An Adaptive Strategy for Sensory Processing

Thesis

An Adaptive Strategy for Sensory Processing

Rishabh Raj, B. Tech

A thesis submitted in fulfillment of the requirement of the Open University for the Degree of Doctor of Philosophy

School of Life, Health, and Chemical Sciences

Stowers Institute for Medical Research

1000 E 50th Street, Kansas City, Missouri, United States of America

An Affiliated Research Center of the Open University, United Kingdom

December 25, 2020

Abstract

Recognizing objects and detecting associations among them is essential for the survival of organisms. The ability to perform these tasks is derived from the representations of objects obtained through processing information along sensory pathways. Our current understanding of sensory processing is based on two sets of foundational theories – The Efficient Coding Hypothesis and hierarchical assembly of object representations. These theories suggest that sensory processing aims to identify independent features of the environment and progressively represent objects in terms of comprehensive combinations of these features. Separately, the two sets of theories have successfully explained the detection of associations and perceptual invariance, respectively; however, reconciling them together in one unified theory has remained challenging. Independent features are deemed essential for detecting association by the Efficient coding hypothesis, but to achieve consistency in representations, multiple comprehensive structures corresponding to the same object must be hierarchically assembled, ignoring independence among such structures.

Here we propose an alternative framework for sensory processing in which the system, instead of finding the truly independent components of the environment, aims to represent objects based on their most informative structures. Using theoretical arguments, we show that following such a strategy allows the system to efficiently represent sensory cues without necessarily acquiring knowledge about statistical properties of all possible inputs. Through mathematical simulations, we find that the framework can describe the known characteristics of early sensory processing stages and permits consistent input representations observed at later stages of processing. We also demonstrate that the framework can be implemented in a biologically plausible neuronal circuit and explain aspects of experience and learning from corrupted inputs. Thus, this framework provides a novel perspective and a unified description of sensory processing in its entirety.

Acknowledgements

First of all, I would like to thank my advisor Dr. C. Ron Yu for giving me this wonderful opportunity to work on this project. His guidance and support have not just been helpful in completing this project but have shaped my thinking as a whole. I cannot imagine doing this project without the long discussions and critical thinking periods I had with him. I would also like to thank past and present members of the lab, who have helped and supported me in various ways throughout my time in the lab.

I would like to thank my committee members Dr. Kausik Si, Dr. Paul Kulesa, and Dr. Jay Unruh for their helpful suggestions and discussions throughout my time in graduate school. I would also like to thank Dr. Paul Trainor for his support and suggestions as a third-party monitor.

My experience with open university procedures would not have been as easy without Dr. Leanne Wiedemann. Her guidance through all the procedures starting from induction till thesis discussion have been extremely helpful. I would also like to thank Lisa Hodges for keeping all my paper works and meeting on schedule.

I would especially like to thank Mr. Dar Dahlen who was working on this project before me and has guided me through the crucial early stages of the project. I would also like to express my gratitude to Mr. Kyle Duyck for sharing the face database, Mr. Malcolm Cook for helping me solve various problems related to MATLAB and Mr. Richard Alexander for helping me with processing of images.

Lastly, I would like to thank my family members for their unconditional love and support. Nothing would have been possible without them.

Table of Contents

List of Figures

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

List of Tables

CHAPTER 1

Theories of sensory processing

Table of Contents

1.1. Introduction

An organism must know its environment and understand the rules by which it functions (Barlow 1991, Barlow 1994, Barlow 1989). For example, an organism needs to distinguish a predator from a potential mate and realize that proximity to a predator is avoided, whereas companionship of mates is preferred. In another situation, it must identify its food sources and learn the cues that indicate its presence. All this knowledge is embedded in two fundamental aspects of its surroundings – the identity of objects and relationships between them. While objects' identities comprise answers to most of the *"what"* questions (like *what is a predator?* or *what is food?*), awareness of the relationship between objects develops intuitions about the rules of the environment. For instance, knowing that a localized movement in grass twigs is indicative of a lion may introduce a law to an antelope that oddly moving grasses are to be avoided. Insights about such rules are necessary for making favorable decisions; deciding to flee before an actual encounter with the lion and can play a decisive role in survival. In this regard, recognizing objects and identifying how they are related to one another can be considered as the two essential tasks that the organism must perform to endure.

Organisms collect the information necessary to perform these essential tasks through their sensory systems. Sensory neurons, which are the basic structural and functional units of sensory systems, pick various information from the surroundings and relay it to centers like the brain and ganglia as electrical impulses (Golgi 1906, Ramon y Cajal 1906). These impulses are analyzed and transformed in different ways to extract relevant pieces of information (Barlow 1972).

However, gathering information pertinent to recognizing objects is a challenge. Owing to differences in lighting, pose, location, surroundings, etc., several different circumstances of encountering objects may arise. These circumstances introduce

inconsistencies in the sensory system's inputs, and specific information about objects available in one situation is often lost in others (Ullman 1996). Internal noise in the system also leads to variations in the impulses that make them less reliable for carrying information (Stein 1967). Nonetheless, the organism must invariably recognize objects and their relationships in different situations. It must avoid the adverse effects of the system's unreliability and compensate for the inconsistencies in inputs. In other words, object recognition must be consistent and robust.

Electrical impulses in sensory neurons carry information about the surroundings (Jacobson 1950, Jacobson 1951, Quastler 1956, Rapoport and Horvath 1960). As they travel down a neuron, these impulses induce activity in subsequent cells in the sensory pathway. Thus, depending on how neurons are connected, one neuron's activity is transformed into the activity of many others. This process of transforming neuronal firings is known as sensory processing. One can imagine that a plausible way to extract information about object identities from these impulses will be to transform them into object-specific patterns, i.e., transform them so that different objects induce distinct activity patterns. These activity patterns can then comprise the *"representations"* of objects, and the system can perform the subsequent task of detecting associations among different objects using them. Indeed, several studies aimed at understanding visual processing in primates and other higher organisms provide pieces of evidence supporting a pattern-based representation of objects (Perrett et al. 1982, Phillips et al. 1984, Baylis et al. 1985, DI Perrett et al. 1985, Young and Yamane 1992, Rolls and Tovee 1995). Studies also show that activities of individual neurons at successive levels of the sensory relay represent increasingly complex combinations of structural features of objects and the highest level neurons represent entire objects (Hubel and Wiesel 1962, Hubel and Wiesel 1968, Gross et al. 1969, Perrett et al. 1982, Schwartz et al. 1983, Miyashita and Chang 1988, Logothetis and Pauls 1995, Logothetis et al. 1995, Logothetis and Sheinberg 1996, Tanaka 1996). However, it is not very clear how the system decides which features or feature combinations to represent, or how it compensates for the

inconsistencies described previously to enable a robust perceptual experience, or how it further utilizes these representations to identify associations between different objects.

Theories proposed over the last several decades bridge the evident gap in our understanding of sensory processing (Attneave 1954, Barlow 1961, Marr and Nishihara 1978, Biederman 1987, Ullman and Basri 1991, Poggio and Edelman 1990). Two sets of theories form the basis of our current understanding of sensory processing: the Efficient Coding Hypothesis (Attneave 1954, Barlow 1961) and hierarchical assembly of object representation (Fukushima 1975, Fukushima and Miyake 1982, Anderson and Van Essen 1987, Wallis et al. 1993, Riesenhuber and Poggio 1999b, Rolls and Milward 2000). Adopted from Shannon's theory of communication (Shannon 1948) and proposed by Barlow and others, the Efficient coding hypothesis suggests that the sensory system should represent objects in ways that minimize loss of information and ensure efficiency in representation (Barlow 1961). While representing objects uniquely preserves information about them, making representations efficient requires individual neurons to represent independent features (Barlow 1987, Barlow 1989, Barlow et al. 1989). Independence among represented features implies that they are not associated with one another in any way, and the presence of any such feature in an object is not indicative of any other feature.

The rationale for representing independent features through individual neurons stems from realizing that any form of association between represented features constraints how neurons get activated. If two neurons represent features that always occur together, then those neurons will always be activated together. Such a representation scheme will restrict the number of different patterns that can be formed in a system and limit its ability to represent various features and objects. To illustrate it with an example, consider the task of representing different animals. As one eye's presence implies the other's existence, the two eyes form a set of features that are not independent. Therefore, having two neurons represent

them individually, renders one of the neurons obsolete. This neuron can represent some other distinguishing features of the animals.

On the other hand, beaks and eyes comprise an independent set of features. One cannot predict the presence of beaks from the knowledge about the eyes. Consequently, according to the Efficient Coding Hypothesis, different neurons should represent these features. In this manner, the hypothesis manifests a scheme of representing objects that enables efficient utilization of a system's capacity to represent objects.

This scheme's usefulness for biological systems arises from its ease in detecting associations among different objects. Representing independent features through individual neurons permits the system to track the occurrence frequencies of individual features. The system can estimate the occurrence frequencies of objects by compounding the occurrence frequencies of independent features. Ultimately, it can determine any association among objects from their occurrence frequencies (Barlow 1987, Barlow 1991). Thus, the Efficient Coding Hypothesis establishes the fundamental nature of object representations that allows identifying associations among objects.

The Efficient Coding Hypothesis has been remarkably successful in providing theoretical explanations of various aspects of sensory processing (Laughlin 1981, Atick 1992, Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997, Lewicki 2002, Smith and Lewicki 2006). In several studies, the statistical properties of the natural scenes have been analyzed to find the independent features (Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997, Van Hateren and van der Schaaf 1998). It has been shown that these independent components conform to the features that neurons in the primary visual cortices of monkeys and cats represent (Hubel and Wiesel 1962, Hubel and Wiesel 1968).

However, the representation scheme proposed under the Efficient Coding Hypothesis, in its current form, is not sufficient to explain the consistency in perceptual experiences. Considering that object recognition is based on their representation, consistent

recognition of objects requires their representations to be invariant, i.e., variation in factors like lighting, pose, location, or surroundings should not alter the representation (Gross 1985, David I Perrett et al. 1985, Hasselmo et al. 1989, Tanaka et al. 1990, Tovee et al. 1994, Tanaka 1996, Hegdé and Van Essen 2000, Hegdé and Van Essen 2003, Ito and Komatsu 2004, Brincat and Connor 2006, Hegdé and Van Essen 2007, Freiwald et al. 2009, Liu et al. 2010). Independent features from natural scenes, on the other hand, are localized edges oriented at different angles (Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997). Their detection in objects is likely to change with factors like viewing position or orientation. Therefore, in a scheme where individual neurons represent these features, the same object will activate different sets of neurons and not maintain invariance when viewed from a different angle or present in a different orientation.

Theories suggesting the hierarchical representation of objects address this invariance issue (Fukushima 1975, Fukushima and Miyake 1982, Anderson and Van Essen 1987, Wallis et al. 1993, Riesenhuber and Poggio 1999b, Rolls and Milward 2000). These theories recommend representing not just the individual features but a hierarchical assembly of features at successive levels of sensory pathway, meaning that neurons at higher levels should represent progressively complex combinations of features that neurons at lower levels represent (Fukushima and Miyake 1982, Wallis et al. 1993, Riesenhuber and Poggio 1999b). The motivation behind such an approach comes from the insight that objects' identities can be derived based on features that comprise them, i.e., one can predict an object based on a specific combination of features. Based on this line of thought, theories of hierarchical representation propose that from exponentially many feature combinations, the system should selectively learn and represent the feature combinations that it encounters (Fukushima and Miyake 1982, Poggio and Edelman 1990). The desired invariance in representing objects is achieved by representing a collection of two-dimensional projections or three-dimensional models of the same object at the higher levels such that the collection

as a whole account for any possible variation in the appearance of the object (Biederman 1987, Ullman and Basri 1991, Ullman 1996). The framework incorporates top-down mechanisms to compensate for discrepancies like missing or uninterpretable features due to occlusion of objects or noisy input to the system. In these mechanisms, higher-order neurons, representing the learned, complex feature combinations, influence the activity of lower-level neurons (Rao and Ballard 1999, Lee and Mumford 2003). Models based on this framework have been successful in explaining the results of psychophysical experiments designed to test the object recognition abilities in humans and monkeys (Bartram 1974, Jolicoeur 1985, Corballis 1988, Tarr and Pinker 1989, Bülthoff and Edelman 1992, Edelman and Bülthoff 1992, Humphrey and Khan 1992, Farah et al. 1994). The framework also explains the gradual increase in the size of the visual field represented in neurons along the visual pathway (Wallis, Rolls et al. 1993). Thus, the hierarchical assembly framework provides a thorough account for the consistent representation of objects and demonstrates a biologically plausible mechanism for perceptual invariance.

Theories of efficient coding and hierarchical assembly, in conjunction with each other, explain how the sensory system can process the information collected from its environment and utilize it to accomplish the essential functions of robustly recognizing objects and detecting associations among them. However, it is difficult to reconcile the two theories together. While the Efficient Coding Hypothesis suggests utilizing independence of represented features to detect association among objects, the hierarchical assembly approach seeks to represent multiple feature combinations originating from the same object and disregards independence among these combinations to achieve invariant representations. Moreover, the hierarchical assembly of complex feature combinations requires experience-based learning. The system needs to detect any association among features to select the set that should be combined and represented at higher levels. In contrast, the representation scheme proposed under the Efficient Coding Hypothesis demands independence among

features, limiting the system's ability to spot any association and represent complex feature combinations.

In addition to these compatibility issues, both theories have their own limitations. For example, an efficient representation of features necessitates the system to have near-complete knowledge of its environment's statistics to find independent features. For biological systems, which gradually understand their environment through experience, having such knowledge is not possible. One can argue that as the system learns about its environment, it eventually acquires a near-complete knowledge; however, this still cannot explain how the system can efficiently represent objects at the early stages of life. Similarly, though hierarchical assembly theories rely on a collection of views or models of the object to achieve invariance, it is never specified which views or models to learn out of infinitely many possible ones. In summary, while the Efficient Coding Hypothesis presents a way to detect associations among objects, it does not provide a basis for robust recognition of the objects. The hierarchical assemblies make robust recognition possible but do not specify a way to detect associations among objects.

In this work, I present a novel framework to represent objects that allows accomplishing both these tasks. In this framework, individual neurons do not represent independent or ordinary features but tend to represent the feature assemblies derived from individual objects that convey most information about them. In particular, I argue that neurons should represent only the structural components that uniquely identify objects because these are the components that convey most information. Without necessarily seeking independence among represented features, this representation framework departs from the one proposed in the classical efficient coding paradigm. Additionally, by specifying the definitive criterion for qualifying a feature combination as representable, it does not follow the traditional hierarchical assembly approach either. Using mathematical simulations, I show that the criterion for selecting representable components based on

maximum information leads to a representation scenario where the number of objects relative to the number of neurons is the determinant of the complexity of the represented component. In other words, relative numbers of objects and neurons determine whether localized features or complex assemblies of features are to be represented. This dependence of represented features' complexity on relative numbers of objects and neurons not only explains the representation of localized features in the early stages of visual processing but also removes the necessity to progressively combine features. Thus, it allows the system to achieve invariant representations of corrupted or occluded inputs without the need of a top-down signal.

Moreover, the features that contain maximum information about an object continue to do so irrespectively of the object's occurrence frequency or its relationship with other objects; therefore, this framework of representing objects eliminates the implausible requirement to know the entire statistics of the environment. In fact, by demanding the system to know only a fraction of its surroundings' possible statistics at any point in its experience, the framework is pertinent to biological systems that seek to represent a finite number of objects and adapt to new inputs. I show that the same framework of representing inputs applies to olfactory processing, enabling biologically plausible and adaptive sensory processing.

In this chapter, I review some early works done in sensory processing before the advent of information theory. Then, with a brief introduction to information theory, I will describe the key concepts in the Efficient Coding Hypothesis and hierarchical assembly framework and will give a brief overview of works based on both sets of theories.

1.2. Pre-Information theory

The study of sensory processing has been a field of great interest, and experts from different fields have contributed to its development. However, the emergence of information

theory has marked an exact inflection point in the development of concepts related to sensory processing. Here, I briefly describe some of the most prominent ideas that prevailed before applying information theory to sensory coding.

1.2.1 Hermann von Helmholtz: A German physician and physicist, Helmholtz contributed to several scientific fields, including physiology and psychology. In physiology, he is most noted for his studies of human vision and auditory systems. Helmholtz's paved the way for scientific studies of relations between measures of physical stimuli and their human perception. In his book, Handbuch der physiologischen Optik (Von Helmholtz 1867) he proposed several theories on the perception of motion, color, and depth. In the third and final volume of the same book, he introduced the idea of *unconscious inference* in which he argued that when encountering a current sensory input (apperception), the organism unconsciously compares the input to the learned concepts of the environment obtained through past experiences. The comparison results in conclusions that are manifested as the perception of the stimulus. Thus, he essentially asserted learning of the environmental structures and forming perceptions based on known structures, an idea central to the current theory of efficient coding.

1.2.2 Ernst Mach: Ernst Mach was an Austrian physicist most noted for his study of shock waves. The ratio of any speed with the speed of sound, popularly known as the Mach number, is named after him. Though a physicist, he has made some significant contributions to the studies of sensory processing. He found out that the sense of balance in humans arises from the movement of fluid inside ears (Blackmore 1972). In theoretical aspects of sensory processing, Mach introduced the concept of *economy of thoughts*. Being a physicist, he interpreted scientific laws as

constructions that make the data from the surroundings more interpretable. Extending the same idea to sensory processing, he asserted that our complex sensory experiences must be stored in our memories in the form of concepts and relations. Therefore, attending to the details of the sensory events is unnecessary, and we can economize the use of mental resources (Mach 1868, Mach 1910). The concept bears similarity with capturing dependencies and using minimum activity to represent them in the sensory system. It must be noted that the concepts of information and redundancy were not developed in the period, and hence the idea did not have suitable measures to quantify.

1.2.3 Kenneth Craik: Regarded as one of the first people to study cognitive sciences, Craik was a Scottish philosopher and physiologist. He introduced the concept of *mental models* in his book The Nature of Explanation (Craik 1943). Mental models are essentially the small-scale symbolic models of the environment the brain stores to predict sensory events. They are internal representations of the various associations that exist in our surroundings. Thus, in his work, Craik pointed to the importance of finding associations and the roles that may play in predicting or anticipating stimuli.

1.2.4 Egon Brunswik: Brunswik was primarily a psychologist and is known for his contributions to probabilistic functionalism. He pointed out that the environment in which an individual grows is as crucial as the individual and should be given equal attention in studies. He realized that the environment is uncertain and probabilistic, and the individual needs to learn and utilize this uncertainty. He studied the characteristics of images and found that portions of an image that belong to an object have different characteristics from randomly selected regions in the image (Brunswik and Kamiya 1953). From this finding, he suggested that if two portions of the image

have similar local characteristics, they likely belong to the same object and should be grouped. It is important to note that local feature detectors in the V1 area of the visual cortex are in close agreement with Brunswick's ideas.

1.2.5 Edward C. Tolman: Tolman was an American psychologist and founded a psychology branch known as purposive behaviorism. Tolman is known for his studies on rats in mazes, in which he wanted to demonstrate the abilities of rats to learn facts about their surroundings and use them in varying situations. Tolman introduced a concept very similar to the mental maps introduced by Craik and called it *cognitive maps*. Like mental maps, cognitive maps are also internal models of the environment where information about sensory events' relative locations is stored. It has a semantic network-like nature. Later discovered place cells in the hippocampus and the grid cells in the entorhinal cortex have been considered the neurological basis of such cognitive maps (O'Keefe and Dostrovsky 1971).

1.2.6 Fred Attneave: Attneave was among the first to bring the concepts from information theory to psychology to quantify information processing in sensory transduction. Information in an event reflected its uncertainty, and redundancy meant a lack of new information. Thus, if an event was predictable, its uncertainty was low, making its information content lower and increasing its redundancy. Attneave pointed out that there is a lot of redundancy in natural images because large portions of them can be predicted (Attneave 1954). However, he argued that the edges in these images are events of high unexpectedness. Therefore, they contain most of the information about the image, and an image can be represented more economically based on the edges. The idea of economic representation of images based on boundaries is essentially redundancy reduction, which both Barlow and Attneave advocated in their theories.

1.2.7 J. J. Gibson: Gibson was a prominent American psychologist known for his contributions to visual perception. He saw the senses as channels for the perception of the stimuli present in the surroundings and tried to find how one maintains a constant perception of the stimuli even when the inputs to the channels, i.e., the sensory inputs, change continuously. He proposed that specific properties of stimuli may remain invariant during the processing of continually changing inputs, and thus they comprised the information about the permanent environment (Marr 1982). The task of sensory processing was to detect these invariants and not to decode signals, or interpret messages, or process data for the fact (Gibson 1966, Gibson 1979). This idea was radically different from the previous notion of actively constructing stimuli representations by encoding the inputs.

1.2.8 David Marr: A British neuroscientist and physiologist, David Marr was one of the most influential figures in computational neuroscience. He studied several fields, including artificial intelligence, psychology, neurophysiology, and developed computational models of visual processing. He developed theories to explain the organization and workings of the cerebellum (Marr 1969), the neocortex (Marr 1970), and the hippocampus (Marr 1971). A significant contribution to the field of object recognition was his work with Nishihara (Marr and Nishihara 1978). In this work, he proposed three criteria, namely accessibility, uniqueness, and stability, and sensitivity, to judge the usefulness of a feature set for object recognition. Accessibility signified that the features should be computable from the sensory input. The uniqueness of features was required to make the representation of the objects distinct. The feature set's stability and sensitivity were indicated by its ability to reflect the similarity between two similar objects. The feature set was expected to be competent in expressing the subtle differences between objects too. An important

point to note here is that Marr approached the problem of recognition with the idea that certain features are identified because they help represent an object.

1.3. Introduction to information theory

The advent of information theory marks a pivotal point in the development of an understanding of sensory processing. Barlow and Attneave were among the first to realize that sensory processing is essentially a way of relaying information along the sensory pathways. Hence, information theory concepts should be readily applicable to it (Attneave 1954, Barlow 1961). Before discussing their arguments about sensory processing, and the utilization of information theory in those arguments, it is essential first to understand some basic concepts of information theory.

In any form of communication, a message is conveyed between two points that are separated in space and time. For example, consider a book. The combinations of letters or symbols in the form of words and sentences printed in the book comprise a message that the book's writer wishes to communicate. This message is transmitted and then received by a reader at a different point in space and time. In terms of communication theory, the book is a *"channel"* that communicates information. A telegraph machine, a telephone, and the worldwide web are all different channels used to convey different messages. However, all channels of communication are not as reliable as others. Several factors tend to corrupt the message. In the example of the book, wear and tear, fading of printing, printing mistakes, etc., are sources of message corruption. Corruptions can make the messages very difficult to interpret, or in worse cases, may convey an entirely different message. Therefore, it becomes imperative to avoid these corruptions. In the words of Shannon (Shannon 1948) *"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."* In his seminal work (Shannon 1948),

he approached this problem by attempting to design a channel that minimizes the chances of error in any message relayed through it. The critical point here is that the intent was to develop a system that reduces errors in all messages of a particular form and not just a single message. In terms of the book, the attempt was to design a book that is resistant to the various forms of corruption that may occur while printing a particular language like English. It did not matter whether the printed text was a fictional story or a scientific finding; all messages printed in English were supposed to have a minimal number of errors.

1.3.1. Entropy and Information

Shannon realized that when a message gets corrupted, then there is some inherent quantity in it that gets reduced. The reduction of this quantity increases the unpredictability of the message (Shannon 1948). For example, an inkblot on the letter *"e"* in the word *"The"* leads to an ambiguity where a three-letter word starting with *"Th"* may be interpreted as *"The"* as well as *"Thy"*. This ambiguity reduces certainty about the word and increases its unpredictability. Thus, essentially Shannon realized that such reduction in certainty could be quantified. This realization originated from the fact that uncertainty is caused by the existence of more than one acceptable message. Continuing with the previous example, if *"The"* was the only three-letter word in the English language that started with *"Th"*, then an inkblot on *"e"* would not have induced any ambiguity. The uncertainly arises because two different words are acceptable as they both comprise of three letters and start with *"Th"*. Therefore, any quantification of uncertainty of a message must be based on how likely or probable that message is. In other words, uncertainly in a message must be a function of its probability. It was suggested that the most natural choice of such function is a logarithmic one (Hartley 1928), and hence, the specific formulation of uncertainty was chosen to be

$$-\log p$$

where p is the probability of occurrence of the message. This quantity was called the entropy of the message. As with single messages, an expected value of entropy of an ensemble of messages can be defined as

$$H = -\sum_{i=1}^{N} p_i \log p_i$$

where p_i is the probability of occurrence of i^{th} message and, N is the total number of messages. Interestingly, as uncertainty can arise not only among messages but among any events that occur around us, one can define entropy for any event based on its probability of occurrence.

The receiver of the message offers another perspective of communication. All messages are equally likely before any message is received, and it is uncertain which particular message will be received. However, as signals are received, this uncertainty is reduced, and we say that the receiver has gained some information. Thus, a reduction in uncertainty of the messages is the information gained. As uncertainty is quantified as entropy (Shannon 1948, MacKay 2003), one can say that a change in entropy corresponds to a gain in information. Hence, they are the measures of the same quantity. In the example of the book, consider a reader who is about to read the book. Before he has opened the book and read any pages, he can expect any possible combinations of words and sentences to be present in the book, but as he reads the book, this uncertainty vanishes, and he gains the information that is present in the book

1.3.2. The communication channel and its capacity

A typical communication requires more than just a channel. It requires what is known as a *communication system* (Shannon 1948). A basic communication system consists of the following components

1. Source
2. Channel
3. Receiver

A *source* is a component that produces the message. In the case of the book, it is the writer. He thinks about the combinations of words and sentences to be written and thus composes the message. A *channel* is any system through which the message is relayed. The book is a channel. Finally, the *receiver* is the component that receives the message, i.e., the reader. Shannon realized that any communication channel could convey only a limited amount of information in a given period. This limitation arises due to the physical properties of the channel. For the book, the number of pages, the quality of paper, the quality of the printing ink, and the bindings are a few factors determining how much information it can carry for how long, hence determining the *"capacity"* of the book.

The implication of having a *"capacity"* is that one cannot expect to transmit information more than the channel's capacity. For example, one cannot convey a message of 250 pages in a 100-page book, and if one wishes to do that then, one must change the form of the message to the one suitable for the channel, i.e., the book.

1.3.3. Source coding

It is important to realize that the same information can be communicated in multiple ways. Consider a situation where one needs to know the position of a car parked in a parking garage with 64 parking spots and write the strategy of finding the vehicle in a book. A possible approach is to check every spot one-by-one and note the car's presence or absence in each spot. In this strategy, a maximum of 64 statements will be written in the book. Another possible method is to check half the parking lot sites and write in the book if the car is present in any of the positions. For example, if the vehicle is in any of the parking spaces in the first half of the parking lot, then the message written in the book should be *"car is in*

1st half". The other half of the parking lot can then be ignored, and the two halves of the 1st half should be checked. If the car is in the 2nd half now, then the subsequent message written in the book can be *"car is in 2nd half of the 1st half"*, and the process should be repeated till each half consists of single parking positions. Due to this form of recording spots, six statements need to be written to find the car's exact position. As we can see, the same information about the position of the car can be conveyed in either 64 statements using the first strategy or in 6 statements using the second one. The process of translating a piece of information into a specific format of statements is known as *source coding*. Any information can be translated into multiple formats, implying that it can be conveyed in numerous ways, each of which may highlight a different aspect of the information. In the above example, while expressing the car's position in terms of 64 statements, the car's position relative to the 1st parking spot is communicated. In contrast, the second strategy's six statements reveal the car's position relative to the set of 1st half parking spots.

Formally, information is quantified in bits where a bit of information can be thought of as a simple statement that answers an equiprobable yes-no question. Shannon showed that the number of such equiprobable yes-no questions that need to be answered to communicate the information content of an ensemble of messages equals the entropy of the ensemble. This is called the source coding theorem (Shannon 1948, MacKay 2003, Cover and Thomas 2006). In simpler terms, the source coding theorem establishes the minimum number of simple statements necessary to convey any information. Consider the above example of communicating the parking spot of a car; there are 64 acceptable messages of the form *"The car is in spot x"*, each of which is equally likely. Thus, the entropy of the ensemble of messages is log64, i.e., 6 bits. Therefore, the source coding theorem determines that at least six simple statements (like those recorded in the 2nd strategy) are needed to covey the position information. Using less than six statements will incur an information loss, and the accurate position cannot be identified. In communicating the information over a channel,

only those channels that have a capacity greater than or equal to the entropy of the message ensemble can be utilized for reliable communications. For instance, to transmit information about the car's position, the channel, i.e., the book, should have the capacity to contain at least six statements. If the book has any less capacity, then the accurate position cannot be noted. One may wish to record the position using 64 statements; in that case, a book of appropriate capacity needs to be chosen.

A common situation is when the entropy of the message ensemble is less than the capacity of the channel. The excess capacity over the entropy is known as *redundancy*. Specifically, it corresponds to statements being communicated with no new information. Suppose the book's capacity described previously is ten statements, and one chooses to describe the position of the car using the 2nd strategy, which requires only six statements. This leaves space for four statements in the book. One can either leave that space blank, conveying no information, or fill them with statements about positions where there is no car. In either case, no new information will be communicated using those four statements; therefore, those statements will be redundant. These concepts of capacity and redundancy are heavily utilized in formulating the theories about sensory processing and will be described in that context in the next section of this chapter.

1.4. The Efficient Coding Hypothesis

The idea of source coding is very appealing in the context of sensory processing. Sensory processing is essentially a transformation of neuronal activity patterns along the sensory pathway. This process can be viewed as a translation of the information about the environment into the language of neuronal firings. The firing patterns correspond to the statements conveying the information, and different transformations of these patterns are equivalent to different formats of statements. Furthermore, this source coding is utilized in a communication system where various objects in the environment serve as the source; the

sensory neurons that interact with these objects comprise the channel, and the subsequent neurons along the sensory pathway which receive input from these sensory neurons act as receivers.

Barlow and others identified these parallels between a communication system and sensory processing (Attneave 1954, Barlow 1961), and based on Shannon's source coding theorem (Shannon 1948), they proposed a set of theories popularly known as *the Efficient Coding Hypothesis* (Barlow 1961). These theories aimed to explain the format of information transmission in the sensory system. They emphasized how a specific format is advantageous for the system in gathering knowledge about the environment's organization. Such knowledge is assumed to be manifested as some form of *"regularity"* in the environment. Consequently, these theories proposed that sensory processing aims to enable the system to identify these *"regularities"* so that it can recognize environmental structure and rules.

The concept of *"regularity"* is analogous to the idea of *predictability*. A geometric shape is *"regular"* if all its edges are equal, i.e., if the length of all its sides can be *predicted* by knowing the length of just one edge. Similarly, a pattern is called *"regular"* if its constituent motifs are repeated *predictably*. Thus, the notions of *regularity* and *predictability* are related in the sense that any form of *regularity* in events allows their *prediction*. Conversely, the *predictability* of events is indicative of their *regularity*.

Examining our surroundings, we find that the natural environment is filled with predictable components. We can predict the shapes of objects, the occurrence of events, changes in conditions, and so on. Such predictability implies that these components are regular. For example, the outlines of shapes are smooth. They are not jagged or randomly broken. If one knows a particular portion of the outline, they can predict the next piece based on the known portion. In similar ways, events like sunrise and the chirping of flocks of birds are also regular. The chirping of birds is often heard in the mornings, and if one knows the

time of sunrise, one can predict the timings of chirping. Such regularities comprise the knowledge about the environment that an organism needs to identify to ensure its survival. For example, if an animal feeds on birds, it must recognize that birds reveal their location around sunrise, and therefore, that is the best time to feed.

However, identifying these regularities is not straightforward. There are no set rules that qualify a regularity, and one has to guess based on the predictability of events. A simple way to predict events can be based on their co-occurrence. If two events occur together, then one can be predicted based on the other. Yet, this approach does not take into account the co-occurrences that arise just by chance. Events that are more likely to happen in our surroundings are more likely to occur together, and therefore, cannot be good predictors of each other.

Another way to predict events can be based on *dependence* among them. In probability theory, the *dependence* between two events is identified when the probability of one event changes with the occurrence of the other. Simply put, two events are *dependent* if the occurrence of one influences the occurrence of the other. The formal definition of dependence is based on two probability measures, namely the *marginal probability* and the *conditional probability*. The *marginal probability* of an event is the quantification of the chance with which it happens. The *conditional probabilit*y is the measure of chance with which it happens when another event has already happened. When the marginal probability of one event differs from its conditional probability, calculated with respect to the other event's occurrence, then the two events are said to be dependent. In terms of notations, if we denote two events as X and Y, then their marginal probabilities can be denoted as $\mathbb{P}(X)$ and $\mathbb{P}(Y)$, respectively. The conditional probability of X with respect to the occurrence of Y is denoted as $\mathbb{P}(X|Y)$. Similarly, the conditional probability of Y with respect to the occurrence of X is denoted as $\mathbb{P}(Y|X)$. Expressed in terms of these notations, X and Y are said to be dependent if

$$\mathbb{P}(X|Y) \neq \mathbb{P}(X)$$

or equivalently,

$$\mathbb{P}(Y|X) \neq \mathbb{P}(Y)$$

An essential aspect of the notion of dependence is that it is not affected by the actual probabilities of events. In contrast, an event depends on the other only when the chances of it happening change with the other's occurrence. Thus, even if the event's occurrence probability is large, it is the change in this large probability that determines dependence. The extent of the probability change is an indicator of the influence that one event has on the other. Therefore, one can predict any event's occurrence by evaluating the probability with which any dependent event occurs.

With such a role of dependence in predicting events and identifying regularities, the goal of sensory processing, as assumed under the Efficient Coding Hypothesis, can be reiterated to be identifying the dependence between different components of the environment. The task can be accomplished by comparing their conditional probabilities with marginal probabilities. However, to make such a comparison, the two probabilities should be made available to the system. As various events elicit responses in the sensory neurons and are represented in the sensory system in the form of distributed activity patterns, the system can use these representations to obtain the probabilities. In this regard, representation of information about the event in an activity pattern, or the coding process, becomes an important aspect of sensory processing. This is because different ways of representing information highlight various aspects of information (Marr 1982) and selectively ease certain operations. For example, representing numbers in decimal form makes arithmetic operations easier, whereas representing numbers in binary form does not. Barlow suggested that a suitable way of representing information about sensory events is to allow individual neurons to be as independent as possible (Barlow 1987, Barlow 1989, Barlow et al. 1989). It makes calculations of probabilities required for identifying the dependencies more manageable.

To understand his rationale behind this suggestion, let us consider the concept of independence. Just like dependence, in probability theory, independence is also defined based on conditional and marginal probabilities. The underlying idea is that if two events are independent, then one should not influence the other's occurrence. Therefore, the probability of one event conditioned on the other should be the same as its marginal probability i.e.

$$\mathbb{P}(X|Y) = \mathbb{P}(X)$$

The mathematical relation can also be expressed in terms of the *joint probability* of events, denoted as $\mathbb{P}(X, Y)$, where *joint probability* is the quantification of chance that the two events occur together. As

$$\mathbb{P}(X, Y) = \mathbb{P}(X|Y)\mathbb{P}(Y)$$

$$\mathbb{P}(X|Y) = \mathbb{P}(X) \Rightarrow \mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y)$$

Following the same line of argument for neurons, two neurons a_1 and a_2 can be said independent if the state of one neuron does not affect the states of the other. In terms of probability, the probability of a_1 being active or inactive does not change depending on the state of a_2. Denoting active state as 1 and inactive state as 0, we can write

$$\mathbb{P}(a_1 = 1|\, a_2 = 0) = \mathbb{P}(a_1 = 1)$$

$$or, \mathbb{P}(a_1 = 1|\, a_2 = 1) = \mathbb{P}(a_1 = 1)$$

$$or, \mathbb{P}(a_1 = 0|\, a_2 = 0) = \mathbb{P}(a_1 = 0)$$

$$or, \mathbb{P}(a_1 = 0|\, a_2 = 1) = \mathbb{P}(a_1 = 0)$$

Similar relationships can be written when a_2 is conditioned on a_1. Summarizing all possible states of neurons in variables a_1 and a_2, we can note the following for independent neurons

$$\mathbb{P}(a_1,\, a_2) = \mathbb{P}(a_1)\mathbb{P}(a_2)$$

An interesting property of independent neurons becomes evident if we consider a third neuron a_3. If the third neuron is independent of the first two, then following the same logic as before, we can write

$$\mathbb{P}(a_1, a_2, a_3) = \mathbb{P}(a_1, a_2)\mathbb{P}(a_3) = \mathbb{P}(a_1)\mathbb{P}(a_2)\mathbb{P}(a_3)$$

In fact, for N independent neurons, one can write

$$\mathbb{P}(a_1, a_2, \ldots a_N) = \prod_{i=1}^{N} \mathbb{P}(a_i)$$

Note that, in the above equation $a_1, a_2, \ldots a_N$ represents any possible combination of states of N neurons, and $\mathbb{P}(a_i)$ represents the probability of the particular state of i^{th} neuron that is considered in $a_1, a_2, \ldots a_N$. Thus, for independent neurons, the probability of any combination of states of a set of neurons factors into probabilities of particular states of individual neurons.

In sensory processing, any combination of states of neurons supposedly represents a sensory event. With independent neurons, the probability of the event can be calculated by multiplying the probabilities of states of individual neurons. Moreover, as a joint event is also an event distinct from its constituents, independence among neurons allows the same ease in calculating their probabilities. In this way, requiring neurons to be independent of one another eases the calculation of probabilities necessary for identifying dependencies. This ease of analysis is the prime motivation behind using such neurons for representations. Note that the representation scheme where the probability of event factors into probabilities of states of neurons is often referred to as *factorial coding* (Barlow 1987, Barlow 1989, Barlow et al. 1989).

Though representing events through independent neurons constitutes an attractive scenario for identifying dependencies, one still has to invent a representation strategy that renders individual neurons independent. In other words, the system has to select which component from its environment it should represent so that individual neurons fire independently. As probabilities of states of individual neurons are determined from occurrence probabilities of events they represent, individual neurons cannot always be completely independent. Depending on the occurrence of events and the representation

scheme, individual neurons may or may not be independent. For example, consider a situation where three events, namely, s_1, s_2 and s_3 are represented by a set of 2 binary neurons (a_1 and a_2) using a representation scheme presented in the figure (**Figure 1.1**).

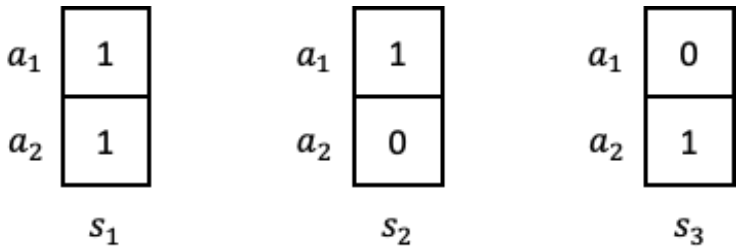


Figure 1. 1: Representations of three events s_1, s_2 and s_3 using two binary neurons a_1 and a_2

Suppose we consider the occurrence probability of the event s_1 to be p_1, s_2 to be p_2, and s_3 to be p_3, then one can calculate probabilities of states of neurons as

$$\mathbb{P}(a_1 = 1) = p_1 + p_2, \text{ and } \mathbb{P}(a_1 = 0) = 1 - p_1 - p_2$$

$$\mathbb{P}(a_2 = 1) = p_1 + p_3, \text{ and } \mathbb{P}(a_2 = 0) = 1 - p_1 - p_3$$

Similarly, joint probabilities of the states of neurons can be obtained. We can check the independence of neurons a_1 and a_2 in this particular representation scheme by comparing the joint and marginal probabilities of the states of the neurons. Considering the particular state $(a_1 = 1, a_2 = 1)$, we find that

$$\mathbb{P}(a_1 = 1, a_2 = 1) = p_1, \mathbb{P}(a_1 = 1) = p_1 + p_2 \text{ and } \mathbb{P}(a_2 = 1) = p_1 + p_3$$

The product of the latter two equals the first in any situation where either p_2 or p_3 is 0, like when $p_1 = 0.7$, $p_2 = 0.3$ and $p_3 = 0$, or when $p_1 = 0.9$, $p_2 = 0$, and $p_3 = 0.1$. Indeed, in these situations, neurons a_1 and a_2 are independent because in these situations, either state of a_1 is always 1 or state of a_2 is always 1, and the probability for attaining any

state by the remaining neurons is not altered. In any other situation, like where $p_1, p_2, p_3 \neq 0$, this representation strategy will not produce independent neurons.

To find a strategy that maximizes neurons' independence, one has to minimize the difference between the probability of a combination of states of neurons, i.e., $\mathbb{P}(a_1, a_2, \dots a_N)$ and the product of probabilities of particular states of neurons. This difference is measured in terms of *KL divergence* (Kullback and Leibler 1951) between probability distributions. Therefore, in information-theoretic terms, the strategy to maximize the independence of neurons can be formulated as a minimization problem where the function to minimize is

$$D_{KL}\left(\mathbb{P}(a_1, a_2, \dots a_N) \middle\| \prod_{i=1}^{N} \mathbb{P}(a_i)\right) \text{ where } D_{KL}(\mathbb{P}\|\mathbb{Q}) = \sum_x \mathbb{P}(x) \log\left(\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}\right)$$

Interestingly, the above function can be decomposed into three terms (J.-F. Cardoso, 2003) as under

$$D_{KL}\left(\mathbb{P}(a_1, a_2, \dots a_N) \middle\| \prod_{i=1}^{N} \mathbb{P}(a_i)\right) = C - \sum_{i=1}^{N} G(a_i) + K$$

Here, the first term is a function of the correlation between the neurons. Formally, correlation is any statistical relationship between two variable quantities. For example, the price of a car is correlated to the miles it has been driven because the higher is the miles on the vehicle, the lower it costs. In the case of neurons, correlation arises when the activity patterns of two neurons are related in any way. They might be firing together, or the peak of a neuron's firing rate may be proportional to the peak of others' firing rate, or any other observed relationship might exist. When neurons do not display any such relation, they can be called uncorrelated. The first term in the above equation vanishes for uncorrelated neurons.

The second term is a measure of non-Gaussianity of the neuronal response profiles. Suppose one knows all the states of a neuron and records each state's probabilities in the form of a histogram. In that case, the similarity of that histogram to a normal bell curve is

quantified as the Gaussianity of the neuron's response profile. Non-Gaussianity, therefore, is the measure of the deviation of the probability distribution histogram from the normal bell curve distribution.

As the third term in the equation is a constant, under this formulation, maximizing the independence of neurons translates into minimizing correlations and maximizing non-Gaussianity of the neuronal response profiles. For neurons with only two states, this increases with the difference between probabilities of inactive and active states; the more is the probability of inactive state, the more is the non-Gaussianity. Therefore, to increase the neurons' independence, the representation strategy should minimize, on average, the activation probability of any neuron. As an illustration of the process, consider the previous example of three events being represented by two neurons. Assuming $p_1 = 0.6$, $p_2 = 0.3$, and $p_3 = 0.1$, and following the strategy suggested in the Efficient Coding Hypothesis; we should represent the events as suggested in the figure (**Figure 1.2**)

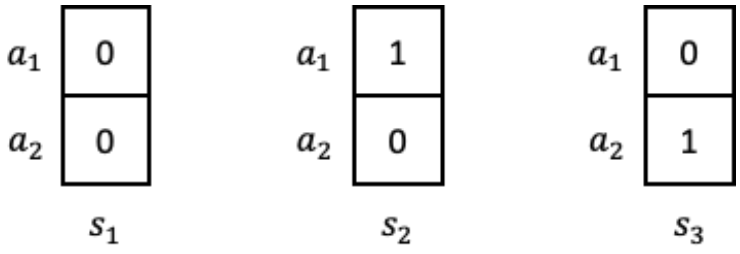


Figure 1. 2: Representations of three events s_1, s_2 and s_3 using two binary neurons a_1 and a_2 following the Efficient Coding Hypothesis

We can calculate the marginal distributions of neurons a_1 and a_2 as

$$\mathbb{P}(a_1 = 1) = 0.3, \quad \mathbb{P}(a_1 = 0) = 0.6$$

$$\mathbb{P}(a_2 = 1) = 0.1, \quad \mathbb{P}(a_2 = 0) = 0.9$$

Clearly, for both the neurons, the probability of inactive state (state 0) is way larger than the probability of active state (state 1); therefore, the distributions are highly non-Gaussian, and hence the encoders are maximally independent. To verify the extent of independence, we can compare the joint probability and the products of marginal probabilities as under

$$\mathbb{P}(a_1 = 0, a_2 = 0) = \mathbb{P}(s_1) = 0.6 \quad \text{and} \quad \mathbb{P}(a_1 = 0)\mathbb{P}(a_1 = 0) = 0.6 \times 0.9 = 0.54$$

$$\mathbb{P}(a_1 = 1, a_2 = 0) = v(s_2) = 0.3 \quad \text{and} \quad \mathbb{P}(a_1 = 1)\mathbb{P}(a_1 = 0) = 0.3 \times 0.9 = 0.27$$

$$\mathbb{P}(a_1 = 0, a_2 = 1) = \mathbb{P}(s_3) = 0.1 \quad \text{and} \quad \mathbb{P}(a_1 = 0)\mathbb{P}(a_1 = 1) = 0.6 \times 0.1 = 0.06$$

In all three cases, we find that the product of marginal probabilities can approximate the probabilities of events and hence are maximally independent as required.

The strategy of minimizing neurons' activity to achieve independence can be interpreted in another way using the information-theoretic arguments. In terms of communication theory, each neuron state can be regarded as a unique message to the sensory system. Therefore, an entropy term can be associated with each neuron, which corresponds to the entropy of the ensemble of messages it is communicating. Interestingly, the knowledge of the actual message from the environment or its occurrence frequency is not required for calculating the neuronal entropy. It can be computed based on the probabilities with which the neuron acquires its states. For example, for a binary neuron a_i, that can take only two possible states, this entropy will take the form

$$H(a_i) = -(\mathbb{P}(a_i = 1)\log\mathbb{P}(a_i = 1) + \mathbb{P}(a_i = 0)\log\mathbb{P}(a_i = 0))$$

A known property of the entropy function is that it takes the maximum value when computed for a uniform probability distribution. Any deviation of the distribution from uniformity decreases the value of entropy associated with it. Thus, a binary neuron achieves maximum entropy when the probability of it being active equals the probability of it being inactive. Consequently, biasing the probability distribution of its states away from this uniformity will lead to a reduction in its entropy. If we identify a collection of neurons as a

representation system and define this system's entropy H_{sys} as the sum of entropies of individual neurons comprising it i.e.

$$H_{sys} = \sum_i H(a_i)$$

then minimizing the average activity of neurons essentially corresponds to minimizing H_{sys}. For this reason, the strategy of efficient coding is also referred to as *minimum entropy coding* (Barlow 1961, Barlow et al. 1989). Note that one can also calculate the entropy H_{inp} associated with probabilities of occurrence of sensory inputs s_j as

$$H_{inp} = -\sum_j \mathbb{P}(s_j) \log \mathbb{P}(s_j)$$

and can compare it with the entropy of the representation system. The comparison term, known as *representation redundancy,* or simply, *redundancy* (MacKay 2003), measures the fractional difference between the entropy of events and the entropy of the system and is defined as

$$R = 1 - \frac{H_{inp}}{H_{sys}}$$

This concept of representation redundancy is the same as the concept of redundancy introduced in the theory of communication. Recall that redundancy in a communication system was defined as the excess capacity of a channel over the amount of information being communicated through it. It corresponds to the portion of the channel that is not conveying any new information. We have also noted that parallels can be drawn between a communication channel and a collection of neurons that relay information about the environment by representing events and objects. In this regard, the entropy of a neuronal ensemble can be regarded as the total amount of information it can represent, given the current distribution of states of its constituting neurons. In other words, if we consider each state of a neuron to be conveying a unique message, then the entropy of the neuronal ensemble H_{sys} equals its capacity to represent information. Furthermore, the entropy of the

sensory inputs H_{inp} is the actual amount of information being relayed through this collection of neurons; therefore, the redundancy in this communication system, i.e., the excess capacity over the amount of information being communicated, can be expressed as

$$redundancy = H_{sys} - H_{inp}$$

If this redundancy is normalized to the capacity of the neuronal collection, we get the expression for representation redundancy i.e.

$$R = \frac{redundancy}{H_{sys}} = \frac{H_{sys} - H_{inp}}{H_{sys}} = 1 - \frac{H_{inp}}{H_{sys}}$$

Thus, representation redundancy is nothing but the redundancy observed while representing information about the environment through the sensory relays and the *minimum entropy coding* that aims to minimize H_{sys}, is an attempt to minimize this redundancy. Due to its nature of reducing redundancy, *minimum entropy coding* is also referred to as the *redundancy reduction* approach towards representing information.

In summary, the Efficient Coding Hypothesis tries to capture different forms of regularities from the surroundings. It suggests that sensory events' representation needs to be factorial in nature and should be based on independent neurons to identify these regularities. The system can obtain such representations by minimizing the average activation probability of neurons or its overall entropy.

1.5. Sensory system studies supporting the efficient coding hypothesis

After Barlow and others proposed the efficient coding hypothesis, several studies found the relevance of the theory in processing sensory information across different modalities. Essentially, there were two types of studies – first, experimental studies that measured the response properties of neurons in the sensory systems and showed that the neurons were representing the information about the surrounding efficiently. Second, models

based on the efficient coding hypothesis were developed that analytically predicted the response properties of neurons. Both these types of studies provided evidence supporting the Efficient Coding Hypothesis and redundancy reduction approach. Here, I will discuss some of those studies.

Among the first studies that showed the Efficient Coding Hypothesis's applicability was Laughlin's study of blowflies. In his study, Laughlin argued that coding efficiency arises when a neuron equally utilizes all its response states in encoding the corresponding stimuli (Laughlin 1981). In a simple case of a neuron responding to a single input parameter, the range of input parameters will be way larger than the neuron's noticeably different response states. In this situation, the optimum will be attained when different response states of the neuron correspond to ranges of the input parameters that occur with the same cumulative frequency so that each response state is equally utilized. In terms of probability distributions, each response level of the neuron encompasses an equal area under the parameter distribution curve. This idea was inspired by a digital image processing technique called *Histogram equalization* (Gonzalez and Wintz 1977). Laughlin, using natural images from scenes such as dry sclerophyll woodland and lakeside vegetation, showed that the large monopolar cells (LMC) in blowfly's compound eye has a response function that matches the cumulative distribution of the contrast levels measured from the images. The idea of equal utilization of response level reflects Barlow's minimal redundancy idea as no response level is utilized for encoding the same stimulus parameter.

The study described above presents a direct way of implementing efficient coding in the sensory systems; however, it does not consider the presence of noise in the system. In another study (Srinivasan et al. 1982), Laughlin and colleagues introduced the concept of *predictive coding* to describe the receptive field properties of the retinal ganglion cells and bipolar cells (Barlow et al. 1957, Hartline and Ratliff 1972), and interneurons in insect compound eye (Laughlin 1981). Under this concept, it was proposed that in the retina, neurons from the surrounding area predict the value of the input at the center and subtract

that value from their current input value so that their dynamic range can be more efficiently utilized. With such subtraction, predictive coding is essentially removing 2nd order correlations, i.e., redundancy between pairs of points observed in the autocorrelation function, and hence, is a model for redundancy reduction. In this study, Laughlin further showed that the nature of the surround in center-surround receptive fields depends on the signal-to-noise ratio. For a low signal-to-noise ratio, a larger surround is necessary to predict the value at the center accurately. In contrast, for a high signal-to-noise ratio, even a confined surround is sufficient. He then demonstrated that the same model could be applied to remove the temporal correlations and could also explain the LMC function in the fly's compound eye

Through studies like Laughlin's, it was evident that efficient coding requires knowledge of the statistical structure of the stimuli. Field was among the first to figure out the statistical properties of natural images. Using various images of natural scenes, he showed that the power spectrum of the natural scenes falls off as $1/f^2$, and the amplitude spectrum falls of as $1/f$ where f is the spatial frequencies in the image (Field 1987). Field argued that such statistics were a natural consequence of the relative contrast energy being scale-invariant and could also be related to the fractal nature of the images' luminance profiles. The $1/f^2$ falloff gives a fractal dimension of 2.5 (Voss 1985). With such power spectrum and stationary statistics of the natural images, Field proposed that the best-suited code for encoding these images is the one where encoders have constant octave bandwidth and constant orientation. These codes allowed the information about stimuli to be evenly distributed across the encoders and presented a way to convert high order redundancy to first-order redundancy.

Later, Ruderman and Bialek also characterized statistical properties of natural scene images using wood images (Ruderman and Bialek 1994). They measured the normalized average contrast in varying sizes of image patches and show that the contrast histograms

overlapped for all sizes of image patches, demonstrating that the contrast distribution is invariant to the angular scales. The distribution, however, was very far from Gaussian. This departure was shown in the deviations of contrast gradients' distributions from the Rayleigh distribution. Typically, following the central limit theorem, one would expect that the distributions will be more Gaussian, but this breakdown of the central limit theorem showed that the pixels were correlated over long distances in the images. They also found that consistent with Field (Field 1987), the power spectrum follows the form

$$S(f) = \frac{A}{f^{2-\eta}}$$

where $\eta = 0.19 \pm 0.01$ and $A = (6.47 \pm 0.13) \times 10^{-3}$deg.

Atick and colleagues were among the first to explain the receptive field properties of retinal ganglion cells using the redundancy reduction principle (Atick and Redlich 1990). Assuming that the input's spatial correlation is known, and the transformation of output from the input is a linear one with noise, they tried to calculate information-theoretical quantities like mutual information between input and output, channel capacity, and redundancy. The probability distributions of both the input and the transformed input with noise were considered to be the ones with maximum entropy displaying assumed correlations. Reducing redundancy under these conditions by reducing the channel capacity resulted in the center-surround type receptive field properties of the encoders that were very similar to the kernels of retinal ganglions measured in experiments on cats and monkeys (Enroth-Cugell and Robson 1966). They also analyzed the effect of noise on properties of the linear transformation from inputs to outputs and the corresponding changes in the receptive fields. It was found that when the signal-to-noise ratio was high, the transformation was decorrelating, as predicted by Barlow's redundancy reduction hypothesis. The receptive fields in such conditions had relatively narrower surround. When the signal-to-noise ratio was low, the transformation approximated a smoothing function, which increased the correlations among the encoders. The receptive field had a larger surround region in these

conditions. These results are very similar to those proposed by Srinivasan and Laughlin (Srinivasan et al. 1982) using predictive coding concepts. Atick and Redlich also noted that these techniques used to derive the optimal coding conditions for the visual signals' spatial properties could be directly applied to their temporal properties. Similar to spatial properties, it would suggest reducing temporal correlations in high signal-to-noise regimes and signal integration when the signal-to-noise ratio was low.

In another attempt to describe the retinal filters noted in experimental studies (Kelly 1972, De Valois et al. 1974), Atick and Redligh utilized the knowledge about the $1/f^2$ powers spectrum of natural scene images (Atick and Redlich 1992). Simple filtering of the natural scenes' amplitude spectrum depicted that the filters are designed to decorate the output at lower frequencies. Deviating from their previous approach of reducing the redundancy and suppressing the noise simultaneously, in this study, they considered the problem in two separate stages – first, they solved the redundancy reduction problem without considering any noise. Then the noise was added, and the obtained solution was modified accordingly. Specifically, they tried to analytically find a retinal filter function that maximally decorrelated the output in a noiseless condition. A $1/f^2$ power spectrum of the input was assumed. The energy function that was minimized to obtain the filters could be interpreted as simultaneous minimization of the bit entropies of outputs and information loss in transforming inputs to outputs. The resulting filter function $K(f)$ was of the form $K(f) \sim k|f|$ which is a whitening filter for $1/f^2$ spectrum. However, when the noise was added, the nature of the filter changed. It maintained its whitening nature at lower frequencies, but the optimal retinal filter was more like a low-pass filter at higher frequencies. To explain the results, Atick and Redlich argued that at lower frequencies, following the $1/f^2$ spectrum, the input signal is larger than that of noise. Therefore, the filters tend to whiten the image, probably to reduce the redundancy as suggested by Barlow. However, at higher frequencies, the signal was lower or comparable to noise, so the filter

adapted to a low-pass filter, which tends to smoothening the noise and removing its effects. Thus, at higher frequencies where noise power does not decrease like the signal power, it was more important to remove the noise than to decorrelate the output because filtering the noise with a $K(f) \sim k|f|$ filter will significantly amplify the noise. Furthermore, different retinal filters for varying levels of mean luminosity were obtained. Under the assumption that the major noise source is quantum noise, mean luminosity is a direct indicator of noise independent of the frequency. The variation in luminosity introduced a transition in the nature of the filters from being bandpass at high luminosity values to a low pass at low luminosities. This transition was consistent with the human contrast sensitivity measurement studies by Van Ness and Bouman (Van Nes et al. 1967).

Following similar techniques, retinal filters using distributions of different colors (Atick 1992, Atick et al. 1992) and temporal correlation (Dong and Atick 1995) were obtained. It was assumed that most of the spatial decorrelation occurred in the retina. The LGN was supposed to be primarily handling temporal decorrelations. With proper rectifications at the retinal and the LGN layers, the study could explain the lagged vs. non-lagged cells in LGN.

J. H. van Hateren carried another set of similar studies to understand the early sensory processing primarily in the visual field. He considered sensory processing to be a combination of filtering and noise addition in the incoming signal. The goal of sensory processing was to maximize the amount of information being transmitted through such noisy channels by optimizing the use of channel capacity (Van Hateren 1992). The neural filters thus obtained were bandpass, i.e., they encoded only a specific range of frequencies. It was argued that the lower frequencies are discarded because they are very strongly present in the stimulus. With high power, they threatened to occupy too much of the channel's dynamic range. Though such removal of frequencies will lead to loss of information, van Hateran argued that it was better to have frequencies of moderate signal-to-noise ratio than to have a mixture of very low and very high ones. The rationale for removing higher frequencies was

that their signal-to-noise ratio was too low to carry any significant information. It was found that the obtained filters sharpened the response histograms of the neurons at a high signal-to-noise ratio, thus reducing the redundancy. At the low signal-to-noise level, the redundancy was increased.

Field realized that in addition to the scale invariance reflected in the $1/f^2$ power spectrum of natural images (Field 1987), a second form of invariance relates to the local structures of the images. A local structure like a line or an edge arises when the phases of the constituting waveforms are locally aligned. These regional structures are encountered across different scales and hence correspond to a type of redundancy between different scales. Field noticed that such redundancy is destroyed when phases are randomized. To have the phases aligned across different scales, the bandwidth of the phase structures, i.e., the frequency window over which the phases are aligned, should be proportional to the frequency (Field 1993). However, what bandwidths must be aligned could not be determined. It was further argued that the natural scenes did not consist of randomly positioned and randomly oriented phase structures. A regularity similar to the one observed in fractal images also exists in the phase structures of images. Field suggested that a wavelet code, with mechanisms to select the local phase structures' orientation and frequency, could be a way to encode the local structures. An advantage of such coding was that it allowed sparse representations of the image structures, which was in line with utilizing the surroundings' redundancy.

Following this line of thought, Field tried to identify the criterion that would allow sparse coding of the natural scenes. He found that a basis set that allows sparse representation of data could be characterized by the kurtosis of data distribution along the basis vector. A sparse representation means that the most likely value the data will take along any basis vector is zero. The probability of it taking any other value is relatively small; hence its distribution is peaked at zero and has elongated tails. Such distributions have high kurtosis,

which is indicative of the non-Gaussianity of a distribution. Field showed that the response profiles of filters resembling simple cell type receptive fields were highly kurtotic. Hence, the system was essentially trying to obtain a sparse code for natural scenes (Field 1994). In the filtering experiments, it was found that response profiles of different filters like the wavelet filters or difference of Gaussian filter had different kurtosis, with maximum kurtosis observed for the wavelet filters. The kurtosis of the response histograms for wavelet filters was then measured as a function of their spatial frequency bandwidth. The results indicated that the filters that produced the maximally kurtortic response profile among the wavelet filters had a bandwidth of 1.0 to 3.0 octaves. This range of bandwidths is most commonly observed in the mammalian visual system (Tolhurst and Thompson 1982). Thus, Field concluded that the system was designed to achieve sparse coding. It is important to note that the sparse coding can be seen as a way of attaining high kurtotic response profiles, which indicate independence among the encoders. The more kurtotic is the response of encoders, the less dependence will be among them. In this way, sparse coding is a way to achieve efficient coding, as suggested by Barlow.

Olshausen and Field further advanced the study of sparse coding in the visual system. They assumed that the transformation of an image to its representation in the sensory system is linear and that the encoders involved in representing the image are independent and sparsely active. The independence and sparsity assumption on encoders implied that the probability distribution defined over its state had high kurtosis. It was expected that a white, gaussian, additive image noise would be introduced during the transformation process; therefore, the likelihood of an image being generated from its coefficient had a Gaussian distribution. Olshausen and Field calculated the probability distributions of images that could be generated from their model using the likelihood of image and the prior distribution of representations. They tried to match this distribution to the probability distribution of images found in the natural scenes by reducing the two distributions' KL divergence. Reducing KL divergence resulted in maximizing the log-likelihood of the image under a

given transformation. It was shown that, when such maximum is found, the receptive field properties that arise for the encoders match the local, oriented, bandpass filter like receptive field properties observed in simple cells (**Figure 1.3**) (Olshausen and Field 1996, Olshausen and Field 1997). Thus, with this study, they demonstrated that the receptive fields of neurons produce sparse response histograms. Conversely, constraining the system to encode natural images sparsely results in generating the neurons' receptive field properties. An essential aspect of this study was the overcomplete nature of the transformation. It was considered that the number of cells involved in encoding the images was larger than the effective dimensions of the image. The rationale behind such consideration was that combined with sparsification, it leads to some deviation from the strictly linear input-output relationship.

In another study, Olshausen and Lewicki demonstrated that using a similar probabilistic inference framework also generates consistent representations of noisy or incomplete images (Lewicki and Olshausen 1999), thus effectively denoising the images or filling the missing parts of the images.

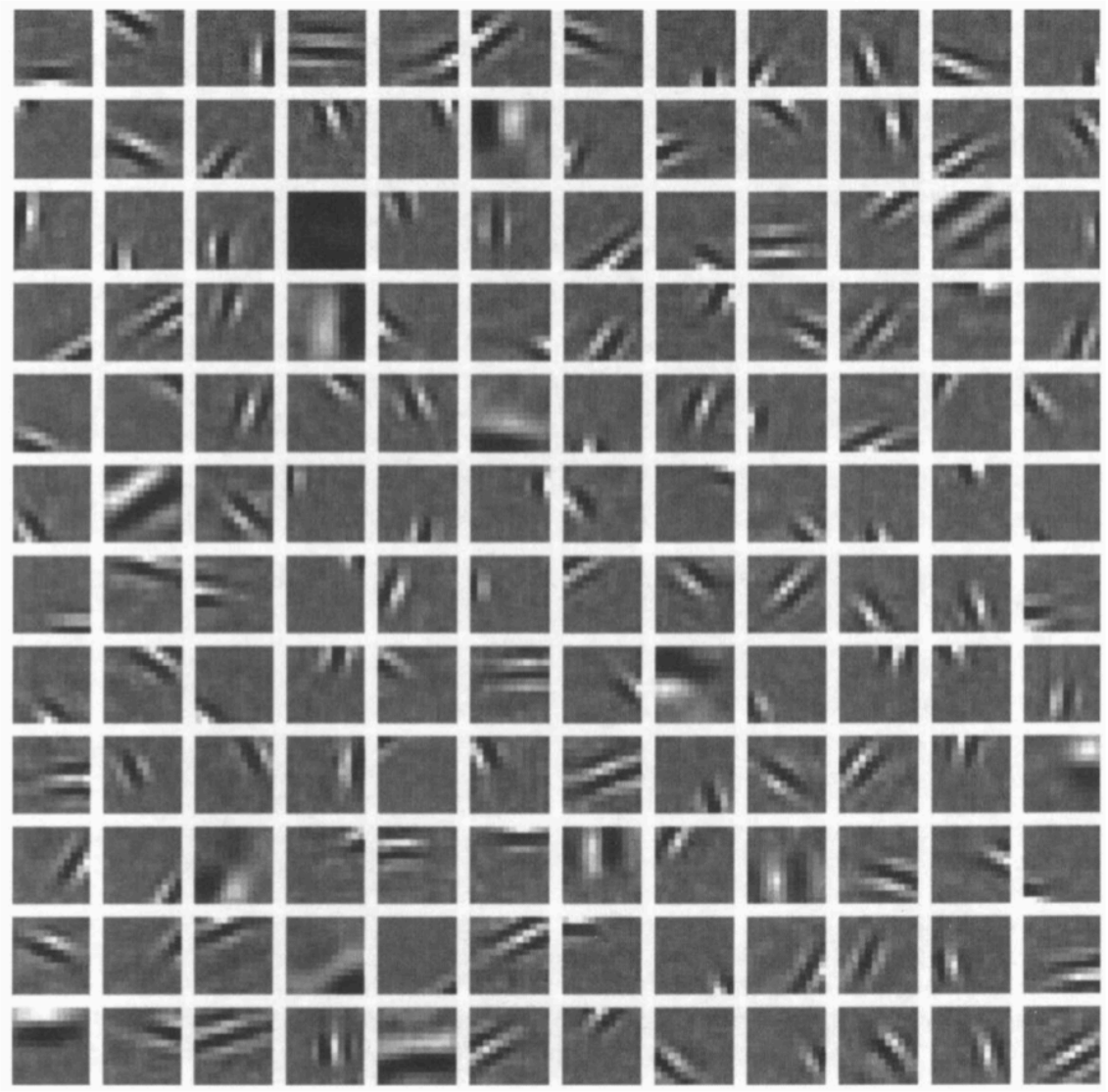


Figure 1. 3: A set of 144 basis functions learned by the sparse coding algorithm (from Olshausen and Field, 1997)

The framework of Olshausen and Field was finding the set of basis images that were independent of each other and could sparsely describe the natural images, thus achieving efficient coding. Several similar algorithms were introduced around the same period to find the independent components of any data. Bell and Sejnowski's approach, also known as ICA or the *infomax approach*, was to maximize the mutual information between the input image and its corresponding representations (Bell and Sejnowski 1995). It could be shown that maximizing such quantity also leads to minimizing the mutual information between

image encoders (Nadal and Parga 1994) thus leading to a factorial code as proposed in the Efficient Coding Hypothesis. A significant difference between Olshausen and Field's approach and Bell and Sejnowski's approach is that the former allows an overcomplete representation of images while the latter does not. Although an overcomplete representation does seem more biologically realistic because of the expansion in the number of cells from the retina to the visual cortex, it may cause the basis images to be linearly dependent. Also, unlike Bell and Sejnowski's algorithm, Olshausen's algorithm does not force the encoders to have low mutual information, which prevents the model from achieving a factorial code.

