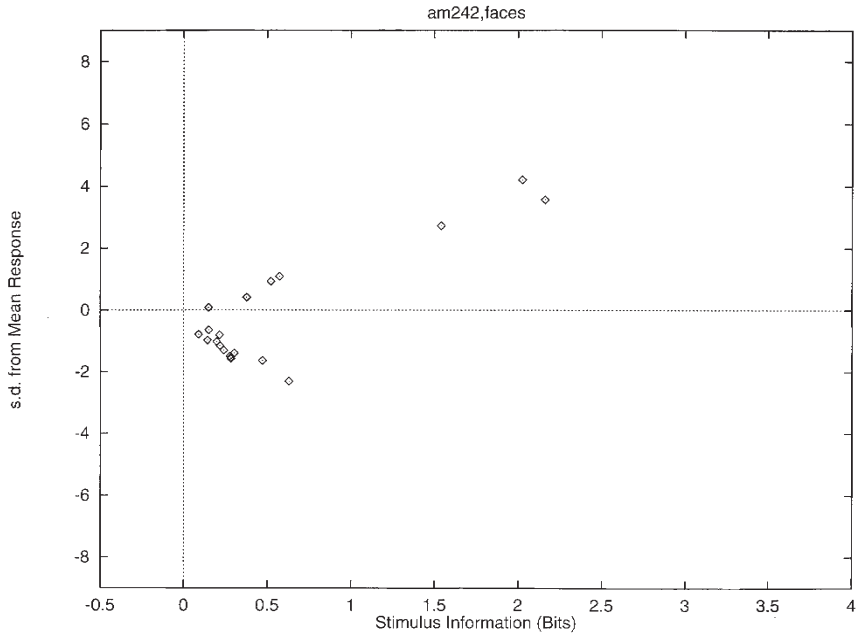


Fig. C.11 The relation for a single cell between the number of standard deviations the response to a stimulus was from the average response to all stimuli (see text, z score) plotted as a function of $I(s, R)$, the information available about the corresponding stimulus, s. (Reproduced from Rolls, E. T., Treves, A., Tovee, M. and Panzeri, S. (1997) Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. Journal of Computational Neuroscience 4: 309–333. © Springer Nature.)

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 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.

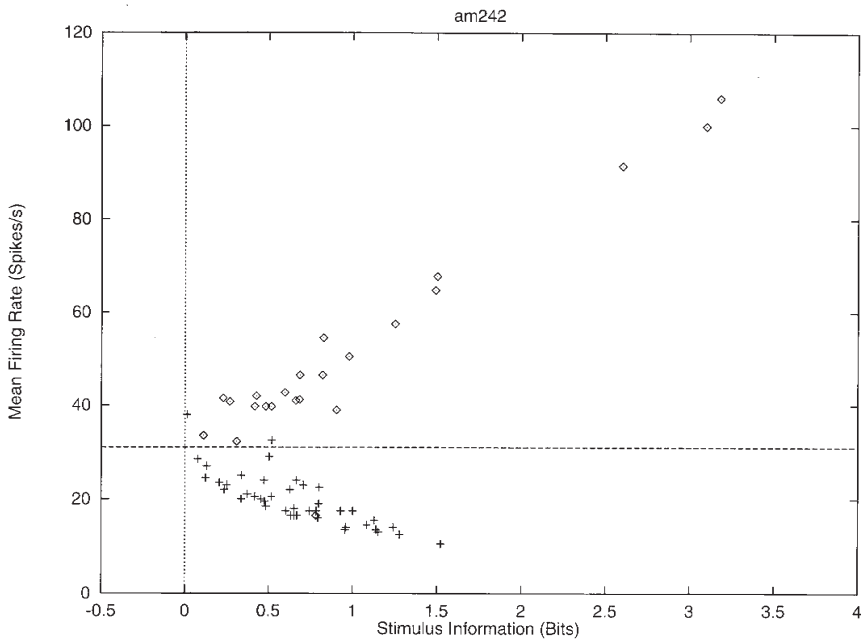


Fig. C.12 The information $I(s, R)$ available in the response of the same neuron about each of the stimuli in the set of 23 face and 42 non-face stimuli (abscissa), with the firing rate of the neuron to the corresponding stimulus plotted as a function of this on the ordinate. The 23 face stimuli in the set are indicated by a diamond, and the 42 non-face stimuli by a cross. The horizontal line shows the mean firing rate across all stimuli. (Reproduced from Rolls, E. T., Treves, A., Tovee, M. and Panzeri, S. (1997) Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. Journal of Computational Neuroscience 4: 309–333. © Springer Nature.)

