# CORTICAL MODELS OF EXPANSIVE NON-LINEAR MIXED SELECTIVITY

**Arvin Gopal Subramaniam**

Thesis Committee:
Prof. Cengiz Pehlevan, Harvard University (Primary Advisor)
Prof. Matthieu Wyart, EPFL

A thesis submitted for partial fullfillment of the degree of
*Master of Physics*
March 2020

# Abstract

Two hallmarks of cortical computation are (i) the integration of multiple sources of information in a non-linear fashion, dubbed *Non-linear Mixed Selectivity* (NLMS), and (ii) expansive feed-forward projections that raise the dimensionality of the input space. Both architectures serve a generic purpose to increase the number of independently classifiable patterns by neural circuits and reduce the readout error of a downstream circuit. This motivates a theoretical analysis of a hybrid architecture: one that increases dimensionality by expansions, but does so whilst being selective to various modalities in the input space.

A canonical model for expansive projections forming sparse representations have been introduced by Babadi and Sompolinsky, with its properties extensively studied. In this thesis, we extend the model to introduce mixed selectivity to various modalities of inputs, with an architecture that readily interpolates between *pure* and *full* mixed selectivity limits. Such a model enables us to study the function performed by a generic cortical circuit as a function of the degree to which the circuit is mixedly selective, and to ask how computational measures in the brain - such as sparseness, expansion, and representational dimensionality - affect learning.

First, we review empirical literature on NLMS. Special attention is paid to electrophysiological studies of such a hybrid architecture in early sensory processing of three distinct brain areas, and to behavioural experiments recording from the prefrontal cortices of monkeys, where the performance of the animal is related to the dimensionality of the neural representation, a quantity that is notoriously difficult to measure experimentally. These experiments will form a motivation for our model and theoretical results that we will obtain.

Then, we study the capacity of a mixedly selective perceptron, where we extend the famous Gardner result, incorporating recent results by Shinzato and Kabashima. Drawing inspiration from behavioural experiments, we call this architecture the *Context-Dependent Perceptron* (CDP). We quantify in detail the bottle-neck of contextual storage of information by the CDP, where we show that under general circumstances only one contextual information can be stored.

Next, we introduce our model, which we call the *Cortical Expansive Non-linear Mixedly Selective Model* (CEMS). We show that its capacity can be extensive with the system size, and is moreover purely determined by the expansion ratio for random projections. We provide a novel measure for the dimensionality of a neural representation as a function of order parameters that are analytically calculable, which we will later study as a function of mixing.

Further, we study the problem of generalization, where a donwstream neural circuit readout information from the cortical layer based on a supervised Hebb rule. Focusing on the covariance structure of independent, random inputs, we extend the result first introduce by Babadi, and study the readout error for CEMS when tested on noisy versions of the training stimuli. We also discuss the subtle change in noise robustness as a function of the degree of mixing.

Finally, we discuss a model of contextual decoding on the CEMS, and a threshold linear expansive model, deriving novel results for the noise amplification, the dimensionality, and the signal-to-noise ratio for these cases. In all our studies on generalization, we aim to highlight the differing roles of sparseness of the mixed representation and the expansion ratio in controlling learning, and how they subtly depend on various model-specific details.

# Acknowledgements

I would firstly like to thank the Bertarelli Foundation for this tremendous opportunity to perform a Masters thesis in theoretical neuroscience at Harvard. I think this is a particularly interesting time for theoretical physicists to pursue first-principles studies in diverse areas in biology, and the field of computational neuroscience is a shining example of such an area. As such, I was even more happy to learn of the fellowship's openness to having (what I think is) the first ever fellow from theoretical physics pursue research in this field - and I hope more will follow. Overall, it has been a pleasure to learn about work done by the fellowship in both basic and applied research, to have attended (and presented at) monthly lunches, and to have made friends and acquaintances along the way. I'm grateful to the friendships made with Kriton, Amanda, Victor, Ilaria, and Constantin. Special mention goes to Gail Townsend who expertly dealt with an unhealthy barrage of annoying questions from my end, and John Assad, for advice on graduate school applications.

A huge deal of credit goes to my supervisor, Cengiz Pehlevan. His patience to deal with my doubts and ramblings, ability to intuit interesting scientific questions, and expertise in a vast number of areas within theoretical neuroscience have been inspiring to learn from. Simply put, this thesis would not have been possible if it were not for him. I'm grateful to Shanshan Qin, Blake Boredon, Jacob Zavatone-Veth, Ricardo Alves, Harsh Sikkha, Yibo Jiang, Will Dorrell, and Qianyi Li who I have been happy to exchange ideas with and call my friends. An special mention is needed for the amazing Dina Obeid who has been a source of inspiration, mentorship, and friendship, and whose insistence that I improve my listening skills I will always remember.

I would also like to pay thanks to those at EPFL who have helped shape my academic interests. I am grateful to Matthieu Wyart, for the opportunity to join his lab and learn about how theoretical physics can contribute towards answering important questions in the world around us. Special thanks is owed to Barbara Bravi, who supervised a wonderful project in my final semester there, in a collaboration I hope to continue in the future. Such exposures I feel have greatly influenced my thinking on problems to pursue in the future, and how physics can contribute.

Finally, I would like to thank my parents and brother, for their support of me to pursue science.

# Contents

# Chapter 1

# Introduction

Imagine you are at the audience of one of your favourite lecture courses. You listen ardently as the the lecturer provides repeated, specific sequences of sentences describing hitherto novel concepts. Simultaneously, you look actively for visual cues on the blackboard, hoping to maybe reinforce the auditory information you have just received with detailed written expressions. Furthermore, you may be even tempted to transcribe what is being heard and seen, in another attempt at reinforcement, this time via the sense of touch. To make the most of this experience (and your tuition fees) your brain has to successfully integrate information across the modalities of vision, audition, and touch. How does the brain cope with this? Indeed, a hallmark of intelligent behaviour in animals is the ability to successfully integrate information from multiple sources to inform behaviour. For instance, famous perceptual experiments in humans display that subjects better discriminate stimuli when information is presented across different modalities, for instance visual coupled with tactile [7], or vestibular [28]. Studies on dolphins for instance have displayed their ability to successfully recognize objects across visual and auditory modalities [43]. Such a mixing of different sources of information is not limited to mammals; experiments on weakly electric fish have displayed their ability to successfully combine visual and electro-sensory information, as a means to guide effective electro-taxis [54]. In this Chapter, we will discuss empirical signatures of multimodal integration in the brain, particularly focusing on: (i) expansive, feed-forward multimodal circuits in early sensory processing, and (ii) non-linear mixed selectivity in mammalian prefrontal cortical areas. These discussions will raise some important theoretical questions that will help motivate our model which we will introduce and study in this thesis.

## 1.1 Multimodal integration in early sensory cortices

We provide here a quick review of the distinct circuits in early sensory systems in the brain, that have been well documented to perform the task of multimodal sensory integration. The findings reviewed here suggest a ubiquitous coding strategy of early sensory

processing, that moreover seems to appear under a common architectural motif. This review here follows along the lines of an excellent recent review by Cayo-gajic, et. al '19 [17]

### 1.1.1 Expansive cortical architectures in early sensory processing

A feature of the above circuits, and a seemingly prevalent circuit motif in early sensory processing, is that of *expansive, feed-forward* - forming representations.

We will mention here three separate circuits, across different sensory systems: the cerebellar cortex, the piriform cortex, and the dentate gyrus. Note that for olfactory circuits in insects, specifically in the mushroom body, similar architectural features are seen, and excellent reviews of this can be found in [17], [86]. Note that our presentation below will ignore some details such as recurrence and feedback inhibition ([38], [95]), and we will comment on these when discussing the limitations of the model.

#### Cerebellar Cortex

A central circuit in the brain that forms motor associations is the cerebellum [92], [82]. At the input layer, mossy fiber (MFs) cells number at about $\sim 10^3$, and project onto a layer of granule cells (GCs) ($\sim 2 \times 10^5$), via divergent parallel fibers. This stunning divergence is also partly why cerebellar GCs are one of the most abundant in the brain. At the readout layer, Purkinje cells (PC) readout the information from this high dimensional cortical layer for the purposes of forming motor associations. Such a basis for associative learning of high-dimensional, random GC activities has been subject to famous studies by Marr [58] and Albus [1], the so-called Marr-Albus theory, where an exact analogy of this readout layer to that of a perceptron was theorized.

#### Olfactory cortex and dentate gyrus

The sense of smell, on the other hand, it formed for mammals in the olfactory cortex. Here, about $\sim 10^3$ cells (glomeruli) in the input layer, the olfactory bulb (OB), project onto a high-dimensional piriform cortex (PCx) layer, which as $\sim 10^6$ neurons. Another documented apparent feature of this circuitry is the apparent random connectivity between the OB to PCx layer [8], [84], which also appears in the corresponding circuitry in the insect mushroom body [59], [42]. Finally, in the hippocampus, an important circuit for the mammalian sense of navigation, a similar feed-forward, expansive architecture is seen from the perforant path axons, originating from the enthorinal cortex [6], to the dentate GCs, with an albeit smaller expansion ratio of about $\sim 5$ [55].

A cartoon illustration of this transformation is illustrated in Figure (1.1). On the theoretical side, the utility of such an expansion, along with the subtle role of sparseness at the cortical layer [1] has been studied by Babadi, Sompolinsky, 2014 [3], and indeed our theoretical modeling here can be seen as an extension of theirs. The purpose of this thesis is more to highlight the computational role of such an architecture in the presence of multi-modal integration, where such circuits in addition mix independent sources of information.

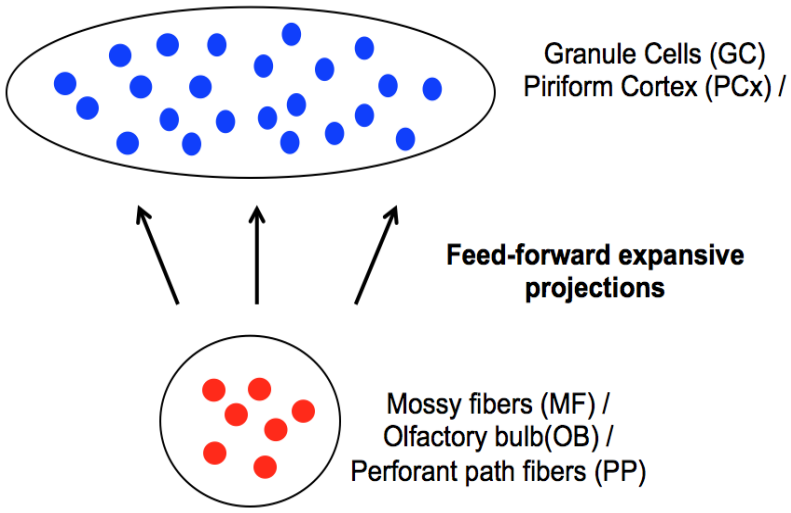


Figure 1.1: Illustration of a canonical computational architecture in early sensory processing, where low dimensional inputs (red) are transformed to high dimensional intermediate representations (blue), via feed-forward, expansive projections, for the purposes of being read out by a downstream neuron/circuit. The names written on each layer indicate those corresponding to the ones reviewed in this Chapter in the areas of (from top to bottom) cerebellar cortex, olfactory cortex, and dentate gyrus.

### 1.1.2 Multi-modal integration in expansive feed-forward circuits

#### The pattern separation and expansion recoding hypotheses

To be able to distinguish such distinct pattern, the brain must somehow implement computations that decorrelate related inputs. "Expansion recoding", central to the Marr-Albus theory of cerebellar function, simply states that low-dimensional, correlated codes can be successfully readout by downstream neurons via divergent expansive projections to form sparse representations [58], [1], [87]. There are the two central tenets of the theory – a large expansion and a sparse code in the expanded representation. The utility of such

[1] We will use the terms "cortical" and "mixed" layer inter-changeably throughout this thesis to refer to the high-dimensional, intermediate layer.

features for the problem of multi-modal integration will be the subject of our theoretical study in this thesis.

**Expansion recoding and mutli-modal integration across different sensory regions**

Hence, it is interesting to note the presence of a somewhat converging line of evidence for multi-modal integration of distinct sensory modalities on the expanded cortical layer, in aforementioned different areas of the brain, as we review below.

**Cerebellar Cortex**

In the cerebellum GCs, recent line of evidence has converged towards the presence of mixed representations in the GC layer formed by multi-modal integration. We mention just three. Firstly, the work by Huang, et. al [47] that found, using a novel mixture of genetic and viral techniques to label pathways enabling them to systematically study the input pathways to neurons across the cerebellar GC layer, that a substantial minority (40%) of neurons mix inputs across different modalities – i.e *mixed selectivity*, with the others being either pure or partially selective. The input streams they looked at in particular were that of sensory feedback from the upper body and another stream conveying motor-information. Another study by Ishikawa, et. al [48], this time using *in vivo* patch clamp recordings, displayed that a minority of GC cells (up to a fifth) were selective to a mixture of all three of auditory, visual, and somatosensory stimulation on anaesthetized rats, although interestingly some evidence for a sub-linear summation of inputs were found. It should be mentioned that electrophysiological studies in this space are not without controversy, following initial in vivo evidence [50], [83] that cerebellar GC cells were largely unimodal (purely selective).

**Olfactory Cortex**

In olfaction, we discuss a recent study from Poo, et. al [71], where neurons in the PCx were found to display mixed selectivity to both odor identity and location sampled. Their setup involved a novel spatial task, where rats received odor cues at random distinct spatial locations (ports), after which a reward was administered at a fixed relative port to the previous one. In this setting, they found again that a substantial minority (16%) of neurons in the PCx were mixedly selective to both odor identity and location, with the majority others displaying pure selectivity. In the insect mushroom body instead, among others, Yagi, et. al [93] have performed a detailed structural analysis of neural pathways, using dye injections and genetic labeling methods, showing a convergence of pathways from distinct sensory regions to the mushroom body cayx.

**Dentate Gyrus**

As noted by [17], mixed representations in the dentate gyrus are more difficult to probe, given that the enthorinal cortex beforehand integrates multiple independent sources of information [40]. However, there are evidences from behavioral ("contextual") discrimination tasks that support evidence for this. Firstly, Morris, et. al [66] used a contextual associative learning task on rats, finding that rats with lesioned dentate gyri were not able to discriminate odorants in the environmental contexts in which they were administered. Another study by Gilbert, et. al [36] found similarly that lesions in the dentate gyrus impairs the ability of rats to perform contextual discrimination.

**Expansive multi-modal integration in the weakly electric fish**

There is another distinct line of evidence for such an expansive, multi-modal integration, not in land dwelling animals discussed above, but in aquatic animals, namely electrosensory circuits in the weakly electric fish. [2] (See [80] for a comprehensive review of this.)

To effectively navigate underwater, circuits in the weakly electric fish have been developed to transform environmental cues in the form of electric signals, itself generated by its own tail upon movement, to neural responses. Such a self-generated electric pulse is known as the electric organ discharge (EOD), could be problematic, as it might be undistinguished from other behaviorally relevant signals, such as those emitted by other animals underwater (e.g prey or predators) [27].

The particular circuit that performs the above transformation that is of importance here is the electrosensory lobe (ELL). This circuit bears a striking resemblance to those of the cerebellar-like structures mentioned earlier, where low-dimensional principal cells that receive input from afferent fibres encoding the discharge signal, is expanded onto a high-dimensional GC layer, before being readout by individual Purkinje cells downstream.

To solve this problem – known as "sensory cancellation" – a body of work has suggested that the weakly electric fish solves this via a multi-modal integration at the GC layer, where electro sensory signals and corollary discharges related to motor commands are integrated [80], [10]. Interesting additional features in this circuit worth mentioning is the role of unsupervised learning [9] an regularization [23] in generating learned responses.

The above examples suggest that widely divergent, expansive feed-forward architectures, might serve another computational purpose beyond what has been previously thought - the integration of independent sources of information across different modalities.

[2] Beyond mechanistic aspects, the presence of such a circuit in the weakly electric fish may be interesting because it seems to have evolved independently of the others.

## 1.2 Non-linear Mixed Selectivity in Prefrontal Cortex

Beyond early sensory processing, the ability to successfully combine multiple sources of information in mammalian prefrontal cortical areas, and its implication in guiding complex cognitive tasks, has been subject to intense neuroscientific studies [64], [65], [12], [90], [73], and subsequent theoretical investigations [31], [5]. For instance, recordings performed on the lateral intraparietal area (LIP) of rhesus monkeys trained to perform a visual motion discrimination task have shown that individual neurons in the LIP are selective for both decision-relevant and irrelevant signals [61], whereas experiments performed on Wisconsin card sorting tests have identified individual neurons that multiplex sensory stimuli, task rules, and motor cues [2]. Of particular interest for this thesis, are paradigmatic studies investigating the ability of individual neurons in the prefrontal cortex (PFC) and related areas to successfully be *selective* to a mixture of *cues* in, various behavioural paradigms. We note from the outset that although no *structural* evidence exists that these brain regions are also feed-forward and expansive, theoretical suggestions on how such problem can be solved ([5], [56]), have used a similar architecture to that reviewed in the previous section.

Let us quickly review an important experiment in this space, specifically that conducted by Rigotti and collaborators, with a series of papers [89],[73], [56], which we will review in detail below. The particular outline of this experimental paradigm, consisting of two tasks, is as follows. In the first task, the *recognition* task, a monkey is trained to remember a sequence of two images after a delay (a.k.a a delayed response task). The monkey is trained to grasp a bar upon achieving fixation. The first object is followed by a brief delay ($\sim 1s$), and then a second object followed by another delay. This phase is then repeated as a test sequence, with either the sequence matching that of the sample sequence or being a nonmatch. There are four possible objects that can be showed, called *cues*. In the second task, known as the *Recall* task, the sample phase is identical to that in the recognition task, but now test sequence requires the monkey to make a correct saccade in the order of cues seen in the sample sequence. These two tasks are then interleaved for about 100-150 trials.

The main findings of the authors can be summarized by the following points:

- **Increasing mixing increases dimensionality.** As is well known in the machine learning community [88], a non-linear mixing leads to an increase in the number of implementable input-output relationships. To display this, the authors fit the measured recordings to linear and non-linear models respectively, where each time series of the observed traces are expressed as a linear/non-linear sum of design matrices corresponding to the tasks, and the four cues. Interestingly, and only noted in the Appendix of [73], the majority of neurons ($\approx 66\%$) recorded and fit with such

a procedure were found to be *linearly* selective to the task and cues , whereas a minority, though substantial ($\approx 34\%$), displayed non-linear mixing.

- **Dimensionality collapse predicts animal behavior.** Another important contribution of the authors is a method to measure the dimensionality of recorded neurons. This is opposed to conventional methods of determining dimensionality, such as PCA, which fails for data sets without an appropriate model of the noise [73]. Their method involved counting the number of linearly separable dichotomies from the recorded neural activities, taken at the middle of the two-object delay period, which they obtained by training a linear classifier.

  A crucial finding from the above method is that the dimensionality measured with the above technique decreases in error trials. To verify this, the neural recordings taken from trials in the recall task where fed into the classifier and the dimension is determined as mentioned above. Such a significant decrease in dimensionality was checked by the authors to not be an artifact of a lower number of sampled neurons, changes in the mean and variability of firing rates, or coding level of the activity [73].

- **Error trials are due to the increase of linear component of mixing.** Given that error trials followed a decrease in dimensionality, the natural question to ask here is what is the source of this decrease? To investigate this, the authors measured the dimensionality of the correct and error trials using the methods mentioned above, in two separate cases, with the linear and non-linear component removed respectively. Their results show that the difference in measured dimensionality between the correct and error trials is appreciable in the case where the linear component is removed, whereas there is hardly a difference in the case where the non-linear component is removed, indicating that a substantial portion of the measured dimensionality of correct trials can be attributed to the non-linear mixing of task and different cues. An interesting further observation noted by the authors is that the identities of the two cues can still be decoded during error trials, an observation made by means of training a population decoder and obtaining its accuracy. This is interesting because it *suggests that errors are not caused by a decrease in the degree of mixing*, but rather on the extent to which the mixing is non-linear versus linear.

These set of observations thus raise a series of important questions regarding the interplay between linear vs. non-linear selectivity, dimensionality of mixed representations and the readout abilities of a downstream neuron, though at a more abstract, behavioural level.

### 1.2.1 Pure and partial neuron selectivity in the PFC

In analogy to those findings mentioned in [47], [48], where a number of measured neurons displayed varying degrees of mixed selectivity, evidence also exists at the behavioural level on pure and partial seelcitivty in prefrontal cortical areas. We discuss two: those by [45] and [96]. Despite this being on a different behavioural paradigm, and measurement taken on a different anatomical brain region, such an experiment will be useful for use later on, when we discuss a generic model that interpolates between different limits of pure vs. mixed selectivity, and what their consequences entail.

#### Hirokawa, et. al ('19) experiment - single neuron selectivity in the rat orbitofrontal cortex

In this experimental paradigm, Long Evans rats are trained on a olfactory two-alternative forced choice (2AFC) discrimination task, with a built-in reward bias. Rats are trained to discriminate between two odours, by first initiating the trial with a nose-poke into a central port. After a variable delay period (between $0.2 - 0.6s$), rewards are administered on either side of the port, depending on the odour, with a bias that the rat is to infer over trials. The neural activities analyzed were those from the anticipation period, which is that between when the odours are presented and when the rewards are administered. The unique feature of this experiment is the method used by the authors to study the neural responses across different contexts, which in this case were more abstract variables such as anticipated reward size, decision confidence, and reward value. To do this, neural activities were combined across the different task contingencies of odour stimulus, reward size, behavioural choice, and previous trial outcomes to have a so-called tuning curve across all these contingencies.

**Clustering analysis reveals pure selectivity.** Due to the nature of the task paradigm, fitting the neural response to linear vs. non-linear models á la Rigotti could not be done. Instead, the authors infer the degree of selectivity from the combined population activity by clustering the neural response and subsequent statistical tests to determine deviation from random mixing. The results of the clustering revealed nine discrete clusters, and found that each cluster encodes putatively for a single decision variable. For instance, the authors would obtain a cluster corresponding to "reward size", where the activities in this cluster do not change with the intensity of odorant, or the previous outcome, with separate clusters instead showing tuning towards these variables. The authors were also able to reproduce the results across four independent cohort of rats. Full details are found in [45] and SI therein.

**Zhang, et. al ('17) experiment - partial selectivity in motor representation of tetraplegic humans with spinal cord injury**

We now move to another experiment, though this time where the representation studied were that of motor responses. Neural recordings were performed on a tetraplegic patient with spinal cord injury, and the combined representations of three variables were tested: body part, body side, and cognitive strategy. Body part referred to either the hand or the shoulder, with which the patients were able to move by either squeezing or shrugging. Body side referred to whether the patient was asked to move either the left or right for either part. Cognitive strategy referred to whether the patient was requested to either initiate a physical movement or imagine a movement. The task paradigm was a standard Go-NoGo paradigm, consisting of either performing a movement, imagining a movement, or speaking.

**Selectivity via a linear model** In contrast to the Rigotti paper mentioned in the above section, the authors here did not study the effect of linear vs. non-linear mixing, but instead study the effect of varying the degree of mixing on the tuning to different conditions. To do this, the authors fit a linear model, consisting of a linear sum of task variables, and measured the fraction of neurons tuned to a particular variable. Though we do not discuss all the details of their experiments and findings here, the main result of their experiment is that neurons in the posterior parietal cortex *cannot be well described by models of random mixing towards all task variables*, as seen for instance in other motor experiments [21]. Instead, some variables, such as body part, were deduced to be full mixed, whereas others, such as body side and strategy, did not exhibit full mixing. Thus, the authors argue that partial mixing could be an appropriate paradigm to explain integration of decision variables in the task.

These two experiments thus raise a similar set of questions as those by Rigotti, et. al, but now for neurons that display different degrees of selectivity towards independent sources of information.

## 1.3 Some other experimental evidence of NLMS

Having individual neurons being selective to multiple sources of information, broadly studied under the umbrella term NLMS, is also well documented in many other brain regions, in circuits performing functions as diverse such as navigation, whisking, locomotion, and we briefly discuss three of them below. We warn the reader upfront, however, that the following examples are by no means supposed to be an exhaustive list.

### 3D selectivity in bat head direction cells

An important recent discovery in systems neuroscience is that of 3D tuning curves in head-direction cells in the bat presubiculum [30], [29], where the cells were found to be selective towards pitch, azimuth, and roll angles. Although tuning of cells to azimuth and pitch angles have been separately reported [94], [79], this work was the first to demonstrate that head-direction cells conjunctively tuned for all three variables, to give the Egyptian fruit bat its 3D sense of direction.

### Multiplexed representations in rat medial enthorinal cortex cells

Another canonical circuit of the mammalian sense of direction is the medial enthorinal cortex, home to grid cells [68], [67]. Cells in the MEC have long been documented to encode a spatially regular code of the surrounding environment [67], whilst individual cells have been previously found to be tuned, amongst others, towards environmental boundaries (so-called "border cells"), and when the animal moves at a particular speed (so-called "speed cells"). However, a recent work by [40] used a novel statistical method of a linear-non-linear Poisson model on recordings of MEC cells of rats in a standard foraging task (analogous to non-linear fits seen in previous sections, but for spiking data), to characterize the tuning curves in the MEC towards a range of directional variables, namely position, head direction, running speed, and phase of theta rhythm. Important findings here include that the decoding error was found to be significantly smaller in cells that exhibited NLMS compared to purely selective cells.

### Dendritic non-linear mixing in layer 5 pyramidal cells

As an illustration of the ubiquity of the phenomenon across the brain, we now move to the layer 5 (L5) pyramidal neurons of the vibrissae cortex, where mixed selectivity between the mouse whisker touch and angle has been recently reported whilst they were recorded performing an active sensing task [72]. Using two photon calcium imaging, intensity of calcium ions via fluorescence indicators were measured, and they were found to modulated by both the position of whisker during a touch, with position here an angular variable with well defined positions relative to the normal of the mouse nose within the whisking cycle.

## 1.4 An expansive, multimodal circuit

The above discussion naturally raises the question of how the aformentioned brain areas might successfully integrate multiple sources of information, and what are the computational benefits of doing so in an expansive, feed-forward manner. The computational properties of such *unimodal* and *bimodal* architectures have been studied in a pair of seminal works by [3] and [5], and we will propose a model in this thesis building on those

works and extending concepts first introduced by them. We will propose the architecture presented in Figure (1.2), with the main architectural features listed below (in Chapter 3 we will give a more formal mathematical account of this):

1. **Independent sources of information that are projected independently onto the mixed layer.** In other words, we follow the setup of [5] (but not for instance [57]), in assuming that independent sources of information are to be modelled by distinct sub-populations at the input layer. This is an important modelling choice, as the resultant correlation structure of the input data will be central to many of the results we present.

2. **Random feed-forward projections of independent sources.** Again following [5], but at the risk of over-simplification, we will assume the feed-forward weights to be purely random, the simplest null model from which we hope more biologically realistic extensions can be made. The exception to this of course would be circuits in the olfactory cortex, which as mentioned above have been shown to be well approximated by such a projection.

3. **A preservation of modality-coverage as a function of mixing.** We need to have a model that has a tunable mixing degree, i.e we need to be able to ask how the computational properties of the network change as we vary from, say a purely selective network, with each neuron selective to only one input modality, to a fully selective network, where each neuron mixes all the input modalities. To smoothly interpolate between the two regimes, we will assume a *task-relevant* stimulus (e.g motor information in the cerebellum), that is integrated with *contextual* information (i.e auditory or visual stimulus). Further, we will assume that across all *mixing degrees*, the network has to at least receive the task-relevant information, and the degree to which the contextual information is integrated then depends on the mixing degree, a parameter that we can tune. We call this a *modality-preserving* feature of the model. As we will explain in Chapter 3, this will lead us to a natural partitioning scheme for which neurons on the cortical layer receive what sources of information, as a function of mixing degree. Note that this is somewhat of a departure from the aforementioned eletrophysiological studies, where one finds that some of the neurons are, say fully selective, and some others are either partially or purely selective. In our model, we assume that the *entire network* is subject to a certain mixing degree, from which we study its computational properties.

4. **A preservation of sparsity as a function of mixing.** Here, we assume that as the mixing degree varies, there is no change *per se* in the sparsity of the mixed representation. In other words, we will consider a model where the only parameter that sets the sparseness of the mixed representation is the biological neuron threshold, independent of what degree of mixing the network exhibits. There is no *a priori* reason of course why the brain should choose to do this, and as we will see such a choice will significantly control the computational properties of our network.

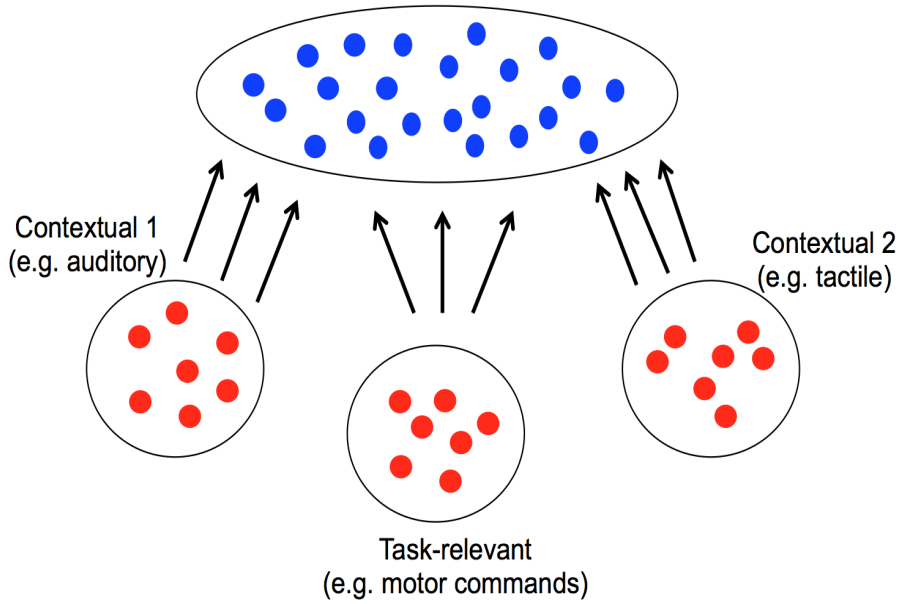


Figure 1.2: An illustration of our proposed multi-modal input-to-mixed architecture, which we call CEMS. As as extension to Figure (1.1), but now with independent sources of information in the input. Other important details of the model are mentioned in the main text and formalized mathematically in Chapter 3. Not shown here is the way the model interpolates between different degrees of mixing, also discussed in Chapter 3.

Specific questions that we will hope to address are: (i) What is the ability of an *arbitrarily* selective architecture to learn and generalize from examples?, (ii) How does noise robustness change as a function of selectivity? [3], and (iii) What is the effect of sparsity and expansion ratio - central pillars of the Marr-Albuss expansion recoding theory - on generalization?

With regards to behavioural experiments, the following questions also remain: (i) Can such an architecture aid in the *inverse* problem of empirically measuring selectivity of recorded populations of neurons, by, for instance, a quantitative relation between the degree of selectivity and the error performed by the subject, without resorting to fitting models with many parameters? (ii) Is there an *internal* model for dimensionality as a function of noise or other sources of randomness in our network, and (iii) Can counter-intuitive outcomes that have been reported - the ability to correctly decode contextual information even during error trials [73] - be recapitulated in a model and understood? In this thesis, we will outline an example of such an architecture, which we call the *cortical, expansive, non-linear mixedly selective* architecture (CEMS), that shares the hybrid features mentioned above, and readily interpolates between different degrees of selectivity. The advantage of such an architecture is that we will be able to study systematically

[3] Another set of terms that we will use inter-changeably in this thesis are "degree of mixing" and "degree of selectivity".

various computational properties such as the capacity to implement input-output associations, the dimensionality of the mixed representation, and the generalization error of a downstream readout neuron as a function of the mixing degree; though we will discuss the potentially over-simplifying model assumptions when discussing the results obtained.

## 1.5 Outline of thesis

The outline of the thesis is as follows. In Chapter 2, we discuss the *context-dependent perceptron*, and study its capacity based on exisiting results by Shinzato and Kabashima [81], which we recap. Then, in Chapter 3, we formally introduce the CEMS, and study its capacity for associative learning, and noise robustness as a function of mixing degree. In Chapter 4, after recapping important theoretical concepts for generalization in a generic model of expansive feed-forward circuits from [3], we will discuss its extension to the CEMS, and discuss the implications of the model in light of the experiments reviewed in this Chapter. Central to this will be the phenomenon of "structured multi-modal overlaps", that we will elucidate, and discuss its effect on generalization on CEMS. We will highlight in particular the role of sparseness of the mixed representation and the expansion ratio on the learning abilities of the network. Finally, in Chapter 5, we will discuss a couple of extensions, including a contextual decoding model, and (threshold)linear selectivity, again highlighting the role of expansion and representation sparseness in those cases.