Hyvarinen and Hoyer further extended the application of ICA in finding features that resembled the receptive field properties of the complex cells. They modeled a complex cell's response as the sum of the simple cells' squared responses and then maximized the independence of complex cells (Hyvärinen and Hoyer 2000). Hyvarinen and Hoyer reproduced phase and translation invariance properties of the complex cells. Phase invariance meant that the response of the cell did not depend on the phase of the stimuli. Translation invariance or limited shift-invariance meant that identical stimuli could elicit the near-maximum response at slightly different locations. These properties could not be found in the previous simple cell models, and this study reported their emergence for the first time. The fact that further dependence could be detected among receptive fields derived from ICA-like algorithms showed that the ICA did not detect completely independent components.

In a further study (Hoyer and Hyvärinen 2002), another layer of cells was added to the existing complex cell model to capture high-order sensory processing. The underlying hypothesis was that the cortex was not just involved in efficiently representing the incoming information but was trying to build a probabilistic model of the surroundings from the information. To capture such processing, they put combined constraints of sparsity and non-negativity on the higher-order cells, and basis patterns consisting of collinear complex cells emerged. These higher cells displayed properties like contour coding and end-stopping.

Hyvarinen and Hoyer noted that the learned basis pattern showed stronger collinearity than present in the data covariance. Therefore, it was concluded that higher-order structures, which could not be explained just from the covariance structure, have been found.

The fact that the response properties of the filters obtained from ICA like approach were not completely independent could be attributed to such approaches' linear nature. In other words, natural image statistics are too complex to be captured by linear models. Simoncelli and Shwartz studied the statistics of the responses of the liner filters to the natural images. Specifically, they plotted the pair-wise joint responses of non-overlapping and orthogonal filters (Schwartz and Simoncelli 2001). If the filters were independent, then any aspect of the response of one filter could not be predicted by the other filter's response. However, they found that the variation in response of one filter was dependent on the other's response.

Moreover, Simoncelli and Schwartz demonstrated that such dependence was found not only in visual stimuli but also in auditory stimuli and vanished in cases of white noise. They then proposed a generic normalization model to remove such dependence. This model included taking the squared sum of all responses and dividing each response with this sum. The model was effective in removing the observed dependencies.

Apart from the studies listed above, several other studies contributed to understanding natural stimulus statistics and the Efficient Coding Hypothesis's relevance. Foldiak developed a neural network that could reduce the statistical dependence between coding elements. The model utilized simple Hebbian units that received anti-Hebbian feedback (Földiak 1990). The network learned the feedback while encountering the inputs, producing sparse, and information preservation codes. This network model was among the first to achieve redundancy reduction in some form.

To demonstrate their theoretical studies' relevance, Dan, Atick, and Reid recorded individual cat LGN neurons' response to natural, time-varying images. The control images were white noise images. Their recordings showed LGN neurons' responses were

decorrelated for natural stimuli but not for white noise (Dan et al. 1996). This study provided a piece of strong evidence in support of the Efficient Coding Hypothesis.

With the advent of independent component analysis (ICA) to describe the emergence of simple cell-like filters, van Hateren and van der Schaff compared macaque simple cells' receptive field properties to the filters obtained through ICA. They reported that the properties like spatial frequency bandwidth, orientation tuning bandwidth, aspect ratio, and length matched very well (Van Hateren and van der Schaaf 1998) which showed that the simple cells were well-tuned to the statistics of natural stimuli.

In his studies, Bialek tried to decode stimuli in real-time using simple linear filters that estimated a time-varying signal based on spike trains (Bialek and Zee 1990, Bialek et al. 1991). He decoded several stimuli that were encoded in firing rates, like motion in blowfly's H1 cells. Later, he found that the filters obtained to decode sound signals from the spike trains are optimized for natural sounds (Rieke et al. 1993). He created stimuli with natural amplitude spectrum but unstructured phase spectrum and showed that the stimuli with natural amplitude spectrum are encoded more efficiently, with coding efficiency as high as 90% of the information transfer's fundamental limit. He also demonstrated that the dynamics of the coding process in primary auditory neurons are matched to the correlation structure of the natural sound. This match was also proposed to be the reason for high efficiency in coding such sounds compared to the white noise. He also predicted that the non-linearities in auditory processing might be increasing the coding efficiency for natural sound.

1.6. Limitations of efficient coding

As described in the previous section, the Efficient Coding Hypothesis has been the foundation for theoretical studies of sensory processing and has successfully explained

various aspects of sensory processing. The principle, however, has certain limitations that restrict its applicability in higher-order sensory processing. In this section, I discuss some of those limitations.

1.6.1 Requiring the knowledge of input statistics: The central idea in the Efficient Coding Hypothesis is to find a representation strategy that maximizes independence and reduces the redundancy among representation neurons. This redundancy reduction can be seen as minimizing the chances of representing the same information about the inputs more than once (Simoncelli and Olshausen 2001). Obtaining such representation, however, is not straightforward. The system must be aware of the entire distribution of inputs so that it can find components that are independent and can be represented by individual neurons. Differently distributed inputs will have different independent components; therefore, the nature of representation will change depending on the distribution. It is very difficult for a biological system to know the entire statistical properties of inputs. One may argue that the system can adapt to the natural environment over the evolutionary period, but for specific classes of objects like faces or non-natural shapes that are more experience-dependent, estimating distribution is not plausible.

1.6.2 Calculating probabilities in sensory circuits: Seeking independence among representation units allows the system to calculate the probabilities of occurrences of sensory events, rather than having them explicitly represented (Barlow 1989, Barlow et al. 1989). Calculating the probabilities has been asserted to be advantageous, as it permits calculations of complex logical functions of these probabilities, which might be essential in determining the associations among these events (Barlow 1989). In addition to this, knowledge of the probabilities helps the system determine the nature of representation for different sensory events (Barlow 1991). Though all the above

arguments are valid, a fundamental problem lies in implementing probability calculations in the neuronal circuits. As discussed previously, the estimation of occurrence probabilities of inputs requires determining the activation probabilities of the representation neurons. This calculation requires pooling the knowledge of activation probabilities of all neurons. Furthermore, calculating other logical functions of the probabilities requires complicated and restrictive connectivity as a substrate. On the other hand, connectivity observed in the brain and other parts of the sensory system lack any clear structure and are generally local. Their ability to pool information from all neurons and perform complex operations is, therefore, limited. However, it must be noted here that such limitation does not imply that independent representation units should not be utilized. It instead emphasizes that the strategy of using independent representation units might not be sufficient just by itself, and additional considerations on sensory representation are necessary to obtain a biologically plausible strategy.

1.6.3 Sparse distributed coding: The necessity to gather information from all neurons can be ignored if one considers different representation strategies. One such approach can be an explicit representation framework where the presence of an input can be determined by simple logical operations performed on a subset of the neurons rather than on the whole set (Barlow 1994). It is argued that if the representation is reasonably sparse, the inactive neurons carry very little information. So, one can obtain sufficient information about the input by considering only the active elements (Barlow 1994). Thus, the occurrence probabilities of inputs can be determined by gathering information from a relatively smaller number of neurons. However, even in this strategy, it is possible that the same representation unit might be involved in representing more than one input, which will make the activation probabilities of

shared neurons higher than the activation probabilities of non-shared ones. For inputs with overlapping representations, this will mean that shared neurons contribute more to determining joint occurrence probabilities than the non-shared neurons. However, such shared neurons by themselves are not very informative about any of the inputs, and therefore, the estimate of the joint occurrence probabilities of inputs will be unreliable. Note that the problem arises due to pooling information from only a subset of neurons. If information from all neurons is pooled, both active and inactive neurons will contribute to determining the joint occurrence probability of inputs. As the entire representation contains all information about the inputs, such estimates will be reliable.

1.6.4 Applicability in higher-order cognitive functions: While the requirement of independence among neurons is often celebrated and is the most sought-after feature in a representation strategy, it can also pose severe problems to a biological system. A fundamental problem for biological systems is to recognize objects presented in different forms and conditions. Presumably, it solves this problem by maintaining consistent representations of the objects in these situations. It is hypothesized that, with its physiological variability, the system needs to infer the representations when only its parts are activated (Clark 2013, Friston 2005). Independence among the neurons, on the other hand, means that a neuron's state does not depend on the states of other neurons. In other words, for independent neurons, one cannot predict the state of any neuron based on other neurons' states. This indeterminacy is directly conflicting with the inference requirement, and the two conditions cannot be met simultaneously. It can happen that because of some unknown reasons, a particular neuron could not fire, or an extra neuron fired. The system now needs to identify this error, but it is impossible to know whether an event like an error has happened or not with the independence condition satisfied. Thus, maintaining consistency in

representations with the inherent variability in the biological system is difficult when the neurons are independent.

1.6.5 Correspondence between neurons and features of inputs: As suggested by Barlow (Barlow 1961, Barlow 1989, Barlow et al. 1989, Barlow 1991), one way to increase independence among neurons is by making sure that the average activation probability of neurons is minimum. This strategy is manifested as minimization of bit entropies of the neurons. In this strategy, the stimuli are represented such that the most frequent ones have the least number of active representation units, and the rare events are often represented with many active units. This strategy's limitation is that the individual neurons' activity does not indicate any recognizable feature of the stimulus. The representation as a whole is meaningful, but its components' activity does not mean the presence of any substructure or identifiable property.

1.7. Hierarchical assemblies and view-based representations of objects

Parallel to the development of the theory of efficient coding, another set of theories was developed to explain invariant recognition of objects. This approach's main focus was to identify ways that could explain the consistent perception of objects when they are presented in different forms or viewed from different perspectives.

As the aim was to explain consistent perceptual experiences, the first studies in this field proposed the theory that the representation of objects should be based on some of their invariant properties (Pitts and McCulloch 1947). To understand the idea behind invariant properties, consider the set of all triangles. As we know, all triangles follow the triangle law, i.e., the sum of lengths of two sides is greater than the length of the third side. The triangle law can be regarded as an invariant property of triangles. One can define a rule that any

object with sides that follows triangle law is a triangle. Consequently, a representation scheme can be designed to indicate the pairwise sum of lengths of sides of objects and compare them to individual sides. In similar ways, other properties like area, perimeter, elongation, etc., were also proposed to be utilized as invariant properties (Bolles and Cain 1982). More complex properties like cross-ratio of four points (Gibson 1950a, Gibson 1950b, Gibson 1979) were also proposed as the basis of invariant representation.

It was easy to realize that such properties do not have wide applications. For example, it is tough to find properties of a 3D object that remain invariant in all its views. It was proposed that instead of using one invariant property, one can consider a group of properties such that some combined measurement of the properties remains invariant across different objects. In such a situation, any object can be considered to be present in an N-dimensional property space, and its different views can be thought to be distributed around it in this space (Tou and Gonzalez 1974). Such images can then be mapped to a unique object on the basis of their distance. However, it could be shown that, if such invariant measurement exists, its value will be constant for all objects (Clemens and Jacobs 1991, Burns et al. 1992, Moses and Ullman 1992). A constant measurement cannot be useful for differentiating objects as all objects will be mapped to the same value.

Another set of proposed approaches to generate the invariant representation of objects was based on the decomposition of objects into their feature components. These approaches aimed to identify simpler constituent parts of the objects first and then, based on the parts and structural relationship among them, identify the objects. Parts-based decomposition was the idea behind the computations performed in a perceptron (Rosenblatt 1957, Rosenblatt 1958) which could be utilized to recognize shapes like a triangle (Minsky and Papert 1969) irrespective of their size or location. The *"pandemonium"* scheme (Selfridge 1959) was also based on the idea of recognizing parts of the objects to invariably identify the entire object.

In addition to parts, later studies also included a description of structural relationships among components in the representation (Grimsdale et al. 1959, Clowes 1967, Winston 1975). In their seminal study, Marr and Nishihara (Marr and Nishihara 1978) also proposed using cylindrical components and their relationships to obtain stable yet sensitive representations of objects. The idea was further followed in the theory of *Recognition by components (RBC)* (Biederman 1987, Biederman 1985, Hummel and Biederman 1992, Cooper et al. 1992, Biederman and Cooper 1992, Biederman and Gerhardstein 1993) which introduced the concept of *"geons."* Geons were cylinder-like 3-dimensional components, and objects could be described as 3D models based on them. Similar 3-dimensional cylindrical features were also utilized in other studies (Binford 1971, Binford 1981, Brooks 1981). Uni-dimensional contour features known as *"codons"* (Hoffman and Richards 1984), or 2-dimensional surface patches (Dane and Bajcsy 1982, Dane 1981, Potmesil 1983, Faugeras 1984, Brady et al. 1985, Faugeras and Hebert 1986) were also used as a basis for describing objects. Though such descriptions successfully generated object models that remained invariant to various transformations, a significant drawback in using them was that they produced complex object models. Moreover, the requirement to include the relationship among parts in the description of objects was not tractable as exponentially many combinations, and relationships are possible.

The idea of hierarchical combinations of features was introduced to circumvent the problem of representing exponentially many feature combinations (Selfridge 1959, Sutherland 1968, Sutherland 1969, Barlow 1972, Milner 1974). In these theories, it was proposed that cells tuned to simpler features can be pooled together and connected to a higher-order cell so that the higher-order cell represents a combination of these features (Fukushima 1975, Fukushima and Miyake 1982, Riesenhuber and Poggio 1999b). Interestingly, pooling features removes information about their exact configuration. Therefore, a major assumption in these approaches is that a feature combination's internal

structure is not very important in its identification (Ullman 1996). Furthermore, as each combination is represented by individual neurons, which serve as parallel processing units, another underlying assumption in hierarchical approaches is that individual configuration of features can be classified independently of other parts and structures (Ullman 1996). However, to avoid representing all combinations of features, the approach required learning only the combinations observed in the objects (Block et al. 1962, Kabrisky 1966, Giebel 1971, Fukushima 1975). The edge detection capabilities of cells in the visual cortices of cats (Hubel and Wiesel 1962) and monkeys (Hubel and Wiesel 1968) motivated the approach, and the simplest features that were further combined into more complex structures were often edge-like. With all such assumptions and motivations, the approach was remarkably successful in attaining shift and scale invariance while representing 2-dimensional images. The first notable hierarchical model, *"Neocognitron"* (Fukushima and Miyake 1982), could produce a shift-invariant representation of 2D objects. Models based on *"shifter circuits"* (Anderson and Van Essen 1987, Olshausen et al. 1993) were used for generating similar location and scale-invariant representations. The hierarchical models were also successful in generating robust representations against occlusion (Shimojo et al. 1989, Fukushima 2005, Fukushima 2003, Johnson and Olshausen 2005). Neural network-based models (Perrett and Oram 1993, Oram and Perrett 1994, Riesenhuber and Poggio 1998, Riesenhuber and Poggio 2002, Koch and Poggio 1999, Riesenhuber and Poggio 1999a, Riesenhuber and Poggio 1999b, Riesenhuber and Poggio 2000, Wallis and Rolls 1997) were also proposed. They provided biologically plausible ways of carrying out the hierarchical assembly of features. Thus, the representation scheme based on the hierarchical assembly of features offered a way for invariant representation of 2-dimensional images. However, the approach just by itself was not useful in generating a consistent representation of 3-dimensional objects.

Invariant representation of 3-dimensional objects was proposed to be based on aligning its views to a model that is stored in the brain (Chien and Aggarwal 1987, Faugeras

and Hebert 1986, Fischler and Bolles 1981, Huttenlocher and Ullman 1990, Huttenlocher and Ullman 1987, Lowe 1985, Thompson and Mundy 1987, Ullman 1989, Linnainmaa et al. 1988, Lamdan et al. 1988). These approaches defined a set of transformations T incorporating changes in scale, position, or orientation for every object model M. The transformations were applied to the model to best align it to the view under consideration; recognition of an object corresponded to finding a suitable model-transformation pair (Ullman 1996). The first set of studies argued that any view or image was aligned to a single stored 3-dimensional object model (Shoham and Ullman 1988, Huttenlocher and Ullman 1990). The alignment was based on identifying a small number of corresponding features between the 3-dimensional object model and a 2-dimensional view and use them as *"anchor points"* to find the appropriate transformation. The model was selected based on the fit between the object's view and its transformed form (Ullman 1996). However, it was realized that a single 3-dimensional model might not be sufficient to recognize all different views of an object, especially in self-occlusion conditions. Therefore, the utilization of multiple 3-dimensional models of a single object was considered to account for its widely distinct views (Koenderink and Van Doorn 1979, David I Perrett et al. 1985, Rock and DiVita 1987, Grimson and Lozano-Perez 1987, Grimson 1990, Huttenlocher and Ullman 1990).

Another approach towards an invariant representation of 3-dimensional objects was based on the alignment of object views, not to its 3-dimensional model but a collection of its 2-dimensional images (Ullman and Basri 1991). The motivation behind such an approach was the realization that any view of an object can be expressed as a linear combination of its 2-dimensional images; therefore, the collection of such images can serve as the object's model. More specifically, a group of N images $\{M_{1j}, M_{2j}, \dots M_{Nj}\}$ can serve as a model for object j as any view V_{kj} of the object can be expressed as

$$V_{kj} = \sum_{i=1}^{N} \alpha_i M_{ij}$$

Here, the set $\{\alpha_i\}$ corresponds to the coefficients of linear combinations of the model images. The number of images required to be stored to account for any view of the object was shown to be as low as three (Ullman and Basri 1991) or two (Ullman and Basri 1991, Poggio 1990) for the 3-dimensional transformation of any general object, and one for symmetric objects (Vetter et al. 1994). The set of coefficients corresponding to each image in the model required to account for any view of the object was determined by searching through the entire space of coefficients (Yuille et al. 1989). Interestingly, it could be shown that explicit recovery of these coefficients could be avoided by mapping object views to a canonical image (Ullman and Basri 1991). Mapping any view to a canonical image $\boldsymbol{Q}$ of the object corresponded to finding a linear transformation matrix $\boldsymbol{C}$ such that for any set of object images $\{M_1, M_2, \ldots, M_N\}$

$$\boldsymbol{C}M_1 = \boldsymbol{C}M_2 = \cdots = \boldsymbol{C}M_N = \boldsymbol{Q}$$

As object views can be described as a linear combination of its images, such transformation implied that any view v could also be mapped to the image $\boldsymbol{Q}$ by the same transformation matrix

$$\boldsymbol{C}v = \boldsymbol{C}\sum_{i=1}^{N}\alpha_i M_i = \sum_{i=1}^{N}\alpha_i \boldsymbol{C}M_i = \sum_{i=1}^{N}\alpha_i \boldsymbol{Q} = \xi\boldsymbol{Q} \quad where\ \xi = \sum_{i=1}^{N}\alpha_i$$

The transformation matrix could be obtained in terms of a matrix $\boldsymbol{M}$ of independent images of the object as

$$\boldsymbol{C} = \boldsymbol{Q}\boldsymbol{M}^{-1}$$

Mapping any view of an object to a canonical image is essentially a framework for producing an invariant representation of the object where all object views are represented through a neuron tuned to the canonical view. Furthermore, the object's independent views or images that comprise matrix $\boldsymbol{M}$ and the transformation $\boldsymbol{C}$, can be generated from its simpler features using a hierarchical approach. Thus, the hierarchical approach and the linear combination of views of objects together comprised a bottom-up framework for producing invariant representations of 3-dimensional objects. The approach was further extended to

utilize a non-linear interpolation between images through a class of functions known as Generalized Radial Basis Functions (GRBFs) (Poggio and Girosi 1989, Poggio and Girosi 1990b, Poggio and Girosi 1990a, Girosi and Poggio 1990, Tikhonov and Arsenin 1977, Poggio et al. 1987, Powel 1987, Broomhead and Lowe 1988, Poggio and Edelman 1990, Edelman and Poggio 1989).

The invariant representation framework based on hierarchical assembly and the linear combination of views of objects is supported by physiological and psychological studies. Psychological studies showing a decreased recognition of objects with changes in their viewing direction (Bartram 1974, Palmer et al. 1981, Jolicoeur 1985, Corballis 1988, Tarr and Pinker 1989, Jolicoeur 1990, McMullen and Jolicoeur 1990, Tarr and Pinker 1990, Tarr and Pinker 1991, Bülthoff and Edelman 1992, Edelman and Bülthoff 1992, Humphrey and Khan 1992, Farah et al. 1994, Tarr 1995, Gauthier and Tarr 1997) provide evidence in favor of this framework. Both humans (Poggio and Edelman 1990) and monkeys (Logothetis et al. 1994) showed similar trends in their performances in such studies. Studies compared human performance with the performance of an *"ideal 2D observer"*. The idea observer stored all previously seen views of the objects and compared any novel view to each of the stored views separately. These studies demonstrated that human use mechanisms that are better than the ideal observer and comparing individual views were insufficient for accounting for human performance (Liu et al. 1995, Moses et al. 1994). While these studies provide evidence in favor of the hierarchical view-based framework, other studies (Biederman and Cooper 1992, Biederman and Gerhardstein 1993) have shown evidence against it. However, these studies have been criticized for lacking generality, evidence, and explanatory power (Tarr and Bülthoff 1995).

In addition to the psychological studies of human performance in object recognition, studies on the physiological properties of neurons in high-level visual processing areas like V4 and IT have revealed the presence of shape-specific cells (Gross 1992, Tanaka et al.

1991, Fujita et al. 1992, Tanaka 1992) and involvement of these cells in object recognition (Damasio et al. 1990, Damasio and Damasio 1993). Studies have also demonstrated that objects and faces are represented by distributed patterns of active neurons in the IT regions of the brain (Perrett et al. 1982, DI Perrett et al. 1985, David I Perrett et al. 1985, Rolls 1984, Baylis et al. 1985, Rolls and Tovee 1995, Young and Yamane 1992). Furthermore, studies have also demonstrated that V4 and posterior IT injuries lead to loss of abilities to compensate for changes in size, orientation, or illumination conditions, rather than recognizing the shape itself (Schiller and Lee 1991, Weiskrantz 1990, Schiller 1995), supporting the alignment-based approaches. View invariant representations have also been reported (Booth and Rolls 1998).

1.8. Limitations of hierarchical assembly and view-based representation framework

1.8.1 Incompatibility with efficient coding: As described in the previous sections, the Efficient Coding Hypothesis, which explains lower-level visual processing, requires independent structural component-based representations. Independence among components means that the presence or absence of one component cannot be determined from others' presence or absence. On the other hand, the hierarchical assembly of features requires learning of association among simple features to build more complex structures. These associations can only be learned if they are made available to the system, and by their definition, independent components cannot reflect such associations. This conflict creates a compatibility issue between the two theories. Most of the studies that demonstrate the working of hierarchical approaches (Fukushima and Miyake 1982, Anderson and Van Essen 1987, Wallis and Rolls 1997, Riesenhuber and Poggio 1999b) do not consider low-level features to be

independent. They build on structurally complex features using features that are structurally similar to independent components but are not necessarily independent.

1.8.2 Selection of views: The view-based representation system proposes that any view of an object can be expressed as a linear combination of its 2-dimensional images (Ullman and Basri 1991). A set of 2-dimensional views or images corresponding to an object can serve as its internal model and can be learned and represented along sensory pathways using hierarchical approaches (Fukushima 1975, Fukushima and Miyake 1982, Riesenhuber and Poggio 1999b). However, as there can be infinitely many 2-dimensional images of any particular object and not all of them can comprise its model, a problem arises in selecting the images that are best suited to serve as the model of the object. None of the proposed theories address this problem. An optimal set of images can be the one that compensates for factors like self occlusions and can be utilized in representing a wide range of object views. Still, none of these theories explain how this optimal set can be learned.

1.8.3 Mapping to a common view: The theory of view-based representation of images proposes that one does not need to explicitly find the coefficient of linear combinations of images to explain any view. Instead, all the views of an object can be mapped to a canonical view using a linear transformation (Ullman and Basri 1991). The object is recognized when a sufficiently good match between its input view and a canonical view is found. However, the approach assumes that the canonical views of objects are unique, and views of different objects can always be mapped to their respective canonical views. In other words, an underlying assumption in this approach is that views that uniquely identify objects can always be found. While a specific way of selecting canonical views has not been discussed

in the approach, it seems less likely that such views exist. Any 2-dimensional image or view of a 3-dimensional object is its projection on a lower, 2-dimensional space. As multiple objects can produce the same projection, any canonical view can likely be matched with views of numerous objects. Such a situation will limit the ability of the system to differentiate between these objects.

1.8.4 Inaccuracy in estimating occurrence frequencies of objects: The use of canonical views to generate invariant representations of objects necessitates that the frequency of occurrence of an object is counted based on its canonical view. For example, suppose a particular neuron is tuned to the canonical view of an object. In that case, the occurrence frequency of the object can only be calculated by estimating the activation frequency of the neuron. However, as discussed in the previous limitation, if the canonical view is matched to multiple objects, its activation probability will reflect a union of occurrences of all these objects. Such a neuron cannot provide a reliable estimate of the occurrence frequency of any individual object, and therefore, cannot be utilized in detecting its association with other objects. Thus, though canonical views provide a way to generate invariant representations of objects, they do not present a reliable way to detect associations between objects.

1.9. Discussion

The organism must rely on the internal representations of the objects formed in its sensory system to achieve competence in invariably recognizing objects and detecting associations among multiple objects. In this chapter, I have described two sets of theories, the Efficient Coding Hypothesis and hierarchical assembly and view-based representation. They have been successful in explaining different aspects of sensory processing yet cannot provide a representation framework that allows the organism to perform the above-

mentioned tasks. The representations of objects obtained through efficient coding allow the detection of association among objects but do not permit invariant recognition. On the other hand, hierarchical and view-based approaches allow invariant representation but do not explain how the organism can detect correct association among objects. A novel approach towards understanding sensory processing is needed.

In this study, I propose a framework that allows the invariant representation of objects that are also efficient. The proposed framework is based on the informativeness of features rather than their independence. It resolves many limitations faced by the current approaches. I show that the framework can successfully explain information processing both at higher and lower levels of the visual pathway, as well as in other sensory modalities. The details of the framework are explained in the next chapter.

This page is intentionally left blank

CHAPTER 2

An adaptive framework for representing objects

Table of Contents

2.1. Introduction

In the previous chapter, I have described two sets of theories that form our understanding of sensory processing. The Efficient Coding Hypothesis, proposed by Barlow and others, suggests that the system should represent sensory inputs in a way that minimizes information loss and reduces redundancy among representation neurons (Attneave 1954, Barlow 1961). The theory recommends adapting to inputs' statistics and representing them based on independent features to reduce redundancy (Barlow 1989, Barlow et al. 1989, Barlow 1991). Inspired from information theory (Shannon 1948), this framework is remarkably successful in describing the early stages of sensory processing across different modalities (Laughlin 1981, Atick 1992, Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997, Lewicki 2002, Smith and Lewicki 2006). In the visual system, the theory successfully explains receptive field properties of retinal ganglion cells (Srinivasan et al. 1982, Atick and Redlich 1990), bipolar cells (Barlow et al. 1957, Hartline and Ratliff 1972), LGN cells (Dan et al. 1996, Dong and Atick 1995) and primary visual cortex neurons (Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997).

Another set of theories that primarily aims to describe the higher-level processing of visual information introduces concepts of hierarchical assembly of features and view-based representations of objects (Fukushima and Miyake 1982, Anderson and Van Essen 1987, Ullman and Basri 1991, Poggio 1990, Vetter et al. 1994, Ullman 1996, Riesenhuber and Poggio 1999b, Ullman 1998). This set of theories proposes that the system achieves perceptual invariance by representing objects as a combination of their multiple views. It learns these views by systematically combining simpler features into progressively complex combinations across multiple processing levels (Fukushima and Miyake 1982, Riesenhuber and Poggio 1999b). Any novel or known view of an object is subsequently mapped to a

specific linear combination of the learned views, and objects are recognized when the map crosses a threshold criterion (Ullman 1996). This framework has successfully explained findings of several psychophysical studies of human recognition abilities (Bartram 1974, Farah et al. 1994, Bülthoff and Edelman 1992, Edelman and Bülthoff 1992, Humphrey and Khan 1992, Jolicoeur 1985, Corballis 1988, Jolicoeur 1990, McMullen and Jolicoeur 1990, Palmer et al. 1981, Tarr and Pinker 1989, Tarr and Pinker 1990, Tarr and Pinker 1991, Tarr 1995, Tarr and Bülthoff 1995, Gauthier and Tarr 1997). Modern neural network-based computer vision studies, which have been remarkably successful in recognizing objects (Sermanet et al. 2013, Girshick 2015, Lin et al. 2017) are also based on it.

Though these theories successfully explain various aspects of visual processing, they only have limited applicability to a biological system. The Efficient Coding Hypothesis proposes encoding independent features to attain a factorial representation of sensory inputs (Barlow 1989, Barlow et al. 1989, Barlow 1991). As described in the previous chapter, such representation allows the detection of association among different objects (Barlow 1987, Barlow 1989, Barlow et al. 1989, Barlow 1991, Barlow 1994). It permits the system to readily calculate the marginal and joint probabilities of sensory events and use them to estimate the conditional probabilities that indicate associations. However, finding independent features of the natural environment requires accurate estimates of its statistical properties. Biological systems, on the other hand, rely on experience to gain knowledge of their surroundings. It is not very clear how they can obtain such estimates. Moreover, statistical analysis of natural scenes has shown that the natural environment's independent features are localized (Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997). These features are sensitive to common transformations and alterations of objects, and hence, do not form an appropriate basis for invariant representations.

Similarly, view-based representation schemes, which explain perceptual invariance, do not specify the views to learn. There can be infinitely many views of individual objects, and any particular view can originate from multiple objects. In such conditions, even

mapping views analytically to a common canonical view corresponding to an object's identity is a challenging task, let alone its biological implementation. Furthermore, view-based representations are likely to be redundant and lack a factorial nature that permits the detection of dependence among objects.

Indeed, the inherent problem in view-based and efficient representation schemes is their compatibility. Efficient representations of objects are based on independent structural components and are necessary for detecting dependence. View-based representations, on the other hand, achieve invariance in object representation using multiple views. As various views from the same object cannot be independent of one another, view-based representations cannot be efficient. Therefore, their applicability in determining dependence among objects is limited. This incompatibility presents a complication where the system needs to compromise either invariant representation or object association detection.

Here, I introduce a novel framework that resolves this complication. I propose that representing objects based on their most informative features can achieve efficiency and invariance in object representation simultaneously. Specifically, in this chapter, I describe the framework's formulation and highlight its several crucial aspects. Starting with an introduction to a more general notion of features and their informativeness, I illustrate differences in the informativeness of features and the most informative feature's uniqueness. I show that individual features' informativeness changes with their occurrence and independent features can be non-informative. I further describe how structurally related groups of features can be more informative than individual features, and adding more features to the most informative group does not change its informativeness. I propose that this particular aspect of feature groups can be utilized in achieving invariance. In the next few sections, I formulate the task of representing objects as a process of basis transformation and highlight how incomplete knowledge of the environment statistics can alter the dependence among features. I derive a limit on the system’s capacity of relaying information

to show further that an efficient way of representing a finite number of objects can be based on unique structures. In the end, I formulate the framework for representing objects based on unique structures and discuss its similarity with the prevailing approaches of implementing sensory processing as basis transformation.

2.2. Definition of features

In everyday language, the *features* of an object are the structural components that constitute it. A triangular object has three straight line features, whereas a dome structure has a semi-circular arc feature. The idea here is to describe any structure in terms of simpler, less complex structures, with lower complexity structures being the features of more complex structures. In practice, structures of lower complexities comprise further smaller fragments. Like smaller point-like structures constitute straight lines and circular arcs. However, these point-like structures are not recognizable geometric shapes like lines or arcs. Therefore, for this study, we do not refer to such structures as features. In other words, in this study, we consider features as the *minimal recognizable structures* of an object.

With this definition of features, an important point to note is that a combination of features, just by itself, is not sufficient to characterize an object. Consider the case of a *triangle* and the English alphabet *"A."* These shapes comprise three different line segments; however, this limited information is insufficient to distinguish them. The aspect of features that differentiates these two structures is their configuration or arrangement. In a *triangle,* all the line segments are joined end to end, whereas in the letter *"A,"* two line segments a joined at their end while the other joins them at an intermediate position. Such a description of feature configuration highlights how features are structurally associated with one another. It contains the information that one needs to identify the shapes. Thus, the identity of a shape or an object is embedded in its constituting features and the structural association among these features.

Interestingly, both these aspects of features, namely, recognizability and structural associations, are instantiated in letters and words from any language. Letters or alphabets in any language correspond to the minimal identifiable symbol utilized in conveying information. Likewise, words are specific arrangements of these symbols that have their meaning or identity. In this regard, an equivalence exists between features and letters, and between words and objects, with the arrangement of letters reflecting the structural association among features. Therefore, in the next sections, I will use words from the English language as examples of objects and corresponding letters as features to bring out several essential aspects of features that play a vital role in object recognition.

As we define features and draw a correspondence between them and letters, it is critical to realize that these definitions only help us understand the aspects of features that can be useful in the recognition task. One should not expect the brain to consider such composition of the objects while representing them. Presumably, the representation of objects is derived based on their statistical properties while satisfying the constraints imposed by factors such as the system size, states of the neurons, and noise conditions. Based on factors like brain state, developmental stage, and health conditions, different representations may arise where individual representation neurons respond to input components of varying complexity. These components may not belong to any recognizable class of shapes, and inherent properties about these shape classes like continuity might be missing. However, we can expect that the aspects of these components that make them suitable for serving as a basis of representation can be explored and identified using the recognizable shape classes. In other words, it is necessary to acknowledge that certain features are represented not because they constitute the inputs, but inputs should be described in terms of specific features because they are represented.

2.3. Informative and non-informative features

In the previous chapter, I have described that *"regularity"* in events can be assessed based on the probabilistic concept of *dependence*. The idea of "*regularity"* in the environment, which is analogous to the concept of *predictability*, brings out the environment's rules and structures. It tells us that a night follows a day, and winter ends in spring, leading to summer. However, to identify these *"regularities,"* or in other words, determine the predictability, one has to know the dependence between events. Dependence between events is the influence that one event's occurrence has on others' occurrence. In more formal terms, two events are dependent when the *marginal occurrence probability* of one event is different from its *conditional occurrence probability,* conditioned on the other event. Simply put, one can see *dependence* as the reliability with which one can guess the next set of events based on the current situation.

Thinking in similar ways, it is not difficult to realize that encountering an object or its features is also an event. Therefore, in line with the idea of dependence among events, different objects, or objects and features can be dependent. The dependence will indicate how an object or a feature influences our guess about the dependent object. It will highlight the structure of objects and the environment in terms of their constituents. For example, our prediction of an object based on a feature changes with the feature's presence or absence in the object. Thus, the dependence between a feature and an object indicates the object's composition. Similarly, the dependence between objects is derived from their predictability and suggests regularity in the environment. For recognition of objects, knowledge of dependence between features and objects allows us to find a set of features that differentiate multiple objects. A representation framework can then be based on such features to produce invariant, unique representations associated with object identities.

Narrowing down to features suitable for representing objects, however, requires more than just dependence. It needs a measure of the degree of dependence based on which

certain features can be selected while others can be discarded. Such a degree of dependence is measured in terms of *informativeness* or *information content* of the feature. *Information content* of a feature about an object is the amount of information that the feature contains about the object. It is the extent to which the object can be accurately identified with just the knowledge of the feature. The more information a feature has about an object, the more definitively the object can be determined. Consider three words, *"am," "an,"* and *"ant,"* for example. As discussed in the previous section, these words can be regarded as three different objects, with the alphabets corresponding to their features. Clearly, these are the examples of objects with some features shared among themselves, while some of the features are unique to individual objects. Suppose one picks the feature *"a"* and tries to guess the object based on it. The chances of a correct guess will be feeble in this situation as all objects have that feature. Similarly, while guessing an object based on the feature *"n,"* the chances of a correct guess, though better than the previous attempt, will still lack precision, and there will be confusion between objects *"an"* and *"ant."* However, objects can be uniquely identified based on features *"m"* or *"t."* We can see that different features, depending on their commonality, reduce the uncertainty about objects differently. We say that these features contain different amounts of information about the objects. In the given example, common features like *"a"* and *"n"* contain less information, and conversely, unique features like *"m"* and *"t"* contain the most information about the object.

The formal definition of information content (Cover and Thomas 1991) of a feature about an object is based on the object's marginal occurrence probability and its joint occurrence probability with the feature. More precisely, if $\mathbb{P}(\mathbf{object})$ and $\mathbb{P}(\mathbf{features})$ are the marginal probabilities of the presence of the object and the feature, and $\mathbb{P}(\mathbf{object}, \mathbf{feature})$ is the joint occurrence probability of object and feature, then information content of the feature about the object, $\boldsymbol{I}(\mathbf{object}; \mathbf{feature})$, is expressed as

$$\boldsymbol{I}(\mathbf{object};\mathbf{feature}) = \sum\sum \mathbb{P}(\mathbf{object},\mathbf{feature})\ \log\frac{\mathbb{P}(\mathbf{object},\mathbf{feature})}{\mathbb{P}(\mathbf{object})\mathbb{P}(\mathbf{feature})}$$

The double summation accounts for the presence and absence of both the object and the feature.

In the previous example, if we consider each object to be equally likely i.e.

$$\mathbb{P}(\text{"am"}) = \mathbb{P}(\text{"an"}) = \mathbb{P}(\text{"ant"}) = \frac{1}{3}$$

then

$$\mathbb{P}(\text{"a"}) = 1 \text{ and } \mathbb{P}(\text{"am","a"}) = \mathbb{P}(\text{"an","a"}) = \mathbb{P}(\text{"ant","a"}) = \frac{1}{3}$$

therefore

$$\boldsymbol{I}(\text{"am";"a"}) = \boldsymbol{I}(\text{"an";"a"}) = \boldsymbol{I}(\text{"ant";"a"}) = \frac{1}{3}\log 1 = 0$$

meaning that *"a"* is not informative about any of the objects; therefore, it is not useful in their recognition. Indeed, guessing objects based on *"a"* does not reduce any uncertainty. Similarly, it can be shown that

$$\boldsymbol{I}(\text{"an";"n"}) = \boldsymbol{I}(\text{"ant";"n"}) = 0.25;\ \boldsymbol{I}(\text{"am";"n"}) = 0.92$$

and

$$\boldsymbol{I}(\text{"am";"m"}) = \boldsymbol{I}(\text{"ant";"t"}) = 0.92$$

which indicates that the information content of *"n"* about any word is larger than *"a"* and thus explains the improvement in chances of correct guess using it. It also shows that, of all letters, unique letters *"m"* and *"t"* contain the most information about individual objects. Thus, this exercise illustrates two important points

1. Different features contain different amounts of information about objects
2. Most informative features are unique to individual objects

An important aspect of selecting features for representing objects is also highlighted in the previous example. Consider a particular object, *"ant,"* for instance. There are three features in the object, namely, *"a," "n,"* and *"t."* While *"a"* contains no information about the object, *"n*" and *"t"* contain more information about it, with *"t"* containing the most

information. In terms of uniqueness, feature *"a"* is the least unique as it is common in all the objects, and features *"n"* and *"t"* are successively more unique as they are shared among one more and no other object, respectively. Now, suppose one chooses to represent the object based on the features that are common among objects. In that case, two different objects with similar features will likely have the same representation. For example, representing *"ant"* based on *"a"* and *"n"* will make its representation identical to *"an."* On the other hand, a representation based on informative, unique features avoids such scenarios, as including *"t"* in the set of features will make the representation of *"an"* different from *"ant."* This shows that any representation framework that aims to produce distinct representations of different objects should be based on unique features. However, such a framework's invariance and efficiency need to be established and will be further discussed in the later sections.

2.4. Information content of independent features

As described previously, the Efficient Coding Hypothesis seeks to achieve factorial object representations. It seeks a representation scheme where the probability of an object's occurrence can be factored into the probabilities of occurrence of its represented features. Such a scheme presents an easy way for the system to estimate the object's occurrence frequencies, which is necessary for determining dependence among objects and understand the environment in terms of such dependence. The factorial nature of code arises when the features that serve as the basis of representation are independent. Interestingly, the Efficient Coding Hypothesis does not consider the information content of the features. Yet, in the previous section, we have discussed that features with high information content can uniquely characterize individual objects and can be particularly useful in their recognition. With such an understanding of the features' information content, it becomes imperative to analyze independent features' information content and assess their usefulness in object recognition.

To assess the information content of independent features, we need to find features that are independent of one another. To obtain such features, let us consider a set of four words, namely *"am," "an," "ant,"* and *"amt."* As in the previous example, these words denote four distinct objects with four distinctive features *"a," "m," "n,"* and *"t."* However, unlike the previous consideration of objects to be equally likely to occur, let us assume that these objects appear in the environment with probabilities x, x, $0.5 - x$, and $0.5 - x$ respectively, where x is a positive number between 0 and 0.5. This distribution allows us to calculate the marginal occurrence probabilities of individual features in terms of variable x. In particular, as the feature *"a"* appears in all three objects, its marginal occurrence probability, $\mathbb{P}("\boldsymbol{a}")$ can be expressed as

$$\mathbb{P}("\mathrm{a}") = x + x + \frac{1}{2} - x + \frac{1}{2} - x = 1$$

In similar ways, marginal probabilities of other features can also be calculated

$$\mathbb{P}("\mathrm{m}") = \frac{1}{2}, \mathbb{P}("\mathrm{n}") = \frac{1}{2}, \text{and } \mathbb{P}("\mathrm{t}") = 1 - 2x$$

The same approach can be applied to calculate the marginal probabilities of pairs of features as well. Without considering the structural arrangement of features, i.e., relative positions of the letters, we have

$$\mathbb{P}("\mathrm{am}") = \frac{1}{2}, \mathbb{P}("\mathrm{an}") = \frac{1}{2}, \mathbb{P}("\mathrm{at}") = 1 - 2x$$

$$\mathbb{P}("\mathrm{mn}") = 0, \mathbb{P}("\mathrm{mt}") = \frac{1}{2} - x$$

$$\mathbb{P}("\mathrm{nt}") = \frac{1}{2} - x$$

Comparing the features' joint occurrence probabilities with the product of their marginal probabilities, we find that

$$\mathbb{P}(\mathrm{am}) = \mathbb{P}(\mathrm{a})\mathbb{P}(\mathrm{m}), \mathbb{P}(\mathrm{an}) = \mathbb{P}(\mathrm{a})\mathbb{P}(\mathrm{n}), \mathbb{P}(\mathrm{at}) = \mathbb{P}(\mathrm{a})\mathbb{P}(\mathrm{t})$$

$$\mathbb{P}(\mathrm{mn}) \neq \mathbb{P}(\mathrm{m})\mathbb{P}(\mathrm{n}), \mathbb{P}(\mathrm{mt}) = \mathbb{P}(\mathrm{m})\mathbb{P}(\mathrm{t})$$

$$\mathbb{P}(\mathrm{nt}) = \mathbb{P}(\mathrm{n})\mathbb{P}(\mathrm{t})$$

These findings illustrate that feature *"a"* is independent of features *"m," "n,"* and *"t,"* and feature *"t"* is independent of features *"m"* and *"n."* On the other hand, features *"m"* and *"n"* are not independent of one another.

If the efficient coding principle is followed in this situation, then the four objects can be represented using a set of three neurons. The neurons may be tuned to the collection of independent features, i.e., either the set {*"a," "m," "t"*}, or the set {*"a," "n," "t"*}. One can also think of using just features *"m"* and *"t"* or features *"n"* and *"t"* for representing all objects. However, this scheme will result in representing objects *"an"* or *"am"* with only inactive neurons, which though theoretically possible, is not applicable for biological systems. Moreover, representing objects with three independent neurons will produce distinct representations for the objects, and the occurrence probabilities of the object can be calculated from the activation probabilities of the neurons representing the features. For example, if we consider neurons n1, n2, and n3 to be tuned to features *"a," "m,"* and *"t"* respectively, then the probability of occurrence of the object *"ant"* can be calculated by multiplying the probabilities of neurons n1 and n3 to be active with the probability of n2 to be inactive i.e.

$$\begin{aligned}\mathbb{P}(\text{"ant"}) &= \mathbb{P}(\text{n1} = \text{active})\mathbb{P}(\text{n2} = \text{inactive})\mathbb{P}(\text{n3} = \text{active}) \\ &= \mathbb{P}(\text{n1} = \text{active})\big(1 - \mathbb{P}(\text{n2} = \text{active})\big)\mathbb{P}(\text{n3} = \text{active}) \\ &= 1.\left(1 - \frac{1}{2}\right).(1 - 2\text{x}) = \frac{1}{2} - \text{x}\end{aligned}$$

It can also be verified that redundancy in such representation is minimal.

However, if we examine the information content of these features about different objects, we find that feature *"a"* is common among all objects, and hence it does not contain information about any object i.e.

$$\boldsymbol{I}(\text{"am"};\text{"a"}) = \boldsymbol{I}(\text{"an"};\text{"a"}) = \boldsymbol{I}(\text{"ant"};\text{"a"}) = \boldsymbol{I}(\text{"amt"};\text{"a"}) = 0$$

Depending on the value of the variable x, features *"m," "n,"* and *"t"* contain more information than *"a"* as they appear in only two of the four possible objects.

Thus, we find that independent features are not necessarily the most informative pieces of the object's structure. Depending on their commonality, they may contain very low information about any individual object. In this particular example, the least informative feature *"a"* was common among all inputs, and therefore, was not informative about any object. Such information content of independent features is not specific to this particular case. Statistical analysis of natural scenes has demonstrated that the natural environment's independent components are oriented localized edges (Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997). These independent edges appear in all possible orientations (Olshausen 2013) and tile the contours outlining the objects. As natural objects have regular contours that do not break or change abruptly, most objects likely have contour portions of all possible orientations. In this situation, any particular orientation is not specific to any object. Therefore, the independent components of natural scenes, like the independent structures in the example above, are not very informative of individual objects.

2.5. Representations based on informative features

We have discussed that different features or components of objects may contain different information about them. Depending on how common a feature is among multiple objects, it can be very informative about a particular object or contain no information about any specific object. In this regard, we have also seen that independent features proposed under the Efficient Coding Hypothesis to serve as a basis for representation may not be very informative about individual objects. Thus, by putting no restriction on the represented features' information content, the efficient coding principle allows object representations to be based on features with minimal information content. In this section, I will describe how representations based on more informative features are more distinct than those based on independent features and how they can be more useful in recognizing objects.

To understand how informative features can be more useful for recognition purposes, let us consider the example from the previous section of four objects depicted as four words *"am," "an," "ant,"* and *"amt."* Assuming the probability of occurrence of these objects to be x, x, $0.5 - x$, and $0.5 - x$ respectively, we have seen that there are two sets of independent features, namely {*"a," "m," "t"*} and {*"a," "n," "t"*} that can serve as the basis for representing these four objects. We have also seen that feature *"a"* is the least informative feature, and features *"m," "n,"* and *"t,"* depending on the value of parameter x, can have more information about individual objects than *"a."* Now, let us consider two representation scenarios, one where objects are represented in terms of independent features *"a," "m,"* and *"t,"* and the other where objects are represented using more informative features *"m," "n,"* and *"t."* Note that in the second scenario, features *"m"* and *"n"* are not independent, and therefore, will never be selected together for representing objects under the Efficient Coding Hypothesis. Assuming that three neurons n1, n2, and n3 are tuned to the three features, the representations of objects in the two scenarios can be depicted as shown below (**Figure 2.1**). We notice that, while the representations of all the objects in both scenarios are distinct, representations of objects in the second scenario have less overlap on average than the representations in the first scenario. Such minimal overlap is a direct reflection of the uniqueness of informative features, which imparts them the ability to distinguish the objects better and will be discussed later in the section.

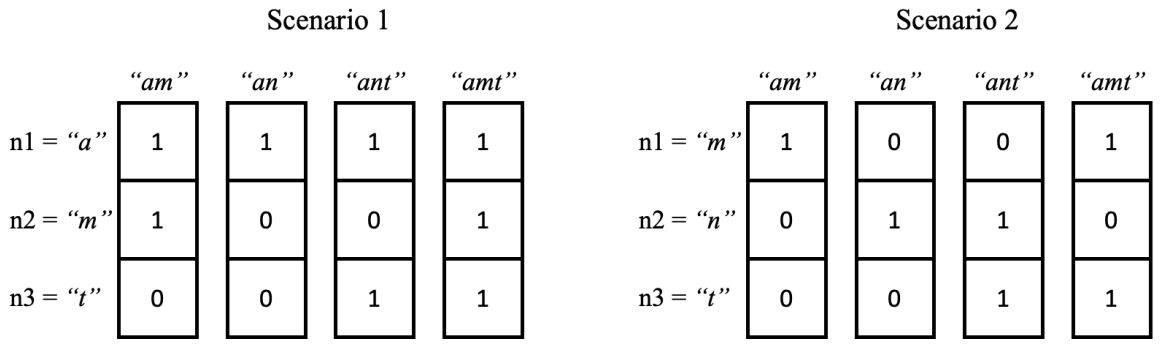


Figure 2. 1: Representation of four objects based on independent features and informative features

In scenario 1, four objects, namely "am," "an," "ant," and "amt," are represented based on their independent features "a," "m," and "t." In scenario 2, the same objects are represented based on informative features, namely "m," "n," and "t." Features "m" and "n" are informative but not independent. Notice that representations based on informative features have less overlap, and therefore are more distinct.

Another critical aspect of the second scenario representations is highlighted if we consider situations where specific neurons are lost or become inactive. Such situations arise in neural circuits due to trauma, injury, or the presence of a refractory period for neurons. In refractory periods neurons do not fire immediately after firing for a certain time, even when the stimulation continues. Considering n1 or n3 to be in a refractory period in both scenarios, we find that loss of n1 does not severely affect the system's ability to represent objects. Three distinct representations of four different objects are formed in both scenarios (**Figure 2.2.1**), which the system can utilize to differentiate the objects and recognize them. Similarly, when n3 is lost, only two distinct representations are formed in both scenarios (**Figure 2.2.2**), and the system losses its ability to differentiate between *"am"* and *"amt"* or between *"an"* and *"ant."*

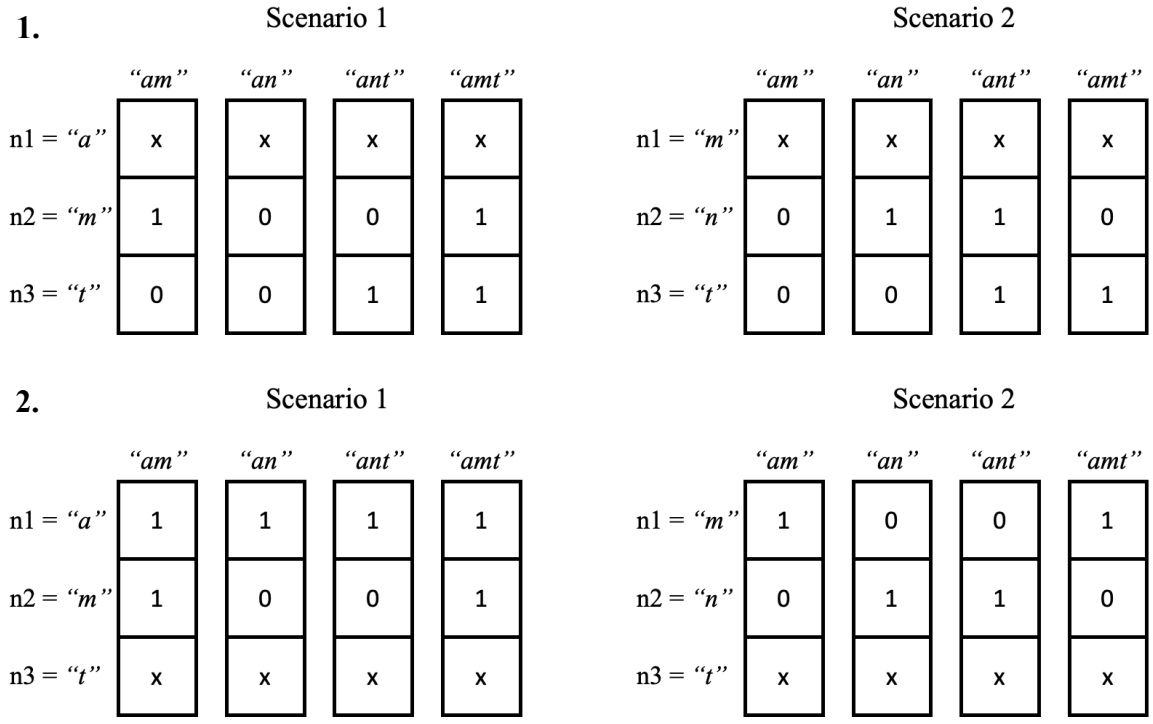


Figure 2. 2: Effect of loss of neurons on representations

Loss of neurons affects the system's ability to represent objects. When either neuron n1 or n3 is lost (denoted by symbol x), the system loses its ability to **1.** *represent objects or* **2.** *differentiate between two objects. This effect persists irrespective of the nature of the feature being utilized for representation. Here in scenario 1, objects are represented based on independent features, and in scenario 2, objects are represented based on informative features.*

However, suppose we consider the loss of n2 (**Figure 2.3**). In that case, we find that three distinct representations are formed in the second scenario, whereas, as before, only two different representations are formed in the first scenario. As a result, the system in the first scenario cannot tell the difference between *"am"* and *"an"* or between *"ant"* and *"amt,"* while the system in the second scenario can still distinguish all objects. Such differences

indicate that representing objects using more informative features makes the system more resistant to the commonly observed corruptions introduced by noise or neuronal loss.

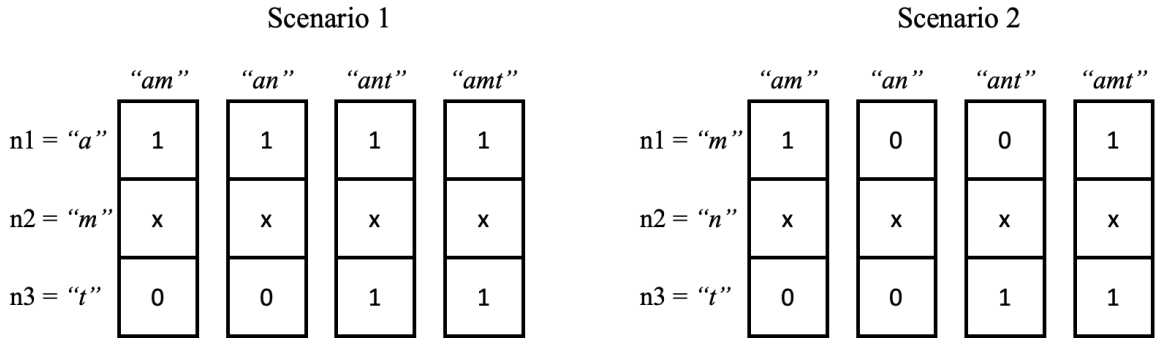


Figure 2. 3: Representation based on informative features preserves distinctiveness

Though the loss of neurons affects the system's ability to maintain distinction among object representations, a representation based on informative features is more likely to preserve distinction than the one based on independent features. After losing neuron n2, representations of objects "ant" and "amt" as well as representations of objects "am" and "an" become identical when they are represented based on independent features in scenario 1. Such representations are likely to prevent the distinction of any of the four objects. On the other hand, in scenario 2, when objects are represented based on informative features, loss of neuron n2 prohibits representing object "an." Still, the ability of the system to represent the other three objects is preserved.

Thus, the above example illustrates that using more informative features makes the representations of objects more distinct and maintains the distinctiveness against corruptions like a neuronal loss. The distinctiveness of representations is beneficial in distinguishing objects. It highlights the differences between similar objects and allows the system to

differentiate between them. Furthermore, more distinct or less overlapping representations can be utilized to estimate individual objects' occurrence frequencies. As neurons' activation indicates the presence of objects they represent, a less overlapping or completely non-overlapping set of representations will produce neurons that are activated only when a particular object or a specific collection of objects is present. The occurrence frequency of such objects can be estimated by comparing individual neurons' activity with the total activity of all the neurons. In other words, normalized activity levels of neurons will be proportional to the occurrence frequencies of objects being encoded by them. Similarly, joint occurrence probabilities of objects can be determined by pooling information from a subset of neurons. This presents a strategy different from factorial coding to calculate the occurrence probabilities of objects.

It is essential to realize that the distinctiveness of representations is a direct consequence of the informativeness of the features. More informative features are unique to individual objects, and when they are utilized in representing objects, the resulting representations are non-overlapping. Therefore, using these features for representing objects restricts the information from being distributed across multiple neurons. It enables the system to retain high information about most objects when some neurons are lost or when the system faces some damage. This concentration of information is reflected in the maintenance of the representations' distinction in the second scenario.

On the other hand, independent features are shared among multiple objects and do not necessarily have high information content. Therefore, while representing objects based on these features, information about individual objects is dispersed across multiple neurons. In this scenario, a few neurons' loss translates into a loss of information about numerous objects. The system fails to preserve information about individual objects, indicating the limited usefulness of such a representation scheme.

2.6. Informativeness of feature combinations

As we have seen, the object's features contain information about them; unique features contain most information, whereas common features contain the least information. Utilizing these unique features in representing objects makes the representation distinct and preserves this distinction against common system-wide corruptions induced due to factors like a neuronal loss. Such distinctions in representations make them useful for recognizing objects and provide an alternative way to determine associations among different objects. However, it is not always possible to find features that are highly informative about individual objects. For example, in the four objects presented in the previous section, none of the individual features are unique to any object. In situations like these, it becomes necessary to address how structural elements that uniquely characterize particular objects can be determined.

A way to increase the information content of a structure is by incorporating more individual features in it. Consider the previous example of four objects *"am," "an," "ant,"* and *"amt*. The individual features *"a," "m," "n,"* and *"t"* are not unique to any object, and therefore, are not sufficient to resolve ambiguity among objects and identify them correctly. Interestingly, instead of individual features, if we consider sets of features, like set {*"m," "t"*} or set {*"n," "t"*}, then we find that these sets of features are only present in objects *"amt"* or *"ant."* Consequently, these sets are expected to have more information about these objects than individual features.

We can formally illustrate that feature sets have more information content about objects than their features. To do so, let us first extend the previous formulation of the information content of features to include feature sets as well i.e.

$$\boldsymbol{I}(\mathbf{object};\mathbf{feature\ set}) = \sum\sum \mathbb{P}(\mathbf{object},\mathbf{feature\ set})\ \log\frac{\mathbb{P}(\mathbf{object},\mathbf{feature\ set})}{\mathbb{P}(\mathbf{object})\mathbb{P}(\mathbf{feature\ set})}$$

Also, continuing with the example of four objects, let us assume that the probability of occurrence of each object is as following

$$\mathbb{P}(\text{"am"}) = 0.2, \mathbb{P}(\text{"an"}) = 0.2, \mathbb{P}(\text{"amt"}) = 0.3, \mathbb{P}(\text{"ant"}) = 0.3$$

With this assumption, the marginal occurrence probabilities of features can be calculated as

$$\mathbb{P}(\text{"a"}) = 1, \mathbb{P}(\text{"m"}) = 0.5, \mathbb{P}(\text{"n"}) = 0.5, \mathbb{P}(\text{"t"}) = 0.6$$

Similarly, marginal occurrence probabilities of feature sets can also be calculated

$$\mathbb{P}(\{\text{"m"}, \text{"t"}\}) = 0.3 \quad \text{and} \quad \mathbb{P}(\{\text{"n"}, \text{"t"}\}) = 0.3$$

Now we can calculate the corresponding information contents of features as under

$$\boldsymbol{I}(\text{"amt"}; \text{"a"}) = 0 \qquad \boldsymbol{I}(\text{"ant"}; \text{"a"}) = 0$$

$$\boldsymbol{I}(\text{"amt"};\text{"m"}) = 0.24 \qquad \boldsymbol{I}(\text{"ant"};\text{"n"}) = 0.24$$

$$\boldsymbol{I}(\text{"amt"};\text{"t"}) = 0.23 \qquad \boldsymbol{I}(\text{"ant"};\text{"t"}) = 0.23$$

and information contents of feature sets can be calculated similarly as

$$\boldsymbol{I}(\text{"amt"}; \{\text{"m"}, \text{"t"}\}) = \boldsymbol{I}(\text{"ant"}; \{\text{"n"}, \text{"t"}\}) = -\log\frac{0.3}{1} = 0.88$$

As we can see, the feature set's information content about respective objects is larger than any of its features. Therefore, using sets of features for representing objects can be a way of producing distinct representations.

Information content of sets of features can be further increased by incorporating features' structural arrangement in the set. For real-world objects, this structural arrangement refers to the configuration of features in 3D space. It corresponds to the relative arrangement of letters for the word examples that we have been using. Incorporating the structural arrangement of features into their set introduces an inequality between the sets of identical features. It makes each set more specific. For example, consider two different objects, *"amt"* and *"atm."* The feature set {*"m," "t"*} is present in both the objects, and therefore, it cannot uniquely characterize any of them. However, if we consider the set of features {*"m," "t"*} to be different from the set {*"t," "m"*}, then the first set is present only in the object *"amt"*

whereas the second set is unique to the object *"atm."* In this way, the same features can uniquely characterize different objects due to their distinctive arrangements. With a similar analysis as before, these varying arrangements can be shown to be more informative about the objects.

An important aspect of differential informativeness of varying feature arrangements is that different neurons can get tuned to the same set of features to represent distinct objects. Such tunings ensure that similar objects are distinctly represented and highlight the contrast between informative features-based representations and independent features-based representations. Independent features, by definition, are not associated with one another in any way. The probability of independent features occurring together factors into the marginal probability of occurrence of individual features, i.e., if $x1$, $x2$, $x3$, and $x4$ are four independent features, the probability of them occurring together $\mathbb{P}(x1, x2, x3, x4)$ is given by

$$\mathbb{P}(x1, x2, x3, x4) = \mathbb{P}(x1)\mathbb{P}(x2)\mathbb{P}(x3)\mathbb{P}(x4)$$

Such a relationship implies that any combination of these features is also independent of any other combination. For example, the combination of features $\{x1, x2\}$ is independent of the combination $\{x3, x4\}$ as

$$\mathbb{P}(\{x1, x2\}, \{x3, x4\}) = \mathbb{P}(x1, x2, x3, x4) = \mathbb{P}(x1)\mathbb{P}(x2)\mathbb{P}(x3)\mathbb{P}(x4)$$

$$= \big(\mathbb{P}(x1)\mathbb{P}(x2)\big)\big(\mathbb{P}(x3)\mathbb{P}(x4)\big) = \mathbb{P}(\{x1, x2\})\mathbb{P}(\{x3, x4\})$$

Notably, these relationships hold irrespective of the consideration of their relative arrangements. Thus, using independence as a criterion for selecting features for representation, neither differentiates between different sets of features nor between different configurations of the same features. Informativeness, on the other hand, as illustrated before, does.

2.7. Properties of informative features and their implications

In this chapter, I have demonstrated that informativeness of features or groups of features can be utilized for selecting them as a basis for representing objects. Representations based on informative features are distinct, which is useful in detecting associations among different objects, and such distinction is maintained even when the system suffers some damage or corruption. In this section, I will describe a few unique properties of informative features that further increase their applicability in representing objects for recognition purposes and discuss their implications.

2.7.1 Experience dependence: While discussing the informativeness of features, it is essential to describe the context in which their informativeness about objects is calculated. For example, whenever I have discussed the information content of features, I have always provided a definite set of objects to be considered while calculating the informativeness. Providing a context is critical because of features' information content change with context. Consider the previously described collection of objects, *"am," "an,"* and *"ant."* In the context of these objects, features *"m"* and *"t"* are unique to objects *"am"* and *"ant,"* respectively. However, if we add another object, *"amt,"* to the set, we find that both the unique features lose their uniqueness. Their information content about the objects is reduced. The collection of objects or the context for which the information content of features is calculated corresponds to the system's experience. As the experience of the system changes, the informativeness or the uniqueness of features also changes. For example, to a person who is only experienced in the English language, the specific arrangement of letters *"dan_e,"* where *"_"* denotes a space, corresponds only to the word "*dance."* On the other hand, to a person who knows both English and German, the same arrangement

is ambiguous as it may correspond to *"dance"* or *"danke,"* which is the German for *"thank you."* Thus, with the experience of a new language, the uniqueness of the feature combination *"dan_e"* changes.

A direct implication of this property of informative features is that a system that relies on these features for representing objects needs to update the set of informative features as it gains more experience. In other words, the system needs to be adaptive. As biological systems rely on their experience for obtaining insights about their environment, informative features form a suitable basis for them to represent objects.