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 (64×64) 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 pro-

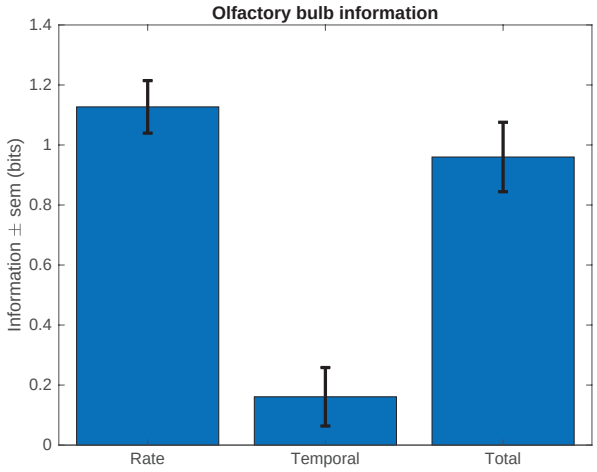


Fig. C.13 The information about which of 6 odours was presented from the mouse olfactory bulb glomeruli rates, latency ('Temporal'), and from both the rates and time courses ('Total'). The mean and standard error of the information was measured in 10 experiments in each of which the information from a population of 9–57 glomeruli was measured with the multiple cell information algorithm described by Rolls, Treves and Tovee (1997b). (Data from Verhagen,J.V., Baker,K.L., Vasan,G., Pieribone,V.A. and Rolls, E. T. (2020) Encoding in the olfactory bulb.)

cedures 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.

Evidence that is consistent with what has just been described for the visual system has

been found in another brain system, the olfactory system. We measured the information available about which of six odors had been presented by measuring the responses of populations of 9–57 glomeruli in the olfactory bulb of the mouse (Verhagen, Baker, Vasan, Pieribone and Rolls, 2020) (see Section 5.2.1). We found that although there was a little information in the time course of the response (effectively the latency), there was much more information in the activity (reflecting the neuronal firing rates) of the glomeruli about which odor had been presented (Fig. C.13). Importantly, the total information from both the rates and the latencies was not greater than that from the rates alone, showing that the latency information is redundant with respect to the rate information. (The total information from the rates and latencies was a little lower than from the rates alone, because the latency values added noise to what could be decoded from the rates.) Thus in this system too, the information is encoded in the rates, and useful extra information is not provided by the latency / time course of the neural response.

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

$$\mathrm{d}I(s)/\mathrm{d}t = r_s \log_2(r_s/ < r >) + (< r > -r_s)/\ln 2, \tag{C.47}$$

where $< r >$ 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.14. 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

$$\chi = \sum_s \mathrm{P}(s)(r_s/ < r >) \log_2(r_s/ < r >) \tag{C.48}$$

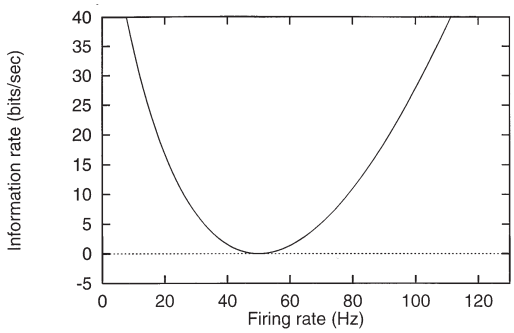


Fig. C.14 Time derivative of the stimulus-specific information as a function of firing rate, for a cell firing at a grand mean rate of 50 Hz. For different grand mean rates, the graph would simply be rescaled.

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 $\log_2(1/a)$

$$0 < \chi < \log_2(1/a), \tag{C.49}$$

where a is the single neuron sparseness a^s defined in Section C.3.1.1. For mean rates r_s distributed in a nearly binary fashion, χ is close to its upper limit $\log_2(1/a)$, whereas for mean rates that are nearly uniform, or at least unimodally distributed, χ 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, 2016b; Rolls and Treves, 2011)), and as a consequence χ 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