# Chapter 2

# The context-dependent perceptron (CDP)

The central tenet of the Marr-Albus theory of cerebellar function is the analogy of the GC-to-Purkinje cell learning to that of a perceptron [25]. Though in the presence of multiple independent sources of information, how does this analogy carry forward?

In this section we study a toy model of multimodal integration in such a circuit, which we call the *context-dependent perceptron* (CDP), and in particular quantify its capacity to store task-relevant stimuli in the presence of contextual information. The study of cerebellar-like circuits as mentioned above is performed in the other chapters. This section essentially extends the results first given by Gardner [34] and more recently by Shinzatio and Kabashima [81], to include mixing of different modalities on the inputs, and the results can be seen as an extension to those first mentioned *en passant* by [5].

We should mention at the outset that the notion of capacity used here is common in the computational neuroscience literature. Our notion of capacity, as we will define below, is a *statistical mechanical* one, first introduced in a pioneering paper by Hopfield [46]. Famous works on capacity in the computational neuroscience community include studies of spike-timing classifiers [39], [76], [62], computational benefits of balanced inputs [75], rationalization of electrophysiology weight distribution from optimality principles in the cerebellum [22], and the cortex [13],[18], and more recently to classification abilities of data manifolds across feedforward hierarchies [20]. These are to be distinguished from other notions of capacity, more commonly encountered in cognitive neuroscience studies, such as working memory capacity [26],[19], or information theoretic measures [78].

## 2.1 Recap of Gardner and Kabashima results

Formally, let us assume a hypothetical linear classifier (a *perceptron* [74]) in the brain, which receives inputs that we will call *stimuli*, $\xi^\mu \in \mathcal{R}^N$, with $\mu = 1, .., P$. This stimuli can be taken to come from a downstream region in the sensory system, in the form of firing rates. In the statistical mechanical sense these are the *quenched randomness* in our system, and we seek to study the *typical* behaviour of the system, independent of a particular realization of the input stimuli. A natural question to ask is then what values can $\xi_i^\mu$ take; indeed most of the works cited above opt for a choice of $\xi \in \{0, 1\}$, which can be taken to correspond to if a neuron is silent or firing. It turns out that for the question of capacity the details of the stimulus space is irrelevant, what is important is that we have a *binary* random variable as a source of quenched disorder, and hence we will stick to the choice $\xi \in \{-1, +1\}$. [1] The task of this linear classifier is to find a set of weights to associate each stimuli to a *binary* label, which we call a *valence*, $\sigma^\mu \in \{-, +\}$. This could correspond to, for instance in the case of vision, classifying images of cats and dogs, or in the case of olfaction, distinguishing appetitive and aversive odorants.

quenched randomness is input date

We thus need to find a set of weights $\mathbf{J}$, such that the following $P$ inequalities are satisfied [25], [34]:

$$\frac{1}{\sqrt{N}}\sigma^\mu \mathbf{J}\boldsymbol{\xi}^\mu \geq 0; \quad \forall \mu \tag{2.1}$$

Without loss of generality, we could further define a *margin* $\kappa$ that can be placed on the right-hand side of Eq. (2.1), indicating how robust the classification is to be towards noise; for this thesis we will focus on the case of $\kappa = 0$. The inequalities above are numerically solved by a linear program whenever theories are tested against simulations. The capacity is defined as the maximal load, $\alpha = \frac{P}{N}$, with $P, N \to \infty$ but their ratio finite[2], above which the set of inequalities in (2.1) cannot be satisfied. The pioneering work by Gardner [34] first mapped this onto a statistical mechanical problem, where the following volume is evaluated:

$$\langle \Omega \rangle_{\boldsymbol{\xi},\sigma} = \langle \int d\mu(\mathbf{J}) \prod_\mu \theta(\frac{1}{\sqrt{N}}\sigma^\mu \mathbf{J}\boldsymbol{\xi}^\mu - \kappa) \rangle_{\boldsymbol{\xi},\sigma} \tag{2.2}$$

where $\theta(x)$ is the Heaviside function, and $d\mu(\mathbf{J})$ is the *measure* over which the integration is performed, corresponding to the choice of normalization of $\mathbf{J}$. The average is appropriately performed via the *replica trick* [63], to obtain the self-averaging volume $\Omega$. The famous result of Gardner, valid for purely uncorrelated stimuli, reads (for $\kappa = 0$):

$$\alpha_c = 2 \tag{2.3}$$

[1] For other computational tasks that we will discuss in later Chapters, such as to improve the readout error of a downstream neuron, this choice is also irrelevant, and merely requires an appropriate re-scaling of the (random) weights.

[2] This is also known as the "thermodynamic limit".

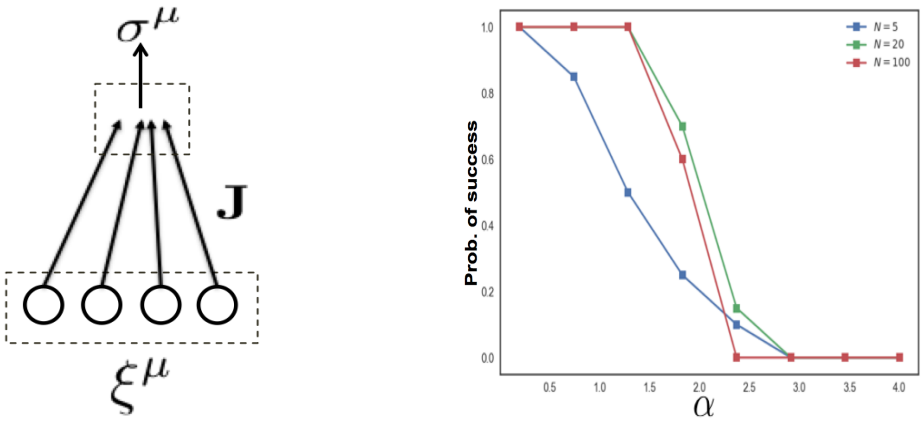


Figure 2.1: *Left:* The architecture of a perceptron - patterns $\xi^\mu$ with an associated label $\sigma^\mu$ are to be classified by a linear classifier, by finding an appropriate set of weights $\mathbf{J}$. *Right:* Recap of Gardner capacity result for different values of $N$ and $P$. The sharp $\alpha_c = 2$ capacity is well defined in the thermodynamic limit. The plot was obtained using a feasibility check on a linear program satisfying Eq. (2.1).

A relatively recent result by Shinzato and Kabashima (herefter ShK) [81] provides a neat generalization of Eq. (2.3) to correlated datasets. Assuming the data matrix $\boldsymbol{\xi}$ has a rank of $cN$, their results reads:

$$\alpha_c = 2c \tag{2.4}$$

which recovers Eq. (2.3) in the limit where the data is fully random and hence $c = 1$.

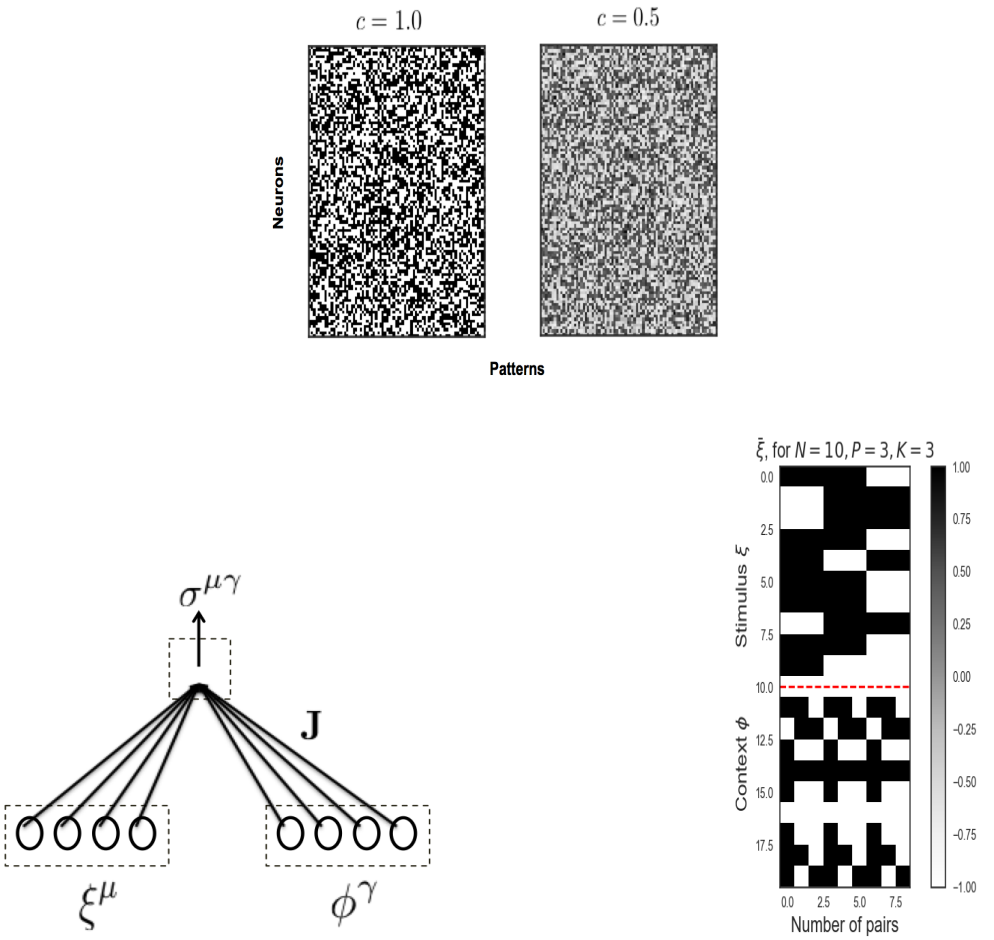


Figure 2.2: *Top*: Example of random data on the left, generated to test feasibility of network, and (on the right), the data can now be correlated in some basis, with the figure shown a low-rank reconstruction of the data set, keeping the first $cN$ largest eigenvalues (see Appendix for how this is done). *Bottom:* On the left, the architecture of the CDP is displayed. On the right, we display the stimuli-context pair used for the CDP. Colorbar for both figures is shown on the bottom right, with (+) indicated in black and (-) in white. Note that in the low-rank reconstruction on the top right, the datum are no longer required to be $\pm$.

## 2.2 The bottleneck of one context for the CDP

We now study the context-dependent perceptron (CDP) (first introduced *en passant* by [5]), and study its capacity. Though the results here are relatively straightforward, the will use the same line of argument in the next Chapter to calculate the capacity of more complicated architectures. In this and throughout the thesis, we formalize the multi-modal inputs as follows. We will always assume that our circuit receives $P$ independent stimuli $\xi^{\mu}$, and $K$ independent contexts *for all other modalities*, and moreover $P \to \infty$ (scaling

with the systems size) and $K$ is finite. We will show that the former indeed implies the latter in the case of the CDP. To do this, we will ask what constraints are imposed on the number of finite contexts to be stored, given that the CDP has to store an extensive number of task-relevant stimuli.

We can now make use of the ShK result to derive the capacity, first focusing on the case where stimuli appear with one contextual modality. In particular, we can easily show that a simple application of our thermodynamic scaling gives a bound on the number of contexts the CDP can store. Let us define the following quantities, retaining the same definition for the load as above:

$$\alpha = \frac{PK}{N+M} \tag{2.5}$$

$$\beta = \frac{P}{N} \tag{2.6}$$

$$\delta = \frac{M}{N} \tag{2.7}$$

$$\tag{2.8}$$

where all of the above are finite and each of the quantities in caps $\to \infty$ in the thermodynamic limit. Let us further define $\beta$ as the *task-relevant load*. From Eq. (2.4), we can evaluate Eq. (2.6), which gives $\alpha = \frac{\beta}{1+\delta} \times K$. Since statistical mechanical analyses require $\alpha \sim O(1)$, this implies that $K$ has to necessarily be finite, as mentioned above. We are now left with the task of evaluating the rank of the composite data matrix. Let us call this $\bar{\boldsymbol{\xi}}^{\mu\gamma} \in \mathcal{R}^{N+M\times PK}$. Explicitly, this looks like

$$(\bar{\boldsymbol{\xi}}^{\mu\gamma})^T = \begin{pmatrix} \xi^{\mu=1} & \phi^{\gamma=1} \\ .. & .. \\ .. & .. \\ \xi^{\mu=P} & \phi^{\gamma=1} \\ \xi^{\mu=1} & \phi^{\gamma=2} \\ .. & .. \\ .. & .. \\ \xi^{\mu=P} & \phi^{\gamma=2} \\ .. & .. \\ .. & .. \\ .. & .. \\ \xi^{\mu=P} & \phi^{\gamma=K} \end{pmatrix} \tag{2.9}$$

We can show that the rank of the above matrix is $P+K-1$ [5](see Appendix). From Eq. (2.4) we have

$$\alpha_c = \frac{2\beta}{1+\delta} \tag{2.10}$$

where we have ignored terms of $O(\frac{K}{N})$ and $O(\frac{1}{N})$. This leads to the a very simple inequality for when we expect perfect storage, namely:

$$\alpha < \alpha_c \implies K < 2 \tag{2.11}$$

Since we cannot have a non-integer value of $K$, our result is in fact $K = 1$. We plot the result of a simulation below, indicating a small region on the $(P, K)$ plane where perfect classification is possible, corresponding to $K = 1$. Eq. (2.10) of course assumes that $N > P$ and $M > K$, and one might wonder whether it holds true for any values of $(N, M, P, K)$. We show in the Appendix, that, remarkably, the result is the same for all choices of $(N, M, P, K)$, i.e the CDP can store *only one context* for all possible choices of architecture parameters.

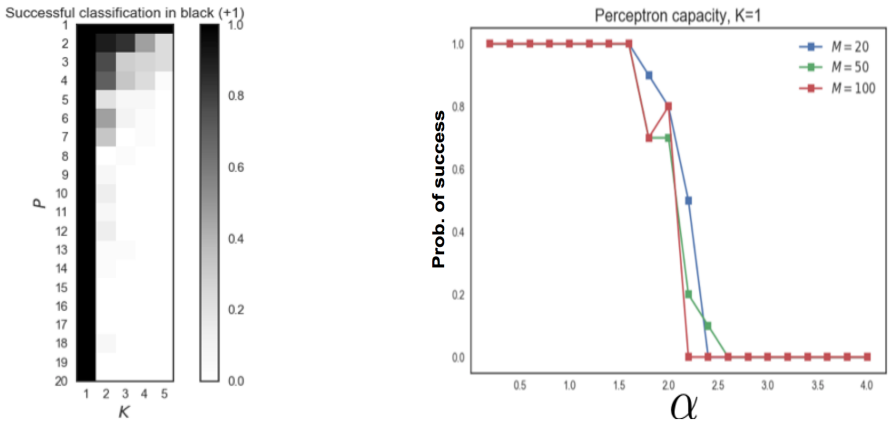


Figure 2.3: *Left*: The finite-size simulation results for the CDP, with $N = 50$ and $M = 50$. Shown is a feasibility matrix, where the entries indicate the probability that the linear inequalities of (2.1) are satisfied, averaged over 50 realizations. The figure shows that there are regions of finite and small $(P, K)$ pairs where the CDP can store perfectly. *Right*: The capacity of $P_c = 2N$ holds true for any choice of $M$ for the CDP, for $K = 1$.

The above results can indeed by generalized to multiple modalities. We can show that the rank of $\bar{\boldsymbol{\xi}} = P + (N_m - 1)K - (N_m - 1)$ for a generic number of modalities integrated [3]. Thus the capacity is significantly worse, and in fact reads

$$\alpha < \alpha_c \implies K < 2^{\frac{1}{N_m - 1}} \tag{2.12}$$

which again means $K = 1$ even for a large number of modalities since in that limit $K_c \approx e^{\frac{\log 2}{N_m - 1}} \approx 1 + \frac{\log 2}{N_m - 1} \to 1$.

[3] Here the modailities include the stimuli $\xi$, thus there are $N_m - 1$ modalities to be stored alongside $\xi$.

## 2.3 Finite stimuli storage in the presence of finite contexts

The above result displays that storage of an extensive number of stimuli is prohibitively difficult in the presence of contextual information. As shown in Fig (2.3) (left), there are values of *both* finite $P$ and $K$ for which the CDP can store perfectly. Hence, here we instead ask what the number of *finite* stimuli can be stored in the presence of a *finite* number of contexts. For this, we can first present an approximate *upper bound* on the number of finite stimuli stored by the suitably manipulating Eq. (2.4). We can make progress by further assuming that the *product* $PK \sim O(N)$, thus $\alpha \sim O(1)$. We will show that this fairly simple assumption leads to an upper bound on numerically evaluated $P_c$. Under the mentioned scaling, we can obtain from Eq. (2.4) self-consistent an equation for $P_c$, which reads

$$P_c K = 2(P_c + K - 1) \tag{2.13}$$

$$\implies P_c = \frac{2(K-1)}{K-2} \tag{2.14}$$

which holds for $K > 2$. [4] In any case, we find numerically that (2.14) is a good upper bound for all $K$. Note that such a discrepancy is expected since Eq. (2.14) is not the *exact* solution. For $K = 1$, the CDP can store an extensive number of $P_c = 2N$ patterns. This is somewhat intuitive since $\bar{\boldsymbol{\xi}}$ in such an instance can be reduced to a basis where there is only one vector $\in \mathcal{R}^{N+M}$, with the rest $(P-1)$ vectors $\in \mathcal{R}^N$. Thus, the problem is well approximated by counting $P$ constraints of random patterns in $\mathcal{R}^N$.

Instead of making uncontrolled approximations that follow from $PK \sim O(N)$ we can now obtain an *exact* lower bound, via Cover's theorem. Recall that the probability of successful linear separability, $\mathcal{P}$, for a generic number of degree of freedoms $N'$ and number of constraints $P'$ by

$$\mathcal{P} = \frac{C(P', N')}{2^{P'}} \tag{2.15}$$

where $C(P', N')$ is the number of dichotomies realized by $P'$ constraints in $N'$ dimensions (see Chapter 6 for exact form of $C$). For all finite $P$, a *lower bound* on $\mathcal{P}$ can be approximated by:

$$\mathcal{P} = \frac{C(P + K - 1, N + M)}{2^{PK}} \tag{2.16}$$

which can be numerically evaluated to give $P_c = 2$ for all finite $K$ (see Figure 2.4). This is plotted in Fig (2.4) in addition to the upper bound obtained by the Gardner (ShK) analysis, and we find the capacity to be in between the two.

[4] Note that for $K < 2$ the solutions are non-sensical. For $K = 2$ the equation has no solution, and for $K = 1$ the solution is negative.

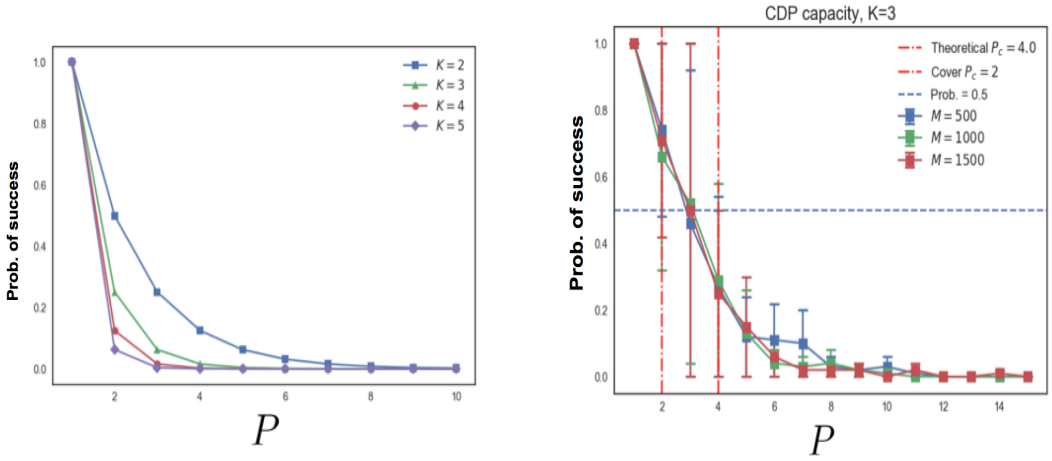


Figure 2.4: *Left*: The numerical evaluation of Eq. (2.16), for different values of K, providing a lower bound on the number of dichotomies we expect to separate. *Right*: The probability of realizing successful dichotomies are now numerically evaluated, with the lower and upper bounds given by (2.16) and (2.14) shown in vertical dotted red lines.

## 2.4 Discussion

To summarize, in this section we have applied the ShK result on the capacity of correlated datasets to the CDP, reproducing a result first given in [5], but in addition with an exact bound on the number of contexts that can be stored, and providing a comprehensive study of its storage capabilities. This will turn out to be a useful warm-up exercise, as we will use this same line of argument to derive the capacity of the CEMS model that we will introduce in the next Chapter.

# 2.5 Appendix

## 2.5.1 Derivation of rank of block data matrix

Consider the data matrix in Eq. (2.9). We can reduce redudancies in $\phi$, to get

$$(\bar{\boldsymbol{\xi}}^{\mu\gamma})^T = \begin{pmatrix} \xi^{\mu=1} & \phi^{\gamma=1} \\ \xi^{\mu=2} - \xi^{\mu=1} & 0 \\ .. & .. \\ \xi^{\mu=P} - \xi^{\mu=1} & 0 \\ \xi^{\mu=1} & \phi^{\gamma=2} \\ \xi^{\mu=2} - \xi^{\mu=1} & 0 \\ .. & .. \\ .. & .. \\ \xi^{\mu=P} - \xi^{\mu=1} & 0 \\ .. & .. \\ .. & .. \\ .. & .. \\ \xi^{\mu=P} - \xi^{\mu=1} & \phi^{\gamma=K} \end{pmatrix} \tag{2.17}$$

which has now $P-1$ redundant rows, repeated $K-1$ times. The rank of the $\hat{\boldsymbol{\xi}}$ is then $PK-(P-1)(K-1) = P+K-1$. We can repeat this argument for a generic block matrix of $N_m$ modalities to get $P+(N_m-1)K-(N_m-1)$ why? might not be true

## 2.5.2 Verification of ShK result

Eq. (2.4) requires the following procedure to generate the dataset:

- For a given $(N, P)$, generate a data matrix $\xi \in \mathcal{R}^{N\times P}$ and perform an SVD $\xi = UDV^T$.
- Choose a parameter $c$ and keep $cN$ largest entries of $D$ non-zero. Then rotate back to the full space to obtain a low-rank approximation, $\tilde{\xi} \in \mathcal{R}^{N\times P}$.
- Present an increasing subset $P'$ of columns to the network, and look for the $P_c$ above which the network cannot store the patterns.

Note that running the algorithm above amounts to choosing values of $c$ that satisfy:

$$c < \frac{1}{2}\min(P, N) \tag{2.18}$$

We present an example of low rank pattern reconstructions below. We can also verify that the ShK result holds true, although our lines and the theoretical prediction differ

slightly, perhaps owing to an insufficient number of realizations averaged over, or some loss of information during the low-rank reconstruction (note that the patterns are no more guaranteed to be $\pm 1$).

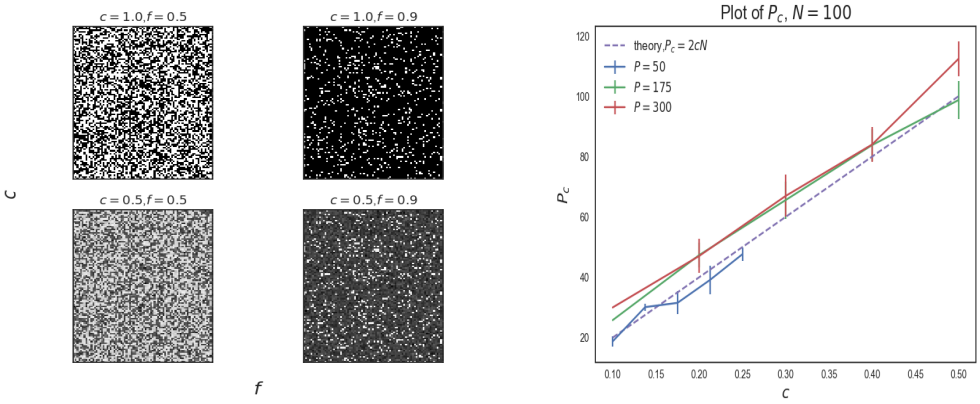


Figure 2.5: *Left*: Visualization of patterns for different values of $c$ and $f$. Note that we are unable to ensure $\pm 1$ entries upon reconstruction. *Right*: Validation of ShK result. Note that the method for extracting $P_c$ by construction will almost always result in an overestimate. Vertical bars indicate error bars of the simulations, averaged over 20 realizations.

### 2.5.3 Effect of sparsity on capacity

Gardner showed us [34] that as patterns are more correlated across rows, classification gets easier and the capacity accordingly diverges in the limit where the pixels are maximally correlated with each other and the outputs. [5]

We can verify that the ShK result in general does not appear to hold for different coding levels [6], repeating the simulation as in Fig. (2.5), but now with a coding level of $f = 0.1$ and $f = 0.9$ respectively. The main point to be made, however, is that sparseness of the data distribution itself does not (noticeably) improve capacity, as is proposed in famous cerebellar theories [58],[1].

[5] She called her parameter a "ferromagnetic bias", whereas we will use the more biologically prevalent term of "coding level", or "sparseness".

[6] We will formally define this in the next Chapter. Here, we can take $f$ to be the fraction of input neurons that are turned on.

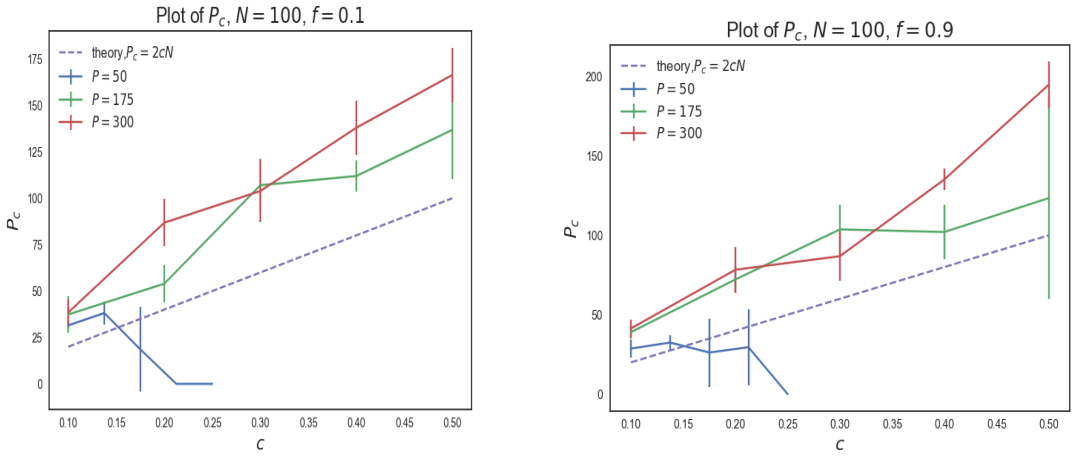


Figure 2.6: *Left* Verification of Eq. (2.6) for a low $f = 0.1$ and (*right*) a high $f = 0.9$. We see a deviation from the results shown above, indicating that the ShK result is incomplete.

### 2.5.4 Proof of capacity for the CDP in the most generic case

Let us stick to the case where $\min(P, K) = K$ (hence we need $K$ to be finite and $P$ to be extensive with $N$). Let us further consider the following four separate cases, where we will evaluate the rank of $\bar{\xi}$, $rk$, separately:

1. $N > P, M > K \implies rk = P + K - 1$

2. $N < P, M > K \implies rk = N + K$

3. $N > P, M < K \implies rk = \min(N, M) + P$

4. $N < P, M < K \implies rk = N + M$

Let us now prove that $K < 2$ follow for each case (Case 1 has already been proven in the main text)

1. See main text.

2. Here $\beta > 1$ and we obtain $\alpha = \frac{\beta}{1+\delta} \times K$, whereas $\alpha_c = \frac{2}{1+\delta}(\delta + O(K))$. Thus $\alpha < \alpha_c \implies K < 2/\beta \implies K < 2$.

3. Here $\alpha \approx \beta K$ and $\alpha_c = \frac{2}{N}(\min(N, M) + P)$. Note that in the limit where $N \to \infty$ we cannot have $M > N$. Using finally that $\beta < 1$ we have that $\alpha < \alpha_c \implies K < 2$.

4. As above we have $\alpha \approx \beta K$ and now $\alpha_c = 2$. Thus, $\alpha < \alpha_c \implies K < 2/\beta \implies K < 2$.

For the case where $\min(P, K) = P$ and $K$ is extensive with $M$, the proof proceeds exactly along the same lines as above, except we now have $P < 2$.

# Chapter 3

# The CEMS Model

We are now in a position to formalize the notion of pattern separation and expansion recoding in a model that interpolates between different degrees of non-linear mixing. In this Chapter, we will study the effect of mixing degree on associative learning and noise amplification of a generic ensemble of clustered stimuli on CEMS. We will present the surprising result that full mixing (and no less) is necessary to achieve an extensive storage capacity, whereas noise amplification is invariant as a function of selectivity for homogeneous input noise across modalities. The effect of these on the generalization of a Hebbian readout will be quantified in the next chapter.

We should note that the model we present here differs in key ways with that studied by Litwin-Kumar et. al [57], who also define a so-called "mixed" layer on a cerebellar-like circuit, and study the effect of the degree distribution of the MF-GC layer on the dimensionality. The key difference here is the notion of distinct modalities represented by *distinct MF neural populations*, whereas this distinction is not made in [57], who instead focus on the aforementioned affect of synaptic degree. In spirit, the model for independent sources of information shares similarities with Barak, et. al [5], and can be seen as an extension of theirs.

## 3.1 Recap of Babadi model

We first recap a canonical model (hereafter called the *unimodal model*) for a cerebellar-like circuit, first introduced by [3], that we will extend. We should mention as well an earlier influential model introduced by [5], that shares a similar structure though is inspired by behavioural experiments.

Formally, we can define an ensemble of clustered inputs, with its centers (centroids) given by $\hat{\xi} \in \mathcal{R}^{N\times P}$, and each members of the clusters are generated by flipping each neuron in the input layer independently with probability $\frac{\Delta\xi}{2}$, where $\Delta\xi$ can be taken to be an *input noise* parameter, that sets the size of the effective data manifold at the input

layer [1]. If $\Delta\xi = 1$, each neuron in each member of the cluster equally likely to be flipped from the centroid, yielding a maximally noisy input data. The input layer is the expanded onto a generic cortical, or *mixed*, layer whose outputs are given by $\mathbf{m}^\mu \in \mathcal{R}^{N_c \times P}$, and we can further define an *expansion ratio* $\mathcal{R} = \frac{N_c}{N}$. In this thesis, we take the feed-forward weights, $\mathbf{J} \in \mathcal{R}^{N_c \times N}$ to be a Gaussian random matrix, thus $J \sim \mathcal{N}(0, \frac{1}{\sqrt{N}})$ [2]. The choice of non-linearity we will use in this thesis is that of a biologically plausible Heaviside function, thus $m_i^\mu = \theta(h_i^\mu - T) \in \{0, 1\}$, with $h_i^\mu = \sum_k J_{ik}\xi_k^\mu$ and $T$ the threshold. Let the average value of the mixed layer activations to be $f$, then we will choose to scale the mixed layer activities to be zero centered, thus we have

$$m_i^\mu \to m_i^\mu - f = \theta(h_i^\mu - T) - f \tag{3.1}$$

The choice of having $\langle m_i - f \rangle = 0$ is simply a convention, and thus not affect the computational properties that we will study - i.e the results obtained will not change up to a global re-scaling. As it turns out, the choice of non-linearity can have drastic difference on the learning trend as a function of sparsity, which is a discussion we will defer to Chapters 4 and 5. Nevertheless, we have that the sparsity, or coding level, $f$ is set by $f = H(T)$, where $H(T) = \int_T^\infty \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz$ [3]. An important, but subtle consequence of this is $f \leq 0.5$, given that $T > 0$.