2.7.2 Dependence on the occurrence frequency of objects: Another intriguing property of features unique to objects is that the information content of such features about the object is maximum irrespective of the object's occurrence frequency. Consider the previous case of three objects, namely *"a," "an,"* and *"ant,"* for example. Here in object *"ant,"* feature *"a"* is the most common, feature *"n"* is less common than *"a,"* but feature *"t"* is unique to it. Consequently, with all objects being equally likely, the information contents of these features about *"ant"* come out to be as

$$\boldsymbol{I}(\text{"ant";"a"}) = 0, \boldsymbol{I}(\text{"ant";"n"}) = 0.25 \text{ and } \boldsymbol{I}(\text{"ant";"t"}) = 0.92$$

Even if we consider some other distribution of occurrence frequency of objects, the information content of *"t"* will be highest. For example, let us assume that the objects occur with probabilities 0.4, 0.4, and 0.2, respectively. In this situation, the information content of individual features is

$$\boldsymbol{I}(\text{"ant";"a"}) = 0, \boldsymbol{I}(\text{"ant";"n"}) = 0.17 \text{ and } \boldsymbol{I}(\text{"ant";"t"}) = 0.72$$

Note that in this situation, though the information contents of *"n"* and *"t"* about the object are different from the previous case, the information content of *"t"* is still the highest. It is essential to realize that this property of unique features' information content arises because of how the information content is defined. Recall

that the information content of a feature about an object, $\boldsymbol{I}(\mathbf{object}; \mathbf{feature})$, is calculated as

$$\boldsymbol{I}(\mathbf{object}; \mathbf{feature}) = \sum\sum \mathbb{P}(\mathbf{object}, \mathbf{feature}) \log \frac{\mathbb{P}(\mathbf{object}, \mathbf{feature})}{\mathbb{P}(\mathbf{object})\mathbb{P}(\mathbf{feature})}$$

The above equation can be re-written as

$$\boldsymbol{I}(\mathbf{object}; \mathbf{feature}) = \sum\sum \mathbb{P}(\mathbf{object}, \mathbf{feature}) \log \frac{\mathbb{P}(\mathbf{object}|\mathbf{feature})\mathbb{P}(\mathbf{feature})}{\mathbb{P}(\mathbf{object})\mathbb{P}(\mathbf{feature})}$$

$$= \sum\sum \mathbb{P}(\mathbf{object}, \mathbf{feature}) \log \frac{\mathbb{P}(\mathbf{object}|\mathbf{feature})}{\mathbb{P}(\mathbf{object})}$$

as $\mathbb{P}(\mathbf{object}|\mathbf{feature}) \leq 1$, we find that the information content is bounded above i.e.

$$\boldsymbol{I}(\mathbf{object}; \mathbf{feature}) \leq \sum\sum \mathbb{P}(\mathbf{object}, \mathbf{feature}) \log \frac{1}{\mathbb{P}(\mathbf{object})}$$

The maximum value of the information content is achieved when

$$\mathbb{P}(\mathbf{object}, \mathbf{feature}) = \mathbb{P}(\mathbf{object}) = \mathbb{P}(\mathbf{feature})$$

which holds only when the feature is unique to the object. Thus, for any occurrence probability of the object, the information content of its unique features is always maximum.

This property of unique features implies that while representing objects based on them, the system does not need to care about the probability distribution of objects. However, knowledge of occurrence probabilities is required while representing objects based on features that are not unique but still have high information content. Later in this chapter, I will demonstrate that complete knowledge or accurate estimate of the distribution is not required even in these situations.

2.7.3 Uniqueness: Another critical property of the most informative features is that they are not singular. Multiple features or feature combinations from the same object can

have equal information content about it. For example, consider the previous example of three objects *"a," "an,"* and *"ant."* We saw that feature *"t"* is most informative about object *"ant,"* followed by features *"n"* and *"a."* Now, if we consider feature combinations *"nt"* or *"at"* and calculate their information content, we find that

$$\boldsymbol{I}(\text{"ant"; "t"}) = \boldsymbol{I}(\text{"ant";"nt"}) = \boldsymbol{I}(\text{"ant";"at"}) = 0.92$$

and similarly, for feature combination *"an"*, we find the

$$\boldsymbol{I}(\text{"ant"; "n"}) = \boldsymbol{I}(\text{"ant";"an"}) = 0.25$$

This is because both feature combinations *"nt"* or *"at"* are unique to the object *"ant"* but feature combination "an" is shared among objects *"ant"* and *"an"*.

Indeed, a system representing objects on the basis of informative features cannot accommodate all the equivalent feature combinations for all objects and has to choose a particular combination over the other. In the next section, I will suggest that choosing the more complex combination of features is preferable in such a situation and will provide a rationale for the suggestion (complex features make it distinct for different objects).

Interestingly, multiple feature combinations may correspond to different views in the case of a three-dimensional object. For these objects, multiple features or feature groups with equivalent information content can correspond to various equally informative views. As discussed in the previous chapter, theories proposing view-based representation of objects suggest that a small number of views may be sufficient to generate invariant representations of 3D objects. Following the same argument from these theories, equally informative views can be utilized to achieve the required invariance. Thus, informativeness of features or feature groups about individual objects can also serve as a criterion for selecting views that lead to an invariant representation of objects.

2.8. Effect of statistics of objects on the informativeness of features

Previously in this chapter, I have highlighted that independence among features does not necessarily correlate with their information content. Using an example of four objects, namely *"am," "an," "amt,"* and *"ant,"* I have shown that though feature sets {*"a," "m," "t"*} and {*"a," "n," "t"*} comprise of independent features, the information content of these features are neither equivalent nor high. Feature *"a"* does not have any information about any object, and its information content is least. Features *"m," "n,"* and *"t"* have higher information than *"a"* about individual objects, and the amount of information depends on the probabilities of occurrence of the objects. Furthermore, with the example of three objects, *"a," "an,"* and *"ant,"* I demonstrated that features that are unique to the object contain maximum information about it irrespective of the occurrence frequency of the object. This result is consistent with the definition of the information content that I have used in this chapter. However, I have not demonstrated how objects' frequency of occurrence affects the information content of individual features. In this section, using the same example as described before, I will highlight that occurrence frequencies of objects play an important role in determining their features' information contents.

Before understanding how the information content of features is affected by the occurrence frequencies of objects, it is essential to realize that dependence among features is also decided by the distribution of objects. Consider the previous example of four objects, namely *"am," "an," "amt,"* and *"ant."* It was assumed that these objects' occurrence probabilities are of the form x, x, $0.5 - x$, and $0.5 - x$ respectively, where x is a positive number less than 0.5. This distribution rendered features sets {*"a," "m," "t"*} and {*"a," "n," "t"*} independent, i.e., it could be shown that

$$\mathbb{P}(\text{"am"}) = \mathbb{P}(\text{"a"})\mathbb{P}(\text{"m"}), \mathbb{P}(\text{"an"}) = \mathbb{P}(\text{"a"})\mathbb{P}(\text{"n"}), \mathbb{P}(\text{"at"}) = \mathbb{P}(\text{"a"})\mathbb{P}(\text{"t"})$$

$$\mathbb{P}(\text{"mt"}) = \mathbb{P}(\text{"m"})\mathbb{P}(\text{"t"}) \text{ and } \mathbb{P}(\text{"nt"}) = \mathbb{P}(\text{"n"})\mathbb{P}(\text{"t"})$$

However, if we consider the distribution to be of a different form, i.e., consider the occurrence probabilities of objects to be 0.1, 0.3, 0.2, and 0.4, respectively for example, then we find that, though

$$\mathbb{P}(\text{"am"}) = \mathbb{P}(\text{"a"})\mathbb{P}(\text{"m"}), \mathbb{P}(\text{"an"}) = \mathbb{P}(\text{"a"})\mathbb{P}(\text{"n"}), \mathbb{P}(\text{"at"}) = \mathbb{P}(\text{"a"})\mathbb{P}(\text{"t"})$$

But

$$\mathbb{P}(\text{"mt"}) \neq \mathbb{P}(\text{"m"})\mathbb{P}(\text{"t"}) \text{ and } \mathbb{P}(\text{"nt"}) \neq \mathbb{P}(\text{"n"})\mathbb{P}(\text{"t"})$$

This relationship indicates that the independence of features depends on the frequency of objects. Therefore, to find truly independent components, the accurate probability distributions of the objects must be known. A partial or wrong estimate of the probability distributions may lead to an erroneous assessment of independence among features.

Now, let us consider the features' informativeness and calculate it for the two different distributions of objects discussed above. Let us first calculate the information content of features about objects *"amt"* and *"ant"* when the objects follow the distribution of form x, x, $0.5 - x$, and $0.5 - x$. For demonstration purposes, set the value of x to be 0.2 so that frequencies of objects are 0.2, 0.2, and 0.3. 0.3. With this distribution, we find that

$$\boldsymbol{I}(\text{"amt"}; \text{"a"}) = 0 \qquad \boldsymbol{I}(\text{"ant"}; \text{"a"}) = 0$$

$$\boldsymbol{I}(\text{"amt"};\text{"m"}) = 0.24 \qquad \boldsymbol{I}(\text{"ant"};\text{"n"}) = 0.24$$

$$\boldsymbol{I}(\text{"amt"};\text{"t"}) = 0.23 \qquad \boldsymbol{I}(\text{"ant"};\text{"t"}) = 0.23$$

As before, these calculations demonstrate that feature *"a"* contains no information about any of the objects. Feature *"t"* contains more information than *"a."* Still, it is lesser than that of features *"m"* or *"n,"* which are unique features of these objects. Again, if we consider the other distribution of occurrence probabilities of objects, i.e., 0.1, 0.3, 0.2, and 0.4, respectively, and calculate the information content of the features of the same set of objects, we find that

$$\boldsymbol{I}(\text{"amt"};\text{"a"}) = 0 \qquad \boldsymbol{I}(\text{"ant"};\text{"a"}) = 0$$

$$\boldsymbol{I}(\text{"amt"};\text{"m"}) = 0.25 \qquad \boldsymbol{I}(\text{"ant"};\text{"n"}) = 0.28$$

$$\boldsymbol{I}(\text{"amt"};\text{"t"}) = 0.17 \qquad \boldsymbol{I}(\text{"ant"};\text{"t"}) = 0.42$$

Here again, we find that feature *"a"* contains the least information about individual objects, and features *"m," "n,"* and *"t"* have more information than that. Interestingly, comparing information contents of features *"m," "n,"* and *"t,"* we find that, as in the previous case, for the object *"amt,"* information content of feature *"t"* is less than feature *"m."* However, for the object *"ant,"* the information content of feature *"n"* is now less than that of the feature *"t."* These values are opposite to the previous case. This reversal in information contents of features *"n"* and *"t"* about object *"ant"* is the result of the change in the distribution of objects because nothing else has changed in the two situations. However, it is not very clear which aspect of change in distribution is causing such reversal.

To understand how changes in the relative occurrence of objects affect the information contents of features, we need to look closely into what has changed about features with the change in the distribution of objects. If we calculate the marginal probabilities of features in the two situations, we find that occurrence probabilities of features when the probability distribution of objects is 0.2, 0.2, 0.3, and 0.3 is given as

$$\mathbb{P}(\text{a}) = 1, \mathbb{P}(\text{m}) = 0.5, \mathbb{P}(\text{n}) = 0.5, and\ \mathbb{P}(\text{t}) = 0.6$$

Whereas, in other situation when the probabilities of occurrence of objects are assumed to be 0.1, 0.3, 0.2, and 0.4, the marginal probabilities of occurrence of features are

$$\mathbb{P}(\text{a}) = 1, \mathbb{P}(\text{m}) = 0.3, \mathbb{P}(\text{n}) = 0.7, and\ \mathbb{P}(\text{t}) = 0.6$$

As we observe, with the change in occurrence probabilities of the objects, the probabilities of occurrence of two features, *"m"* and *"n,"* have changed, and the probability of occurrence of feature *"t"* has remained unchanged. However, what has not changed for feature *"m"* and has changed for the feature *"n"* is its probability of occurrence relative to

feature *"t."* In the first situation, both features *"m"* and *"n"* are equally likely to occur, and their probabilities of occurrence are 0.5 each. In this situation, feature *"t"* with marginal probability 0.6 is more likely to occur than any of them. On the other hand, in the second situation, while feature *"t"* (marginal probability 0.6) is still more likely to occur than feature *"m"* (marginal probability 0.3), it is less likely to occur than feature *"n"* (marginal probability 0.7). The marginal occurrence probabilities of features can be regarded as their relative abundance in the environment. Thus, these observations indicate that informativeness of features about an object changes with their relative abundance. More abundant features convey less information about any object, whereas rare or sparsely occurring features are more informative. This finding makes intuitive sense as well because rare events have more information content than common ones.

Two important points must be noted here. First, as dependence among features does not affect their information content, such change in the features' informativeness is not affected by their dependence either. For example, if we set the value of x to be 0.4 in the first situation so that the probabilities of occurrence of objects become 0.4, 0.4, 0.1, and 0.1, respectively, then information content of features for the objects come out to be

$$\boldsymbol{I}(\text{"amt"};\text{"a"}) = 0 \qquad \boldsymbol{I}(\text{"ant"};\text{"a"}) = 0$$

$$\boldsymbol{I}(\text{"amt"};\text{"m"}) = 0.11 \qquad \boldsymbol{I}(\text{"ant"};\text{"n"}) = 0.11$$

$$\boldsymbol{I}(\text{"amt"};\text{"t"}) = 0.27 \qquad \boldsymbol{I}(\text{"ant"};\text{"t"}) = 0.27$$

Again, in this particular situation, as the relative abundance of feature *"t"* is less than both features *"m"* and *"n,"* we find that it is more informative about both *"amt"* and *"ant."* However, as in the first case, features *"m"* and *"n"* are independent of the feature *"t."*

The second and probably more crucial point is that such changes in information content are observed when the features under consideration are not unique to individual objects. If a feature uniquely characterizes an object, its information content is maximum, irrespective of the object's frequency of occurrence. As described previously, this follows

from the positive correlation between the feature's information content about an object and the conditional occurrence probability of the object conditioned on the feature. For features that are unique to the object, this conditional probability is 1, which is its maximum possible value, and hence, such features have maximum information content about the object. If these unique features are utilized to represent the object, then only one active neuron that indicates the presence of the unique feature will form its representation. However, finding such unique features for all objects is not possible, and a common scenario is that features are shared among objects. The above observation of changes in the information content then implies that in such a situation abundance of features can be used as a measure of their information content for any object. Therefore, sparsely occurring features can be selected as a basis for representing objects.

This realization brings out a critical difference between representing objects based on independent features and informative features. We see that, though the distribution of objects affects both independence of features and their information content about objects, assessing their independence requires knowledge of the absolute occurrence frequencies of objects. In contrast, their information content can be guessed based on their relative abundance. For a finite set of objects, a particular set of occurrence frequencies can render the constituting features independent, while other frequency distributions can develop dependence among them. The two situations can be distinguished only by identifying the distributions of objects and using them to calculate the marginal and joint probabilities of features. However, the information content of features about individual objects changes with the relative abundance of features in the environment. One does not need to know anything about the objects from which the features come from, and as long as certain features are less common in experience, they can be guessed to have higher information content. This process eliminates the necessity to estimate the exact distribution of objects, which is an extremely

challenging task for a biological system that gathers information about its environment through experience.

2.9. An adaptive framework for representing objects

The usefulness of a higher-order representation of an object is determined based on two criteria. First, it should allow invariable recognition of the object, and second, it should permit the detection of association among multiple objects. Invariable recognition of objects requires their representation to remain stable in varying conditions yet be sensitive enough to distinguish similar objects. Detection of associations among objects, on the other hand, necessitates that representations can be utilized in assessing their marginal and joint occurrence probabilities so that any *"regularity"* between objects can be detected. As discussed in the previous chapter, various theories that have been proposed so far have been successful in satisfying one of the two criteria. Still, none can explain how both the requirements can be simultaneously fulfilled.

I realized that informative features or sets of features, owing to their properties discussed above, can form a suitable basis for representing objects. As we have seen, object representations based on the most informative, unique features can be maximally distinct. Such distinction is desirable because it highlights differences between objects, which is required for distinguishing them. Non-overlapping representations also permit associating separate identities to the representations of very similar objects. Furthermore, the uniqueness of represented features or feature groups makes individual neurons maximally informative about individual objects. As information is not shared among multiple neurons, the entire representation's informativeness is maintained even in conditions when the system gets corrupted or suffers a neuronal loss. Thus, informative features impart both desired qualities, namely stability and sensitivity, to the representations. Likewise, more distinct or less overlapping representations can also be utilized to estimate individual objects' occurrence

frequencies. As the activities of individual neurons indicate the presence of specific objects, the average activation frequency of neurons correlates directly with the occurrence probability of individual objects. Furthermore, joint occurrence probabilities of objects can be readily assessed by pooling information from multiple neurons. In this way, representations based on informative features simultaneously satisfy the two criteria on which the usefulness of a higher-order representation of objects is judged.

Accordingly, I propose that sensory processing should aim to identify the most informative components from the environment and use them for representing objects. In particular, the system should represent objects so that individual neurons convey maximum information about individual objects, i.e., the mutual information between individual neurons and specific objects are maximized. Doing so will tune each neuron to respond to the most informative components of the objects, which may consist of individual features or groups of features. Though such a procedure will not ensure neurons' independence, it will allow them to get tuned to informative features of objects in the surroundings. Thus, enabling the system to adapt to its environment in an experience-dependent manner. Even without independent neurons, estimation of occurrence frequencies of objects based on average activity will be possible, and the detection of dependencies between them will be facilitated.

It is important to realize that several aspects of this framework of representing objects are significantly different from the classical efficient coding framework. Firstly, Informative components are not independent, and conversely, independent features are not always informative. Consequently, the features utilized in the two frameworks are fundamentally different. Secondly, though the occurrence frequency of objects affects both informativeness and independence of features, relative informativeness of features can be judged based on their relative abundance. Assessing independence, on the other hand, requires knowledge of the occurrence frequencies of the object. Lastly, as feature groups can be more informative than individual features, using informativeness as a criterion allows using individual features

and feature groups as a basis of representation. This equivalence eliminates the need for a separate scheme for hierarchically assembling feature groups. On the contrary, groups of independent features are also independent, and therefore, independence as a criterion is not sufficient to distinguish features from feature groups. As a result, a separate scheme of assembling feature groups is necessary while using independent features.

2.10. Efficiency of representation framework based on informative features

In the previous section, I proposed that representing objects at higher levels of visual processing should be based on informative features. These features allow the system to form highly distinct representations of the objects associated with their identity. The representations can tolerate corruptions in the system and enable the system to detect associations among objects, thus satisfying the two criteria used to judge the usefulness of higher-order object representations. In this section, I will show that representation based on informative features will also be efficient in communicating information about the objects. In particular, maximizing the efficiency in communicating information for a finite number of objects, I will show that the representations of objects should be maximally distinct to achieve maximum efficiency. As such representations arise when they are based on the unique, most informative features, the exercise essentially demonstrates that these features form a basis for efficient representation

Let us consider a system of K binary neurons, where each neuron can be in only one of the two possible states, namely, an active (1) or a non-active (0) state. The system is supposed to represent sensory inputs as patterns of neuronal activity. We further assume that any representation will consist of at least one active neuron. This assumption is necessary because, for any biological system, inactivity does not mean a response. In other words,

there exists no situation in which an input can elicit no response in the system and is still detected by it. To be detected, the stimuli properties must interact, either directly or indirectly, with receptor proteins present in the neurons and cause activity in them. This limitation is an important distinction of biological systems from a mechanistic communication system where the absence of any signal can also be considered a representation. Keeping this distinction in mind, if a representation of any input comprises of r active neurons, then

$$1 \leq r \leq K$$

As there can exist several different patterns in which r units are active, and each of those patterns can appear several times, we can consider the number of active neurons at any instance as a measure of the level of activity. If we assume that there are n_r instances of the same activation level, i.e., the same number of neurons (r) constituting either the same activity pattern or different patterns of activity are active at n_r instances, then the total activity in K neurons will be rn_r. Moreover, if there are m different levels of activations, the total observed activity a_{tot} is given by

$$a_{\text{tot}} = \sum_{r=1}^{m} rn_r \tag{1}$$

The probability p_i of a neuron i to be active can then be determined from the number of instances a_i when the neuron responds and the total number of instances of activation N i.e.

$$p_i = \frac{a_i}{N}$$

The average probability $\langle p \rangle$ of a unit being active can be expressed as

$$\langle p \rangle = \frac{1}{K}\sum_{i=1}^{K} p_i = \frac{1}{K}\sum_{i=1}^{K} \frac{a_i}{N} = \frac{1}{KN}\sum_{i=1}^{K} a_i = \frac{a_{\text{tot}}}{KN}$$

The last equality follows from the fact that the sum of all instances of activation in all neurons is the same as the sum of activities of all neurons in all instances of activation. This gives us

$$\langle p \rangle = \frac{1}{K}\left(\frac{\sum_{r=1}^{m} rn_r}{\sum_{r=1}^{m} n_r}\right) \text{where} \sum_{r=1}^{m} n_r = N \tag{2}$$

and as $1 \leq r \leq m$

$$\sum_{r=1}^{m} n_r \leq \sum_{r=1}^{m} rn_r < \sum_{r=1}^{m} mn_r$$

From the above set of inequalities, we can obtain bounds on the average probability of a neuron being active, i.e.

$$\frac{1}{K}\left(\frac{\sum_{r=1}^{m} n_r}{\sum_{r=1}^{m} n_r}\right) \leq \langle p \rangle < \frac{1}{K}\left(\frac{\sum_{r=1}^{m} mn_r}{\sum_{r=1}^{m} n_r}\right)$$

$$or, \frac{1}{K} \leq \langle p \rangle < \frac{m}{K} \tag{3}$$

In his seminal work (Barlow 1961), Barlow considered a similar encoding system of F nerve fibers and, using the similar measure of average activation probability of nerve fibers, calculated the bit entropy of the nerve fibers as

$$C = FH_2(\langle p \rangle)$$

where

$$H_2(x) = -x\log x - (1-x)\log(1-x)$$

Using similar arguments, we can define the bit entropy of our system as

$$C_R = KH_2(\langle p \rangle)$$

and following inequalities in relation (3), we get

$$KH_2\left(\frac{1}{K}\right) \leq C_R < KH_2\left(\frac{m}{K}\right)$$

$$or, \quad -K\left(\frac{1}{K}\log\frac{1}{K}+\left(1-\frac{1}{K}\right)\log\left(1-\frac{1}{K}\right)\right) \leq C_R$$

$$< -K\left(\frac{m}{K}\log\frac{m}{K}+\left(1-\frac{m}{K}\right)\log\left(1-\frac{m}{K}\right)\right)$$

$$or, \quad \log K - K\left(1-\frac{1}{K}\right)\log\left(1-\frac{1}{K}\right) \leq C_R < m\log\frac{K}{m} - K\left(1-\frac{m}{K}\right)\log\left(1-\frac{m}{K}\right)$$

$$or, \quad \log K + \left(1-\frac{1}{K}\right)K\log\left(\frac{K}{K-1}\right) \leq C_R$$

$$< m\log\frac{K}{m} + \left(1-\frac{}{K}\right)K\log\left(\frac{K}{K-m}\right)$$

If we consider K to be very large compared to 1, then

$$\lim_{K\to\infty} K\log\left(\frac{K}{K-1}\right) = \lim_{K\to\infty}\frac{\log\left(\frac{K}{K-1}\right)}{\frac{1}{K}} = \lim_{K\to\infty}\frac{K^2}{K(K-1)} = 1 \quad [\text{byL}'\text{Hospital}'\text{srule}]$$

And, if we consider K to be very large compared to m too, then

$$\lim_{K\to\infty} K\log\left(\frac{K}{K-m}\right) = \lim_{K\to\infty}\frac{\log\left(\frac{K}{K-m}\right)}{\frac{1}{K}} = \lim_{K\to\infty}\frac{K^2 m}{K(K-m)} = m \quad [\text{byL}'\text{Hospital}'\text{srule}]$$

which means that for large K and $m \ll K$

$$\log K + \left(1-\frac{1}{K}\right)K\log\left(\frac{K}{K-1}\right) \approx \log K$$

and

$$m\log\frac{K}{m} + \left(1-\frac{m}{K}\right)K\log\left(\frac{K}{K-m}\right) \approx m\log\frac{K}{m}$$

hence

$$\log K \leq C_R < m\log\frac{K}{m} \tag{4}$$

The last set of inequalities gives bounds on the bit entropy of our representation system. It shows that, with our considerations, the minimum bit entropy for a system with K neurons is $\log K$, which is independent of the total instances of activation N and is achieved when in no instance more than one neuron is active. Furthermore, when the maximum

number of neurons active at any instance is minimal compared to the total number of neurons, then the bit entropy also has a maximum limit.

Examining the derivation of the bounds, one can realize that such bounds, specifically the lower bound, exist because we have assumed that any representation must contain at least one active neuron. This assumption constrained the average probability of a neuron being active to be at least $1/K$, which made the lower bound on the bit entropy to be $\log K$. Without this assumption, the lower bound would have been zero because the minimum value of the average probability of activity would have been zero.

A system's efficiency in representing inputs is measured by comparing its bit entropy with the entropy of the set of inputs. Formally, if H_{inp} is the entropy of the inputs, and C_R is the bit entropy of the representation system, then the efficiency of the system E is expressed as

$$E = \frac{H_{inp}}{C_R}$$

It is generally assumed that due presence of millions of neurons, and their multiple states, the system can always represent all the input, i.e., $C_R > H_{inp}$. Efficiency is said to be achieved when the capacity, which is equivalent to bit entropy, matches the entropy of the inputs. This assumption holds even in cases like ours, where only a finite set of inputs are considered. Consequently, efficiency is achieved in our case when C_R is reduced to match H_{inp}, which corresponds to reducing the average probability of activation of neurons. However, because there is a lower bound on the level to which C_R can be reduced, more than one ways exist to match it with H_{inp} and achieve efficiency. Three scenarios must be considered to highlight different ways in which efficiency can be attained

1. **H_{inp} is greater than $\log K$:** This scenario corresponds to the general assumption that the system has the capacity to represent all inputs. As H_{inp} is greater than the system's minimum possible capacity to represent inputs, the capacity can always be adjusted to

match it. The required adjustment is made by ensuring that the average probability of activation of neurons is at appropriate levels and is not too high. These adjustments translate into tuning neurons to rare, less abundant components of the environment and utilizing fewer neurons to represent any input.

2. **H_{inp} is equal to log$K$:** This scenario arises in a particular situation when the entropy of inputs exactly matches the minimum capacity of the system to represent inputs. Efficiency can only be attained in this situation by reducing the average probability of a neuron being active to $1/K$. As at least one neuron must be active to indicate an input's presence, such average activation probability implies that more than 1 neuron is active in no instance. This condition, in turn, means that a distinct neuron represents each input. In other words, the representations of objects need to be maximally distinct to achieve efficiency in this scenario.

3. **H_{inp} is less than log$K$:** This scenario, though unlikely, emerges when only a small number of inputs have a significantly higher probability of occurrence. In these situations, H_{inp} is very low compared to the system's capacity, and the only way to achieve efficiency is by reducing the number of neurons. Reducing the number of neurons from K to K', reduces the system's minimum capacity $\log K$ to $\log K'$, which can be matched to the low value of H_{inp}. It is important to note that reducing the number of neurons transforms this situation into one of the previously described scenarios, and efficiency can be achieved either by tuning individual neurons to rare features of the inputs or by making the representations maximally distinct.

Thus, we see that achieving efficiency in representing a finite number of inputs corresponds to ensuring that neurons are tuned to rare, less abundant components of the environment and by making the representation of individual objects maximally distinct. In

the previous sections of this chapter, we have seen that both these procedures correspond to representing objects in terms of more informative features. Less abundant features are more informative than common features, and representations based on unique features of objects are maximally distinct. Therefore, this demonstrates that representations based on unique features can be efficient.

Two points must be noted here. First, in the derivation of the system's entropy, we have considered only a finite number of patterns (N) that are experienced over a limited time. This consideration highlights that for a biological system, it is not possible to know all the environment's statistical properties. It relies on its experience to obtain knowledge about its environment, and at any given instance, has insights only of its immediate surroundings. As the system gains more knowledge with experience, or as the environment changes with time, it needs to update its view of the environment and adapt to it.

The system having limited experience with the environment implies that its efficiency should not be judged based on the environment's actual statistics. One cannot expect the system to efficiently represent the aspects of its surroundings that it has never experienced. This realization brings out the second noteworthy point about the above derivation. Here, the entropy of the inputs considered is not the actual entropy but the observed entropy of the objects in the surroundings. In other words, if there are $\mathcal{N}$ distinct objects in the environment at a given instance, with the actual frequency of occurrence of object j being f_j, and the observed frequency of occurrence of the same object being g_j, then the entropy of the inputs, H_{inp}, that is considered while determining efficiency, should be expressed as

$$H_{inp} = \sum_{j=1}^{\mathcal{N}} -f_j \log g_j$$

whereas the actual entropy of the objects based on the exact occurrence probabilities has the form

$$\widetilde{H_{inp}} = \sum_{j=1}^{\mathcal{N}} - f_j \log f_j$$

Note that the difference in two entropies, $H_{inp} - \widetilde{H_{inp}}$, can be expressed as

$$H_{inp} - \widetilde{H_{inp}} = \sum_{j=1}^{\mathcal{N}} f_j \log \frac{f_j}{g_j}$$

which is a measure of the divergence between the actual and observed distributions of the objects and is known as KL divergence (Kullback and Leibler 1951). It can be shown that KL divergence between any two distributions is always positive, meaning that H_{inp} is always greater than $\widetilde{H_{inp}}$ or in other words, the observed entropy of the inputs will always be higher than their actual entropy. Given that the bit entropy of the system has a lower bound, a higher observed entropy will benefit the system as it will bring the input entropies closer to this lower bound. This effect will allow the system to represent very skewed distributed objects efficiently. Such distributions arise when the system has a biased experience of a few objects, which is a common scenario. Furthermore, as described in the previous sections, it is impossible to obtain truly independent components of the environment with such experience of the environment. Thus, with the above derivation, we can conclude that using informative features allows the system to be efficient and maintain this efficiency while being adaptive to its environment.

2.11. Object representation using informative features

In the previous sections, I have proposed that informative features can be utilized as a basis for object representation. The representations based on unique, informative features are distinct, robust, and efficient. Moreover, these features are suited for representing objects by a biological system. They do not require complete knowledge of the environment's statistical properties and can be identified by the system in an experience-dependent manner.

However, a question that remains to be answered is how the system can extract these features. In other words, given a set of objects, how a system identifies which features or feature groups are more informative. The theoretical calculations performed previously in this chapter cannot be performed by a biological system. First, it presumably has only partial knowledge of the object distributions. Second, the required computations are too complicated to be carried out in a neuronal circuit. In this chapter, I will first introduce the problem of selecting the most informative features as a problem of basis transformation. Then I will formulate an optimization problem that can be solved for obtaining the transformation. The biological plausibility of this method will be discussed in later chapters.

An organism must obtain knowledge about its environment, which is manifested in the form of associations and dependencies observed in the surroundings. The information about the surroundings is made available to the organism through its senses. However, the format in which the information is presented to the system is not always suitable for identifying the necessary associations (Marr 1982). A natural step to overcome this limitation is reorganizing the sensory information in a format that brings out these associations. In other words, the sensory information must be represented in an appropriate form that makes identifying the associations in the surroundings easier.

Any representation process is essentially a transformation of how information is conveyed. Consider the English language, for example. Representing English words into Morse code implies transforming English letters into some series of dots and dashes. The symbols that we recognize as letters are converted to specific combinations of dots and dashes, which are further combined into words and sentences. One can imagine the English language's entire text as data points that was spread in a space defined by the letter symbols. Representing these words in Morse codes implies transforming the space of letters and characters into a space defined by the combination of dots and dashes (**Figure 2.4**) so that each data point can now be described in terms of combinations of these elements. In this

way, a representation process transforms the basis in which data is represented. It acts as a function that maps one basis set to another.

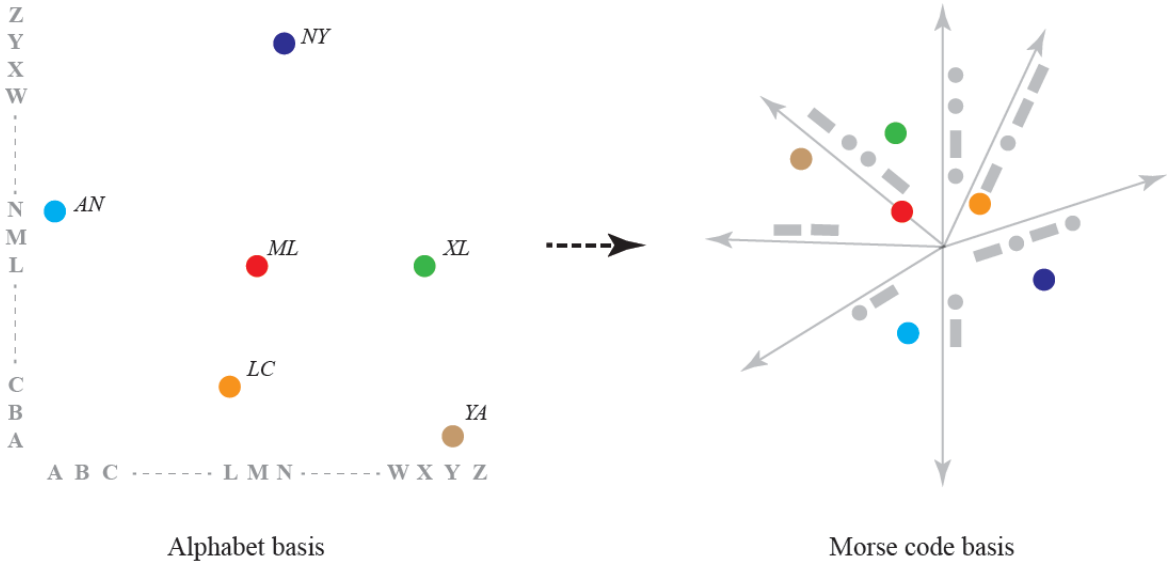


Figure 2. 4: Illustration of coding as basis transformation

Combinations of English alphabets can be represented in a basis defined in symbols corresponding to alphabets. When the alphabets are coded into Morse code, each alphabet is described by a combination of dots and dashes. These combinations constitute the new basis components, and the letter combinations can now be described in this basis. Note that the arrangement of data points changes when the basis is changed.

In this regard, one can also view sensory processing as a basis transformation where the basis set comprises the physical properties of stimuli like frequency, intensity, and chemical structure is transformed into a basis set of neuronal activities. To understand this, imagine the details involved in seeing something. When we see any object, light reflected from the object enters our eyes and is focused on our retina through the eye lens. The cells

in the retina contain photoreceptors, the proteins that change their conformation when they encounter light. When light focused onto the retina interacts with these photoreceptors, a change in their conformation induces electrical activity in retinal cells, which is further relayed to other parts of our brains through neuronal circuits originating in the retina (**Figure 2.5.1**). As the object's properties, like its shape, size, color, and surface, all contribute to determining its reflectance properties, the light being reflected must contain information about all these properties. One can imagine the object as a point in the space defined by properties that determine how light is reflected from any object in the environment. When the reflected light activates the retinal cells, the energy of light is transformed into the electrical activities of the cells, and hence, the information about the object that was present in light would now be present in the neuronal activities. Different objects will reflect light in different ways, and light reflected in different ways will elicit different responses in the neurons. Therefore, the object can correspond to a pattern of neuronal activity, which can be described in the basis defined by individual neurons. As this activity is further relayed along the sensory pathway, the objects' information is represented as activities of different sets of neurons. New basis sets defined by these new sets of neurons emerge (**Figure 2.5.2**). Some of these bases will highlight some aspect of information about the object, and some others will highlight some other aspects; thus, effectively reorganizing the information.

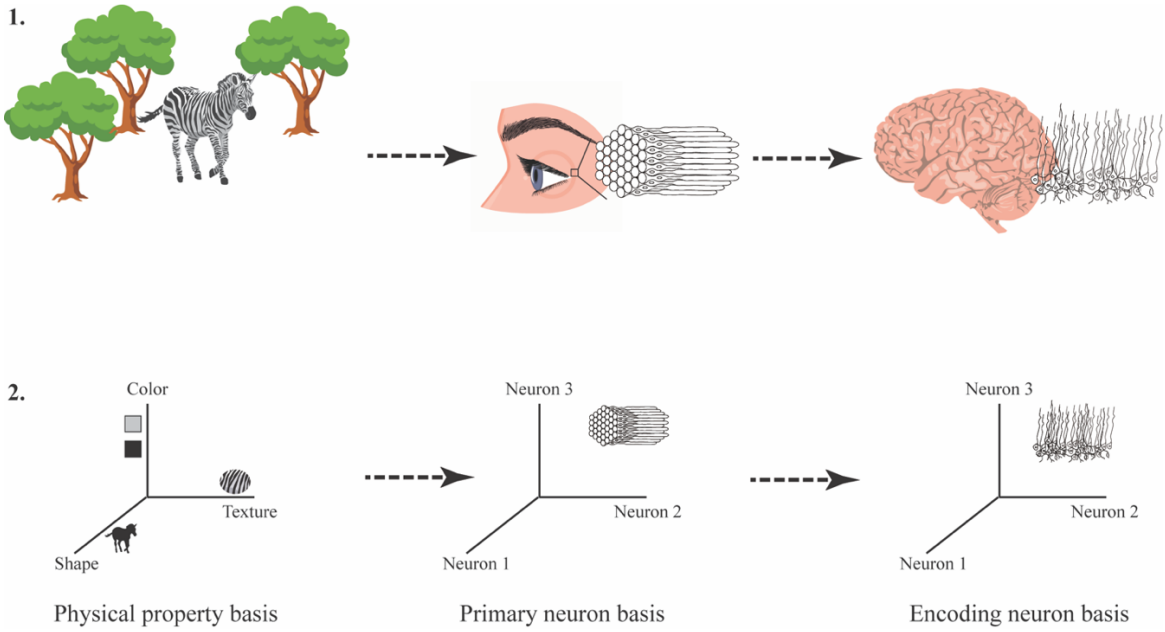


Figure 2. 5: Sensory processing as a basis transformation

1. *Light from an object carrying information about it enters our eyes and activates a set of photosensitive neurons in the retina. Their activity is relayed further along the sensory pathway to activate cells in higher cortical regions.* **2.** *This entire process can be seen as a set of successive basis transformations. The object can be described in a basis set defined by its physical properties, which is transformed into a basis set defined by photosensitive neurons in the retina, which is further transformed into a basis set defined by higher cortical neurons, referred to here as encoding neurons. Different aspects of information about the object may be highlighted in a different basis.*

Considered accordingly, representing objects based on the most informative components from the environment essentially requires transforming the representation basis from individual pixels in the retina to informative features or sets of features at higher levels of visual processing. However, obtaining such a transformation is not an easy task because there can be infinitely many ways of transforming pixels into structural elements. Not all

such transformations will lead to a basis of informative structures. An algorithmic implementation of the transformation requires constraints that improve the chances of obtaining the basis of the most informative structures. These constraints are generally designed to narrow down the nature of representations to those that one expects in the desired basis, which forces the basis to have characteristics necessary for producing the required nature of representation.

In the previous section, we have established that while representing objects based on informative features, only a few neurons tuned to the unique features from the object should be active. Moreover, as these unique elements should be rarely observed in the environment, the activation probability of a neuron tuned to such features must be small. Consequently, a representation based on informative features should be sparse and non-overlapping. Thus, the first constraint that the representations based on informative structures must satisfy is sparseness and minimal overlap. However, it is important to realize that there is no guarantee to have completely non-overlapping sparse representations, particularly when the representation dimensions, which corresponds to the number of neurons utilized in representing inputs, are small compared to the number of inputs. In these situations, a single neuron should participate in representing multiple objects, and therefore, must be tuned to features that are unique to this set of objects.

Another aspect that must be considered while utilizing unique, informative structures for representing objects is that these structures may comprise only a few features or a minimal set of features from an object. In such cases, a possibility arises that only these few structural elements are utilized in representation. A system trying to base its representation only on informative structures can completely ignore non-unique structures from the objects. However, while the representation must remain sparse when based on these unique, informative structures, it is equally important that the representations are not based on partial structures. Basing representations only on partial structures may lead to ambiguity when these partial structures are invisible due to partial occlusion. Therefore, the second constraint

that the representation based on informative structures must satisfy is that it should account for as much input as possible. In other words, the representations must preserve the structures of the input.

Formally, the set of all unique structures that comprise the basis of representation can be expressed in a matrix form. If there are K representation neurons, K different structures corresponding to these neurons' tuning can be organized into a matrix Φ of dimensions $M \times K$, where M is the number of neurons in the retina and correspond to the maximum number of pixels that can be present in a structure. Note that any informative structure will be composed of only a subset of these neurons which become active when that particular structure is encountered; the other neurons remain inactive in the presence of this structure. We call this matrix of unique structures the *dictionary of structures* or simply the *dictionary*.

Furthermore, as M is the number of neurons in the retina, encountering any object will elicit a response in these cells, and a pattern of activity will emerge corresponding to the object. Such activity pattern can be expressed as an M-dimensional vector, which can be used to denote an input to the system. Similarly, the representation of any object corresponds to the activity pattern in K representation neurons, and hence, can be expressed as a K-dimensional vector of activity.

With such definition of input, its representation and dictionary, if s is an M-dimensional input to the system, and a is its K-dimensional representation based on dictionary Φ, then preserving structure implies that these quantities satisfy the linear relationship expressed as

$$s = \Phi a$$

The linear form of relation arises because we have considered that representations are combinatorial, meaning that if multiple structures present in the input are also captured in the dictionary, then while representing the input, all neurons tuned to its structures will

get activated. The relationship can be extended to include a finite set of N objects by replacing the input and representation vectors corresponding to a single object with matrices of inputs and representations. In particular, if $S \in \mathbb{R}^{M \times N}$ is a matrix of N input patterns and $A \in \mathbb{R}^{K \times N}$ is a matrix of their representation patterns, then preserving input structure implies

$$S = \Phi A \tag{1}$$

In the above equation, the system encounters only the matrix S and has to obtain matrices Φ and A through learning and experience. Analytically, this corresponds to solving a matrix factorization problem. The problem can be readily solved when $M > K$, i.e., the number of neurons in the retina is greater than the number of representation neurons. However, in most animal species, the number of high-order neurons in the sensory circuits far exceeds that of the primary neurons, and therefore, $M < K$. The standard matrix factorization methods cannot be used in these situations. An optimization problem that minimizes the difference between the original matrix and the factor matrices' product must be solved. The particular form of such an optimization problem is stated below

$$\underset{\Phi, A}{\operatorname{argmin}} \frac{1}{2} \|S - \Phi A\|_2^2$$

where $\|.\|_2$ corresponds to the *Frobenius norm* of a matrix, which in this case serves as a measure of the difference between matrices.

Note that even by solving this optimization problem, there is no guarantee that the dictionary Φ will contain the most informative structures because it can consist of any feature combination that allows representation to preserve the input structure. To ensure that only the most informative structures are captured, we have to put constraints on this optimization problem. In the previous section, I have shown that to achieve efficiency in representing a finite number of objects; the representations need to comprise a fewer number of neurons and should be based on features that are sparsely encountered. Both these requirements can be satisfied if the representations are based on the most informative features. Conversely, if the representation is constrained to fulfill both these requirements,

then it is likely that it will be based on the most informative structures, and the same will be captured in the dictionary. In other words, a sparsity constraint needs to be put on representations to ensure that the dictionary contains the most informative features. Ideally, we need a restriction on the number of active neurons in any representation so that only a few neurons are active while representing any input. For vectors, the l_0 norm quantifies the number of non-zero elements; therefore, constraining the number of active neurons corresponds to reducing the l_0 norm of the representation vectors. However, it is not possible to analytically derive representation vectors with minimum l_0 norm and satisfy the linear relationship stated in equation (1). One has to examine all potential vectors of dimension K that satisfy the above relationship to find the minimal l_0 norm vector. With $M < K$, this becomes particularly challenging because there are infinitely many representation vectors that satisfy the relationship. Incidentally, under certain conditions, vectors with minimal l_1 norms (sum of absolute values of the components of a vector) form a good approximation of the vector with minimal l_0 norm (Candes and Tao 2005, Donoho 2006a, Argaez et al. 2011, Candes et al. 2006, Elad 2010). Therefore, a unique solution to the problem can be achieved by seeking the minimal l_1 norm of the solution. Accordingly, the above optimization problem can be updated to include the sparsity constraint as under

$$\underset{\Phi,A}{\operatorname{argmin}} \ \frac{1}{2}\|S - \Phi A\|_2^2 + \lambda\|A\|_1 \tag{2}$$

where

$$\|A\|_1 = \sum_i \sum_j |A_{ij}|$$

In general, minimizing the l_1 norm leads to an increase in the difference between the input matrix and the product of its factor matrices. In these situations, one has to decide how much of this difference can be tolerated in favor of achieving sparse representations. Here, the free parameter λ is indicative of such tolerance. Its values typically lie between 0 and 1,

with the value 0 indicating that getting sparse representations is not preferred at all. Value 1 indicates that the sparseness of the representation is equally important to obtaining correct factor matrices. With this formulation, the above optimization problem becomes similar to the optimization performed in Compressed Sensing (Donoho 2006a, Candes and Romberg 2007). Still, it is different in critical ways, as we will discuss later in this chapter.

Enforcing sparseness and structure preservation constraints individual neurons to get tuned to the sets of features unique to individual objects. Consequently, the dictionary captures the most informative structures. However, as the number of objects increases, a single representation neuron will get involved in representing multiple objects. It will be required that the structure that this neuron is tuned to contains maximum information about all these objects. This requirement can be satisfied when the neuron becomes tuned to the features shared within the group of objects and are unique to this particular set. As such structures can only be localized in nature, representing more objects will make the feature combinations captured in the dictionary increasingly localized. In other words, the dictionary elements will get pruned from the most complex to less complicated forms while gradually adapting to more objects. Such pruning can be carried out by performing set difference operations on the inputs. With such a process, the system can eliminate the less informative structures from the set of represented inputs and maintain only the group's most informative structures. However, the system cannot utilize these set difference operations because while representing fewer objects, such operations will hinder preserving input structure. Interestingly, while using a single neuron in representing multiple objects, shared structures can be extracted through superposition followed by normalization of the encoded inputs without explicitly removing the structures not shared by the objects. In other words, a superposition followed by the normalization of multiple inputs will make the shared components more prominent and identifiable than the unshared ones. The uncommon features will “fade away” rather than being eliminated by subtraction. Thus, to represent objects based on the most informative structures in all conditions, a non-negativity constraint

is necessary. This constraint prevents the formation of arbitrarily and unnecessarily complicated dictionary elements by preventing the use of negative coefficients. This constraint is also justified because, in biological brains, information is carried by action potentials and cannot be negative. As such, the optimization problem can be written as the following with the non-negative constraints:

$$\underset{\Phi,A}{\operatorname{argmin}} \ \frac{1}{2}\|S - \Phi A\|_2^2 + \lambda\|A\|_1, \qquad A \geq 0;\ \Phi \geq 0 \tag{3}$$

Solving the optimization problem stated in equation (3) produces a dictionary (Φ) that generates the sparsest representation (output A) patterns for a finite set of input signals (S).

2.12. The probabilistic approach towards basis transformation

In the previous section, I discussed an analytical approach to capture the most informative structures and obtain representations of objects based on those structures. I introduced the task of obtaining object representation as a basis transformation problem and derived constraints to get a basis of informative features. Notably, interpreting sensory processing as a form of basis transformation is not new. Several previous studies that follow efficient coding principles (Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997, Lewicki and Olshausen 1999) have tried to estimate such transformation. However, unlike our approach, these studies adopt a probabilistic view of the transformation process. In particular, of all possible features that can form a representation basis, they aim to find the sparsely occurring independent features whose combinations are most likely present in the natural environment. With these considerations, these approaches arrive at an optimization problem that is similar to ours. In this section, I will explain their probabilistic approach towards basis transformation. Interestingly, though our approach's optimization

problem bears a resemblance to the function optimized in these studies, critical differences exist between the two approaches. I will use the next section to highlight these differences.

The technique utilized in these previous approaches to estimate the representation basis is known as *maximum aposteriori (MAP)* estimation. In this technique, it is assumed that a sensory input identified by the population response of *M* neurons, which can be described by an *M*-dimensional vector s, is transformed into another level of neuronal response described by a *K*-dimensional vector a through the transformation process. The *M*-dimensional input lies in a basis defined by individual pixels, whereas the *K*-dimensional representation is assumed to be based on a set of features. The transformation of the pixel basis to feature set is considered to be linear so that the transformation process can be approximated as

$$s = \Phi a + z$$

Here Φ denotes the basis of features or the dictionary, and z marks the error term in the transformation. The error in transformation is supposed to have a Gaussian distribution given by

$$\mathbb{P}(z) = \frac{1}{Z_c} e^{-\frac{\|z\|^2}{\sigma^2}}$$

where Z_c is a normalizing constant, σ^2 is the variance of the error term, and $\|z\|$ is the norm of the error. Arguments from the probability theory state that the conditional probability of an input s giving rise to a particular response a is also the same, i.e.

$$\mathbb{P}(s|a, \Phi) = \frac{1}{Z_c} e^{-\frac{\|s - \Phi a\|^2}{\sigma^2}}$$

Following the Efficient Coding Hypothesis, the resulting basis set components are expected to be utilized independently; therefore, the representation neurons, a_i, are constrained to be independent in these methods. The constraint is put in place by assuming the representation to be factorial, i.e., the probability of the population response is obtained from the product of individual responses' probabilities. Furthermore, the distribution of

response values of neurons is chosen to have high kurtosis. Such distributions lead to a sparse representation of the inputs, which helps achieve efficiency (Field 1994). In general,

$$\mathbb{P}(a_i) = \frac{1}{Z_\lambda} e^{-\lambda |a_i|} \quad \text{and} \quad \mathbb{P}(a) = \prod_i \mathbb{P}(a_i)$$

Here, as before, Z_λ is a normalizing constant, and $|a_i|$ denotes the absolute value of the response. With these assumptions, the posterior probability of sensory input for a given type of transformation is calculated using the Bayesian rule

$$\mathbb{P}(s|\Phi) \propto \mathbb{P}(s|a, \Phi)\mathbb{P}(a)$$

The posterior probability is the measure of the chance of mapping back a representation to the corresponding sensory input given a particular basis of representation. Naturally, one needs to maximize these probabilities. As it is a function of the representation a, maximizing it gives the input's representation in the transformed basis. A common practice is to minimize the negative of the natural logarithm of the posterior distribution, which is equivalent to minimizing the objective function given as

$$E(a) = \|s - \Phi a\|^2 + \lambda \sum_{i=1}^{k} |a_i|$$

Thus, the optimization problem to solve is

$$\underset{\Phi, a}{\text{argmin}} \ \frac{1}{2}\|s - \Phi a\|_2^2 + \lambda \|a\|_1 \tag{4}$$

which is very similar to the optimization problem that I have introduced in equation (3).

2.13. Differences with previous approaches

As described in the previous section, conventional approaches seeking efficient representation of sensory stimuli have followed a probabilistic approach towards determining a suitable basis. The optimization problem that these approaches propose to

solve is similar to the optimization problem that needs to be solved for obtaining a basis of informative features. However, critical differences exist between the conventional approach and the framework proposed in this study. In this section, I will highlight these critical differences

2.13.1. Sparseness

The idea of sparseness in the representation of sensory stimuli has been proposed for several different reasons. Early works on associative memory in neural networks suggest that sparse representations enhance the memory capacity (Willshaw et al. 1969, Tsodyks and Feigelman 1988, Baum et al. 1988, Cortes and Vapnik 1995) and allow learning associations (Palm 1980, Kanerva 1988, Zetzsche 1990, Field 1994, Palm and Sommer 1996, Schwenker et al. 1996, Földiak and Young 1995, Baddeley 1996). Studies have also illustrated that sparse representations are energy efficient (Levy and Baxter 1996, Attwell and Laughlin 2001, Lennie 2003). However, the conventional approaches, which are based on the Efficient Coding Hypothesis, rationalize sparse representation based on arguments put forth by Barlow (Barlow 1994) and Field (Field 1993, Field 1994). Barlow has argued that, while efficiently representing inputs, the system should seek sparse representations because it presents a convenient way to calculate the inputs' marginal occurrence probabilities (Barlow 1994). He reasoned that if individual inputs' representations are sparse and comprise independent neurons, any neuron will spend extended time in an inactive state. In other words, its probability of being inactive will be close to one. In this condition, utilizing the factorial nature of representation, the inputs' marginal occurrence probabilities can be calculated by just taking the product of activation probabilities of active neurons. The inactive neurons, owing to large probabilities of being idle, will not contribute to calculating inputs' marginal occurrence probabilities. This nature of representations will allow the system to perform such calculations without necessarily confirming each neuron's state.

Later, Field showed that receptive field properties of the simple cells in the visual cortex (Hubel and Wiesel 1962, Hubel and Wiesel 1968) are indeed sparsely distributed in natural scenes (Field 1993, Field 1994). This finding, together with Barlow's argument, supported the classical approaches to seek sparseness in representations.

In contrast to the conventional approaches, we do not seek sparsity in representation because it allows a more straightforward calculation of the marginal properties or because features comprising receptive fields are sparsely distributed. In fact, the framework proposed in this study does not require representations to be based on independent features, and the nature of representation is not factorial. Therefore, in this framework, the sparsity in representation does not help calculate the inputs' marginal occurrence probabilities. Here, sparsity in representations is essentially a consequence rather than a requirement. As most informative features are unique structural components of the objects that rarely occur, the representations based on these structures tend to be maximally distinct and naturally sparse. Therefore, as discussed in previous sections, the sparsity constraint in the optimization problem ensures that representations are based on the most informative structures. Furthermore, the specific form of sparsity that we seek is l_0 sparsity, which corresponds to reducing the number of neurons representing any object. In contrast, the sparse distribution of features that Field showed in his study (Field 1993, Field 1994) corresponds to minimal l_1 sparsity, which translates into minimizing each neuron's activity.

It is important to note that the notion of sparse representations is not purely theoretical. Several experimental studies of the olfactory system of the mouse (Poo and Isaacson 2009, Stettler and Axel 2009) and fly (Turner et al. 2008), auditory system (DeWeese et al. 2003, Lewicki 2002, Smith and Lewicki 2006), and somatosensory system (Brecht and Sakmann 2002) have shown that brain representation of inputs is indeed sparse. Thus, justifying their requirement in theoretical models of sensory processing.

2.13.2. Non-negativity

In our approach, the non-negativity constraint plays a very crucial role. As pointed out in the previous sections, the informativeness of a group of structurally related features is either greater than or equal to the individual features' information content. This difference arises because complex feature groups are more specific to objects. Therefore, while selecting structures for representing objects based on their informativeness, a simple strategy can be to choose structures based on their complexity, where structurally complex features are more likely to represent objects. However, in cases where a large number of objects need to be represented by fewer neurons, each neuron has to participate in representing multiple objects. It can only be tuned to features that are shared among the set of objects that it is involved in representing. Such features will not be specific to any particular object in the group, and their complexity will be lower than the specific structures. Thus, while gradually encountering more objects, our framework demands the system to capture complex as well as simple structures. More specifically, complex structures that are more informative about individual objects could be used for representing them when a few objects need to be represented. As the number of represented objects grows, these complex structures need to be pruned down to simpler structures by removing features that are not specific to an entire group of objects. We reasoned that this pruning could not be reflected in the dictionary by carrying out feature subtraction. A non-negativity constraint was required, so that features specific to the group objects can be highlighted as non-specific features are faded away.

The conventional approaches (Olshausen and Field 1996, Olshausen and Field 1997, Lewicki and Olshausen 1999, Olshausen 2013), on the other hand, use independence as a criterion for selecting features for representations. As groups of independent features are also independent of one another, independence as a criterion

does not differentiate between individual features and complex feature groups. Therefore, a non-negativity constraint is not required in these approaches to adjust feature complexity in an experience-dependent manner. However, without non-negativity constraint, neurons may assume negative states which do not have a biological basis or meaning.

Interestingly, in our approach, requiring non-negativity restricts the neurons' states to be positive and makes them more realistic. Furthermore, together with the sparsity constraint, it makes the activity states of neurons follow an exponential distribution (l_1 sparsity constraint translates into a high kurtosis Laplace distribution and non-negativity constraint restricts this distribution to positive states resulting in an exponential distribution). Exponential distributions are followed by events that rarely occur in the environment. Constraining the states of neurons to such distribution forces them to be tuned to rare, sparsely occurring structures. Thus, the non-negativity constraint, together with the sparsity constraint, forces the neurons to get tuned to more informative structures.

2.13.3. Learning the dictionary

Another significant difference between our approach and the classical approaches is how the dictionary of features is updated and learned. Conventional approaches aim to capture independent features. Therefore, they need to know the entire input space's statistical properties irrespective of what is experienced by the system. This requirement results in a learning rule where the dictionary is updated with respect to multiple sets of objects to incorporate in it as many statistics of input as possible.

On the other hand, our approach tries to optimize the representations for the structure of data encountered by the system in a limited period of experience. Therefore,

the goal is to obtain a dictionary that contains the most informative features of a finite set of objects. This aim leads to an update strategy where the dictionary is updated with respect to the same set of objects. In this way, we assume that in a limited period of sensory experiences, the system encounters a limited set of inputs and intends to efficiently represent only this set by identifying the most informative features from them. This set is supposed to change with experience, and the system adapts the representations to the revised set.

Taken together, though the optimization procedure followed in our approach is similar in form to the procedure followed in classical approaches, the significant differences in underlying assumptions change the motivation towards carrying the optimization. They change the meaning of constraints and their effects on optimization.

2.14. Comparison with Infomax principle

The idea of maximizing information between input and output in neural networks is not new. Linsker, in his seminal studies (Linsker 1987, Linsker 1989a, Linsker 1989b, Linsker 1990), proposed the principle of *"maximum information preservation"* or the *"infomax"* principle. It states that a set of inputs X should be mapped to the set of outputs Y in a neural network so that mutual information between X and Y is maximized. Mutual information, $I(X;Y)$, between the ensemble of inputs X and the corresponding ensemble of outputs Y is defined as

$$I(X;Y) = H(Y) - H(Y|X) \tag{5}$$

where $H(Y)$ is the entropy associated with the ensemble of outputs, and $H(Y|X)$ is the entropy associated with the conditional distribution of outputs conditioned on the inputs. In simpler terms, the entropy of outputs, $H(Y)$, is a measure of average uncertainty in outputs. Similarly, the entropy of the conditional distribution of outputs, $H(Y|X)$, is a measure of

average uncertainty in outputs when the inputs are known. With such definitions, mutual information corresponds to the average uncertainty reduction or the average gain in information about outputs that one obtains by knowing the inputs. One can gain maximum information about outputs from inputs if they become certain about output by knowing the corresponding input. In other words, reducing the average uncertainty about outputs when inputs are known maximizes the mutual information. It corresponds to reducing the conditional entropy term, $H(Y|X)$, in the above equation.

Linsker derived the set of learning rules that will allow the network to maximize information between inputs and outputs when they were related to one another through a linear function. His work was extended by Bell and Sejnowski (Bell and Sejnowski 1995) who derived the learning rules for a network when the inputs and outputs were related through a non-linear function. Interestingly, if the outputs Y comprise of activity patterns in a set of K output neurons, namely y1, y2… yk, then the entropy of outputs, $H(Y)$, can be written as

$$H(Y) = \sum_{i=1}^{K} H(y_i) - D(y_1, y_2, \dots y_k) \tag{6}$$

where $H(y_i)$ is the entropy associated with the distribution of i^{th} representation neuron's state and $D(y_1, y_2, \dots y_k)$ is a term measuring neurons' dependence. Maximizing $H(Y)$, which also corresponds to maximizing mutual information, $I(X; Y)$ implies reducing the term $D(y_1, y_2, \dots y_k)$, or the dependence between neurons. Thus, the infomax principle is equivalent to Barlow's redundancy reduction principle (Barlow 1961). It is also the fundamental principle behind *Independent Component Analysis (ICA)* methods (Jutten and Herault 1991, Hyvärinen and Oja 2000, Hyvarinen et al. 2001, Hyvärinen et al. 2001, Comon 1994, Comon and Jutten 2010, Amari et al. 1996, Nadal and Parga 1994). These methods aim to find independent components in any data distribution.

The framework for representing objects proposed in this study is also based on maximizing information between inputs, that are the objects, and the corresponding outputs, that are their representations. However, this approach does not follow the infomax principle in the form that is described above. To understand the differences between our approach and the infomax principle, let us first combine equations (5) and (6), to express mutual information between objects and their representation as a function of entropy of individual neurons i.e.

$$I(X;Y) = \sum_{i=1}^{K} H(y_i) - D(y_1, y_2, \dots y_k) \; - H(Y|X) \tag{7}$$

In equation (7), the *infomax principle* aims to maximize mutual information by reducing the term $H(Y|X)$ and the term $D(y_1, y_2, \dots y_k)$ in case of efficient coding. On the other hand, our approach focuses more on maximizing the term $\sum_{i=1}^{K} H(y_i)$ which is the sum of the entropies of the individual neurons. This shift in focus is brought into our approach by the combined constraints of sparsity and non-negativity. As described previously, this combination of constraints in our approach translates into an exponential distribution of states of neurons. By seeking sparse and non-negative representations, we force the neurons to remain in inactive states most of the time. They take any other active states rarely so that the probability of acquiring a state of higher activity falls off exponentially with the level of activity. This kind of probability distribution has higher entropy than any other probability distribution defined over the neuronal states. Thus, with the combination of sparsity and non-negativity constraints, we force each $H(y_i)$ to attain maximum possible value, and consequently, maximize $\sum_{i=1}^{K} H(y_i)$. Simply put, the *infomax principle* maximizes the mutual information between objects and their representation by reducing the uncertainty in representations and obtaining maximum information about the input from the representation as a whole. On the other hand, we aim to obtain maximum information about the input from the individual neurons. Indeed, by doing so, we maximize the information that we get from the entire representation as well, but the strategy of getting this information is more specific

in our approach. It must be noted here that with this strategy, we do not seek to reduce $D(y_1, y_2, \ldots y_k)$, which is the term denoting dependence among neurons. Thus, our approach does not necessarily render individual neurons independent.

2.15. Comparison with Compressed Sensing

The optimization problem (equation (4)) that we propose to solve to obtain the most informative structural components of objects also involves solving for a sparse representation of objects. Moreover, as the number of cells comprising the representation is probably greater than the number of cells in the retina, the sparse representations are supposed to have a higher dimension than the input. Thus, the optimization problem proposed in this study requires solving for a sparse high dimensional output from a non-sparse low dimensional input, and hence, bears similarity with the approach of Compressed sensing(Candès et al. 2006, Candes and Tao 2006, Donoho 2006a, Candes and Romberg 2007).

Compressed sensing is a signal processing technique that exploits sparsity in signals to recover them using far fewer measurements than traditionally required (Candès et al. 2006, Candes and Tao 2006, Donoho 2006b). To understand this technique, consider the example of movie rating. Any movie has various aspects like story, screenplay, acting, music, etc., which can be thought of as its components that have values corresponding to the viewers' approval associated with them. A good movie has higher approval of viewers for its certain aspects, and therefore, its corresponding components have higher values. In this regard, any film can be described as a list of values of its various components, and one can decide which movies to watch based on this list of values. However, movies are not described in terms of these components but are rated out of a certain score. This rating score of the movie is determined by taking into account its various aspects. It can be thought of as

a compressed description of the movie derived from the combination of its component values. Several different combinations of component values can give rise to the same score, like a person who enjoys comedy can give a higher rating to a movie with a poor storyline, and conversely, a person who prefers good stories can give the same rating to a movie with no comical aspect. Therefore, it is desirable to know how good the movie is in each of its different aspects from its ratings. The Compressed sensing technique solves this problem. It uses compressed measurements of signals, like compressed movie scores, to determine the actual signal, i.e., movie components' actual values. Many excellent articles explain the detailed concepts of Compressed sensing (Ganguli and Sompolinsky 2012, Eldar and Kutyniok 2012, Duarte and Eldar 2011). Briefly, it uses an l_1 minimization approach (Candes et al. 2008, Candes and Romberg 2005) to estimate the signal, which in the example of the movie is the list of values of all movie components, from very few measurements like different ratings of the movie made by different viewers. Interestingly, it has been shown that inaccurate or noisy measurements also lead to accurate signal recovery, thus enabling correct identification of the signals using the Compressed sensing technique even when information about them is inaccurate and incomplete (Candès et al. 2006, Candes and Tao 2006).

As noted previously, the optimization problem that we derive for obtaining the dictionary and representation of objects bears similarity to Compressed sensing in terms of dimensional expansion and l_1 minimization. However, the problems being addressed are fundamentally different. In Compressed sensing, the measurements that correspond to the dictionary of features Φ in our approach are constructed to meet the restricted isometry property (Candès et al. 2006) and are known beforehand. In contrast, in our representation framework, the Φ term is not constructed. It is a transforming matrix that is learned and contains all informative features observed in the set of encountered objects. Furthermore, the dictionary changes in form as more objects are experienced. The measurement of signals in the Compressed sensing approach remains fixed. Lastly, the dictionary in our approach is

not required to follow properties like incoherence (Candes and Tao 2006, Candes and Romberg 2007) and restricted isometry property (Candès et al. 2006) The measurements in the Compressed sensing approach must have these properties.

2.16. Discussion

To survive and gain insights about its surrounding, an organism must recognize objects and detect associations among them. Its performance in both these tasks relies on internal representations of objects that it forms through sensory processing. However, the two sets of theories, namely efficient coding and view-based representation frameworks, that aim to describe sensory processing, do not explain how brain representation of objects must be formed to enable the organism to perform these tasks. The motivation behind the efficient coding and redundancy reduction principle is to arrive at a factorial representation of objects that permits straightforward detection of object associations. In contrast, view-based frameworks are focused on attaining invariance in representation. While the two sets of theories have successfully explained the lower and higher levels of visual processing, they are not compatible with each other. A factorial representation based on independent features with minimal redundancy is not suitable for invariant coding, especially in cases involving corruption or occlusion. Conversely, the view-based framework relies on hierarchical models to learn feature associations. Multiple instances of feature combinations originating from the same object are learned to achieve robustness. This learning disrupts the factorial nature of representations.