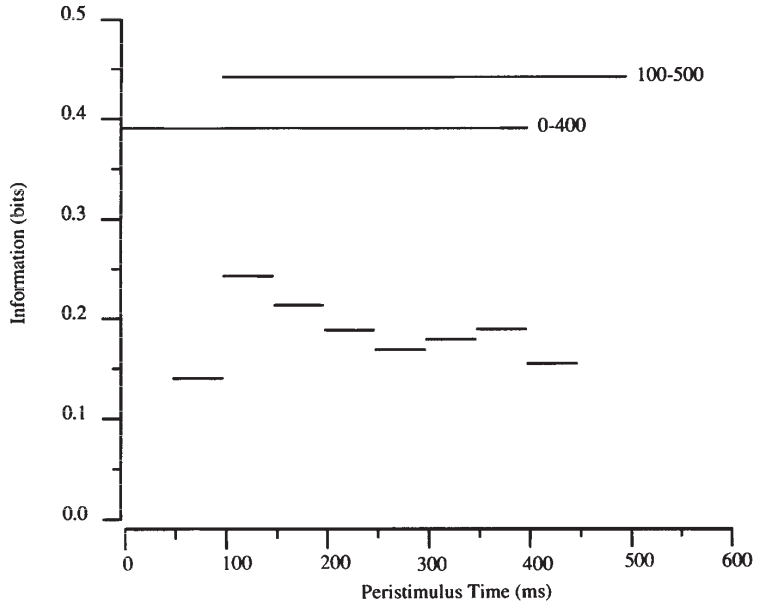


Fig. C.15 The average information I(S,R) available in short temporal epochs (50 ms as compared to 400 ms) of the spike trains of single inferior temporal cortex neurons about which face had been shown. (From Tovee,M.J. and Rolls, E.T.(1995) Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. Visual Cognition 2: 35–58. © Informa Ltd.)

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

$$\mathrm{d}^2 I/\mathrm{d}t^2 = [\ln a + (1-a)] < r >^2 /(a \ln 2), \tag{C.50}$$

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, ± 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

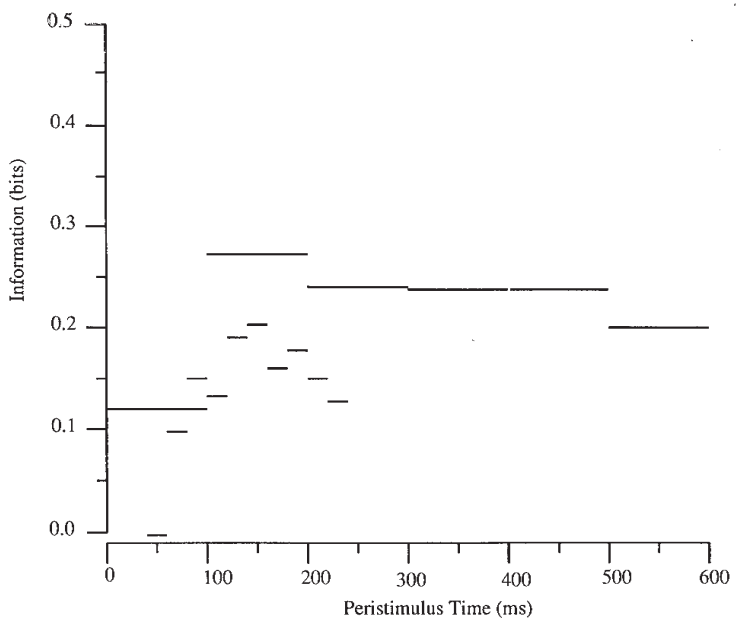


Fig. C.16 The average information I(S,R) available in short temporal epochs (20 ms and 100 ms) of the spike trains of single inferior temporal cortex neurons about which face had been shown. (From Tovee,M.J. and Rolls, E.T.(1995) Information encoding in short firing rate epochs by single neurons in the primate temporal visual cortex. Visual Cognition 2: 35–58. © Informa Ltd.)

information was measured near the start of the spike train, when the firing rate of the neuron tended to be highest (see Figs. C.15 and C.16). 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.15, C.16 and 2.16 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.17). 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

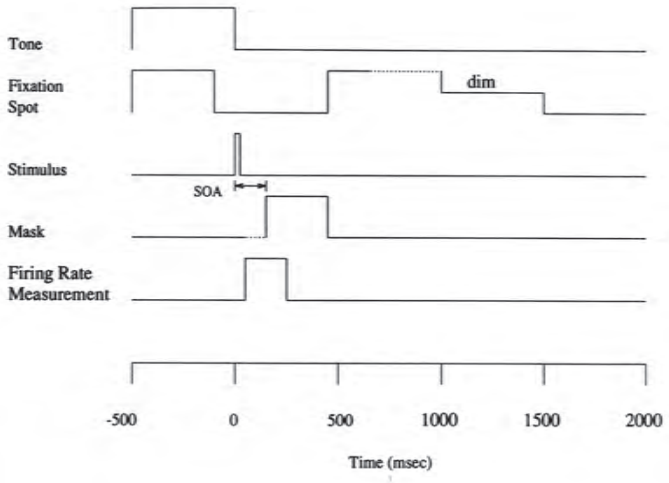


Fig. C.17 Backward masking paradigm. The visual stimulus appeared at time 0 for 16 ms. The time between the start of the visual stimulus and the masking image is the Stimulus Onset Asynchrony (SOA). A visual fixation task was being performed to ensure correct fixation of the stimulus. In the fixation 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, E. T. and Tovee,M.J. (1994) Processing speed in the cerebral cortex and the neurophysiology of visual masking. Proceedings of the Royal Society, B, 257: 9–15. © Royal Society.)