At the mixed layer, the distribution of neural activities will also have an inherent variability as a consequence of random projections of random input activities. We call this $\Delta m$, and we want to define $\Delta m$ such that the normalized statistics render the most noisy clusters with $\Delta m = 1$. The correct choice will turn out to be [3]

$$\Delta m = \frac{1}{2N_c f(1-f)} \sum_i^{N_c} \langle |\hat{m}_i - m_i| \rangle \tag{3.2}$$

where $\hat{\mathbf{m}}$ corresponds to the mixed layer activations of the projected centroids $\hat{\boldsymbol{\xi}}$. It has been shown previously [3] that such an expression for $\Delta m$ is a monotonically increasing function of $\Delta\xi$ and $f$, and we will represent this by

$$\Delta m = \mathcal{G}(T, \Delta\xi) \tag{3.3}$$

with the exact closed form expression for $\mathcal{G}$ given in the Appendix. Note that $\Delta\xi$ is always amplified in our model because of the assumption of random feed-forward weights; other

[1] We note that the exact number of members in a *given* cluster is irrelevant. It is only necessary for our computational measures (clusters size, readout error), that we have a the centroid along with another *typical* member of the cluster. And so for simulation purposes we do not specify the number of members of a cluster.

[2] Such a normalization of course assumes that the inputs are $\{+, -\}$, which is what we use in this thesis without loss of generality.

[3] Another potentially confusing terminology that is worth mentioning is that *high sparseness* refers to a *small* f.

mechanisms such as structured weights may indeed suppress the input noise [3]. We also note that the readout weights, as will be discussed in the next Chapter, is imposed to obey a supervised Hebbian rule.

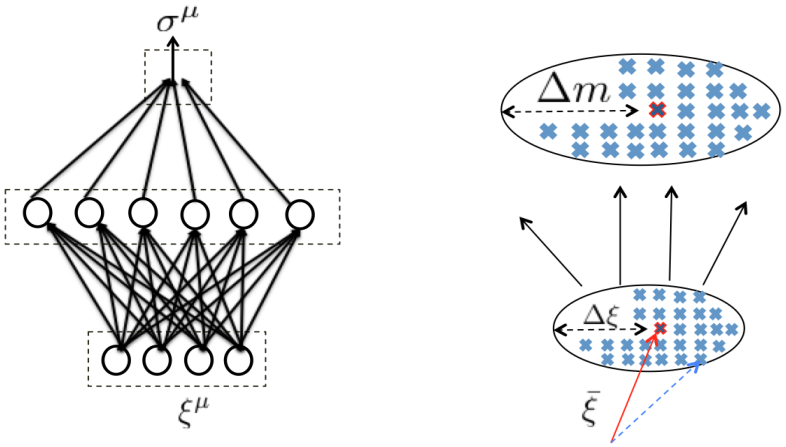


Figure 3.1: A recap of the unimodal model. *Left:* Shown is the generic architecture of such a model, where random inputs are projected onto a higher dimensional cortical layer, which is then read out by a Hebbian weights at the readout layer. *Right:* A schematic of noise amplification across the processing hierarchy. At the input layer (axes denoting input state space), the cluster has a size has size $\Delta\xi$, which then gets amplified to a size $\Delta m$ at the cortical layer.

## 3.2 CEMS as a generic model of mixing

Let us formalize the modelling choice of the CEMS in the following subsections, following the introduction in Chapter 1.

### 3.2.1 Independent sources of information projected by independent random weights

A central assumption of our model builds on that used by Barak, et. al [5], where we assume different modalities are modelled by independent random inputs $\{+,-\}$, similar to the inputs received by the CDP in the previous Chapter (hence the data matrix effectively looks like that in Figure (2.2), bottom right). In this setting, the expansion ratio is now defined as

$$\mathcal{R} = \frac{N_c}{(1+(N_m-1)\delta)N} \tag{3.4}$$

Formally, let us consider the case $N_m = 3$, the input data matrix $\bar{\boldsymbol{\xi}} \in \mathcal{R}^{N(1+2\delta),PK^2}$ will be of the form

$$\bar{\boldsymbol{\xi}}^{(\mu,\gamma,\rho)} = \begin{pmatrix} \xi^{\mu} & \phi^{\gamma} & \eta^{\rho} \end{pmatrix} \tag{3.5}$$

where each of $\mu = 1, .., P$, $\gamma = 1, ..., K$, $\rho = 1, ..., K$. Each of these sub-blocks have are independently drawn $\{+, -\}^{N_{mod}}$, with $N_{mod}$ denoting the number of input neurons for each modality (i.e $N$ for the task relevant one, $M$ for the contextual ones). As before, we will define $\beta = \frac{P}{N}$ as the *task-relevant* load, which we will use to calculate the capacity of CEMS. The labels (valences) will now have the same composite number of indices as the input data, i.e for $N_m = 3$ we have $\sigma^{\mu\gamma\rho}$.

This input data, will then be projected onto a mixed layer by independent, *randomly distributed* weights [15],[41], normalized such that sparsity is maintained as a function of mixing, as explained further below. This choice of random weights is admittedly arbitrary an over-simplification, though it will enable us to produce analytical results as a function of mixing as a null model. Such simple assumptions, however, appear to be pertinent to olfactory circuits in mammalians [59] and insects [84].

### 3.2.2 A sparsity and modality-coverage preserving model of mixing

As an extension to the unimodal model, we now have each neuron in the mixed layer that is selective to a mixture of different modalities on the input layer, the degree to which is quantified by a *mixing index* $\mathcal{M}$, with $1 \leq \mathcal{M} \leq N_m$, $\mathcal{M} \in \mathcal{Z}^+$. Crucially, for the sake of modelling, and partly inspired by theoretical works of [49] [4], we will assume all the neurons on the mixed layer are subject to the same level of mixed selectivity thus we will not consider cases where some neurons mix inputs whilst others do not, as is seen from many empirical results. This is a simplification, but it will help us make theoretical progress, and to systematically probe how the computational properties of the network change as a function of $\mathcal{M}$. We will discuss some limitations of the model based on results obtained at the end of Chapter 4.

#### Partitioning scheme

We refer to architectures where $\mathcal{M} = N_m$ as *fully mixed*, such that each neuron on the cortical layer non-linearly mixes all modalities, where $1 < \mathcal{M} < N_m$ as *partially mixed*, such that each neuron mixes the task-relevant stimulus $\boldsymbol{\xi}$ with at least one other modality, and where $\mathcal{M} = 1$ as *purely* mixed, where $N_c/N_m$ neurons receive distinct inputs on the cortical layer with no mixing. In general, the number of partitions to be constructed on

[4] An example of a model that smoothly interpolates between different degrees of mixing.

Not True

the mixed layer for $N_m$ modalities for a mixing index $\mathcal{M}$ is $N_m - \mathcal{M} + 1$.

Such a partitioning scheme can be regarded as one possible mechanism to impose a fixed *modality-coverage*, such that for all values of $\mathcal{M}$ all modalities are ensured to be received by the network. Some notable experiments in multi-modal integration in cerebellar-like circuits display that a minority of neurons on the cortical layer may mix different sources of information, but others might not [47], [71]. An extension of our theory to more biologically realistic models will be a key further step.

### Feed-forward weight distribution

Given the above interpolation scheme, we now have to ensure the feed-forward random weights are properly normalized, such that sparseness is fixed for different values of $\mathcal{M}$. It is straightforward to show (see Appendix), that we require

$$J \sim \mathcal{N}(0, \frac{1}{\sqrt{\mathcal{M} N_{mod}}}) \tag{3.6}$$

### An example

Let us for illustration purposes demonstrate the architectures for when $N_m = 3$ and the cases where $\mathcal{M} = 1$ and $\mathcal{M} = 3$. For the former case, we have three sub-populations on the mixed layer, each with $N_c/3$ number of neurons, and each selective to a distinct modality at the input layer. Thus, each of the activations of each sub-population is given by

$$h_i^{\mu} = \sum_j J_{ij}^{(\xi)} \xi_j^{\mu}, \quad h_i^{\gamma} = \sum_j J_{ij}^{(\phi)} \phi_j^{\gamma}, \quad h_i^{\rho} = \sum_k J_{ik}^{(\eta)} \eta_k^{\rho} \tag{3.7}$$

with each of the $J \sim \frac{1}{\sqrt{N_{mod}}}$. The effective input matrix $\in \mathcal{R}^{N_c \times PK^2}$ given by

$$\mathbf{h}^{(\mu,\gamma,\rho)} = \begin{pmatrix} h^{(\xi_1)} & \dots & h^{(\xi_P)} & h^{(\xi_1)} & \dots & h^{(\xi_P)} & \dots & \dots & h^{(\xi_P)} \\ h^{(\phi_1)} & h^{(\phi_1)} & h^{(\phi_1)} & h^{(\phi_2)} & h^{(\phi_2)} & h^{(\phi_2)} & \dots & \dots & h^{(\phi_K)} \\ h^{(\eta_1)} & h^{(\eta_1)} & h^{(\eta_1)} & h^{(\eta_1)} & h^{(\eta_1)} & h^{(\eta_1)} & \dots & \dots & h^{(\eta_K)} \end{pmatrix} \tag{3.8}$$

where we have enumerated all $PK^2$ possible combinations of input columns. For the other extreme of $\mathcal{M} = 3$, we have that

$$h_i^{(\mu,\gamma,\rho)} = \sum_j J_{ij}^{(\xi)} \xi_j^{\mu} + \sum_j J_{ij}^{(\phi)} \phi_j^{\gamma} + \sum_k J_{ik}^{(\eta)} \eta_k^{\rho} \tag{3.9}$$

with $J \sim \frac{1}{\sqrt{3N_{mod}}}$, and the full input matrix (again $\in \mathcal{R}^{N_c \times PK^2}$) given by

$$\mathbf{h}^{(\mu,\gamma,\rho)} = \begin{pmatrix} h^{(\xi_1,\phi_1,\rho_1)} & h^{(\xi_2,\phi_1,\rho_1)} & \dots & h^{(\xi_P,\phi_K,\rho_K)} \end{pmatrix} \tag{3.10}$$

An illustration of the different architectures for the above cases is shown in Figure (3.2), top and middle row.

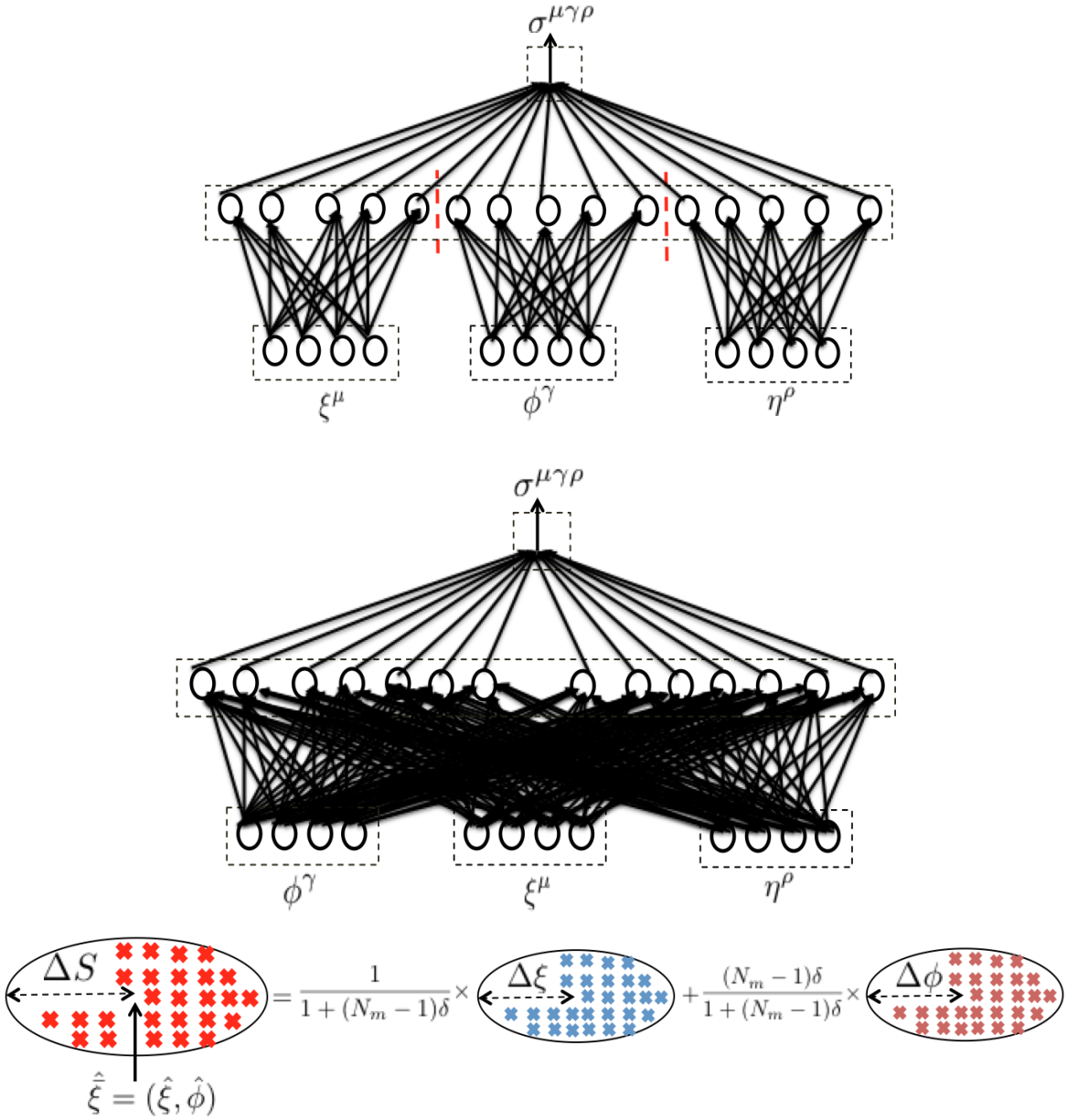


Figure 3.2: *Top:* Example of the interpolation scheme (partitioning scheme) of CEMS, for $N_m = 3$, for $\mathcal{M} = 1$. *Middle:* The corresponding partitioning scheme for $\mathcal{M} = 3$ is shown. Note that the spatial structure at the top (dashed vertical red lines) is artificial - we simply emphasize that one-third of the neurons receive inputs from one distinct modality. Note as well that the valences have a three-indexed label, unique for each composite input combination. *Bottom:* Noise model of the CEMS, for $N_m = 2$. The net cluster size $\Delta S$ is given by a weighted sum of the task-relevant variability $\Delta\xi$, and a contextual variability, $\Delta\phi$, which is the same across all contextual modalities.

## 3.3 Capacity of CEMS

We can now calculate the capacity of the CEMS model as a function of mixing index. To do this, we simply re-state the steps used in the derivation of the capacity of the CDP, as explained in the previous section. We first consider the case of $\mathcal{M} = N_m$ (fully mixed), in which case the matrix $\mathbf{m} \in \mathcal{R}^{N_c \times PK^{N_m-1}}$ is random and its rank is given by $\min(N_c, PK^{N_m-1})$. The line of critical capacity given by the ShK theory of Eq. (2.4), which is when the critical number of composite patterns $PK^{N_m-1} = 2N_c$, above which associative learning is impossible. On the other hand, for a large enough expansion ratio, any number of composite input-output mappings are implementable [88].

This trend can be summarized by a putative "phase-diagram" in Fig (3.3), with the red line of $PK^{N_m-1} = 2N_c$ indicating the delimiting line given by Eq. (2.4), above which (in region **I**) the CEMS cannot store any stimuli-context pairs, and below which (region **II**) the CEMS has perfect storage. The blue line indicates the point at which the effective dimensionality of the problem changes. Focusing on regions **I** and **II**, however, we have that

$$N_c > \frac{PK^{N_m-1}}{2} \tag{3.11}$$

as the minimum number of mixed layer neurons required for perfect associative learning. We can now as what the task-relevant capacity is and re-arrange (3.11) to get

$$\beta_c = \frac{2\mathcal{R}\Big(1 + (N_m - 1)\delta\Big)}{K^{N_m-1}} \qquad \text{True but trivial} \tag{3.12}$$

as the critical capacity for a given $K$ (see red vertical dashed lines in Figure (3.3)). To obtain the rank of the partially selective case, let us specialize to the case where $N_m = 3$, and consider the case of $\mathcal{M} = 2$. Here, by the same line of argument in Chapter 2 we have that the rank of the matrix [5] $\mathbf{m}$ is $P + K - 1$ and thus

$$\alpha < \alpha_c \implies K < (2)^{\frac{1}{3-1}} \implies K_c = 1 \tag{3.13}$$

in exact analogy to the derivation of the capacity of the CDP . This result is the same for $\mathcal{M} = 1$, where the rank is now $P + 2K - 2$, resulting in the same computational capacity as the tri-modal CDP. These results are shown in Figure (3.3), bottom, where we display the results of the feasibility check for $N_m = 3$ and $\mathcal{M} = 2$, where for $K = 1$ (on the left) and $K = 2$ (on the right). An illustration of how the rank changes with mixing is given by simulation results in Figure (3.4), bottom left, for $K = 9$, and $P = 100$. On the rank of the fully mixed layer increases with the expansion ratio, and saturates above $N_c = PK^2$. Such a result generalizes for any $\mathcal{M} < N_m$ - i.e anything less than full mixed selectivity

[5] Note that $\mathcal{M} = 2$ the intermediate case of (3.7) and (3.9)

on the CEMS leads to a storage bottleneck of one contextual variable.

Such a scaling law for $N_m = 3$ is displayed in Figure (3.4), top and bottom right. On the top, we run the linear program for different values of $K$, and plot the theoretical line of capacity in the dotted horizontal lines. On the bottom right, we numerically calculate $\beta_c$ from the previous linear program check, but now on the bimodal CEMS, and plot its value for different values of $K$, with the theoretical scaling law shown in dotted lines for different $\mathcal{R}$. These results display a simple interplay between the expansion ratio, $\mathcal{R}$, and the finite number of contexts $K$ to be stored across modalities, in determining the task-relevant capacity.

This scaling law readily makes for comparisons between the computational capabilities of different circuits by simply comparing $\mathcal{R}$, the expansion ratio. Comparing to typical expansion ratios found in the cerebellum ([10]) and piriform cortex ([84]) of $\sim \mathcal{R} = 10^3$, and assuming $\delta = 1$ with $N_m = 3$ (e.g auditory and visual inputs in addition to the task-relevant one as in the [48] experiment), we have that $\beta_c = \frac{6}{K^{N_m - 1}} \times 10^3$. Asking for how many finite contexts can the network attain a task-relevant capacity as well as the classical Gardner result [34], we have that $K \approx (3 \times 10^3)^{\frac{1}{N_m - 1}}$, which $\approx 55$ distinct contextual inputs across each modality. Whereas for more modest expansion ratios as found in the dentate gyrus [6] of $\mathcal{R} \approx 5 \implies K = \sqrt{15} \implies K \approx 4$ for the same level of performance, indicating a significantly smaller capability for storing contextual information in the CEMS model for such a circuit. [6] A systematic comparison of the capacity of the CEMS model with biologically plausible estimates of parameters, and in addition to known associative learning capabilities, could be interesting further work.

[6] Both of these of course assume that we are in Region (**I** - **II**) of our phase-diagram, which along with the assumption that $\delta = 1$, are hypotheses that in principle could be empirically checked.

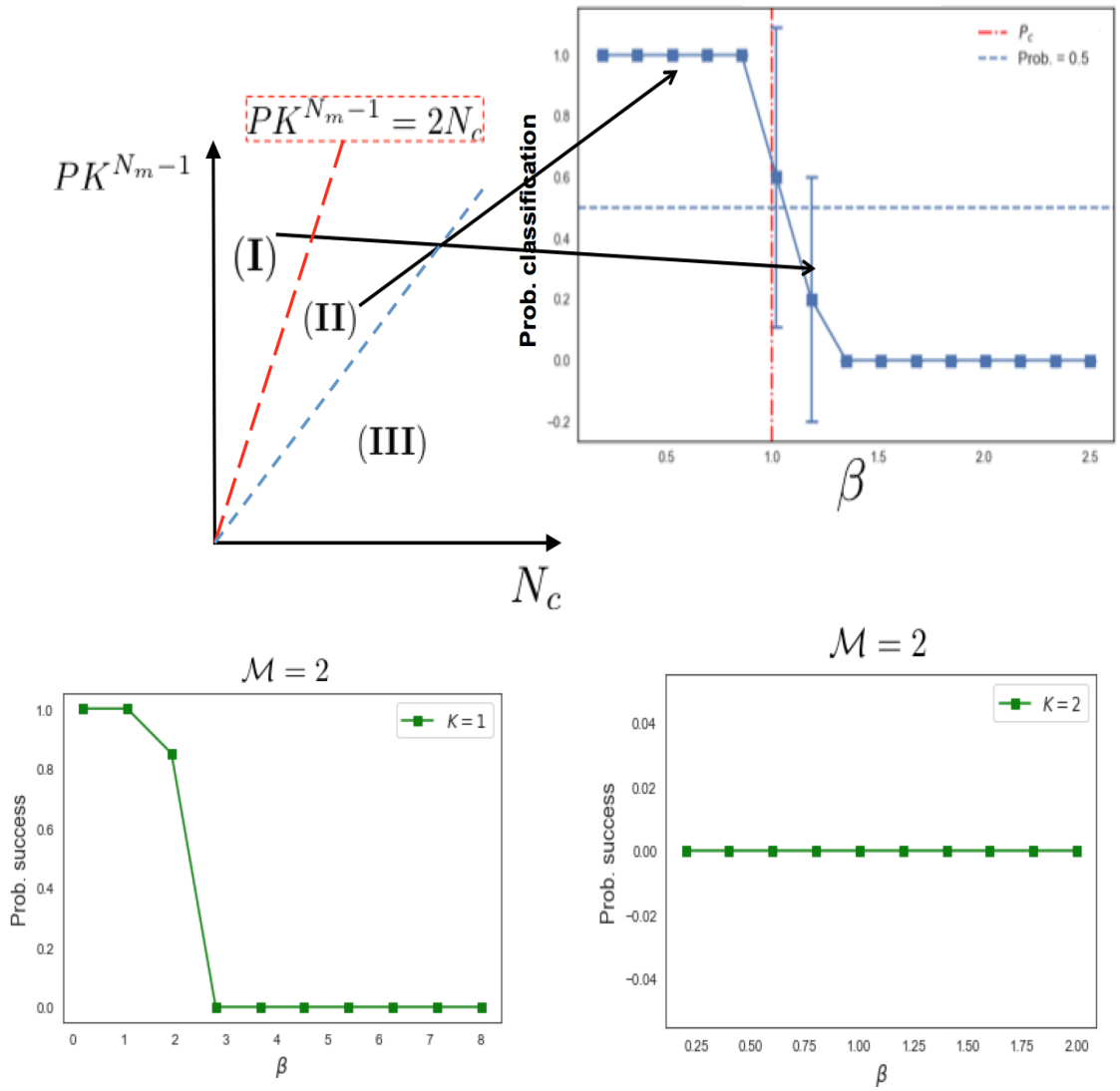


Figure 3.3: Capacity of CEMS. *Top:* The "phase-diagram" of CEMS' $\mathcal{M} = 3$ capacity is presented. The blue dotted line indicates the line at which the rank of the effective mixed layer inputs changes; below this line capacity is extensive and above this a scaling law is predicted. The red dotted line indicates the critical line corresponding to the ShK result. Inset shows linear feasibility check on corresponding regions; here we have $\delta = 0.5$, $\mathcal{R} = 1$ and $K = 2$ therefore $\beta_c = 1$ from (3.12). Note that in the region (**III**), any input-output mappings can be associated. *Bottom:* Linear program feasibility check is run on the $\mathcal{M} = 2$ CEMS, showing that for $K = 1$ (left), the Gardner result is reproduced, whereas for $K = 2$ (right), extensive storage of task-relevant stimuli is not possible. Simulation parameters for the linear program are $N = 100, N_c = 200$.

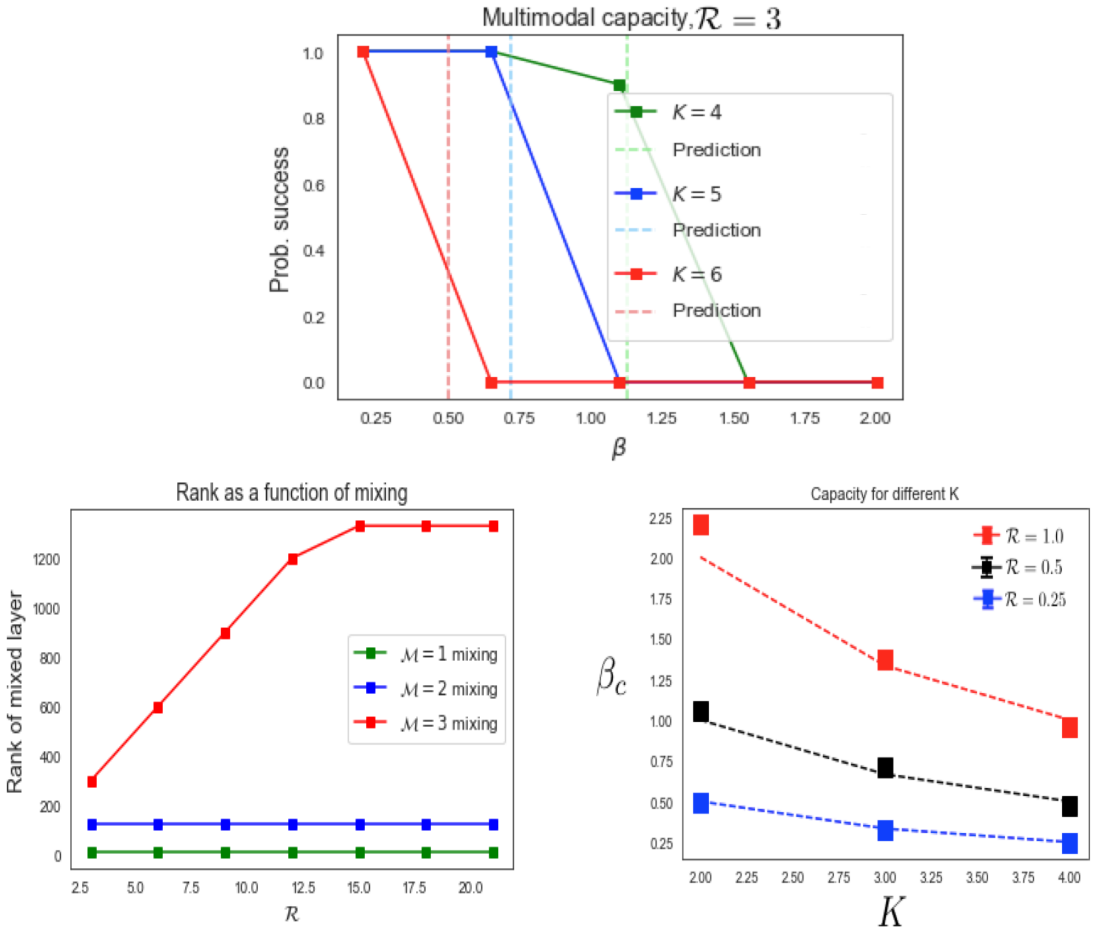


Figure 3.4: *Top:* The feasibility check is performed for the $\mathcal{M} = 3$ case for different $K$ for a fixed $\mathcal{R}$, with the theoretical predictions displayed in vertical dotted lines for each $K$, showing an agreement with the empirical results. Simulation details are $N = 100$,$M = 50$,$\mathcal{R} = 3$. *Bottom:* On the left, we show the evolution of the rank of the mixed layer activities as a function of the expansion ratio. Unless there is full mixing, the rank does not grow as a function of $\mathcal{R}$, whereas with full mixing the rank saturates at $N_c$. On the right, the simulation results for the $N_m = 2$ CEMS is shown, and $\beta_c$ is estimated retrospectively from the linear check, and we see that (3.12) is reproduced.

## 3.4 Mixed layer clusters size as a function of selectivity

We can also extend the unimodal model with regards to the noise structure on the input layer, and ask how this gets transformed at the mixed layer.

### Noise model

Moreover, we introduce separate noise parameters, $\Delta\xi$ for the variability of the stimulus, and $\Delta\phi$ for variability for all other modalities. This is done for simplicity without loss of generality. Is it not too hard to show that the net input noise (see Appendix), which we call $\Delta S$ is given by the following formula:

$$\Delta S(N_m, \delta) = \frac{1}{1+(N_m-1)\delta}\Delta\xi + (N_m-1)\frac{\delta}{1+(N_m-1)\delta}\Delta\phi \tag{3.14}$$

See Fig (3.3), bottom for a cartoon illustration of this.

### A behaviorally distinct noise amplification for the partially selective CEMS

We can now ask what the effect of mixing is on the noise amplification at the mixed layer, given the statistics of the inputs and the random feed-forward weights. Let us call the effective input noise parameters into the closed form function $\mathcal{G}$ defined in (4.40) as $\Delta S_{eff}(\mathcal{M})$, since it depends on the mixing index. In the unimodal case, recall we had that $\Delta S$ was trivially given by $1-\frac{1}{N_c}\langle \mathbf{h}^\mu \hat{\mathbf{h}}^\mu \rangle$, with the average performed over all patterns $\mu$. We can extend this to the CEMS, and give normalized distance distance each *composite input* (i.e the linear projections of the centroids) $\hat{\mathbf{h}}^{(\mu,\gamma,\rho)}$ is from the corresponding perturbed input $\mathbf{h}^{(\mu,\gamma,\rho)}$ (i.e the linear projections of the other cluster members). In particular,

$$\Delta S_{eff} := 1 - \frac{1}{N_c}\langle \mathbf{h}^{(\mu,\gamma,\rho)} \hat{\mathbf{h}}^{(\mu,\gamma,\rho)} \rangle \tag{3.15}$$

where the average is performed over all $PK^2$ composite inputs. When $\Delta S_{eff} = 1$ the training and test inputs are maximally uncorrelated. Note that in the unimodal case, the distance between inputs and between training and test stimuli are equal; here they are not. Then, it is not too hard to show (see Appendix) that

$$\Delta S_{eff}(\mathcal{M}=1) = \Delta S_{eff}(\mathcal{M}=N_m) = \Delta S(\delta=1) \tag{3.16}$$

whereas for any partially selective case

$$\Delta S_{eff}(1<\mathcal{M}<N_m) = \Delta S(\mathcal{M}, \delta=1) \tag{3.17}$$

with $\Delta S$ defined in (3.14) [7] For the case of $N_m = 3$, we have respectively $\Delta S_{eff}(\mathcal{M}=1) = \Delta S_{eff}(\mathcal{M}=3) = \frac{1}{3}(\Delta\xi + 2\Delta\phi)$, $\Delta S_{eff}(\mathcal{M}=2) = \frac{1}{2}(\Delta\xi + \Delta\phi)$.

[7] A couple of notational points are in order. Firstly, in (3.17), the first argument implies that $N_m$ is to be replaced with $\mathcal{M}$ wherever it appears in (3.14). Secondly, (3.16) and (3.17) are both valid for any value of $\delta$, although the expression is given by (3.14) evaluated at $\delta = 1$. Finally, note of course that there strictly speaking should be arguments of $\Delta\xi$ and $\Delta\phi$ in Eq. (3.14). Here, we assume that is implicit as we want to emphasize the dependencies on $N_m$ and $\delta$.

A couple of interesting consequences follow from this. Firstly, we have that for a homogeneous noise across modalities, $\Delta m$ is invariant across $\mathcal{M}$, since (3.16) and (3.17) are equal, suggesting a regime where the noise robustness of CEMS is independent of the degree of selectivity. Secondly, for a fixed task-relevant variability $\Delta\xi$, (3.16) and (3.17) suggest that partial selectivity, somewhat counter-intuitively, is the best strategy, since (3.16) is more sensitive to variations in $\Delta\phi$. Finally, for a fixed contextual variability $\Delta\phi$, being either purely or fully selective is instead better. Note that in all cases, the sensitivity to $f$ as in the unimodal case (see Appendix for plot) is recapitulated.