Here, I have proposed an alternative approach to represent objects based on their most informative features. We find that informativeness of features is a suitable criterion for selecting features for representing objects. Object representations based on informative features are sensitive to similar objects and stable for different forms of the same object.

Furthermore, these features allow representations to be efficient in conveying information about inputs and can be learned in an experience-dependent manner. Importantly, representing objects based on informative features does not require the system to know the surroundings' accurate statistics. A rarely observed feature is likely to be more informative than a commonly observed one. Therefore, an estimate of only the currently observed statistical properties of the environment is sufficient to determine its informative components.

Informative features are different from independent features that were proposed to be the basis of representing objects in the Efficient Coding Hypothesis (Barlow 1961, Barlow 1989, Barlow et al. 1989). While informative features need not be independent, independent features are not necessarily the most informative structures. In fact, the very idea of determining rarely occurring structural components is antithetical to Barlow's idea of detecting *suspicious coincidences* in the environment (Barlow 1987). A combination of features was called a *suspicious coincidence* if its occurrence was more probable than what could be predicted from its constituent features' occurrences, assuming that the constituent features are independent. Thus, detection of a *suspicious coincidence* was an indication of dependence. Therefore, Barlow proposed that the way to discover the environment's associative structure was to detect these *suspicious coincidences* (Barlow 1987, Barlow 1989, Barlow et al. 1989).

In contrast, rarely occurring structures are feature combinations that likely appear less than their constituting elements, and detecting one such structure does not reveal the environment's associative structure. Instead, they are likely to provide information about individual objects, which can be further processed to detect the associative structure. Detecting *suspicious coincidences* does not provide information about individual objects.

A biological system is peculiar in the sense that it never really encounters all objects at once. It gradually learns about its environment with experience, and it is likely that properties of features, like informativeness and independence, change as new objects are

discovered. Utilizing rarely occurring, informative structural components for representing objects allows the system to adapt to its environment. The system can determine which structures and features it is experiencing relatively less than the others and can use them to represent specific objects. As its experience changes, or as new objects are encountered, these sets of features can be updated. Thus, this framework enables an adaptive nature in representation. However, this adaptive nature is a departure from an ideal input representation scenario, where representations are constructed after considering the entire statistics of the inputs. The consideration is important primarily to minimize information loss and ensure efficiency in representation. However, it can be argued that for a biological system, a goal more important than efficient information transmission is the utilization of that information in the decision-making process. In this regard, efficiency in communicating information about all existing inputs may be ignored in favor of communicating only selected information that has ethological relevance for the system.

The transformation of sensory inputs into brain representation is essentially achieved by transforming the basis of representation. At the peripheral levels of sensory processing, the image of an object elicits responses in the neurons. Thus, the object is represented in terms of the pixels that constitute its image. At higher visual processing levels, the representation is based on complex structural features. I have shown that solving a constrained optimization problem provides an analytical way to transform the representation basis from pixels to informative structures. The approach bears resemblance with the previous approaches of transforming pixel basis into independent features. However, critical differences exist between this approach and the classical approaches in terms of constraints and their interpretation. Specifically, a combination of non-negativity and sparsity constraint in this framework forces individual neurons to get tuned to rarely occurring features and extract informative structures from the environment. Similar critical differences also exist between this approach and the infomax approaches utilized to find independent structures in

any data. Thus, this framework puts sensory processing in a completely different light and offers a new perspective for understanding it.

CHAPTER 3

Obtaining object representation through sparse non-negative matrix factorization

Table of Contents

3.1. Introduction

As described in the previous chapter, we introduce a novel framework of representing objects based on their most informative features in this study. The framework enables stable representations of objects and efficient communication of information. It does not require complete knowledge of the objects' statistical properties and allows the system to adapt to its environment in an experience-dependent manner. Thus, it is suited to biological systems. Moreover, considering the representation process as a linear transformation of basis, solving the following optimization problem can extract informative features from the inputs

$$\underset{\Phi,A}{\operatorname{argmin}} \ \frac{1}{2}\|S-\Phi A\|_2^2+\lambda\|A\|_1, \qquad \text{subject to} \ \ A\geq 0;\ \Phi\geq 0 \qquad (P)$$

The optimization of the problem, P, is not unique to our approach. It has been extensively investigated in the signal processing field to obtain individual signals from a mixture of signals. The problem of unmixing signals is popularly known as *blind source separation* or *BSS* problem (Ans et al. 1985, Bar-Ness et al. 1982, Herault and Ans 1984, Herault and Jutten 1986, Hérault et al. 1985).

In this chapter, with a brief introduction to blind source separation approaches, I will discuss how to solve P in its current formulation. I will then describe the methods utilized to characterize the dictionary, Φ, and the representation, A, obtained through the optimization process. The results and their discussions follow after that.

3.2. Blind source separation and Non-negative matrix factorization

The blind source separation technique originated from attempting to solve a biological problem (Comon and Jutten 2010). It was observed that in the case of motion in

a joint, though two types of sensory neurons carried information about stretching and speed, both the types conveyed a mixture of stretching and speed information (Roll 1981). In other words, the activities in both neuron types were found to be dependent on both the joint speed and its stretch. The problem was to understand how each of these properties contributed to the activity of neurons. Formally, if $f_1(t)$ and $f_2(t)$ were activities of two sensory neurons as functions of time, then the goal was to assess time-varying speed $v(t)$ and stretch $s(t)$ in the joint, along with factors a_{1v}, a_{1s}, a_{2v} and a_{2s} that corresponded to the contributions of speed (v) and stretch (s) in the neurons' activities. Here, a_{1v} denotes the contribution factor of speed in the activity of neuron type 1 and a_{1s} represents the contribution factor of stretch in the same neuron. The factors for neuron type 2 are similarly indicated. When speed and stretch are assumed to be contributing linearly to the activities, the problem is reduced to solving the system of equations

$$f_1(t) = a_{1v}v(t) + a_{1s}s(t)$$

$$f_2(t) = a_{2v}v(t) + a_{2s}s(t)$$

where $f_1(t)$ and $f_2(t)$ are the only known quantities

With this linear assumption, each sensory neuron's activity originating from the joint could be considered a linear mixture of time-varying speed and stretch signals. The objective is then translated into determining the constituent signals of the mixture and the process of mixing. In this formulation, the problem could be readily recognized as an unmixing problem encountered in the signal processing field. In these problems, A set of N signal sources are assumed to be generating n-dimensional signals s_i, which are linearly mixed to create a signal mixture m_j as

$$m_j = \sum_{i=1}^{N} \alpha_{ji} s_i$$

Here, α_{ji} denotes the coefficients of the linear combinations of the source signals. The aim is to use k different mixtures of the same set of source signals and determine both the sources and the mixing process. The signal unmixing problem can be imagined as the *cocktail party*

problem (Cherry 1953). In this problem, one attempts to recognize what a person is saying in a cocktail party where everyone is chatting (Bronkhorst 2000, Yost 1997).

In mathematical terms, if k different mixture signals are sampled, then the overall mixing process can be described by a matrix A, where the $(p, q)^{th}$ element of A is the coefficient of q^{th} source signal in the p^{th} mixture sample, i.e., α_{pq}. Representing all signals in a $n \times N$ matrix S, and all mixed samples in a $n \times k$ matrix M, the mixing process can be described in the matrix forms as

$$M = SA$$

The goal here is to find matrices A and S from matrix M. The mixing process is often assumed not to be clean, and some noise is introduced in the mixed signal. The overall mixing process can then be formulated as

$$M = SA + R$$

where R is an $n \times k$ noise matrix. In this formulation, the goal is to determine R together with A and S.

The technique utilized to do this unmixing or separation is called *Blind source separation* (*BSS*) and has been extensively studied (Bar-Ness et al. 1982, Herault and Ans 1984, Ans et al. 1985, Hérault et al. 1985, Herault and Jutten 1986). Evidently, blind source separation is an ill-posed problem that does not have a unique solution. Infinitely many solutions to the problem exist, specifically in cases where matrix A is underdetermined, i.e., the number of mixture samples, k, is less than the number of sources, N. Even when matrix A is complete, i.e., k equals N, the source signals can only be estimated up to a permutation or a scale. Therefore, certain assumptions need to be made about the sources to determine them uniquely.

One assumption that has been widely considered in *BSS* approaches is the independence of sources. It is assumed that the sources generating the signal are not influenced by one another, and the statistics of signals generated from all the sources are the

same. In other words, it is presumed that the sources are *i.i.d,* i.e., they are independent and identically distributed (Comon 1994). This set of assumptions leads to a separation approach that is commonly referred to as *Independent Component Analysis* (*ICA*) (Herault and Ans 1984, Ans et al. 1985, Hérault et al. 1985, Herault and Jutten 1986, Bell and Sejnowski 1995, Hyvärinen and Oja 2000, Hyvarinen et al. 2001, Hyvärinen et al. 2001, Hyvärinen 1998, Stone 2004).

In another set of approaches, the sources are either assumed to be only identically distributed (Belouchrani et al. 1993, Molgedey and Schuster 1994, Tong et al. 1990) or assumed to be only independent (Matsuoka et al. 1995, Pham and Cardoso 2001). As these approaches are more straightforward, a significant advantage of using them over the *ICA* approach is that they are faster and efficient to implement (Comon and Jutten 2010). Apart from independence and distribution of sources, geometrical properties of their joint distribution (Pham and Vrins 2006, Puntonet et al. 1995, Theis et al. 2003a, Theis et al. 2003b), discreteness of signal values (Castella 2008, Grellier and Comon 1998, Jallon et al. 2004), and other correlated properties like their coherence with other signals (Rivet et al. 2005, Sodoyer et al. 2004) are also utilized for identifying them uniquely.

For underdetermined *BSS*, the assumption that each mixed signal is a sparse combination of sources is particularly useful (Bofill and Zibulevsky 2001, Jourjine et al. 2000, Lee et al. 1999, Lewicki and Sejnowski 2000, Lin et al. 1997, Van Hulle 1999, Yilmaz and Rickard 2004, Zibulevsky and Pearlmutter 2001). In these approaches, each source signal's coefficient is modeled with a sparse or super-Gaussian distribution, which has a peak at zero and heavy tails everywhere else. A standard model for such distribution is the family of generalized Gaussian distribution (Charkani and Deville 1999a, Charkani and Deville 1999b, Vincent 2007, Wu and Principe 1999, Comon and Jutten 2010) that is formulated as

$$p(x) \propto exp(-\eta|x|^{\tau})$$

Here η and τ are the parameters of the distribution. Assuming that all coefficients follow this distribution, and the sources are independent of one another renders the joint distribution of coefficients as

$$p(A) \propto exp\left(-\eta \sum_{i,j} |\alpha_{ij}|^{\tau}\right)$$

which can be expressed as

$$p(A) \propto exp(-\eta \|A\|_{\tau}^{\tau}) \quad \text{where} \quad \|A\|_{\tau} = \left(\sum_{i,j} |\alpha_{ij}|^{\tau}\right)^{\frac{1}{\tau}}$$

Assuming that the noise in mixing is Gaussian, the underdetermined BSS problem reduces to solving the optimization problem expressed as

$$\underset{A,S}{\operatorname{argmin}} \|M - AS\|_2^2 + \lambda \|A\|_{\tau}$$

using the method of *maximum aposteriori* (*MAP*) estimation. Here $\|.\|_2$ is defined similarly to $\|.\|_{\tau}$ with $\tau = 2$.

The optimization can be carried out using several methods like M-FOCUSS (Cotter et al. 2005), Basis Pursuit (Chen et al. 2001), Least angle regression (LARS) (Efron et al. 2004), Iterative thresholding (Daubechies et al. 2004, Elad and Aharon 2006, Figueiredo and Nowak 2003), and Matching pursuit (Mallat and Zhang 1993, Leviatan and Temlyakov 2006, Gribonval 2002, Gribonval and Nielsen 2003)

In many situations, signals can only be additive. Consider the intensity of pixels in an image or the amplitudes of sound waves as examples. The values of these signals are all non-negative, and the mixing process, like a superposition of images, can only add one signal to another. If one performs *BSS* on these signals, their additive nature can be utilized to separate the sources better. The approach used for such separation is called *Non-negative Matrix Factorization* (*NMF*) (Lee and Seung 1999, Lee and Seung 2000, Leggett 1977,

Paatero and Tapper 1994). As evident by its name, the method aims at finding only additive sources by factoring the mixture matrix into a non-negative matrix and a general matrix i.e.

$$M = SA \quad \text{where} \quad S \geq 0$$

Note that in the above problem, no constraint is imposed on the mixing matrix A. However, in certain situations, like in cases where all positive mixtures are observed for all positive source signals, it is safe to assume that the mixing process is all additive. In these situations, the mixing matrix will be non-negative as well, and the above problem can be restated as,

$$M = SA \quad \text{where} \quad S \geq 0;\ A \geq 0$$

the separation can be achieved by solving the optimization problem stated as

$$\underset{A,S}{\operatorname{argmin}} \|M - SA\|_F^2 \quad \text{subject to} \quad S \geq 0 \ \text{ and } \ A \geq 0$$

using methods like Gradient descent (Curry 1944), Gradient descent with multiplicative updates (Lee and Seung 2000), or Alternative Least Squares (ALS) (Berry et al. 2007, Bro and De Jong 1997, Cichocki et al. 2009, Cichocki and Phan 2009, Cichocki et al. 2008, Tauler et al. 1991).

Recent studies have also introduced a sparsity constraint in the *NMF* (Hoyer 2002, Hoyer 2004, Hoyer and Hyvärinen 2002) so that the optimization problem becomes

$$\underset{A,S}{\operatorname{argmin}} \|M - SA\|_F^2 + \lambda \|A\|_1 \quad \text{subject to} \quad S \geq 0 \ \text{ and } \ A \geq 0 \qquad (P')$$

where $\|A\|_1$ is the l_1 norm of the source signals and is a convex measure of the signals' sparsity. The sparsity assumption essentially translates into assuming a non-Gaussian prior distribution of source signals, which can be better estimated through iterative algorithms (Hoyer 2004). These studies demonstrated that *NMF* could better capture the source signals' complex structures with this extension and could be utilized in image processing (Hoyer 2004) or text mining (Pauca et al. 2004).

BSS approaches like *ICA* have found extensive application in the field of biomedical sciences. *ICA*, in particular, has been utilized in analyzing EEG/MEG data (Flexer et al.

2005, Hu et al. 2007, Makeig et al. 1996, Onton et al. 2005, Ossadtchi et al. 2004, Porée et al. 2006, Tang et al. 2002a, Tang et al. 2002b, Vigário et al. 1997). Artifact detection and removal (Iriarte et al. 2003, James and Gibson 2003, Joyce et al. 2004, Jung et al. 2000, Jung et al. 2001, Vigário 1997, Vigário et al. 2000), analysis of event-related response averages (Makeig et al. 1996, Makeig et al. 1997), and single-trial EEG/MEG (Debener et al. 2005, Jung et al. 2001, Onton et al. 2005) have all benefitted from the *ICA* approach. *ICA* has also been applicable in fMRI (Calhoun et al. 2003, McKeown et al. 1998, McKeown and Sejnowski 1998), Magnetic Resonance Spectroscopy (MRS) (Pulkkinen et al. 2005, Sajda et al. 2004), and EMG (Farina et al. 2004). BSS has also been employed in Medical acoustics, ultrasound, infrasound techniques (Gallippi and Trahey 2002, Ham et al. 1999, Pietilä et al. 2006, Xinhua et al. 2000).

A relatively straightforward application of BSS is in separating different sound mixtures. However, as sound signals extend over time, most suitable approaches for their separation regard mixed sounds as convolutive mixtures rather than linear mixtures (Comon 1990, Gorokhov and Loubaton 1997, Yellin and Weinstein 1994, Yellin and Weinstein 1996). The main difference between convolutive and linear mixtures is that the same source explains delayed mixture portions in the former. In other words, the same source is utilized to explain parts of the mixture observed at different time points. Several *ICA* procedures have been extended to include such convolutions and are used in separating recorded or synthesized mixtures of sounds (Albouy and Deville 2001, Charkani and Deville 1999b, Choi and Cichocki 1997, Douglas et al. 2004, Ehlers and Schuster 1997, Ham et al. 1999, Ito et al. 2002, Thomas et al. 2006, Wehr et al. 2007). Sparse BSS technique has also been utilized broadly for separating sound mixtures (Mitianoudis and Stathaki 2007, Abrard and Deville 2005, Arberet et al. 2006, Bofill 2008).

Similarly, NMF approaches have also found wide applications in fields of Air quality analysis and chemometrics (Henry 1997, Henry 2002), text analysis (Novak and Mammone

2001, Tsuge et al. 2001, Xu et al. 2003), image processing (Buchsbaum and Bloch 2002, Lee and Seung 1999, Lee et al. 2001), audio analysis (Schmidt and Mørup 2006, Smaragdis 2004, Virtanen 2004, Smaragdis 2006), and gene expression analysis (Devarajan 2008).

The similarity between problems P and P' indicates that one can perform sparse *NMF* on objects to obtain their most informative structures. Accordingly, we utilized the *naïve Generalized Morphological Component Analysis* (*nGMCA*) algorithm (Rapin et al. 2013) to solve our optimization problem. The details of this algorithm are discussed in the following section.

3.3. Methods

3.3.1. nGMCA

The form of the optimization problem, P, that we intend to solve is similar to the non-negative blind source separation problem P'. Therefore, it is natural to utilize an algorithm that solves the latter. However, the motivation for using this particular objective function in our approach, and the interpretation of the optimal solution is different. For example, we do not intend to unmix the signals originating from non-Gaussian sources. Instead, we seek a transformation that results in object representations being based on their most informative structures. We do not have a mixed signal to unmix. Similarly, the non-negativity constraint does not indicate prior knowledge of the signal properties, but its implication is to derive interpretable structures of varying complexity in different representation scenarios.

To solve the optimization problem stated above, I utilized a specific algorithm called *naïve Generalized Morphological Component Analysis* (*nGMCA*) (Rapin et al. 2013). I used the algorithm's MATLAB implementation publicly available at https://www.cosmostat.org/ngmca/. The algorithm is essentially an extension of the sparse

BSS algorithm known as *Generalized Morphological component Analysis* (*GMCA*) (Bobin et al. 2007, Bobin et al. 2008) to include non-negativity constraint. It solves the optimization P by iteratively solving two subproblems P_A and P_Φ as described below

$$\min_A \|S - \Phi A\|^2 + \lambda \sum_{i,j} |A_{i,j}| + f^+(A) \qquad (P_A)$$

$$\min_\Phi \|S - \Phi A\|^2 + f^+(\Phi) \qquad (P_\Phi)$$

Here $f^+(x)$ is a function that takes the value of $+\infty$ when $x < 0$ and equals x otherwise. Solutions to P_A and P_Φ are obtained by iteratively updating A and Φ by the rules given below

$$A_{t+1} \leftarrow \left[A_t - \frac{1}{L_A}\left({\Phi_t}^T(\Phi_t A_t - \Phi) - \lambda_t \hat{1}\right)\right]_+$$

$$\Phi_{t+1} \leftarrow \left[\Phi_t - \frac{1}{L_\Phi}(\Phi_t A_{t+1} - S)\Phi_t^T\right]_+$$

where $[.]_+$ denotes positive thresholding, L_A and L_Φ are the maximum eigenvalues of the matrices AA^T and $\Phi^T\Phi$, respectively, and $\hat{1}$ is a vector of all 1s.

3.3.2. Sparse recovery of input representations

Any input $\tilde{s}$ is expected to be related to its representation $\tilde{a}$ in the representation basis Φ in the following way

$$\tilde{s} = \Phi\tilde{a}$$

The above relation is a linear system of equations that can be solved for any input $\tilde{s}$ to obtain its representation $\tilde{a}$. The system can be uniquely solved if balanced, i.e., when Φ is a square matrix and all its columns are linearly independent. However, in this study Φ represents a k-dimensional representation basis of informative structures transformed from an m-dimensional primary response space. So, Φ can be a square only if k and m are equal. A common observation in sensory systems is that the number of higher-level neurons is far

more than the number of primary neurons, making k larger than m. Such a system of equations is called an underdetermined system, and a unique solution cannot be obtained for it.

However, recent theories developed independently by Donoho and by Candes and Tao (Candes and Romberg 2005, Candès et al. 2006, Candes and Tao 2006, Donoho 2006b, Donoho 2006a, Donoho and Elad 2003) show that a unique solution can be obtained by requiring the solution to be sparse. In our approach, the transformation process explicitly seeks sparseness because it ensures that object representations are based on the most informative structures and are efficient. I obtained the sparse solutions using a sparse recovery approach that solves the following optimization problem

$$\min_{\tilde{a}} \|\tilde{a}\|_1 \quad \text{subject to } \tilde{s} = \Phi\tilde{a}, \text{ where, } \|\tilde{a}\|_1 = \sum_i |\tilde{a}_i|$$

The optimization can be implemented as a standard convex optimization procedure (Candes and Romberg 2005). I used the MATLAB implementation of the procedure is publicly available at https://statweb.stanford.edu/~candes/software/l1magic/.

3.3.3. Information-theoretic analysis

It is essential to understand the information-theoretic aspect of any representation framework. However, many information-theoretic quantities are defined over the distributions of random variables. In representing a finite set of inputs, the entire distribution of data is not known. Nevertheless, one can use basic definitions of the information-related quantities. Here, I describe these definitions and the intuitions behind them using a simplistic example.

Consider a world where only three types of inputs A, B, and C, exist, and each of them is encoded by a set of 9 encoders (**Figure 3.1**). All inputs occur with equal frequency, and each encoder can only have two states –on, and off. The goal here is to identify the

statistics of this stimulus space and characterize the information that any encoder's activity gives about the input.

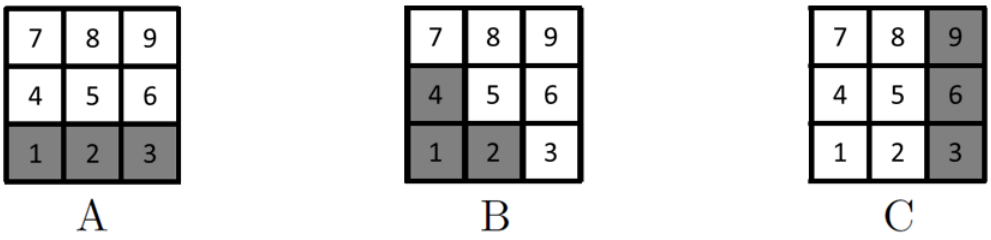


Figure 3. 1: Illustration of inputs and calculation of mutual information

Three inputs are encoded with nine encoders. The black boxes denote the "on" values, and the white boxes indicate the "off" values.

An important point to note here is that an encoder is only active when input is present. Mathematically, this means that a probability distribution can be defined for the occurrence of all observed inputs so that

$$\sum_{X} \mathbb{P}_X = 1$$

where X represents the set of all inputs, which in this case are three. It is important to note that this probability distribution gives the joint probabilities of encoders' activities.

$$\mathbb{P}_X(X = A) = \mathbb{P}_X(1, 2, 3) = \frac{1}{3}$$

$$\mathbb{P}_X(X = A) = \mathbb{P}_X(1, 2, 4) = \frac{1}{3}$$

$$\mathbb{P}_X(X = A) = \mathbb{P}_X(3, 6, 9) = \frac{1}{3}$$

From these probabilities, we can calculate the marginal probabilities of activations of individual encoders as under

$$\mathbb{P}_X = \sum_X \mathbb{P}_X(\text{encoder 1 is active})$$

$$= \mathbb{P}_X(1,2,3) + \mathbb{P}_X(1,2,4) = \frac{2}{3}$$

Similarly

$$\mathbb{P}_X(2) = \mathbb{P}_X(3) = \frac{2}{3};\ \mathbb{P}_X(4) = \mathbb{P}_X(6) = \mathbb{P}_X(9) = \frac{1}{3}$$

Note that, if we represent encoder numbers by n, then

$$\sum_n \mathbb{P}_X(n) \neq 1$$

Once we have marginal probabilities of encoder activities, we can calculate conditional probabilities using Bayes' theorem. For instance, the probability that the detected input is *A* given encoder 1 is active, i.e., $\mathbb{P}_X(X = A|1)$ can be calculated using the Bayes' theorem

$$\mathbb{P}_X(X = A|1) = \frac{\mathbb{P}_X(1|X = A)\mathbb{P}_X(X = A)}{\mathbb{P}_X(1)}$$

where $\mathbb{P}_X(1|X = A)$ is the probability that encoder 1 is active given the code corresponding to input A was detected, $\mathbb{P}_X(X = A)$ is the probability of detecting input A, and $\mathbb{P}_X(1)$ is the marginal probability of encoder 1 being active. Putting in the corresponding numbers, we get

$$\mathbb{P}_X(X = A|1) = \frac{1.\left(\frac{1}{3}\right)}{\left(\frac{2}{3}\right)} = \frac{1}{2}$$

Similarly, other conditional probabilities can be calculated

$$\mathbb{P}_X(X = A|2) = \frac{\mathbb{P}_X(2|X = A)\mathbb{P}_X(X = A)}{\mathbb{P}_X(2)} = \frac{1.\left(\frac{1}{3}\right)}{\left(\frac{2}{3}\right)} = \frac{1}{2}$$

$$\mathbb{P}_X(X = A|3) = \frac{\mathbb{P}_X(3|X = A)\mathbb{P}_X(X = A)}{\mathbb{P}_X} = \frac{1.\left(\frac{1}{3}\right)}{\left(\frac{2}{3}\right)} = \frac{1}{2}$$

$$\mathbb{P}_X(X = A|1,2) = \frac{\mathbb{P}_X(1,2|X = A)\mathbb{P}_X(X = A)}{\mathbb{P}_X(1,2)} = \frac{1.\left(\frac{1}{3}\right)}{\left(\frac{2}{3}\right)} = \frac{1}{2}$$

$$\mathbb{P}_X(X = A|1,3) = \frac{\mathbb{P}_X(1,3|X = A)\mathbb{P}_X(X = A)}{\mathbb{P}_X(1,3)} = \frac{1.\left(\frac{1}{3}\right)}{\left(\frac{1}{3}\right)} = 1$$

One can see that probability that the detected input was A given encoder 1 and 3 are active is 1 because no other codeword contains both the encoders in an active state.

Further, these marginal and conditional probabilities can be utilized to calculate the information content of various events. For example, the information content of occurrence of input A is given by

$$-\log_2 \mathbb{P}_X(X = A) = -\log_2 \frac{1}{3} = 1.6 \text{ bits}$$

Also, the information content of the occurrence of A when encoder 1 is known to be active is

$$-\log_2 \mathbb{P}_X(X = A|1) = -\log_2 \frac{1}{2} = 1 \text{ bit}$$

The difference in the information contents is the information about A that is conveyed by the activity of encoder 1, i.e., 1.6 bits - 1 bit = 0.6 bits.

I have used similar information content calculations to determine the information contents of structures about objects. In most cases, the objects were 2-dimensional binary images, similar to the inputs described above. The informative structures were obtained through performing *NMF* on these inputs.

3.3.4. Bit entropies and redundancy

One measure of information communicating capacity of the encoders is their bit entropy (Barlow 1989, Barlow et al. 1989, Barlow 1991, Barlow 1994). If they are

considered as binary random variables, the entropy of the distribution defined over their states can be calculated as under

$$H(a_i) = -(\mathbb{P}(a_i = 1)\log\mathbb{P}(a_i = 1) + \mathbb{P}(a_i = 0)\log\mathbb{P}(a_i = 0))$$

The bit entropy of the encoders is then defined as the sum of the entropies of the individual encoders i.e.

$$H_a = \sum_{i=1}^{m} H(a_i)$$

Representing stimuli such that the bit entropy of neurons matches the inputs' entropy leads to efficient coding (Barlow 1989, Barlow et al. 1989, Barlow 1991, Barlow 1994). The efficiency of representation can be assessed using redundancy, which is defined as under

$$R = 1 - \frac{H_s}{H_a}$$

where H_s is the entropy of the sensory inputs.

I used these definitions to measure both bit entropy and the neurons' redundancy in various circumstances to assess representation efficiency.

3.3.5. Data sets

To simulate different forms of sensory inputs, I utilized different data sets. These data sets are listed below

1. Symbol data: A set of simplistic binary sensory inputs were modeled as a set of 1000 symbols from different languages. Each symbol was a 16-pixels by 16-pixel image where each pixel corresponded to a neuron. The neurons had only two possible states – on (1) and off (0). As quantifying information-theoretical terms was convenient using this type of binary inputs, this data set was chiefly used to characterize this representation framework's details.

2. Face data: To simulate non-binary sensory input different from natural scenes, we used a set of 2000 grayscale faces. There were 1000 face images each of males and females of size 100-pixels by 100-pixels. These images were resized to 25-pixels by 25-pixel grayscale images. Again, each pixel corresponded to a neuron. This dataset was utilized to assess if the representations obtained under this framework could be used for higher-order cognitive functions like recognition.
3. Natural scenes data: We also tested if our adaptive strategy of coding could explain certain aspects of the traditional efficient coding. A set of 2995 16-pixels by 16-pixels patches from natural scenes, assembled in Van Hateran data set (Van Hateren and van der Schaaf 1998), were utilized in these simulations. The images were grayscale but were not whitened before simulations.

4. Olfactory response data: A response of 94 glomeruli located on the dorsal surface of the mouse brain to 40 different odors was considered as sensory input. The data used was previously published calcium imaging of the olfactory bulb, in which we had imaged the response of dorsal olfactory bulb of GCaMP2 mice to 189 chemicals (Ma et al. 2012). Of these chemicals, ~150 did not elicit significant responses in the glomeruli. Since non-responding stimuli provide no information for our analyses, we removed them from further analysis. To accomplish this, we calculated the Euclidean length of each response and plotted a histogram of response amplitude. 40 chemicals elicited responses that crossed the threshold length of 0.1. These odor-evoked responses were used as inputs to the system.

3.3.6. Corruption of inputs

The consistency of input representation was assessed using corrupted inputs. Inputs were corrupted in different ways, and the sparse representations of the corrupted inputs were compared to the representations of their uncorrupted forms. Following three types of corruption were made:

Noise-added corruption: we introduce noise by adding a Gaussian i.i.d. matrix $\mathcal{N}$ of varying standard deviation to the input matrix S, i.e., $S_{\mathcal{N}} = S + \mathcal{N}$, where $S_{\mathcal{N}} \in \mathbb{R}^{M \times N}$, is a matrix representation of noisy input.

Pixel corruption: A fraction of the *M* pixels (glomeruli) was selected from the inputs. Their values are maintained, whereas the coefficients of the rest were set to zero.

Occlusion: For images, a contiguous set of pixels were selected, and their values were set to zero.

3.3.7. Monte Carlo analysis

Monte Carlo simulations were performed by choosing 100 random sets (numbers varied from 2 to *M*) of pixels (glomeruli) and using each of these randomly chosen sets to obtain the sparse representation in the representation basis. The consistency of the obtained representations was assessed.

3.4. Results

3.4.1. Analysis of symbols

To portray the representation framework based on the most informative features, we utilized the symbols data set (**Figure 3.2.1**). Each image in the dataset was regarded as an input with pixels corresponding to neurons in the early stages of visual processing. The most informative structures from the inputs and input representations based on these structures were obtained using the *nGMCA* algorithm (see methods). Examples of three inputs and their representations are shown. Representations are shown as grayscale images where each pixel corresponds to a representation neuron (**Figure 3.2.2(i)**). The pixels' grayscale values indicate the response levels of representation neurons, depicted as a bar plot (**Figure 3.2.2(ii)**). The set of features to which the representation neurons are tuned is shown in the form of an image array, which we call the dictionary (**Figure 3.2.3**). There is a correspondence between neurons' position in the representation image and the location of tuning properties in the dictionary. It is important to realize that two free parameters exist while representing a fixed number of sensory inputs. These parameters are the number of inputs and the number of representation neurons. We analyzed how the representations change when either of these parameters is varied. Two sets of simulations were performed. In the first set, the number of inputs was varied while keeping the number of neurons fixed, and in the second set of simulations, the number of neurons was varied while keeping the number of inputs fixed. Different characteristics of the representations that emerged in the two sets of simulations were analyzed.

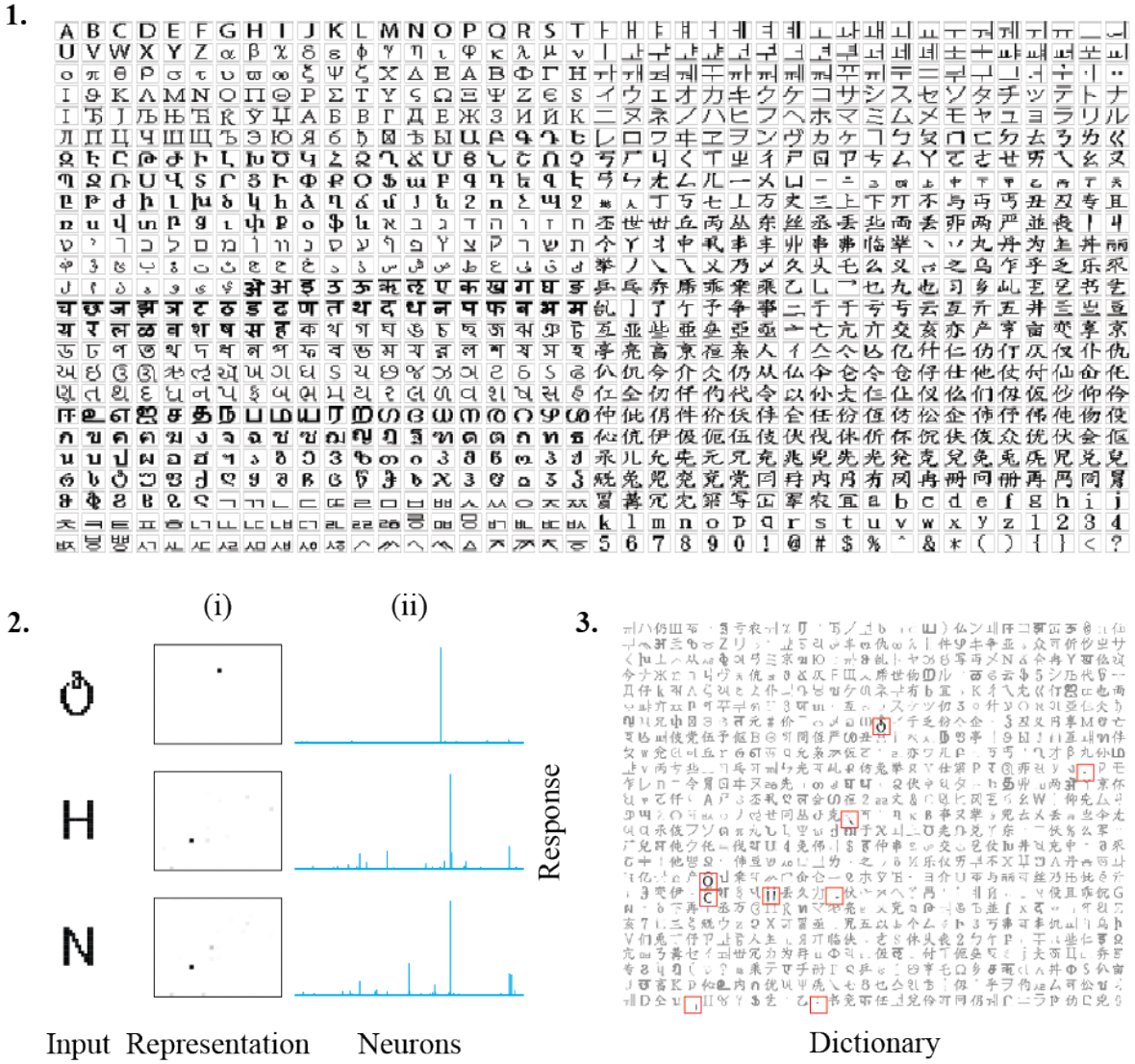


Figure 3. 2: Analysis of symbols

1. *The set of 1000 symbol images (16 pixels by 16 pixels) with binary pixel values that were used as inputs.* **2.** *A set of 800 most informative features of the images and representations of inputs in a basis defined by these features were obtained through nGMCA algorithm. Three example inputs and their representations in image form are illustrated (i). Each pixel in the representation image corresponds to a representation neuron, and the grayscale pixel values in the image correspond to neuronal activity levels. (ii) The population responses of all neurons for these specific inputs are also plotted as bar graphs.* **3.** *The set of 800 obtained informative features or the "dictionary." Each feature corresponds to the tuning property of a representation neuron. Tuning properties of neurons that are active in*

representations of example inputs in panel 2 are highlighted with red boxes. There is a positional correspondence between the tuning property's location in the dictionary and the neuron's location in the representation image.

3.4.1.1. Information content of extracted features

To understand the tuning properties' informativeness about objects, we calculated mutual information between the captured features and objects (see methods for details). However, in *nGMCA* algorithm, convergence to a unique dictionary requires the normalization of columns. In other words, the feature vectors extracted as dictionary columns are required to have unit length. This requirement prevents the extracted features from having a binary nature like the inputs.

To mitigate this difference, we turned to the definition of tuning properties of neurons. For a neuron, tuning corresponds to the structure in the input that elicits a maximal response in it. Responses of a neuron to several different stimuli are recorded, and from those stimuli, the one producing maximum response is selected as its tuning property. However, in our case, the representation process's linear nature eliminated the necessity of using different stimuli. Consider a situation in which one of the neurons, p, has a very high response value compared to all other neurons. In this case, for a given dictionary Φ with unit length columns, the product Φa can be approximated as

$$\Phi a = \sum_{i=1}^{k} \widehat{\Phi_i} a_i \approx \widehat{\Phi_p} a_p$$

where $\widehat{\Phi_i}$ denotes the i^{th} column of Φ. Interestingly, the particular input x_p that might have caused such response in the neurons must satisfy $x_p = \Phi a$ i.e.

$$x_p \approx \widehat{\Phi_p} a_p$$

Since a_p is a scalar denoting the response of the highest active neuron, the input vector x_p will be coincident with the column vector $\widehat{\Phi_p}$. In other words, x_p will predominantly be a very similar structure to the column of the dictionary that corresponds to the most active neuron. Looking back at the definition of tuning of neurons, one can realize that x_p, and hence $\widehat{\Phi_p} a_p$ is the structure that elicits the maximum response in neuron p.

With this understanding, we obtained the binary tuning properties of representation neurons by multiplying the maximum response that they produce for any stimulus with the corresponding column of the dictionary (**Figure 3.3.1**). The product was then transformed into binary values using a Heaviside step function to obtain the closest structure equivalent to the inputs. In mathematical terms, the binary tuning property $\boldsymbol{\psi_p}$ of the neuron p can be given as

$$\boldsymbol{\psi_p} = \mathcal{H}_\alpha\left(\widehat{\Phi_p}\left(\max_i A_{p,i}\right)\right)$$

here $\mathcal{H}_\alpha(.)$ is the Heaviside step function defined as

$$\mathcal{H}_\alpha(x) = 0 \; if \; x < \alpha \quad and \quad \mathcal{H}_\alpha(x) = 1 \; if \; x \geq \alpha$$

and $A_{p,i}$ is the $(p,i)^{th}$ element of the matrix A whose columns are representations of stimuli. It is important to note here that as tuning properties are essentially the inputs causing the maximum response in a neuron, the dimensions of any neuron's tuning property are the same as the input. In other words, the dimensions of a neuron's tuning property equal the number of neurons in the primary layer. The binary tuning properties obtained from the dictionary in **Figure 3.2.3** are shown (**Figure 3.3.2**).

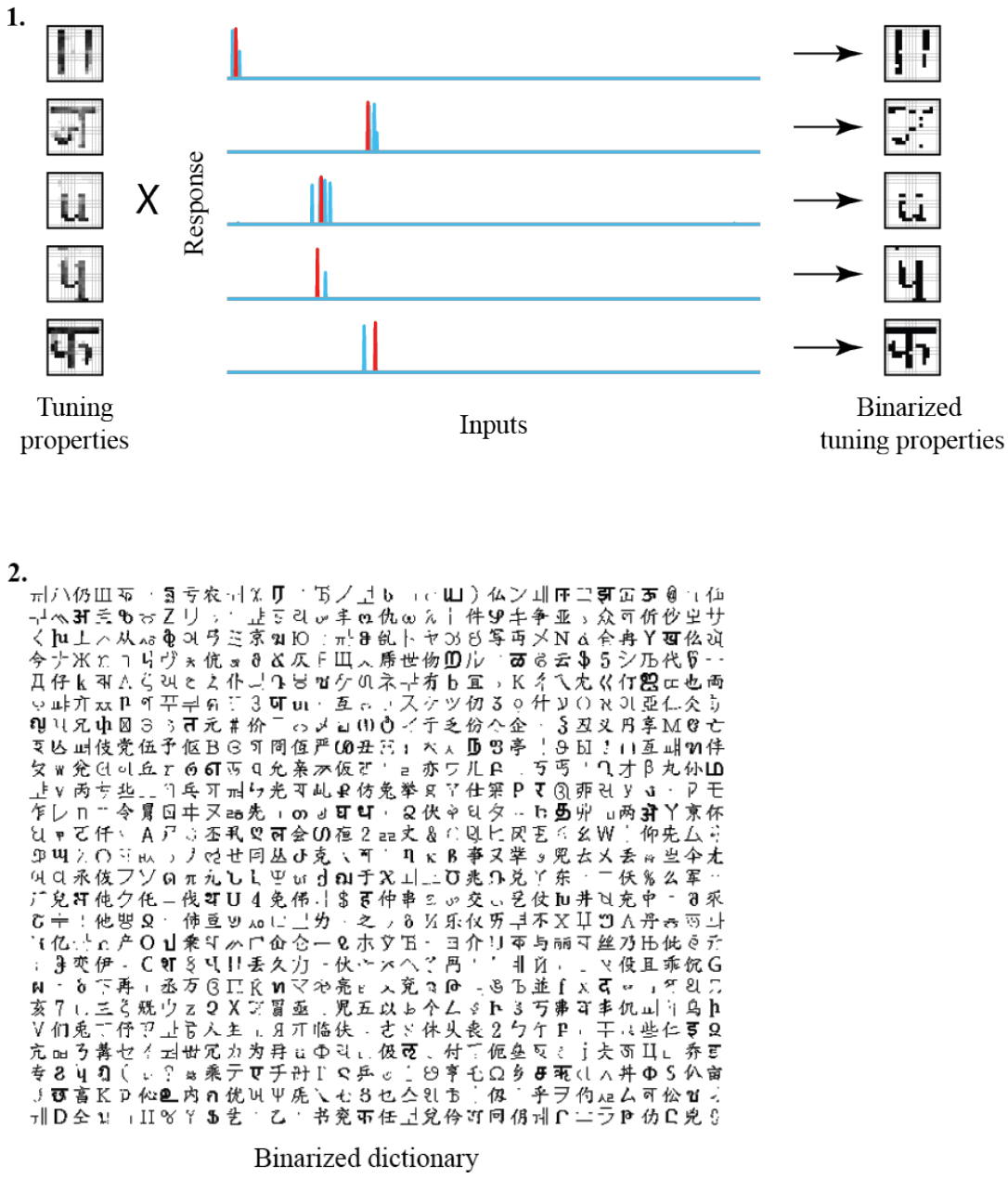


Figure 3. 3: Binarizing tuning properties

1. *Tuning properties of neurons obtained through the non-negative matrix factorization method were converted to binary values to characterize their informativeness. Shown are tuning properties of 5 neurons obtained through the nGMCA algorithm. The gray pixels indicate that the features extracted as tuning properties do not have binary nature like the inputs. To binarize the tuning properties, responses of these neurons to all inputs were obtained. The red peaks indicate the maximum observed responses of neurons across all*

inputs. Maximum responses were multiplied with tuning properties to reconstruct the inputs that caused the maximum response. Reconstructed inputs were binarized using a Heaviside step function to produce binarized tuning properties. **2.** *Binarized form of the dictionary of 800 features shown in* **Figure 3.2.3.**

Once the binary tuning properties were obtained, we calculated each neuron's information content about individual objects (see methods for details). We found that the dictionary captured structures that were most informative about individual objects. However, the common and less informative structures about any object were also captured (**Figure 3.4.1**). We obtained a histogram of the maximum information content of the captured structures. It was found that most of the structures were highly informative about one of the inputs (**Figure 3.4.2**). This analysis demonstrated that the *nGMCA* algorithm could capture highly informative and unique structures from the inputs.

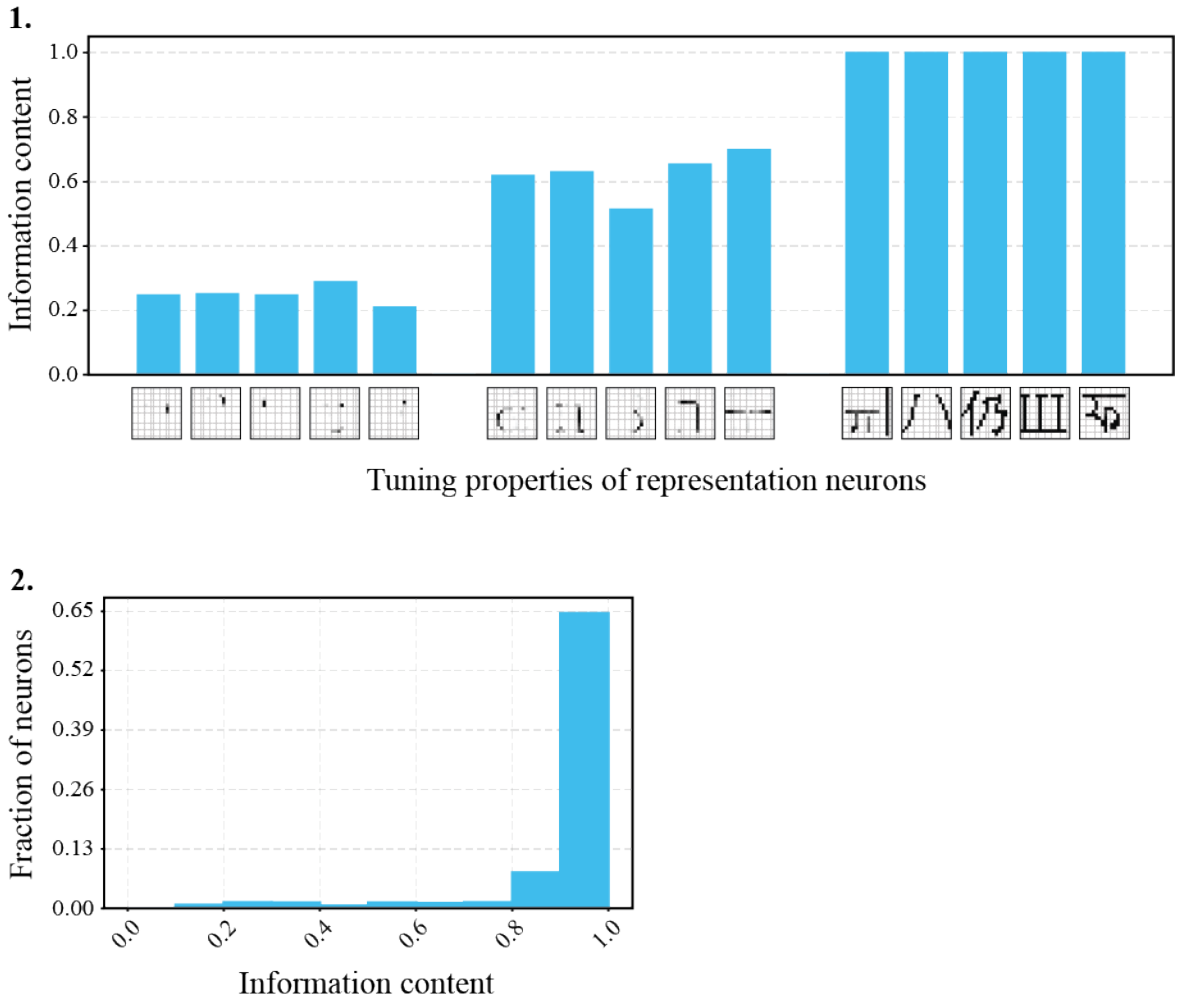


Figure 3. 4: Informativeness of obtained features

1. *After binarization, the features' information content about the inputs was calculated. Each extracted feature contained different information about different inputs. The maximum information content of few features about any input is plotted. Note that very localized features have minimal information about any input, whereas comprehensive features are most informative.* **2.** *Distribution of maximum information content of all features is shown. Most features are highly informative about individual inputs.*

3.4.1.2. Efficiency of representations

In the previous chapter, we discussed that the representation's efficiency could be established by ensuring their sparseness. Therefore, to assess the efficiency of representations obtained under our framework, we characterized their sparseness. We realized that sparseness of representations could be characterized in three different ways

1. **The number of active neurons and total activity:** A direct measure of the sparseness of representations is the number of active neurons in representations. A smaller number of active neurons means more sparseness. This measure corresponds to the l_0 norms of the representations. Sparseness is also measured in terms of the total activity of all the neurons in a representation. Limited total activity is an indication that fewer neurons are active. The total activity corresponds to the l_1 norms of the representation vectors. We utilized both l_0 and l_1 norm to characterize sparseness.

2. **Kurtosis of response distribution:** While measures like the number of active neurons and the total activity in neurons indicate the sparseness of individual representations, the overall sparseness across all representations can be characterized from the distribution of individual neurons' states. If all representations are sparse, individual neurons spend most of their time in inactive states and rarely take higher activity states. Such a response characteristic results in a probability distribution of states that is peaked at zero and falls off with heavy tails at higher activity levels. This *"tailedness"* of distributions is measured in terms of their kurtosis. Kurtosis is the fourth standardized moment of a distribution, which is calculated as

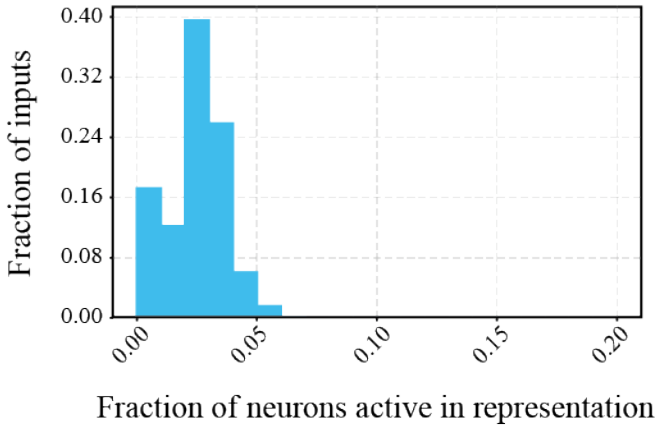
$$K = \frac{\sum_{i=1}^{N}(X - \bar{X})^4/N}{\sigma^4} - 3$$

here $\bar{X}$ is the mean, and σ is the standard deviation of the data.

3. **Correlation:** Correlation among neurons is a combined indicative of overall sparseness and uniqueness of representations. If representations of individual inputs are sparse and non-overlapping, then the correlation among neurons becomes minimal. As the overlap between representations of different inputs increases, or the number of neurons in individual representations increases, the overall correlation between neurons also increases. In this regard, a diagonal correlation matrix indicates that representations are sparse as well as non-overlapping.

We utilized all these three measures to assess sparseness and hence the efficiency of representations. The distribution of the number of active neurons in representations (**Figure 3.5.1**) and total activities of neurons (**Figure 3.5.2**) is shown. The distributions are skewed towards fewer neurons and lesser activity.

1.



2.

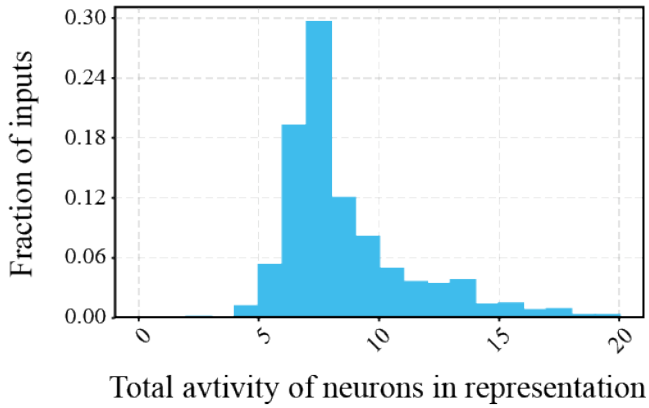


Figure 3. 5: Sparseness of representations

1. *Sparseness of representation was measured in terms of the total number of active neurons in a representation (L0 norm). The plot shows the distribution of normalized L0 norms (fraction of active neurons in a representation). Most representations have less than 5% active neurons (equals 40 neurons out of 800 neurons).* **2.** *The sparseness measured as total activity of all neurons in a representation (L1 norm) is also skewed towards lower values.*

To measure the kurtosis of the distribution of states for individual neurons, we analyzed individual neurons' response profiles across all inputs (**Figure 3.6.1(i)**). From these response profiles, we obtained the histograms of activity states binned over intervals of 0.1 (**Figure 3.6.1(ii)**). The kurtosis of the response distribution was then measured from these

histograms. The distribution of kurtosis values for all neurons is shown (**Figure 3.6.2**). We can see that the kurtosis of neurons is high, indicating high overall sparseness of representations.

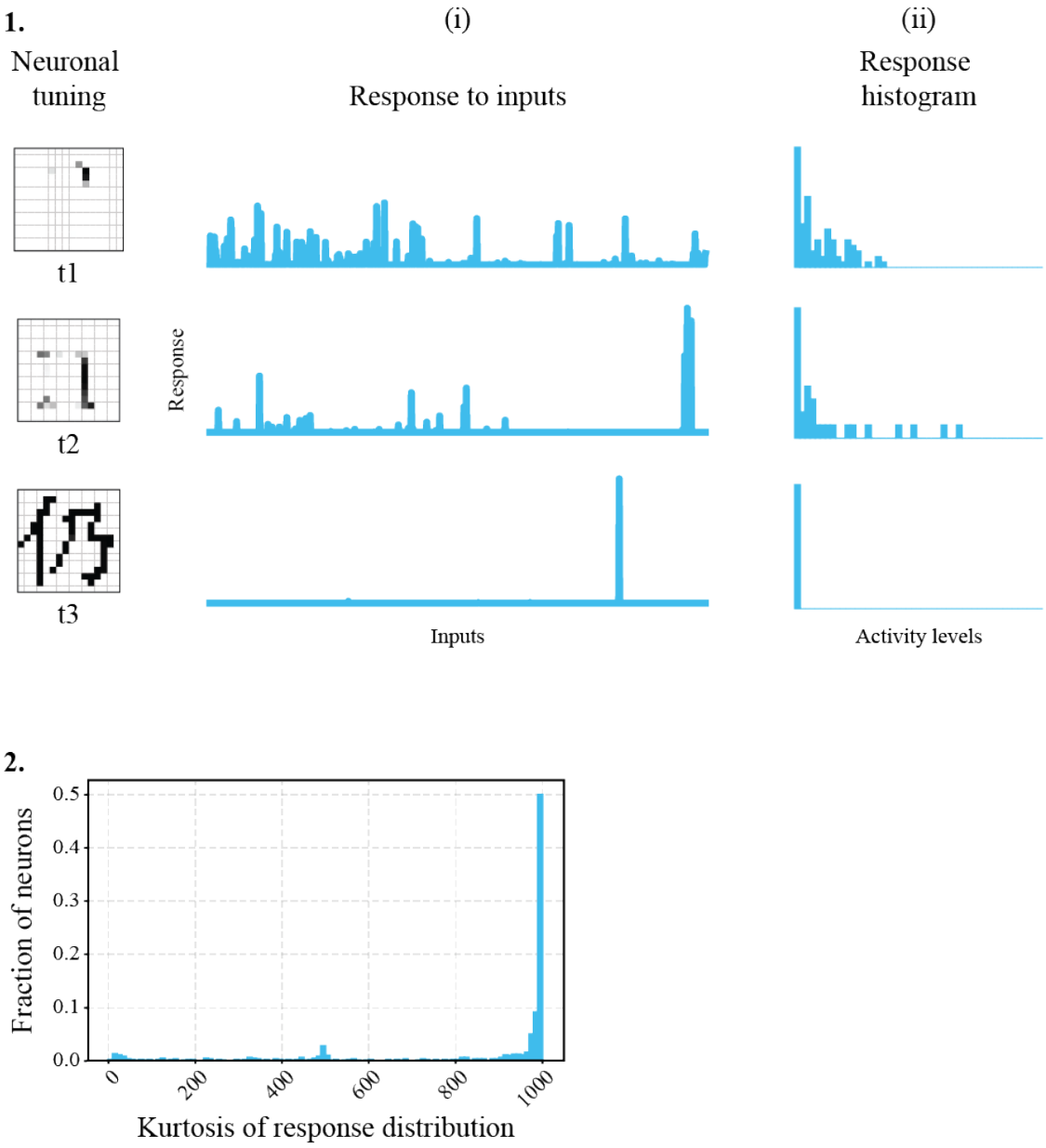


Figure 3. 6: Kurtosis of response distributions

The efficiency of representations was assessed from the kurtosis of response distributions of individual neurons. Response histograms of individual neurons were obtained from their

responses to all inputs, and kurtosis was calculated from each histogram. **1.** *Shown are tuning properties (t1, t2, and t3) of three different neurons. The tuning properties have varying information content about any input, with t1 being least informative and t3 being most informative. Responses of these neurons for all the inputs (i) are shown. From these responses, response histograms (ii) were plotted. Note that the response histogram of the neuron with the least informative tuning (t1) lacks heavy tails, whereas response profiles of more informative neurons have heavy tails.* **2.** *For all neurons, kurtosis was calculated from the response profiles and was plotted as a histogram. Most of the neurons have high kurtosis values indicating a sparse response distribution.*

The correlation among neurons was obtained from the same response profiles, and the correlation matrix was plotted (**Figure 3.7**). The diagonal structure of the correlation matrix indicates that the representations obtained are both sparse and unique.

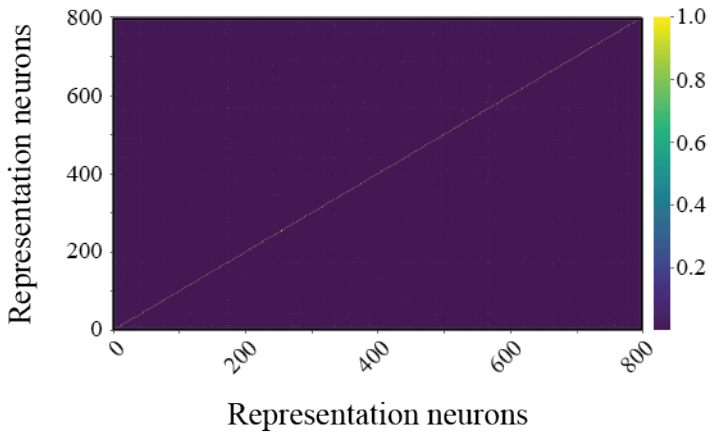


Figure 3. 7: Correlation among neurons

The responses of neurons across all inputs were utilized to obtain pairwise correlations between neurons. The absolute values of pairwise correlation coefficients are plotted in the

form of a heatmap. The diagonal nature of the plot indicates that representations are sparse and non-overlapping

3.4.1.3. Simulations with a fixed number of neurons and a varying number of inputs

As pointed before, there are two free parameters in these simulations, namely the number of inputs and the number of neurons. While the previous results show different aspects of the framework while representing a fixed number of inputs with a fixed number of neurons, we also analyzed how the framework behaved when either the number of inputs or the number of neurons is varied.

To assess the nature of representation with variation in the number of inputs, we fixed the number of representation neurons to be 500. There were two reasons for choosing these many neurons

1. The number was greater than 256, which was the number of pixels (considered primary neurons) in the symbol dataset. In this way, we tried to make the simulations consistent with the observation that the number of neurons in higher-order brain centers is greater than the number of neurons present early in the sensory pathway.
2. The number was less than the total number of distinct inputs in the dataset. Thus, it was made sure that the number of distinct stimuli represented in the system is larger than the number of cells in the system. It is important to note that maintaining this consistency limits the minimum number of inputs that can be considered in the simulations.

The numbers of inputs were varied from 500 to 1000 to understand the system's adaptation to different input numbers (**Figure 3.8.1**). The inputs were chosen at random with replacements from the data set to avoid any sampling bias. The occurrence frequency of each input was considered the same.

First, we analyzed the change in the nature of neurons' tuning properties as the number of inputs increased. Plotting a few neurons' tuning properties, we found that the extracted features were structurally similar to the complete inputs. However, the features got increasingly local as the number of inputs increased (**Figure 3.8.2**).

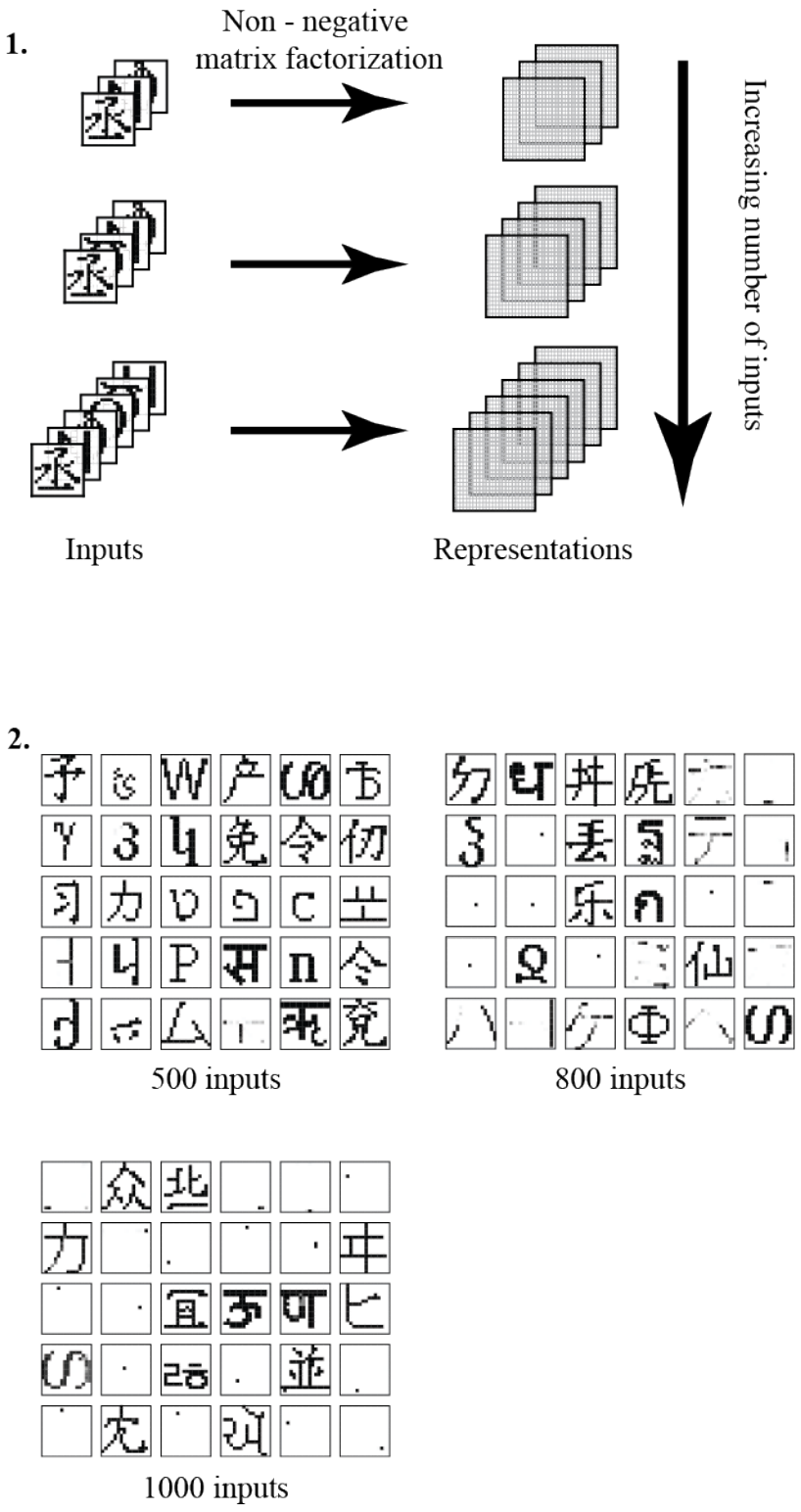


Figure 3. 8: Effects of variation in the number of inputs

1. *To assess changes in representations with the number of inputs, the number of inputs was varied from 500 to 1000 (depicted in the figure by an increasing number of inputs across multiple trials) while keeping the number of neurons fixed at 500 (depicted in the figure by*

the constant size of all representation images). **2.** *Examples of tuning properties of neurons when 500, 800, and 1000 inputs were represented. Note that as more inputs were represented, the tuning properties of neurons got more localized.*

We compared the binarized tunings of neurons to the inputs to estimate the uniqueness of the tuning properties. The tuning properties of neurons were matched in a component-wise sense to the inputs that elicited maximum responses. The match is plotted as a fraction of the input dimensions (**Figure 3.9.1**). We find that when the number of inputs is low, the structure of the input eliciting a maximum response in any neuron matches completely with its tuning property. In other words, the complete structure of the input is captured (greater than 98% on average). However, when a larger number of inputs are represented, the average matching gets lowered (around 95% average match), indicating that only partial input structures are captured.

To further test the commonality of tuning properties among the inputs, we matched the structure of all the inputs to the neurons' tuning properties in a component-wise manner. A match was considered when an input incorporated more than 90% of the tuning property's structure. We found that, while representing a lower number of inputs, only specific inputs matched any tuning property (less than 10% of inputs matched any tuning), indicating that the neurons were tuned to specific inputs. The tunings' uniqueness declined as the number of represented inputs increased (more than 12% of inputs matched any tuning) (**Figure 3.9.2**). To further quantify the number of neurons whose tuning properties were not unique, we plotted the number of neurons whose tuning properties were detected in more than 10% of the inputs. This number increased with an increasing number of inputs (**Figure 3.9.3**).