ms (see Fig. C.18). 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, 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 the code from the input neurons to initiate retrieval, whether in a pattern associator or auto-association network (see Appendix B). This is also a crucial issue for understanding how any stage of cortical processing operates, given that each stage includes associative or competitive network processes that require the code to be read before it can pass useful output to the next stage of processing (see Chapter 2; Rolls and Deco (2002); Rolls (2016b); and Panzeri, Rolls, Battaglia and Lavis (2001)). For this reason, we have performed further analyses of the speed of availability of information from neuronal firing, and the neuronal code. A rapid readout of information from any one stage of for example visual processing is important, for the ventral visual system is organized as a hierarchy of cortical areas, and the neuronal response latencies are approximately 100 ms in the inferior temporal visual cortex, and 40–50 ms in the primary visual cortex, allowing only approximately 50–60 ms of processing time for V1–V2–V4–inferior temporal cortex (Baylis, Rolls and Leonard, 1987; Nowak and Bullier, 1997; Rolls and Deco, 2002). There is much evidence that the time required for each stage

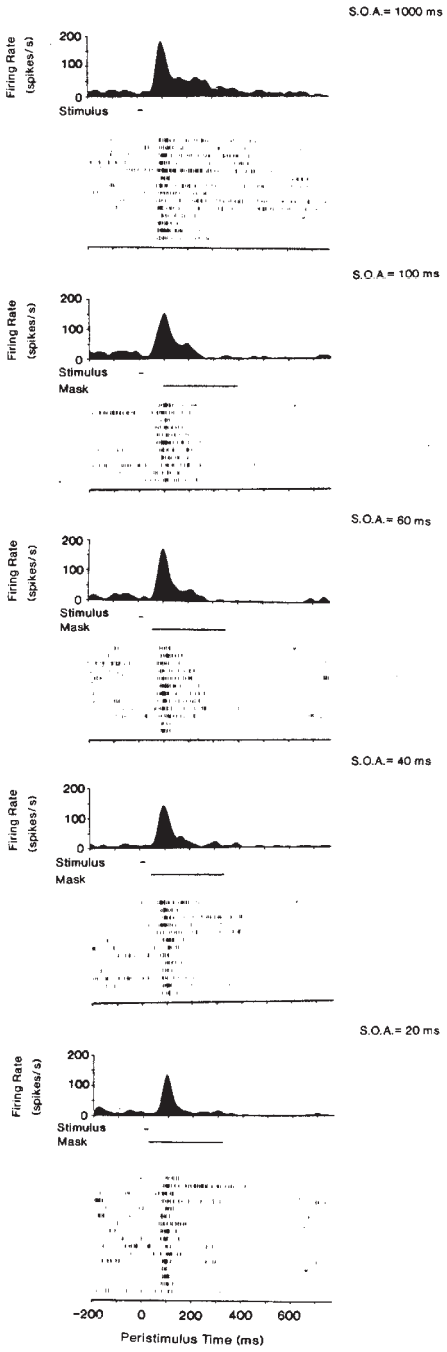


Fig. C.18 Firing of a temporal cortex cell to a 20 ms presentation of a face stimulus when the face was followed with different stimulus onset asynchronies (SOAs) by a masking visual stimulus. At an SOA of 20 ms, when the mask immediately followed the face, the neuron fired for only approximately 30 ms, yet identification above change (by 'guessing') of the face at this SOA by human observers was possible. (After Rolls, E. T. and Tovee,M.J. (1994) Processing speed in the cerebral cortex and the neurophysiology of visual masking. Proceedings of the Royal Society, B, 257: 9–15; and Rolls, E. T., Tovee,M.J., Purcell,D.G., Stewart,A.L. and Azzopardi,P. (1994) The responses of neurons in the temporal cortex of primates, and face identification and detection. Experimental Brain Research 101: 473–484. © Springer Nature.)