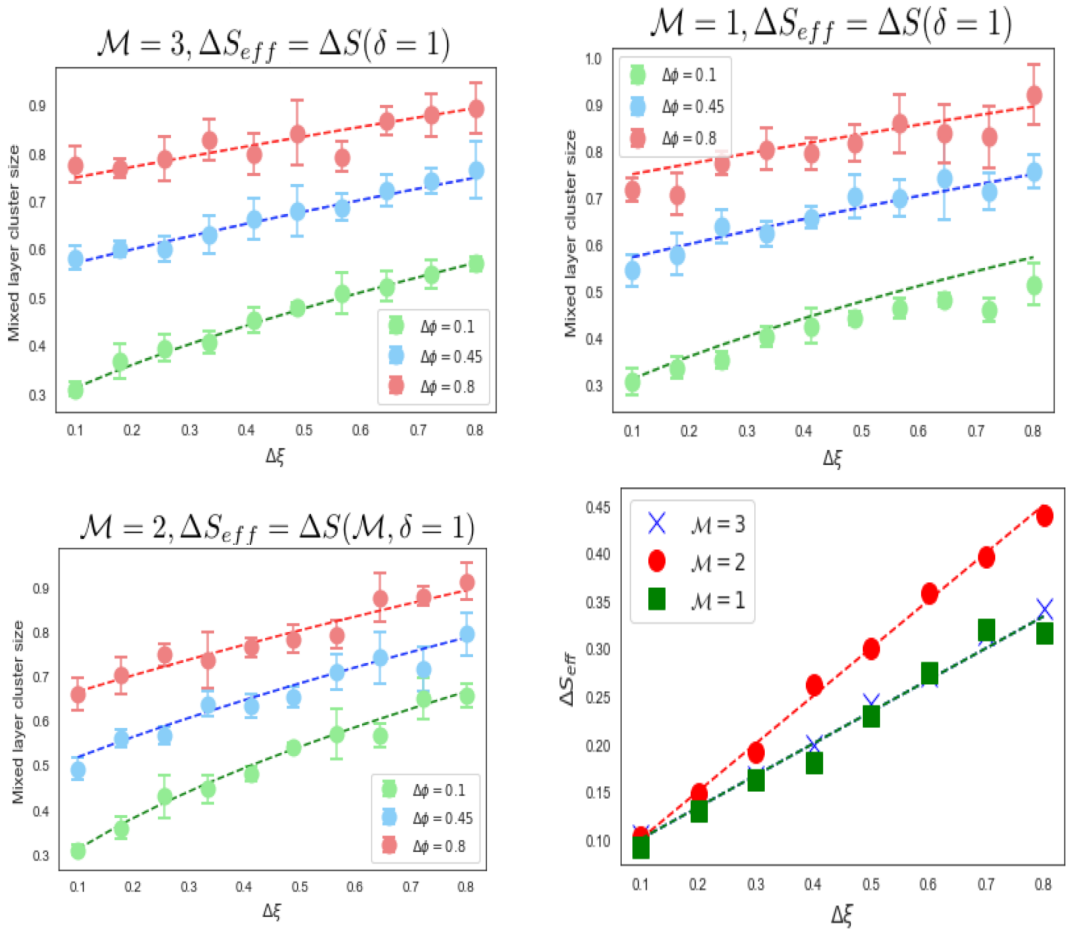


Figure 3.5: *Top*: Fixed noise amplification for the fully (*left*) and purely (*right*) selective cases. *Bottom*: On the left is displayed the mixed layer cluster size for the partially selective CEMS. On the right we display the numerical evaluation of $\Delta S_{eff}$, by subtracting 1 from the computed normalized $\langle \mathbf{h}^{(\mu,\gamma,\rho)} \hat{\mathbf{h}}^{(\mu,\gamma,\rho)} \rangle$, for a fixed $\Delta\phi = 0.1$, with the theoretical predictions in dotted lines. This suggests that for a fixed $\Delta\phi$, the partially selective CEMS is less robust towards behaviourally relevant variability.

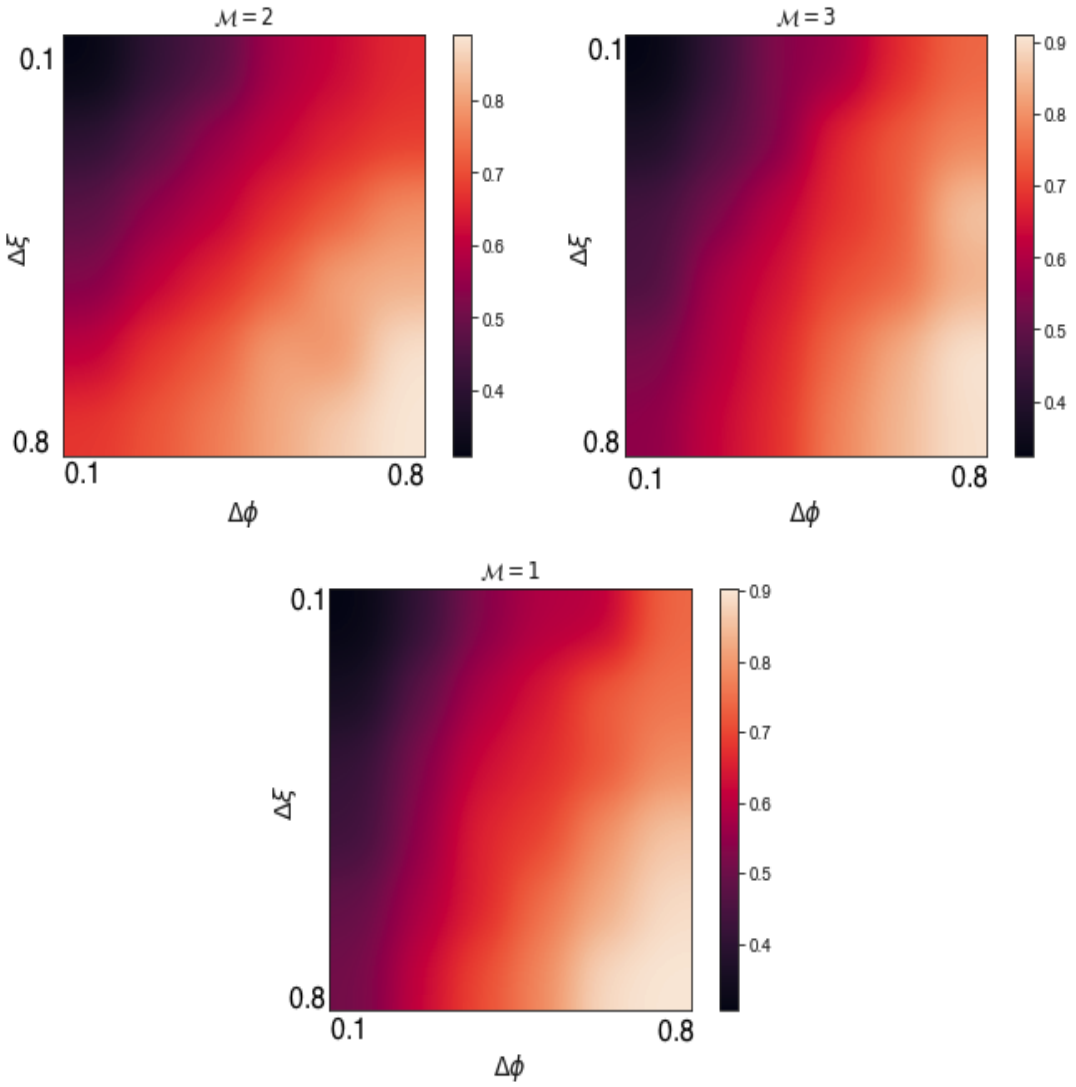


Figure 3.6: Illustration of different noise robustness for different selectivity indices $\mathcal{M}$, via heat maps of $\Delta m = \mathcal{G}(\Delta S_{eff}, T)$ across different $\Delta\phi$ and $\Delta\xi$. *Left*: The partially selective CEMS is more robust towards contextual noise, as can be seen by the growth of $\Delta m$ across $\Delta\phi$ for small $\Delta\xi$ (i.e top left of figure, moving to the right), but less robust towards task-relevant noise, as is seen by now tracking the growth of $\Delta m$ across $\Delta\xi$ for small $\Delta\phi$. The reverse effect occurs for the full selective CEMS (*right*), which is now more robust towards variations in $\Delta\xi$, but more sensitive towards those in $\Delta\phi$. *Bottom:* The heat map for $\mathcal{M} = 1$ is the essentially the same as on the top right.

The origin of this counter-intuitive non-monoticity in $\Delta m$ as a function of $\mathcal{M}$ is solely a consequence of our sparsity-conserving choice of weight scaling. This suggests a basic inter-dependence between the variation of sparseness as a function of selectivity and the computational properties of noise amplification at the mixed layer. Exploring further

consequences of such an interplay may be interesting future work.

## 3.5 Discussion

Under a model that: (i) Smoothly interpolates between different degrees of mixing whilst preserving modality-coverage on the mixed layer, and (ii) Does so whilst maintaining sparsity of the representation, we have uncovered some basic results regarding its computational properties. We have shown that the capacity can only be extensive for the fully mixed model, whereas any smaller degree of selectivity leads to a storage bottleneck of *exactly* one contextual variable. Then, we have shown that sparsity preserving constraint on the model leads to non-monoticity in the noise amplification at the mixed layer with respect to mixing degree; in particular that the fully and purely mixed network are more sensitive to contextual input variability and more robust to task-relevant input variability, whereas the partially mixed network displays the opposite effect. Further, that the noise robustness is invariant across selectivity levels if the input variability is homogeneous across modalities.

The results here provide what we believe to be a first quantitative theory of Marr's expansion recoding hypothesis [58], under a multi-modal paradigm. Importantly, the CEMS also constrains the scope of the utility of expansion recoding for different degrees of mixing; in particular anything less that full mixed selectivity will lead to the aforementioned prohibitive storage bottleneck. In the next section, we will build on these and other novel results to study the generalization abilities of a readout neuron as a function of mixing degree.

## 3.6 Appendix

### 3.6.1 Normalization of weights in CEMS

Here we will pedagogically explain how we appropriately normalize the weights, such that the CEMS is sparsity-conserving as a function of $\mathcal{M}$. As in the main text, we focus here on the case where $N_m = 3$, with three separate possible architectures shown in Fig (3.7) below.

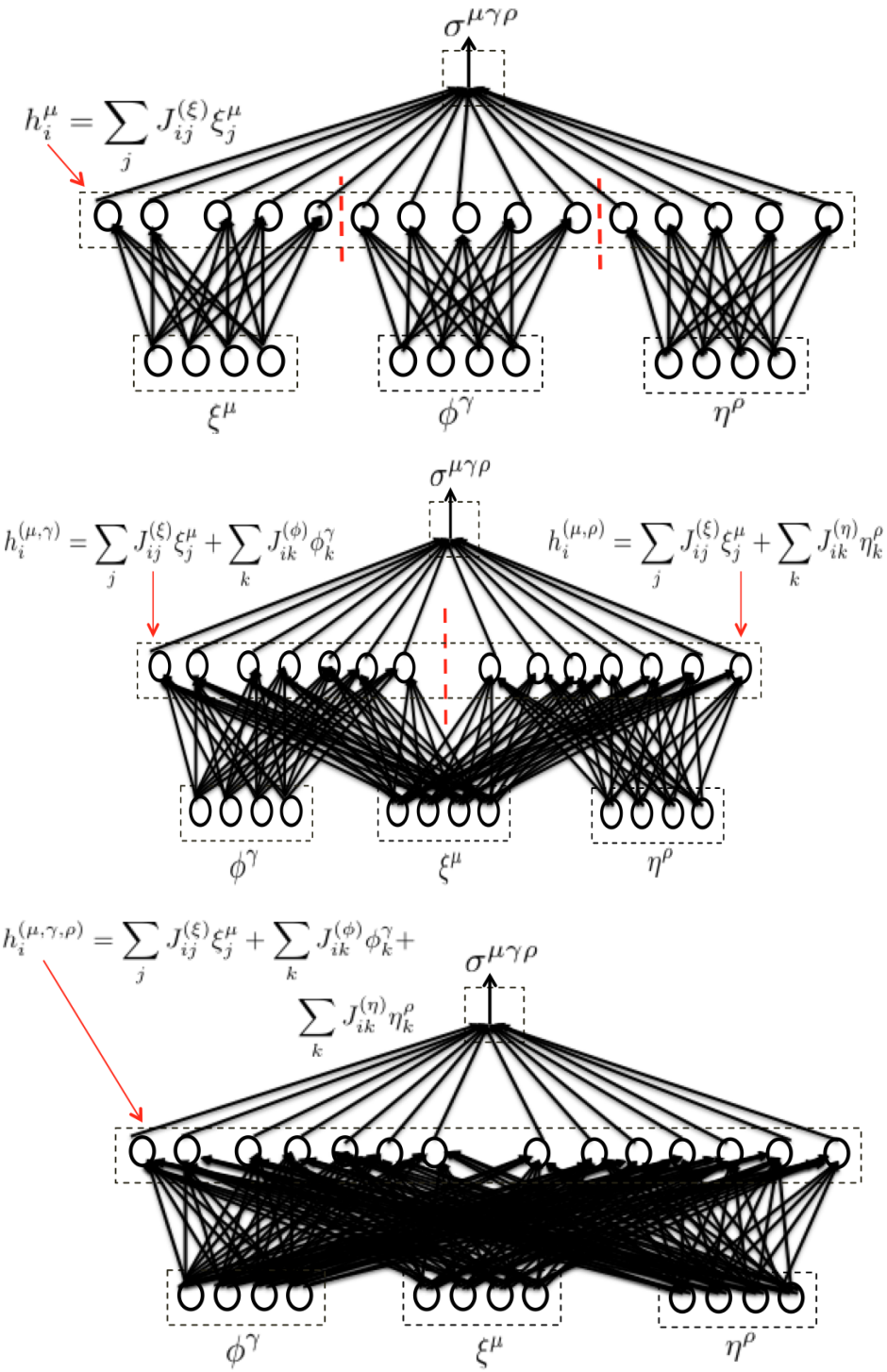


Figure 3.7: From top to bottom, an illustration of the CEMS architecture as a function of mixing. *Top:* Purely selective case of $\mathcal{M} = 1$, *Middle:* Partially selective case of $\mathcal{M} = 2$, and *Bottom:* Fully selective case of $\mathcal{M} = 3$. Again, any spatial partitioning is artificial and is simply to illustrate what number of neurons on the mixed layer receive which inputs.

We want to scale the weights appropriately such that the sparsity is conserved across different values of $\mathcal{M}$. First, we consider the unimodal case for comparison, where we want to normalize the inputs at the cortical layer to havae unit norm on average, thus $\langle||\mathbf{h}||_2\rangle = \frac{1}{N_c}\sum_i^{N_c}\langle h_i^2\rangle = \frac{1}{N_c}\sum_i^{N_c}\sum_{j,j'}^{N}\langle J_{i,j}J_{i,j'}\rangle\langle\xi_j\xi_{j'}\rangle$. Hence, we set $J \sim \mathcal{N}(0, \frac{1}{\sqrt{N}})$, since $\langle\xi_j\xi_{j'}\rangle = \delta_{j,j'}$. Now, we consider the case where $\mathcal{M} = 1$, shown a the top of Figure (3.7). The input matrix given in (3.7). We have that

$$\langle||\mathbf{h}||_2\rangle = \frac{1}{N_c}\sum_i^{N_c/3}\langle\sum_{j,j'}^{N} J_{ij}^{(\xi)}J_{ij'}^{(\xi)}\rangle\langle\xi_j\xi_{j'}\rangle + \frac{1}{N_c}\sum_i^{N_c/3}\sum_{k,k'}^{M}\langle J_{ik}^{(\phi)}J_{ik'}^{(\phi)}\rangle\langle\phi_k\phi_{k'}\rangle + \frac{1}{N_c}\sum_i^{N_c/3}\sum_{l,l'}^{M}\langle J_{il}^{(\eta)}J_{il'}^{(\eta)}\rangle\langle\eta_l\eta_{l'}\rangle$$
$$= \frac{1}{3}\Big(N\langle J^{(\xi)2}\rangle + M\langle J^{(\phi)2}\rangle + M\langle J^{(\eta)2}\rangle\Big) \quad (3.18)$$

To have a unit norm, we see that the only possible choice of weight scalings are

$$J^{(\xi)} \sim \mathcal{N}(0, \frac{1}{\sqrt{N}}), \quad J^{(\phi)} \sim \mathcal{N}(0, \frac{1}{\sqrt{M}}), \quad J^{(\eta)} \sim \mathcal{N}(0, \frac{1}{\sqrt{M}}) \quad (3.19)$$

Next, for $\mathcal{M} = 2$, on the middle of Figure (3.7). The input norm now reads

$$\langle||\mathbf{h}||_2\rangle = \frac{1}{2}\Big(\sum_{j,j'}^{N}\langle J_{ij}^{(\xi)}J_{ij'}^{(\xi)}\rangle\langle\xi_j\xi_{j'}\rangle + \sum_{k,k'}^{M}\langle J_{ik}^{(\phi)}J_{ik'}^{(\phi)}\rangle\langle\phi_k\phi_{k'}\rangle\Big) \quad (3.20)$$
$$+\frac{1}{2}\Big(\sum_{j,j'}^{N}\langle J_{ij}^{(\xi)}J_{ij'}^{(\xi)}\rangle\langle\xi_j\xi_{j'}\rangle + \sum_{k,k'}^{M}\langle J_{ik}^{(\eta)}J_{ik'}^{(\eta)}\rangle\langle\eta_k\eta_{k'}\rangle\Big) \quad (3.21)$$
$$= \frac{1}{2}\Big(N\langle J^{(\xi)2}\rangle + M\langle J^{(\phi)2}\rangle\Big) + \frac{1}{2}\Big(N\langle J^{(\xi)2}\rangle + M\langle J^{(\phi)2}\rangle\Big) \quad (3.22)$$

where we have used the fact that cross-terms of the form $\langle J^{(\xi)}J^{(\phi)}\rangle$ vanish. Here, we have that the only possible scaling to have unit norm is

$$J^{(\xi)} \sim \mathcal{N}(0, \frac{1}{\sqrt{2N}}), \quad J^{(\phi)} \sim \mathcal{N}(0, \frac{1}{\sqrt{2M}}), \quad J^{(\eta)} \sim \mathcal{N}(0, \frac{1}{\sqrt{2M}}) \quad (3.23)$$

Finally, we can repeat this for $\mathcal{M} = 3$, where we obtain

$$\langle||\mathbf{h}||_2\rangle = N\langle J^{(\xi)2}\rangle + M\langle J^{(\phi)2}\rangle + M\langle J^{(\eta)2}\rangle \quad (3.24)$$

From which the scaling then has to be

$$J^{(\xi)} \sim \mathcal{N}(0, \frac{1}{\sqrt{3N}}), \quad J^{(\phi)} \sim \mathcal{N}(0, \frac{1}{\sqrt{3M}}), \quad J^{(\eta)} \sim \mathcal{N}(0, \frac{1}{\sqrt{3M}}) \quad (3.25)$$

And so in general, we conclude that the appropriate scaling of weights is

$$J \sim \mathcal{N}(0, \frac{1}{\sqrt{\mathcal{M}N_{mod}}}) \quad (3.26)$$

where $N_{mod}$ refers to the number of neurons on the input layer for the corresponding modality.

### 3.6.2 Noise model and generalized cluster ensemble

It is straightforward to derive Eq. (3.14). Note that $\Delta S$ is the probability that the net block vector $\hat{\bar{\xi}}$ is distinct from $\bar{\xi}$, given that within each sub-block the probability of differing from the other members of the cluster by $\frac{\Delta\xi}{2}$ and $\frac{\Delta\phi}{2}$ respectively for the task-relevant and contextual modalities. Thus we simply add the probabilities on each sub-block to get Eq. (3.14) in the main text. The interesting result here is the dependence of $\Delta S$ of $\delta$, in particular $\delta$ controls both the sensitivity of the net cluster size with respect to the input, with a smaller $\delta$ leading to a larger sensitivity (see Figure (3.8), top right), but also with respect to $\Delta\phi$, with a larger $\delta$ leading to a greater sensitivity with respect to $\Delta\phi$ (Figure(3.8), top left).

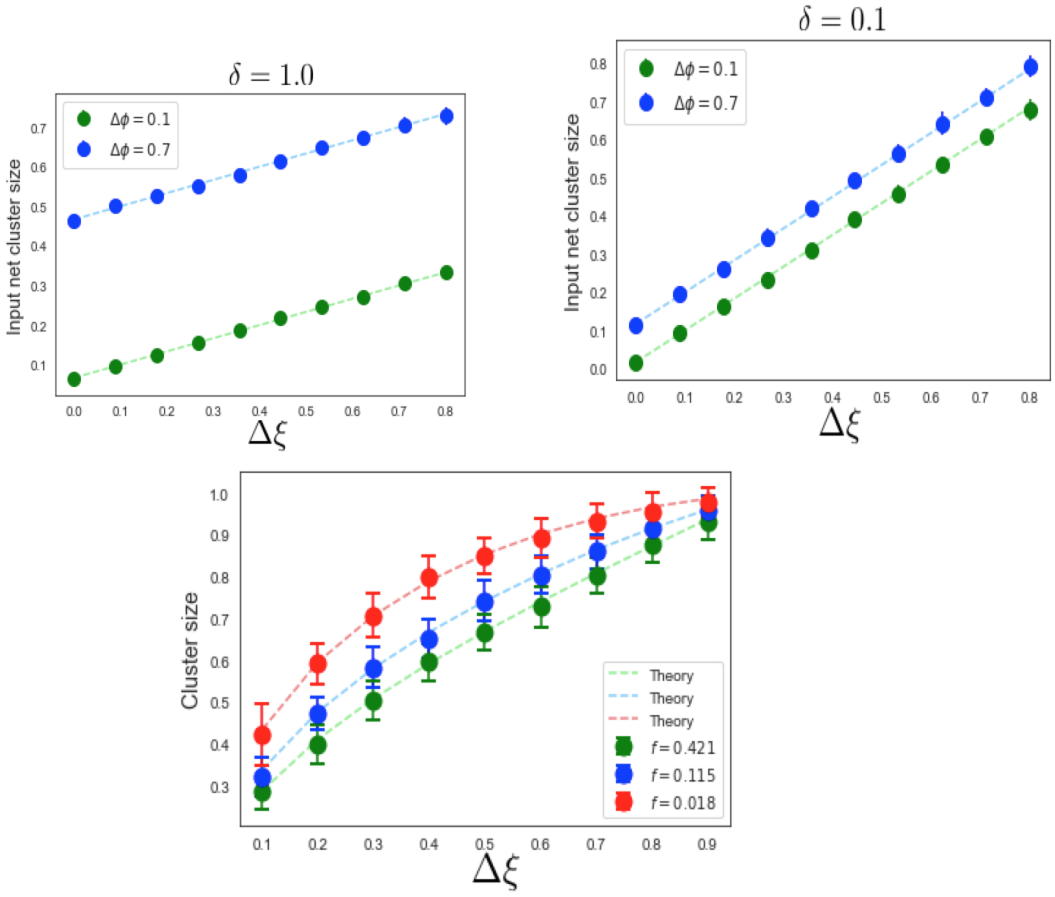


Figure 3.8: *Top:* Simulation results to verify Eq. (3.14), where on the left we compute $\Delta S$ with a small $\delta = 0.1$, and on the right for $\delta = 1.0$, both for different values of $\Delta\phi$. The theoretical lines of Eq. (3.14) are shown in the dotted lines. *Bottom:* Plot of the mixed layer cluster size $\Delta m$ as a function of $\Delta\xi$ for different $f$, with the theoretical lines (dotted) given by Eq. (3.29). Simulation details are $N = P = 100$, and for the bottom $N_c = 1200$.

### 3.6.3 Effect of sparsity of capacity

An important component of Marr's expansion recoding hypothesis is the utility of the sparseness of the cortical representation in implementing associative learning. For the unimodal case, a more detailed study [35], [39] shows that sparseness of the representation by itself indeed worsens the capacity for associative learning; and requires a concomitant sparseness of labels (or valences). Here, we show simulation results for the (trimodal) CEMS that corroborates this. We simulate a CEMS network with $N_m = 3$, $K = 2$, $N = M = 100$ and $\mathcal{R} = 1$ [8], leading to a $\beta_c = 0.5$ In Figure (3.9) below, we run the linear program check with increasing threshold (from left to right). For a coding level of $f \approx 0.04$, the CEMS can effectively no longer implement any associations. Note that, importantly, the labels in this case are still left random.

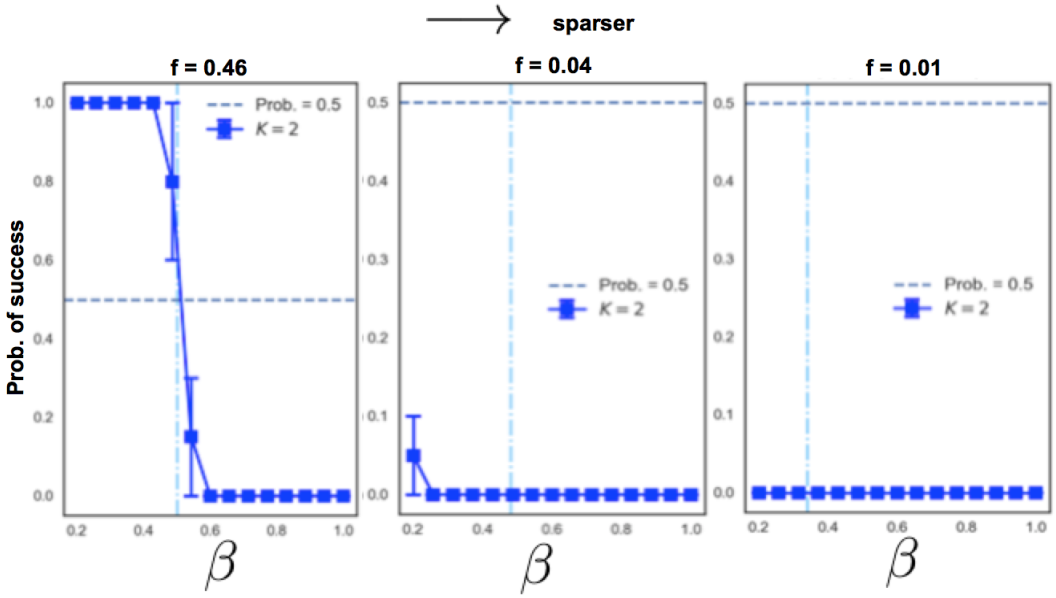


Figure 3.9: The effect of sparsity of the mixed layer on the capacity is numerically checked. The coding level is shown at the top of each figure, and $f$ decreases from left to right. The results indicate that sparseness by itself worsens associative capacity at the mixed layer, contradicting early theories of Marr-Albus. Simulation details are given in the text.

### 3.6.4 Derivation of mixed layer cluster size

#### Unimodal result

We recap here results that have been shown elsewhere in [3], which we will then generalize to the CEMS model. First, we solve the unimodal case, and note that from (3.3), we need

[8] The particular value of $\mathcal{R}$ is unimportant, we only want to show that for a given $\mathcal{R}$, capacity worsens with sparsity.

to evaluate

$$\frac{1}{f(1-f)}\,\mathrm{Prob}(h_i > T, \hat{h}_i < T) \tag{3.27}$$

(Note that the symmetry of $\mathrm{Prob}(h_i > T, \hat{h}_i < T) = \mathrm{Prob}(h_i < T, \hat{h}_i > T)$ has been exploited.) To evaluate this, we need the covariance structure of the $2 \times 2$ matrix $h_i^\mu \hat{h}_i^\mu$, which is given by

$$h\hat{h} = \begin{pmatrix} 1 & 1-\Delta\xi \\ 1-\Delta\xi & 1 \end{pmatrix} \tag{3.28}$$

and

$$\begin{aligned} \Delta m &= \frac{1}{f(1-f)}\,\mathrm{Prob}(h_i > T, \hat{h}_i < T) \\ &= \frac{1}{f(1-f)} \int_T^\infty \int_{-\infty}^T \frac{dh_1 dh_2}{(2\pi)\sqrt{\Delta\xi(2-\Delta\xi)}} \exp\Big[-\frac{1}{2\Delta\xi(2-\Delta\xi)}\Big(h_1^2 + h_2^2 - 2(1-\Delta\xi)h_1 h_2\Big)\Big] \\ &= \frac{1}{f(1-f)} \int_T^\infty \frac{e^{-h_1^2/2}}{\sqrt{2\pi}} H\Big(\frac{(1-\Delta\xi)h_1 - T}{\sqrt{\Delta\xi(2-\Delta\xi)}}\Big) dh_1 \\ &= \mathcal{G}(T, \Delta\xi) \end{aligned} \tag{3.29}$$

which is a monotonically increasing function of both $\Delta\xi$ and the sparsity $f = H(T)$. This is shown on Figure (3.8), bottom. Note that as $f$ is decreased the mixed layer cluster size is amplified.

### Cluster size of CEMS

We now generalize the above result to the CEMS model. Again, without loss of generality, we explicitly state the results for $N_m = 3$ for each $\mathcal{M}$, after which we will state the most general result. Following the line of argument above, we need to compute the covariance matrix as in (3.28). The diagonal terms are simply equal to unity by the appropriate scaling in the weights, as argued in equations (3.18) to (3.26), thus we are left with computing the off-diagonal elements $\langle \mathbf{h}^{(\mu,\gamma,\rho)} \hat{\mathbf{h}}^{(\mu,\gamma,\rho)} \rangle$. Let us start with $\mathcal{M} = 1$, where

$$\begin{aligned} \langle h_i^{(\mu,\gamma,\rho)} h_i^{(\mu,\gamma,\rho)} \rangle_{\mathcal{M}=1} &= \frac{1}{3}\Big( \langle \sum_{j,j'}^N J_{ij}^{(\xi)} J_{ij'}^{(\xi)} \rangle \langle \xi_j \hat{\xi}_{j'} \rangle + \langle \sum_{k,k'}^M J_{ik}^{(\phi)} J_{ik'}^{(\phi)} \rangle \langle \phi_k \hat{\phi}_{k'} \rangle + \langle \sum_{l,l'}^M J_{il}^{(\eta)} J_{il'}^{(\eta)} \rangle \langle \eta_l \hat{\eta}_{l'} \rangle \Big) \\ &= \frac{1}{3}\Big(1 - \Delta\xi + 2(1-\Delta\phi)\Big) = 1 - \frac{1}{3}(\Delta\xi + 2\Delta\phi) \end{aligned} \tag{3.30}$$

Likewise, for $\mathcal{M} = 2$,

$$\langle h_i^{(\mu,\gamma,\rho)} h_i^{(\mu,\gamma,\rho)} \rangle_{\mathcal{M}=2} = \frac{1}{2}\Big( \langle \sum_{j,j'}^{N} J_{ij}^{(\xi)} J_{ij'}^{(\xi)} \rangle \langle \xi_j \hat{\xi}_{j'} \rangle + \langle \sum_{k,k'}^{M} J_{ik}^{(\phi)} J_{ik'}^{(\phi)} \rangle \langle \phi_k \hat{\phi}_{k'} \rangle \Big) \tag{3.31}$$

$$+ \frac{1}{2}\Big( \langle \sum_{j,j'}^{N} J_{ij}^{(\xi)} J_{ij'}^{(\xi)} \rangle \langle \xi_j \hat{\xi}_{j'} \rangle + \langle \sum_{l,l'}^{M} J_{il}^{(\eta)} J_{il'}^{(\eta)} \rangle \langle \eta_l \hat{\eta}_{l'} \rangle \Big) \tag{3.32}$$

$$= \frac{1}{2}\Big( \frac{1}{2}(1 - \Delta\xi) + \frac{1}{2}(1 - \Delta\phi) \Big) \times 2 = 1 - \frac{1}{2}(\Delta\xi + \Delta\phi) \tag{3.33}$$

and for $\mathcal{M} = 3$

$$\langle h_i^{(\mu,\gamma,\rho)} h_i^{(\mu,\gamma,\rho)} \rangle_{\mathcal{M}=3} = \langle \sum_{j,j'}^{N} J_{ij}^{(\xi)} J_{ij'}^{(\xi)} \rangle \langle \xi_j \hat{\xi}_{j'} \rangle + \langle \sum_{k,k'}^{M} J_{ik}^{(\phi)} J_{ik'}^{(\phi)} \rangle \langle \phi_k \hat{\phi}_{k'} \rangle + \langle \sum_{l,l'}^{M} J_{il}^{(\phi)} J_{il'}^{(\phi)} \rangle \langle \phi_l \hat{\phi}_{l'} \rangle \tag{3.34}$$

$$= 1 - \frac{1}{3}(\Delta\xi + 2\Delta\phi) \tag{3.35}$$

Note that in general, it is not too hard to show that the partitioning scheme of CEMS implies that for a genaral $N_m$, the $\Delta S_{eff}$ as defined in the main text is given by

$$\Delta S_{eff}(\mathcal{M} = N_m) = \Delta S_{eff}(\mathcal{M} = 1) = \Delta S(\delta = 1) \tag{3.36}$$

$$\Delta S_{eff}(1 < \mathcal{M} < N_m) = \Delta S(\mathcal{M}, \delta) \tag{3.37}$$

# Chapter 4

# Hebbian readout of CEMS

In this Chapter, we will present the results on the generalization abilities of the CEMS model, assuming a supervised Hebbian rule at the readout layer. First, we discuss the dimensionality of the mixed layer representation, and present novel results on its sole dependence on the values of order parameters that capture the notion of disentanglement of the representation. Then, we will briefly recap the main results of [3], highlighting the role of the dimensionality and in particular that of the excess overlaps. We will then discuss the subtle role of non-linearity choice on learning, particularly that of the *sign* non-linearity used in [5] vs. the *Heaviside* non-linearity used in [3], highlighting the distinctive role of the excess overlaps and interference terms for both cases. Then, we will generalize the notion of excess overlaps for an arbitrary mixing degree, before discussing its implication on the generalization abilities of the CEMS. We will highlight the role of an effect we call *structured multi-modal overlaps*, a basic statistical property of correlated inputs of independent modalities on the mixed layer - and we will tease apart its distinct behavior from the excess overlaps in the unimodal case.