We wondered if the neurons tuned to many inputs are the same that captured the partial structures. An alternative possibility is that the neurons capture the complete input structures, but the inputs themselves have a considerable degree of similarity. To test these,

we examined the set of neurons that captured less than 90% of the input structure and the set of neurons whose tuning properties were detected in more than 10% of the inputs. The ratio of the two sets' intersection with the later set was plotted (**Figure 3.9.4**). We found that in all conditions except for 700 inputs, more than 60% of the neurons with more commonly detected tuning properties captured inputs' partial structures. In the case of 700 inputs, the lower number could be attributed to the overall lower number of partial structure-capturing neurons. These findings indicated that as the number of inputs increased, some of the neurons got tuned to partial input structures common among several inputs.

As neurons' tuning properties have the same dimensions as the inputs, one can compare the two to assess the tuning properties' overall structure. We measured pairwise correlations between the components of input vectors and tuning property vectors. The Frobenius norm of the differences in the two correlations was obtained (**Figure 3.9.5**). As expected from the previous analysis, the norm increased with the number of inputs, indicating that the tuning properties' overall structure deviated from the overall input structure. This deviation could be attributed to the partial structures of the tuning properties.

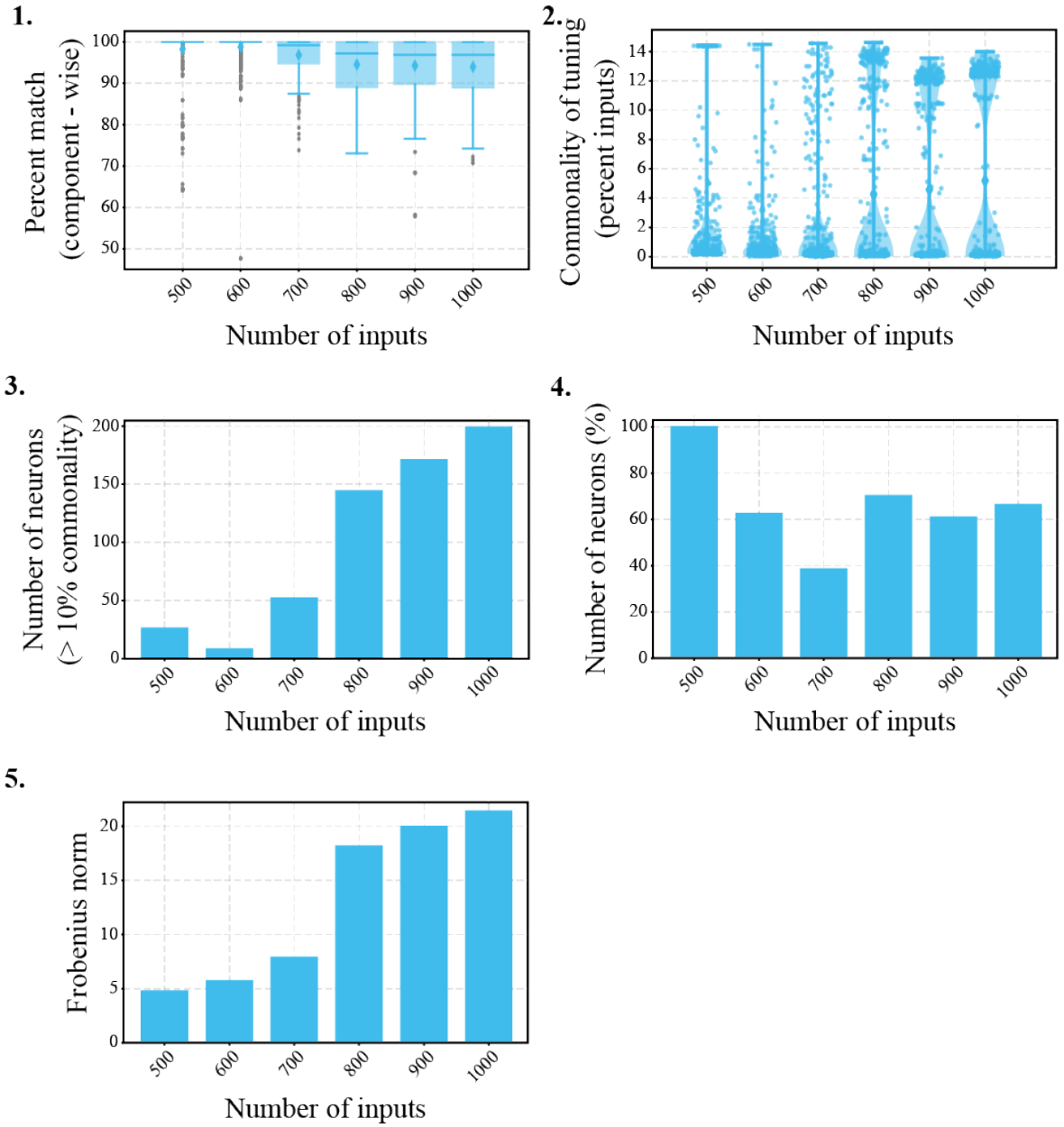


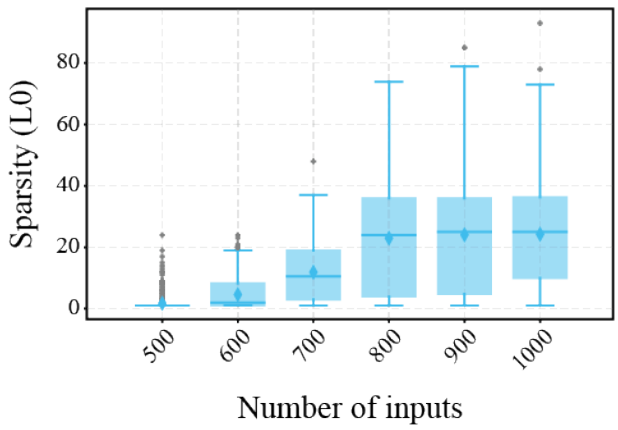
Figure 3. 9: Analysis of uniqueness of tuning properties of neurons while varying number of inputs

1. *The binarized tuning properties were compared in a component-wise manner, with the inputs causing the maximum response. The match percent decreased on average as the number of inputs increased.* **2.** *The commonality of a neuron's tuning property was determined based on the frequency with which it was encountered in inputs. A feature was considered present in input if more than 90% of its structure matched some input portion. As the number of represented inputs increased, more common features were captured as*

tuning properties. **3.** *The number of neurons with commonly occurring tuning properties (greater than 10% commonality) increased when more inputs were represented.* **4.** *To test if the partial tuning properties were more common, the number of neurons with more common tuning properties (more than 10% commonality) was plotted as a fraction of the neurons tuned to partial input structures (tuning property matched less than 90% to input). In most cases, the fraction was more than 60%.* **5.** *To compare the overall structure of inputs with the tuning properties' overall structure, the correlation between input vectors' components was compared with the correlation among tuning property vectors' components. The Frobenius norm of the difference in two correlations indicated that the tuning properties' overall structure deviated as the number of represented inputs increased.*

Next, we characterized the efficiency of representations using the three measures described previously. As before, the sparsity of representations was measured in two different ways, namely, L0 norm and L1 norm. Plots of both the sparsity measures are shown (**Figure 3.10.1**; **Figure 3.10.2**). As expected from the decrease in uniqueness of the captured features, we find that both measures increase with the increasing number of inputs; hence, the sparsity decreases.

1.



2.

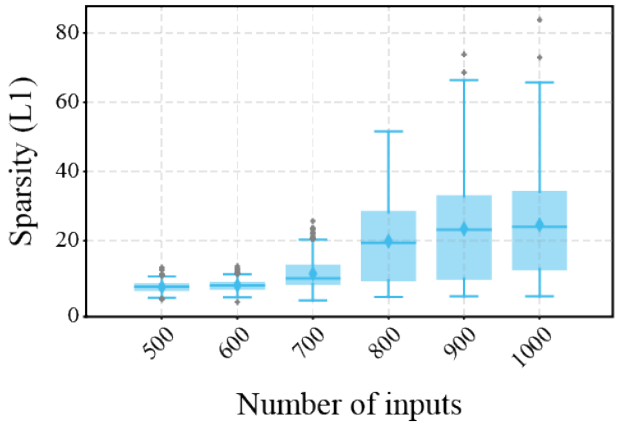


Figure 3. 10: Analysis of sparsity of representations

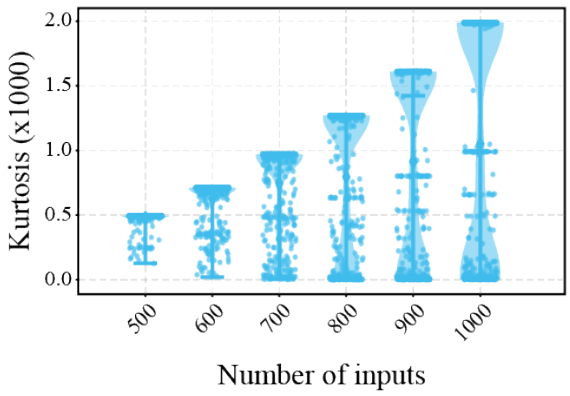
1. *The sparsity of representations measured as L0 norm. An increase in the L0 norm indicated that many neurons were active in representing any input.* **2.** *Sparsity of representations measured as L1 norm.*

The distribution of kurtosis values is shown (**Figure 3.11.1**). It is evident that as the number of inputs increase, the kurtosis of neurons increases on average. This increase is expected because, with increasing input numbers, the chances of a neuron being in a non-active state also increases. However, the distribution of kurtosis values becomes increasingly bimodal when the number of inputs becomes larger, i.e., some of the neurons have very high kurtosis values, while others have very low values. The cumulative fraction of neurons at

different kurtosis values was determined to better visualize this change. It was observed that, in cases where a higher number of inputs were represented, 15 to 20 percent of neurons had kurtosis values less than 10.

In contrast, while representing fewer inputs, this percent was significantly low (< 3%) (**Figure 3.11.2**). Again, these trends can be explained with the commonality of tuning properties of neurons. As more inputs are represented, an increasing number of neurons get tuned to common features. A neuron tuned to more common features is expected to be more active, and therefore, the *"tailedness"* of its response distribution is expected to be small.

1.



2.

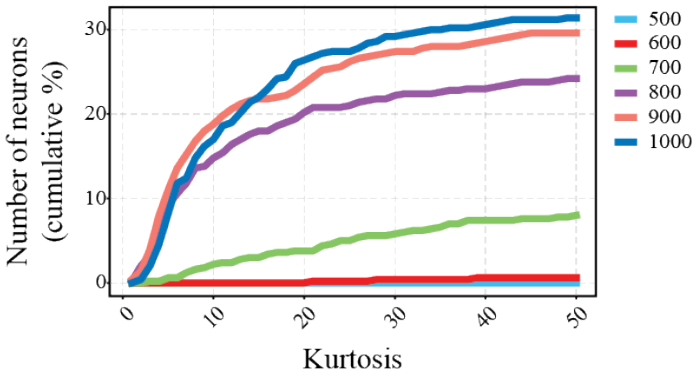


Figure 3. 11: Analysis of kurtosis of neuronal response profile

1. *Kurtosis of response distributions of neurons when different numbers of inputs were represented. A higher kurtosis corresponded to more sparsely active neurons. Note that as*

the number of inputs increased, the kurtosis of response profiles of some of the neurons dropped. **2.** *The plot of the cumulative number of neurons having kurtosis of response profiles below 50 indicated that as many as 20% of neurons had response distributions with kurtosis below the value of 10 when the number of represented inputs was more than 700.*

We also obtained a pairwise correlation between all neurons in all different representation circumstances. The deviation of a correlation matrix from an identity matrix was measured as the Frobenius norm of their difference. Consistent with the previous sparseness measures, we found that when the number of inputs was small, the correlation matrices were closer to the identity matrix. The difference increases with the number of inputs (**Figure 3.12)**.

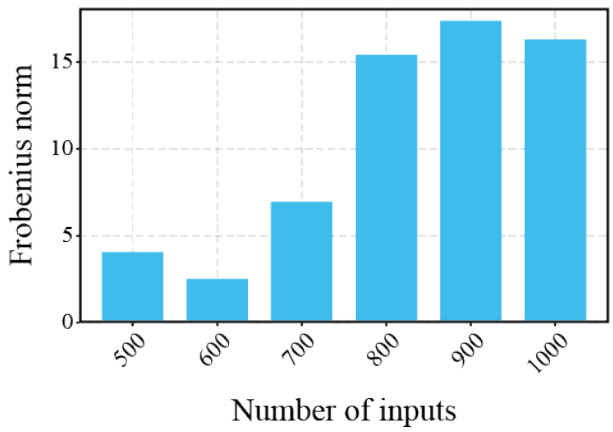


Figure 3. 12: Analysis of response correlation

Matrices of pairwise correlation coefficients of neuronal response profiles were obtained for each representation scenario with varying inputs. Frobenius norm of the difference between the correlation matrix and identity matrix was used to quantify its difference from a diagonal matrix. A lower difference indicated that neurons' response profiles were uncorrelated; therefore, input representations were sparse and non-overlapping.

Previous results indicated that using the most informative structures as a basis for representations results in an efficient representation of inputs. As the number of represented inputs increases, more neurons get tuned to common localized structures. We can further establish the efficiency of representation by comparing the entropy of the stimulus space with the entropy of representation neurons. The entropy associated with an ensemble of N inputs, which are equally likely or uniformly distributed, is given by

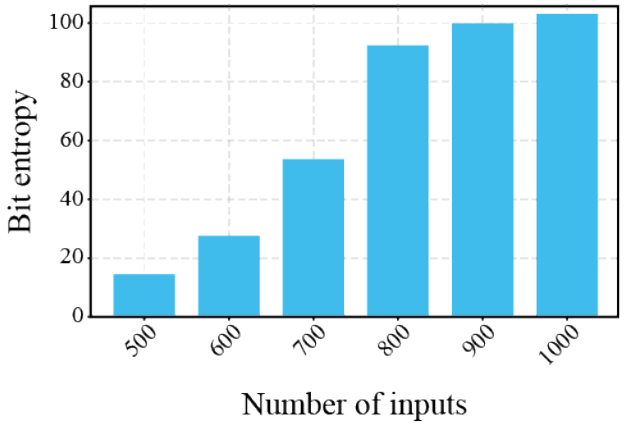
$$H_x = \sum_{i=1}^{N} -\frac{1}{N}\log\frac{1}{N} = \log N$$

The bit entropies of the representation neurons, H_a, were calculated by converting their response to a binary form using a Heaviside function (see methods) (**Figure 3.13.1**). We compared both these entropies and measured the redundancy in representation (**Figure 3.13.2**) R as

$$R = 1 - \frac{H_x}{H_a}$$

As expected from the analysis before, the redundancy was minimal when the number of inputs being represented was low and increased with the number of inputs. This reduction indicated a decrease in representation efficiency.

1.



2.

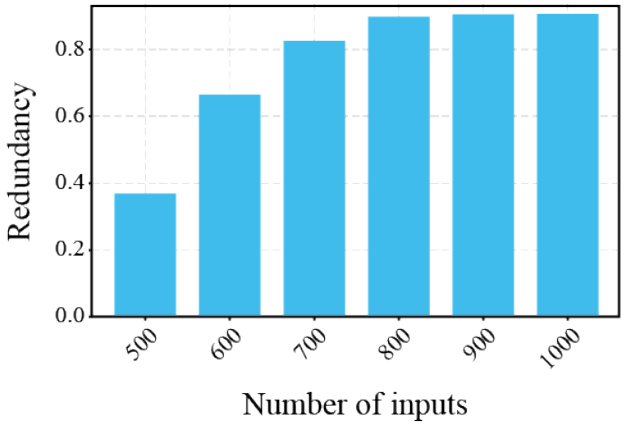


Figure 3. 13: Representation redundancy with a varying number of inputs

1. *The bit entropy of the neurons was calculated from their probabilities of being active. A lower bit entropy indicated a lower overall probability of activation. The bit entropy increased with the number of represented inputs* **2.** *From bit entropies, redundancy in representation was also determined. The redundancy increased with the number of represented inputs indicating lesser representation efficiency.*

Overall, from these simulations, it was found that, while representing a varying number of inputs using a fixed number of neurons, the uniqueness of captured features decreases with the number of inputs. Neurons get tuned to common structures when more inputs need to be represented. The efficiency of representation also varies with the

uniqueness of captured features. While representing fewer inputs, the efficiency remains maximal, but it decreases as the number of inputs that need to be represented increases. This decrease in efficiency can be attributed to a shift in the tuning properties of the representation neurons. As neurons get tuned to more common structures, their response profiles become less sparse, and hence the representations become less efficient.

3.4.1.4. Simulations with a fixed number of symbols and a varying number of neurons

In the next set of simulations, we kept the numbers of inputs fixed and varied the number of representation neurons (**Figure 3.14.1**). Following the observed relation between the number of inputs and the number of neurons, 1000 inputs were represented while varying the number of neurons from 500 to 1000. Like before, we analyzed the changes in neurons' tuning properties as more neurons were employed in representing the same number of inputs (**Figure 3.14.2**). We found that the neurons' tuning properties got more specific and less local as more neurons were utilized.

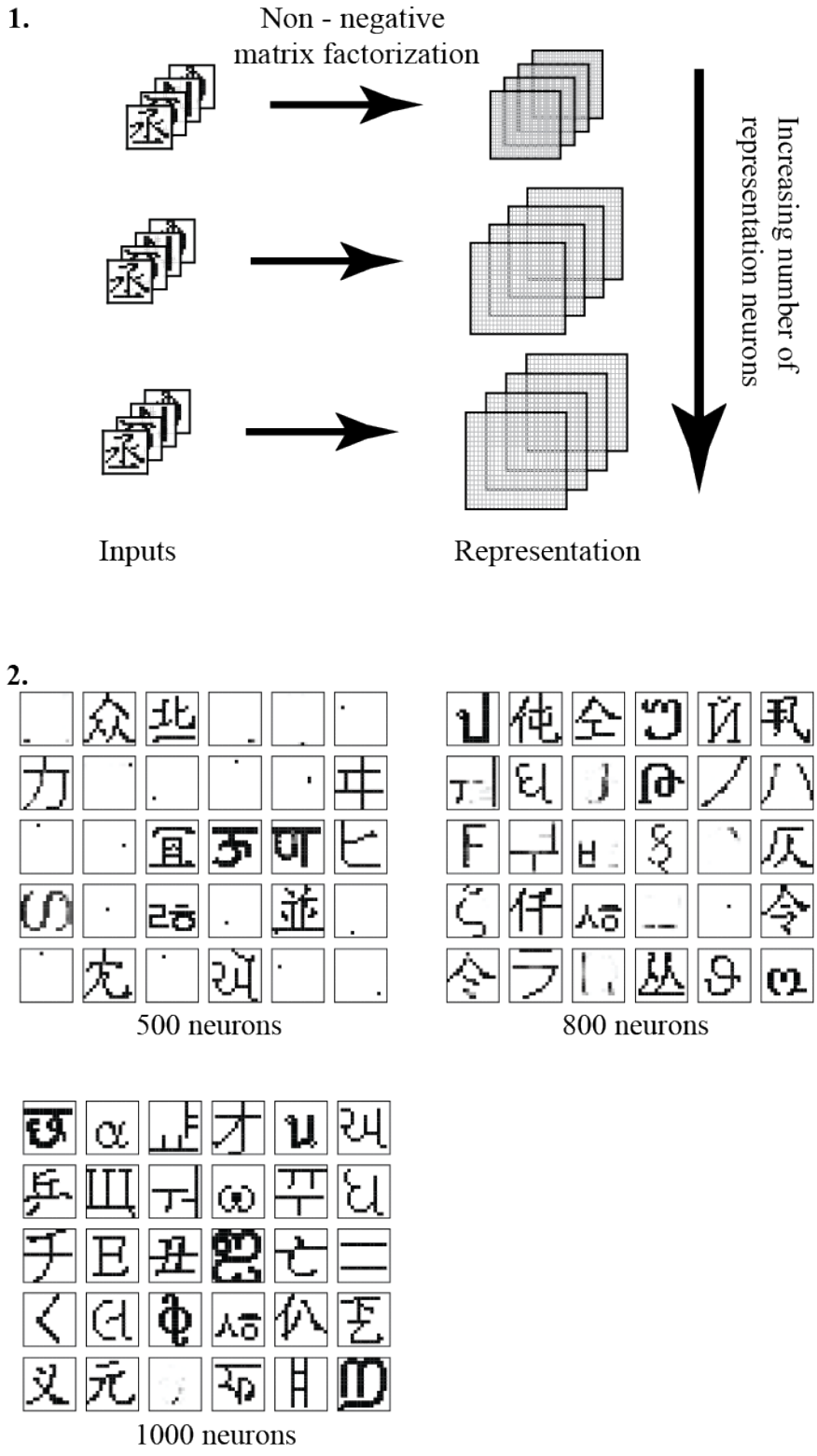


Figure 3. 14: Effects of change in the number of neurons on representation

1. *To assess changes in representations with the number of neurons, the number of neurons was varied from 500 to 1000 (depicted in figure by the increasing size of representation images) while keeping the number of inputs fixed at 1000 (depicted in the figure by the equal*

number of inputs across multiple trials). **2.** *Examples of tuning properties of neurons when 500, 800, and 1000 neurons were utilized in representation. Note that as more neurons were employed, the tuning properties of neurons became more comprehensive.*

To further confirm this, binarized tuning properties of all the neurons were obtained using the previously described approach and compared with inputs. First, individual neurons' tuning properties were matched in a component-wise sense to the inputs that elicited the maximum response. The average percent match increased with an increasing number of neurons. In particular, the percent match increased from 90% at 500 neurons to 95% at 800 neurons and reached almost 100% at 1000 neurons (**Figure 3.15.1**). Next, the commonality of the tuning properties in the input set was assessed. As expected, the commonality of a 500-neuron system's tuning properties was 12 – 13%, and the commonality for an 800-neuron system's tuning feature was 2 – 3%. The decrease can be attributed to the comprehensive structure of the tuning properties (**Figure 3.15.2**). The decline in the number of neurons with more than 10% commonality also confirmed the fact that the tuning properties were getting unique with the increase in the number of neurons (**Figure 3.15.3**). Finally, the tuning properties' overall structure was compared to inputs' overall structure in terms of component correlations. As before, the Frobenius norm of the correlation matrices' difference was plotted. The norm decreased to minimal values after 800 neurons were utilized (**Figure 3.15.4**), confirming the overall structural similarity of tuning properties with the input set.

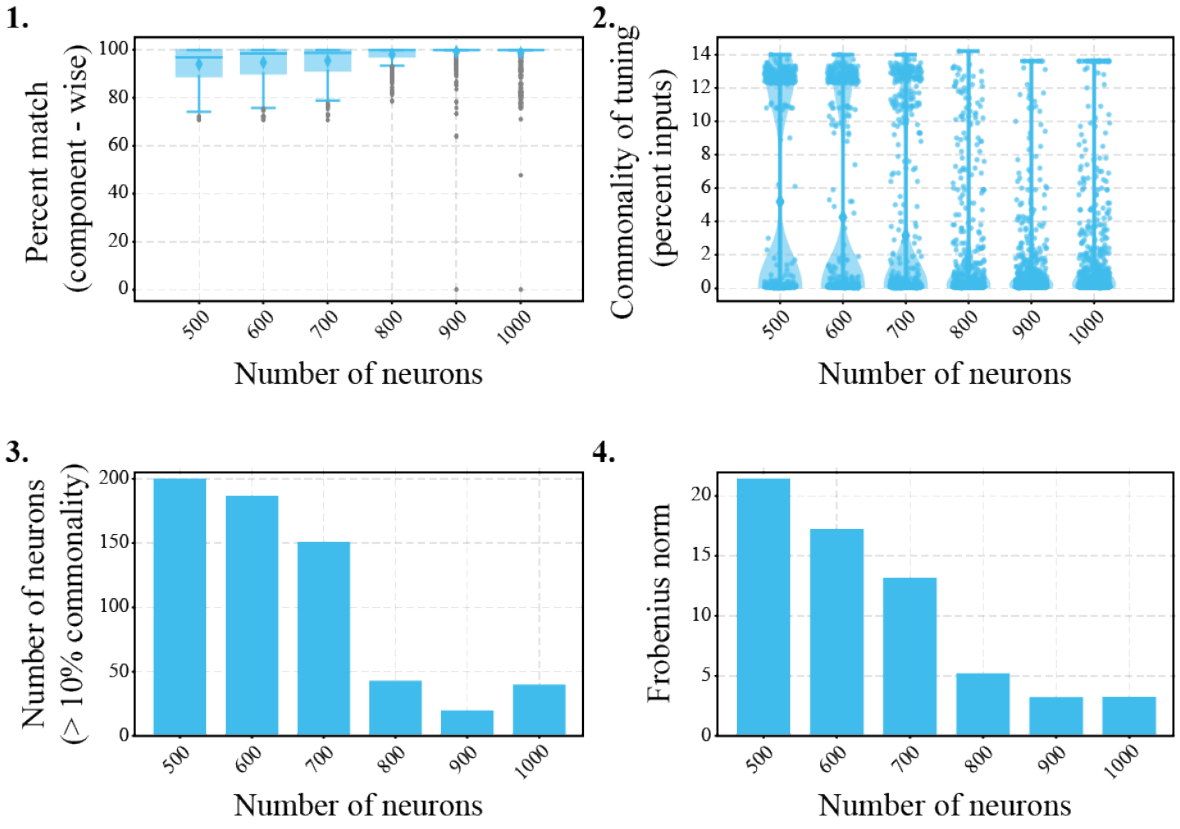


Figure 3. 15: Analysis of uniqueness of tuning properties of neurons while varying the number of neurons

1. *As before, the neurons' binarized tuning properties were compared with the inputs causing the maximum response in a component-wise manner. The match percent increased on average as the number of neurons increased.* **2.** *The commonality of tuning properties was determined in terms of the fraction of inputs that incorporated more than 90% of the tuning property's structure. It was observed that with the increasing number of representation neurons, each neuron got gradually tuned to less common, more unique structures from the input set.* **3.** *The number of neurons with more than 10% common tuning decreased with the total number of neurons.* **4.** *To compare the overall structure of inputs with the tuning properties' overall structure, the correlation between components of the input set and tuning property set was measured. The Frobenius norm of the difference of two matrices indicated*

that the tuning properties' overall structure matched more to the overall structure of the inputs as the number of neurons increased.

Further, we analyzed the efficiency in representation by comparing different measures of the sparseness of representations. We found that both the L0 and L1 norms of the representations decreased with the increasing number of neurons. The average L0 norm of the representations took a value close to one at high neuron numbers, which meant that only one neuron was involved in representing a single input (**Figure 3.16.1**; **Figure 3.16.2**).

Similarly, the changes in kurtosis of response profiles of neurons were also striking. Previously, representing 1000 inputs using 500 neurons had resulted in lower kurtosis values of the response profiles. However, increasing the number of neurons from 500 lead to an overall increase in the kurtosis values (**Figure 3.16.3**). The cumulative fraction of neurons with small kurtosis values was plotted. In contrast to the previous analysis, the fraction of neurons having response distribution with low kurtosis values dropped with increasing neurons. The fraction of neurons whose response had kurtosis less than 15 dropped from 30% at 500 neurons to less than 3% at 800 neurons and 0% at 1000 neurons (**Figure 3.16.4**).

Corresponding trends were also observed in the correlation among representation neurons. The correlation decreased as we increased the dimensions of representation. The trend was indicated by the Frobenius norm of the difference between the correlation matrices of the representation neurons and the corresponding identity matrices (**Figure 3.16.5**).

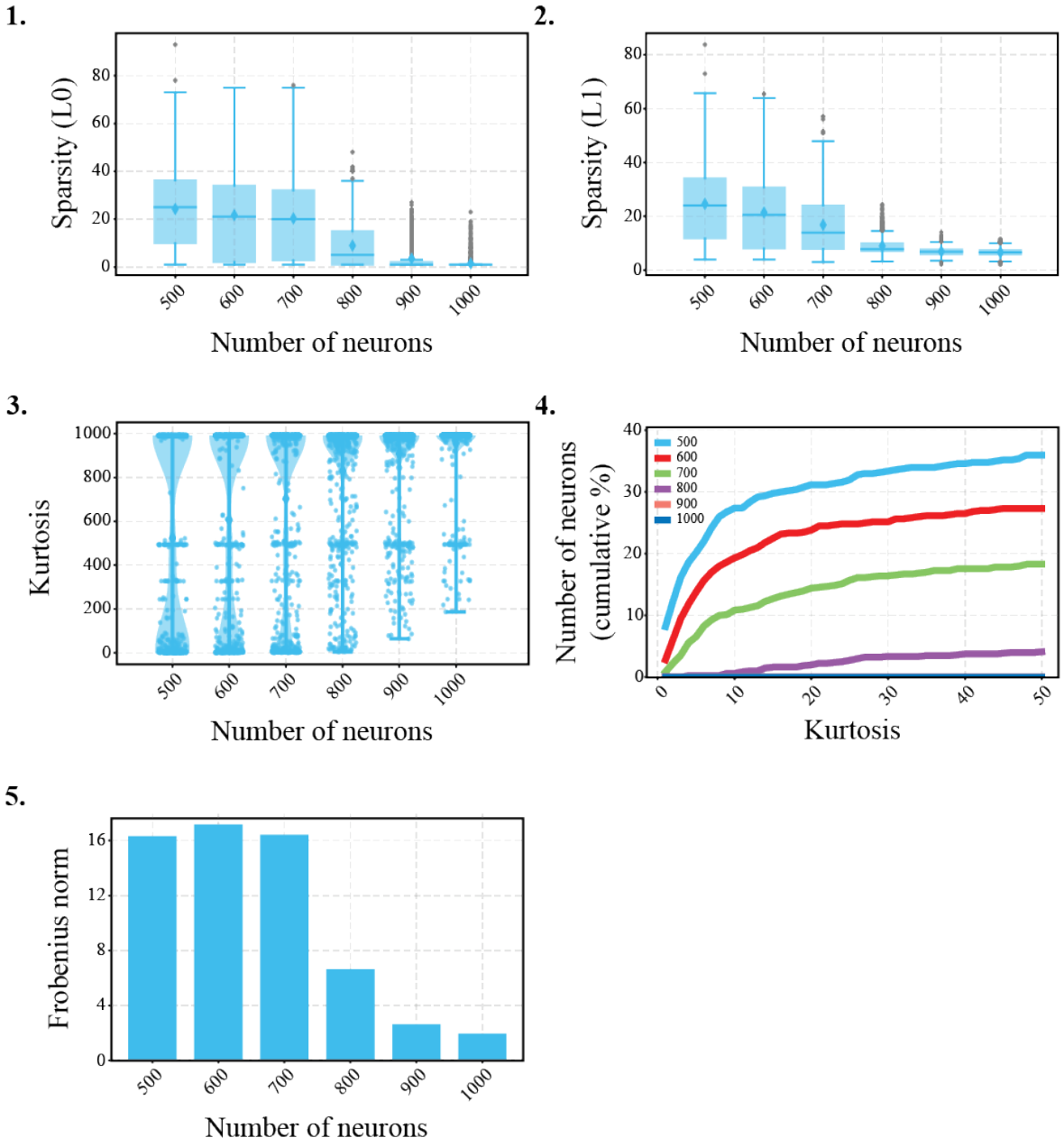


Figure 3. 16: Analysis of representation efficiency with varying number of neurons

1. *As the number of neurons involved in representation increased, the sparsity of representation measured as the L0 norm decreased.* **2.** *The trend was the same for sparsity measured in terms of L1 norm.* **3.** *The kurtosis of response profiles of neurons was measured. The overall kurtosis increased with the number of neurons.* **4.** *Cumulative fraction of neurons with response distributions having kurtosis below 50 also decreased with the increasing number of neurons.* **5.** *The correlation between responses of neurons in different*

conditions of representation was determined. The Frobenius norm of the response correlation matrix and identity matrix difference was calculated. The norm decreased with an increasing number of neurons, indicating that the neurons got decorrelated as the number of neurons increased.

Combined with the previous sets of results, these results indicate that unique structures from inputs can be captured by employing more neurons to represent more inputs. They show that the number of neurons relative to the number of inputs plays a crucial role in determining the uniqueness and, hence, the captured structures' informativeness. If the number of neurons is comparable to the number of inputs, more informative structures can be captured. Informativeness decreases as the number of inputs grow relative to the number of representation neurons. The representation efficiency follows the same trend as uniqueness. Efficiency in representations decreases as captured structures become less specific and more common.

We also calculated the redundancy among representation neurons by comparing their bit entropies (**Figure 3.17**) to the entropy of uniformly distributed stimuli. Consistent with the previous simulations, the redundancy decreased with the increasing number of neurons.

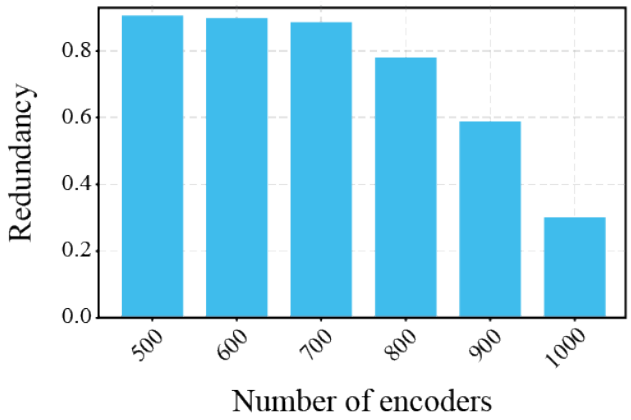


Figure 3. 17: Representation redundancy with a varying number of neurons

As before, redundancy in representation was calculated by comparing the total entropy of neurons with the object ensemble's entropy. Redundancy among neurons decreased as more neurons were used in representing objects.

3.4.2. Relationship between mutual information and response values of neurons

So far, in different analyses like calculating redundancies and bit entropies, we have transformed neurons' responses to binary values. However, neither in actual biological systems nor in our framework, the response values are binary. It is expected that the response values of the neurons are meaningful to the system in some way. Moreover, one of the reasons to have the non-negativity constrained in our framework was to impart a meaning to the neurons' response. Therefore, a proper understanding of the meaning of the response values was necessary. Two different interpretations of response values were possible in this regard

1. **The neuron's response value corresponded to its tuning property's similarity to the input structure.** With this interpretation, the response values would only indicate the

presence or absence of a particular input structure. It would not be, in any way, an indication of any form of inference that the system could draw.

2. **The response value indicated the mutual information between the tuning property and the input.** Under this interpretation, the response value would suggest the degree of inference of the input's identity. It would mean a form of confidence that the system has in identifying an input. This form of information would be beneficial in higher-order cognitive functions like recognition.

To find which of the two interpretations were valid, we first calculated the similarity between individual neurons' tuning properties and the inputs (**Figure 3.18.1**). We then measured the correlation of the similarity with the response values of neurons (**Figure 3.18.2**). We find that the responses neurons to various inputs and similarity of their tuning property to these inputs are not correlated. Next, we calculated the mutual information between the tuning properties and the inputs (see methods) (**Figure 3.18.1**) and again performed the correlation analysis (**Figure 3.18.2**). Interestingly, the response values showed a much higher degree of correlation with the mutual information, supporting the second interpretation.

Interestingly, though the similarities between the inputs and the neurons' tuning properties do not directly resolve the input identities, the system can utilize these similarity values to draw inference about the input. However, drawing such inference will require the system to know the distribution of similarities between all represented inputs and all neurons' tuning properties. This knowledge is hard to get and difficult to store. Thus, with a direct correspondence between the mutual information values and the neuronal activity, the presented framework not only removes the need for drawing inferences but also allows the system to determine the identity of objects without storing all information about all objects.

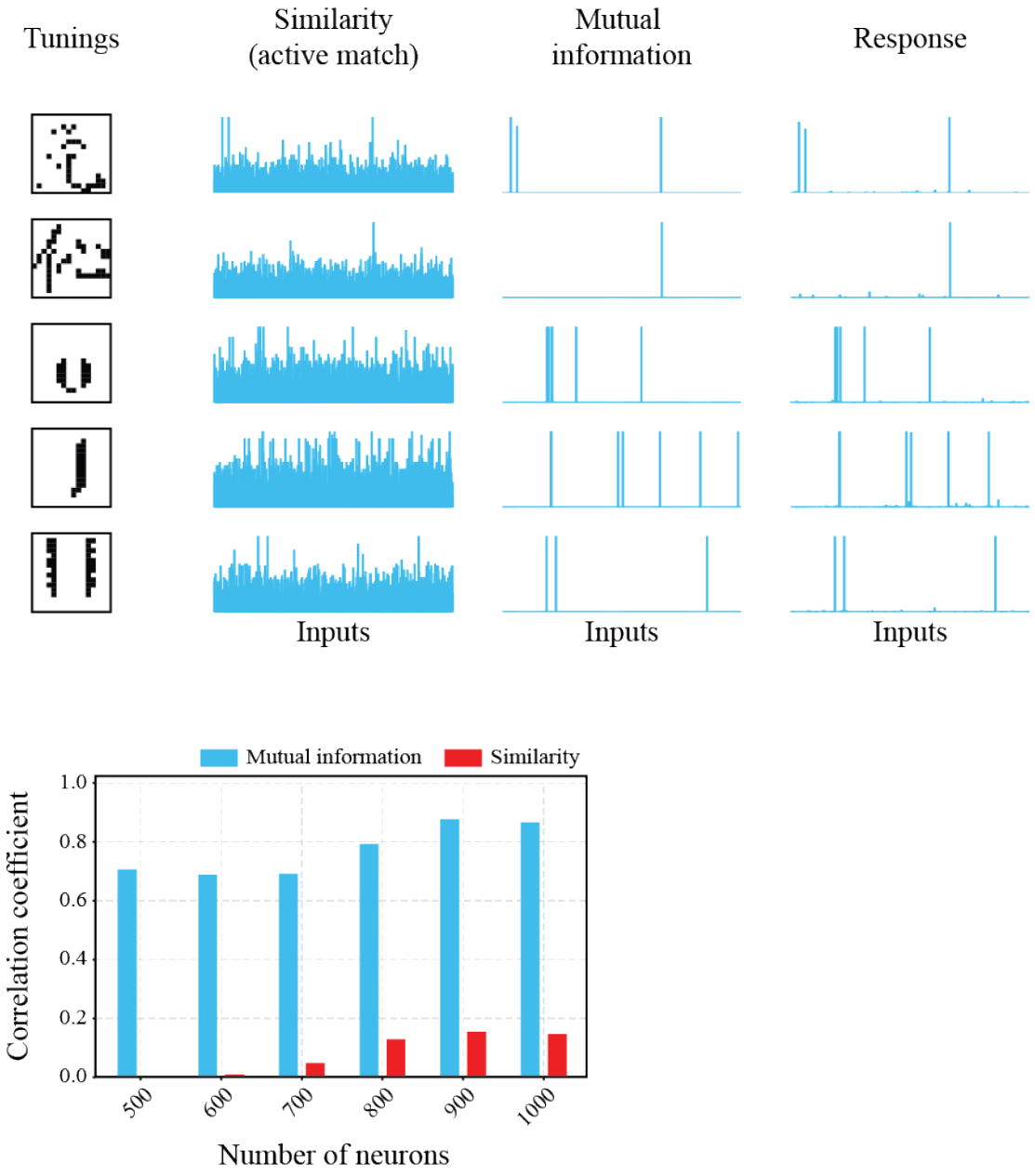


Figure 3. 18: Information-theoretic characterization of tuning properties

1. *To understand the meaning of neurons' response levels, two quantities were considered – the similarity of the tuning with the inputs and the mutual information between the tuning and the inputs. A few of the tuning properties are shown in the left column. Their similarity to inputs, the mutual information between them and the inputs, and their response to different inputs are plotted.* **2.** *The similarity of the tuning with the inputs did not correlate with the response value, but the mutual information between the input and the tunings correlated*

strongly, suggesting that the response values of neurons were indicative of the mutual information between the input and the tunings.

3.4.3. Consistency in representing sensory inputs

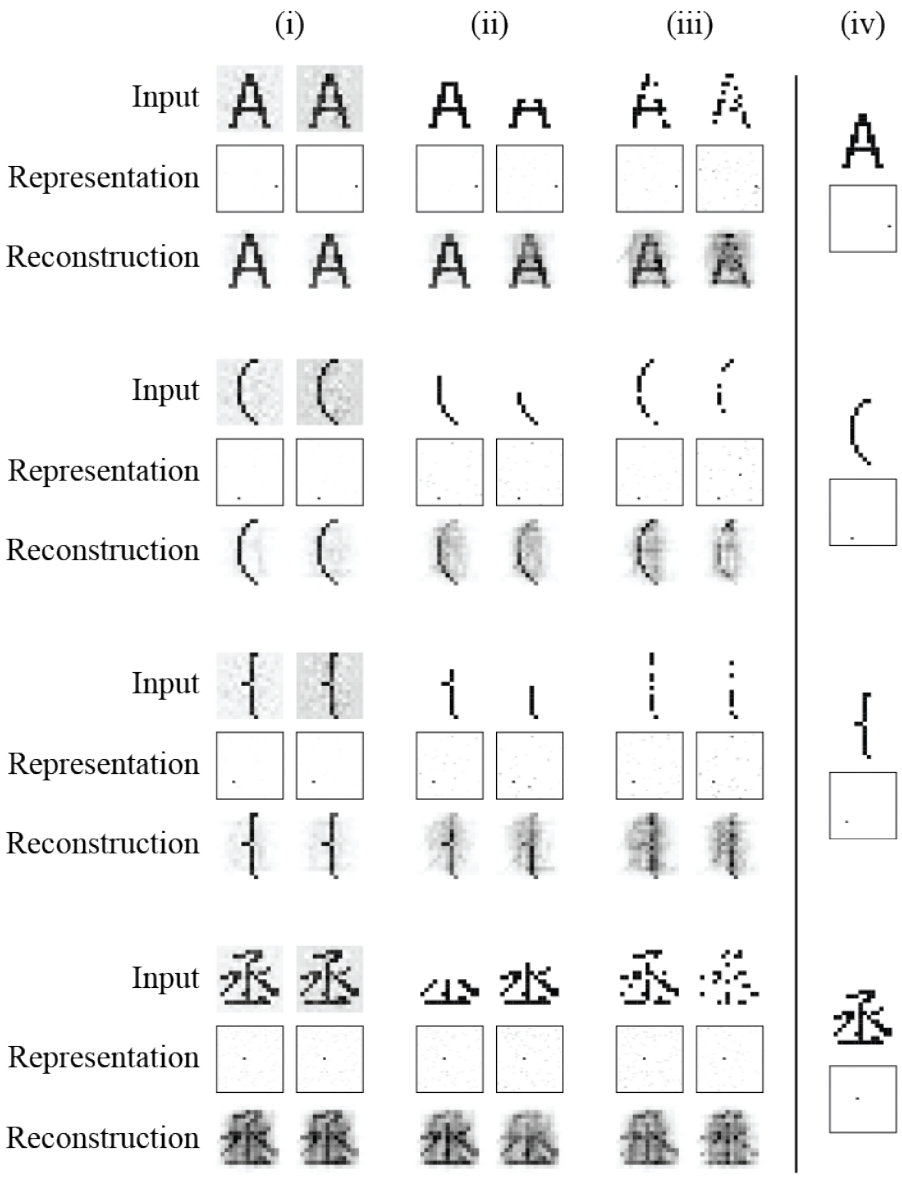
So far, we have seen how the informativeness of captured features is affected by the number of neurons relative to the inputs. We have also analyzed the efficiency of representations in these circumstances and established the meaning of individual neurons' response levels. However, a crucial aspect of representations that still needs to be analyzed is their consistency. In a process like recognition, the inference that the system makes about an input should not change much with certain input corruption or transformation. Hence, the representation should remain consistent. We estimated the consistency in representations when inputs were corrupted in different ways (see methods). The representations of the corrupted input were obtained using the sparse recovery approach (see methods).

As noted, the efficiency of representations depended on the number of inputs relative to the neurons. We tested the consistency while representing 1000 symbols using a varying number of neurons (500, 800, and 1000). We found that remarkable consistency in representing inputs could be achieved when the representations were most efficient, i.e., when comprehensive structures unique to the inputs were captured. Without adapting the system to different forms of corruption, we found that inputs corrupted by Gaussian noise (**Figure 3.19.1.i**), missing an extended (**Figure 3.19.1.ii**), or randomly silenced components (**Figure 3.19.1.iii**) produced representations that were nearly identical to those of uncorrupted inputs (**Figure 3.19.1.iv**). Reconstruction of the inputs using the dictionary and neuronal responses resembled the entire inputs rather than its parts (**Figure 3.19.1**). To quantify the degree of consistency of responses, we calculated pairwise cosine similarity

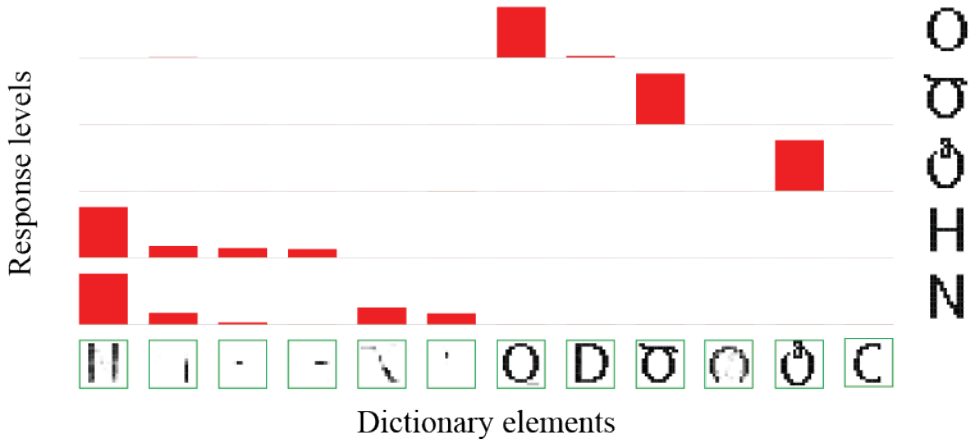
between representations of corrupted inputs and all other inputs. The similarity values were z-scored to estimate how different a given similarity value is from the average observed similarity. If a z-scored similarity value was closer to zero, it meant that the pair of representations were as similar to each other as similar was any pair on average. A higher z-score value, on the other hand, meant that the representation pair had a similarity value that was not commonly observed. In other words, the recovered representation of the corrupted input was particularly similar to a specific input, thus making z-scored similarity a measure of the specificity of representations. We found high specificity for the correct input-representation pairs in all corruption cases (**Figure 3.19.3**; **Figure 3.19.4**; **Figure 3.19.5**), indicating that the framework generated highly specific representations. In Monte Carlo simulations with randomly silenced early neurons, representations with high specificity were obtained with as few as 60 (23.4% of the 256) neurons (**Figure 3.19.6**).

However, as we decreased the number of representation neurons, the neurons' informativeness decreased, and the extracted features became more localized. The responses of locally tuned neurons could not differentiate between the inputs (**Figure 3.19.2**), whereas highly similar inputs could be readily distinguished based on responses of neurons having complex tuning properties. These locally tuned neurons' overall effect was reflected in lower specificity values achieved in Monte Carlo simulations (**Figure 3.19.6**). The presence of these locally tuned neurons diminished the specificity of representations in other cases of corruption as well (**Figure 3.19.3**; **Figure 3.19.4**; **Figure 3.19.5**).

1.


(i)
(ii)
(iii)
(iv)
Input
Representation
Reconstruction
Input
Representation
Reconstruction
Input
Representation
Reconstruction
Input
Representation
Reconstruction

2.


Response levels
Dictionary elements

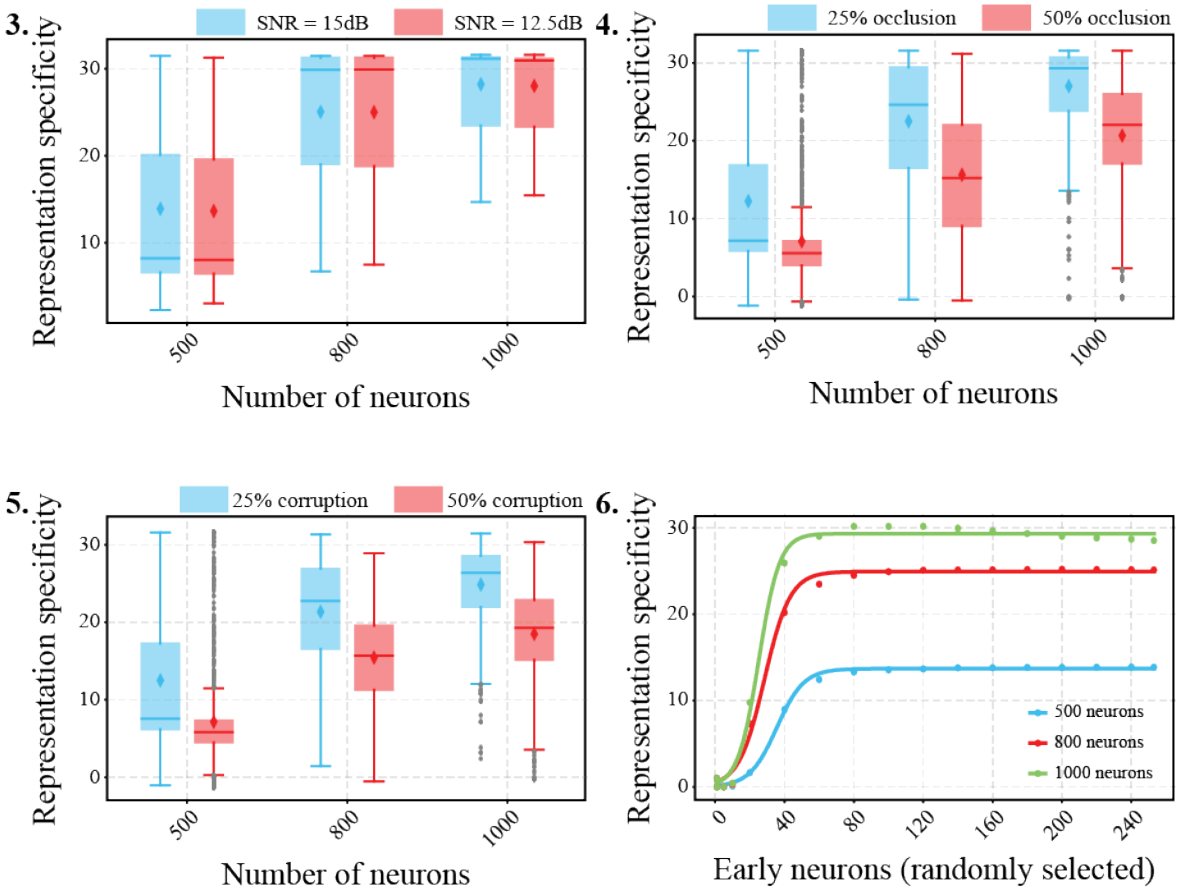


Figure 3. 19: Consistency in representing symbols

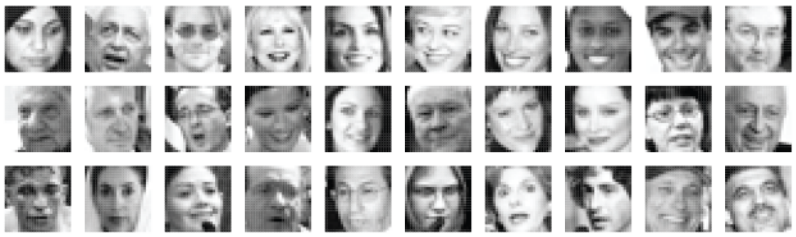
1. *Examples of representations obtained when inputs corrupted in different ways like (i) addition of noise, (ii) removal or occlusion of a portion of early neurons, and (iii) random removal of early neurons, were represented. Note the similarity of representations to the uncorrupted input (iv). The representations were utilized to reconstruct the inputs using the dictionary. The reconstructed inputs were similar to uncorrupted inputs.* **2.** *Response levels of neurons tuned to different features are shown. As indicated by similar response levels, localized features could not be utilized to differentiate similar inputs well. In contrast, structurally similar inputs could be well distinguished using neurons tuned to features that were very similar in structure to the inputs. Localized tunings were observed when the number of represented inputs was high relative to the number of representation neurons. They reduced the overall specificity of representations in those conditions.* **3 – 5.** *The obtained representations were very specific to the original uncorrupted representations*

across all forms of corruption. However, specificity increased with an increasing number of neurons. **6.** *Results of Monte Carlo analysis performed by randomly selecting a varying number of early neurons. Note that representation specificity saturates after 80 early neurons indicating that only 80 out of 256 early neurons are sufficient to produce highly specific representations. The specificity increases with the number of representation neurons in the system.*

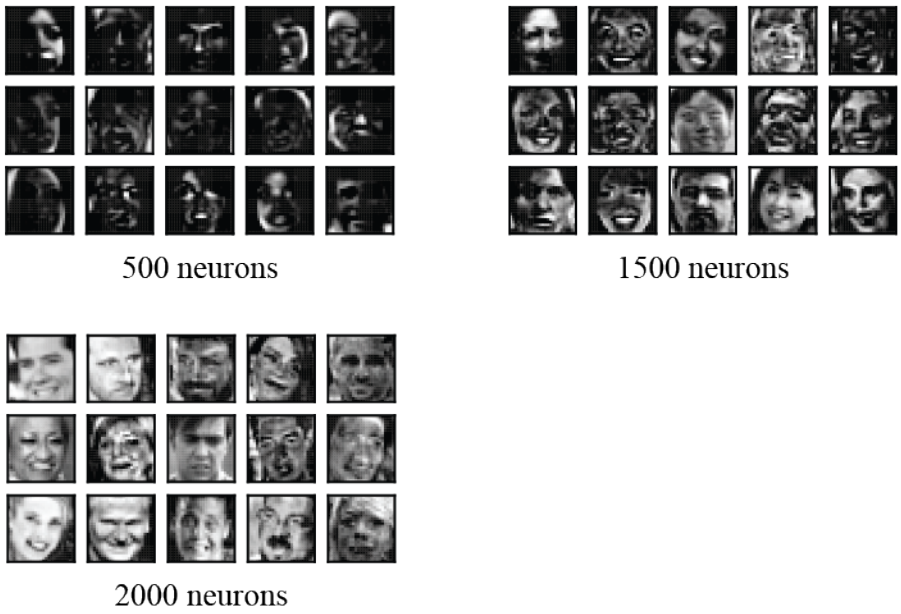
3.4.4. Analysis of faces

We next tested our framework in representing complex, non-binary inputs such as human faces (**Figure 3.20.1**). A set of 2000 human faces were represented using a varying number of neurons. Analyzing the tuning properties of neurons in different situations revealed that when fewer neurons were employed, the tuning properties were a complex assemblage of local facial features. The tuning became unique and face-like when the number of neurons was increased (**Figure 3.20.2**). Analyzing the kurtosis of response distributions neurons (**Figure 3.20.3.i**) and the correlation among neurons (**Figure 3.20.3.ii**) showed that maximum efficiency was achieved when the number of neurons matched the number of inputs.

1.



2.



3.

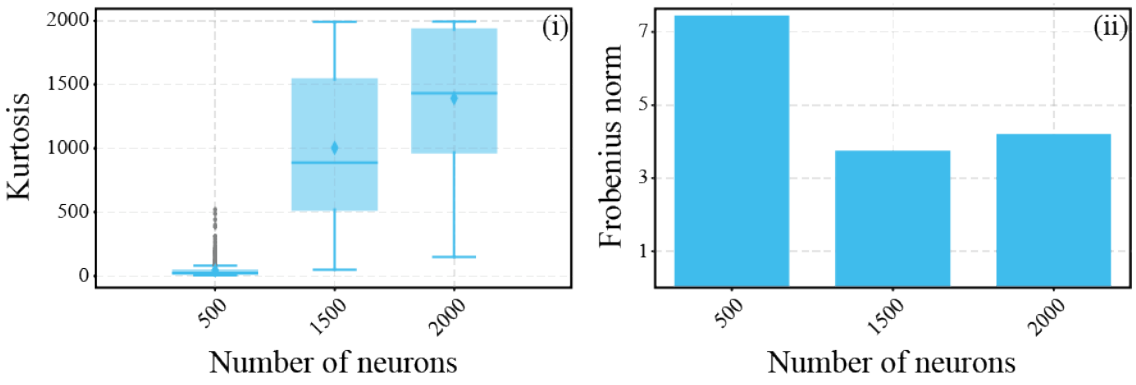


Figure 3. 20: Analysis of faces

1. *A few examples of faces that were considered in the analysis. A total of 2000 faces with 1000 male and 1000 female faces were included in the dataset of faces.* **2.** *The tuning properties of the neurons. Variation in tuning properties with number of neurons was*

analyzed, and it was found that at the low number of neurons, the tunings were similar to local features of the faces. However, as the number of neurons grew, the tuning became more face-like. **2.** *Faces were represented with a varying number of neurons. The kurtosis of the response profiles of the face neurons increased with their number (i), and the Frobenius norm of the difference between correlation matrix of neurons and identity matrix decreased (ii), indicating that the face representations were efficient.*

As face recognition comprises one of the most significant cognitive tasks, face representations' consistency was also analyzed. The faces' representations were stable, unique, and robust against common alterations such as the addition of headwear, facial hair, or eyewear **(Figure 3.21.1)**. The same face was represented nearly identically when a mustache, a pair of sunglasses, or both were added. Even when half of a face was blocked in different positions, the framework produced the same representation (**Figure 3.21.1**). Inversely reconstructed inputs from the representations were similar to those of unadulterated faces even when the faces were half blocked (**Figure 3.21.1**).

We compared our *"face code"* against a recently proposed code based on principal components (Chang and Tsao 2017). In the basis set resulting from the faces' principal component analysis (PCA), the same face with different parts occluded generated different representations. Input recovery resulted in occluded but not uncorrupted inputs (**Figure 3.21.2**). Quantification of specificity using similarity z-scores of 50 different faces occluded in different locations shows that our framework generates representations that are highly specific in matching the original input (**Figure 3.21.3)**. Recovered inputs from representations of corrupted inputs were highly similar to the original faces (**Figure 3.21.4**). PCA-based representations did not exhibit such selectivity or similarity (**Figure 3.21.3**; **Figure 3.21.4**). Thus, our study presented a robust combinatorial face-code distinct from the

one proposed before (Chang and Tsao 2017, Stevens 2018).

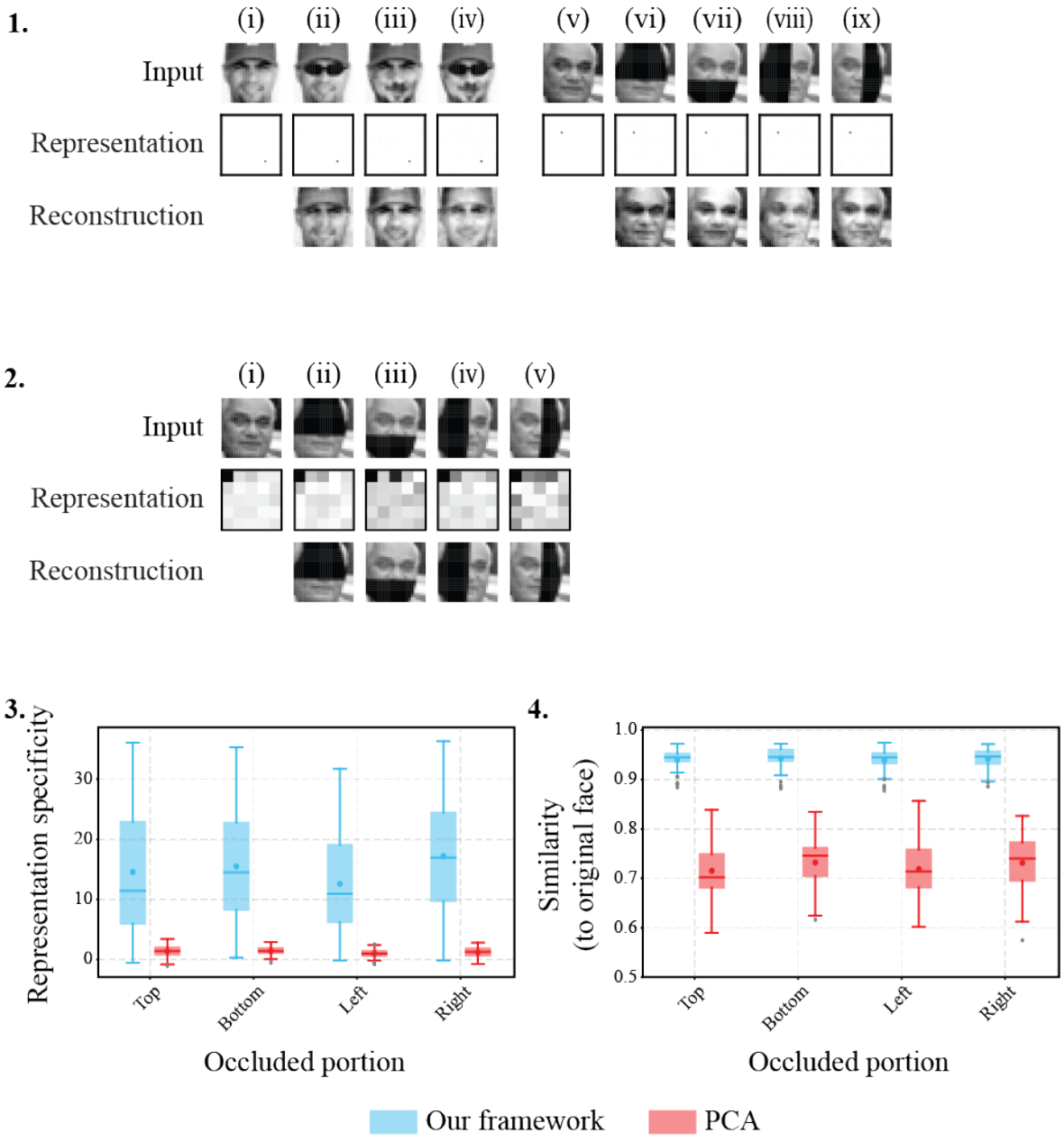


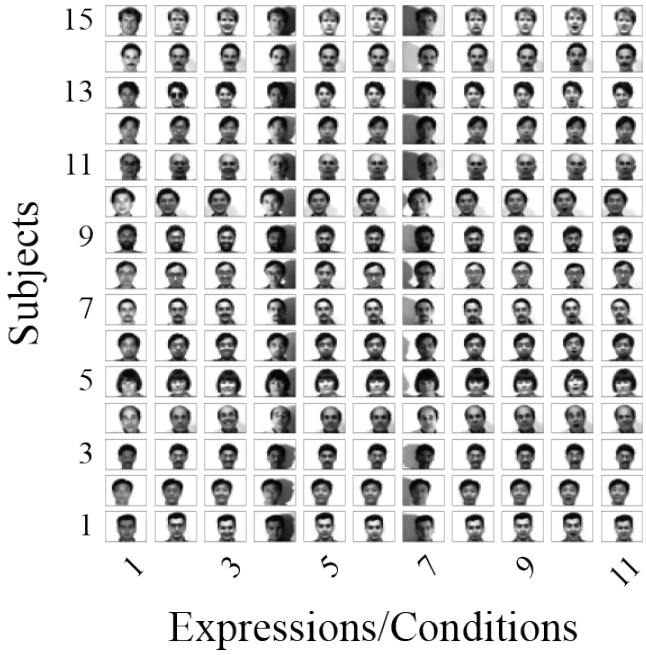
Figure 3. 21: Consistency in face representations

1. *Consistency of representations was analyzed using corrupted faces. Glasses (ii), beard and mustache (iii), or both (iv) were added to a face, and the representations of the altered faces were obtained. The representations were consistent with the original face (i). Moreover, the faces reconstructed from the tunings and response values of neurons resembled the original faces. In another example of corruption, half of a face was occluded*

from different locations (top (vi), bottom (vii), left (viii), right (ix)). In all cases, consistent representations of occluded faces were obtained, and the reconstructed faces matched the original one (v). **2.** *The specificity of representations in our framework was compared with the representations obtained through the faces' principal component analysis. The representations in the PCA basis were not consistent for the occluded faces ((ii) – (iv)). The reconstructed faces matched the corrupted ones and not the original faces (i).* **3.** *Using a set of 50 different faces occluded in different locations, it was shown that the overall specificity of representations of occluded faces in the PCA basis was very low.* **4.** *The faces reconstructed from PCA representations were not similar to the original face.*

We also tested if the basis set resulting from capturing unique structures from a specific set of faces can be utilized for consistently representing new face inputs. A different set of face inputs from the Yale face database containing faces of 15 individuals in 11 different lighting conditions and facial expressions (**Figure 3.22.1**) was acquired and represented (https://www.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html). Using specificity scores, as described previously, we found that the new faces' representations could be accurately categorized according to the individuals (**Figure 3.22.2**).

1.



2.