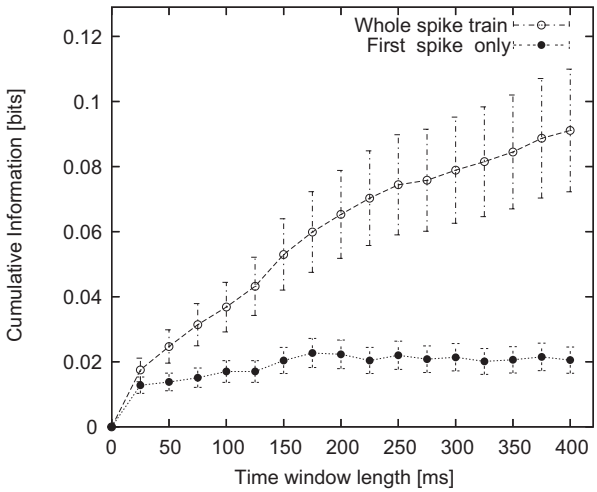


Fig. C.19 Speed of information availability in the inferior temporal visual cortex from single neurons. Cumulative single cell information from all spikes and from the first spike with the analysis starting at 100 ms after stimulus onset. The mean and sem over 21 neurons are shown. (After Rolls, E. T., Franco,L., Aggelopoulos,N.C. and Perez,J.M. (2006) Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. Vision Research 46: 4193–4205. © Elsevier Ltd.)

of processing is relatively short. For example, in addition to the evidence already presented, visual stimuli presented in succession approximately 15 ms apart can be separately identified (Keysers and Perrett, 2002); and the reaction time for identifying visual stimuli is relatively short and requires a relatively short cortical processing time (Rolls, 2003; Bacon-Mace et al., 2005).

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 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, 2016b). 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.19. 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

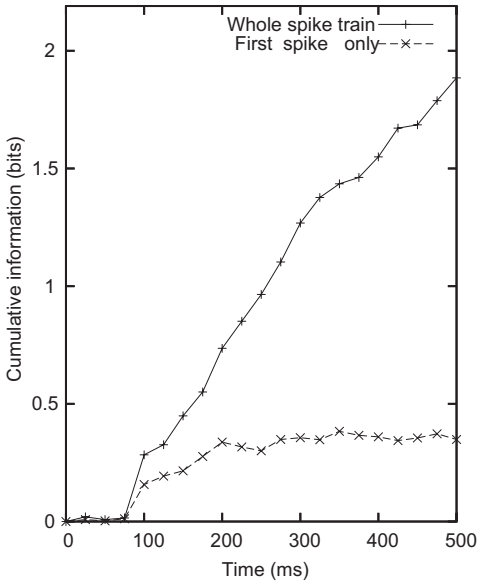


Fig. C.20 Speed of information availability in the inferior temporal visual cortex from a population of neurons. Cumulative multiple cell information from all spikes and first spike starting at the time of stimulus onset (0 ms) for the population of 21 neurons about the set of 20 stimuli. (After Rolls, E. T., Franco,L., Aggelopoulos,N.C. and Perez,J.M. (2006) Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. Vision Research 46: 4193–4205. © Elsevier Ltd.)

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.20 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 11.4.1), and might produce the advantage shown in Fig. C.20 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 2). 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.21 (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.22 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

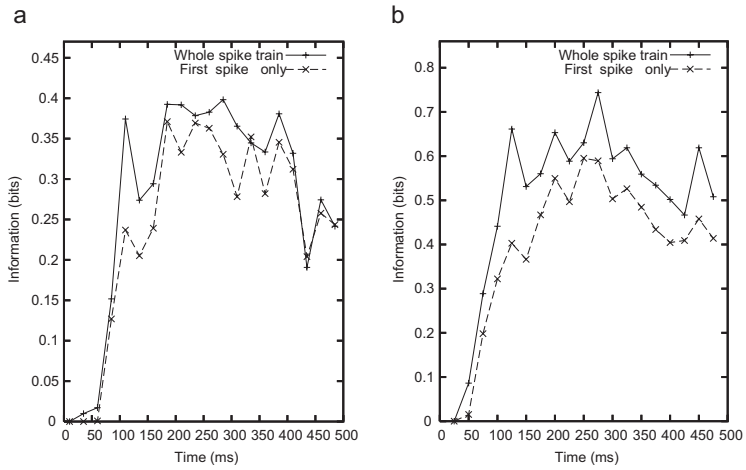


Fig. C.21 Speed of information availability in the inferior temporal visual cortex from a population of neurons. (a) Multiple cell information from all spikes and 1 spike in 20 ms time windows taken at different post-stimulus times starting at time 0. (b) Multiple cell information from all spikes and 1 spike in 50 ms time windows taken at different post-stimulus times starting at time 0. (After Rolls, E. T., Franco,L., Aggelopoulos,N.C. and Perez,J.M. (2006) Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. Vision Research 46: 4193–4205. © Elsevier Ltd.)

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.21).