## 4.1 Dimensionality of the mixed layer

In the previous sections, we utilized the ShK result to study the effect of the rank of the mixed layer on the capacity to implement input-output mappings. The rank of the data matrix at the mixed layer is given by the number of non-zero eigenvalues of the activities at this layer, after the inputs are propagated through a layer of random weights. Here, we study another complementary measure of dimensionality, namely the *participation ratio* [33], [57], $\mathcal{D}$ ($\mathcal{D}$ for dimensionality), which is given by

$$\mathcal{D} = \frac{\mathrm{Tr}(\mathbf{C})^2}{\mathrm{Tr}(\mathbf{C}^2)} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \tag{4.1}$$

where $\mathbf{C} = \mathbf{m}\mathbf{m}^T \in \mathcal{R}^{N_c \times N_c}$, the covariance of neural responses at the mixed cortical layer. As in the previous chapter (see Eq. (3.1)), we assume implicitly that $\mathbf{m}$ is ap-

propriately re-scaled such that the mean is subtracted [1]. If the neural activities at the cortical layer were maximally random, we would expect all the eigenvalues to be equally likely and hence $\mathcal{D} \to N_c$ - this is shown in Fig. (4.1), top left [2]. However, as arbitrary thresholding is performed on random input data, the behaviour of (4.1) is much less obvious (see Figure (4.1), top right).

To study the behaviour of $\mathcal{D}$ in this regime, we define the following order parameters:

$$q^2 := m_i^2 \tag{4.2}$$

$$q^4 := m_i^4 \tag{4.3}$$

$$\mathcal{I}_2 = \mathcal{I}^{\mu\nu} := m_i^\mu m_i^\nu \tag{4.4}$$

$$\mathcal{I}_4 = \mathcal{I}_{ij}^{\mu\nu} := m_i^\mu m_j^\mu m_i^\nu m_j^\nu \tag{4.5}$$

As we will show, statistical averages of these order parameters are intimately tied to the dimensionality of mixed layer representations and the readout accuracy for a Hebbian choice of readout weights. The first order parameter, $q^2$, defines the squared norm of the signal at the mixed layer outputs; likewise the second defines the quartic norm. We call the third and fourth order parameters the *interference* terms [3]. The relevant quantities will turn out to be statistical averages, $\langle\ \rangle$ of these with respect to *both* neurons at the mixed layer, and also the random input statistics. Intuitively, as $\langle\mathcal{I}\rangle$ increases, the representation at the cortical layer is maximally *entangled*, since there is more cross-talk between random cortical states originating from different input patterns, whereas as $\langle\mathcal{I}\rangle$ decreases, these responses are said to be maximally *disentangled*. The range of these order parameters will be bounded by the respective range of their activation functions - for instance for $m \in \{0, 1\}$ all the order parameters in equations (4.2) - (4.5) will also lie in that range.

It is not too hard to show that in the limit of large systems size, $N_c, P \to \infty$, (4.1) is well approximated by the following expression:

$$\mathcal{D} \approx \frac{1}{\frac{1}{N_c P}\frac{\langle q^4\rangle}{\langle q^2\rangle^2} + \frac{1}{N_c} + \frac{1}{P} + \frac{\langle\mathcal{I}_4\rangle}{\langle q^2\rangle^2}} \tag{4.6}$$

We hence wish to understand the behaviour of $\mathcal{D}$ as a function of these order parameters. In the limit where there is no interference contribution at the mixed layer, and for mixed layers with statistics with no skewness, i.e $\langle q^4\rangle = \langle q^2\rangle^2$ - this is the limit of good old

[1] In Chapter 5 when discussing the threshold linear model, we will relax this assumption.
[2] Strictly speaking, this also requires $P >> N_c$, see Eq. (4.6).
[3] Borrowing terminology from [25].

random data clouds - one can show that $\mathcal{D} \to N_c$, if $P >> N_c$.

It is interesting as well to study the behaviour of $\mathcal{D}$ as the interference term, or the re-scaled $\frac{\langle \mathcal{I}_4 \rangle}{\langle q^2 \rangle^2}$, grows - i.e its asymptotics. In this limit the first term cannot be necessarily neglected, since the ratio of $\frac{\langle q^4 \rangle}{\langle q^2 \rangle^2}$ is not necessarily small compared to $N_c P$; nevertheless for most cases considered in this thesis this term can be ignored (though see our discussion on the ReLU activation function in Chapter 5). The second and third terms in the denominator can be ignored in the thermodynamic limit, though in practice this would require simulations of very large systems sizes; we put them here for the sake of completion.

Let us give concrete examples of this and study $\mathcal{D}$ for two choices of non-linearities commonly used in the computational neuroscience literature, namely the *Heaviside* non-linearity, used elsewhere in this thesis, and the *sign* non-linearity [4]. For the former, we have that $\langle q^2 \rangle = f(1-f)$ and $\langle q^4 \rangle = \langle q^2 \rangle^2 = f^2(1-f)^2$, which means that the first term in the denominator of (4.6) can be ignored as long as $\langle \mathcal{I}_4 \rangle$ is not exceedingly small. For the latter, we have that all $\langle q^4 \rangle = \langle q^2 \rangle = 1$, and hence the first term can also be ignored. Overall, we have asymptotically that

$$\mathcal{D}_{\{0,1\}} \sim \frac{f^2(1-f)^2}{\langle \mathcal{I}_4 \rangle}, \quad \mathcal{D}_{\{+,-\}} \sim \frac{1}{\langle \mathcal{I}_4 \rangle} \tag{4.7}$$

which means that there is a direct inverse relationship between the dimensionality of the representation and the interference term, away form saturation. This effect is illustrated in Figure (4.1) below. Note that for the theory plots the terms $\frac{1}{N_c P}$, $\frac{1}{N_c}$,$\frac{1}{P}$ are kept such that the function is analytic as $\langle \mathcal{I}_4 \rangle \to 0$, though the approximations of (4.7) hols asymptotically as $\langle \mathcal{I}_4 \rangle$ grows. This directly captures the notion of disentanglement, since $\mathcal{D}$ is a decreasing function of the $\langle \mathcal{I}_4 \rangle$. Such scaling plots will be used throughout this thesis when discussing the dimensionality of the mixed layer representation for various different models. The form of $\langle \mathcal{I}_4 \rangle$ is analytically calculable for random input data and random weights - we discuss this later on (see also Appendix) - and so the behaviour of $\langle \mathcal{I}_4 \rangle$ is understood as a function of $T$.

[4] It is implicit here that the choice of the Heaviside non-linearity implies that the mixed layer activations are re-scaled such that the mean of $f = H(T)$ is subtracted; for the sign non-linearity the mean is zero.

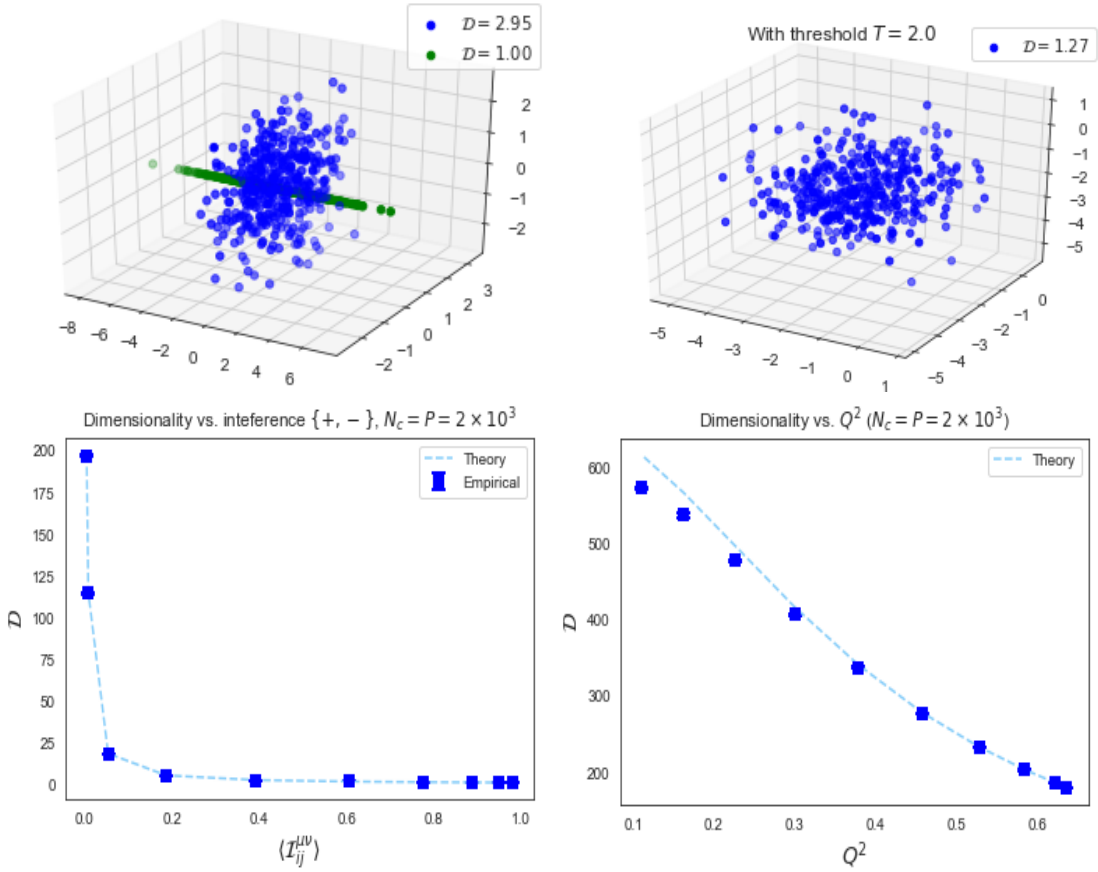


Figure 4.1: *Top:* An illustration of the dimensionality of the representation for different cases. On the left, we have plotted a random 3-dimensional data cloud ($P = 400$) in blue, which has $\mathcal{D} \approx 3$, along with a localized data cloud along one direction; this has $\mathcal{D} = 1$. On the right, we plot the data cloud after being passed through a threshold at the mixed layer; here $\mathcal{D} = 1.27$. *Bottom*: An illustration of the direct inverse relationship between the dimensionality of the representation, $\mathcal{D}$, and the interference term, for two separate choices of non-linearities at the mixed layer. Theory curves are given by (4.6), with analytical expressions for $\langle \mathcal{I}_4 \rangle$ given in the Appendix. On the right we have the $\{0, 1\}$ mixed layer activities, which we have labelled differently, with $Q^2$ defined in (4.11) below. Note that on the right the dimensionality is significantly larger, since the interference term grows inversely with the size of the input layer, as discussed in the text preceeding to (4.11). Note that in both cases, the scalings of Eq. (4.7) hold away from zero on the horizontal axis, near which the saturation effects take over.

### 4.1.1 Excess overlaps

In the previous section, we have explained the role of an interference term at the mixed layer on the dimensionality of its representation. Here, we explain an important and subtle sub-leading effect, namely the *excess overlaps*, first explained in [3]. Here, we derive explicitly the sub-leading effect, and elucidate its signature via the structure of

the co-variance matrix of inputs onto the mixed layer. As shown in Figure (4.2) (top left), the co-variance matrix of inputs onto the mixed layer for the unimodal case has a unit diagonal, as a consequence of the appropriate normalization of the weights, with the off-diagonal terms on average equal to zero but with a deviation of order $O(\frac{1}{\sqrt{N}})$. This is shown on the top right of Figure (4.2), with the bulk of the histogram of the entries centered at zero, with variance $O(\frac{1}{\sqrt{N}})$; for a derivation via the law of large numbers see the Appendix. We can write this as

$$h^{\mu}h^{\nu} = \delta^{\mu\nu} + (1 - \delta^{\mu\nu})\frac{x}{\sqrt{N}} \tag{4.8}$$

with $x \sim \mathcal{N}(0, 1)$. The effect of excess overlaps will then turn out to be the sub-leading effect due to the covariance structure in (4.8) on the nonlinear averages that depend on the covariance $h^{\mu}h^{\nu}$. For two (and three) point correlations, averaging over the randomness in the inputs and the neurons at the cortical layer lead to no sub-leading contribution, but for higher-order correlation functions, such as a four point correlation, sub-leading terms of $O(\frac{1}{N})$ appear. The effect of this on the activation on the mixed layer is illustrated in the bottom of Figure (4.2), where the projected centroids are illustrated with crosses. On average, the distance between two centroids are equal to each other, apart from $O(\frac{1}{\sqrt{N}})$ fluctuations. However, non-negligible higher order correlations tend to lead to a so-called "clustering of the clusters", shown schematically with the dotted circles. Intuitively, as will be made concrete in the proceeding parts of this Chapter, this will worsen the learning abilities of our network, since distinct clusters to be learnt might get too close to one another for a downstream neuron to successfully discriminate between them.

The exact analytical form of these excess overlaps will be presented in the next Section.

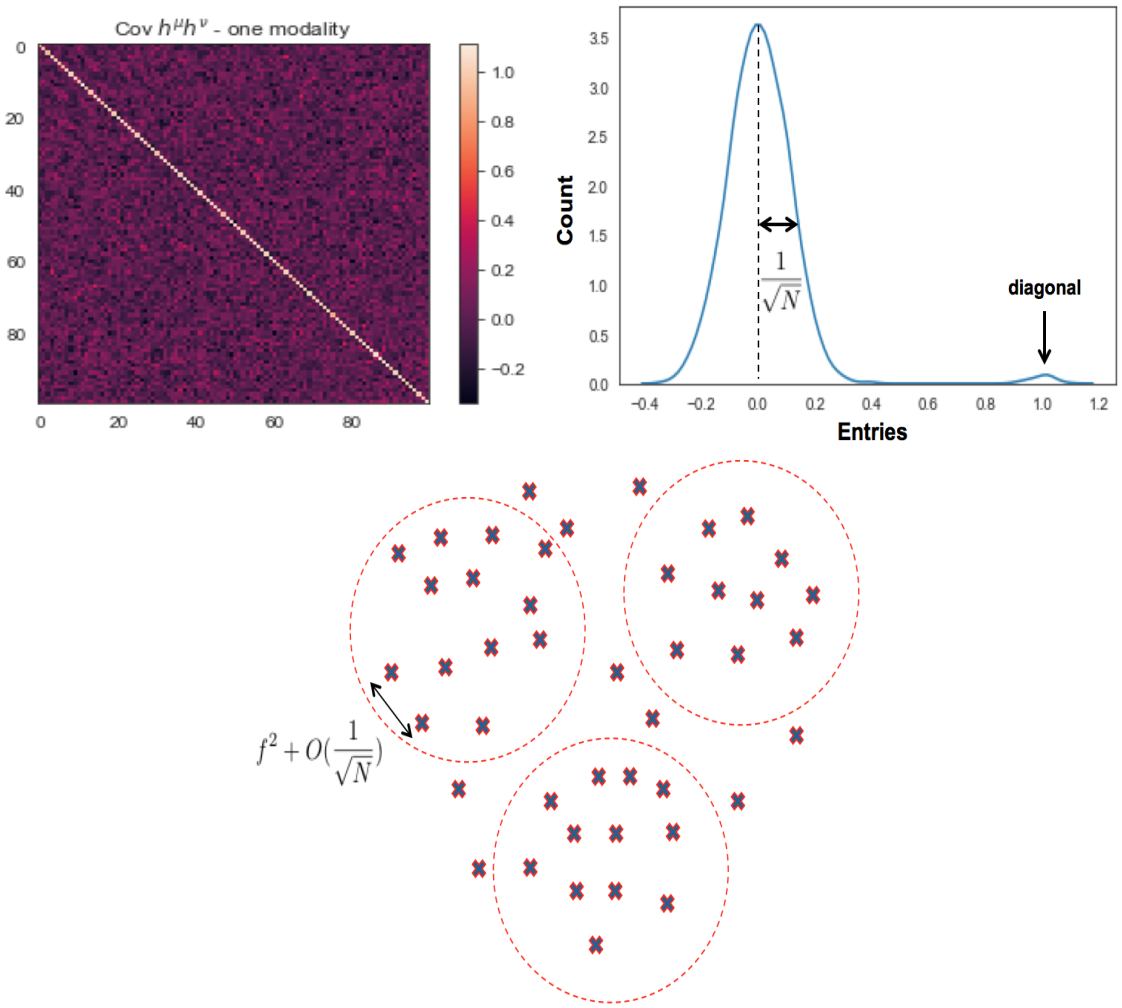


Figure 4.2: Illustration of the phenomenon of excess overlaps. *Top:* On the left we plot the covariance matrix of linear inputs onto the cortical layer, where we see that the diagonal entries are equal to one, but the off-diagonal are random variables centered at zero with a width $\frac{1}{\sqrt{N}}$. This is also illustrated on the right, where now we plot the histogram of the entries of the covariance matrix. Such plots will be used later on in this thesis as we study more complicated covariance structures. *Bottom:* A cartoon illustration of the phenomenon of excess overlaps. On average random distinct centroids projected on the mixed layer will have a distance $f^2$ from each other [5], with fluctuations of $O(\frac{1}{\sqrt{N}})$. There, however can be cases - in particular for a dense code and for small $N$ - where higher-order correlations between different centroids, leading to a so-called "clustering of clusters", illustrated with the dotted red lines.

The results in this section quantify the relationship between sparseness, decorrelation, and dimensionality at the mixed layer - for random inputs projected by random weights. This has been the subject of some theoretical interest (see [57], [17], [16]). In particular, the authors of [17] have proposed it as a unified measure to explain diverse concepts such

as sparse coding, expansion, mixing, and decorrelation. Here, we *analytically* establish such a connection - for random input statistics [6], in what we believe to be novel results establishing a relationship between these quantities. The extension to more biologically realistic models would be interesting further work - but in principle would simply entail studying the behaviour of the order parameters in (4.2) - (4.5). Later in Chapter 5, we will also precisely make this link for a threshold-linear model.

## 4.2 Readout error

We are now in a position to make explicit the relationship between the accuracy of a Hebbian readout to the noise cluster size at the mixed layer and the dimensionality of the representation. Let us consider the centers of the clusters at the layer, $\hat{\mathbf{m}}^{\mu}$ as training data and the other members of cluster, $\mathbf{m}^{\mu}$ as noisy realizations of the centroids which will be presented to the downstream readout neuron. The Hebb rule at the readout layer is given by

$$w_i^H = \sum_{\mu} \sigma^{\mu}(m_i^{\mu} - f) \tag{4.9}$$

Specifically, it can be shown ([25],[3], [44]) that the readout error, $\epsilon = H(\sqrt{SNR})$, with

$$SNR = \frac{(1-\Delta m)^2}{\frac{\alpha}{\mathcal{R}} + P\frac{\langle \mathcal{I}_4 \rangle}{f^2(1-f)^2}} \sim \frac{(1-\Delta m)^2}{\frac{\alpha}{\mathcal{R}} + \frac{P}{\mathcal{D}}} \tag{4.10}$$

[7] with $\Delta m$ given by Eq. (3.3) and explained in the previous chapter. Thus, the generalization abilities of our neural network depends on the interplay between three quantities. Firstly, and most importantly, the signal is a decreasing function of $\Delta m$, such that in the limit where $\Delta m = 1$ when the mixed layer activations are maximally noisy, we have a zero $SNR$ and thus $\epsilon = 0.5$. The second quantity is the ratio between the load $\alpha$ and the expansion ratio $\mathcal{R}$. Intuitively, this contribution decreases (hence the noise contribution is decreased) as $\mathcal{R}$ increases, recapitulating the intuitive effect that the generalization abilities of our network, very much like the storage abilities discussed in the previous Chapter, improves as a function of $\mathcal{R}$. Finally, we have the the re-scaled interference term $\frac{\langle \mathcal{I}_4 \rangle}{f^2(1-f)^2}$, which decreases the $SNR$. Note that in this term we have written explicitly the $P$ dependence, and not re-scaled it with respect to the number of neurons in the input layer, as done in the previous term. The consequence of this is important, since it means that unless $\langle \mathcal{I}_4 \rangle \sim O(\frac{1}{N})$, the interference term is effectively the sole contributor to the denominator (the noise contribution) of the SNR.

[6] The connection with mixing will be explained later on in this Chapter.

[7] In the limit of very large expansion ratio, we have $SNR = \mathcal{D}\frac{(1-\Delta m)^2}{P}$ as first reported by [57], though they were considering a setup of purely random test data.

### 4.2.1 The readout error for the unimodal model

We are now in a position to recap results first presented in [3], which we will use for comparison when we discuss the generalization abilities of the CEMS. Here, we have the important result that (see Appendix)

$$\langle \mathcal{I}_4 \rangle = \frac{1}{N} \frac{e^{-2T^2}}{(2\pi)^2} = \frac{1}{N} G^4(T) \tag{4.11}$$

where $G(T) = \int_T^\infty \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz$. Following the notation in [3], hence defining $Q^2 = \frac{e^{-2T^2}}{(2\pi)^2 f^2 (1-f)^2}$, (4.10) reads $\frac{(1-\Delta m)^2}{\frac{\alpha}{\mathcal{R}} + \alpha Q^2}$. The results are shown in Fig (4.3) below. The two important consequences of this are: (i) The interplay between the expansion ratio and excess overlaps in the noise term, and (ii) the existence of an *optimal sparseness* for the accuracy of a Hebbian readout, due to the simultaneous noise amplification and suppression of excess overlaps for sparse representations. From the Figure (4.3) below, we see that the optimal sparseness is $f_{opt} \approx 0.01$ (for our choice of expansion ratio and $\Delta\xi$), below which the effect of the noise amplification dominates and above which the amplification of the excess overlaps dominate. A rough relationship between an optimal sparseness and the expansion ratio has been previously reported in [3], and it is important to note that such as optimal sparseness only exists given the interplay between the two terms in the denominator of (4.10). For cases where there is no competing contributions to the noise term, as in the next section, such an optimal sparseness will not necessarily exist if the numerator (signal contribution to the SNR) is a monotonic function of sparseness.

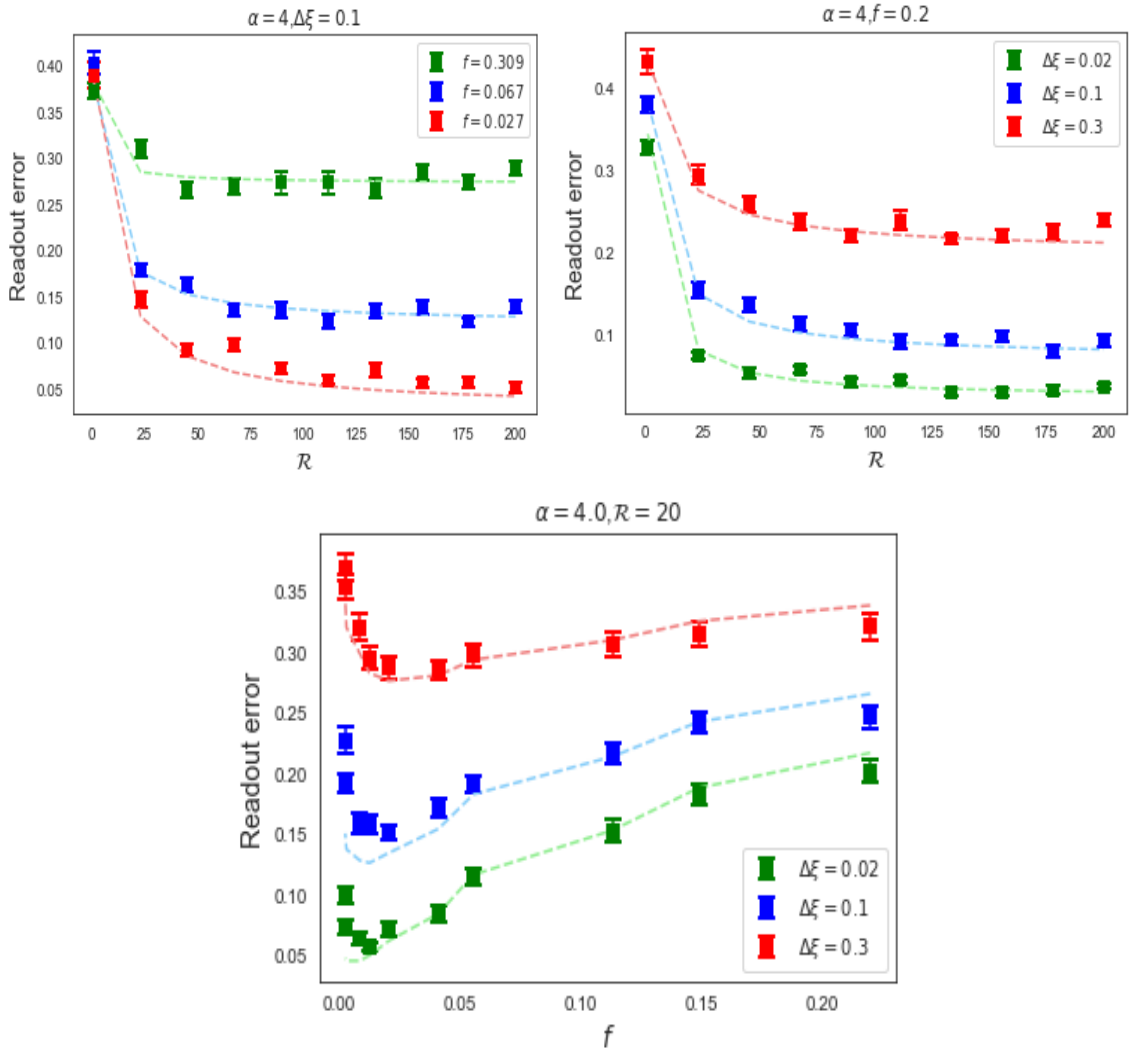


Figure 4.3: Readout error for the unimodal model. *Top:* On the left, the variation of readout error with expansion ratio $\mathcal{R}$ for different values of $f$, indicating a significant advantage of a large expansion of the generalization abilities of the network. On the right, the variation with $\mathcal{R}$ is studied for different $\Delta\xi$. *Bottom:* An optimal sparsity is discerned, for different values of $\Delta\xi$. Simulation parameters are $N = 100$, $\alpha = 4$, with the others stated in the subheadings, and theoretical predictions are presented in dashed lines.

### 4.2.2 Choice of non-linearity on readout error and optimal sparseness

The previous section demonstrated an optimal sparseness for a Hebbian readout, a consequence of having a monotonic function of sparseness in the signal contribution to the SNR, whilst having an interplay between a constant term and sparsity-dependent term in the noise contribution. Here, inspired by a famous model proposed by [5], we display the

simplest counter example, where the interference terms on the mixed layer can be shown to be of $O(1)$, thus is dominant in the noise contribution and no optimal sparseness is found. This section can be skipped on first reading. Thus, the two papers to be compared in this section are that of [5] and [3]. The first paper uses a *sign* non-linearity on the mixed layer, where the outputs are now $\{+,-\}$, whereas the second uses the Heaviside function, with the outputs $\{0,1\}$ as used in the rest of this thesis. Both papers differ in their choice of readout weights, with the former paper using a perceptron readout, whereas the latter the more biologically plausible Hebb rule, that we use in this thesis. The purpose of this subsection is to display that, with a Hebbian readout, the learning ability of the network as a function of sparsity is drastically different for both choices.

The SNR for $m_i \in \{+,-\}$ reads

$$SNR_{\{+,-\}} = \frac{(1-\Delta m)^2}{\frac{\alpha}{\mathcal{R}} + P\langle \mathcal{I}_4 \rangle} \sim \mathcal{D}\frac{(1-\Delta m)^2}{P} \tag{4.12}$$

where the asymptotics hold for a sufficiently large $\mathcal{R}$, and a sufficiently large $\langle \mathcal{I}_4 \rangle$, which means that, as opposed to the generalization abilities being dominantly determined by $\Delta m$ as in the previous section, here its dominant contribution is $\langle \mathcal{I}_4 \rangle$. We first study the form of $\langle \mathcal{I}_4 \rangle$, which can be shown (see Appendix) to be

$$\langle \mathcal{I}_{ij}^{\mu\nu} \rangle = 1 - 8\Big[H(T)^3(1-H(T) + (1-H(T))^3 H(T)\Big] + \frac{16}{N}G^4(T) \tag{4.13}$$

Note that the first term is finite (see Appendix for detailed study of its behaviour) bounded between zero and one, whereas the second term is $O(\frac{1}{N})$, in contrast to $m_i \in \{0,1\}$, where the only effect is $O(\frac{1}{N})$. This means that the second term in the denominator of (4.12) is of $O(1)$, rendering the effect of the expansion ratio somewhat negligible on learning (see Figure (4.3), top left). The second quantity of interest is $\Delta m$, which in contrast to the results in Chapter 3, is now an *decreasing* function of sparsity (see Figure (4.3), top right, and also Appendix for derivation). Finally, we can ask whether there is an optimal sparseness for the generalization abilities of this network, and the answer is no (Figure (4.3), bottom). This can be rationalized by noting that, although for any value of $\Delta\xi$ the numerator is an increasing function of sparseness, the effect on the SNR is dominated by the denominator, which grows with the number of patterns $P$. In contrast to the $\{0,1\}$ case, there is no competing term dependent on the expansion ratio, which is finite as the number of patterns learned by the network grows, meaning that, in the limit of large systems size, the behaviour of the SNR is dominated by the interference term. The interference term itself in (4.13) is an increasing function of sparsity (see Appendix for detailed plots), thus as $f$ decreases, the change in the SNR is controlled by the growing interference. The final important point to note is that the learning abilities for this choice of non-linearity *improve as the representation gets denser*, a rather contrastive trend to other theoretical results reported [3], [5]. The origin of this is clear - the aforementioned monotonic decrease of the interference term as the coding level gets denser - thus the

growth of the dimensionality - in the dominant contribution in (4.12).

This illustrates an important point about non-linearity choice on claims of optimal sparsity in computational neuroscience studies. For instance, the optimal sparseness claimed in [5] is solely due to the choice of learning rule on the readout layer, and not a general statement about optimal trade-offs employed by the brain. In this section we hope to illustrate the simplest counter-example of this, where an optimal sparseness is absent. A systematic study of effect of non-linearity choice on different learning rules (such as the Pseudo-inverse rule [70], [51]) at the readout layer would be interesting further study.