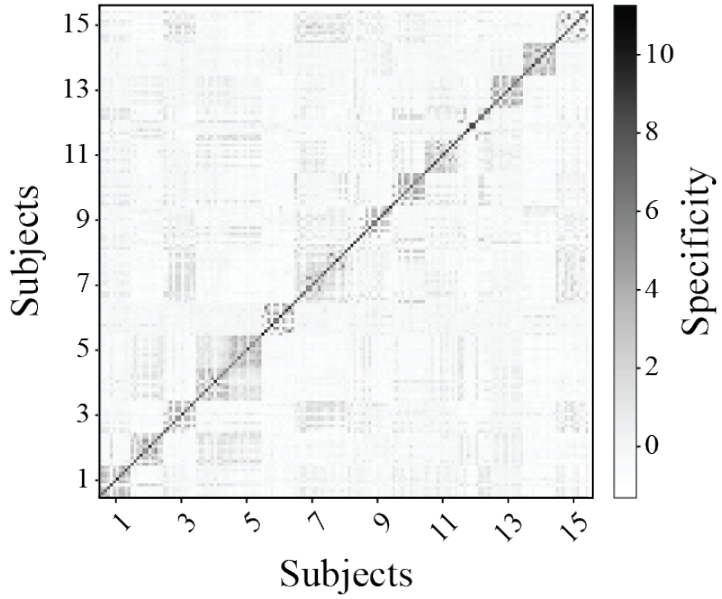


Figure 3. 22: Analysis of faces with different expressions and lighting conditions

1. *A separate set of faces, consisting of 15 different individuals with 11 different expressions or lighting conditions. The neurons were not tuned to these examples of faces, yet they were used to test the robustness of face representation in our framework.* **2.** *The obtained representations of the new data set of faces were specific to individuals.*

3.4.5. Analysis of odor response in the mouse olfactory system

Using the mouse olfactory system, we also tested if the framework proposed in this study can be extended to sensory modalities that do not detect external stimuli with pixel-like spatial segregation of the input patterns. In the mouse olfactory system, an odor activates a disparate set of glomeruli in the olfactory bulb (Fantana et al. 2008, Ma et al. 2012, Mombaerts 2006, Ressler et al. 1993, Vassar et al. 1993, Mombaerts et al. 1996, Treloar et al. 2002). The pattern is transformed into sparse activities in the piriform cortex, where odor identities presumably are decoded (**Figure 3.23.1**) (Poo and Isaacson 2009, Stettler and Axel 2009, Willhite et al. 2006). Without explicitly solving for the elemental features of chemicals, this two-stage system has a remarkable ability in identifying individual odorants (Ma et al. 2012). The olfactory system is also resilient against neuronal loss. Even when large portions of the olfactory bulb have been removed, rodents can still recognize trained odors (Lu and Slotnick 1998).

In a previous study (Ma et al. 2012), we collected responses of 94 glomeruli to 40 odorants from the mouse olfactory bulb's dorsal surface. Using this odor response data as our finite set of inputs, we obtained a 150-dimensional basis set for representing odors. Note that this case is peculiar because we are constrained by the olfactory system's anatomical organization to choose the number of representation neurons greater than the number of glomeruli. This choice makes the number of inputs far less than the number of neurons. Such situations are less likely to arise in a natural system; nonetheless, we decided to analyze them. We found that if the number of representation neurons was larger than the number of inputs, the representation neurons' correlation increased with their number (**Figure 3.23.2**). However, their response profiles' kurtosis remained high in all conditions of representation (**Figure 3.23.3**). This scenario could only arise when multiple neurons' tuning properties were similar but rarely detected in the input sets. Indeed, the number of neurons whose

tuning properties were more than 80% similar increased as the number of neurons was increased (**Figure 3.23.4**). Next, we decided to test the consistency of representations of odors in this peculiar situation. We found that nearly identical representations could be generated from the responses of small subsets of glomeruli (**Figure 3.23.5**). For example, odor representations generated from a random set of 16 glomeruli were nearly identical to those from the full set. This consistency suggested that odor recognition could be achieved with far fewer glomeruli (**Figure 3.23.5**). Moreover, a nearly identical representation of the same odor was achieved using different, arbitrary glomeruli sets (**Figure 3.23.6**).

We also performed Monte Carlo analyses using the responses of different numbers of randomly selected glomeruli. We found that the odor identification error rate decreased rapidly when the glomeruli number increased (**Figure 3.23.7**). 100% of odorants could be correctly identified with an average of 15 or more glomeruli randomly selected from the set (**Figure 3.23.7**). Note that an odor was correctly identified when the response evoked by it in a partial set of glomeruli could be mapped to a representation that was maximally similar to its representation obtained from the complete glomerular response. Representations of glomerular patterns were also consistent against noise. Gaussian noise was added to the glomerular responses, and odor identification rates were measured from Monte Carlo analyses. Increasing noise level reduced performance, and accurate identification required more glomeruli (**Figure 3.23.8**).

Nevertheless, odor identification was resilient against noise. At 10% noise level, nearly perfect identification was achieved with 20 glomeruli. Even when the noise level reached 40% of the signal, 60% of odorants could be identified using the responses from 30 glomeruli.

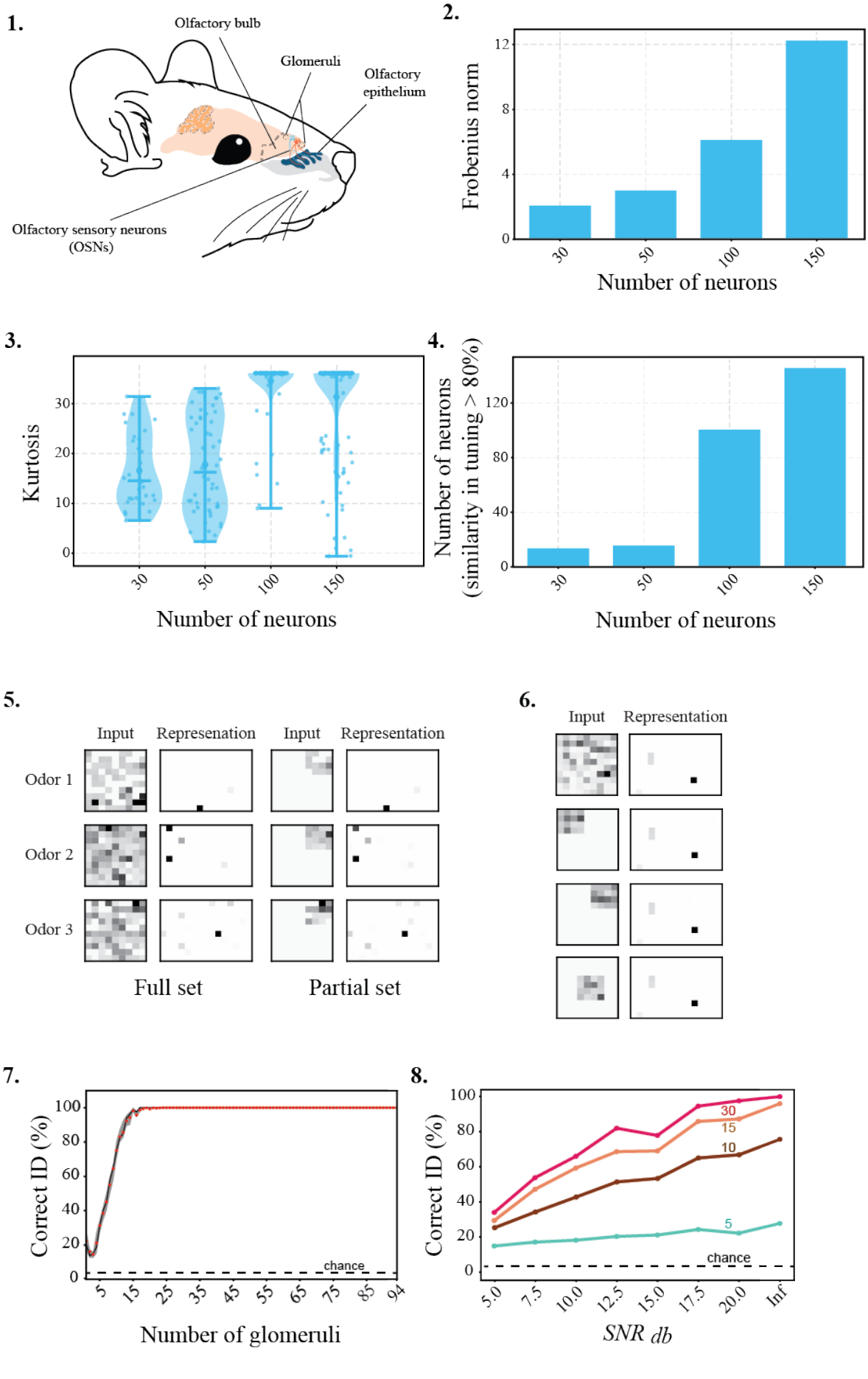

1.
Olfactory bulb
Glomeruli
Olfactory epithelium
Olfactory sensory neurons (OSNs)
2.
Frobenius norm
Number of neurons
3.
Kurtosis
Number of neurons
4.
Number of neurons (similarity in tuning > 80%)
Number of neurons
5.
Input
Represenation
Input
Representation
Odor 1
Odor 2
Odor 3
Full set
Partial set
6.
Input
Representation
7.
Correct ID (%)
chance
Number of glomeruli
8.
Correct ID (%)
30
15
10
5
chance
SNR db

Figure 3. 23: Analysis of odor response in the mouse olfactory system

1. *Schematic of the olfactory system in the mouse. Olfactory sensory neurons, which detect odorant molecules, project to stereotypic positions called glomeruli in the olfactory bulb from the olfactory epithelium. The projection is such that the neurons expressing the same odorant receptors converge to the same glomerulus. Activity from glomeruli is relayed to higher-order centers in the brain.* **2.** *Odor response of glomeruli in the mouse olfactory system was represented using a varying number of neurons. As the number of odors being represented in this case was lower than the number of neurons, an increase in the correlation between neurons was observed (indicated by increased Frobenius norm of the difference between correlation matrix of neurons and identity matrix).* **3.** *The kurtosis, on the other hand, increased with the number of neurons. These results meant that though the neurons were correlated, they still had relatively sparse response profiles. This situation could arise when multiple neurons were tuned to similar structures rarely found in the input set.* **4.** *Similarity of the tuning properties of neurons was analyzed. With the increasing number of neurons, the number of neurons that had more than 80% structural similarity between their tuning properties increased.* **5.** *Nonetheless, the consistency in representation was remarkable. Consistent representations of different odors could be obtained using responses from the same set of glomeruli.* **6.** *Different sets of glomeruli also produced consistent representations for a given odor.* **7.** *In Monte Carlo analyses performed with varying numbers of glomeruli, nearly all odors could be correctly identified with as few as 16 glomeruli.* **8.** *The correct identification of odors was affected by the addition of noise in the glomerular responses. Around 80% of odors could be identified with 15 glomeruli at 17.5db SNR. The percent of identified odors decreased with noise. However, using more glomeruli in the identification process improved performance. (Colored lines indicate the fraction of odors correctly identified with a constant number of glomeruli. The number of glomeruli is listed near the line)*

3.4.6. Analysis of natural images

A major portion of the past studies has shown that the receptive field properties of neurons in visual processing pathways can be explained by the efficient coding (Srinivasan et al. 1982, Atick and Redlich 1990, Olshausen and Field 1996, Bell and Sejnowski 1997). Independent components obtained from the statistical analysis of natural scenes conform to the oriented edge like receptive fields of V1 neurons (Hubel and Wiesel 1962, Hubel and Wiesel 1968). On the other hand, our framework is based on the most informative structures of the objects that need not be independent. As a detailed account of statistical properties of inputs is not required to obtain these structures, a question arises that how can the receptive field properties of neurons that have been determined using experimental studies can be produced in this framework. To test if our framework could explain receptive field properties of neurons in visual cortices, we decided to generate representations of a finite set of natural scene patches. Two channels were created in the input stream to be consistent with the physiology of the visual system. The "on" channel responded to the bright portions of the image, and its activity corresponded to the input intensity. The "off" channel detected the darker portions in the image, and its activity corresponded to the intensity of inverted input. Representing a finite set of image patches with a fixed number of representation neurons resulted in neurons with localized and orientation-selective tunings when the number of neurons was relatively low compared to the number of inputs. These tuning properties were similar to the receptive fields of V1 simple cells (Hubel and Wiesel 1962, Hubel and Wiesel 1968) (**Figure 3.24.1).** Despite high correlations among the images, the neurons were highly decorrelated (**Figure 3.24.3**). We quantified the fraction of tuning properties of neurons that resembled V1 receptive field using the Fourier transforms of the tuning properties. An increased number of input images increased the fraction of simple cell-like tuning properties among neurons (**Figure 3.24.2**). Thus, in this framework, localized tuning features naturally

emerge when large numbers of objects are represented. It can be argued that the requirement to represent an extraordinarily large number of stimuli from the natural environment forces areas such as V1 to produce localized tuning features.

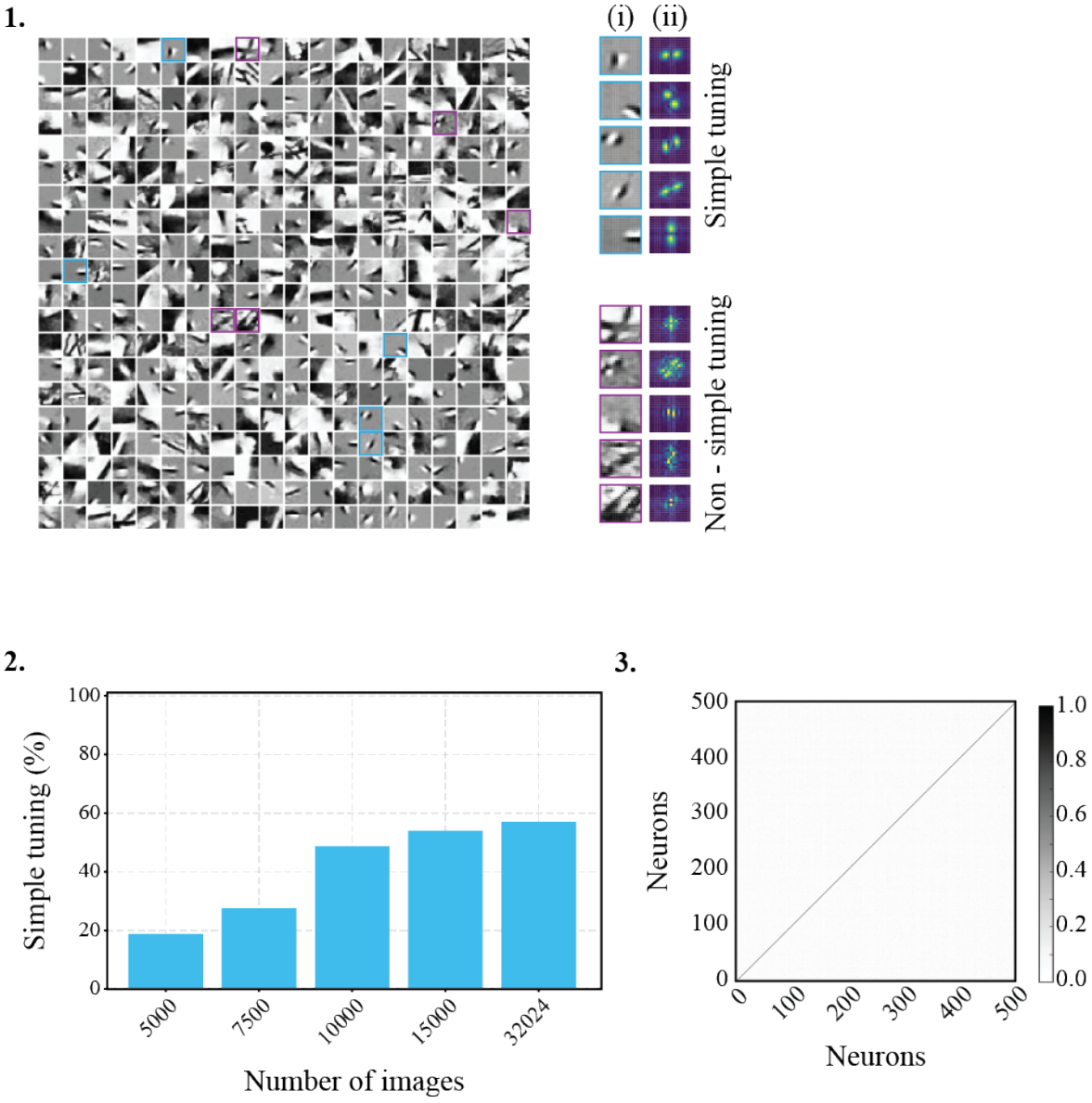


Figure 3. 24: Analysis of natural images

1. *2993 image patches were obtained from natural scenes and were represented by 500 neurons. The neurons displayed tuning properties that were like simple cells found in the V1 area of the visual cortex. Some of the neurons had more complex tunings (i). Fourier*

transforms of the tunings were utilized to distinguish the two types (ii). **2.** *The fraction of neurons with simple cell-like tunings increased with the number of images being represented.* **3.** *The correlation matrix of the representation neurons was nearly an identity matrix, indicating uncorrelated neurons.*

3.5. Discussion

Utilizing non-negative matrix factorization techniques, we have approximated the process of capturing the most informative features from a finite set of inputs in a linear transformation paradigm. Though limited by the difficulty of its implementation in the sensory circuit, the approach highlights some exciting characteristics of representations based on elements that uniquely characterize objects.

While representing inputs based on informative features, an intriguing relationship was observed between the number of inputs and the number of representation neurons. An increase in the number of inputs decreased the uniqueness of neurons' tuning properties when the number of neurons was held fixed. Conversely, an increase in the number of neurons increased the uniqueness of captured features when the number of inputs was kept constant. Such trends indicated that neurons became most informative about individual objects when the number of neurons matched the number of represented inputs. The efficiency of representation was also maximal in this situation. A common observation across different sensory modalities is that the number of neurons increases multiple folds in the brain's higher cortical regions. In light of the relationship between the number of neurons and the number of inputs, this increase might efficiently accommodate a larger number of inputs. Efficient representation of a greater number of inputs will allow identifying more accurate dependence among those inputs and help the organism understand its environment better. Thus, this can

explain the correlation between the cortex sizes observed among organisms and the complexity of tasks they can perform (Reader and Laland 2002).

The change in representation efficiency while representing a larger number of inputs was attributed to neurons getting tuned to more localized features of the input. Getting tuned to localized features decreased the sparseness of representations and hence informativeness of individual neurons. However, in certain situations, such representations might be useful. For example, it was shown that simple cell-like receptive field properties arise in neurons when encoding many natural image patches. The emergence of simple cell-like properties has been demonstrated by several studies (Olshausen and Field 1996, Bell and Sejnowski 1997, Olshausen and Field 1997) which have considered efficient encoding of statistics of natural images. These findings can be reconciled if one considers the saturating nature of the framework proposed in this study. The analysis with an increasing number of inputs has shown that more neurons get tuned to localized features with increasing input numbers. In this regard, the system is shown to adapt to inputs that it has encountered, and as more inputs are encountered, it gradually adapts the total statistics inputs. Note that this consideration is different from assuming that the system in all situations will adapt to the entire statistics because the portion of the statistics that the system will adapt to will depend upon its experience. In the matrix factorization approach, however, the component corresponding to the system's experience is missing. All inputs are presented at once as one input matrix to the system, which is then utilized to extract informative structures. This method is inadequate for studying the system's saturation states, which it presumably attains after experiencing a significant portion of its environment. A network simulation of the process which adapts to inputs in sequence with experience will be more appropriate for such studies.

The gradual emergence of localized tunings of neurons can also be explained in terms of the spread of representation vectors in high-dimensional space. Seeking representations based on unique structures renders the representations maximally distinct. This approach is equivalent to seeking representations that are maximally separated in high-dimensional

space. However, as more inputs are represented, the separation between representations must be decreased to accommodate new representations. Less separated representations will result from neurons that are tuned to commonly observed features. In other words, in less separated representations, the neurons will be tuned to local features, which is what we observe in our system.

Finally, we observe remarkable consistency in representing corrupted inputs. As noted in the simulations, the neurons' response values to individual inputs correspond to the mutual information between their tuning property and the input. Therefore, consistency in the representations implies that corruption has not vastly diminished the mutual information between the input and the neuron's tuning property. Such situations arise when a neuron is tuned specifically to a particular input, i.e., its tuning property comprehensively accounts for the input's entire structure. Indeed, the specificity of representations of corrupted inputs increases in representation scenarios where non-localized tunings are observed. An important point to note here is that correspondence between the mutual information and the response values of neurons is also a departure from the classical efficient coding paradigm where neurons' response values were supposed to indicate the unexpectedness of the input (Barlow 1987, Barlow 1989). Here, it indicates the confidence of the system in identifying the input.

This page is intentionally left blank.

CHAPTER 4

A network implementation of the adaptive strategy for sensory coding

Table of Contents

4.1. Introduction

In the previous chapter, I demonstrated the informativeness of features extracted from a finite set of objects using a sparse *NMF* approach, depended on the number of inputs relative to the neurons. When the number of inputs was comparable to the number of representation neurons, captured features were most informative and unique to individual objects. As the number of represented inputs grew, the features became localized and less informative about any input. Consequently, the efficiency of representations, measured in terms of their sparsity, was also affected. The representations were sparse for a relatively low number of inputs, and redundancy among neurons was small. When the number of inputs relative to the number of neurons increased, sparsity decreased, and redundancy increased. With this nature, the framework could successfully account for the localized receptive fields of the V1 neurons and the tunings of higher-order neurons to comprehensive structure, indicating that it can explain both early and high order visual processing. However, the biological plausibility of the framework still needs to be established.

Any theoretical framework trying to describe a biological process must consider the constraint faced by a biological system. For example, a framework that aims to explain sensory processing must address how neurons and their connections might serve as a substrate to carry out the proposed computations. Furthermore, as a biological system learns from gradually experiencing variegated inputs, aspects of experience, and learning from different input forms should be included in the framework. In this chapter, I describe how capturing the most informative structures can be implemented in neuronal circuits. Starting with a discussion on the limitations of the matrix factorization approach, I explain how biologically inspired neuronal networks have been utilized to generate inputs' sparse representations. Building on the understanding of these networks, I design a network to capture the unique, informative structures from inputs in an experience-dependent manner.

Using symbols dataset, I show the network's working and its efficiency and consistency in representing inputs. Finally, I show that the network can learn from the corrupted form of inputs as well.

4.2. Limitation of the matrix factorization approach

Though the framework based on informative features has successfully generated invariant and efficient representations of inputs, the sparse non-negative matrix factorization-based approach used in obtaining the informative features is not biologically plausible in its current form. The limitations arise because the mathematical algorithm utilized here does not incorporate the physiological constraints faced by a biological system. Here, we discuss a few aspects of a biological system that are desirable in any sensory coding process but are absent in this novel approach of sensory processing.

4.2.1 Learning as a continuous process: An essential aspect of a biological system is its development. Organisms grow and develop with time, reach maturation, and eventually die. During the span of their lives, they experience their surroundings and learn to adapt to them. From the perspective of sensory processing, this constitutes a continuous period of sensory experiences, and it allows the organisms to learn and re-learn sensory events. As a corollary, the system does not at once encounter all the events and stimuli to which it adapts. It gradually discovers these events, determines their relevance with experience, and then conforms accordingly to represent them. The blind source separation approach taken so far in this work cannot account for this facet of biological systems. In all the simulations, the input set is modeled as one input matrix that does not change anywhere in simulations. Moreover, the algorithm does not allow any change in the input, limiting its applicability in explaining sensory processing.

4.2.2 Ignoring the frequencies: As mentioned before, the informativeness of features is determined by their relative abundance. Though a framework set to capture informative features does not need to know the exact occurrence frequency of objects, it must take the relative abundance of features into account. The current approach based on blind source separation techniques is not capable of doing so. Changing the input matrix to include multiple occurrences of the same input cannot change the dictionary's nature. The multiple occurrences lead to repeated representations with the same level of sparsity and reconstruction error. Therefore, the dictionary and the representations remain similar to those obtained while considering each input only once. In other words, there is no constraint on the dictionary that forces it to change according to the inputs' relative occurrence. Thus, the current approach fails to utilize the environment's statistical properties for its benefits and ignores information relevant to biological systems.

4.2.3 A unified approach: So far, in our approach, we have considered capturing the most informative structures from inputs as a different process than obtaining input representations. While the former is achieved through the non-negative blind source separation technique (Rapin et al. 2013), the latter is done through a sparse recovery approach (Candes and Romberg 2005). The two methods are different in their formulation and implementation. On the other hand, a biological system does not have separate circuits to capture features and generate representations. The same circuit adapts to a set of inputs and represents them. Moreover, the input representations are expected to guide the process of adaptation. The current approach fails to recapitulate these critical sensory processing aspects and does not integrate the two processes.

4.3. A neuronal network for capturing informative structures

Realizing these limitations of the matrix factorization approach, we decided to utilize a neuronal network to capture the informative structures from inputs. This part of the study aimed to design a network of model neurons that could extract unique input structures and efficiently represent inputs. In other words, we sought a single network model that incorporated the functionality of both blind source separation and sparse recovery.

However, we realized that both these functionalities correspond to different properties of the network. The capturing of the informative structures is reflected in the tuning properties of the representation neurons. The representation neurons' tuning properties are determined by how they are connected to the early-stage neurons in the sensory pathway. Therefore, the adaptation to inputs pertains to changes in the connections of the network.

On the other hand, an input's representation is the population response pattern of the representation neurons. Hence, achieving efficiency in representation corresponds to appropriately shaping this response pattern.

With these considerations, the optimization problems that were being solved by the combination of blind sources separation and sparse recovery could be broadly divided into two subproblems stated below

1. **Given connectivity among neurons, find a sparse response pattern for any input encountered:** Essentially, this problem is about finding sparse representations of inputs in any given network. The possible solutions to the problem have been proposed in previous studies (Földiak 1990, Rozell et al. 2008). We utilize the same approach as these studies but with an additional constraint that the representations must be non-negative.

2. **Given the neurons' response pattern, change the connectivity appropriately to adapt to the encountered inputs:** This subproblem corresponds to updating the network's connectivity. As the connectivity of neurons changes, their tuning property also changes. Appropriate changes in the connectivity can guide the neurons to be tuned to the most informative structures. As a connection between two neurons can be both excitatory and inhibitory, the changes in these connections can similarly be of either nature. Therefore, the updates in different connections can have different signs. Such updates may appear contradictory to the non-negativity constraint that has been essential for capturing informative structures. However, it is critical to realize that though the connectivity changes can be bidirectional, the inhibitory connections only reduce neurons' activity and do not push it below zero. In this setting, the network cannot subtract the neurons' tuning properties from one another. Thus, the non-negativity constraint can be satisfied even though the neurons receive both excitatory and inhibitory inputs.

A specific architecture of neuronal networks chosen to solve these two subproblems is described in the next few sections.

4.3.1. Hopfield network and locally competitive algorithm for sparse recovery

The intuition to develop a network that solves the first subproblem comes from one of the most popular forms of the artificial neural network developed by John Hopfield (Hopfield 1982). Hopfield network is essentially a recurrent network of binary threshold units which, at any point in time, can take only one of the two possible values (-1 and 1, or 0 and 1). The network comprises layers of these units, with each unit receiving input from all other units except itself (**Figure 4.1**).

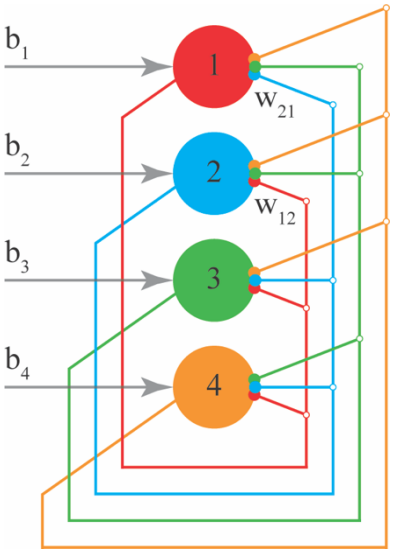


Figure 4. 1: A schematic diagram of a Hopfield network

An example of a Hopfield network in which four units (numbered and presented in red, blue, green, and orange) form recurrent connections with each other such that each unit receives input from all units except itself. Connections from a particular unit have the same color as the unit itself. Mutual connections between units 1 and 2, which are listed as weights w_{21} *and* w_{12}, *are equal. Apart from recurrent connections, units may also receive inputs from an input layer* (b_1-b_4) *in the network (denoted by grey arrows).*

The connection strength of these units is described by a set of parameters called weights w_{ij}. These parameters are chosen such that the strength of connection from unit i to unit j is the same as the strength of connection form unit j to unit i i.e.

$$w_{ij} = w_{ji}$$

Thus, if one arranges these weights in a square matrix $\boldsymbol{W}$, the resulting matrix will be symmetric, with all diagonal entries being zeros indicating that the units do not receive inputs from themselves. At any instant t, the input to a unit is the weighted sum of other units' states, where the weights correspond to the connection strengths. In mathematical

form, if there are N units in the network, and the state of unit i at time t is denoted by $x_i(t)$, then input to a particular unit k at time t, can be expressed as

$$b_k(t) = \sum_{i=1}^{N} w_{ik} x_i(t)$$

Following this rule, inputs to all N units can be expressed in terms of connection matrix $\boldsymbol{W}$ as

$$\hat{b}(t) = \boldsymbol{W}\hat{x}(t)$$

where $\hat{b}(t)$ is an N-dimensional vector of inputs at time t to all N units, and $\hat{x}(t)$ is another N-dimensional vector of states of the units. The units' states at the next instant $(t + 1)$ are determined from these inputs by the following rule

$$x_k(t+1) = 1 \ if \ b_k(t) \geq 0; \quad x_k(t+1) = -1 \ if \ b_k(t) < 0$$

An attractive property of these networks is that their units tend to *pull in or push away*. For example, consider a connection between two units i and j. If $w_{ij} = w_{ji} > 0$, then irrespective of the value of x_i, the state update will bring the value of x_j closer to x_i. (If x_i is 1 then $w_{ij}x_i > 0$, this means that in the next update, the value x_j will tend to be positive too, and hence, the value of x_j is *pulled in* towards the value x_i and vice versa.) Similarly, if $w_{ij} = w_{ji} < 0$, the units tend to *push away* each other.

This property led to the realization that if the connection weights between the encoders are chosen in certain ways, then the network can be made to *"remember"* specific patterns in its unit. In the previous example, assume that one wishes to store a pattern $x_i = 1 \ and \ x_j = -1$ in the network, and retrieve it in situations when only a partial form of the pattern is available. Since the values of units in the pattern are already pushed away from each other, setting $w_{ij} < 0$ will ensure that even when only the state of one of the units is known, the network pushes the other unit's state towards the correct state and recalls the complete pattern. In general, to store p patterns in a network with N units, setting weights by the following rule can ensure recollection of patterns

$$w_{ij} = \frac{1}{N}\sum_{r=1}^{p} x_i^r x_j^r$$

where x_i^r denotes the state of the i^{th} unit in the r^{th} pattern. Such a way of settings the weights is often referred to as the *Hebbian rule of learning* (Hebb 1949). It imparts the network a form of *associative memory* known as *content-addressable memory CAM* (Kohonen 2012). Due to these properties, the network has found wide applications in pattern recognition and explaining associative memory (Paik and Katsaggelos 1992, Young et al. 1997, Zhu and Yan 1997).

In his later work (Hopfield 1984), Hopfield extended these networks to include units with graded response profiles rather than binary values. Each such unit was viewed as an individual neuron, and parameters like membrane potential u, membrane capacitance C, transmembrane resistance R, and firing rate V were defined. The dynamics of the states of each model neuron was described with the following equations

$$C_k \frac{du_k}{dt} = \sum_{i=1}^{N} W_{ik} V_i - \frac{u_k}{R_k} + b_k$$

$$u_k = g_k^{-1}(V_k)$$

where b_k was input and g_k denoted an invertible function relating membrane potential u_k to average firing rate V_k of the unit k. Hopfield could show that if the weight matrix $\boldsymbol{W}$ was designed using the Hebbian rule, this network functioned as *CAM* (Hopfield 1984).

Later, Rozell (Rozell et al. 2008) demonstrated that if a set of linear model neurons having tuning properties $\widehat{\phi_\iota}$ were connected in Hopfield network architecture, with weight matrix defined as

$$\boldsymbol{W} = -(\boldsymbol{\phi}^T\boldsymbol{\phi} - \boldsymbol{I})$$

where $\boldsymbol{\phi}$ was a matrix whose columns were tuning properties of neurons, and $\boldsymbol{I}$ was an identity matrix, then appropriately choosing the function g resulted in the network solving the sparse recovery problem (Rozell et al. 2008). In particular, if an input $\hat{y}$ was presented to the network, then input to the individual linear neurons was defined as

$$\hat{b} = \boldsymbol{\phi}^T \hat{y}$$

and the dynamics of the network could be described as

$$\tau \frac{d\hat{u}}{dt} = -\hat{u} + \boldsymbol{\phi}^T \hat{y} - (\boldsymbol{\phi}^T \boldsymbol{\phi} - \boldsymbol{I})\hat{V}$$

$$\hat{u} = g^{-1}(\hat{V})$$

He further showed that if g was of the form

$$g_{(\alpha,\gamma,\lambda)}(u_k) = \frac{u_k - \alpha\lambda}{1 - e^{-\gamma(u_k - \lambda)}}$$

then, with $\alpha = 1$, and in limits $\lim_{\gamma \to \infty} g_{(\alpha,\gamma,\lambda)}$, the evolving dynamics of the system minimized the energy function given as

$$E = \frac{1}{2}\left\|\hat{y} - \boldsymbol{\phi}\hat{V}\right\|^2 + \lambda\left\|\hat{V}\right\|_1$$

which is the same as the optimization function for sparse recovery problems.

4.3.2. Network design

As described above, Hopfield networks (Hopfield 1982, Hopfield 1984), with certain alterations, can solve the sparse recovery problem. The first subproblem that we intend to solve in our network is also to find sparse representations for inputs. Therefore, we designed a two-layered network of neurons based on the Hopfield network architecture. The first layer (activity denoted by $\hat{y}$), which we call the primary layer, corresponded to the layers present early in the sensory pathway and presented input patterns to the system. The second layer (membrane potential denoted by $\hat{u}$ and firing rate or the representation pattern denoted by $\hat{V}$) comprised representation neurons that received input from the first layer and had recurrent connections among themselves based on the Hopfield architecture (**Figure 4.2**). The primary layer was connected to the representation layer through a connection matrix $\boldsymbol{W}$. The shape of the connection matrix depended on the number of neurons in the primary and

representation layers and was not symmetric. The recurrent connections, on the other hand, were described by a symmetric matrix $\boldsymbol{S}$. The symmetry of the matrix implied that, like the Hopfield network (Hopfield 1982, Hopfield 1984), the connection strength from neuron i to j was the same as the connection strength from neuron j to neuron i.

In the Rozell model (Rozell et al. 2008), the connection strengths of recurrent connections were formulated as the similarity between the tuning properties of the neurons. However, our network was expected to be adapting to the inputs. The neurons' tuning properties were supposed to change with experience. In this sense, prior knowledge about tuning properties was not available. We realized that the tuning properties of neurons arise due to their connections to the primary layers. Therefore, a suitable measure for the strength of recurrent connections could be the similarity of representation neurons' connections to the primary neurons.

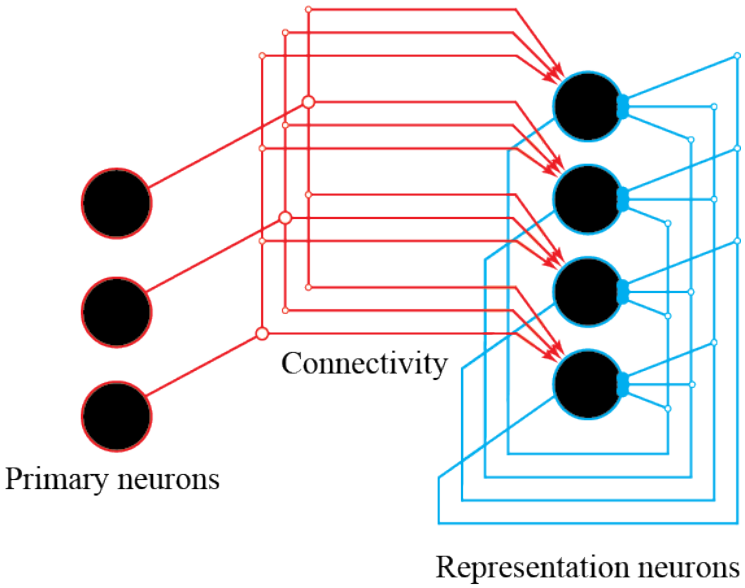


Figure 4. 2: A diagram of the network designed to extract the most informative features form inputs

A set of primary neurons (presented as dark circles with red outlines) are connected to a set of representation neurons (presented as dark circles with blue outlines). Each representation neuron is connected to all primary neurons (connections are shown in red). Besides, the representation neurons receive recurrent inputs (connections shown in blue) in ways similar to a Hopfield network.

If two neurons were similarly connected to the primary layers, any given input would similarly activate them. Hence, based on their activities, their recurrent interactions would be similar as well. In formal terms,

$$\boldsymbol{S} = -(\boldsymbol{W^T W} - \boldsymbol{I})$$

With these considerations, the dynamics of our network was given as

$$\tau \frac{d\hat{u}}{dt} = -\hat{u} + \boldsymbol{W}^T \hat{y} - (\boldsymbol{W}^T \boldsymbol{W} - \boldsymbol{I})\hat{V}$$

$$\hat{u} = g^{-1}(\hat{V})$$

Here, the function g relating the membrane potential to the firing rate was the same as the Rozell model (Rozell et al. 2008).

An important point that needs to be noted is that as our network adapts to inputs, the connections between the first layer of neurons and the representation neurons are expected to change. This change will be reflected in the recurrent connections' strengths because they are defined based on the similarity of representation neurons' connections to the primary layers. The dependence of this form makes our network completely dynamic. It is adapting to the inputs not only through the changes in connections between the primary and the representation layers but also through updating recurrent connections' strengths. Hopfield and Rozell's networks lacked such dynamic nature and hence were significantly different from our network model.

Considering the second subproblem, we realized that an appropriate way of quantifying the goodness of adaptations was to measure the difference between an input and its reconstruction obtained from the neurons' tuning properties and response values. If $\hat{V}$ was the representation of an input $\hat{y}$, and $\boldsymbol{\phi}$ was the matrix of tuning properties of the neurons, then this measure could be defined as

$$E = \left\| \hat{y} - \boldsymbol{\phi}\hat{V} \right\|^2$$

The strategy for updating the connectivity should be such that the above term is reduced with each update. For linear neurons, as the activity is a function of the weighted sum of its inputs, a change in tuning properties directly corresponds to a change in its connectivity i.e.

$$\Delta \boldsymbol{W} \propto \Delta \boldsymbol{\phi}$$

Therefore, a change in connectivity that reduces the above error should correspond to a change in $\boldsymbol{\phi}$. Following this rationale, we devised a three-step method for updating the connectivity. First, for each state of connectivity, the tuning properties were determined. Second, a change in tuning property that would reduce the error was then calculated from the representations, and lastly, a change proportional to that was made in the connectivity. This method will be further discussed in detail in a later section of this chapter.

4.4. Methods

4.4.1. Creating a bias in the connectivity

In the adaptive strategy of representing inputs based on the most informative structures, to adapt to different forms of inputs, the system must be competent in differentiating the inputs in the first place. If the system cannot distinguish two different inputs, then the whole adaptation process will be flawed, and the system can only achieve

selective adaptation. As the neuronal response caused by the inputs are their only possible identifiers for the system, the system must be set in ways that make it capable of differentiating inputs based on the response they elicit. In this regard, setting up the initial connectivity for the network is one of the crucial steps. Without proper initial connectivity, different inputs may cause similar network responses and can be regarded as the same.

Even from an evolutionary perspective, it can be argued that a bias in connectivity is selected over complete randomness. A system set to identify expected threats early in life will have a better chance of survival.

Considering that evolution has selected specific connectivity adapted to environmental stimuli, we proceeded with an assumption that the system is set to minimize the chances of getting two representation neurons activated by the same input. Such a constraint will ensure that different inputs activate different neurons and do not get mapped to the same representation. Stated formally, with this constraint, we demanded the expected value of the variance-covariance matrix of the response profiles of neurons to be an identity matrix i.e.

$$\mathbb{E}[\boldsymbol{V}\boldsymbol{V}^T] = \boldsymbol{I}$$

where $\boldsymbol{V}$ is the matrix of representations of different inputs and $\boldsymbol{I}$ is an identity matrix. Ignoring the non-linearity conferred to the system by the function g, we can approximate $\boldsymbol{V}$ in terms of input matrix $\boldsymbol{Y}$ and weight matrix $\boldsymbol{W}$ as

$$\boldsymbol{V} = \boldsymbol{W}^T\boldsymbol{Y}$$

This relation gives

$$\mathbb{E}[\boldsymbol{V}\boldsymbol{V}^T] = \mathbb{E}[(\boldsymbol{W}^T\boldsymbol{Y})(\boldsymbol{W}^T\boldsymbol{Y})^T] = \mathbb{E}[\boldsymbol{W}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{W}] = \boldsymbol{W}^T\mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^T]\boldsymbol{W}$$

Clearly, $\mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}^T]$ is the variance-covariance matrix of response profiles of early neurons (denoted by $\boldsymbol{\Sigma_{YY}}$) based on the set of inputs. With this relation, the above requirement of matching variance-covariance matrix of representation neurons to the identity matrix reduces to solving the following equation

$$\boldsymbol{W^T \Sigma_{YY} W = I}$$

Since variance-covariance matrix of any set of random variables is symmetric, it can be diagonalized using the orthogonal matrix $\boldsymbol{Q}$ of its eigenvectors i.e.

$$\boldsymbol{\Sigma_{YY} = Q \Lambda Q^T}$$

where $\boldsymbol{Q}$ is the matrix of orthogonal eigenvectors of $\boldsymbol{\Sigma_{YY}}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues of $\boldsymbol{\Sigma_{YY}}$. Using this transformation to solve the problem at hand, we have

$$\boldsymbol{W^T Q \Lambda Q^T W = I}$$

$$or, \quad \boldsymbol{W^T Q} = \Lambda^{-1/2}$$

$$or, \quad \boldsymbol{W^T = \Lambda^{-1/2} Q^T}$$

In general, any matrix $\boldsymbol{\eta} \in \mathbb{R}^{N \times M}$ with orthogonal columns can be multiplied with the above solution, i.e.

$$\boldsymbol{W^T = \eta \Lambda^{-1/2} Q^T}$$

Thus, a connectivity matrix $\boldsymbol{W}$ as derived above will make the variance-covariance matrix of representation neurons' response profiles match the identity matrix. Two important points need to be noted here

1. Finding appropriate $\boldsymbol{W}$ requires eigenvectors of the variance-covariance matrix of the input matrix. However, complete knowledge of inputs is not required. A subsample of the inputs that are more likely to be encountered will also set up the network such that the expected inputs are not mapped to the same representation.
2. The generalizing matrix $\boldsymbol{\eta} \in \mathbb{R}^{N \times M}$ is supposed to have all orthogonal columns, which is possible only in cases where $N \geq M$. As M and N are the numbers of primary and representation neurons, respectively, such a generalizing matrix implies that our network's connectivity can be generalized only when the number of representation neurons is larger than the number of primary neurons. Architectures, where higher-order neurons exceed early neurons by several folds,

are common in biological networks. Thus, this form of connectivity can be achieved in biological systems.

4.4.2. Updating the connectivity between the primary layer and the representation layer

As discussed in the previous section, the update in the connectivity matrix is derived from the updates in the neurons' tuning properties. The update in the tuning properties, in turn, should be designed to reduce the measure of adaptation of the system to inputs given by

$$E = \left\| \hat{y} - \boldsymbol{\phi}\hat{V} \right\|^2$$

In this regard, for a particular input $\hat{y}$ and its corresponding representation $\hat{V}$, the optimization problem in $\boldsymbol{\phi}$ can be stated as

$$\underset{\boldsymbol{\phi}}{\text{minimize}} \quad f(\boldsymbol{\phi}) = \frac{1}{2}\left\| \hat{y} - \boldsymbol{\phi}\hat{V} \right\|^2 \qquad (P)$$

The problem can be solved by taking a gradient descent approach. In this approach, a function's value is iteratively reduced by updating its variables along its gradient. In other words, for every variable, the value which further reduces the function is found by moving along functions' negative gradient with respect to the variable. Eventually, a minimum of the function is reached. In our case, the gradient descent steps can be formulated as

$$\boldsymbol{\phi_{k+1}} = \boldsymbol{\phi_k} - \alpha \nabla f$$

$$= \boldsymbol{\phi_k} - \alpha\left(\boldsymbol{\phi_k}\hat{V} - \hat{y}\right)\hat{V}^T \quad \boldsymbol{M} = \left(\boldsymbol{I} - \alpha\hat{V}\hat{V}^T\right)$$

$$= \boldsymbol{\phi_k} - \alpha\boldsymbol{\phi_k}\hat{V}\hat{V}^T + \alpha\hat{y}\hat{V}^T$$

$$= \boldsymbol{\phi_k}\left(I - \alpha\hat{V}\hat{V}^T\right) + \alpha\hat{y}\hat{V}^T$$

$$= \boldsymbol{\phi_k}\boldsymbol{M} + \boldsymbol{C}$$

where $\boldsymbol{\phi_k}$ is the value for $\boldsymbol{\phi}$ after the k^{th} iteration, $\boldsymbol{M} = \left(\boldsymbol{I} - \alpha\hat{V}\hat{V}^T\right)$ $\boldsymbol{C} = \alpha\hat{y}\hat{V}^T$, and α is the step size. After n such descent steps $\boldsymbol{\phi_n}$ can be calculated in terms of initial $\boldsymbol{\phi_0}$ as

$$\boldsymbol{\phi_n} = \boldsymbol{\phi_0}\boldsymbol{M}^n + \boldsymbol{C}\left(\sum_{k=0}^{n-1} \boldsymbol{M}^k\right) \tag{1}$$

We observe that $\boldsymbol{M}$ is a rank one perturbation in the identity matrix and hence has the following property

$$\boldsymbol{M}\hat{V} = \left(\boldsymbol{I} - \alpha\hat{V}\hat{V}^T\right)\hat{V} = \hat{V} - \alpha\hat{V}\hat{V}^T\hat{V} = \hat{V} - \alpha\left\|\hat{V}\right\|^2\hat{V} = \left(1 - \alpha\left\|\hat{V}\right\|^2\right)\hat{V}$$

putting $\alpha\left\|\hat{V}\right\|^2 = \tilde{\alpha}$, we get

$$\boldsymbol{M}\hat{V} = (1 - \tilde{\alpha})\hat{V} \tag{2}$$

similarly,

$$\hat{V}^T\boldsymbol{M} = \hat{V}^T\left(\boldsymbol{I} - \alpha\hat{V}\hat{V}^T\right) = \hat{V}^T - \alpha\left\|\hat{V}\right\|^2\hat{V}^T = \left(1 - \alpha\left\|\hat{V}\right\|^2\right)\hat{V}^T = (1 - \tilde{\alpha})\hat{V}^T \tag{3}$$

and for any other vector $\hat{x} \neq \hat{V}$

$$\boldsymbol{M}\hat{x} = \left(\boldsymbol{I} - \alpha\hat{V}\hat{V}^T\right)\hat{x} = \hat{x} - \alpha\hat{V}\hat{V}^T\hat{x} = \hat{x} - \alpha\left(\hat{V}^T\hat{x}\right)\hat{V} = \hat{x} - \beta_{\hat{x}}\hat{V} \text{ where } \beta_{\hat{x}} = \alpha\hat{V}^T y \tag{4}$$

Also, as $\boldsymbol{M}$ is symmetric, it can be diagonalized as under

$$\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T \quad \text{where} \quad \boldsymbol{\Lambda} = \boldsymbol{\mathcal{D}}\begin{pmatrix} 1 - \tilde{\alpha} \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

here, $\boldsymbol{\mathcal{D}}$ represents a diagonal matrix, with diagonal elements given by the column vector as the argument. Following diagonalization, we can calculate $\boldsymbol{M}^p$ as

$$\boldsymbol{M}^p = \boldsymbol{Q}\boldsymbol{\Lambda}^p\boldsymbol{Q}^T \quad \text{where} \quad \boldsymbol{\Lambda}^p = \boldsymbol{\mathcal{D}}\begin{pmatrix} (1 - \tilde{\alpha})^p \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{5}$$

and as $\boldsymbol{C} = \alpha\hat{y}\hat{V}^T$, using (3) we can say that

$$\boldsymbol{C}\boldsymbol{M}^k = \alpha\hat{y}\hat{V}^T\boldsymbol{M}^k = \alpha\hat{y}(1 - \tilde{\alpha})^k\hat{V}^T = (1 - \tilde{\alpha})^k\alpha\hat{y}\hat{V}^T = (1 - \tilde{\alpha})^k\boldsymbol{C} \tag{6}$$

Using (6) in (1), we get

$$\boldsymbol{\phi_n} = \boldsymbol{\phi_0}\boldsymbol{M}^n + \sum_{k=0}^{n-1} \boldsymbol{C}\boldsymbol{M}^k$$

$$= \boldsymbol{\phi_0 M}^n + \sum_{k=0}^{n-1} (1-\tilde{\alpha})^k \boldsymbol{C}$$

$$= \boldsymbol{\phi_0 M}^n + \left(\sum_{k=0}^{n-1} (1-\tilde{\alpha})^k\right) \boldsymbol{C}$$

$$= \boldsymbol{\phi_0 M}^n + \left(\frac{1-(1-\tilde{\alpha})^n}{\tilde{\alpha}}\right) \boldsymbol{C}$$

$$= \boldsymbol{\phi_0 M}^n + \left(\frac{1-(1-\tilde{\alpha})^n}{\alpha \left\|\hat{V}\right\|^2}\right) \alpha \hat{y} \hat{V}^T$$

$$= \boldsymbol{\phi_0 M}^n + (1-(1-\tilde{\alpha})^n) \frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2} \tag{7}$$

We can also calculate $\boldsymbol{\Delta \phi_n = \phi_n - \phi_{n-1}}$ for $n \geq 2$ using (7) as follows

$$\boldsymbol{\Delta \phi_n} = \left(\boldsymbol{\phi_0 M}^n + (1-(1-\tilde{\alpha})^n) \frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2}\right) - \left(\boldsymbol{\phi_0 M}^{n-1} + (1-(1-\tilde{\alpha})^{n-1}) \frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2}\right)$$

$$= \boldsymbol{\phi_0}(\boldsymbol{M}^n - \boldsymbol{M}^{n-1}) + \left((1-(1-\tilde{\alpha})^n) - (1-(1-\tilde{\alpha})^{n-1})\right) \frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2}$$

$$= \boldsymbol{\phi_0}(\boldsymbol{M}^n - \boldsymbol{M}^{n-1}) + ((1-\tilde{\alpha})^{n-1} - (1-\tilde{\alpha})^n) \frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2}$$

$$= \boldsymbol{\phi_0 M}^{n-1}(\boldsymbol{M} - \boldsymbol{I}) + (1-\tilde{\alpha})^{n-1}\left(1-(1-\tilde{\alpha})\right) \frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2}$$

$$= \tilde{\alpha}(1-\tilde{\alpha})^{n-1} \frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2} - \frac{\tilde{\alpha}}{\left\|\hat{V}\right\|^2} \ \boldsymbol{\phi_0 M}^{n-1} \hat{V}\hat{V}^T$$

$$= \frac{\tilde{\alpha}}{\left\|\hat{V}\right\|^2}\left((1-\tilde{\alpha})^{n-1} \hat{y}\hat{V}^T - \boldsymbol{\phi_0}(1-\tilde{\alpha})^{n-1} \hat{V}\hat{V}^T\right) \quad \text{using(2)}$$

This means that

$$\boldsymbol{\Delta \phi_n} = \frac{\tilde{\alpha}(1-\tilde{\alpha})^{n-1}}{\left\|\hat{V}\right\|^2}\left(\hat{y} - \boldsymbol{\phi_0}\hat{V}\right)\hat{V}^T \tag{8}$$

The following observations can be made from (7) and (8)

1. As $\tilde{\alpha} \to 0$, $\boldsymbol{\phi_n} \to \boldsymbol{\phi_0}$ and $\boldsymbol{\Delta \phi_n} \to 0$ for any value of n, confirming no descent

2. If we choose $\tilde{\alpha} > 1$ then $\boldsymbol{\Delta\phi_n}$ starts oscillating with n
3. At $\tilde{\alpha} = 1$, $\boldsymbol{\Delta\phi_n} = 0$ and $\boldsymbol{\phi_n} = \boldsymbol{\phi_0 M}$ ($\boldsymbol{\Lambda}^p = \boldsymbol{\Lambda}\ \forall\ p$) which is stationary (no descent)

thus, from the above observations, we can conclude that the feasible $\tilde{\alpha} \in (0,1)$

An interesting situation arises if $\tilde{\alpha} \in (0,1)$. In this region, as $(1-\tilde{\alpha})^p$ falls faster than $(1-\tilde{\alpha})$ for any $p > 1$, assuming $(1-\tilde{\alpha}) = \epsilon$ will imply $(1-\tilde{\alpha})^p = \epsilon - \omega_p^2$ where ω_p^2 is a finite positive number whose value depends on p. Putting these values in (5) gives

$$\boldsymbol{M}^p = \boldsymbol{Q}\boldsymbol{\mathcal{D}}\begin{pmatrix}\epsilon-\omega_p^2 & & \\ 1 & & \\ \vdots & & \\ 1 & & \end{pmatrix}\boldsymbol{Q}^T$$

$$= \boldsymbol{Q}\left(\boldsymbol{\mathcal{D}}\begin{pmatrix}\epsilon\\1\\\vdots\\1\end{pmatrix} - \boldsymbol{\mathcal{D}}\begin{pmatrix}\omega_p^2\\0\\\vdots\\0\end{pmatrix}\right)\boldsymbol{Q}^T$$

$$= \boldsymbol{Q}\boldsymbol{\mathcal{D}}\begin{pmatrix}\epsilon\\1\\\vdots\\1\end{pmatrix}\boldsymbol{Q}^T - \boldsymbol{Q}\boldsymbol{\mathcal{D}}\begin{pmatrix}\omega_p^2\\0\\\vdots\\0\end{pmatrix}\boldsymbol{Q}^T$$

$$= \boldsymbol{M} - \omega_p^2 q q^T$$

here q is the eigenvector of $\boldsymbol{M}$ corresponding to the eigenvalue $(1-\tilde{\alpha})$ which we know from (2) is $\hat{V}$, thus

$$\boldsymbol{M}^p = \boldsymbol{M} - \omega_p^2 \hat{V}\hat{V}^T \tag{9}$$

putting (9) in (7), we get

$$\boldsymbol{\phi_n} = \boldsymbol{\phi_0}\left(\boldsymbol{M} - \omega_n^2\hat{V}\hat{V}^T\right) + \left(1-(\epsilon-\omega_n^2)\right)\frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2}$$

$$= \boldsymbol{\phi_0 M} - \omega_n^2\boldsymbol{\phi_0}\hat{V}\hat{V}^T + (\tilde{\alpha}+\omega_n^2)\frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2}$$

or,

$$\boldsymbol{\phi_n} = \boldsymbol{\phi_0 M} + \tilde{\alpha}\frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2} - \omega_n^2\left(\boldsymbol{\phi_0}\hat{V} - \frac{\hat{y}}{\left\|\hat{V}\right\|^2}\right)\hat{V}^T \tag{10}$$

To find which value of $\tilde{\alpha}$ which best solves (P), we can calculate $\boldsymbol{\phi_n}\hat{V} - \hat{y}$ using (10)

$$\boldsymbol{\phi_n}\hat{V} - \hat{y} = \left(\boldsymbol{\phi_0 M} + \tilde{\alpha}\frac{\hat{y}\hat{V}^T}{\left\|\hat{V}\right\|^2} - \omega_n^2\left(\boldsymbol{\phi_0}\hat{V} - \frac{\hat{y}}{\left\|\hat{V}\right\|^2}\right)\hat{V}^T\right)\hat{V} - \hat{y}$$

$$= \boldsymbol{\phi_0 M}\hat{V} + \tilde{\alpha}\hat{y}\frac{\hat{V}^T\hat{V}}{\left\|\hat{V}\right\|^2} - \omega_n^2\left(\boldsymbol{\phi_0}\hat{V} - \frac{\hat{y}}{\left\|\hat{V}\right\|^2}\right)\hat{V}^T\hat{V} - \hat{y}$$

$$= (1-\tilde{\alpha})\boldsymbol{\phi_0}\hat{V} + \tilde{\alpha}\hat{y} - \omega_n^2\left(\left\|\hat{V}\right\|^2\boldsymbol{\phi_0}\hat{V} - \hat{y}\right) - \hat{y}$$

or,

$$\boldsymbol{\phi_n}\hat{V} - \hat{y} = \left(1 - \tilde{\alpha} - \omega_n^2\left\|\hat{V}\right\|^2\right)\boldsymbol{\phi_0}\hat{V} + (\tilde{\alpha} + \omega_n^2 - 1)\hat{y} \tag{11}$$

Writing ω_n^2 as $(1-\tilde{\alpha}) - (1-\tilde{\alpha})^n$ in (11), we get

$$\boldsymbol{\phi_n}\hat{V} - \hat{y} = \left((1-\tilde{\alpha}) - \left\|\hat{V}\right\|^2\left((1-\tilde{\alpha}) - (1-\tilde{\alpha})^n\right)\right)\boldsymbol{\phi_0}\hat{V}$$

$$+ (\tilde{\alpha} + (1-\tilde{\alpha}) - (1-\tilde{\alpha})^n - 1)\hat{y}$$

$$= \left((1-\tilde{\alpha})\left(1 - \left\|\hat{V}\right\|^2\right) + (1-\tilde{\alpha})^n\left\|\hat{V}\right\|^2\right)\boldsymbol{\phi_0}\hat{V} - (1-\tilde{\alpha})^n\hat{y}$$

and constraining $\left\|\hat{V}\right\|^2 = 1$, will give

$$\boldsymbol{\phi_n}\hat{V} - \hat{y} = (1-\alpha)^n\left(\boldsymbol{\phi_0}\hat{V} - \hat{y}\right) \tag{12}$$

which can become infinitesimally small with $\alpha \in (0,1)$. This shows that, with an appropriate value of α, the network can become highly adapted to any particular input. Also, under this constraint, from (8), we have

$$\boldsymbol{\Delta\phi_n} = (\alpha(1-\alpha)^{n-1})\left(\hat{y} - \boldsymbol{\phi_0}\hat{V}\right)\hat{V}^T \tag{13}$$

and from (10)

$$\boldsymbol{\phi_n} = \boldsymbol{\phi_0}\widetilde{\boldsymbol{M}} + \alpha\hat{y}\hat{V}^T - \omega_n^2\left(\boldsymbol{\phi_0}\hat{V} - \hat{y}\right)\hat{V}^T \tag{14}$$

where $\omega_n^2 = (1-\alpha) - (1-\alpha)^n$ and $\widetilde{\boldsymbol{M}} = \boldsymbol{Q\mathcal{D}}\begin{pmatrix}1-\alpha & & \\ & 1 & \\ & \vdots & \\ & & 1\end{pmatrix}\boldsymbol{Q}^T$.

It is interesting to look into matrix $\boldsymbol{Q}$ under this constraint. As it is a matrix of orthonormal eigenvectors of $\boldsymbol{M}$, it is evident that one of the columns of $\boldsymbol{Q}$ is $\hat{V}$ ($\left\|\hat{V}\right\|^2 = 1$).

The other vectors are q_i such that $\|q_i\|^2 = 1,\ \hat{V}^T q_i = 0\ \forall\ i$ and $q_i^T q_j = 0\ \forall\ i \neq j$. This allows us to express $\widetilde{\boldsymbol{M}}$ as a sum of rank-one matrices as under

$$\widetilde{\boldsymbol{M}} = (1-\alpha)\hat{V}\hat{V}^T + \sum_i q_i q_i^T \tag{15}$$

using (15) in (14) and writing $\omega_n^2 = (1-\alpha) - (1-\alpha)^n$, we get

$$\boldsymbol{\phi_n} = \boldsymbol{\phi_0}\left((1-\alpha)\hat{V}\hat{V}^T + \sum_i q_i q_i^T\right) + \alpha\hat{y}\hat{V}^T - \left((1-\alpha) - (1-\alpha)^n\right)\left(\boldsymbol{\phi_0}\hat{V} - \hat{y}\right)\hat{V}^T$$

$$= \boldsymbol{\phi_0}\left(\hat{V}\hat{V}^T + \sum_i q_i q_i^T\right) - \alpha\left(\boldsymbol{\phi_0}\hat{V} - \hat{y}\right)\hat{V}^T - \left((1-\alpha) - (1-\alpha)^n\right)\left(\boldsymbol{\phi_0}\hat{V} - \hat{y}\right)\hat{V}^T$$

$$= \boldsymbol{\phi_0}\left(\hat{V}\hat{V}^T + \sum_i q_i q_i^T\right) - (1-(1-\alpha)^n)\left(\boldsymbol{\phi_0}\hat{V} - \hat{y}\right)\hat{V}^T$$

the matrix $\left(\hat{V}\hat{V}^T + \sum_i q_i q_i^T\right)$ is nothing but $\boldsymbol{QQ}^T$ and $(\boldsymbol{QQ}^T)^p = \boldsymbol{QQ}^T$ for any $p \geq 1$ implies that $\boldsymbol{QQ}^T = \boldsymbol{I}$, this makes the above expression of $\boldsymbol{\phi_n}$ as under

$$\boldsymbol{\phi_n} = \boldsymbol{\phi_0} - (1-(1-\alpha)^n)\left(\boldsymbol{\phi_0}\hat{V} - \hat{y}\right)\hat{V}^T$$

$$= \boldsymbol{\phi_0} + (1-(1-\alpha)^n)\hat{y}\hat{V}^T - (1-(1-\alpha)^n)\boldsymbol{\phi_0}\hat{V}\hat{V}^T$$

or,

$$\boldsymbol{\phi_n} = \boldsymbol{\phi_0} + \mathcal{C}\left(\hat{y}\hat{V}^T - \boldsymbol{\phi_0}\hat{V}\hat{V}^T\right) \tag{16}$$

where $\mathcal{C}$ is a constant which equals $(1-(1-\alpha)^n)$.

Thus, after n steps of gradient descent, the change in $\boldsymbol{\phi}$ has two components, an additive component given by the rank one matrix $\hat{y}\hat{V}^T$, and a subtractive component given by the rank one matrix $\boldsymbol{\phi_0}\hat{V}\hat{V}^T$. If we analyze the matrix $\hat{y}\hat{V}^T$, we notice that the matrix will have positive entries at the location (i,j) if and only if y_i and V_j are both positive. Thus, this matrix corresponds to the Hebbian update rule that strengthens the connection when a primary neuron and a representation neuron fire together. Similarly, matrix $\hat{V}\hat{V}^T$ can be positive only when V_i and V_j are both positive. However, the negative sign before this update component makes it anti-Hebbian in nature, i.e., the update reduces all the connections between primary neurons and two similarly active representation neurons. In other words, if

two representation neurons are firing together, their input is reduced so that they can be decoupled. Overall, an update in connectivity strengthens the connections between simultaneously firing primary neurons and representation neurons but reduces the chances of two representation neurons firing together.

4.4.3. Stochastic gradient descent: Adapting to multiple stimuli in sequence

Using the update procedure described above, we could update the connectivity to adapt to a particular input. One can assume that if such an update is carried out in sequence for different inputs, the network will gradually get tuned to features from multiple inputs presented to it. This kind of adaptation is what we intend to achieve through the network. However, the task is more complicated than it appears. Updating the connections to adapt to a novel input, in the way described above, often disrupts the system's adaptation to the previously encountered inputs.

Simultaneous re-learning of features from all the previous inputs is one way to minimize the effects of such disruptions. However, this approach cannot be utilized because it increases the number of learning iterations for the system. Furthermore, it is also an overcomplicated version of the matrix factorization approach as it necessitates the system to re-learn from the entire input set simultaneously while extracting features from the last input.

The stochastic gradient descent (SGD) (Robbins and Monro 1951) is another method that can be utilized to solve this problem. As evident from the name, it is a stochastic approximation of gradient descent optimization. In this method, instead of optimizing the objective function for all the training data, one optimizes the function for only a randomly selected subset of the data. To understand this approach, imagine any optimization problem as a finite-sum problem, where the value of the objective function can be expressed as a sum of losses for each data point, i.e.,

$$f(x) = \sum_{i=1}^{N} f_i(x)$$

Here f is the objective function, f_i is the loss at the i^{th} data point and x is the optimization variable. The gradient of the objective function, then, is the gradient of this finite-sum, which is calculated with respect to every training data point.

$$\frac{\partial f(x)}{\partial x} = \sum_{i=1}^{N} \frac{\partial f_i(x)}{\partial x}$$

In contrast, in SGD, each step of descent is decided using only a subset of training data points, and hence, the gradient is decided based only on a portion of this finite-sum

$$\frac{\partial f(x)}{\partial x} \approx \sum_{j \in S} \frac{\partial f_j(x)}{\partial x} \quad \text{where} \quad S \subset [1, N]$$

Though this strategy does not reach optimum, it has been shown to reach very close to the objective function's optimum value (Bottou 1998, Kiwiel 2001).

In our network model, the objective is to update the network's connectivity so that it learns to efficiently represent a finite set of inputs based on their most informative structures. In this regard, the objective function is the measure of adaptiveness, the optimization variable is the matrix of tuning properties, and the training data points are the pairs of inputs and their corresponding representations. As a single input can be a subset of data points, we realized that the SGD method could train the network for all the inputs presented in a sequence. However, there were two points of concern while utilizing this method

1. As SGD does not reach the optimum (Kiwiel 2001, Bottou 1998), using it in our network will mean that the network is never completely adapted to any input. Although this seems troublesome, it is unlikely that brain and sensory systems adapt entirely either. In this light, this limitation might make our model closer to the biological networks of sensory processing.
2. SGD method is sensitive to step size taken during gradient descent (Goodfellow et al. 2016). As only a subset of data points are considered while estimating the

gradient, taking larger gradient steps in SGD may throw the updated point very far from the optimum. Therefore, it is advised to use only small step sizes (Goodfellow et al. 2016). In contrast, the adaptation process requires the connectivities to be updated to a particular strength to make the adaptation effective (a smaller update in connectivity may not be differentiated from unadapted connectivity), which means that a minimum step size or a minimal update is necessary. To solve this problem, we decided to update the connectivity using smaller step sizes and utilize multiple presentations of the same input to reach the desired adaptation level. These kinds of updates were more realistic and provided us with a way to understand how the frequency of inputs affected the adaptation process.