To show how the information increases with the number of neurons in the ensemble in these short epochs, we show in Fig. C.22 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, Franco, Aggelopoulos and Reece, 2003b,b). It is also clear from Fig. C.22 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.22 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 ($p_i 0.05$) responses to the stimulus set where it

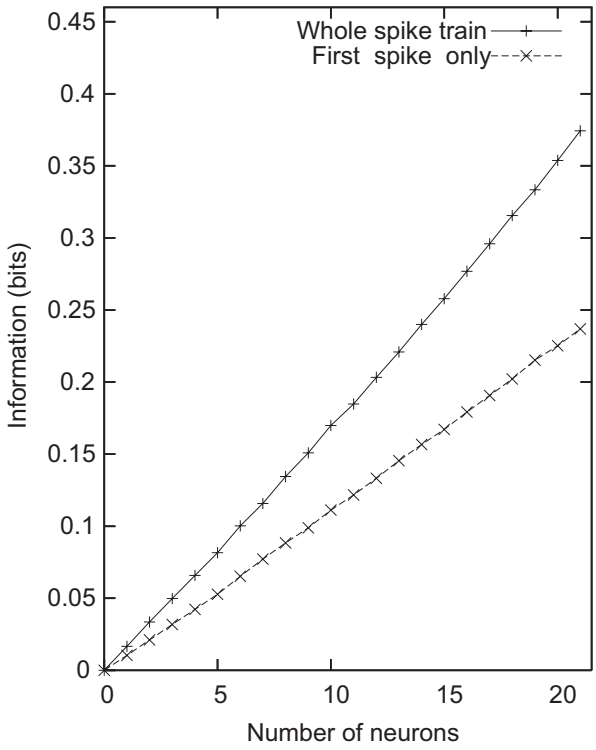


Fig. C.22 Speed of information availability in the inferior temporal visual cortex from populations of neurons. Multiple cell information from all spikes and 1 spike in a 20 ms time window starting at 100 ms after stimulus onset as a function of the number of neurons in the ensemble. (After Rolls, E. T., Franco,L., Aggelopoulos,N.C. and Perez,J.M. (2006) Information in the first spike, the order of spikes, and the number of spikes provided by neurons in the inferior temporal visual cortex. Vision Research 46: 4193–4205. © Elsevier Ltd.)

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 $\pm$ 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.22 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, 2005d; Deco, Rolls and Horwitz, 2004; Deco and Rolls, 2005b; Rolls and Deco, 2002; Rolls, 2016b). 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 11.4.1), 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.19 and C.20 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. 2.18 on page 64, and in Fig. C.23 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

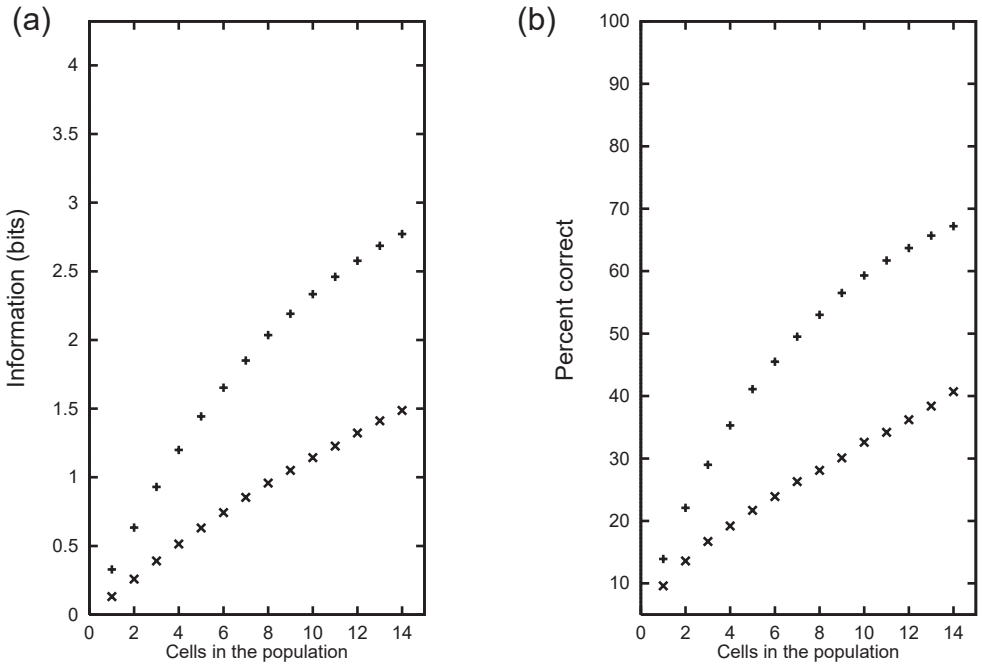


Fig. C.23 The information available from populations of neurons. (a) The information available about which of 20 faces had been seen that is available from the responses measured by the firing rates in a time period of 500 ms (+) or a shorter time period of 50 ms (x) of different numbers of temporal cortex cells. (b) The corresponding percentage correct from different numbers of cells. (From Rolls, E. T., Treves, A. and Tovee, M. J. (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. Experimental Brain Research 114: 149–162. © Springer Nature.)