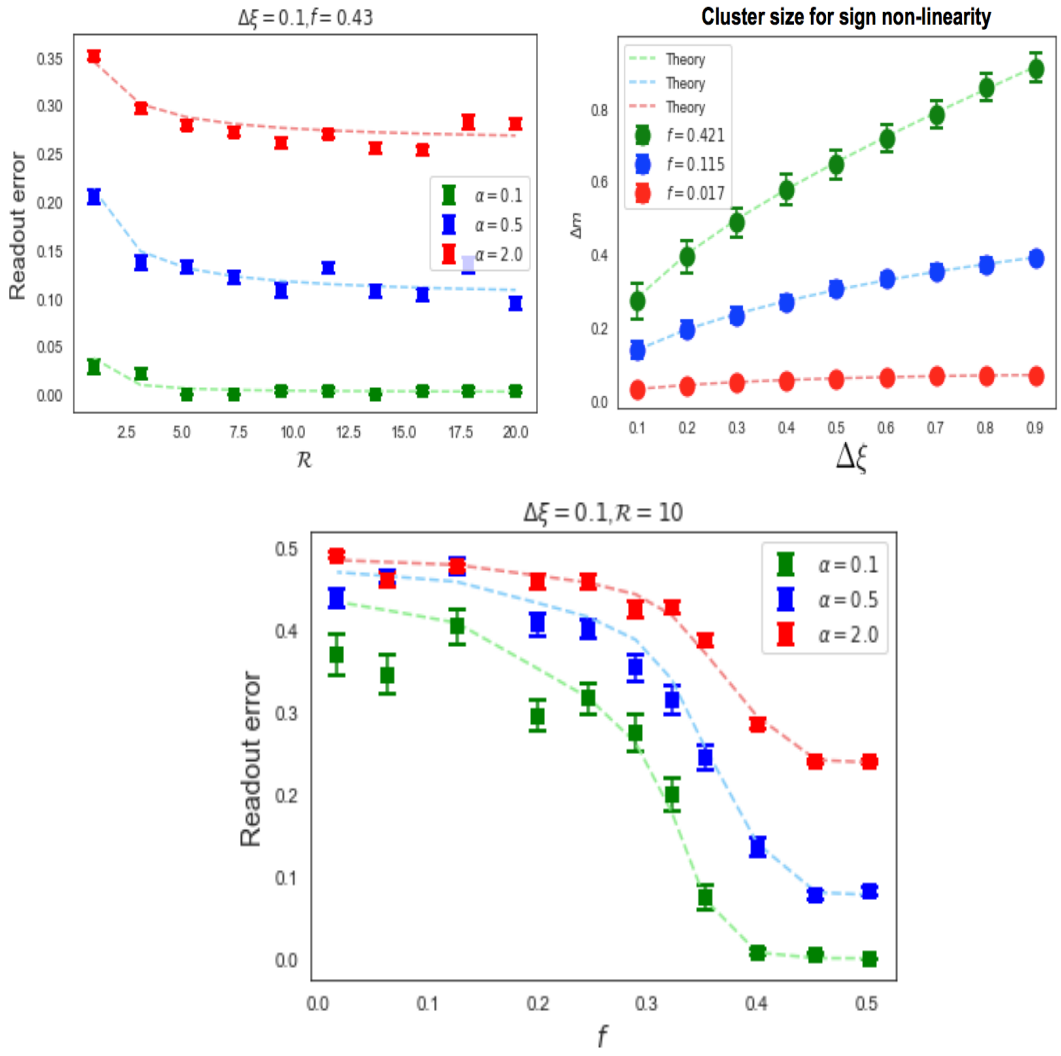


Figure 4.4: The readout error as in (4.3), but for the sign non-linearity, $m_i \in \{+, -\}$. *Top left:* The readout error as a function of $\mathcal{R}$ for different $\alpha$, showing the relative unimportance of the expansion ratio for this choice of non-linearity. *Top right:* We see that the mixed layer cluster size is a monotonically increasing function of $\Delta\xi$, but now a decreasing function of sparsity. *Bottom:* The variation of the readout error with $f$ shows that a denser representation is more advantageous to learning, whereas an optimal sparseness cannot be discerned. Simulations parameters are $N = 400$, with others given in the subheadings, and theoretical predictions are shown in dotted lines.

## 4.3 Structured multi-modal overlaps

We now will generalize the concept of excess overlaps introduced earlier, to the CEMS model with an arbitrary mixing degree $\mathcal{M}$. We should note from the outset that the

results presented in this section again solely arise from the two main features in the CEMS model - sparsity and modality-coverage conservation. Nevertheless, they shed light on an important mechanisms that determine the dimensionality and hence the learning abilities of a multi-modal cortical layer.

### 4.3.1 The various structured contributions of multi-modal overlaps

Recall that the effect of excess overlaps introduced earlier on derived from fluctuations of $O(\frac{1}{\sqrt{N}})$ on the off-diagonal entries in the covariance of the linear inputs, $h^\mu h^\nu$ [8] . Under the CEMS model, for arbitrary mixing indices, we now have in addition contributions that are not centered at zero, but have a non-zero, structured entry. This is displayed in Figure (4.5), where we plot the $h^\mu h^\nu$ (left) for $N = M = 100$, $\mathcal{R} = 10$, $P = 25$, $K = 10$, for $\mathcal{M} = 2$ (top) and $\mathcal{M} = 3$ (bottom) respectively. The striking observation here is the presence of structured elements on the off-diagonal entries, previously absent in the unimodal case (see Fig (4.2)). We can rationalize the origin of these structured peaks, which we denote $p(\mathcal{M})$ by appealing to the LLN. First, let us consider the case of $\mathcal{M} = 3$, where (see Section 3.6.1), we have

$$\langle h^{(\mu,\gamma,\rho)} h^{(\mu',\gamma',\rho')} \rangle = \frac{1}{3}\delta^{\mu\mu'} + \frac{x}{3\sqrt{N}}(1 - \delta^{\mu,\mu'}) + \frac{1}{3}\delta^{\gamma\gamma'} + \frac{y}{3\sqrt{M}}(1 - \delta^{\gamma,\gamma'}) \tag{4.14}$$

$$+\frac{1}{3}\delta^{\rho\rho'} + \frac{z}{3\sqrt{M}}(1 - \delta^{\rho,\rho'}) \tag{4.15}$$

with $x$,$y$, and $z \sim \mathcal{N}(0, 1)$, and we have $p(3) = \frac{1}{3}, \frac{2}{3}$. We can also straightforwardly show that (see Appendix) $p(3) = p(1)$. For $\mathcal{M} = 2$, we have $p(2) = \frac{1}{2}, \frac{1}{4}$. From Figure (4.5), we see that the dominant effect is that of the $p = 0$ peak, with the other components inducing a small relative contribution. The width of these residual peaks also scale as $O(\frac{1}{\sqrt{N}})$, but with specific pre-factors that we outline in detail in the Appendix.

[8] For the multi-modal case we will now implicitly assume, for the sake of simplicity, that whenever a covariance of $h$ is written, each superscript carries a 3-object index, such as $(\mu, \gamma, \rho)$.

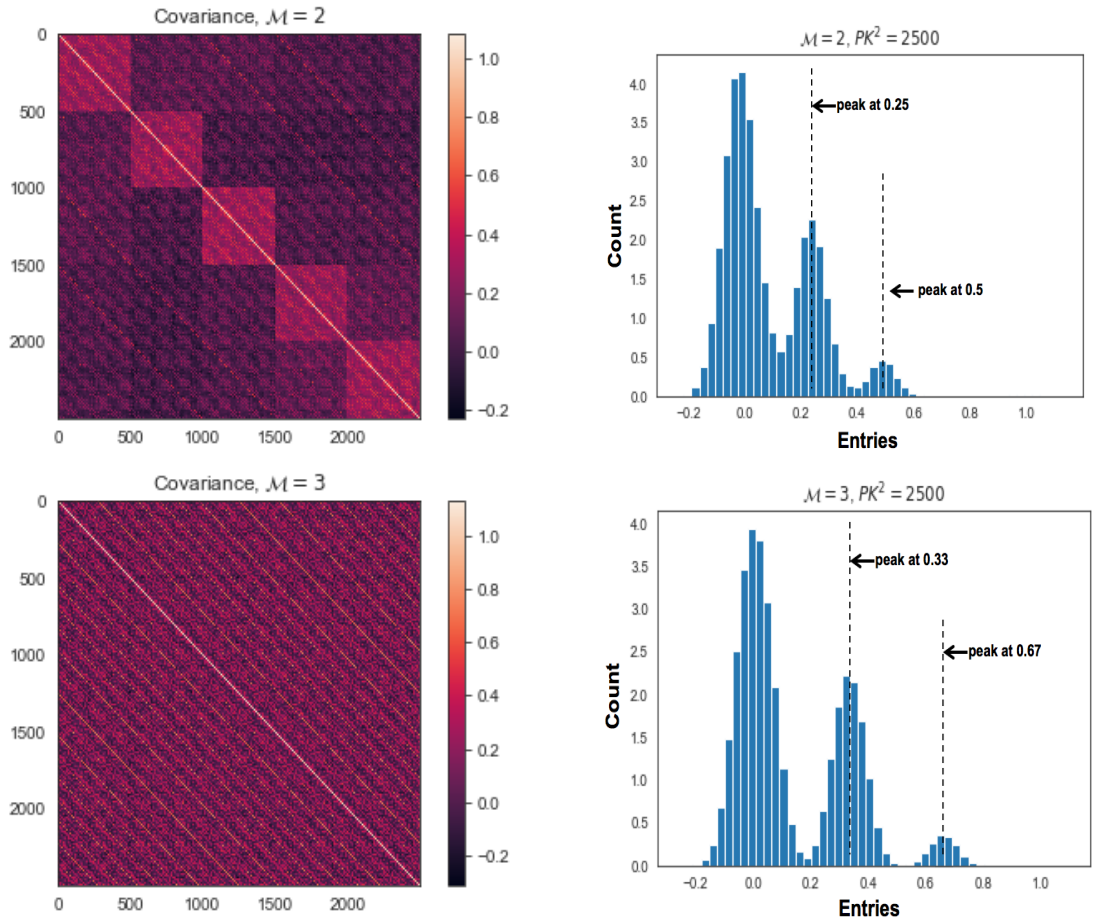


Figure 4.5: Illustration of the phenomenon of structured multi-modal overlaps. As opposed to the unimodal case, the off-diagonal terms in the input covariance matrix $h^\mu h^\nu$, have additional structure in the CEMS model. We show this for $\mathcal{M} = 2$ (top) and $\mathcal{M} = 3$ (bottom), with the corresponding histogram of entries of the covariance matrix displayed on the right. The location of peaks are determined by our weight normalization (sparsity-conserving) choice for the feed-forward random weights.

### 4.3.2 Weight of structured overlap contributions depend on absolute number of stimuli and contexts

An important observation on the behaviour of the structured overlaps are their variation with the number of composite inputs to the network, in this case $PK^2$. To see this, we simulate the same network as in Figure 4.5), but now with $P = 40$, $K = 20$. We observe that the relative weights of the residual peaks compared to that of the peak at 0 is decreased. The source of this load-dependent structured overlaps can be understood by studying the combinatorics of the structured entries in $h^\mu h^\nu$.

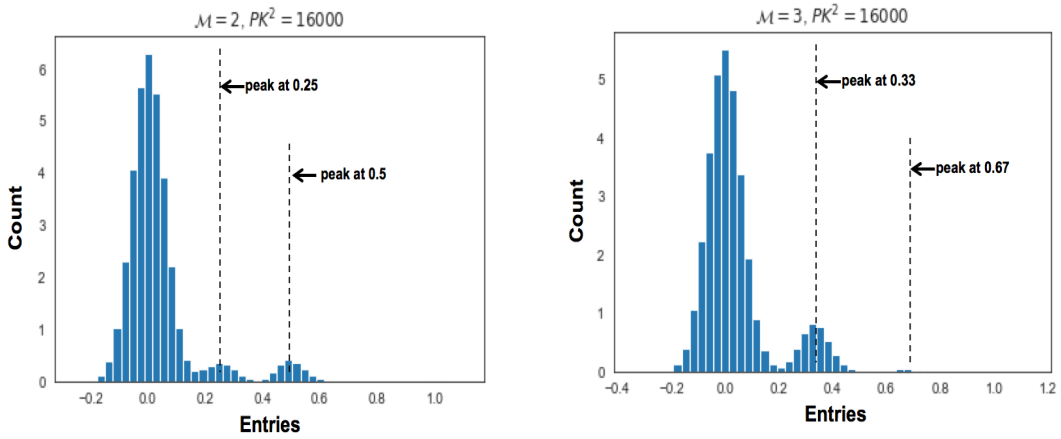


Figure 4.6: An illustration of the dependence of the relative weights of the structured peaks. The same peaks as in Fig. (4.5), but with with a larger $P$ and $K$ (compare the value of $PK^2$ in the headings of both figures), showing that the weight of the structured peaks are a decreasing function of the composite load number. The combinatorics of this are presented in the text.

The combinatorics are evaluated in Appendix. We list here the results for $\mathcal{M} = 3$, which we label according to peaks, Prob($p$). They read

$$\text{Prob}(p = 0) = (1 - \frac{1}{P})(1 - \frac{1}{K})^2 \tag{4.16}$$

$$\text{Prob}_I(p = 1/3) = (1 - \frac{1}{P})(\frac{1}{K})(1 - \frac{1}{K}) \tag{4.17}$$

$$\text{Prob}_{II}(p = 1/3) = (1 - \frac{1}{P})(\frac{1}{K})^2 \tag{4.18}$$

$$\text{Prob}_I(p = 2/3) = (1 - \frac{1}{P})(\frac{1}{K})^2 \tag{4.19}$$

$$\text{Prob}_{II}(p = 2/3) = (1 - \frac{1}{K})(\frac{1}{PK}) \tag{4.20}$$

$$\text{Prob}(p = 1) = \frac{1}{PK^2} \tag{4.21}$$

where the subscripts $I$ and $II$ denote the two separate possible ways to give the same peak For instance, the peak $p = 1/3$ can be formed (c.f. Eq. (4.15)) either by mixing the task-relevant stimulus with one contextual modality, or by mixing within contextual modalities. These probabilities - and hence its net effect on the resultant interference term (see Appendix) - thus depend on the *absolute* values of $P$ and $K$, indicating a subtle influence on the absolute value of the number of stimuli fed to the network on the dimensionality - and hence potentially the learning abilities - of the network.

### 4.3.3 Structured overlaps lead to a reduced dimensionality of the mixed representation

The effect of the structured overlaps and the residual peaks manifests itself notably through the dimensionality of the mixed representation. To do this (again from Eq. (4.6)), we need to compute the $\langle \mathcal{I}_4 \rangle$, which now has contributions consisting of a sum of correlations of the form $\langle m_i^\mu m_i^\nu \rangle$, evaluated at the respective residual peaks, and weighted their respective probabilities. The full details of such calculations are given in Appendix.

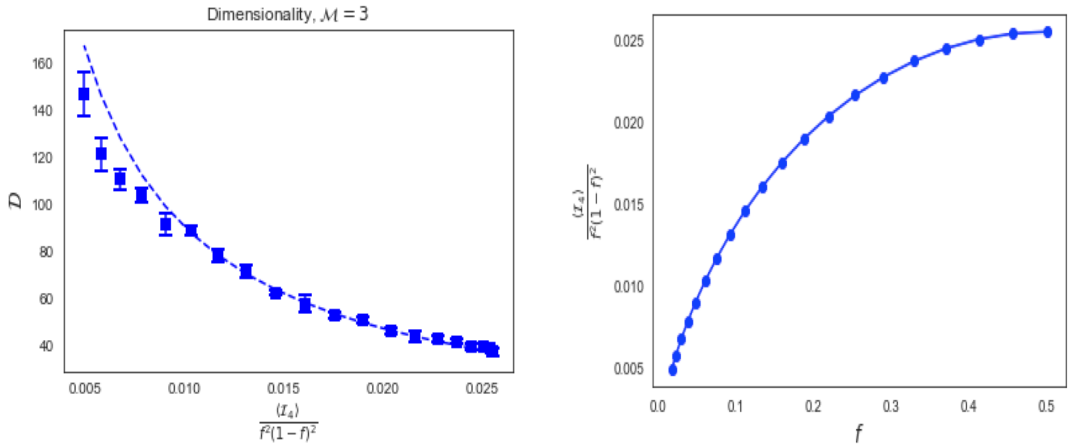


Figure 4.7: *Left:* Behaviour of $\mathcal{D}$ for CEMS, $\mathcal{M} = 3$, as a function of the re-scaled interference term. Theoretical predictions are given in dotted lines. *Right:* The re-scaled interference term is an increasing function of $f$, and is larger than the value of $Q^2/N$ in the unimodal case.

Comparing the behaviour of $\langle \mathcal{I}_4 \rangle$ for the CEMS of $\mathcal{M} = 3$ versus the unimodal case (equation for $Q^2$ in (4.11)) reveals a few interesting features - plots of these are to be found in Appendix. Firstly, the dimensionality of the representation formed in the unimodal case is significantly higher, owing to the fact that the excess overlaps $Q^2$ picks up a $\frac{1}{N}$ pre-factor in the unimodal case, and noting the relationship between the two away from saturation effects from Eq. (4.7). As we will see below, this will accordingly result in a worse readout error for the CEMS. Thus, we can conclude that the dimensionality of the representation in the multi-modal case, due to structured multi-modal overlaps, are greatly suppressed relative to the unimodal case. We mention here that this reduced dimensionality relative to the unimodal case also holds for $\mathcal{M} = 2$ (see Appendix) and $\mathcal{M} = 1$ (not shown in this thesis), although full theoretical descriptions of their interference terms are yet lacking, and left as important future work.

The effect of these structured overlaps shares some qualitative similarities with the effect of *structured projections*, studied by [3], where the input-to-cortical layer is no longer random, but set by a basic covariance rule [85]. The net effect of such a rule is that the histogram of entries per neuron has bimodal distribution, leading also to additional

modes in the covariance $h^\mu h^\nu$. There are also some similarities in that the interference term is relatively suppressed in the presence of structure, however in the multi-modal case this leads to a smaller dimensionality and hence worsens learning (see next section) - in the unimodal case learning is improved. Another key difference is the dependence of this effect on the individual number of patterns and contexts fed to the network, $(P, K)$, since it controls the relative weights of the structured peaks. In the unimodal structured case, the effect of the absolute number of patterns is irrelevant.

## 4.4 Generalization in the presence of multi-modal inputs

A central tenet in the Marr-Albus recoding hypothesis is the utility of sparseness of a representation in performing pattern separation. Some previous theoretical studies have downplayed its effect [16],[57], though none of them have studied its interplay with mixing degree, which we study here. Motivated by our analysis of the effect of the structured peaks on the interference term before, we study here in particular the role of (i) sparsity, (ii) expansion ratio, and (iii) composite load on the generalization abilities of CEMS. In particular, we will argue that the first two are relatively unimportant, whereas the latter controls learning similar to the unimodal case. These are the main results of this section, besides the broad observation that an increased mixing leads to a better generalization.

### 4.4.1 The role of sparsity and expansion ratio

In this section we display the relative *unimportance* of sparsity and the expansion ratio. We plot in Figure (4.8) the readout error as a function of $f$ for $\mathcal{M} = 3$ (top left) and $\mathcal{M} = 2$ (bottom left), for different values of $\beta$, with a fixed $K$, $\Delta\xi$, $\Delta\phi$ and $\mathcal{R}$. In general, although an optimal sparsity can be discerned, its effect is relatively unimportant on the generalization properties of the network, as the relative magnitude of change in $\epsilon$ across different values of $f$ is small. A similar effect is seen for variation of the expansion ratio at a fixed $f = 0.2$, where the relative change of $\epsilon$ across a large range of $\mathcal{R}$ is small, unlike what is seen in the unimodal case [9].

What is the source of this relative unimportance of $f$ and $\mathcal{R}$? The answer lies in the aforementioned re-scaled interference term, which we have shown to be of $O(1)$. Let us first study the effect of variations in $\mathcal{R}$. This effect was also previously seen in the unimodal sign non-linearity model studied in a previous section where having an interference

[9] One may wonder why simulations for different values of $\Delta\xi$, instead of $\beta$, were not shown, as in Fig. (4.3). We can report that these simulations exist, and that the results are qualitatively similar - a relatively small change across either $f$ or $\mathcal{R}$.

term of $O(1)$ resulted in no interplay in the denominator of the SNR between the expansion ratio dependent term and the sparsity dependent interference term, with the latter being the dominant contribution - rendering the effect of the expansion ratio negligible. The effect of variations in $f$ however, are much more subtle. Unlike the $\{+,-\}$ case, an optimal sparsity can be discerned, although the relative change in readout error as a function of $f$ is much smaller than in the unimodal case.

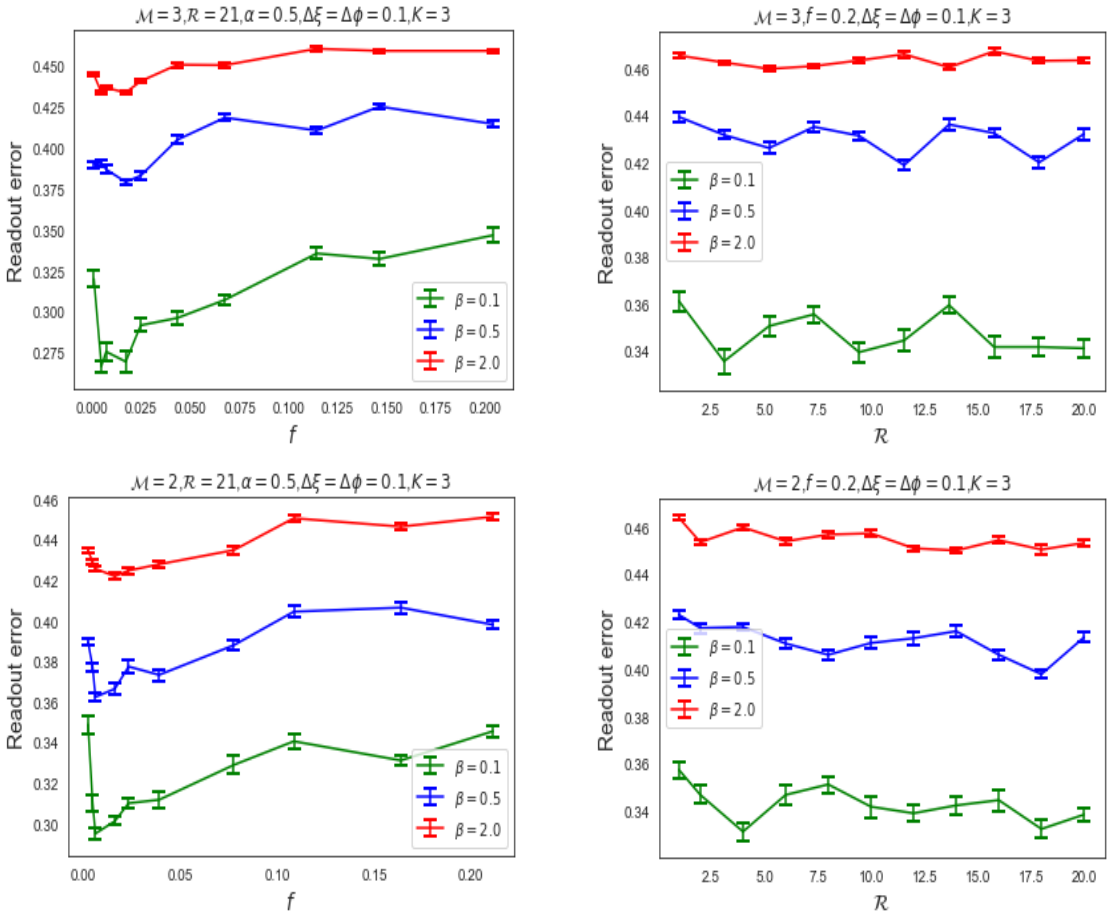


Figure 4.8: The generalization abilities of the CEMS model. The effect of structured multi-model overlaps render the effect of the sparsity somewhat unimportant in controlling the behaviour of the CEMS. $\mathcal{M}=3$ is illustrated on the top, whereas $\mathcal{M}=2$ is on the bottom. On the left column we display results of variations with respect to $f$, whereas on the right we display results of variations with respect to $\mathcal{R}$. The one parameter that does seem to control the behaviour, however, is the task-relevant load $\beta$. Simulation parameters are $N=M=100$, with the others stated in the respective headings.

It is interesting to contrast the result of CEMS to existing theoretical results [16], [11] and empirical data on sparseness in various cerebellar-like circuits, where cerebellar gran-

ule cells and the piriform cortex tend to be denser (say $f \approx 0.1$) [84] than the ultra-sparse representation seen in for instance the mushroom body and dentate gyrus [42] - all of which are thought to implement some for of expansion recoding. One possible solution to this is that cerebellar-like circuits do not have to preserve sparsity as a function of mixing - which may be a naive solution that the brain employs. In any case these results here suggest that the various simplifications used in CEMS may be worth revisiting.

### 4.4.2 The role of composite load

Given the relative unimportance of sparsity and the expansion ratio in CEMS in controlling the generalization properties of the network, it is important to ask what does. An obvious candidate here is $(P, K)$, the total composite number of stimuli presented to the network; with a slight abuse of language we call this the *composite load*. As a consequence of the relative probabilities in the structured multimodal overlaps, we have an interesting interplay between the two terms in the noise contribution due to the composite load. The first effect arises from an amplification of the $O(1)$ interference terms, which concomitantly increase the noise contribution. However, this is offset by the relative suppression of the multimodal peaks for large $(P, K)$, leading to now more dominant contributions of $O(\frac{1}{N})$ on the interference term - hence learning abilities approaching that of the unimodal case.

What is the effect of the composite load on the dimensionality itself? We see from Fig (4.9), top left, that the dimensionality increases with $K$ for a fixed $P$ - this is simply an effect of the necessary relative saturation values from Eq. (4.6) [10]. However, we also see (from Fig (4.9), top right), that the relative change in peak probabilities from Eq. (4.21) does not show up at all in the net re-scaled interference term - suggesting that there is no significant interplay from the relative suppression of $O(1)$ structured overlaps. The net effect on learning, thus, is a sizeable change in the learning abilities of the network upon variations in $\beta$ (hence $P$), and independent variations of $K$ (Figure (4.9) - bottom). The interplay of the competing effects of the composite load on learning is not seen - we only recapitulate the well known result that learning gets worse with the task-relevant load.

[10] Note that wherever we have a $P$ in Eq. (4.6) for the unimodal case, we will now replace it with $P_{eff} = PK^2$ for the CEMS.

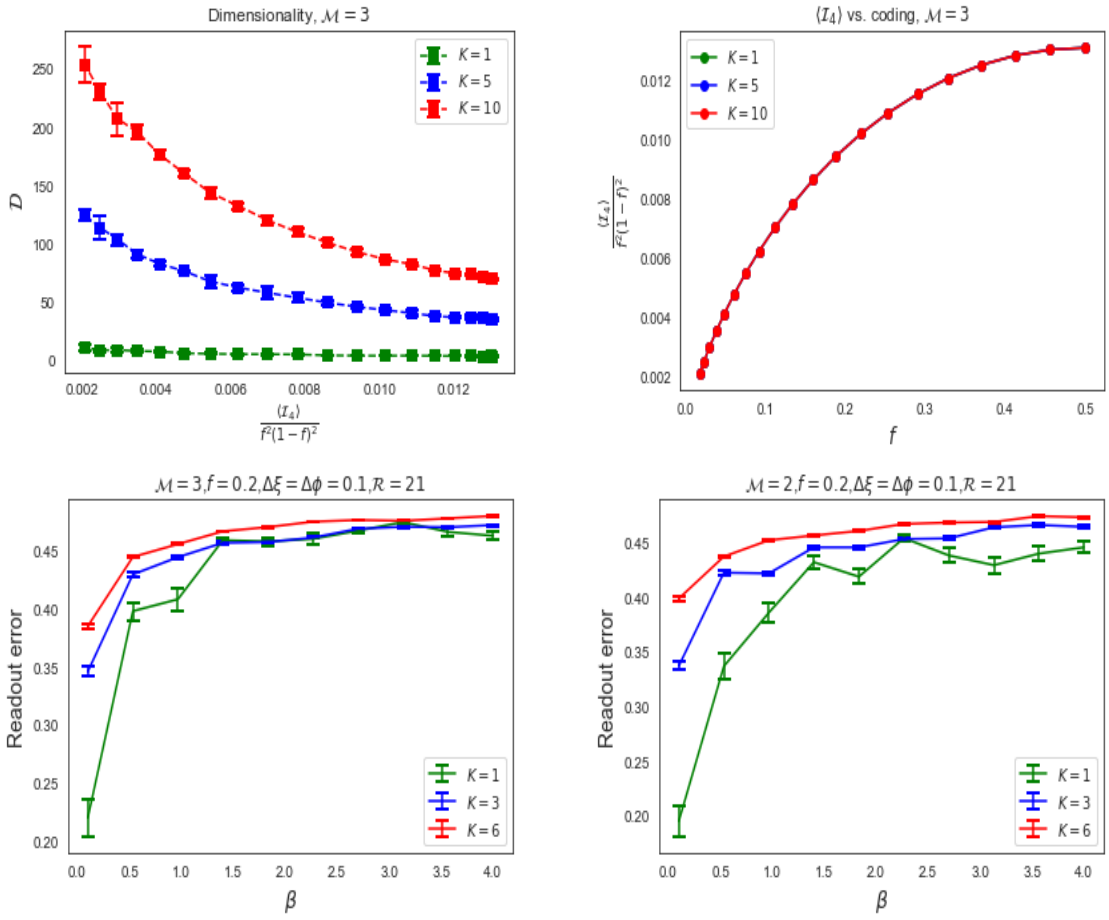


Figure 4.9: Role of the composite load on the generalization abilities of CEMS. *Top:* We see that (on the left) the dimensionality of the mixed, multi-modal representation increases with $K$; the origin of these is indeed the effect of varying the total composite number of patterns in Eq. (4.25) on $\mathcal{D}$. On the right, we verify this, as we see that the relative change in peak probabilties with $K$ from (4.21) does not effect the magnitude of the re-scaled interference term. *Bottom:* The behaviour of the readout error is studied as a function of the task-relevant load $\beta$, for different $K$, where we find, for both $\mathcal{M}=3$ (left) and $\mathcal{M}=2$ (right), it significantly controls learning in CEMS. System size is the same as in Fig. (4.8).

## 4.5 Discussion

In this Chapter we have uncovered a fundamental problem encountered by multi-modal expansive circuits in the brain that mix different sources of information - the phenomenon of structured multi-modal overlaps. The consequences of these are mainly a decrease on the dimensionality of the mixed representation compared to the unimodal case. Moreover, the $O(1)$ contributions to the interference term from these overlaps resulted in a relative unimportance of the expansion ratio and sparsity on the generalization abilities

of CEMS - in stark contrast to empirically known sparsity levels in various pattern separating cerebellar-like circuits. The statistics of these structured overlaps are rich, however, and displayed a interesting behaviour dependent on the absolute values of $P$ and $K$, previously unimportant in controlling the unimodal excess overlaps. However, such an effect does not show up in the study of the network's generalization abilities - we find the interference term in our simulations to be largely unaffected by variations in $(P, K)$, whereas the learning abilities in general show a marked decrease with either $P$ or $K$, similar to what is seen in the unimodal case.

The main result from this Chapter can thus be said to be the finding that the learning abilities of CEMS are realtively unchanged across variations in sparseness of the mixed layer and the expansion ratio. Such a result can be interpreted two ways. The first could be the conclusion that random mixing not only worsens learning in expansive feed-forward architectures relative to the unimodal case, but that it also worsens the ability of the network to improve its generalization abilities by either preserving the number of neurons that fire (maintaining a sparseness), or by arbitrarily increasing the dimensionality of the input space (increasing expansion ratio). This could in principle be empirically tested in behavioural paradigms in circuits that are thought to exhibit such random mixing, such as in olfaction. The second could be the conclusion that the CEMS model itself is an over-simplification - and that more biological realism is needed to capture the effect of sparsity and the expansion ratio on learning.

## 4.6 Appendix

### 4.6.1 Dimensionality of the mixed layer - derivation

The participation ratio is defined as in (4.1), which explicitly reads

$$\mathcal{D} = \frac{(\sum_i C_{ii})^2}{\sum_{i,j} C_{ij}^2} \tag{4.22}$$

Noting that the entries of the covariance matrix read $C_{ij} = \sum_\mu m_i^\mu m_j^\mu$, we have that the numerator simply reads $NP\langle (q_i)^2 \rangle$ on average. The denominator is instead

$$\sum_{i,j} (\sum_\mu m_i^\mu m_j^\nu)^2 = P\Big[N_c \langle q_i^4 \rangle + N_c(N_c-1)\langle q_i^2 \rangle^2\Big] + P(P-1)\Big[N_c \langle q_i^2 \rangle^2 + N_c(N_c-1)\langle \mathcal{I}_{ij}^{\mu\nu} \rangle\Big] \tag{4.23}$$

where we have first separated the sum in $\mu, \nu$ and then in $i, j$ for each. The result then gives us

$$\mathcal{D} = \frac{1}{\frac{1}{N_c P}\frac{\langle q^4 \rangle}{\langle q^2 \rangle^2} + \frac{1}{P}\frac{N_c(N_c-1)}{N_c^2}\langle q^2 \rangle^2 + \frac{1}{N_c}\frac{P(P-1)}{P^2}\langle q^2 \rangle^2 + \frac{P(P-1)}{P^2}\frac{N_c(N_c-1)}{N_c^2}\frac{\langle \mathcal{I}_4 \rangle}{\langle q^2 \rangle^2}} \tag{4.24}$$

In the limit where $N_c$ and $P$ are large, we have that $N_c(N_c-1) \approx N_c^2$ and likewise for $P$, and the expression in (4.24) reads

$$\mathcal{D} \approx \frac{1}{\frac{1}{N_c P}\frac{\langle q^4\rangle}{\langle q^2\rangle^2} + \frac{1}{N_c} + \frac{1}{P} + \frac{\langle \mathcal{I}_4\rangle}{\langle q^2\rangle^2}} \tag{4.25}$$

as in (4.6) in the main text.