4.4.4. Simulating the network

One of the limitations of the matrix factorization approach was its inability to represent inputs not included in the input matrix. Separate algorithms were utilized for the sparse recovery of inputs. In contrast, our network could perform both these tasks. Hopfield network-like architecture allowed it to solve sparse recovery problems for a given input, and the connectivity between primary and representation neurons could be updated using the SGD method. These two tasks were performed in two modes of the network described below

1. Mode 0: In this mode, the network only performed a sparse recovery. The connectivity between the primary and representation neurons and the input were given as arguments to the network. It produced the desired representation. No update in connectivity was performed in this mode.
2. Mode 1: This was the mode in which the network performed both sparse recovery and basis adaptation. Initial connectivity and input were given as arguments to

the network. The network produced a sparse representation of the input. The connections between various neurons in the network were updated using the obtained representation and the corresponding input to ensure learning.

The network was written as a MATLAB function, and both these modes were used as per the simulation requirements.

4.4.5. Data set

For studying the network implementation of our framework, we utilized the same set of binary symbols that were previously used with the matrix factorization approach.

4.4.6. Image corruption

The set of symbols was corrupted to different extents by flipping different fractions of pixels. One hundred different forms of corruption of the same level were produced by flipping random subsets of pixels.

4.5. Results

To test the working of the network and its adaptiveness to the inputs compared to the matrix factorization approach, we decided to use a simplistic network of 256 primary neurons and 500 representation neurons. Like before, as the binary symbols data set provided an easy method to perform quantifications, the first set of analyses were performed using them only. The results of these analyses are described in the next few sections.

4.5.1. Effects of biasing the network

As discussed in the previous section, to have an adaptive nature, it is imperative for the system (**Figure 4.3.1**) to differentiate between the inputs. Moreover, how well the system adapts to two different inputs depends on how well it can differentiate the inputs before adaptation. We first decided to test how biasing the connectivity, discussed in the previous section, shapes the representation neurons' correlation. To test that and compare it against the connectivity generally used in prevailing neuronal networks, we utilized three different models of connectivity as listed below

1. **Non-negative uniform connectivity:** In this model of connectivity, the connection strengths between the primary and representation neurons were chosen to be values between 0 and 1. The probability of a connection strength attaining any value was the same, i.e., the connection weights were derived from a uniform distribution over (0, 1) (**Figure 4.3.2.i**). The weights were normalized such that the length of the weight vector corresponding to any representation neuron was 1.

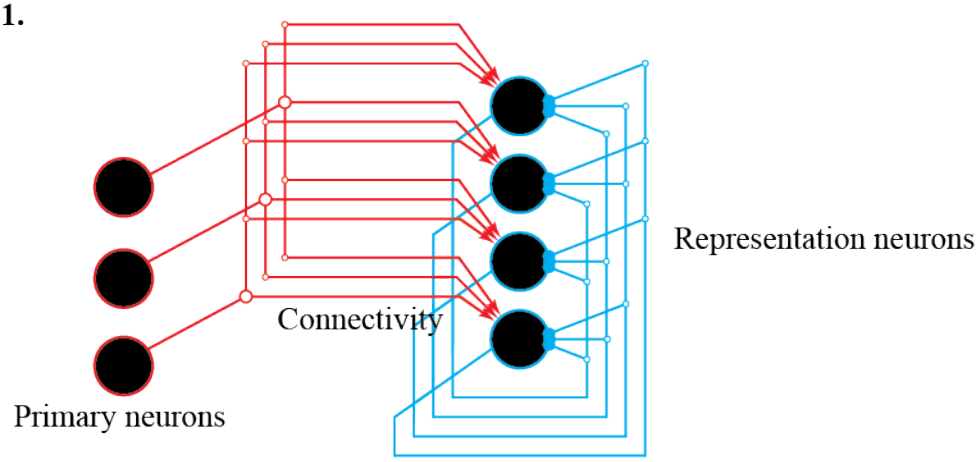
2. **Normally distributed connectivity:** In this model of connectivity, the weights were derived from a normal distribution with mean 0 and standard deviation 1. Like the uniformly distributed weights, these weights were then normalized to have length 1 (**Figure 4.3.2.ii**).

3. **Decorrelating connectivity:** We refer to our biased connectivity as decorrelating connectivity. The weights were normalized in this case too to have length 1. The decorrelation was based on the eigenvectors of the variance-covariance matrix of the inputs. It was observed that the variance of the input space along these vectors saturated

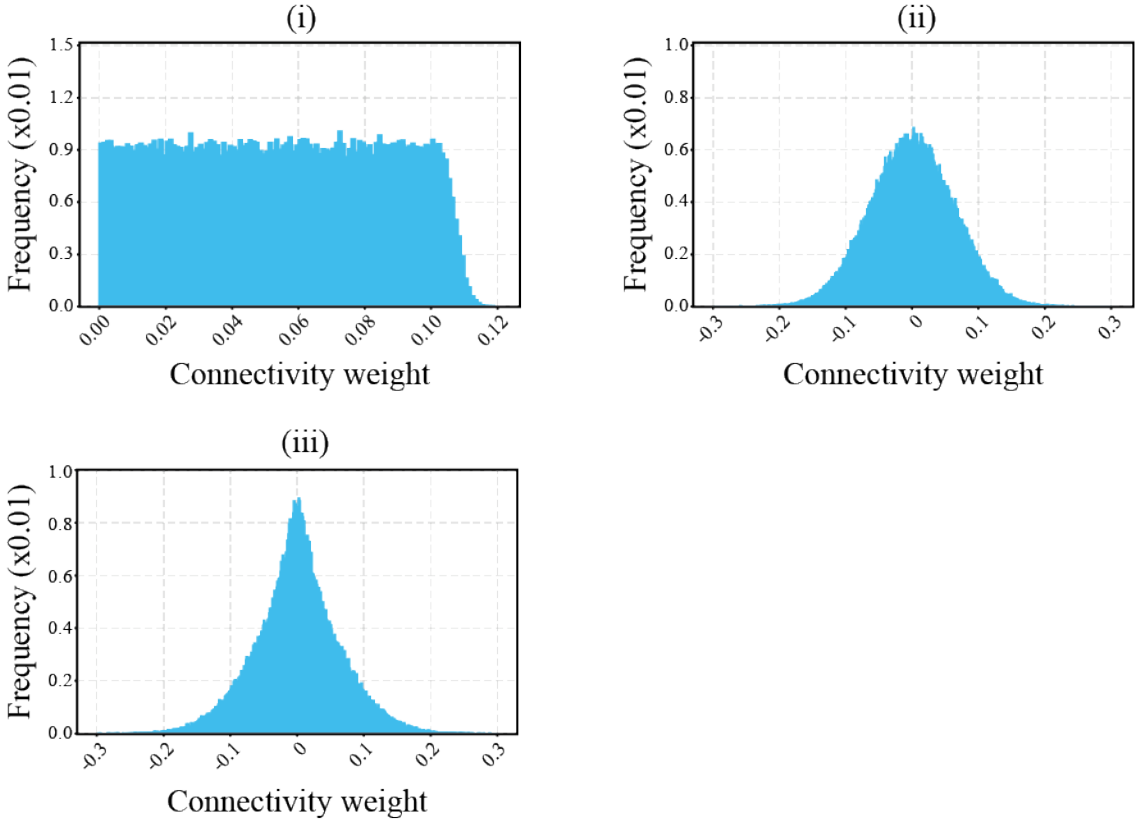
after 150 dimensions; therefore, only 150 eigenvectors were utilized as affective dimensions of the input space (**Figure 4.3.2.iii**).

We obtained symbols' representations in all these networks with different connections without any adaptations. We then used the response profiles of representation neurons to calculate pair-wise correlations between them. The Frobenius norm of the correlation and identity matrices' difference was calculated to measure the difference between the two matrices (**Figure 4.3.3**). A lower norm indicated that the biased connectivity produced better decorrelation than the other models of connectivity. A sample of the decorrelating connectivity showed that the connections did not have any apparent structure (**Figure 4.3.4**).

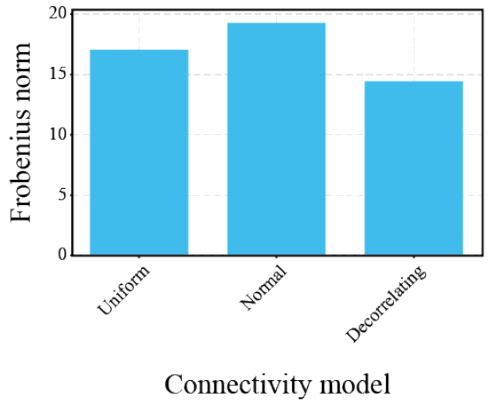
1.


Representation neurons
Connectivity
Primary neurons

2.


(i)
Frequency (x0.01)
Connectivity weight
(ii)
Frequency (x0.01)
Connectivity weight
(iii)
Frequency (x0.01)
Connectivity weight

3.



4.

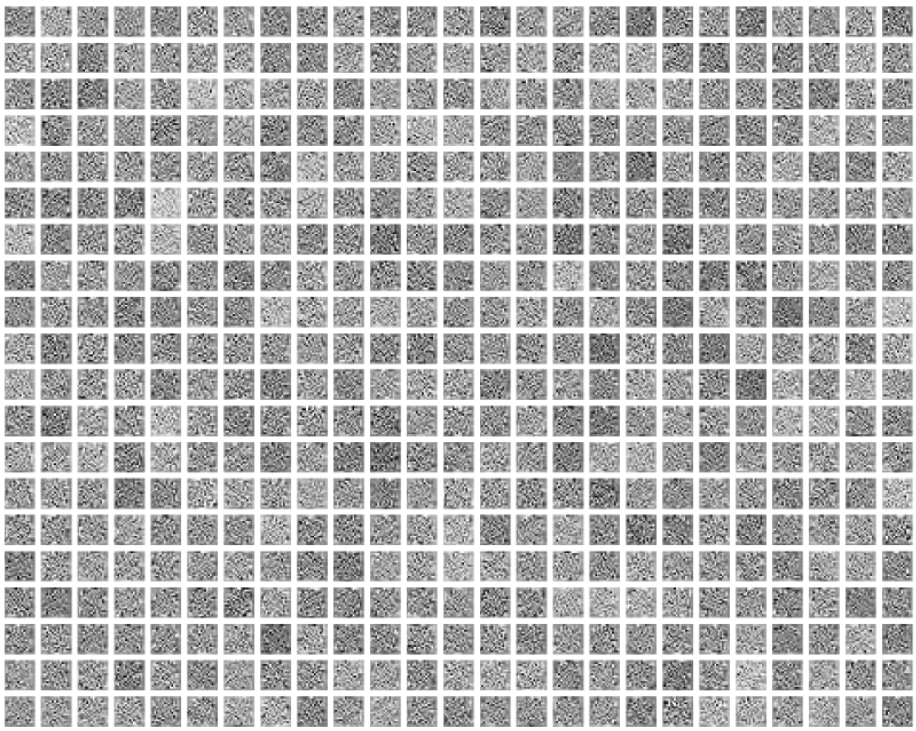


Figure 4. 3: Effects of biasing the connectivity of the network

1. *A diagram of the model neuronal circuit. The primary neuron units are shown as black circles with red outlines, and the representation neurons are shown as black circles with blue outlines. All primary neurons connect to all representation neurons (red connections), which also have recurrent inhibitory connections (blue connection). The primary neurons do not have recurrent connections.* **2.** *Three different ways of connecting primary neurons to representation neurons were considered. Shown are distributions of connectivity weights under various connectivity schemes (i) uniformly distributed weights, (ii) normally*

distributed weights, (iii) decorrelating weights. **3.** *With initial weights set according to the three models, representations of different inputs were obtained, and the correlation among representation neurons was measured. The Frobenius norm of the difference between the correlation matrix and the identity matrix was lowest for the decorrelating model of connectivity, indicating that it could decorrelate the neurons most.* **4.** *The connection weights to all 500 representation neurons in the decorrelation model are shown in the form of a grayscale image array. 500 images in the panel correspond to 500 representation neurons, and pixels in each image correspond to the primary neurons. The pixel's grayscale value is proportional to the connection strength between the primary and the representation neuron. Note that even though connections are designed to decorrelate the neurons, no apparent structure emerges in the connectivity.*

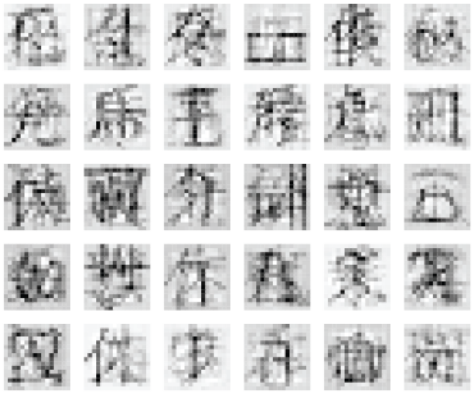
4.5.2. The adapting nature of the network

To test the network's adaptive nature, we allowed the network to learn features from a varying number of inputs while keeping the number of neurons fixed. In particular, three overlapping sets containing 500, 800, and 1000 inputs were presented to the network, and the network adapted to these inputs. As discussed in the previous section, each input was presented repeatedly (100 times at maximum) to allow for SGD type adaptation. Note that the inputs were presented one at a time in a sequence. The order of their presentation was randomly chosen every time. The state of the network was recorded after the presentation of the entire set of inputs. The different states of adaptations of the network to 500 inputs are shown as the changes in the networks' connectivity (**Figure 4.4.1**). These changes were calculated with respect to the initial decorrelating connectivity and represented how strongly a particular neuron is connected to primary layer neurons. As an input neuron strongly

connected to a representational unit will elicit a maximum response in that representation neuron, these connections essentially reflected the representation neurons' tuning properties. The first thing to notice is that different neurons get tuned to different structures from the inputs. We plotted the distribution of cosine similarity of the connectivity changes for different neurons across different states and found that connectivity similarity was maintained while repeatedly encountering symbols (**Figure 4.4.2**). A sustained similarity level indicated that the distinctiveness of neuronal tunings remained unaltered. However, these similarity measures gave an idea only of the overall connectivity changes in a particular state. They did not provide information about how connectivity changed for individual neurons across different states. To assess that, we analyzed the changes in connectivity to individual neurons across different states of adaptation. We found that while connectivity structure did not change for individual neurons, the similarity of connectivity to neurons increased slightly over states and then saturated (**Figure 4.4.3**). This change illustrated that the connections to individual representation neurons were slightly changing as inputs were encountered repeatedly and then reached a stable state after a certain number of encounters. Attainment of such a stable state in neurons' connectivity demonstrated how the adaptation of the network saturated. As only the first few encounters of any input changed the structure of connectivity, it could be inferred that the representations of the inputs changed based on the immediate experience of the network and saturated afterward. This saturation highlights the critical difference between our framework and the classical efficient coding paradigm, where the representations of inputs depend upon their overall statistical and not just immediate encounters.

1.

(i)

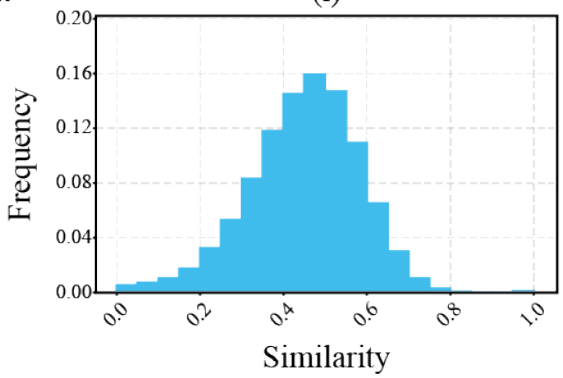


(ii)

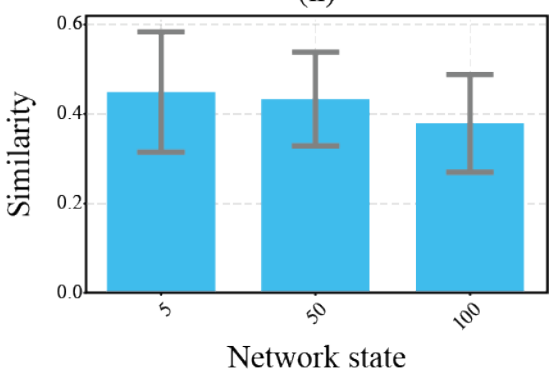
(iii)

2.

(i)



(ii)



3.

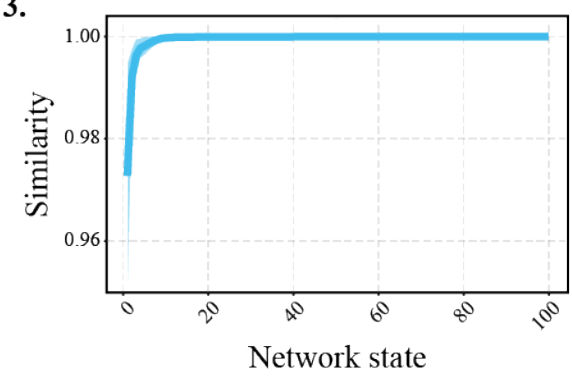


Figure 4. 4: The adaptation properties of the network

1. *A set of 500 inputs was repeatedly presented to the network, and the network's state was recorded after the presentation of the complete set. The change in connectivity of network at different states ((i) 5th state, (ii) 50th state, and (iii) 100th state) is shown.* **2.** *The similarity between changes in connectivity across all representation neurons was analyzed. (i) The distribution of similarity is shown. A low average similarity (< 0.5) indicated that the connections of different neurons changed differently. (ii) The similarity in the changes in connectivity was monitored at different states of the network. The average similarity remained consistently small and slightly decreased with the state.* **3.** *How connectivity changed for a single neuron across different states was also observed. The plot shows the similarity between connectivity changes to all representation neurons as the simulation progressed. The similarity increased a bit and reached a saturation state, indicating that the network was saturated after encountering a certain number of the inputs' repeats.*

Next, we analyzed the structural changes in the connectivity. We compared the structure of changes in connectivity to the input patterns. We found that with an increasing number of input encounters, the structures became more input-like (**Figure 4.4.1**). We further analyzed the changes in connectivity to representation neurons with a varying number of symbols. Sample connectivity changes when the network adapted to 500, 800 and 1000 symbols are shown (**Figure 4.5.1**). We measured the cosine similarity between the inputs and the changes in connectivity structure with varying numbers of inputs. We found that the similarity increased with increasing the number of inputs across all network states (**Figure 4.5.2**). Such similarities indicated that connectivity change structures became less like unique inputs and more local when the number of inputs was increased. In terms of informativeness of captured features, these results showed that as the network encountered the same inputs repeatedly, it successfully identified comprehensive, unique structures from

the inputs. Increasing the number of inputs, however, resulted in neurons getting tuned to more localized structures.

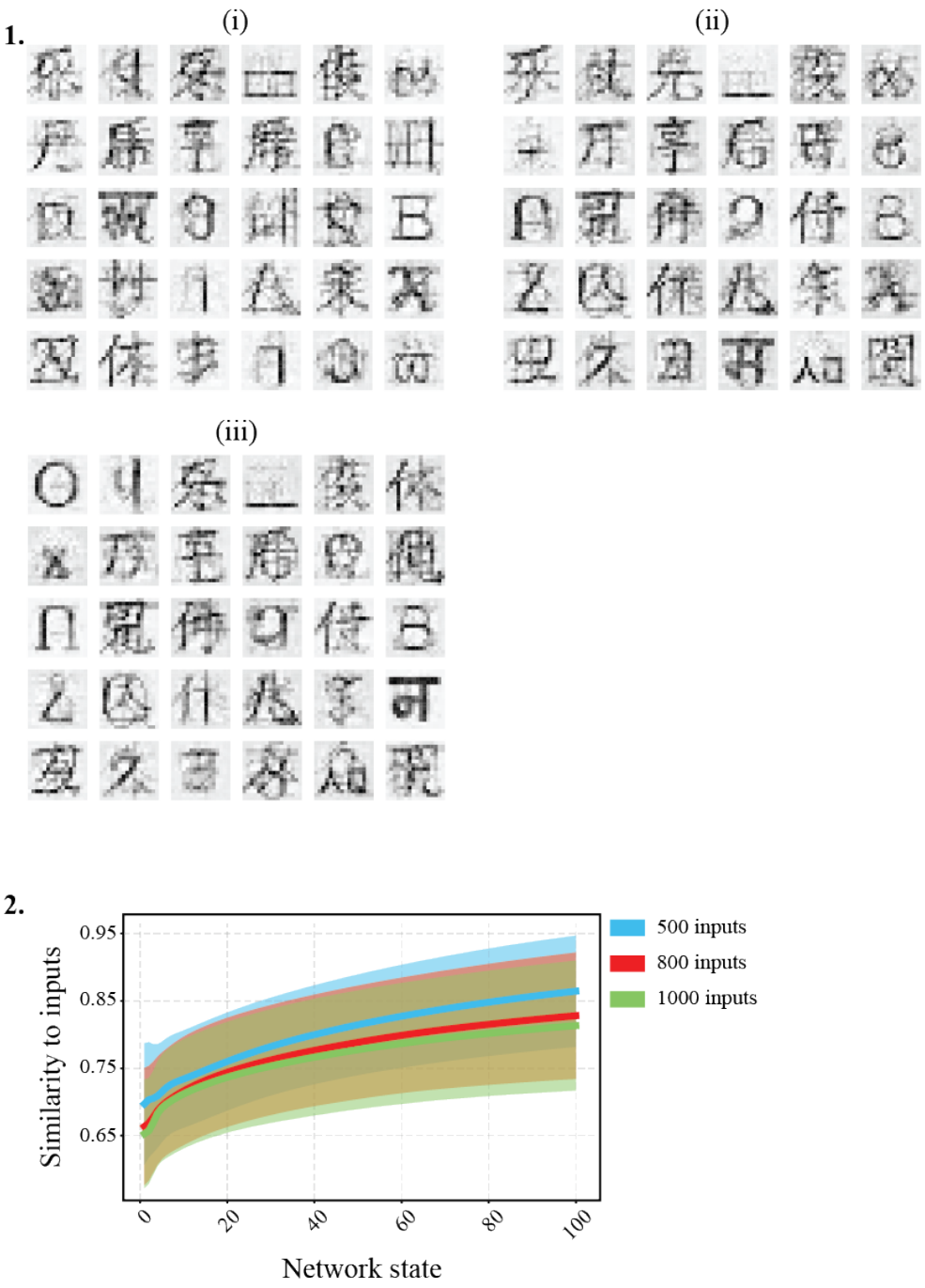


Figure 4. 5: Analysis of the structure of connectivity changes

1. *To analyze the changes in the network structure with network states and the input numbers, three different sets of inputs containing 500, 800, and 1000 symbols were*

presented repeatedly to the network. The changes in the connections of the network at the 50th state are shown ((i) 500 inputs, (ii) 800 inputs, and (iii) 1000 inputs). **2.** *It was observed that the changes in connectivity were similar to the structure of the inputs. The cosine similarity between changes in connectivity and input was measured at different stages. The similarity increased with the network state but decreased with the increasing number of inputs. The line represents the average similarity, and the band is the standard deviation in the observed similarity.*

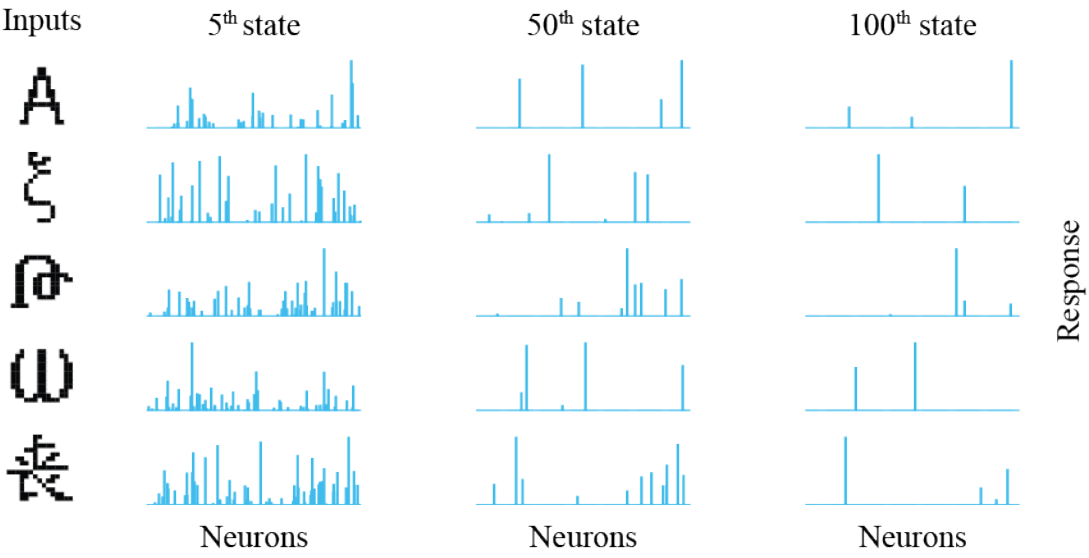
4.5.3. Efficiency of representations

We analyzed the network's efficiency at different adaptation states while representing a varying number of inputs. Responses of all representation neurons to a set of selected inputs are shown at different adaptation states (**Figure 4.6.1**) and when the network adapted to a different number of inputs (**Figure 4.6.2**). We noticed that with more encounters of the inputs, the representations became sparser. Similarly, with the increasing number of inputs, the responses got confined to a smaller number of neurons. We quantified the representation efficiency to further highlight the changes that occurred while adapting to a varying number of inputs. Three quantities, namely, response profiles' correlation, kurtosis, and sparsity, were measured across different states of the network, as well as across the different numbers of inputs. We found that as the network experienced more inputs, the neurons' response became increasingly non-Gaussian (**Figure 4.6.3**). Increasing the number of input presentations also increased the kurtosis of neuronal response profiles. These trends indicated that both experience and sampling of inputs increased representation efficiency. The correlation among the neurons further confirmed the increase in representation efficiency. Following the same trend as kurtosis, it decreased (indicated by the smaller Frobenius norm of the difference of correlation and identity matrices) with more encounters

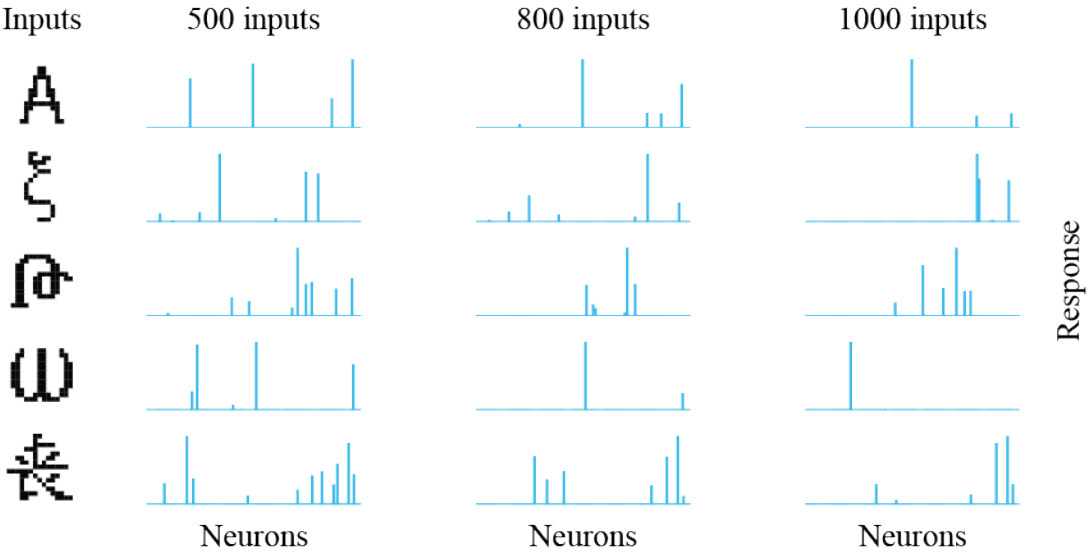
of the same set of inputs, as well as encounters of new inputs (**Figure 4.6.4**). Similar trends were observed in the L0 and L1 sparsity measures (**Figure 4.6.5**; **Figure 4.6.6**).

For a biological system, repeatedly facing the same inputs is equivalent to the increased practicing of identification. Encountering additional inputs corresponds to newer experiences. One expects that the proficiency of an organism in performing any task increases with practice and experience. Results described above indicate that the network produces increasingly sparse and unique input representations as it gets more practiced and experienced. As distinct representations demonstrate enhanced recognition abilities, such a network's behavior is more similar to a biological system. Interestingly, the results do not match the ones obtained through the matrix factorization approach, where the efficiency in representation dropped with increasing inputs.

1. Inputs



2. Inputs



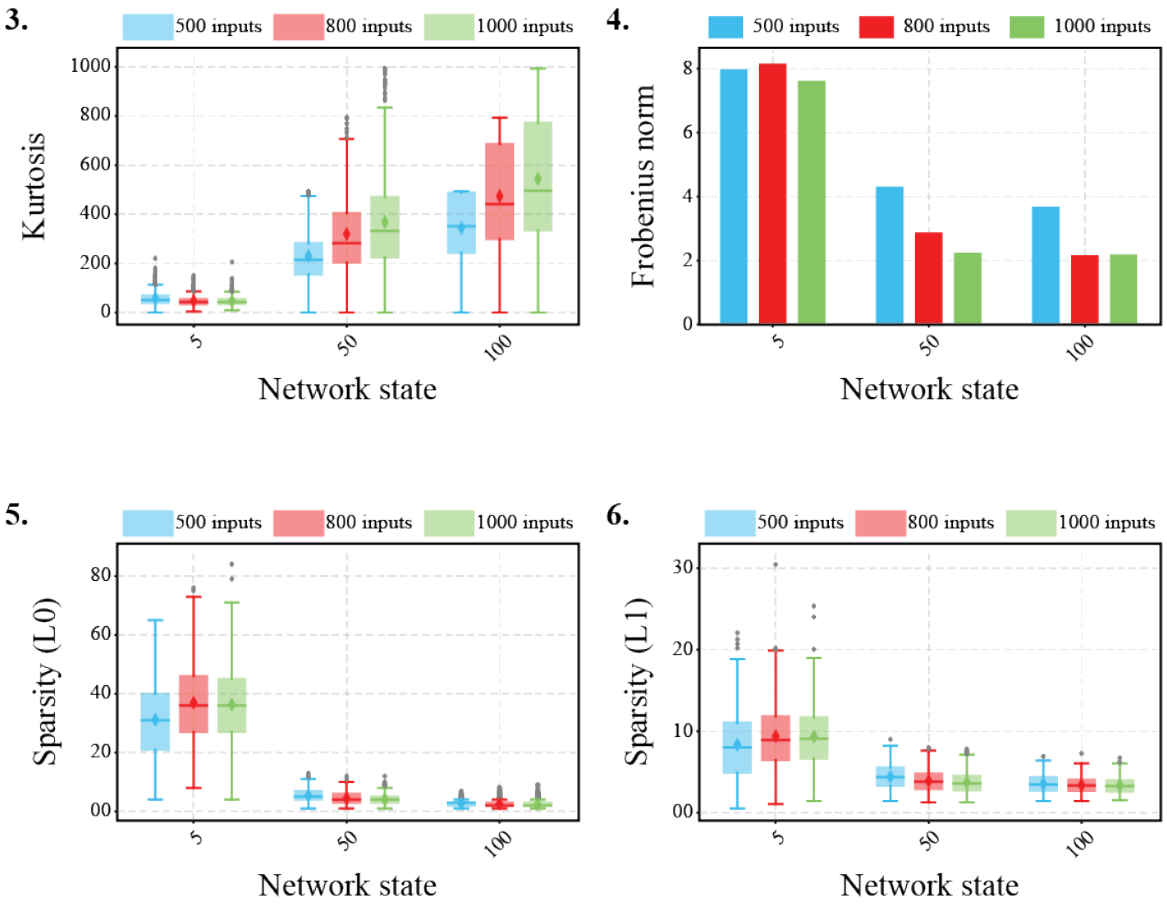


Figure 4. 6: Efficiency of representation

1 – 2. *Changes in representation efficiency with network states and a varying number of inputs were analyzed.* **1.** *Responses of 500 neurons to a few sample inputs at different adaptation states are shown.* **2.** *The neurons' responses to the same inputs when the number of inputs was varied. Note that the responses get sparser with the adaptation states as well as with the number of inputs.* **3.** *To further assess the efficiency in terms of representation sparseness, neuronal response profiles' kurtosis was calculated. Kurtosis increased with the network states as well as the number of inputs.* **4.** *The correlation among neurons was measured, and the Frobenius norm of the difference between correlation and identity matrices was calculated. The norm too decreased with the states and the number of inputs, indicating a decorrelation trend.* **5 – 6.** *The sparsity of representations also showed similar*

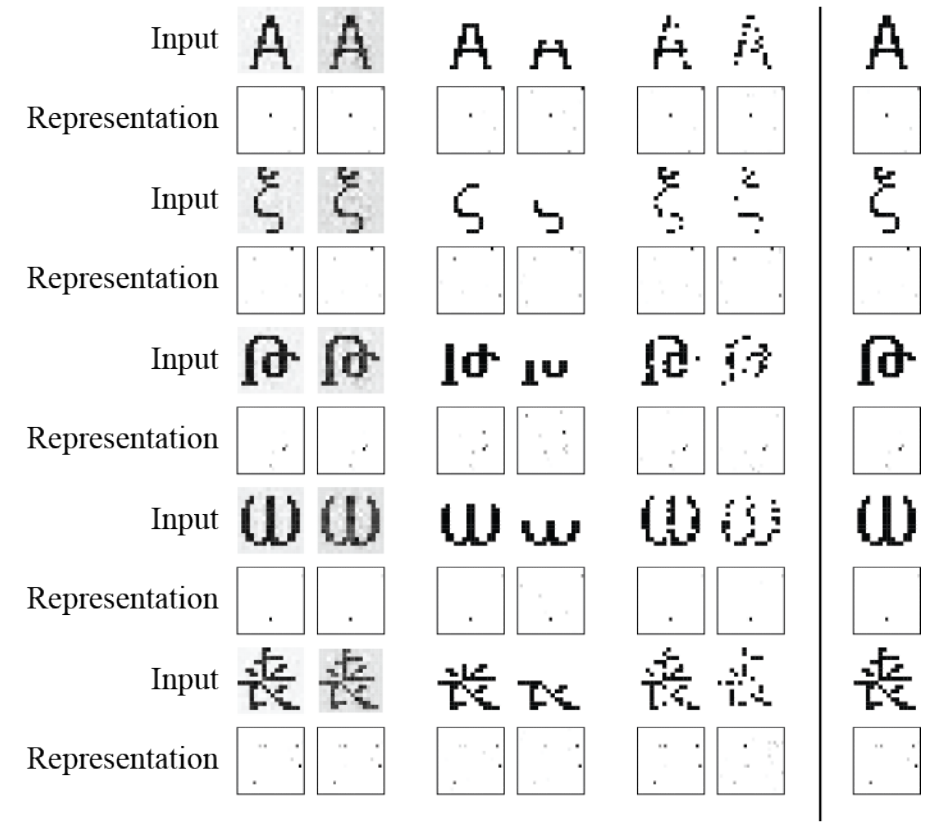
trends. Both the L0 and L1 sparsity measures decreased with the network state while maintaining the levels across the number of inputs.

4.5.4. Consistency in representations

With networks performing in more biologically realistic ways, we wanted to know how consistently the input's corrupted forms can be represented. Different forms of corrupted inputs that were used during the analysis of the matrix factorization approach were chosen, and their corresponding representations were obtained using a network adapted to 800 inputs. We observe that across all types of corruption, consistent representations could be obtained at different network states. The examples show representations of 5 different inputs and their corrupted forms (**Figure 4.7.1 and Figure 4.7.2**). Note that the representations are consistent across different forms of corruption and across different states of the network. Using the z-scored cosine similarity between the representations of uncorrupted and corrupted inputs, we calculated the specificity of representations for different forms of corruption (**Figure 4.7.3**; **Figure 4.7.4**; **Figure 4.7.5**). We found that the specificity increased slightly with practice, i.e., after encountering the inputs a greater number of times. This trend was observed consistently for all forms of corruption. The specificity decreased with increasing levels of corruption, occlusion, or addition of noise. These results indicated that the representations' consistency increased with the representation neurons getting more specific by getting tuned to unique features from the inputs.

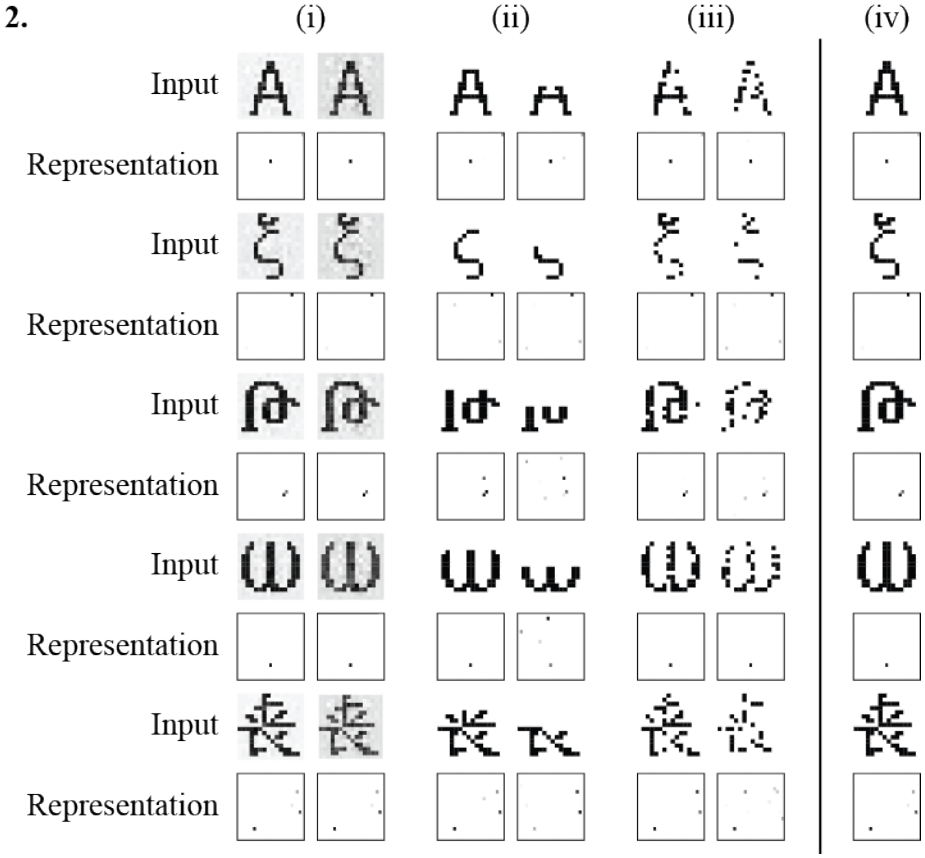
50th state

1.



100th state

2.



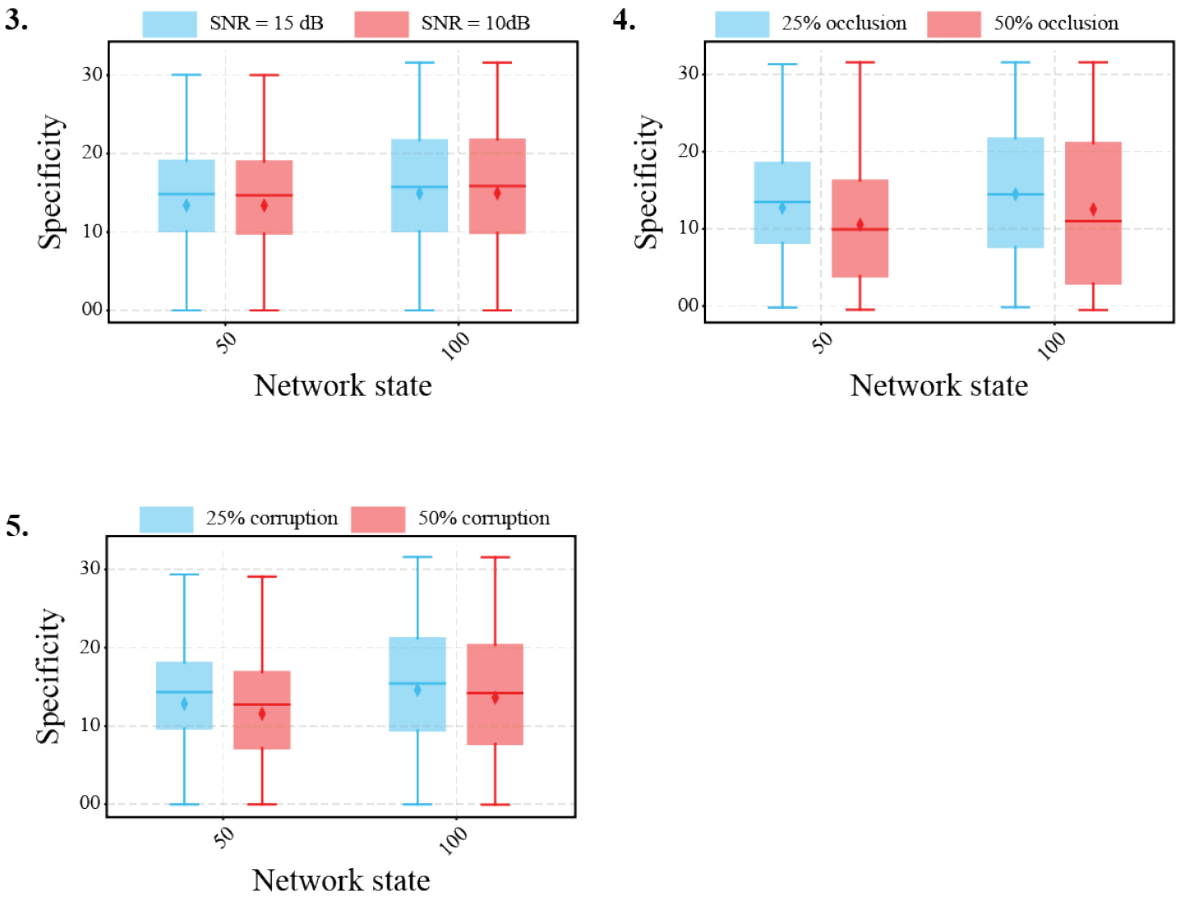


Figure 4. 7: Consistency in representations

1 – 2. *To test the consistency of representations, we presented corrupted forms of inputs to a network adapted to 800 inputs. The corrupted forms of inputs and their corresponding representations are shown at two different states of the network. For all forms of corruption, namely, (i) addition of noise, (ii) removal of a portion of primary neurons, and (iii) randomly silencing primary neurons, representations consistent with the uncorrupted input (iv) could be obtained in both states of the network. Note that the representations in the 100th state are sparser than the representations in the 50th state.* **3 – 5.** *The specificity of representations was measured by using z-scored similarity as described previously. Across all forms of corruption, the high specificity of representations was observed with a slight increase in the network's 100th state. The specificity scores dropped with an increase in corruption.*

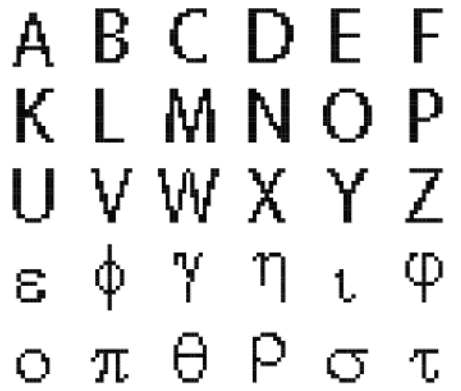
4.5.5. Learning from corrupted examples

In the previous results, we have shown that an adaptive network that sequentially encounters inputs can adapt to them and produce efficient representations. It does not need to know the entire input space's statistics to be efficient and can produce consistent representations of inputs under varying circumstances. However, for a real biological system, experience does not only mean sequentially encountering inputs; it also includes encountering inputs in different forms. For example, consider a cup. One encounters cups almost every day in different shapes and sizes. All of them could be considered variations of an *"ideal cup"* that is probably never seen, yet we can all draw an *"ideal cup"* when asked to do so. This ability means that our sensory system generalizes the concept of a cup by looking at different variants of it. We decided to test whether the network can similarly generalize concepts. To try this, we produced different variations of the input symbols by randomly flipping the values of a fraction of its pixels. These corrupted symbols were now used as input sets to allow the network to adapt. We used two different flipping extents (10% and 20%) to produce the corrupted inputs (**Figure 4.8.1**). Different corrupted forms at the same level of flipping were presented to the network for each adaptation session. Again, we examined the network's adaptation as the change in the connectivity of the representation neurons. To our surprise, we found that the change in connectivity resembled uncorrupted inputs just as observed in the case of adaptation to non-corrupted symbols (**Figure 4.8.2**). We further quantified the similarity between connectivity changes and uncorrupted inputs. While the similarity varied from input to input, the maximum similarity observed with any input was considerably high (**Figure 4.8.3**). Thus, the network was able to find the consistency that existed across the input variants and adapt to it. Such adaptation is rare, and only complex deep or convoluted neural networks have been shown to perform in this manner (Vogelsang et al. 2018). However, these networks are very complex, contain

multiple layers, and require numerous examples. On the other hand, our adaptive network consisting of only two layers, and learning from 800 examples can perform similarly.

1.

Original inputs

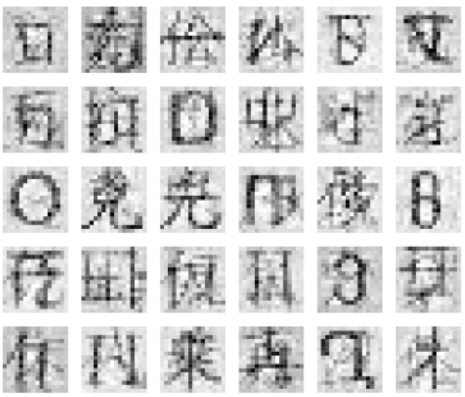


Corrupted inputs
(10% flipping)

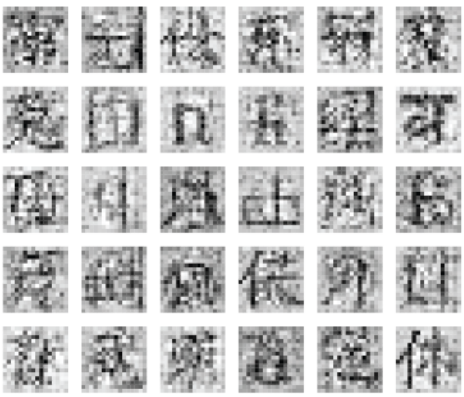
Corrupted inputs
(20% flipping)

2.

(i)



(ii)



3.

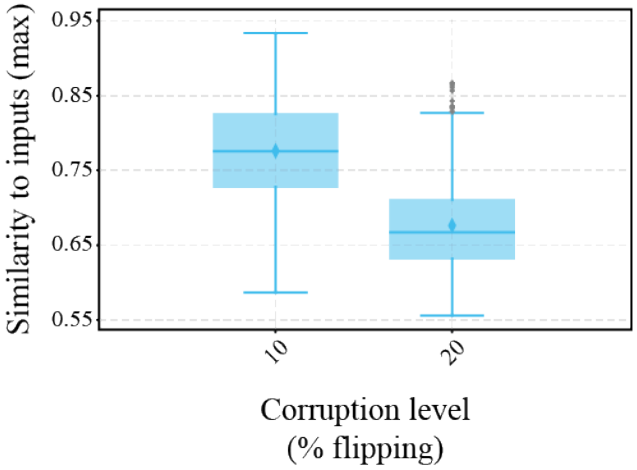


Figure 4. 8: Adapting to corrupted forms of inputs

1. *Original inputs were corrupted to different extents by flipping fractions of their pixels. Examples of original symbols and their corrupted forms obtained after flipping 10% and 20% of the pixels are shown.* **2.** *These corrupted forms were used as inputs to the network. Each symbol was presented 100 times to the network; however, each presentation had a different corruption pattern. The changes in connectivity observed resembled the structure of uncorrupted inputs. (i). Changes in connectivity after adapting to inputs having 10% of their pixels flipped. (ii). Changes in connectivity after adapting to inputs having 20% of their pixels flipped.* **3.** *The structure of changes in connectivity was compared with the uncorrupted inputs. The maximum observed similarity of connectivity of each neuron to any symbol is plotted.*

4.6. Discussion

In this chapter, I introduced a network that could extract unique features from inputs in an experience-dependent manner and generate sparse, efficient representations of the inputs based on such structures. The network was based on the previously developed class

of Hopfield networks that could perform sparse recovery of signals. However, in contrast to those networks, ours was designed to be adaptive. In other words, unlike other networks, the connectivity between the input layer and the representation layer was allowed to change based on the input to optimize its representation. A crucial aspect of the network was using the stochastic gradient descent (SGD) (Robbins and Monro 1951) type approach. Theoretically, while adapting to a finite set of inputs, the aim is to reduce some measure of non-adaptiveness of the network for the entire set of inputs. Achieving such a goal is challenging, particularly in an experience-dependent manner. The adaptation to previously encountered inputs is influenced while adapting to the current input. Using the SGD-like approach, we allowed the network to slowly adapt to new inputs so that its adaptation to other inputs was not affected. With repeated encounters, the network could adapt to all different inputs. A limitation of such an approach is that an optimal adaptation may not be achieved. However, as argued previously, the goal for sensory systems might not be to achieve the optimal but to adapt to an extent and extract enough information that ensures survival.

The variation in the efficiency of the network with repeated encounters and the number of inputs were analyzed. We found that both these parameters increase efficiency. This aspect of the network was particularly intriguing. We had seen with the matrix factorization approach that the efficiency decreased with the number of inputs. Such results were expected in that approach because the method did not consider the inputs' occurrence frequencies. However, the behavior of the network is more similar to a real biological system for which both repetition and new encounters are expected to enhance the recognition abilities.

The increase in the network's representation efficiency with the increasing number of inputs can be explained if one considers its efficiency to be suboptimal in all coding scenarios. In this situation, adapting to a larger number of inputs can cause the network to contain more information about the inputs. Accommodating more information will lead to

proper utilization of the network's capacity and increase its efficiency. Interestingly, such a scenario implies the system does not achieve the efficiency sought in the Efficient Coding Hypothesis. Indeed, for a biological system achieving such efficiency may not be critical. It may be geared more towards extracting the relevant information than relaying all information.

Lastly, we demonstrated that corrupted inputs could be utilized to guide the learning process of the network. However, the changes observed in the network connectivity were similar to the changes observed while adapting to uncorrupted forms of inputs. As corrupted forms of the inputs were created by introducing noise in the form of random silencing and activation of early neurons, this result indicated the network's ability to extract consistency from the inputs while identifying individual differences among them. Such a capability is desired in any system trying to achieve competence in recognizing inputs.

Our network is based on the Hopfield networks (Hopfield 1982, Hopfield 1984) and their variants performing sparse recovery (Rozell et al. 2008). However, it is significantly distinct in its form and function. The following table highlights critical differences among these networks

	Hopfield Networks	**Sparse recovery network**	**Our network**
Recurrent connections	only Hebbian	Only anti-Hebbian but non – adaptive	Only anti-Hebbian, Adaptive
Connection from primary neurons	Non – adaptive	Non – adaptive	Adaptive

Table 4. 1: Differences between our network, Hopfield networks, and sparse recovery network

This page is intentionally left blank

CHAPTER 5

Conclusions

Table of Contents

5.1. Conclusions

In this study, I have studied how information about the surroundings can be conveyed through the sensory systems. Organisms can utilize this information to their advantage and generate appropriate responses to different stimuli. As highlighted in the study, two fundamental tasks that a biological system must perform to survive are invariably recognizing objects and identifying relationships among objects. Accomplishing these tasks is by no means straightforward, especially for a biological system that faces several physiological and anatomical constraints. Yet, organisms have evolved into information processing systems that remain to be matched by any human-made machine. The previous studies on sensory processing have primarily focused on the theoretical aspects of relaying information. They have thought of the sensory system as an optimal information transmitting device. This picture of a biological system might not be very accurate. A biological system's primary goal is not conveying information but is utilizing the information to increase the chances of its survival. In my thesis work, I have presented a novel sensory processing strategy, suggesting that the system should adapt specifically to a finite set of inputs that it experiences and represent them using their most informative components. Using mathematical simulations, I have analyzed various aspects of this representation framework. Some of the key conclusions are listed in this chapter.

5.1.1 An adaptive strategy of representing inputs should be based on informativeness:

Based on the Efficient Coding Hypothesis, a significant section of the previous studies have argued that the optimal strategy for representing sensory inputs should be redundancy reducing (Barlow 1961). The system needs to know its environment's statistical properties to realize this strategy. It should identify the independent components of natural stimuli and use them as the basis for representing objects

(Barlow 1987, Barlow 1989, Barlow et al. 1989). It has been assumed that knowledge about the surroundings' statistics is incorporated into the system through the development process guided by the organism's genetics. The organism learns only the components that have not been incorporated into its developmental process (Barlow 1987). In contrast, I have shown that a representation framework based on the most informative components of objects allows an organism to continuously learns about its surroundings in an experience-dependent manner. The organism does not need to know its environment completely for adaptation, allowing it to accommodate unexpected changes. Moreover, the framework efficiently represents information about objects, and minimal redundancy is observed in input representations. Thus, the representation framework based on informative features allows the system to be genuinely adaptive while being efficient.

5.1.2 The number of inputs relative to neurons determines representation efficiency:

I have argued that the objects' most informative components can be extracted using non-negative matrix factorization. Representations based on these components are maximally efficient when the number of representation neurons matches the inputs. Any deviation from this balance results in a decrease in representations' sparseness, resulting in inefficient information transmission. By analyzing the neurons' tuning properties, I found that a critical difference between most sparse and less sparse representations is that in less sparse representations, neurons are tuned to local features of the input. In contrast, most sparse representations arise when neurons are tuned to complete structures of the input. In this context, I have demonstrated that the localized receptive field properties observed in the visual cortices (Hubel and Wiesel 1962, Hubel and Wiesel 1968) can be accounted for by the necessity of representing a large number of natural scene images with a relatively smaller number of neurons.

5.1.3 Inputs' absolute occurrence frequencies can be ignored: The frequency of occurrence of inputs constitutes the statistics of the inputs, and in all previous studies, it has played a crucial role in determining the representation strategy. Informativeness of features, on the other hand, does not necessarily rely on their absolute occurrence. Unique features are most informative irrespective of the occurrence frequency, and informativeness of other features can be estimated from their relative abundance. In this regard, the system does not need to know the actual occurrence frequencies in any situation. The absolute frequencies can be ignored, and efficient representations can be obtained based on the relative abundances.

5.1.4 Representations based on informative features are consistent: Utilizing sparse recovery approaches to derive representations of the inputs, I showed that consistent representations of corrupted forms of the inputs could be obtained when representations are sufficiently sparse. Inputs corrupted by the addition of noise, removal, or occlusion of primary neurons or random silencing of primary neurons all produced representations that were highly similar to the non-corrupted inputs' representations. Even for complex inputs like faces, occlusion of different portions, or common alterations like the addition of glasses or beards did not change the representations. Furthermore, the odor representations obtained from the glomeruli recordings of the mouse olfactory system remained unaltered when a portion of glomeruli was removed or when a noise was added into the system. Interestingly, in all situations, the responses of representation neurons corresponded to the mutual information between their tuning properties and the input. This relation indicated that the representations could be utilized in higher-order cognitive functions like recognition and identifying associations between inputs.

5.1.5 A neuronal network can implement the adaptive strategy of encoding: Finally, I designed a neuronal network to show that this adaptive strategy of representation based on the most informative features could be achieved in a biologically relevant network. The analysis of representations obtained in the network showed that the network could efficiently represent inputs. Consistent representations were also obtained from different forms of corrupted inputs. Moreover, the network could also utilize the corrupted input forms for adaptation, and even while using the corrupted inputs, the network neurons got tuned to structures from uncorrupted inputs.

References

Abrard, F. and Deville, Y. (2005) 'A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources', *Signal processing,* 85(7), pp 1389-1403.

Albouy, B. and Deville, Y. (2001) *Improving noisy speech recognition with blind source separation methods: validation with artificial mixtures,* In Proceedings of the 5th International Workshop on Electronics, Control, Modeling, Measurement and Signals, Toulouse, France: vol 30.

Amari, S.-i., Cichocki, A. and Yang, H. H. (1996) *A new learning algorithm for blind signal separation,* In Advances in neural information processing systems: pp 757-763.

Anderson, C. H. and Van Essen, D. C. (1987) 'Shifter circuits: a computational strategy for dynamic aspects of visual processing', *Proceedings of the National Academy of Sciences,* 84(17), pp 6297-6301.

Ans, B., Herault, J. and Jutten, C. (1985) *Adaptive neural architectures: Detection of primitives,* In Proceedings of COGNITIVA'85, Paris, France: pp 593-597.

Arberet, S., Gribonval, R. and Bimbot, F. (2006) *A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture,* In Rosca, J., Erdogmus, D., Príncipe, J. C. and Haykin, S. eds., International Conference on Independent Component Analysis and Signal Separation - ICA 2006, Charleston, SC, USA Lecture Notes in Computer Science, vol 3889: Springer, Berlin, Heidelberg, pp 536-543.

Argaez, M., Ramirez, C. and Sanchez, R. (2011) *An ℓ 1-algorithm for underdetermined systems and applications,* In 2011 Annual Meeting of the North American Fuzzy Information Processing Society, El Paso, TX, USA: IEEE, pp 1-6.

Atick, J. J. (1992) 'Could information theory provide an ecological theory of sensory processing?', *Network: Computation in neural systems,* 3(2), pp 213-251.

Atick, J. J., Li, Z. and Redlich, A. N. (1992) 'Understanding retinal color coding from first principles', *Neural computation,* 4(4), pp 559-572.

Atick, J. J. and Redlich, A. N. (1990) 'Towards a theory of early visual processing', *Neural computation,* 2(3), pp 308-320.

Atick, J. J. and Redlich, A. N. (1992) 'What does the retina know about natural scenes?', *Neural computation,* 4(2), pp 196-210.

Attneave, F. (1954) 'Some informational aspects of visual perception', *Psychological review,* 61(3), pp 183.

Attwell, D. and Laughlin, S. B. (2001) 'An energy budget for signaling in the grey matter of the brain', *Journal of Cerebral Blood Flow & Metabolism,* 21(10), pp 1133-1145.

Baddeley, R. (1996) ‘An efficient code in V1?’, *Nature,* 381(6583), pp 560-561.

Bar-Ness, Y., Carlin, J. and Steinberger, M. (1982) *Bootstrapping adaptive interference cancelers: some practical limitations,* In Globecom’82-Global Telecommunications Conference, Miami, FL, USA: vol 3 pp 1251-1255.

Barlow, H. B. (1961) 'Possible principles underlying the transformation of sensory messages' in Rosenblith, W. A., ed. *Sensory communication*, vol 1, MIT Press, pp 217-234.

Barlow, H. B. (1972) ‘Single units and sensation: a neuron doctrine for perceptual psychology?’, *Perception,* 1(4), pp 371-394.

Barlow, H. B. (1987) 'Cerebral cortex as model builder' in Vania, L. M., ed. *Matters of intelligence,* Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science), vol 188, Springer, Dordrecht, pp 395-406.

Barlow, H. B. (1989) ‘Unsupervised learning’, *Neural computation,* 1(3), pp 295-311.

Barlow, H. **B.** (1991) 'Vision tells you more than “what is where”' in Gorea, A., Fregnac, Y., Kapoula, Z. and Findlay, J., eds., *Representations of Vision: Trends and Tacit Assumptions in Vision Research*, Cambridge University Press, pp 319-330.

Barlow, H. **B.** (1994) 'What is the computational goal of the neocortex' in Koch, C. and Davis, J. L., eds., *Large-scale neuronal theories of the brain,* Computational Neuroscience Series, MIT Press pp 1-22.

Barlow, H. B., Fitzhugh, R. and Kuffler, S. (1957) ‘Change of organization in the receptive fields of the cat's retina during dark adaptation’, *The Journal of physiology,* 137(3), pp 338-354.

Barlow, H. B., Kaushal, T. P. and Mitchison, G. J. (1989) ‘Finding minimum entropy codes’, *Neural computation,* 1(3), pp 412-423.

Bartram, D. J. (1974) ‘The role of visual and semantic codes in object naming’, *Cognitive Psychology,* 6(3), pp 325-356.

Baum, E. B., Moody, J. and Wilczek, F. (1988) ‘Internal representations for associative memory’, *Biological cybernetics,* 59(4-5), pp 217-228.

Baylis, G. C., Rolls, E. T. and Leonard, C. (1985) ‘Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey’, *Brain research,* 342(1), pp 91-102.

Bell, A. J. and Sejnowski, T. J. (1995) ‘An information-maximization approach to blind separation and blind deconvolution’, *Neural computation,* 7(6), pp 1129-1159.

Bell, A. J. and Sejnowski, T. J. (1997) ‘The “independent components” of natural scenes are edge filters’, *Vision research,* 37(23), pp 3327-3338.

Belouchrani, A., Abed-Meraim, K., Cardoso, J. and Moulines, E. (1993) *Second-order blind separation of temporally correlated sources,* In Proceeding of the International Conference on Digital Signal Processing: Citeseer, pp 346-351.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. and Plemmons, R. J. (2007) 'Algorithms and applications for approximate nonnegative matrix factorization', *Computational statistics & data analysis,* 52(1), pp 155-173.

Bialek, W., Rieke, F., Van Steveninck, R. D. R. and Warland, D. (1991) 'Reading a neural code', *Science,* 252(5014), pp 1854-1857.

Bialek, W. and Zee, A. (1990) 'Coding and computation with neural spike trains', *Journal of Statistical Physics,* 59(1-2), pp 103-115.

Biederman, I. (1985) 'Human image understanding: Recent research and a theory', *Computer vision, graphics, and image processing,* 32(1), pp 29-73.

Biederman, I. (1987) 'Recognition-by-components: a theory of human image understanding', *Psychological review,* 94(2), pp 115-147.

Biederman, I. and Cooper, E. E. (1992) 'Size invariance in visual object priming', *Journal of Experimental Psychology: Human Perception and Performance,* 18(1), pp 121-133.

Biederman, I. and Gerhardstein, P. C. (1993) 'Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance', *Journal of Experimental Psychology: Human Perception and Performance,* 19(6), pp 1162-1182.

Binford, T. O. (1971) *Visual Perception by Computer,* In Proceeding of IEEE Conference on Systems and Control, 1971, Miami, FL, USA.

Binford, T. O. (1981) 'Inferring surfaces from images', *Artificial Intelligence,* 17(1-3), pp 205-244.

Blackmore, J. T. (1972) *Ernst Mach; his work, life, and influence,* Univ of California Press.

Block, H. D., Knight Jr, B. and Rosenblatt, F. (1962) 'Analysis of a four-layer series-coupled perceptron. II', *Reviews of Modern Physics,* 34(1), pp 135.

Bobin, J., Moudden, Y., Starck, J.-L., Fadili, J. and Aghanim, N. (2008) 'SZ and CMB reconstruction using generalized morphological component analysis', *Statistical Methodology,* 5(4), pp 307-317.

Bobin, J., Starck, J.-L., Fadili, J. M., Moudden, Y. and Donoho, D. L. (2007) 'Morphological component analysis: An adaptive thresholding strategy', *IEEE Transactions on Image processing,* 16(11), pp 2675-2681.

Bofill, P. (2008) *Identifying single source data for mixing matrix estimation in instantaneous blind source separation,* In International Conference on Artificial Neural Networks: Springer, pp 759-767.

Bofill, P. and Zibulevsky, M. (2001) 'Underdetermined blind source separation using sparse representations', *Signal processing,* 81(11), pp 2353-2362.

Bolles, R. C. and Cain, R. A. (1982) 'Recognizing and locating partially visible objects: The local-feature-focus method', *The international journal of robotics research,* 1(3), pp 57-82.

Booth, M. C. and Rolls, E. T. (1998) 'View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex', *Cerebral cortex (New York, NY: 1991),* 8(6), pp 510-523.

Bottou, L. (1998) 'Online learning and stochastic approximations' in Saad, D., ed. *On-line learning in neural networks*, vol 17, Cambridge University Press, pp 142-175.

Brady, M., Ponce, J., Yuille, A. and Asada, H. (1985) 'Describing surfaces', *Computer vision, graphics, and image processing,* 32(1), pp 1-28.

Brecht, M. and Sakmann, B. (2002) 'Dynamic representation of whisker deflection by synaptic potentials in spiny stellate and pyramidal cells in the barrels and septa of layer 4 rat somatosensory cortex', *The Journal of physiology,* 543(1), pp 49-70.

Brincat, S. L. and Connor, C. E. (2006) 'Dynamic shape synthesis in posterior inferotemporal cortex', *Neuron,* 49(1), pp 17-24.

Bro, R. and De Jong, S. (1997) 'A fast non-negativity-constrained least squares algorithm', *Journal of Chemometrics: A Journal of the Chemometrics Society,* 11(5), pp 393-401.

Bronkhorst, A. W. (2000) 'The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions', *Acta Acustica united with Acustica,* 86(1), pp 117-128.

Brooks, R. A. (1981) 'Symbolic reasoning among 3-D models and 2-D images', *Artificial Intelligence,* 17(1-3), pp 285-348.

Broomhead, D. S. and Lowe, D. (1988) *Radial basis functions, multi-variable functional interpolation and adaptive networks*, (RSRE-MEMO-4148), Royal Signals and Radar Establishment Malvern (United Kingdom).

Brunswik, E. and Kamiya, J. (1953) 'Ecological cue-validity of'proximity'and of other Gestalt factors', *The American journal of psychology,* 66(1), pp 20-32.