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 approximately exponentially with the number of neurons in the population, as illustrated in Figs. 2.19 and C.23.

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 $T_i, i = 1, ..., C$:

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

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

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

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

$$N = 1 - R' \tag{C.53}$$

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

$$I(C) \approx H\{1 - \exp[C \ln(1 - y)]\} \tag{C.54}$$

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 2. 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 sig-

nificantly 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. 2.18, C.23 and 2.19) 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.24. 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

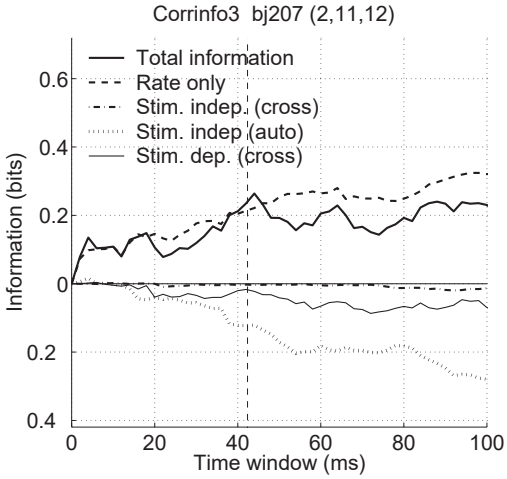


Fig. C.24 Most of the information is present in the firing rates, with little in stimulus-dependent cross-correlations between the spikes of different neurons. A typical result from the information analysis described in Section C.2.5 on a set of 3 simultaneously recorded inferior temporal cortex neurons 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 796, 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. (From Rolls, E. T., Aggelopoulos,N.C., Franco,L., and Treves,A. (2004) Information encoding in the inferior temporal cortex: contributions of the firing rates and correlations between the firing of neurons. Biological Cybernetics 90: 19–32. © Springer Nature.)

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 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 *simul-*

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.

rate	0.26
stimulus–dependent "noise" correlation-related, cross term	0.04
stimulus–independent "noise" correlation-related, cross term	-0.05
total information	0.31

taneously 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, 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.25. 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, 2016b) (Chapter 2), 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