### 4.6.2 Derivation of excess overlaps

The sub-leading terms in the correlations are a simple consequence of the law of large numbers, as we will show here. Note that the covariance matrix explicitly reads

$$h^\mu h^\nu = \sum_{j,j'} J_j J_{j'} \xi_j^\mu \xi_{j'}^\nu \tag{4.26}$$

We first compute the average of (4.26) with respect to the random inputs, yielding $\langle \xi_j \xi_{j'}\rangle = \delta_{jj'}$, and noting that $J^2 = \frac{1}{N}$, we have that [11]

$$h^\mu h^\nu = \frac{1}{N}\sum_j \langle \xi_j^\mu \xi_j^\nu\rangle \tag{4.27}$$

Now we invoke the law of large numbers (LLN) argument, which states that the sum of $N$ uncorrelated random variables converges to $\sqrt{N}x$, with $x$ being a unit Gaussian random variable, $x \sim \mathcal{N}(0,1)$. As a result we can split (4.27) into two terms: those with $\mu = \nu$, yielding a sum of *correlated* terms [12], and the case where $\mu \neq \nu$, giving us Eq. (4.8).

We wish to compute the non-linear average of $m_i^\mu m_i^\nu$ in a closed form, first with respect to the random inputs and then with respect to the neurons on the cortical layer, including sub-leading terms of $O(\frac{1}{\sqrt{N}})$. We note that the covariance matrix between the inputs read

$$h^\mu h^\nu = \begin{pmatrix} 1 & \frac{x}{\sqrt{N}} \\ \frac{x}{\sqrt{N}} & 1 \end{pmatrix} \tag{4.28}$$

where $x \sim \mathcal{N}(0,1)$. Thus the non-linear average that we need to compute reads:

$$m_i^\mu m_i^\nu = \int_{T,T}^{\infty,\infty} \frac{dh_1 dh_2}{2\pi\sqrt{1-\epsilon^2}} \exp\Big[-\frac{1}{2(1-\epsilon)}\Big(h_1^2 + h_2^2 - 2h_1 h_2 \epsilon\Big)\Big] \tag{4.29}$$

[11] Note that although the average with respect to the random input data is performed, the $\langle\ \rangle$ symbol is not used on the LHS. We will reserve this symbol for when averages over the random weight matrices are performed.

[12] which individually are just equal to one

where $\epsilon \sim O(\frac{x^2}{N})$ and is hence ignored in the following expansion. To a leading order, this will of course give $H^2(T) = f^2$. The subtle effect here is the higher-order correlations obtained by keeping terms that are $O(\frac{1}{\sqrt{N}})$, such that they give a non-trivial additional contribution when evaluating $\langle \mathcal{I}_4 \rangle$. We this expand the intergrand in $\frac{x}{\sqrt{N}}$, with the first sub-leading term reading

$$\frac{x}{\sqrt{N}} \int_{T,T}^{\infty,\infty} \frac{dh_1 dh_2}{2\pi} h_1 h_2 \exp\Big[ -\frac{1}{2}h_1^2 - \frac{1}{2}h_2^2 \Big] \tag{4.30}$$

(4.30) is trivially integrated to give $\frac{x}{\sqrt{N}} \frac{e^{-T^2}}{2\pi}$. Thus, on average (averaged over all the realizations of weights on the mixed layer), $\langle m_i^\mu m_i^\nu \rangle = f^2$. The interesting behaviour occurs when we compute higher order correlations, specifically $\mathcal{I}_4$. Upon performing the relevant calculation, we find that $\langle \mathcal{I}_4 \rangle = \langle \mathcal{I}^{\mu\nu} \rangle^2$. By the above line of argument, we find that

$$\langle \mathcal{I}_4 \rangle = f^4 + \frac{1}{N} \frac{e^{-2T^2}}{(2\pi)^2} \tag{4.31}$$

From this, it is straightforward to compute centered correlations such as $\langle (m_i^\mu - f)(m_i^\nu - f)(m_j^\mu - f)(m_j^\nu - f) \rangle$ to give $\frac{1}{N} \frac{e^{-2T^2}}{(2\pi)^2}$ as the only contribution [3].

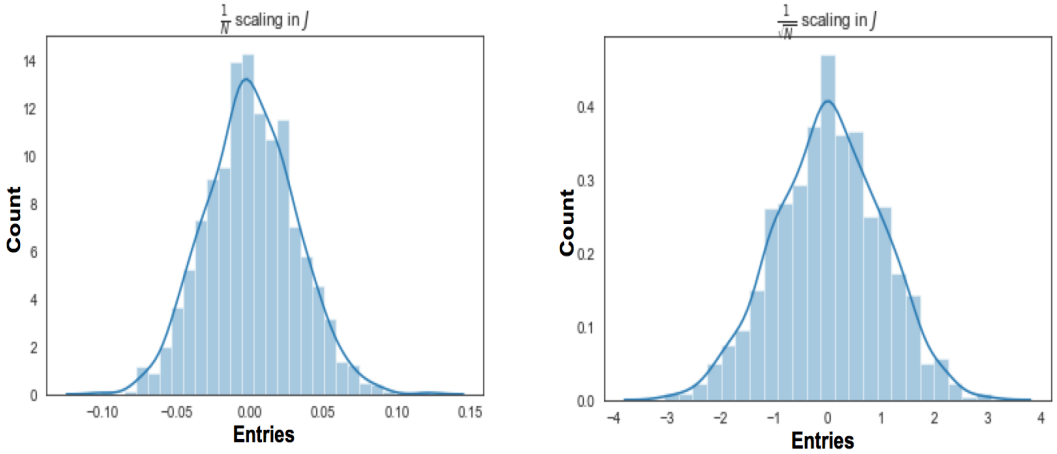


Figure 4.10: Illustration of the effect of the law of large numbers. Shown are different scaling choices for the weights on the distribution of entries of $h$. *Left*: When the weights are scaled as $J \sim \frac{1}{N}$, the mean of the entries are zero with fluctuations of $O(\frac{1}{\sqrt{N}})$, whereas for (*right*:) $J \sim \frac{1}{\sqrt{N}}$, the fluctuations are of $O(1)$.

Note that excess overlaps only appear when considering correlations between the same weight matrix projecting distinct patterns, i.e $\langle h_i^\mu h_i^\nu \rangle$, as it is when the LLN gives sub-leading terms $O(\frac{1}{\sqrt{N}})$. In other cases, such as computing $\langle h_i^\mu h_j^\mu \rangle$, it is straightforward to show that no such effect occurs.

### 4.6.3 Order parameters for a sign non-linearity

Let us now derive the above expressions for our order parameters for the sign non-linearity. This will help For this choice of $m_i \in \{-1, +1\}$, the *one-point* correlation functions $\langle q^2 \rangle$ and $\langle q^4 \rangle$ both equal 1. The two point correlation function $\langle \mathcal{I}_{ij} \rangle = \langle m_i m_j \rangle$ can be evaluated as

$$\langle \mathcal{I}_{ij} \rangle = \text{Prob}(h_i > T, h_j > T) + \text{Prob}(h_i < T, h_j < T) - 2\,\text{Prob}(h_i < T, h_j > T) \quad (4.32)$$

$$= 1 - 4\,\text{Prob}(h_i < T, h_j > T) \quad (4.33)$$

where in the last line we have simply used the normalization condition of the probabilities. Recall that for a generic $h_i$ and $h_j$ the covariance between the two is simply $\delta_{ij}$. As a result, we have that $\text{Prob}(h_i < T, h_j > T) = H(T)H(-T)$ and hence we can evaluate (4.33) to give

$$\langle \mathcal{I}_{ij} \rangle = (1 - 2H(T))^2 \quad (4.34)$$

Note that as per the previous discussion we have a term that is sub-leading of $O(\frac{1}{\sqrt{N}})$, which is zero upon averaging over all neurons on the mixed layer.

For the four point correlation $\langle \mathcal{I}_{ij}^{\mu\nu} \rangle$, we have the same statistics as above, but must now evaluate possibilities two rows further down Pascal's triangle. We have that

$$\langle \mathcal{I}_{ij}^{\mu\nu} \rangle = \text{Prob}(h_{i,j,k,l} > T) + \text{Prob}(h_{i,j,k,l} < T) - 4\,\text{Prob}(h_i > T, h_{j,k,l} < T) - 4\,\text{Prob}(h_{i,j,k} > T, h_l < T)$$
$$+ 6\,\text{Prob}(h_{i,j} > T, h_{k,l} < T)$$
$$= 1 - 8\Big[\,\text{Prob}(h_i > T, h_{j,k,l} < T) + \text{Prob}(h_{i,j,k} > T, h_l < T)\Big] \quad (4.35)$$

where we have simply indexed $i, j, k, l$ to the four distinct variables in the correlation function. This can be evaluated to give

$$\langle \mathcal{I}_4 \rangle = \langle \Big( (1 - 2H(T))^2 + 4\frac{x}{\sqrt{N}} \frac{e^{-T^2}}{2\pi} \Big)^2 \rangle \quad (4.36)$$

$$= 1 - 8\Big[ H^3(T)(1 - H(T)) + (1 - H(T))^3 H(T) \Big] + \frac{16}{N} \frac{e^{-2T^2}}{(2\pi)^2} \quad (4.37)$$

as in (4.13) in the main text. This is used for the theoretical curves in Fig (4.1) (for the dimensionality) and Fig. (4.4) (for the readout error) respectively.

### 4.6.4 Hebbian readout

#### Derivation of SNR for simple cases

Let us recap the simplest case of associations between $\{\xi^\mu, \sigma^\mu\}$, and study the typical error of a Hebbian readout [25]. The weights are defined by $w_i^H = \sum_\mu \sigma^\mu \xi^\mu$ upon learning

the above associations, with $\xi_i \in \{+,-\}^N$ and $\sigma \in \{+,-\}$. Now, let us present a test pattern from the above training set, $\xi^\nu$ along with its associated label, $\sigma^\nu$. To obtain the signal term, we need to project the output of the weights upon presentation of this test example, $g^\nu$, to the corresponding label. The signal term hence reads

$$\langle g^\nu \sigma^\nu \rangle = \sum_{i,\mu} \langle \sigma^\mu \sigma^\nu \xi_i^\mu \xi_i^\mu \rangle = N \tag{4.38}$$

since the terms with $\mu \neq \nu$ average to zero. The noise term, is given by

$$\mathrm{Var}(g^\nu \sigma^\nu) = \langle (\sum_{i,\mu} \sigma^\mu \sigma^\nu \xi_i^\mu \xi_i^\nu)^2 \rangle \tag{4.39}$$

$$= NP + N^2 P \langle \mathcal{I}_{ij}^{\mu\nu} \rangle \tag{4.40}$$

where we have decomposed the sum into terms with $i = j$ and $i \neq j$, and defined an *inteference* order parameter $\langle \mathcal{I}_{ij}^{\mu\nu} \rangle$ (see main text for definition). The SNR, which we recall is defined as

$$\mathrm{SNR} = \frac{\langle g^\nu \sigma^\nu \rangle^2}{\mathrm{Var}(g^\nu \sigma^\nu)} \tag{4.41}$$

can be written as

$$\mathrm{SNR} = \frac{1}{P[\frac{1}{N} + \langle \mathcal{I}_{ij}^{\mu\nu} \rangle]} = \frac{1}{\alpha + P \langle \mathcal{I}_{ij}^{\mu\nu} \rangle} \tag{4.42}$$

and the readout error is given by $\epsilon = H(\sqrt{SNR})$. For the simplest case we consider here, $\langle \mathcal{I}_{ij}^{\mu\nu} \rangle = 0$ (plugging in for $T = 0.5$ since that is the effective threshold for random input data), thus $\epsilon \sim H(\sqrt{1/\alpha}) \sim \sqrt{\frac{\alpha}{2\pi}}$ in the limit of large $\alpha$.

At the next level of complexity, we now consider a test pattern that is slightly distorted from the one taken form the training set, which we call $\hat{\xi}^\nu$. This distortion is performed by flipping at random each entry in $\xi^\nu$ with a probability $\frac{\Delta\xi}{2}$. As a consequence, we have that

$$\langle g^\nu \sigma^\nu \rangle = \sum_{i,\mu} \langle \sigma^\mu \sigma^\nu \xi_i^\mu \hat{\xi}_i^\mu \rangle \tag{4.43}$$

$$= N(1 - \Delta\xi) \tag{4.44}$$

and

$$\mathrm{Var}(g^\nu \sigma^\nu) = \langle (\sum_{i,\mu} \sigma^\mu \sigma^\nu \xi_i^\mu \hat{\xi}_i^\nu)^2 \rangle \tag{4.45}$$

$$= N + N(N-1)(1 - \Delta\xi)^2 + NP + N(N-1)P \langle \mathcal{I}_{ij}^{\mu\nu} \rangle \tag{4.46}$$

Note that since the test patterns also have the same probability of being turned on or silent, the statistics of $\langle \xi_i^\mu \hat{\xi}_i^\nu \xi_j^\mu \hat{\xi}_j^\nu \rangle$ and $\langle \xi_i^\mu \xi_i^\nu \xi_j^\mu \xi_j^\nu \rangle$ are identical. In the limits where $N$ is large $N(N-1) \approx N^2$. The SNR hence reads

$$\text{SNR} = \frac{(1-\Delta\xi)^2}{P(\frac{1}{N} + \langle \mathcal{I}_{ij}^{\mu\nu} \rangle)} = \frac{(1-\Delta\xi)^2}{\alpha + P\langle \mathcal{I}_{ij}^{\mu\nu} \rangle} \tag{4.47}$$

Here, we also have that $\langle \mathcal{I}_4 \rangle = 0$ since $\langle \mathcal{I}_4 \rangle$ for random uncorrelated data, and we have instead that $\epsilon = H(\frac{1-\Delta\xi}{\sqrt{\alpha}})$.

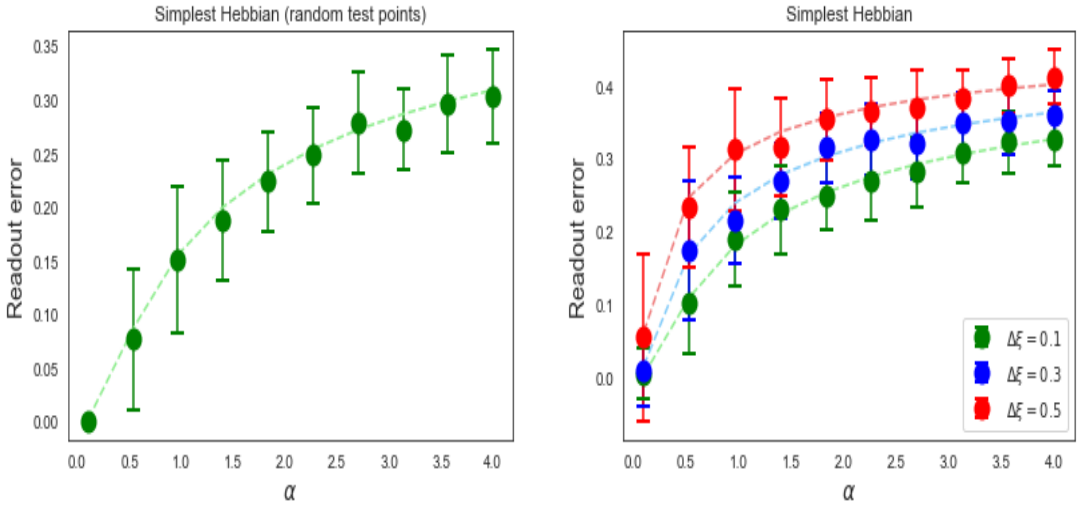


Figure 4.11: An illustration of generalization curves for simplest possible setups under a Hebb rule. On the left, we have the readout error for random training *and* test points, whereas on the right we have the generalization error for when the test data has a correlation structure $\langle \xi\hat{\xi} \rangle = 1 - \Delta\xi$ from the training data. Theoretical predictions are given in dotted lines, and simulation parameters are $N = 100$.

### Derivation of the SNR for the unimodal model

The results above generalize to the unimodal model. On the mixed layer, noting that the Hebb rule is now $w_i^H = \sum_\mu \sigma^\mu(m_i{}^\mu - f)$, we have that the SNR is given by

$$SNR = \frac{(1-\Delta m)^2}{\frac{\alpha}{\mathcal{R}} + P\frac{\langle \mathcal{I}_4 \rangle}{f^2(1-f)^2}} \tag{4.48}$$

where the normalization by $f^2(1-f)^2$ follows from the appropriate mean centering of the mixed layer outputs. Explicitly, let us first consider the signal term. Having a distance $d$ between the training and test cortical states imposes $2f - 2\langle m\hat{m} \rangle = d$, from which we

obtain (along the same lines as (4.44))

$$\langle g^\nu \sigma^\nu \rangle = N_c \langle (m^\nu - f)(\hat{m}^\nu - f) \rangle = N_c \Big[ \langle m^\nu \hat{m}^\nu \rangle - f^2 \Big] \tag{4.49}$$

$$= N_c f(1-f)(1 - \frac{d}{2f(1-f)}) = N_c f(1-f)(1-\Delta m) \tag{4.50}$$

where we have used the definition of $\Delta m$ in (3.2). The other terms in the denominator straightforwardly follow from noting that $\langle q^2 \rangle = f(1-f)$ and $\langle q^4 \rangle = f^2(1-f)^2$.

In general $\langle \mathcal{I}_4 \rangle$ is now non-zero, as correlation functions at the mixed layer involve non-linear averages of functions that depend on the statistics of the input layer. Furthermore, the effect of the second term in the denominator will dominate the noise contribution, unless $\langle \mathcal{I}_4 \rangle$ is of $O(\frac{1}{N})$, which as we argued in the main text, it is. Note that for a different choice of non-linearity and for the CEMS model, this is no longer true, as explained below (see also Chapter 5).

### 4.6.5 Effect of non-linearity choice on SNR

#### Mixed layer cluster size

Here, we have that the mixed layer cluster size $\Delta m$ as plotted in Fig. (3.8) is now a *decreasing* function of $f$, as opposed to the $\{0,1\}$ case. To evaluate $\Delta m$, we note first that the signal term in the SNR now does not require normalizations by $f$, thus we simply need to evaluate

$$\langle |m - \hat{m}| \rangle = 2 \times \Big[ \text{Prob}(h > T, \hat{h} < T) + \text{Prob}(\hat{h} > T, h < T) \Big] = 4f(1-f)\mathcal{G}(T, \Delta S) \tag{4.51}$$

#### Interference term comparison with $\{0,1\}$

A central reason for the marked difference in learning behaviour for the $\{+,-\}$ network is the difference in the behaviour of the order interference order parameters. Firstly, a sizeable difference is seen as a function of sparsity. We plot in Fig. (4.12) both interference term - on the left the excess overlaps for the $\{0,1\}$ network (see Eq. 4.8)), and on the right $\langle \mathcal{I}_4 \rangle$ for the $\{+,-\}$ network. Thus, the former is a monotonically increasing function of sparsity, whereas the opposite is true for the latter. The result of this is the difference in the difference in dimensionality of the representation at the mixed layer - in the former the dimensionality increases with sparsity, whereas in the latter it decreases with sparsity. The second difference, of importance to the readout error is the relative scaling of the contributions to the noise contribution in the SNR. The plot of the left of Fig. (4.12) will have an additional factor of $\frac{1}{N}$, thus will give an $O(1)$ term when multiplied by $P$ in Eq. (4.10). The latter, however, has no such scaling, and thus is the dominant term in the noise contribution to the SNR.

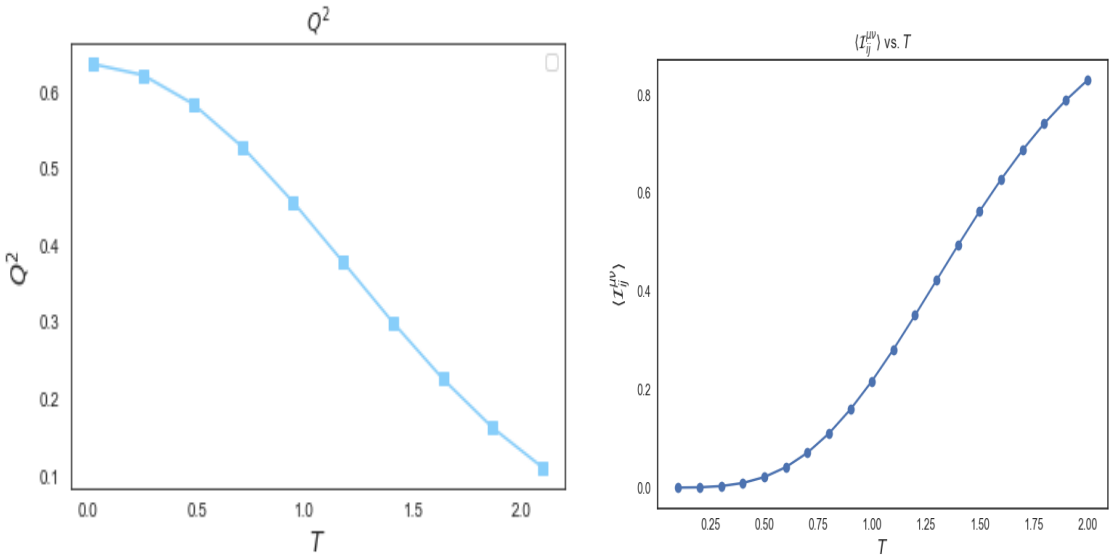


Figure 4.12: Comparison of the behaviour of the interference term for the $\{0, 1\}$ mixed layer activities and the $\{+, -\}$ activities. On the left, we display the interference term, as a function of $T$ for the $\{0, 1\}$ mixed layer activites, whereas on the right we display the same term for the $\{+, -\}$ activities. Note that $Q^2$ on the left comes in to the expression for the SNR in Eq. (4.10) with a pre-factor of $\frac{1}{N}$, whereas on the right there is no such pre-factor.

### 4.6.6 Derivation of structured multimodal overlaps

Here, we first provide the detailed derivation of the location of the structured entries in $h^\mu h^\nu$ via the LLN, in analogy to that performed above in (4.26) - (4.27). Then, we will provide the derivations for probability weight from each of the peaks in the structured overlaps, and then explain how their numerical pre-factor is calculated. Finally, we provide supplementary plots of the dimensionality of the mixed layer and the interference term for the unimodal case (of same systems size to that used in Fig. (4.7)) and $\mathcal{M} = 2$ CEMS, for comparison against the $\mathcal{M} = 3$ CEMS.

#### Location of structured peaks via LLN

We can again use the LLN to study the off-diagonal entries in $h^\mu h^\nu$, but now for the sparsity-conserving CEMS. Let us first consider the case of $\mathcal{M} = 3$. Here, we have that

(along the same lines of those in the Appendix of Chapter 3)

$$\langle h_i^{(\mu,\gamma,\rho)} h_i^{(\mu',\gamma',\rho')} \rangle = \frac{1}{3N} \sum_{j,j'} \langle J_{i,j}^{(\xi)} J_{i,j'}^{(\xi)} \rangle \langle \xi_j^{\mu} \xi_{j'}^{\mu'} \rangle + \frac{1}{3M} \sum_{k,k'} \langle J_{i,k}^{(\phi)} J_{i,k'}^{(\phi)} \rangle \langle \phi_k^{\gamma} \phi_{k'}^{\gamma'} \rangle + \tag{4.52}$$

$$\frac{1}{3M} \sum_{l,l'} \langle J_{i,l}^{(\eta)} J_{i,l'}^{(\eta)} \rangle \langle \xi_l^{\rho} \xi_{l'}^{\rho'} \rangle \tag{4.53}$$

$$= \frac{1}{3N} \Big( N\delta^{\mu,\mu'} + (1-\delta^{\mu,\mu'})\sqrt{N}x \Big) + \frac{1}{3M} \Big( M\delta^{\gamma,\gamma'} + (1-\delta^{\gamma,\gamma'})\sqrt{M}y \Big) + \tag{4.54}$$

$$\frac{1}{3M} \Big( M\delta^{\rho,\rho'} + (1-\delta^{\rho,\rho'})\sqrt{M}z \Big) \tag{4.55}$$

where $x$, $y$ and $z$ are independent random variables distributed as $\mathcal{N}(0,1)$, applying the LLN argument to each term. The derivation for $\mathcal{M} = 1$ is essentially identical, leading to $p(3) = p(1) = \frac{1}{3}, \frac{2}{3}$.

For $\mathcal{M} = 2$, the same argument leads to $p(2) = \frac{1}{2}, \frac{1}{4}$,

**Derivation of peak probabilities**

We now derive the peak probabilities, Prob $\Big(p(\mathcal{M})\Big)$ = Prob$(P, K)$, which will will turn out to be the same for any $\mathcal{M}$, and only dependent on $P$ and $K$. What is the probability of obtaining either one of the structured entries in $h^{\mu}h^{\nu}$? This problem has a simple combinatoric solution, akin to the problem of throwing darts on the block covariance matrix and asking with what probability it sticks on any one of the sub-blocks. For instance, consider the simplest case of a $P \times P$ matrix with $P^2$ entries. The probability that we obtain an element on the diagonal is $\frac{P}{P^2} = \frac{1}{P}$, whereas the probability that we obtain one of the off-diagonal elements is $\frac{P(P-1)}{P^2} = 1 - \frac{1}{P}$.

On the next level of complexity, consider now a covariance matrix with two-object index, $(\mu, \gamma)$, $\mu = 1, ..., P$, $\gamma = 1, ..., K$ (i.e that of a CEMS $N_m = 2$ model). The probability of obtaining a diagonal element, Prob$(\mu = \mu', \gamma = \gamma') = \frac{1}{PK}$. To obtain the probability of one on the structured elements of this matrix, say Prob$(\mu = \mu', \gamma \neq \gamma')$, we need to perform the marginalization Prob$(\mu = \mu', \gamma \neq \gamma')$ = Prob$(\mu = \mu' | \gamma = \gamma') \times$ Prob$(\gamma \neq \gamma')$, where we have that Prob$(\gamma \neq \gamma') = 1 - \frac{1}{K}$ and Prob$(\mu = \mu' | \gamma = \gamma') = \frac{1}{P}$, i.e the probability of obtaining a diagonal term within a given block is the same as the probability of obtaining a diagonal term in the simplest $P \times P$ matrix. Thus we have the

following four possible probabilities

$$\text{Prob}(\mu = \mu', \gamma = \gamma') = \frac{1}{PK} \tag{4.56}$$

$$\text{Prob}(\mu = \mu', \gamma \neq \gamma') = (\frac{1}{P})(1 - \frac{1}{K}) \tag{4.57}$$

$$\text{Prob}(\mu \neq \mu', \gamma = \gamma') = (1 - \frac{1}{P})(\frac{1}{K}) \tag{4.58}$$

$$\text{Prob}(\mu \neq \mu', \gamma \neq \gamma') = (1 - \frac{1}{P})(1 - \frac{1}{K}) \tag{4.59}$$

which can be shown to add to 1 as required. The above argument generalizes to CEMS $\mathcal{M} = 3$. We have the following probabilities:

$$\text{Prob}(\mu \neq \mu', \gamma \neq \gamma', \rho \neq \rho') = (1 - \frac{1}{P})(1 - \frac{1}{K})^2 \tag{4.60}$$

$$\text{Prob}(\mu \neq \mu', \gamma \neq \gamma', \rho = \rho') = \text{Prob}(\mu \neq \mu', \gamma = \gamma', \rho \neq \rho') = (1 - \frac{1}{P})(\frac{1}{K})(1 - \frac{1}{K}) \tag{4.61}$$

$$\text{Prob}(\mu \neq \mu', \gamma = \gamma', \rho = \rho') = (1 - \frac{1}{P})(\frac{1}{K})^2 \tag{4.62}$$

$$\text{Prob}(\mu \neq \mu', \gamma = \gamma', \rho = \rho') = (1 - \frac{1}{P})(\frac{1}{K})^2 \tag{4.63}$$

$$\text{Prob}(\mu = \mu', \gamma \neq \gamma', \rho = \rho') = \text{Prob}(\mu = \mu', \gamma = \gamma', \rho \neq \rho') = (1 - \frac{1}{K})(\frac{1}{PK}) \tag{4.64}$$

$$\text{Prob}(\mu = \mu', \gamma = \gamma', \rho = \rho') = \frac{1}{PK^2} \tag{4.65}$$

where the final line indicates the probability of attaining any of the diagonal entries (this is prohibitively small, thus can't be seen in for instance Figure (4.5)). Once again, we can check that the sum of equations in (4.65) is equal to unity.

### First order approximation to structured excess overlaps

Given the above peak probabilities, we now have to compute their net effect on correlation functions of the form $\langle \mathcal{I}^{\mu\nu} \rangle = \langle m_i^\mu m_i^\nu \rangle$. To do this, we first write

$$\langle \mathcal{I}^{\mu\nu} \rangle = \sum_p \text{Prob}(p) \langle \mathcal{I}^{\mu\nu}(p) \rangle \tag{4.66}$$

where $\langle \mathcal{I}^{\mu\nu}(p) \rangle$ further consists of the $O(1)$ correlation function plus a sub-leading $O(\frac{1}{\sqrt{N}})$. Note that in the unimodal case we simply had $p = 0$, with $\text{Prob}(0) = 1$, thus the $O(1)$ term was $f^2$ [13] and the sub-leading term was (4.11). Explicitly, $\langle \mathcal{I}^{\mu\nu}(p) \rangle$ looks like

$$\langle \mathcal{I}^{\mu\nu}(p) \rangle = \int_{T,T}^{\infty,\infty} \frac{dh_1 dh_2}{(2\pi)\sqrt{1 - (p+\alpha)^2}} \exp\Big[ - \frac{1}{2(1 - (p+\alpha)^2)} \Big( h_1^2 + h_2^2 - 2h_1 h_2 (p + \alpha) \Big) \Big] \tag{4.67}$$

[13] In the case where we do not subtract the mean $f$ from $m$.

where $\alpha$ is small term given by the ratio of the Gaussian random variable to the square root of the respective number of inputs. For the case $p = 0$, it is straightforward to derive such a decomposition to a leading term plus an $O(\frac{1}{\sqrt{N}})$ one - i.e the excess overlap. In the case where $p \neq 0$, such an approximation need not hold, and a full, comprehensive asymptotic expansion is needed.

We do not provide here such a full expansion; we instead choose to approximate (4.67) by

$$\langle \mathcal{I}^{\mu\nu}(p) \rangle \approx \langle \mathcal{I}^{\mu,\nu}_{lead}(p) \rangle + \frac{Q(p)}{\sqrt{N}} x \tag{4.68}$$

which we call the *Linearity Lemma*, where

$$\langle \mathcal{I}^{\mu,\nu}_{lead}(p) \rangle = \int_{T,T}^{\infty,\infty} \frac{dh_1 dh_2}{(2\pi)\sqrt{1-(p)^2}} \exp\Big[ - \frac{1}{2(1-(p)^2)} \Big( h_1^2 + h_2^2 - 2h_1 h_2 p \Big) \Big] \tag{4.69}$$

and

$$Q(p) = \mathcal{F}(p, \mathcal{M}) \times \int_{T,T}^{\infty,\infty} \frac{dh_1 dh_2}{(2\pi)\sqrt{1-(p)^2}} h_1 h_2 \exp\Big[ - \frac{1}{2(1-(p)^2)} \Big( h_1^2 + h_2^2 - 2h_1 h_2 p \Big) \Big] \tag{4.70}$$

where $\mathcal{F}(p, \mathcal{M})$ a numerical pre-factor explained below. Intuitively, we expect the Linearity Lemma to give a good approximation to the correct answer when $p$ is small. Both (4.69) and (4.70) calculated numerically when making theoretical predictions. Evaluating $\mathcal{F}(p, \mathcal{M})$ requires a case-by-case evaluation. For instance, consider the $\mathcal{M} = 3$ case at $p = \frac{1}{3}$, corresponding to $\mu = \mu'$, but $\gamma \neq \gamma'$ and $\rho \neq \rho'$. Here, note that the averaged squared spread around $p = \frac{1}{3}$ (for this possibility, thus we include a subscript $I$) will be given by

$$\mathcal{F}_I(1/3, \mathcal{M} = 3) = \langle \Big( \frac{1}{3\sqrt{M}} (y + z) \Big)^2 \rangle = \frac{2}{9M} \tag{4.71}$$

such a procedure is used to calculate numerical pre-factors for all peaks for different values of $\mathcal{M}$, which can we exhaustively list for $\mathcal{M} = 3$ below:

$$\mathcal{F}(0, \mathcal{M} = 3) = \langle \Big( \frac{1}{3\sqrt{N}} (x + \frac{y}{\sqrt{\delta}} + \frac{z}{\sqrt{\delta}}) \Big)^2 \rangle = \frac{1}{9N} \Big( \frac{\delta + 2}{\delta} \Big) \tag{4.72}$$

$$\mathcal{F}_{II}(1/3, \mathcal{M} = 3) = \langle \Big( \frac{1}{3\sqrt{N}} (x + \frac{y}{\sqrt{\delta}}) \Big)^2 \rangle = \frac{2}{9N} \Big( \frac{\delta + 1}{\delta} \Big) \tag{4.73}$$

$$\mathcal{F}_I(2/3, \mathcal{M} = 3) = \langle \Big( \frac{1}{3\sqrt{N}} (x) \Big)^2 \rangle = \frac{1}{9N} \tag{4.74}$$

$$\mathcal{F}_{II}(2/3, \mathcal{M} = 3) = \langle \Big( \frac{1}{3\sqrt{M}} (y) \Big)^2 \rangle = \frac{1}{9M} \tag{4.75}$$

which are the same for $\mathcal{M} = 1$. Applying Eq. (4.75), computing Eq. (4.69) and Eq. (4.70), and weighting the peaks by that given in Eq. (4.65) leads to theoretical lines given in Fig. (4.7). It should be noted that such an approximation for $\mathcal{M} = 2$ and $\mathcal{M} = 1$ (with different respective expressions for $\mathcal{F}(p)$) does not seem to work (plots not shown in this thesis), and thus a full theoretical description of the interference term in the presence of structured multi-modal overlaps for all $\mathcal{M}$ would be important future work.