Buchsbaum, G. and Bloch, O. (2002) 'Color categories revealed by non-negative matrix factorization of Munsell color spectra', *Vision research,* 42(5), pp 559-563.

Bülthoff, H. H. and Edelman, S. (1992) 'Psychophysical support for a two-dimensional view interpolation theory of object recognition', *Proceedings of the National Academy of Sciences,* 89(1), pp 60-64.

Burns, J. B., Weiss, R. S. and Riseman, E. M. (1992) 'The non-existence of general-case view-invariants' in Mundy, J. L. and Zisserman, A., eds., *Geometric invariance in computer vision,* Artificial Intelligence Series, vol 1, MIT Press, pp 554-559.

Calhoun, V. D., Adali, T., Hansen, L. K., Larsen, J. and Pekar, J. J. (2003) *ICA of functional MRI data: An overview,* In Amari, S.-i., Cichocki, A., Makino, S. and Murata, N.

eds., Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Signal Separation-ICA'03, Nara, Japan: pp 909-914.

Candes, E. J. and Romberg, J. (2005) 'l1-magic: Recovery of sparse signals via convex programming', *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf,* 4, pp 14.

Candes, E. J. and Romberg, J. (2007) 'Sparsity and incoherence in compressive sampling', *Inverse problems,* 23(3), pp 969.

Candes, E. J., Romberg, J. and Tao, T. (2006) 'Stable signal recovery from incomplete and inaccurate measurements', *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences,* 59(8), pp 1207-1223.

Candès, E. J., Romberg, J. and Tao, T. (2006) 'Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information', *IEEE transactions on Information theory,* 52(2), pp 489-509.

Candes, E. J. and Tao, T. (2005) 'Decoding by linear programming', *IEEE transactions on Information theory,* 51(12), pp 4203-4215.

Candes, E. J. and Tao, T. (2006) 'Near-optimal signal recovery from random projections: Universal encoding strategies?', *IEEE transactions on Information theory,* 52(12), pp 5406-5425.

Candes, E. J., Wakin, M. B. and Boyd, S. P. (2008) 'Enhancing sparsity by reweighted ℓ1 minimization', *Journal of Fourier Analysis and Applications,* 14, pp 877-905.

Castella, M. (2008) 'Inversion of polynomial systems and separation of nonlinear mixtures of finite-alphabet sources', *IEEE Transactions on signal processing,* 56(8), pp 3905-3917.

Chang, L. and Tsao, D. Y. (2017) 'The code for facial identity in the primate brain', *Cell,* 169(6), pp 1013-1028.

Charkani, N. and Deville, Y. (1999a) 'Self-adaptive separation of convolutively mixed signals with a recursive structure. Part I: Stability analysis and optimization of asymptotic behaviour', *Signal processing,* 73(3), pp 225-254.

Charkani, N. and Deville, Y. (1999b) 'Self-adaptive separation of convolutively mixed signals with a recursive structure. Part II: Theoretical extensions and application to synthetic and real signals', *Signal processing,* 75(2), pp 117-140.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001) 'Atomic decomposition by basis pursuit', *SIAM review,* 43(1), pp 129-159.

Cherry, E. C. (1953) 'Some experiments on the recognition of speech, with one and with two ears', *The Journal of the Acoustical Society of America,* 25(5), pp 975-979.

Chien, C. and Aggarwal, J. (1987) *Shape recognition from single silhouette,* In Proceedings of International Conference on Computer Vision, London, UK: Computer Society Press of IEEE, pp 481-490.

Choi, S. and Cichocki, A. (1997) *Adaptive blind separation of speech signals: Cocktail party problem,* In Proccedings of the International Conference on Speech Processing, ICSP'97: pp 617-622.

Cichocki, A. and Phan, A. H. (2009) 'Fast local algorithms for large scale nonnegative matrix and tensor factorizations', *IEICE transactions on fundamentals of electronics, communications and computer sciences,* 92(3), pp 708-721.

Cichocki, A., Phan, A. H. and Caiafa, C. (2008) *Flexible HALS algorithms for sparse non-negative matrix/tensor factorization,* In 2008 !EEE Workshop on Machine Learning for Signal Processing: IEEE, pp 73-78.

Cichocki, A., Zdunek, R., Phan, A. H. and Amari, S.-i. (2009) *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation,* John Wiley & Sons.

Clark, A. (2013) 'Whatever next? Predictive brains, situated agents, and the future of cognitive science', *Behavioral and brain sciences,* 36(3), pp 181-204.

Clemens, D. T. and Jacobs, D. W. (1991) 'Space and time bounds on indexing 3d models from 2d images', *IEEE Transactions on Pattern Analysis & Machine Intelligence,* 13(10), pp 1007-1017.

Clowes, M. B. (1967) 'Perception, picture processing and computers' in Collins, N. L. and Michie, D., eds., *Machine intelligence*, vol 1, Oliver and Boyd, London, pp 181-197.

Comon, P. (1990) 'Analyse en composantes indépendantes et identification aveugle', *Traitement du signal,* 7(5), pp 435-450.

Comon, P. (1994) 'Independent component analysis, a new concept?', *Signal processing,* 36(3), pp 287-314.

Comon, P. and Jutten, C. (2010) *Handbook of Blind Source Separation: Independent component analysis and applications,* Academic press.

Cooper, E. E., Biederman, I. and Hummel, J. E. (1992) 'Metric invariance in object recognition: a review and further evidence', *Canadian Journal of Psychology/Revue canadienne de psychologie,* 46(2), pp 191.

Corballis, M. C. (1988) 'Recognition of Disoriented Shapes', *Psychological review,* 95(1), pp 115-123.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine learning,* 20(3), pp 273-297.

Cotter, S. F., Rao, B. D., Engan, K. and Kreutz-Delgado, K. (2005) 'Sparse solutions to linear inverse problems with multiple measurement vectors', *IEEE Transactions on signal processing,* 53(7), pp 2477-2488.

Cover, T. M. and Thomas, J. A. (1991) 'Information Theory and Statistics' in Shilling, D. L., ed. *Elements of Information Theory,* Chapter 12, 1 ed., John Wiley and Sons, Inc., pp 279-335.

Cover, T. M. and Thomas, J. A. (2006) 'Data Compression' in *Elements of Information Theory,* Chapter 5, 2 ed., Jhon Wiley and Sons, Inc, pp 103-158.

Craik, K. J. W. (1943) *The Nature of Explanation,* vol 445, Cambridge University Press Archive.

Curry, H. B. (1944) 'The method of steepest descent for non-linear minimization problems', *Quarterly of Applied Mathematics,* 2(3), pp 258-261.

Damasio, A. R. and Damasio, H. (1993) 'Cortical systems underlying knowledge retrieval: Evidence from human lesion studies' in Poggio, T. A. and Glaser, D. A., eds., *Exploring Brain Functions: Models in Neuroscience*, John Wiley and Sons, pp 233-233.

Damasio, A. R., Tranel, D. and Damasio, H. (1990) 'Face agnosia and the neural substrates of memory', *Annual review of neuroscience,* 13(1), pp 89-109.

Dan, Y., Atick, J. J. and Reid, R. C. (1996) 'Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory', *Journal of Neuroscience,* 16(10), pp 3351-3362.

Dane, C. A. (1981) 'Three-dimensional segmentation using the Gaussian image and spatial information', in *Conference on Pattern Recognition and Image Processing*, Dallas, TX, USA, IEEE, pp 54-56

Dane, C. A. and Bajcsy, R. (1982) *An Object-Centered Three-Dimensional Model Builder,* In Proceedings of the 6th International Conference on Pattern Recognition, Munich, Germany: IEEE Computer Society, pp 348-350.

Daubechies, I., Defrise, M. and De Mol, C. (2004) 'An iterative thresholding algorithm for linear inverse problems with a sparsity constraint', *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences,* 57(11), pp 1413-1457.

De Valois, R. L., Morgan, H. and Snodderly, D. M. (1974) 'Psychophysical studies of monkey vision-III. Spatial luminance contrast sensitivity tests of macaque and human observers', *Vision research,* 14(1), pp 75-81.

Debener, S., Makeig, S., Delorme, A. and Engel, A. K. (2005) 'What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis', *Cognitive Brain Research,* 22(3), pp 309-321.

Devarajan, K. (2008) 'Nonnegative matrix factorization: an analytical and interpretive tool in computational biology', *PLoS Comput Biol,* 4(7), pp e1000029.

DeWeese, M. R., Wehr, M. and Zador, A. M. (2003) 'Binary spiking in auditory cortex', *Journal of Neuroscience,* 23(21), pp 7940-7949.

Dong, D. W. and Atick, J. J. (1995) 'Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus', *Network: Computation in Neural Systems,* 6(2), pp 159-178.

Donoho, D. L. (2006a) 'Compressed sensing', *IEEE transactions on Information theory,* 52(4), pp 1289-1306.

Donoho, D. L. (2006b) 'For most large underdetermined systems of linear equations the minimal ', *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences,* 59(6), pp 797-829.

Donoho, D. L. and Elad, M. (2003) 'Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ1 minimization', *Proceedings of the National Academy of Sciences,* 100(5), pp 2197-2202.

Douglas, S. C., Sawada, H. and Makino, S. (2004) 'Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters', *IEEE Transactions on Speech and Audio Processing,* 13(1), pp 92-104.

Duarte, M. F. and Eldar, Y. C. (2011) 'Structured compressed sensing: From theory to applications', *IEEE Transactions on signal processing,* 59(9), pp 4053-4085.

Edelman, S. and Bülthoff, H. H. (1992) 'Orientation dependence in the recognition of familiar and novel views of three-dimensional objects', *Vision research,* 32(12), pp 2385-2400.

Edelman, S. and Poggio, T. (1989) 'Integrating visual cues for object segmentation and recognition', *Optics News,* 15(5), pp 8.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) 'Least angle regression', *The Annals of statistics,* 32(2), pp 407-499.

Ehlers, F. and Schuster, H. G. (1997) 'Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment', *IEEE Transactions on signal processing,* 45(10), pp 2608-2612.

Elad, M. (2010) *Sparse and Redundant Representations: from Theory to Applications in Signal and Image Processing,* Springer Science & Business Media.

Elad, M. and Aharon, M. (2006) 'Image denoising via sparse and redundant representations over learned dictionaries', *IEEE Transactions on Image processing,* 15(12), pp 3736-3745.

Eldar, Y. C. and Kutyniok, G. (2012) *Compressed Sensing: Theory and Applications,* Cambridge university press.

Enroth-Cugell, C. and Robson, J. G. (1966) 'The contrast sensitivity of retinal ganglion cells of the cat', *The Journal of physiology,* 187(3), pp 517-552.

Fantana, A. L., Soucy, E. R. and Meister, M. (2008) 'Rat olfactory bulb mitral cells receive sparse glomerular inputs', *Neuron,* 59(5), pp 802-814.

Farah, M. J., Rochlin, R. and Klein, K. L. (1994) 'Orientation invariance and geometric primitives in shape recognition', *Cognitive Science,* 18(2), pp 325-344.

Farina, D., Merletti, R. and Enoka, R. M. (2004) 'The extraction of neural strategies from the surface EMG', *Journal of applied physiology,* 96(4), pp 1486-1495.

Faugeras, O. D. (1984) *New steps toward a flexible 3-D vision system for robotics,* In Proceedings of IEEE Seventh International Conference on Pattern Recognition, Montreal, Canada: vol 2 pp 796-805.

Faugeras, O. D. and Hebert, M. (1986) 'The representation, recognition, and locating of 3-D objects', *The international journal of robotics research,* 5(3), pp 27-52.

Field, D. J. (1987) 'Relations between the statistics of natural images and the response properties of cortical cells', *JOSA A,* 4(12), pp 2379-2394.

Field, D. J. (1993) *Scale-invariance and self-similar wavelet'transforms: an analysis of natural scenes and mammalian visual systems,* In Farge, M., Hunt, J. C. R. and Vassilicos, J. C. eds., Wavelets, fractals, and Fourier transforms: Clarendon Press, pp 151-193.

Field, D. J. (1994) 'What is the goal of sensory coding?', *Neural computation,* 6(4), pp 559-601.

Figueiredo, M. A. and Nowak, R. D. (2003) 'An EM algorithm for wavelet-based image restoration', *IEEE Transactions on Image processing,* 12(8), pp 906-916.

Fischler, M. A. and Bolles, R. C. (1981) 'Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography', *Communications of the ACM,* 24(6), pp 381-395.

Flexer, A., Bauer, H., Pripfl, J. and Dorffner, G. (2005) 'Using ICA for removal of ocular artifacts in EEG recorded from blind subjects', *Neural Networks,* 18(7), pp 998-1005.

Földiak, P. (1990) 'Forming sparse representations by local anti-Hebbian learning', *Biological cybernetics,* 64(2), pp 165-170.

Földiak, P. and Young, M. P. (1995) 'Sparse coding in the primate cortex' in Arbib, M. A., ed. *The Handbook of Brain Theory and Neural Networks,* A Bradford Book, 1 ed., MIT Press, pp 895-898.

Freiwald, W. A., Tsao, D. Y. and Livingstone, M. S. (2009) 'A face feature space in the macaque temporal lobe', *Nature neuroscience,* 12(9), pp 1187-1196.

Friston, K. (2005) 'A theory of cortical responses', *Philosophical transactions of the Royal Society B: Biological sciences,* 360(1456), pp 815-836.

Fujita, I., Tanaka, K., Ito, M. and Cheng, K. (1992) 'Columns for visual features of objects in monkey inferotemporal cortex', *Nature,* 360(6402), pp 343-346.

Fukushima, K. (1975) ‘Cognitron: A self-organizing multilayered neural network’, *Biological cybernetics,* 20(3-4), pp 121-136.

Fukushima, K. (2003) 'Restoring Partly Occluded Patterns: A Neural Network Model with Backward Paths', in Kaynak, O., Alpaydin, E., Oja, E. and Xu, L., eds., *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003*, ICANN 2003, ICONIP 2003, Lecture Notes in Computer Science, vol 2714: Springer, Berlin, Heidelberg, pp 393-400

Fukushima, K. (2005) ‘Restoring partly occluded patterns: a neural network model’, *Neural Networks,* 18(1), pp 33-43.

Fukushima, K. and Miyake, S. (1982) 'Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition', in Amari, S.-i. and Arbib, M. A., eds., *Competition and cooperation in neural nets*, Lecture Notes in Biomathematics, vol 45: Springer, Berlin, Heidelberg, pp 267-285

Gallippi, C. M. and Trahey, G. E. (2002) ‘Adaptive clutter filtering via blind source separation for two-dimensional ultrasonic blood velocity measurement’, *Ultrasonic imaging,* 24(4), pp 193-214.

Ganguli, S. and Sompolinsky, H. (2012) ‘Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis’, *Annual review of neuroscience,* 35, pp 485-508.

Gauthier, I. and Tarr, M. J. (1997) ‘Becoming a “Greeble” expert: Exploring mechanisms for face recognition’, *Vision research,* 37(12), pp 1673-1682.

Gibson, J. J. (1950a) *The perception of the visual world,* Boston, MA, USA: Houghton Mifflin and Company.

Gibson, J. J. (1950b) ‘The perception of visual surfaces’, *The American journal of psychology,* 63(3), pp 367-384.

Gibson, J. J. (1966) *The senses considered as perceptual systems,* Boston, MA, USA: Houghton Mifflin and Company.

Gibson, J. J. (1979) *The ecological approach to visual perception,* Boston, MA, USA: Houghton, Mifflin and Company.

Giebel, H. (1971) 'Feature extraction and recognition of handwritten characters by homogeneous layers', in Grusser, O. J. and R, K., eds., *Zeichenerkennung durch biologische und technische systeme/Pattern recognition in biological and technical systems*, Springer, Berlin, Heidelberg, pp 162-169

Girosi, F. and Poggio, T. (1990) ‘Networks and the best approximation property’, *Biological cybernetics,* 63(3), pp 169-176.

Girshick, R. (2015) *Fast r-cnn,* In Proceedings of the IEEE international conference on computer vision (ICVV): pp 1440-1448.

Golgi, C. (1906) ‘The neuron doctrine: theory and facts’, *Nobel lecture,* 1921, pp 189-217.

Gonzalez, R. C. and Wintz, P. (1977) *Digital Image Processing, Applied mathematics and computations,* vol 13, Addison Wesley Publication Company, Reading, Massachusetts.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning, Adaptive computation and machine learning,* MIT press.

Gorokhov, A. and Loubaton, P. (1997) 'Subspace-based techniques for blind separation of convolutive mixtures with temporally correlated sources', *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications,* 44(9), pp 813-820.

Grellier, O. and Comon, P. (1998) 'Blind separation of discrete sources', *IEEE Signal Processing Letters,* 5(8), pp 212-214.

Gribonval, R. (2002) 'Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture', in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, IEEE, pp III-3057-III-3060 3.

Gribonval, R. and Nielsen, M. (2003) 'Sparse representations in unions of bases', *IEEE transactions on Information theory,* 49(12), pp 3320-3325.

Grimsdale, R., Sumner, F., Tunis, C. and Kilburn, T. (1959) 'A system for the automatic recognition of patterns', *Proceedings of the IEE-Part B: Radio and Electronic Engineering,* 106(26), pp 210-221.

Grimson, W. E. L. (1990) *The combinatorics of heuristic search termination for object recognition in cluttered environments,* In Faugeras, O. D. eds., European Conference on Computer Vision-ECCV 90, ECCV 1990: Springer, Berlin, Heidelberg, pp 552-556.

Grimson, W. E. L. and Lozano-Perez, T. (1987) 'Localizing overlapping parts by searching the interpretation tree', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PAMI-9(4), pp 469-482.

Gross, C. G. (1985) 'Inferior temporal cortex and pattern recognition', *Experimental Brain Research Supplement,* 11, pp 179-201.

Gross, C. G. (1992) 'Representation of visual stimuli in inferior temporal cortex', *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 335(1273), pp 3-10.

Gross, C. G., Bender, D. B. and Rocha-Miranda, C. E. (1969) 'Visual receptive fields of neurons in inferotemporal cortex of the monkey', *Science,* 166(3910), pp 1303-1306.

Ham, F. M., Faour, N. A. and Wheeler, J. C. (1999) *Infrasound signal separation using independent component analysis,* In Warren, N. J. eds., 21st Seismic Research Symposium, SRS'99: Technologies for Monitoring The Comprehensive Nuclear-Test-Ban Treaty, Las Vegas, NV, USA: pp II-133-II-140.

Hartley, R. V. L. (1928) 'Transmission of Information ', *Bell System Technical Journal,* 7(3), pp 535-563.

Hartline, H. K. and Ratliff, F. (1972) 'Inhibitory interaction in the retina of Limulus' in Fuortes, M. G. F., ed. *Physiology of Photoreceptor Organs*, Handbook of Sensory Physiologu, vol 7/2: Springer, Berlin, Heidelberg, pp 381-447.

Hasselmo, M. E., Rolls, E. T., Baylis, G. and Nalwa, V. (1989) 'Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey', *Experimental brain research,* 75(2), pp 417-429.

Hebb, D. O. (1949) *The Organization of Behavior,* New York: Wiley.

Hegdé, J. and Van Essen, D. C. (2000) 'Selectivity for complex shapes in primate visual area V2', *Journal of Neuroscience,* 20(5), pp RC61-1-6.

Hegdé, J. and Van Essen, D. C. (2003) 'Strategies of shape representation in macaque visual area V2', *Visual neuroscience,* 20(3), pp 313-328.

Hegdé, J. and Van Essen, D. C. (2007) 'A comparative study of shape representation in macaque visual areas V2 and V4', *Cerebral cortex,* 17(5), pp 1100-1116.

Henry, R. C. (1997) 'History and fundamentals of multivariate air quality receptor models', *Chemometrics and intelligent laboratory systems,* 1(37), pp 37-42.

Henry, R. C. (2002) 'Multivariate receptor models—current practice and future trends', *Chemometrics and intelligent laboratory systems,* 60(1-2), pp 43-48.

Herault, J. and Ans, B. (1984) 'Circuits neuronaux synapses modifiables: dcodage de messages composites par apprentissage non supervis', *Comptes rendus des séances de l'Académie des sciences. Série 3, Sciences de la vie.1984,* 299(13), pp 525-528.

Herault, J. and Jutten, C. (1986) *Space or time adaptive signal processing by neural network models,* In AIP conference proceedings: American Institute of Physics, vol 151 pp 206-211.

Hérault, J., Jutten, C. and Ans, B. (1985) *Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé,* In 10 Colloque sur le traitement du signal et des images, FRA, 1985, France: GRETSI, Groupe d'Etudes du Traitement du Signal et des Images.

Hoffman, D. D. and Richards, W. A. (1984) 'Parts of recognition', *Cognition,* 18(1-3), pp 65-96.

Hopfield, J. J. (1982) 'Neural networks and physical systems with emergent collective computational abilities', *Proceedings of the National Academy of Sciences,* 79(8), pp 2554-2558.

Hopfield, J. J. (1984) 'Neurons with graded response have collective computational properties like those of two-state neurons', *Proceedings of the National Academy of Sciences,* 81(10), pp 3088-3092.

Hoyer, P. O. (2002) *Non-negative sparse coding,* In Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing, Martigny, Switzerland: IEEE, pp 557-565.

Hoyer, P. O. (2004) ‘Non-negative matrix factorization with sparseness constraints’, *Journal of machine learning research,* 5(Nov), pp 1457-1469.

Hoyer, P. O. and Hyvärinen, A. (2002) ‘A multi-layer sparse coding network learns contour coding from natural images’, *Vision research,* 42(12), pp 1593-1605.

Hu, S., Stead, M. and Worrell, G. A. (2007) ‘Automatic identification and removal of scalp reference signal for intracranial EEGs based on independent component analysis’, *IEEE Transactions on Biomedical Engineering,* 54(9), pp 1560-1572.

Hubel, D. H. and Wiesel, T. N. (1962) ‘Receptive fields, binocular interaction and functional architecture in the cat's visual cortex’, *The Journal of physiology,* 160(1), pp 106-154.

Hubel, D. H. and Wiesel, T. N. (1968) ‘Receptive fields and functional architecture of monkey striate cortex’, *The Journal of physiology,* 195(1), pp 215-243.

Hummel, J. E. and Biederman, I. (1992) ‘Dynamic binding in a neural network for shape recognition’, *Psychological review,* 99(3), pp 480-517.

Humphrey, G. K. and Khan, S. C. (1992) ‘Recognizing novel views of three-dimensional objects’, *Canadian Journal of Psychology/Revue canadienne de psychologie,* 46(2), pp 170-190.

Huttenlocher, D. P. and Ullman, S. (1987) *Object recognition using alignment,* In Proceedings of the international conference on computer vision, 1987, London, UK: Computer Society Press of the IEEE, pp 102-111.

Huttenlocher, D. P. and Ullman, S. (1990) ‘Recognizing solid objects by alignment with an image’, *International journal of computer vision,* 5(2), pp 195-212.

Hyvärinen, A. (1998) *New approximations of differential entropy for independent component analysis and projection pursuit,* In Advances in neural information processing systems: MIT Press, vol 10 pp 273-279.

Hyvärinen, A. and Hoyer, P. (2000) ‘Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces’, *Neural computation,* 12(7), pp 1705-1720.

Hyvarinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis, Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control, vol 26,* New York, 20: John Wiley & Sons, Inc.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001) 'Introduction' in *Independent Component Analysis,* Chapter 1 Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control, vol 26, John Wiley & Sons, Inc, pp 11-14.

Hyvärinen, A. and Oja, E. (2000) ‘Independent component analysis: algorithms and applications’, *Neural Networks,* 13(4-5), pp 411-430.

Iriarte, J., Urrestarazu, E., Valencia, M., Alegre, M., Malanda, A., Viteri, C. and Artieda, J. (2003) 'Independent component analysis as a tool to eliminate artifacts in EEG: a quantitative study', *Journal of clinical neurophysiology,* 20(4), pp 249-257.

Ito, M. and Komatsu, H. (2004) 'Representation of angles embedded within contour stimuli in area V2 of macaque monkeys', *Journal of Neuroscience,* 24(13), pp 3313-3324.

Ito, M., Takeuchi, Y., Matsumoto, T., Kudo, H., Kawamoto, M., Mukai, T. and Ohnishi, N. (2002) *Moving-source separation using directional microphones,* In Proceedings of the 2nd IEEE International Symposium on Signal Processing and Information Technology, ISSPIT'02, Marrakech, Morroco: pp 523-526.

Jacobson, H. (1950) 'The information capacity of the human ear', *Science,* 112(2901), pp 143-144.

Jacobson, H. (1951) 'Information and the human ear', *The Journal of the Acoustical Society of America,* 23(4), pp 463-471.

Jallon, P., Chevreuil, A., Loubaton, P. and Chevalier, P. (2004) *Separation of convolutive mixtures of cyclostationary sources: a contrast function based approach,* In Puntonet, C. G. and Prieto, A. eds., International Conference on Independent Component Analysis and Signal Separation-ICA 2004: Springer, Berlin, Heidelberg, pp 508-515.

James, C. J. and Gibson, O. J. (2003) 'Temporally constrained ICA: an application to artifact rejection in electromagnetic brain signal analysis', *IEEE Transactions on Biomedical Engineering,* 50(9), pp 1108-1116.

Johnson, J. S. and Olshausen, B. A. (2005) 'The recognition of partially visible natural objects in the presence and absence of their occluders', *Vision research,* 45(25-26), pp 3262-3276.

Jolicoeur, P. (1985) 'The time to name disoriented natural objects', *Memory & cognition,* 13(4), pp 289-303.

Jolicoeur, P. (1990) 'Orientation congruency effects on the identification of disoriented shapes', *Journal of Experimental Psychology: Human Perception and Performance,* 16(2), pp 351.

Jourjine, A., Rickard, S. and Yilmaz, O. (2000) *Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures,* In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), Istanbul, Turkey: IEEE, vol 5 pp 2985-2988.

Joyce, C. A., Gorodnitsky, I. F. and Kutas, M. (2004) 'Automatic removal of eye movement and blink artifacts from EEG data using blind component separation', *Psychophysiology,* 41(2), pp 313-325.

Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E. and Sejnowski, T. J. (2000) 'Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects', *Clinical Neurophysiology,* 111(10), pp 1745-1758.

Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E. and Sejnowski, T. J. (2001) 'Analysis and visualization of single-trial event-related potentials', *Human brain mapping,* 14(3), pp 166-185.

Jutten, C. and Herault, J. (1991) 'Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture', *Signal processing,* 24(1), pp 1-10.

Kabrisky, M. (1966) *A proposed model for visual information processing in the human brain*, PhD thesis, Electical Engineering, University of Illinois at Urbana-Champaign.

Kanerva, P. (1988) *Sparse distributed memory, Bradford Books,* MIT press.

Kelly, D. (1972) 'Adaptation effects on spatio-temporal sine-wave thresholds', *Vision research,* 12(1), pp 89-101.

Kiwiel, K. C. (2001) 'Convergence and efficiency of subgradient methods for quasiconvex minimization', *Mathematical programming,* 90(1), pp 1-25.

Koch, C. and Poggio, T. (1999) 'Predicting the visual world: silence is golden', *Nature neuroscience,* 2(1), pp 9-10.

Koenderink, J. J. and Van Doorn, A. J. (1979) 'The internal representation of solid shape with respect to vision', *Biological cybernetics,* 32(4), pp 211-216.

Kohonen, T. (2012) *Content-Addressable Memories, Springer Series in Information Sciences,* vol 1, Springer Science & Business Media.

Kullback, S. and Leibler, R. A. (1951) 'On information and sufficiency', *The annals of mathematical statistics,* 22(1), pp 79-86.

Lamdan, Y., Schwatrtz, J. and Wolfson, H. J. (1988) *On recognition of 3-D objects from 2-D images,* In Proceedings. 1988 IEEE International Conference on Robotics and Automation, Philadelphia, PA, USA: IEEE, vol 3 pp 1407-1413.

Laughlin, S. (1981) 'A simple coding procedure enhances a neuron's information capacity', *Zeitschrift für Naturforschung C,* 36(9-10), pp 910-912.

Lee, D. D. and Seung, H. S. (1999) 'Learning the parts of objects by non-negative matrix factorization', *Nature,* 401(6755), pp 788-791.

Lee, D. D. and Seung, H. S. (2000) 'Algorithms for non-negative matrix factorization', in *Advances in neural information processing systems (Proceedings of NIPS 2000)*, MIT Press, pp 556-562 13.

Lee, J. S., Lee, D. D., Choi, S. and Lee, D. S. (2001) *Application of nonnegative matrix factorization to dynamic positron emission tomography,* In Lee, T. W., Jung, T. P., Makeig, S. and Sejnowski, T. J. eds., Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation, San Diego, CA, USA: Springer, Berlin, Heidelberg, pp 629-632.

Lee, T.-W., Lewicki, M. S., Girolami, M. and Sejnowski, T. J. (1999) 'Blind source separation of more sources than mixtures using overcomplete representations', *IEEE Signal Processing Letters,* 6(4), pp 87-90.

Lee, T. S. and Mumford, D. (2003) 'Hierarchical Bayesian inference in the visual cortex', *JOSA A,* 20(7), pp 1434-1448.

Leggett, D. J. (1977) 'Numerical analysis of multicomponent spectra', *Analytical Chemistry,* 49(2), pp 276-281.

Lennie, P. (2003) 'The cost of cortical computation', *Current Biology,* 13(6), pp 493-497.

Leviatan, D. and Temlyakov, V. N. (2006) 'Simultaneous approximation by greedy algorithms', *Advances in Computational Mathematics,* 25(1-3), pp 73-90.

Levy, W. B. and Baxter, R. A. (1996) 'Energy efficient neural codes', *Neural computation,* 8(3), pp 531-543.

Lewicki, M. S. (2002) 'Efficient coding of natural sounds', *Nature neuroscience,* 5(4), pp 356-363.

Lewicki, M. S. and Olshausen, B. A. (1999) 'Probabilistic framework for the adaptation and comparison of image codes', *JOSA A,* 16(7), pp 1587-1601.

Lewicki, M. S. and Sejnowski, T. J. (2000) 'Learning overcomplete representations', *Neural computation,* 12(2), pp 337-365.

Lin, J. K., Grier, D. G. and Cowan, J. D. (1997) 'Faithful representation of separable distributions', *Neural computation,* 9(6), pp 1305-1320.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) *Feature pyramid networks for object detection,* In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, Hawaii, USA: pp 2117-2125.

Linnainmaa, S., Harwood, D. and Davis, L. S. (1988) 'Pose determination of a three-dimensional object using triangle pairs', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 10(5), pp 634-647.

Linsker, R. (1987) 'Towards an organizing principle for perception: Hebbian synapses and the principal of Optimal neural coding', in *Technical Report-RC12830, IBM Research Division*, Yorktown Heights NY, IBM TJ Watson Research Center,

Linsker, R. (1989a) 'An application of the principle of maximum information preservation to linear systems', in Touretxky, D. S., ed. *Advances in neural information processing systems*, Morgan Kaufmann Publishers Inc., pp 186-194 1.

Linsker, R. (1989b) 'How to generate ordered maps by maximizing the mutual information between input and output signals', *Neural computation,* 1(3), pp 402-411.

Linsker, R. (1990) 'Perceptual neural organization: some approaches based on network models and information theory', *Annual review of neuroscience,* 13(1), pp 257-281.

Liu, J., Harris, A. and Kanwisher, N. (2010) 'Perception of face parts and face configurations: an fMRI study', *Journal of cognitive neuroscience,* 22(1), pp 203-211.

Liu, Z., Knill, D. C. and Kersten, D. (1995) 'Object classification for human and ideal observers', *Vision research,* 35(4), pp 549-568.

Logothetis, N. K. and Pauls, J. (1995) 'Psychophysical and physiological evidence for viewer-centered object representations in the primate', *Cerebral cortex,* 5(3), pp 270-288.

Logothetis, N. K., Pauls, J., Bülthoff, H. and Poggio, T. (1994) 'View-dependent object recognition by monkeys', *Current Biology,* 4(5), pp 401-414.

Logothetis, N. K., Pauls, J. and Poggio, T. (1995) 'Shape representation in the inferior temporal cortex of monkeys', *Current Biology,* 5(5), pp 552-563.

Logothetis, N. K. and Sheinberg, D. L. (1996) 'Visual object recognition', *Annual review of neuroscience,* 19(1), pp 577-621.

Lowe, D. G. (1985) *Perceptual organization and visual recognition,* Kluwer Academic Publishers, MA, USA.

Lu, X.-C. M. and Slotnick, B. M. (1998) 'Olfaction in rats with extensive lesions of the olfactory bulbs: implications for odor coding', *Neuroscience,* 84(3), pp 849-866.

Ma, L., Qiu, Q., Gradwohl, S., Scott, A., Elden, Q. Y., Alexander, R., Wiegraebe, W. and Yu, C. R. (2012) 'Distributed representation of chemical features and tunotopic organization of glomeruli in the mouse olfactory bulb', *Proceedings of the National Academy of Sciences,* 109(14), pp 5481-5486.

Mach, E. (1868) 'On the physiological effect of spatially distributed light stimuli' in *Translated in: Ratliff, F, Mach Bands: Quantitative Studies on Neural Networks in the Retina (1965),* Holden-Day series in psychology., San Francisco, CA, USA: Holden-Day, pp 299-306.

Mach, E. (1910) *History and Root of the Principle of the Conservation of Energy,* Open Court Publishing Company.

MacKay, D. J. (2003) *Information theory, inference and learning algorithms,* Cambridge university press.

Makeig, S., Bell, A. J., Jung, T.-P. and Sejnowski, T. J. (1996) 'Independent component analysis of electroencephalographic data', in Mozer, M. C., Jordan, M. I. and Petsche, T., eds., *Advances in neural information processing systems*, Denver, CO, USA, MIT Press, pp 145-151

Makeig, S., Jung, T.-P., Bell, A. J., Ghahremani, D. and Sejnowski, T. J. (1997) *Blind separation of auditory event-related brain responses into independent components,* In Proceedings of the National Academy of Sciences: vol 94 pp 10979-10984.

Mallat, S. G. and Zhang, Z. (1993) ‘Matching pursuits with time-frequency dictionaries’, *IEEE Transactions on signal processing,* 41(12), pp 3397-3415.

Marr, D. (1969) ‘"A theory of cerebellar cortex"’, *J. Physiol.,* 202(2), pp 437–70.

Marr, D. (1970) ‘A theory for cerebral neocortex’, *Proceedings of the Royal society of London. Series B. Biological sciences,* 176(1043), pp 161-234.

Marr, D. (1971) ‘ "Simple memory: a theory for archicortex". ’, *Philos. Trans. R. Soc. Lond. B Biol. Sci.,* 262 (841), pp 23-81.

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information,* Henry Holt and Company.

Marr, D. and Nishihara, H. K. (1978) ‘Representation and recognition of the spatial organization of three-dimensional shapes’, *Proceedings of the Royal society of London. Series B. Biological sciences,* 200(1140), pp 269-294.

Matsuoka, K., Ohya, M. and Kawamoto, M. (1995) ‘A neural net for blind separation of nonstationary signals. Neural Networks, 8 (3): 411-419’.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J. and Sejnowski, T. J. (1998) ‘Analysis of fMRI data by blind separation into independent spatial components’, *Human brain mapping,* 6(3), pp 160-188.

McKeown, M. J. and Sejnowski, T. J. (1998) ‘Independent component analysis of fMRI data: examining the assumptions’, *Human brain mapping,* 6(5-6), pp 368-372.

McMullen, P. A. and Jolicoeur, P. (1990) ‘The spatial frame of reference in object naming and discrimination of left-right reflections’, *Memory & cognition,* 18(1), pp 99-115.

Milner, P. M. (1974) ‘A model for visual shape recognition’, *Psychological review,* 81(6), pp 521.

Minsky, M. and Papert, S. (1969) ‘Perceptrons: An essay in computational geometry’, *MIT Press.*

Mitianoudis, N. and Stathaki, T. (2007) 'Underdetermined source separation using mixtures of warped Laplacians', in Davis, M. E., James, C. J., Abdallah, S. A. and Plumbley, M. D., eds., *International Conference on Independent Component Analysis and Signal Separation*, ICA 2007, Lectures in Computer Science, vol 4666: Springer, Berlin, Heidelberg, pp 236-243

Miyashita, Y. and Chang, H. S. (1988) ‘Neuronal correlate of pictorial short-term memory in the primate temporal cortexYasushi Miyashita’, *Nature,* 331(6151), pp 68-70.

Molgedey, L. and Schuster, H. G. (1994) ‘Separation of a mixture of independent signals using time delayed correlations’, *Physical review letters,* 72(23), pp 3634.

Mombaerts, P. (2006) ‘Axonal wiring in the mouse olfactory system’, *Annu. Rev. Cell Dev. Biol.,* 22, pp 713-737.

Mombaerts, P., Wang, F., Dulac, C., Chao, S. K., Nemes, A., Mendelsohn, M., Edmondson, J. and Axel, R. (1996) 'Visualizing an olfactory sensory map', *Cell,* 87(4), pp 675-686.

Moses, Y., Adini, Y. and Ullman, S. (1994) *Face recognition: The problem of compensating for changes in illumination direction,* In Eklundh, J. O. eds., Computer vision-ECCV'94, ECCV 1994 Lectures Notes in Computer Science, vol 800: Springer, Berlin, Heidelberg, pp 286-296.

Moses, Y. and Ullman, S. (1992) 'Limitations of non model-based schemes', in Sandini, G., ed. *Computer Vision-ECCV'92*, ECCV 1992, Lecture Notes in Computer Science, vol 588: Springer, Berlin, Heidelberg, pp 820-828

Nadal, J.-P. and Parga, N. (1994) 'Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer', *Network: Computation in neural systems,* 5(4), pp 565-581.

Novak, M. and Mammone, R. (2001) *Use of non-negative matrix factorization for language model adaptation in a lecture transcription task,* In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA: IEEE, vol 1 pp 541-544.

O'Keefe, J. and Dostrovsky, J. (1971) 'The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat', *Brain research,* 34, pp 171-175.

Olshausen, B. A. (2013) *Highly overcomplete sparse coding,* In Proceedings of SPIE 8651: Human Vision and Electronic Imaging XVIII, 86510S.

Olshausen, B. A., Anderson, C. H. and Van Essen, D. C. (1993) 'A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information', *Journal of Neuroscience,* 13(11), pp 4700-4719.

Olshausen, B. A. and Field, D. J. (1996) 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images', *Nature,* 381(6583), pp 607-609.

Olshausen, B. A. and Field, D. J. (1997) 'Sparse coding with an overcomplete basis set: A strategy employed by V1?', *Vision research,* 37(23), pp 3311-3325.

Onton, J., Delorme, A. and Makeig, S. (2005) 'Frontal midline EEG dynamics during working memory', *Neuroimage,* 27(2), pp 341-356.

Oram, M. W. and Perrett, D. I. (1994) 'Modeling visual recognition from neurobiological constraints', *Neural Networks,* 7(6-7), pp 945-972.

Ossadtchi, A., Baillet, S., Mosher, J., Thyerlei, D., Sutherling, W. and Leahy, R. (2004) 'Automated interictal spike detection and source localization in magnetoencephalography using independent components analysis and spatio-temporal clustering', *Clinical Neurophysiology,* 115(3), pp 508-522.

Paatero, P. and Tapper, U. (1994) 'Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values', *Environmetrics,* 5(2), pp 111-126.

Paik, J. K. and Katsaggelos, A. K. (1992) 'Image restoration using a modified Hopfield network', *IEEE Transactions on Image processing,* 1(1), pp 49-63.

Palm, G. (1980) 'On associative memory', *Biological cybernetics,* 36(1), pp 19-31.

Palm, G. and Sommer, F. T. (1996) 'Associative data storage and retrieval in neural networks' in *Models of neural networks III*, Springer, pp 79-118.

Palmer, S., Rosch, E. and Chase, P. (1981) *Canonical perspective and the perception of objects,* In Long, J. and Baddeley, A. eds., Attention & performance IX, Cambridge, UK: Lawrence, Erlbaum, pp 135-151.

Pauca, V. P., Shahnaz, F., Berry, M. W. and Plemmons, R. J. (2004) *Text mining using non-negative matrix factorizations,* In Proceedings of the 2004 SIAM International Conference on Data Mining, Orlando, FL, USA: SIAM, pp 452-456.

Perrett, D., Smith, P., Mistlin, A., Chitty, A., Head, A., Potter, D., Broennimann, R., Milner, A. and Jeeves, M. A. (1985) 'Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: a preliminary report', *Behavioural brain research,* 16(2-3), pp 153-170.

Perrett, D. I. and Oram, M. W. (1993) 'Neurophysiology of shape processing', *Image and Vision Computing,* 11(6), pp 317-333.

Perrett, D. I., Rolls, E. T. and Caan, W. (1982) 'Visual neurones responsive to faces in the monkey temporal cortex', *Experimental brain research,* 47(3), pp 329-342.

Perrett, D. I., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, A. D. and Jeeves, M. (1985) 'Visual cells in the temporal cortex sensitive to face view and gaze direction', *Proceedings of the Royal society of London. Series B. Biological sciences,* 223(1232), pp 293-317.

Pham, D.-T. and Cardoso, J.-F. (2001) 'Blind separation of instantaneous mixtures of nonstationary sources', *IEEE Transactions on signal processing,* 49(9), pp 1837-1848.

Pham, D.-T. and Vrins, F. (2006) *Discriminacy of the minimum range approach to blind separation of bounded sources,* In Verleysen, M. eds., 14th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning-ESANN'06, Bruges, Belgium Advances in Computational Intelligence and Learning: pp 377-382.

Phillips, P. A., Rolls, B. J., Ledingham, J. G., Forsling, M. L., Morton, J. J., Crowe, M. J. and Wollner, L. (1984) 'Reduced thirst after water deprivation in healthy elderly men', *New England Journal of Medicine,* 311(12), pp 753-759.

Pietilä, A., El-Segaier, M., Vigário, R. and Pesonen, E. (2006) *Blind source separation of cardiac murmurs from heart recordings,* In Rosca, J., Erdogmus, D., Principe, J. C. and Haykin, S. eds., Independent Component Analysis and Signal Separation, ICA

2006 Lecture Notes in Computer Science, vol 3889: Springer, Berlin, Heidelberg, pp 470-477.

Pitts, W. and McCulloch, W. S. (1947) 'How we know universals the perception of auditory and visual forms', *The Bulletin of mathematical biophysics,* 9(3), pp 127-147.

Poggio, T. (1990) *A theory of how the brain might work,* In Cold Spring Harbor symposia on quantitative biology: Cold Spring Harbor Laboratory Press, vol 55 pp 899-910.

Poggio, T. and Edelman, S. (1990) 'A network that learns to recognize three-dimensional objects', *Nature,* 343(6255), pp 263.

Poggio, T. and Girosi, F. (1989) *A theory of networks for approximation and learning*, A.I. Memo No 1140, Artificial Intelligence Laboratory: Massachusetts Institute of Technology.

Poggio, T. and Girosi, F. (1990a) *Networks for approximation and learning,* In Proceedings of the IEEE: vol 78(9) pp 1481-1497.

Poggio, T. and Girosi, F. (1990b) 'Regularization algorithms for learning that are equivalent to multilayer networks', *Science,* 247(4945), pp 978-982.

Poggio, T., Torre, V. and Koch, C. (1987) 'Computational vision and regularization theory' in Fischler, M. A. and Firschein, O., eds., *Readings in computer vision*, Morgan Kauffman, pp 638-643.

Poo, C. and Isaacson, J. S. (2009) 'Odor representations in olfactory cortex:"sparse" coding, global inhibition, and oscillations', *Neuron,* 62(6), pp 850-861.

Porée, F., Kachenoura, A., Gauvrit, H., Morvan, C., Carrault, G. and Senhadji, L. (2006) 'Blind source separation for ambulatory sleep recording', *IEEE Transactions on Information Technology in Biomedicine,* 10(2), pp 293-301.

Potmesil, M. (1983) *Generating models of solid objects by matching 3D surface segments,* In Proceedings of the eighth International Joint Conference on Artifiical Intelligence-IJCAI'83 Karlsruhe, Germany: Morgan Kaufmann Publishers Inc., vol 2 pp 1089-1093.

Powel, M. J. D. (1987) 'Radial basis function for multivariable interpolations: a review' in Mason, J. C. and Cox, M. G., eds., *Algorithms for Approximation*, Clarendon Press, Oxford, pp 143-167.

Pulkkinen, J., Häkkinen, A.-M., Lundbom, N., Paetau, A., Kauppinen, R. and Hiltunen, Y. (2005) 'Independent component analysis to proton spectroscopic imaging data of human brain tumours', *European journal of radiology,* 56(2), pp 160-164.

Puntonet, C., Mansour, A. and Jutten, C. (1995) *A geometrical algorithm for blind separation of sources,* In Actes du XV`eme Colloque GRETSI 95, Juan-Les-Pins, France: pp 273-276.

Quastler, H. (1956) *Studies of human channel capacity,* In Cherry, E. C. eds., Information Theory: papers read at a symposium on information theory held at the Royal

Institution, London, September 12th to 16th, 1955., London, UK: Academic Press Inc., pp 361-371.

Ramon y Cajal, S. (1906) 'Nobel lecture—The structure and connexions of neurons',

Rao, R. P. and Ballard, D. H. (1999) 'Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects', *Nature neuroscience,* 2(1), pp 79-87.

Rapin, J., Bobin, J., Larue, A. and Starck, J.-L. (2013) 'Sparse and non-negative BSS for noisy data', *IEEE Transactions on signal processing,* 61(22), pp 5620-5632.

Rapoport, A. and Horvath, W. J. (1960) 'The theoretical channel capacity of a single neuron as determined by various coding systems', *Information and control,* 3(4), pp 335-350.

Reader, S. M. and Laland, K. N. (2002) 'Social intelligence, innovation, and enhanced brain size in primates', *Proceedings of the National Academy of Sciences,* 99(7), pp 4436-4441.

Ressler, K. J., Sullivan, S. L. and Buck, L. B. (1993) 'A zonal organization of odorant receptor gene expression in the olfactory epithelium', *Cell,* 73(3), pp 597-609.

Rieke, F., Warland, D. and Bialek, W. (1993) 'Coding efficiency and information rates in sensory neurons', *EPL (Europhysics Letters),* 22(2), pp 151.

Riesenhuber, M. and Poggio, T. (1998) *Modeling invariances in inferotemporal cell tuning*, A.I. Memo 1629, CBCL Paper 160, MIT Artificial Intelligence Lab and CBCL: MIT, USA.

Riesenhuber, M. and Poggio, T. (1999a) 'Are cortical models really bound by the "binding problem"?', *Neuron,* 24(1), pp 87-93.

Riesenhuber, M. and Poggio, T. (1999b) 'Hierarchical models of object recognition in cortex', *Nature neuroscience,* 2(11), pp 1019-1025.

Riesenhuber, M. and Poggio, T. (2000) 'Models of object recognition', *Nature neuroscience,* 3(11), pp 1199-1204.

Riesenhuber, M. and Poggio, T. (2002) 'Neural mechanisms of object recognition', *Current opinion in neurobiology,* 12(2), pp 162-168.

Rivet, B., Girin, L. and Jutten, C. (2005) *Solving the indeterminations of blind source separation of convolutive speech mixtures,* In Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Philadelphia, PA: IEEE, vol 5 pp v/533-v/536.

Robbins, H. and Monro, S. (1951) 'A stochastic approximation method', *The annals of mathematical statistics,* 22(3), pp 400-407.

Rock, I. and DiVita, J. (1987) 'A case of viewer-centered object perception', *Cognitive Psychology,* 19(2), pp 280-293.

Roll, J.-P. (1981) *Contribution à la Proprioception Musculaire, à la Perception et au Contrôle du Mouvement Chez l'Homme*, Thèse de doctorat d'état (PhD) thesis, Science, University of Aix-Marseille I.

Rolls, E. T. (1984) 'Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces', *Human neurobiology,* 3(4), pp 209-222.

Rolls, E. T. and Milward, T. (2000) 'A model of invariant object recognition in the visual system: learning rules, activation functions, lateral inhibition, and information-based performance measures', *Neural computation,* 12(11), pp 2547-2572.

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(2), pp 713-726.

Rosenblatt, F. (1957) *The perceptron, a perceiving and recognizing automaton (Project Para)*, Report No 85-460-1, Cornell Aeronautical Laboratory Inc.

Rosenblatt, F. (1958) 'The perceptron: a probabilistic model for information storage and organization in the brain', *Psychological review,* 65(6), pp 386.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G. and Olshausen, B. A. (2008) 'Sparse coding via thresholding and local competition in neural circuits', *Neural computation,* 20(10), pp 2526-2563.

Ruderman, D. L. and Bialek, W. (1994) 'Statistics of natural images: Scaling in the woods', *Physical review letters,* 73(6), pp 814-817.

Sajda, P., Du, S., Brown, T. R., Stoyanova, R., Shungu, D. C., Mao, X. and Parra, L. C. (2004) 'Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain', *IEEE transactions on medical imaging,* 23(12), pp 1453-1465.

Schiller, P. H. (1995) 'Effect of lesions in visual cortical area V4 on the recognition of transformed objects', *Nature,* 376(6538), pp 342-344.

Schiller, P. H. and Lee, K. (1991) 'The role of the primate extrastriate area V4 in vision', *Science,* 251(4998), pp 1251-1253.

Schmidt, M. N. and Mørup, M. (2006) 'Nonnegative matrix factor 2-D deconvolution for blind single channel source separation', in Rosca, J., Erdogmus, D., Principe, J. C. and Haykin, S., eds., *Independent Component Analysis and Blind Signal Separation*, ICA 2006, Lecture Notes in Computer Science, vol 3889: Springer, Berlin, Heidelberg, pp 700-707

Schwartz, E. L., Desimone, R., Albright, T. D. and Gross, C. G. (1983) 'Shape recognition and inferior temporal neurons', *Proceedings of the National Academy of Sciences,* 80(18), pp 5776-5778.

Schwartz, O. and Simoncelli, E. P. (2001) 'Natural signal statistics and sensory gain control', *Nature neuroscience,* 4(8), pp 819-825.

Schwenker, F., Sommer, F. T. and Palm, G. (1996) 'Iterative retrieval of sparsely coded associative memory patterns', *Neural Networks,* 9(3), pp 445-455.

Selfridge, O. G. (1959) *Pandemonium: A paradigm for learning,* In The mechanism of thought processes (Proceedings of a symposium, National Physical Laboratory, Teddington, England), London, UK: Her Majesty's Stationery Office.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y. (2013) 'Overfeat: Integrated recognition, localization and detection using convolutional networks', *arXiv preprint, arXiv:1312.6229[cs.CV]*.

Shannon, C. E. (1948) 'A mathematical theory of communication', *Bell System Technical Journal,* 27(3), pp 379-423.

Shimojo, S., Silverman, G. H. and Nakayama, K. (1989) 'Occlusion and the solution to the aperture problem for motion', *Vision research,* 29(5), pp 619-626.

Shoham, D. and Ullman, S. (1988) *Aligning a model to an image using minimal information,* In Proceedings of the 2nd International Conference on Computer Vision-ICCV'88, Tampa, FL, USA: pp 259-263.

Simoncelli, E. P. and Olshausen, B. A. (2001) 'Natural image statistics and neural representation', *Annual review of neuroscience,* 24(1), pp 1193-1216.

Smaragdis, P. (2004) *Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,* In Puntonet, C. G. and Prieto, A. eds., Independent Component Analysis and Blind Signal Separation, ICA 2004 Lecture Notes in Computer Science, vol 3195: Springer, Berlin, Heidelberg, pp 494-499.

Smaragdis, P. (2006) 'Convolutive speech bases and their application to supervised speech separation', *IEEE Transactions on Audio, Speech, and Language Processing,* 15(1), pp 1-12.

Smith, E. C. and Lewicki, M. S. (2006) 'Efficient auditory coding', *Nature,* 439(7079), pp 978-982.

Sodoyer, D., Girin, L., Jutten, C. and Schwartz, J.-L. (2004) 'Developing an audio-visual speech source separation algorithm', *Speech Communication,* 44(1-4), pp 113-125.

Srinivasan, M. V., Laughlin, S. B. and Dubs, A. (1982) 'Predictive coding: a fresh view of inhibition in the retina', *Proceedings of the Royal society of London. Series B. Biological sciences,* 216(1205), pp 427-459.

Stein, R. B. (1967) 'The information capacity of nerve cells using a frequency code', *Biophysical journal,* 7(6), pp 797.

Stettler, D. D. and Axel, R. (2009) 'Representations of odor in the piriform cortex', *Neuron,* 63(6), pp 854-864.

Stevens, C. F. (2018) 'Conserved features of the primate face code', *Proceedings of the National Academy of Sciences,* 115(3), pp 584-588.

Stone, J. V. (2004) *Independent component analysis: a tutorial introduction, A Bradford Book,* MIT press.

Sutherland, N. S. (1968) 'Outlines of a theory of visual pattern recognition in animals and man', *Proceedings of the Royal society of London. Series B. Biological sciences,* 171(1024), pp 297-317.

Sutherland, N. S. (1969) *Stimulus Analyzing Mechanisms,* In Mechanization oj thought processes: National physical laboratory symposium, London, UK: Her Majesty's Stationery Office, vol 2(10) pp 575-609.

Tanaka, K. (1992) 'Inferotemporal cortex and higher visual functions', *Current opinion in neurobiology,* 2(4), pp 502-505.

Tanaka, K. (1996) 'Inferotemporal cortex and object vision', *Annual review of neuroscience,* 19(1), pp 109-139.

Tanaka, K., Saito, C., Fukada, Y. and Moriya, M. (1990) *Integration of form, texture, and color information in the inferotemporal cortex of the macaque,* In Iwai, E. and Mishkin, M. eds., Vision, memory and the temporal lobe, Tokyo, Japan: New York: Elsevier, pp 101-109.

Tanaka, K., Saito, H.-A., Fukada, Y. and Moriya, M. (1991) 'Coding visual images of objects in the inferotemporal cortex of the macaque monkey', *Journal of neurophysiology,* 66(1), pp 170-189.

Tang, A. C., Pearlmutter, B. A., Malaszenko, N. A. and Phung, D. B. (2002a) 'Independent components of magnetoencephalography: single-trial response onset times', *Neuroimage,* 17(4), pp 1773-1789.

Tang, A. C., Pearlmutter, B. A., Malaszenko, N. A., Phung, D. B. and Reeb, B. C. (2002b) 'Independent components of magnetoencephalography: localization', *Neural computation,* 14(8), pp 1827-1858.

Tarr, M. J. (1995) 'Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects', *Psychonomic Bulletin & Review,* 2(1), pp 55-82.

Tarr, M. J. and Bülthoff, H. H. (1995) 'Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993)', *Journal of Experimental Psychology: Human Perception and Performance,* 21(6), pp 1494-1505.

Tarr, M. J. and Pinker, S. (1989) 'Mental rotation and orientation-dependence in shape recognition', *Cognitive Psychology,* 21(2), pp 233-282.

Tarr, M. J. and Pinker, S. (1990) 'When does human object recognition use a viewer-centered reference frame?', *Psychological Science,* 1(4), pp 253-256.

Tarr, M. J. and Pinker, S. (1991) 'Article Commentary: Orientation-Dependent Mechanisms in Shape Recognition: Further Issues', *Psychological Science,* 2(3), pp 207-209.

Tauler, R., Casassas, E. and Izquierdo-Ridorsa, A. (1991) 'Self-modelling curve resolution in studies of spectrometric titrations of multi-equilibria systems by factor analysis', *Analytica chimica acta,* 248(2), pp 447-458.

Theis, F. J., Jung, A., Puntonet, C. G. and Lang, E. W. (2003a) 'Linear geometric ICA: Fundamentals and algorithms', *Neural computation,* 15(2), pp 419-439.

Theis, F. J., Puntonet, C. G. and Lang, E. W. (2003b) *Nonlinear geometric ICA,* In Amari, S.-i., Cichocki, A., Makino, S. and Murata, N. eds., Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation-ICA'03, Nara, Japan: pp 275-280.

Thomas, J., Deville, Y. and Hosseini, S. (2006) 'Time-domain fast fixed-point algorithms for convolutive ICA', *IEEE Signal Processing Letters,* 13(4), pp 228-231.

Thompson, D. and Mundy, J. (1987) *Three-dimensional model matching from an unconstrained viewpoint,* In Proceedings. 1987 IEEE International Conference on Robotics and Automation, Raleigh, NC, USA: IEEE, vol 4 pp 208-220.

Tikhonov, A. N. and Arsenin, V. Y. (1977) *Solutions of ill-posed problems (Translated by Jhon, F.), Halsted Press book, Scripta series in mathematics,* Winston.

Tolhurst, D. and Thompson, I. (1982) 'Organization of neurones preferring similar spatial frequencies in cat striate cortex', *Experimental brain research,* 48(2), pp 217-227.

Tong, L., Soon, V. C., Huang, Y. F. and Liu, R. (1990) *AMUSE: a new blind identification algorithm,* In IEEE International Symposium on Circuits and Systems, New Orleans, LA, USA: IEEE, vol 3 pp 1784-1787.

Tou, J. T. and Gonzalez, R. C. (1974) *Pattern recognition principles,* Reading, MA, USA: Addison-Wesley.

Tovee, M. J., Rolls, E. T. and Azzopardi, P. (1994) 'Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque', *Journal of neurophysiology,* 72(3), pp 1049-1060.

Treloar, H. B., Feinstein, P., Mombaerts, P. and Greer, C. A. (2002) 'Specificity of glomerular targeting by olfactory sensory axons', *Journal of Neuroscience,* 22(7), pp 2469-2477.

Tsodyks, M. and Feigelman, M. (1988) 'Enhanced storage capacity in neural networks with low level of activity', *Europhysics Letters (EPL),* 6(2), pp 101-105.

Tsuge, S., Shishibori, M., Kuroiwa, S. and Kita, K. (2001) *Dimensionality reduction using non-negative matrix factorization for information retrieval,* In 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236), Tucson, AZ, USA: IEEE, vol 2 pp 960-965.

Turner, G. C., Bazhenov, M. and Laurent, G. (2008) 'Olfactory representations by Drosophila mushroom body neurons', *Journal of neurophysiology,* 99(2), pp 734-746.

Ullman, S. (1989) ‘Aligning pictorial descriptions: An approach to object recognition’, *Cognition,* 32(3), pp 193-254.

Ullman, S. (1996) *High-level vision: Object recognition and visual cognition, A Bradford Book,* vol 2, MIT press Cambridge, MA.

Ullman, S. (1998) ‘Three-dimensional object recognition based on the combination of views’, *Cognition,* 67(1-2), pp 21-44.

Ullman, S. and Basri, R. (1991) ‘Recognition by linear combination of models’, *IEEE Transactions on Pattern Analysis & Machine Intelligence,* 13(10), pp 992-1006.

Van Hateren, J. H. (1992) ‘A theory of maximizing sensory information’, *Biological cybernetics,* 68(1), pp 23-29.

Van Hateren, J. H. and van der Schaaf, A. (1998) ‘Independent component filters of natural images compared with simple cells in primary visual cortex’, *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 265(1394), pp 359-366.

Van Hulle, M. M. (1999) *Clustering approach to square and non-square blind source separation,* In Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No. 98TH8468), Madison, WI, USA: IEEE, pp 315-323.

Van Nes, F. L., Koenderink, J. J., Nas, H. and Bouman, M. A. (1967) ‘Spatiotemporal modulation transfer in the human eye’, *Journal of the Optical Society of America (JOSA),* 57(9), pp 1082-1088.

Vassar, R., Ngai, J. and Axel, R. (1993) ‘Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium’, *Cell,* 74(2), pp 309-318.

Vetter, T., Poggio, T. and Bülthoff, H. (1994) ‘The importance of symmetry and virtual views in three-dimensional object recognition’, *Current Biology,* 4(1), pp 18-23.

Vigário, R., Jousmäki, V., Hämäläinen, M., Hari, R. and Oja, E. (1997) *Independent component analysis for identification of artifacts in magnetoencephalographic recordings,* In Jordan, M. I., Kearns, M. J. and Solla, S. A. eds., Advances in Neural Information Processing Systems (Proceedings of NIPS’97), Denver, CO, USA: MIT Press, vol 10 pp 229-235.

Vigário, R., Sarela, J., Jousmiki, V., Hamalainen, M. and Oja, E. (2000) ‘Independent component approach to the analysis of EEG and MEG recordings’, *IEEE Transactions on Biomedical Engineering,* 47(5), pp 589-593.

Vigário, R. N. (1997) ‘Extraction of ocular artefacts from EEG using independent component analysis’, *Electroencephalography and clinical neurophysiology,* 103(3), pp 395-404.

Vincent, E. (2007) 'Complex nonconvex l p norm minimization for underdetermined source separation', in Abdallah, S. A. and Plumbley, M. D., eds., *Independent Component Analysis and Signal Separation*, ICA 2007, Lecture Notes in Computer Science, vol 4666: Springer, Berlin, Heidelberg, pp 430-437

Virtanen, T. (2004) *Separation of sound sources by convolutive sparse coding,* In ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing(SAPA-2004), paper 55, ICC Jeju, Korea.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R. and Sinha, P. (2018) 'Potential downside of high initial visual acuity', *Proceedings of the National Academy of Sciences,* 115(44), pp 11333-11338.

Von Helmholtz, H. (1867) *Handbuch der physiologischen Optik, Allgemeine Encyklopädie der Physik,* vol 9, Voss.

Voss, R. F. (1985) 'Random fractal forgeries' in Earnshaw, R. A., ed. *Fundamental Algorithms for Computer Graphics,* NATO ASI Series (Series F: Computer and Systems Sciences), vol 17, Springer, Berlin, Heidelberg, pp 805-835.

Wallis, G. and Rolls, E. T. (1997) 'Invariant face and object recognition in the visual system', *Progress in neurobiology,* 51(2), pp 167-194.

Wallis, G., Rolls, E. T. and Foldiak, P. (1993) *Learning invariant responses to the natural transformations of objects,* In Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan), Nagoya, Japan: IEEE, vol 2 pp 1087-1090.

Wehr, S., Lombard, A., Buchner, H. and Kellermann, W. (2007) *'Shadow BSS'for Blind Source Separation in Rapidly Time-Varying Acoustic Scenes,* In Davies, M. E., James, C. J., Abdallah, S. A. and Plumbley, M. D. eds., Independent Component Analysis and Signal Separation, ICA 2007 Lecture Notes in Computer Science, vol 4666: Springer, Berlin, Heidelberg, pp 560-568.

Weiskrantz, L. (1990) 'Visual prototypes, memory, and the inferotemporal cortex' in Iwai, E. and Mishkin, M., eds., *Vision, memory and the temporal lobe*, Elsevier, New York, pp 13-28.

Willhite, D. C., Nguyen, K. T., Masurkar, A. V., Greer, C. A., Shepherd, G. M. and Chen, W. R. (2006) 'Viral tracing identifies distributed columnar organization in the olfactory bulb', *Proceedings of the National Academy of Sciences,* 103(33), pp 12592-12597.

Willshaw, D. J., Buneman, O. P. and Longuet-Higgins, H. C. (1969) 'Non-holographic associative memory', *Nature,* 222(5197), pp 960-962.

Winston, P. H. (1975) 'Learning structural descriptions from examples' in Winston, P. H., ed. *The Psychology of Computer Vision*, New York: McGraw-Hill, pp 157-209.

Wu, H.-C. and Principe, J. C. (1999) *Generalized anti-Hebbian learning for source separation,* In 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258), Phoenix, AZ, USA: IEEE, vol 2 pp 1073-1076.

Xinhua, Z., Anqing, Z., Jianping, F. and Shaoqing, Y. (2000) *Study on blind separation of underwater acoustic signals,* In WCC 2000-ICSP 2000. 2000 5th International

Conference on Signal Processing Proceedings. 16th World Computer Congress 2000, Beijing, China: IEEE, vol 3 pp 1802-1805.

Xu, W., Liu, X. and Gong, Y. (2003) *Document clustering based on non-negative matrix factorization,* In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval:SIGIR'03, Toronto, Canada: pp 267-273.

Yellin, D. and Weinstein, E. (1994) 'Criteria for multichannel signal separation', *IEEE Transactions on signal processing,* 42(8), pp 2158-2168.

Yellin, D. and Weinstein, E. (1996) 'Multichannel signal separation: Methods and analysis', *IEEE Transactions on signal processing,* 44(1), pp 106-118.

Yilmaz, O. and Rickard, S. (2004) 'Blind separation of speech mixtures via time-frequency masking', *IEEE Transactions on signal processing,* 52(7), pp 1830-1847.

Yost, W. A. (1997) 'The cocktail party problem: Forty years later' in Gilkey, R. and Anderson, T. R., eds., *Binaural and spatial hearing in real and virtual environments,* Chapter 17, Psychology Press, pp 329-347.

Young, M. P. and Yamane, S. (1992) 'Sparse population coding of faces in the inferotemporal cortex', *Science,* 256(5061), pp 1327-1331.

Young, S. S., Scott, P. D. and Nasrabadi, N. M. (1997) 'Object recognition using multilayer Hopfield neural network', *IEEE Transactions on Image processing,* 6(3), pp 357-372.

Yuille, A. L., Cohen, D. and Hallinan, P. (1989) *Facial feature extraction by deformable templates, Tech. Rep. 88-2,* Harvard Robotics Lab, Cambridge, MA, USA.

Zetzsche, C. (1990) *Sparse coding: the link between low level vision and associative memory,* In Eckmiller, R., Hartmann, G. and Hauske, G. eds., Parallel processing in neural systems and computers, North Holland, Amsterdam: Elsevier Science, pp 273-276.

Zhu, Y. and Yan, Z. (1997) 'Computerized tumor boundary detection using a Hopfield neural network', *IEEE transactions on medical imaging,* 16(1), pp 55-67.

Zibulevsky, M. and Pearlmutter, B. A. (2001) 'Blind source separation by sparse decomposition in a signal dictionary', *Neural computation,* 13(4), pp 863-882.