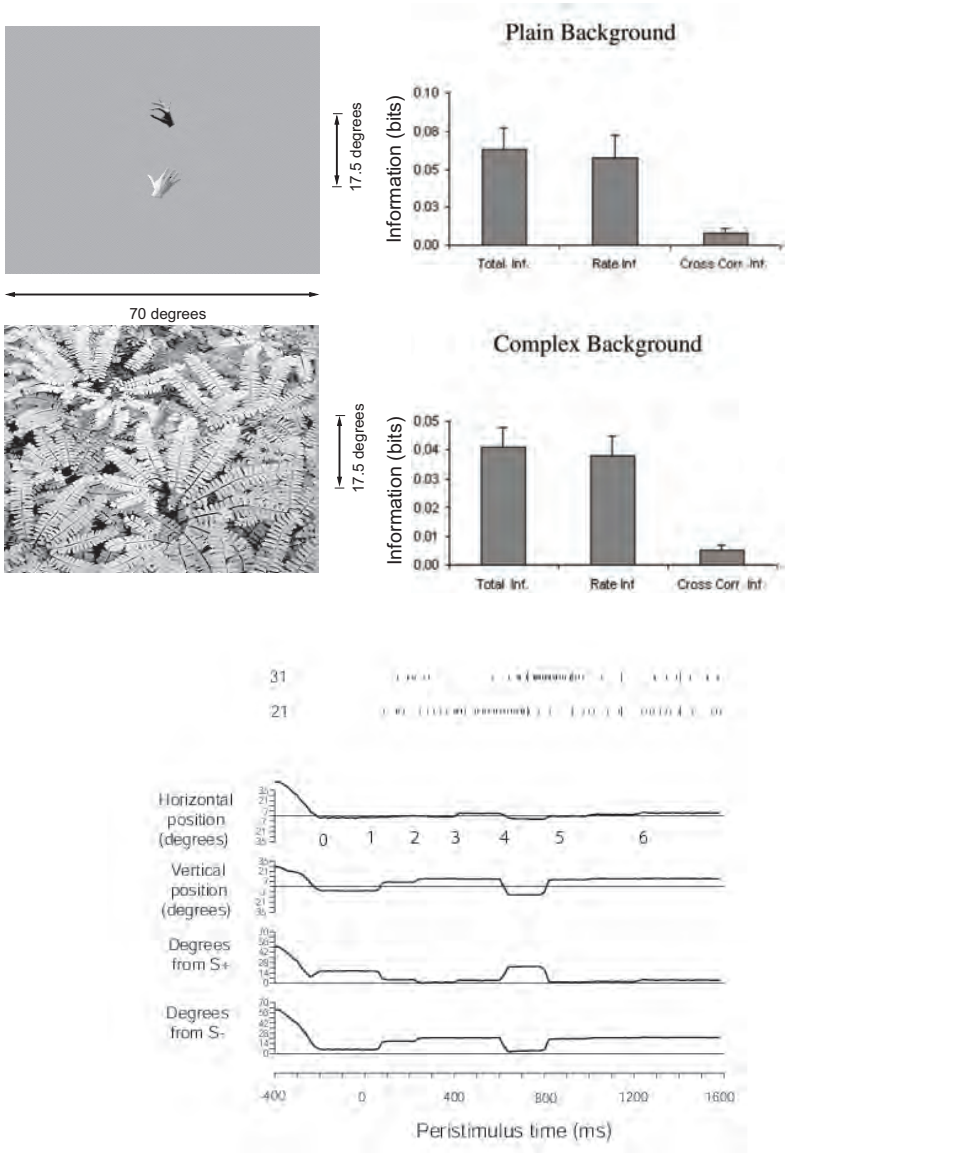


Fig. C.25 Even in complex natural scenes, most of the information is in the firing rates and not in the cross-correlations between the neurons. Left: the objects against the plain background, and in a natural scene. Right: the information available from the firing rates (Rate Inf) or from stimulus-dependent 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,N.C., Franco,L. and Rolls, E. T. (2005) Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. Journal of Neurophysiology 93: 1342–1357. © American Physiological Society.)

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). Further, it was shown that there was little redundancy (less than 6%) between the information 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, 2016b, 2012d) (Chapter 2), and not by temporal synchrony and attention (Malsburg, 1990; Singer, Gray, Engel, Konig, Artola and Brocher, 1990; Abeles, 1991; Hummel and Biederman, 1992; Singer and Gray, 1995; Singer, 1999, 2000; Rolls, 2016b). In any case, the computational point made in Section 2.7.5.1 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 2.7.10).

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.15 and C.16 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. 2.16 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 9.2.6 and Chapter 2). Moreover, information theory has helped to show that for many neurons considerable invariance in the representa-

tions 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 2 and Section 9.2.6).

Information theory has also enabled the information available in neuronal representations to be compared with that available to the whole animal in its behaviour (Zohary, Shadlen and Newsome, 1994) (but see Section C.3.6).

Finally, information theory also provides a metric for directly comparing the information available from neurons in the brain (see Chapter 2 and this Appendix) with that available from single neurons and populations of neurons in simulations of visual information processing (see Chapter 2).

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, 2015c, 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 $\mathrm{P}(s_i)$ is

$$I(s_i) = \log_2(1/\mathrm{P}(s_i)) = -\log_2 \mathrm{P}(s_i). \tag{C.55}$$

(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**,

$$H(S) = -\sum_i \mathrm{P}(s_i) \log_2 \mathrm{P}(s_i) \tag{C.56}$$

(or *a priori* entropy).

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

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

$$\mathrm{P}(s|s') = \frac{\mathrm{P}(s'|s)\mathrm{P}(s)}{\mathrm{P}(s')} \tag{C.57}$$

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

5. **Mutual information**. Prior to reception of s', the probability of the input symbol s was $\mathrm{P}(s)$. This is the *a priori* probability of s. After reception of s', the probability that the input symbol was s becomes $\mathrm{P}(s|s')$, 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**

$$\begin{aligned} I(S,S') &= \sum_{s,s'} \mathrm{P}(s,s')\{\log_2[1/\mathrm{P}(s)] - \log_2[1/\mathrm{P}(s|s')]\} \\ &= \sum_{s,s'} \mathrm{P}(s,s') \log_2[\mathrm{P}(s|s')/\mathrm{P}(s)]. \end{aligned} \tag{C.58}$$

Alternatively,

$$I(S,S') = H(S) - H(S|S'). \tag{C.59}$$

$H(S|S')$ 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.
4. Code translated into Matlab that was used for the single neuron information analysis (Rolls, Treves, Tovee and Panzeri, 1997d) and multiple single neuron information analysis Rolls, Treves and Tovee (1997b) is described in Section D.7 and is available at https://www.oxcns.org/software.