**Comparison of dimensionality across mixing degree and against unimodal model**

Firstly, let us comment of the comparison between the CEMS $\mathcal{M} = 3$ (Fig (4.7) main text) and the unimodal case, in the top row of Figure (4.13). We immediately note that $\mathcal{D}$ is substantially larger in the unimodal case, owing to the $Q^2 \times \frac{1}{N}$ contributions, and so is the value of $Q^2$ across different coding levels, since there is a single $p = 0$ peak with no fudging probabilities $\mathcal{F}(p)$. Second, we can compare the CEMS $\mathcal{M} = 3$ to that of $\mathcal{M} = 2$, where we find that the dimensionality and re-scaled interference term to be of the same order is magnitude. Our simulation results show that the $\mathcal{M} = 2$ performs accordingly roughly as well as $\mathcal{M} = 3$.

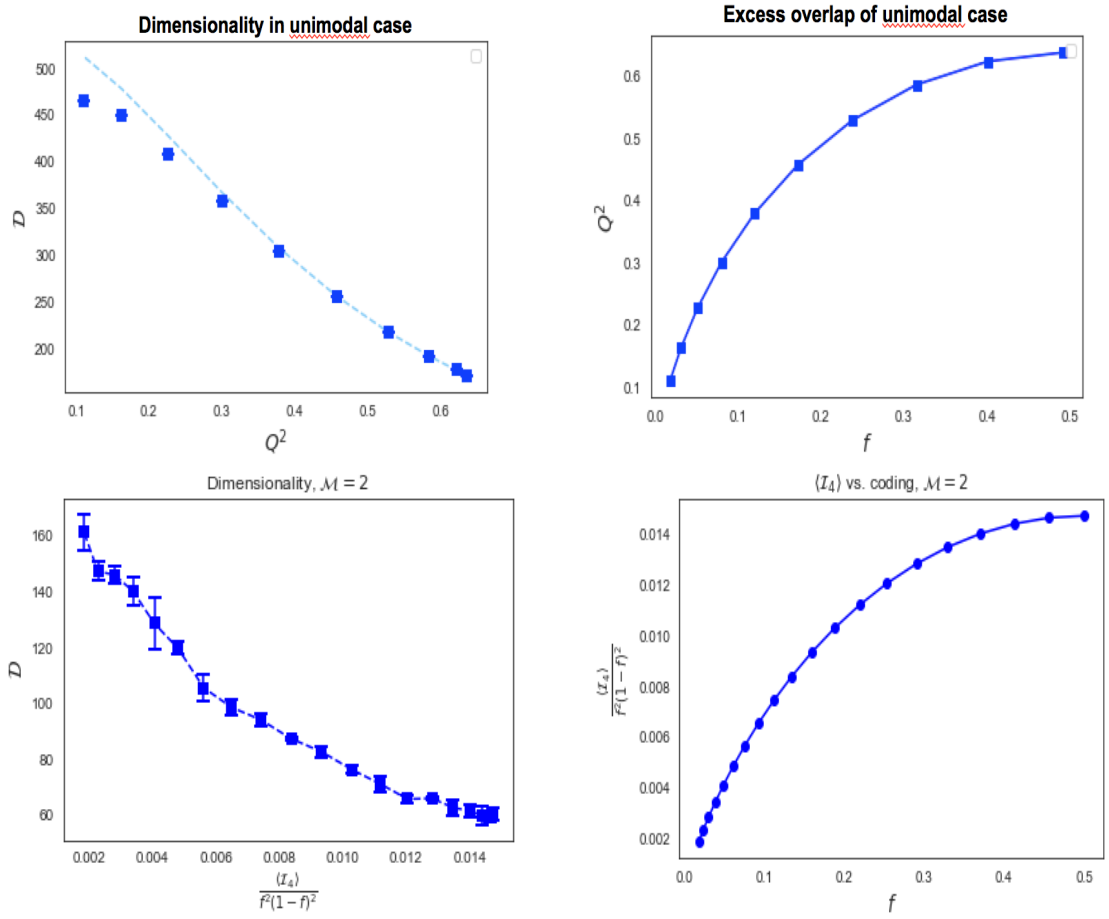


Figure 4.13: *Top:* Dimensionality and excess overlaps for unimodal model for comparison. *Bottom:* Dimensionality and re-scaled $\langle \mathcal{I}_4 \rangle$ for $\mathcal{M} = 2$ for comparison. Simulation details are $P_{eff} = 1000$ and $N_c = 1200$.

# Chapter 5

# Further directions

In this Chapter, we will build on the insights developed in our study of the CEMS model to the problem of contextual decoding, a widely studied phenomenon in mixedly selective circuits. Then, we will take a step back from CEMS, and discuss an extension on the unimodal case, to model (threshold)linear mixed selectivity. Finally, we will summarise the findings of our thesis and suggest potential future directions.

## 5.1 Mixed contextual decoding

In some experiments, particularly lesioning experiments in the hippocampus, behavioural assays have been used as a proxy for pattern separation, and empirical findings have implicated the role of the cerebellar-like architecture in the dentate-gyrus CA1/3 region in discrimination in different contexts. For instance, in a standard contextual fear conditioning paradigm [60] (see also [36]), mice trained in a fear conditioning paradigm would not be able to discriminate a different contexts [1] upon lesion of the dentate gyrus, suggesting its role for pattern separation of contextual information. Another example would be famous behavioural experiments, by Rigotti, et. al [73], where a monkey is trained to guess the sequence of visual stimuli presented, but in the different contexts of it being either a recognition or a recall task. We will study such a model under the CEMS architecture, and we will argue for the utility of a dense representation for the task of decoding contexts.

To model such a contextual discrimination, we study in this subsection the CEMS model under a different statistics of the training and test patterns than that used in the previous Chapter. In particular, we consider a model where the CEMS is to learn associations between $P$ independent stimuli and valences, with the following statistics

$$\bar{\boldsymbol{\xi}}^{(\mu,1,1)} = \begin{pmatrix} \xi^\mu & \phi^1 & \eta^1 \end{pmatrix}^T, \quad \hat{\bar{\boldsymbol{\xi}}}^{(\mu,2,2)} = \begin{pmatrix} \xi^\mu & \phi^2 & \eta^2 \end{pmatrix}^T \tag{5.1}$$

[1] Here, context refers to abstract environmental cues that the shock is administered in.

thus the generalization abilities for the contextual decoding task is to be able to distinguish the same task-relevant stimulus, but in the presence of distinct independent sources of information. In the case of the first experiment metioned above, $\xi^\mu$ can be taken to be the fear conditioning stimulus that the mice has to associate, whereas in the second case it could be the sequence of items presented that the monkey has to remember. This correlation structure of the training and test patterns may be seen to be a drastic over-simplification, where different contexts give rise to orthogonal sub-spaces in the input datum, but nevertheless serves as a useful null model.

### 5.1.1 Mixed layer cluster size and representation dimensionality

In contrast to the unimodal and CEMS models, we now do not have an input noise parameter that determines the cluster size. In fact, we can show that the cluster size is purely determined by the mixing index, and in general (though not for $\mathcal{M} = 1$), is a monotonically decreasing function of $f$. To understand the behaviour of these clusters, we note that the statistics of the inputs in (5.1) render the following results for the cluster sizes for different $\mathcal{M}$

$$\Delta m_{\mathcal{M}=3} = \mathcal{G}(T, 2/3) \tag{5.2}$$

$$\Delta m_{\mathcal{M}=2} = \mathcal{G}(T, 1/2) \tag{5.3}$$

$$\Delta m_{\mathcal{M}=1} = \frac{2}{3}\mathcal{G}(T, 1) = \frac{2}{3} \tag{5.4}$$

$$\tag{5.5}$$

which we derive in Appendix. Note that in the purely selective case we have that the mixed layer cluster size is in fact a constant, independent of $T$, and is more robust to noise that the other cases. The origin of these results are due to the appearance again of structured overlaps, which now appear in the statistics of $h\bar{h}$ (see Appendix). Because of the simplified statistics in (5.1), there is only one peak in our contextual decoding model for each $\mathcal{M}$. The origin of the relative advantage of pure selectivity for noise amplification in this model is intuitive - since one-third of the neurons on the mixed layer receive noiseless inputs, whereas for partial and full selectivity all neurons will integrate the contextual inputs that are orthogonal to each other when computing the cluster size.

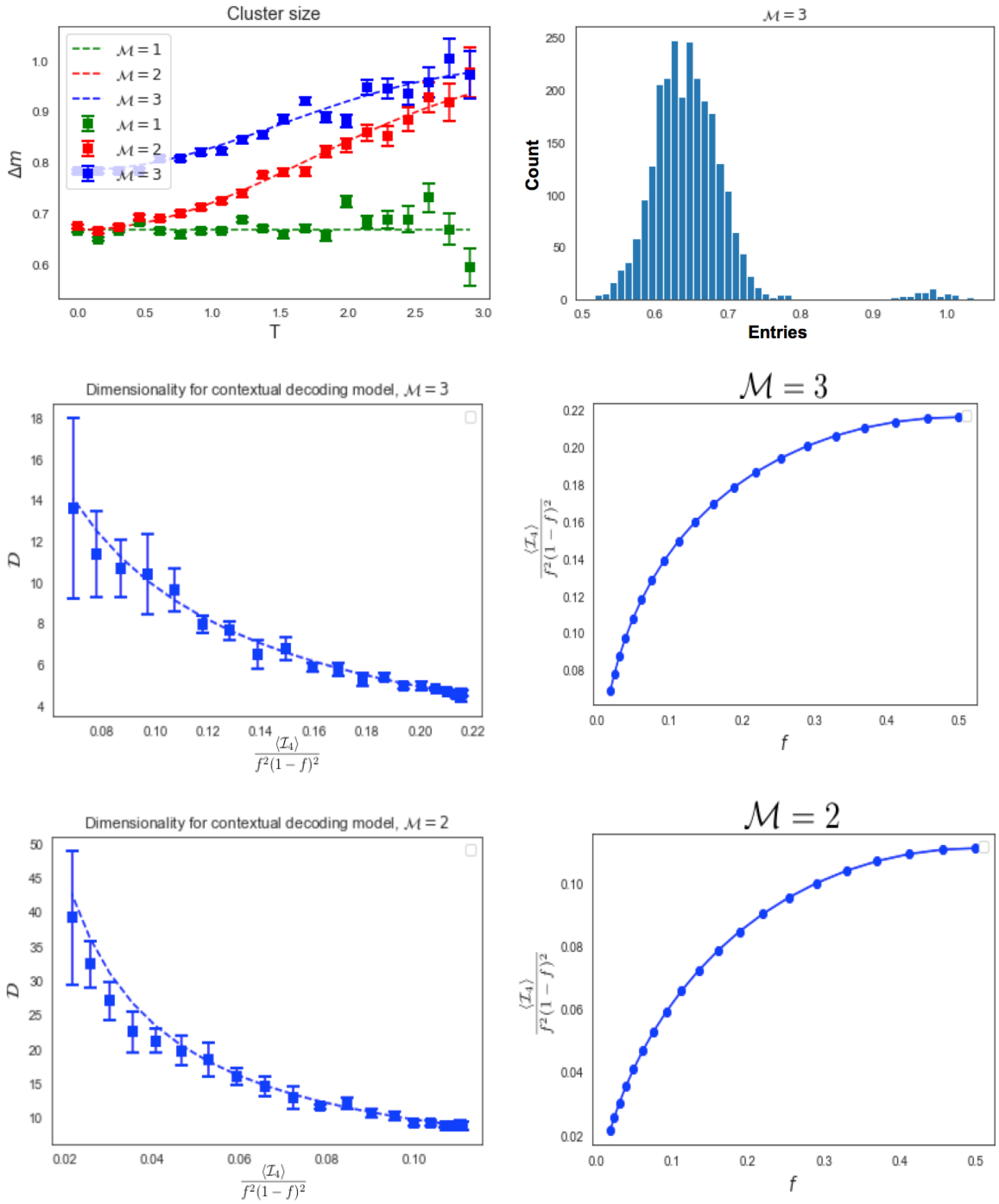


Figure 5.1: *Top left*: The cluster size at the mixed layer is now determined for different $\mathcal{M}$ and $T$, with the simulation results compared with theoretical predictions shown in the dashed lines. *Top right*: An example of structured multimodal overlaps in the contextual decoding model for $\mathcal{M} = 3$; displayed in the histogram of entries of the covariance matrix of distinct inputs. *Midddle and bottom rows:* The dimensionality of the representation is displayed along with the re-scaled interference term, comparing the $\mathcal{M} = 3$ (middle) and $\mathcal{M} = 2$ (bottom) cases, with theoretical predictions given again in dashed lines. Simulation parameters are $N = M = P = 100$, $N_c = 1200$.

We can now study $\mathcal{D}$ for this model, by first deriving the structured multi-modal overlaps for the input statistics given in (5.1). The result is simpler than those obtained in the previous Chapter, the structured peaks are

$$p(\mathcal{M}=1) = p(\mathcal{M}=3) = \frac{2}{3} \tag{5.6}$$

$$p(\mathcal{M}=2) = \frac{1}{2} \tag{5.7}$$

Thus, as in the previous Chapter, we expect correlations evaluated at these peaks to give $O(1)$ contributions to the interference term, with sub-leading $O(\frac{1}{\sqrt{N}})$ terms. However, unlike the previous Chapter, owing to our simplified input statistics, we have (with measure 1) no $(P, K)$ dependence on the probability of different structured peaks.

Nevertheless, we can study the dimensionality of the mixed representation, where we find that (again) that the dimensionality decreases as a function of mixing (compare left of Figure (5.1), middle and bottom), owing to an increased interference term (right of Figure (5.1), middle and bottom). The origin of these are now more straightforwardly understood, since correlations computed at the smaller peak (for smaller $\mathcal{M}=2$ relative to $\mathcal{M}=3$) will lead to a smaller interference terms.

### 5.1.2 Utility of dense coding and a large expansion

We can now provide simulation results for our contextual decoding model, where we argue for the utility of dense coding for the task of contextual decoding. We simulate the CEMS for $\mathcal{M}=2$ and $\mathcal{M}=3$ with $N=M=100$, $\mathcal{R}=27$ (see Figure (5.2)). We see the general trend that learning improves as the task relevant load is decreased. However, we note that, similar to what was observed in the unimodal $\{+,-\}$ case, a dense code is advantageous. The relative role of the expansion ratio is detailed in the Appendix; we find that expansions are advantageous. We leave the corresponding study for $\mathcal{M}=1$ as future work.

These results may be consistent with recent observations from cerebellar GCs activities that display a highly dense code, when measured over longer timescales than physiologically relevant ones [53], [37]. Though it is not clear what the utility of such a dense code is, this results above could be relevant to behavioural paradigms involving sensory discrimination under, in for instance, eyeblink conditioning tasks, where the role of the cerebellum has been widely documented [32], [77].

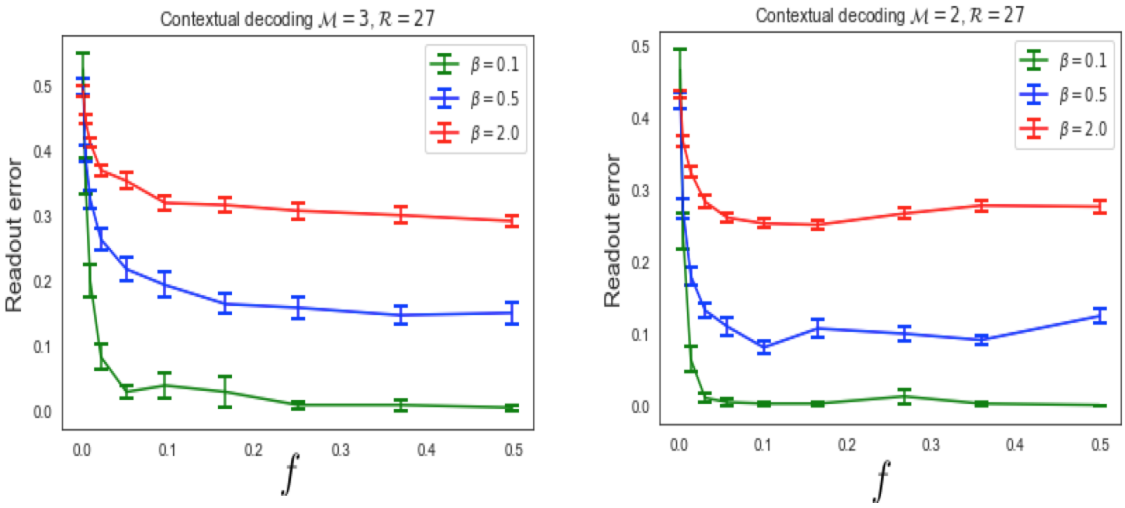


Figure 5.2: Readout error as a function of sparseness for the contextual decoding model, comparing $\mathcal{M} = 3$ (left) and $\mathcal{M} = 2$ (right). Results are plotted against $f$ for different values of the task-relevant load $\beta$. Simulation details are given in main text.

## 5.2 Linear vs. non-linear selectivity

In this section, we study a model of (threshold)linear selectivity, focusing on the the unimodal model, and leaving the corresponding mixed model as further work. Apart from being an important paradigm in computational neuroscience studies (see [16], [52]) for just two examples from diverse areas), such studies may also be pertinent for the study of its utility in artificial networks [4]. We note as well that famous experiments [73] have shown that a majority of neurons recorded linearly mix different sources of information, thus theoretical results here might shed light of the computational properties of such a phenomenon.

The mixed layer activities for the ReLU activation function read

$$m_i^\mu = h_i^\mu \theta(h_i^\mu - T) \tag{5.8}$$

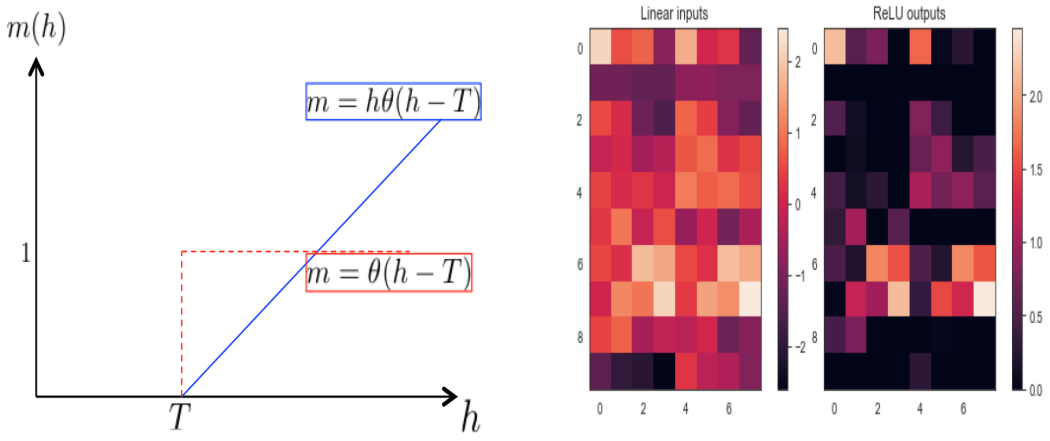


Figure 5.3: An illustration of the rectified linear unit activation function (ReLU) used in our threshold linear model. *Left*: The exact form of the activation function is shown in blue, with a comparison with the other non-linearity used throughout this thesis, the Heaviside non-linearity in red. *Right*: For a small network, we display on the left the $N_c \times P$ matrix of linear (random) inputs onto the cortical layer, followed by the outputs having passed through the non-linearity on the right.

Note that we have not centered $m$ by its mean for this model. In this setting, we will tease apart the relative importance of the hitherto neglected order parameters appearing in Eq. (4.6), which will turn out to control the behaviour of the dimensionality and hence the generalization abilities of a Hebbian readout.

### 5.2.1 Order parameters and dimensionality at cortical layer

As it turns out for the (threshold)linear model, order parameters such as $\langle q^2 \rangle$ and $\langle q^4 \rangle$, which yielded simple expressions in the previous problems, need to be explicitly determined, and as well $\langle \mathcal{I}_4 \rangle$. We do these in the Appendix. We have that

$$\langle q^2 \rangle = TG(T) + H(T) \tag{5.9}$$
$$\langle q^4 \rangle = (T^2 + 3T)G(T) + 3H(T) \tag{5.10}$$
$$\langle \mathcal{I}_4 \rangle = G^4(T) + \frac{1}{N}(\langle q^2 \rangle)^4 \tag{5.11}$$

where $G(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$. The mathematical structure of Eq. (5.11) is interesting - our excess overlaps in the case is $\langle q^2 \rangle^4$, as opposed to $G^4(T)$ in the unimodal case, whereas the leading term is $G^4(T)$ instead of $H^4(T)$. All of these order parameters are a monotonically decreasing function of $T$, except that $\langle q^4 \rangle$ has a slightly non-monotonic decrease.

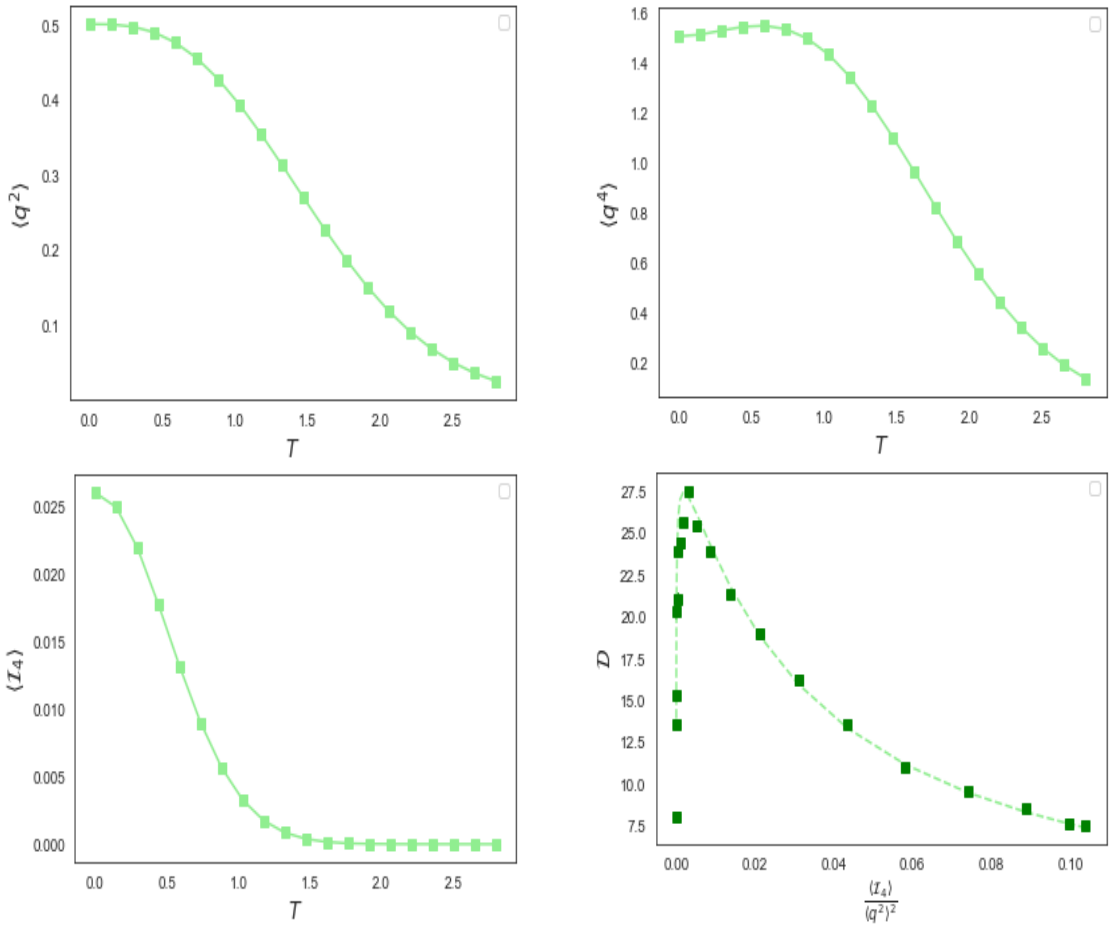


Figure 5.4: An illustration of our analytical form of the various order parameters needed to describe the behaviour of the threshold linear cortical model. *Top:* On the left, we display the form of $\langle q^2 \rangle$ as a function of $T$, whereas on the right we display the corresponding behaviour of $\langle q^4 \rangle$, with the interesting point being its slight non-monoticity with respect to $T$. Note that, unlike in the models studied in other Chapters, the ratio $\frac{\langle q^2 \rangle^2}{\langle q^4 \rangle}$ is not equal to one. *Bottom:* We display the behaviour of the interference term $\langle \mathcal{I}_4 \rangle$ on the left, and the dimensionality $\mathcal{D}$ on the right. The dotted lines are the corresponding analytical theory given by Eq. (4.6), using the value of the corresponding order parameters. Simulation details are $N = M = P = 100$, $N_c = 2100$.

As a result of the behaviour of does not display the aforementioned asymptotic scaling law as a function of $\langle \mathcal{I}_4 \rangle$. Indeed we see from Figure (5.4), bottom right, that there is an optimal rescaled $\langle \mathcal{I}_4 \rangle$ that maximizes dimensionality.

### 5.2.2 Readout error

We note here that the setup for generalization is slightly different from that in the unimodal model of [3]. In particular, we do not have a well defined (properly normalized)

cluster size, at the mixed layer, and indeed the expression for the SNR is slightly different than in Eq. (4.10), reading

$$SNR \approx \frac{\langle m\hat{m}\rangle^2}{\alpha\langle q^2\rangle^2 + P\langle \mathcal{I}_4\rangle} \tag{5.12}$$

as derived in the Appendix. There are two differences between the SNR here and those in the previous Chapter (e.g. Eq. (4.10)). Firstly, there is no properly normalized cluster size in this model, thus the signal contribution needs a separate treatment. This will turn out to be analogous to what was done when computing overlaps for at structured peaks in the previous Chapter, and we can provide an approximation to this, as we explain in the Appendix. Secondly, the $\langle q^2\rangle$ order parameter now controls learning by the first term in the noise contribution, and are given by those shown in Fig (5.4).

### Role of sparsity and expansion ratio

Here, we can report on two features: the *importance* of sparsity, and once again the relative unimportance of the expansion ratio. This is shown in Figure (5.5). On the left, we shown the behaviour of the readout error as a function of sparseness $f$, whereas on the right we display it as a function of the expansion ratio. Note here, that unlike in most other models, the behaviour with respect to $f$ and $\mathcal{R}$ is not fully analytically accessible, since the numerator in (5.12) is also a non-trivial function of $f$. Nevertheless, we see that the generic trend of the relative unimportance of the expansion ratio and the lack of an optimal sparseness for cases when the re-scaled interference term in of $O(1)$ is recapitulated.

It is interesting to compare the results here with those from [16] (although their model has important details the non-linearity choice, the generic expansion + linear readout architecture is the same). It was found in [16], that sparsity of the representation tends to not affect the *speed* of learning. Here, we display that for the problem of generalization on a similar setup, it is very beneficial.

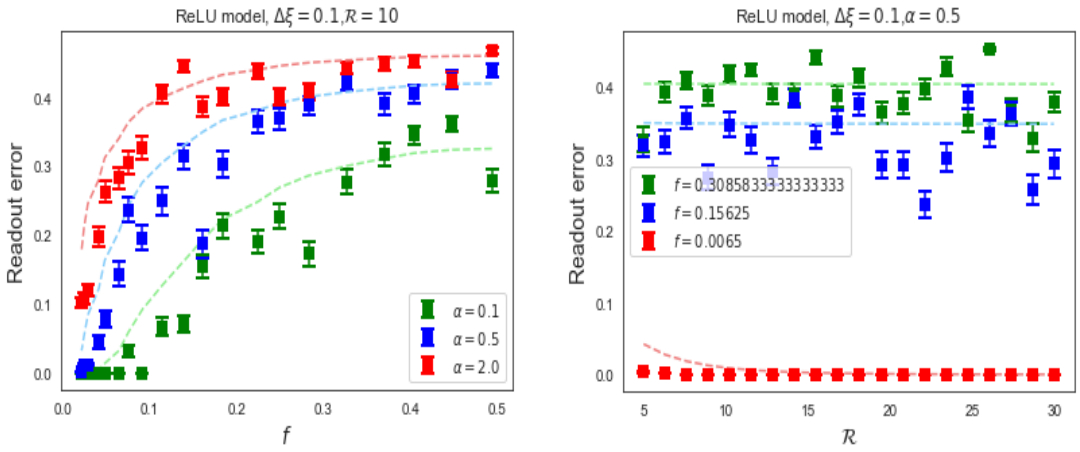


Figure 5.5: Simulations results for the generalization error of the unimodal threshold linear model. *Left*: The variation with $f$ for different $\alpha$ is displayed, indicating the utility of a sparse representation for the threshold linear model. On the right, we display that the expansion ratio, $\mathcal{R}$, is again relatively unimportant, and in fact for all cases with the exception of the ultra-sparse case, the error is effectively constant with $\mathcal{R}$. $N = 400$ is used simulations, with other parameters given in subheadings, and theoretical predictions from Eq. (5.12).

## 5.3 Summary of thesis

In this thesis we have extended a previously introduced model of expansive feed-forward circuitry in early sensory systems to the case of when such a network is multi-modal, integrating information from task-relevant and contextual modalities, whose statistics are modelled as independent sources of information. Our model, called CEMS, had three important features - random feed-forward weights, sparsity-preservation across mixing degree, and modality-coverage preservation across mixing degree. The phenomenology of this model then turned out to be rich - we found that there is a severe storage bottleneck of contextual information in cases where the network is less than fully mixed, whilst in the case of full mixing a simple scaling law describes the task-relevant capacity. Then, we saw that there is a distinct non-monoticity of noise amplification as a function of mixing degree, and that for the problem of generalization, the abilities of the network are largely unaffected by variations in the sparseness of the mixed representation and the expansion ratio - parameters that severely controlled learning in the unimodal case ([3]). We then saw, in this Chapter, that for a contextual decoding model, a dense code and a large expansion are useful for the problem of generalization - similar to the unimodal case but with a sign non-linearity at the mixed layer, used in other famous works on mixed selectivity [5]. Finally for a (threshold)linear unimodal model, we saw that a sparse code is advantageous for the same problem, whilst the expansion ratio does not appreciably control learning.

Central to the findings in the multi-modal and contextual decoding models is the notion of *structured multi-modal overlaps*, a simple consequence of randomly mixing independent sources of information on the statistical structure of inputs onto the mixed layer. This extends the notion of excess overlaps first introduced in [3] - except here instead there are $O(1)$ contributions to relevant order parameters, as opposed to $O(\frac{1}{N})$ in the case of excess overlaps. Such contributions share some similarities with the result of structured projections studied in [3] - in that they introduced structure to an otherwise random input statistic, but their effects on learning are very different - in the case of structured projections the SNR and noise variability is found to be greatly suppressed, and hence learning enhanced; in the case of structured multi-modal overlaps learning is worsened relative to the unimodal case. An interesting question in this regard would be to study how structured multi-modal overlaps change in the presence of non-random feed-forward projections, and how that might affect the generalization abilities of CEMS.

A recurring theme in this thesis is the subtle dependence of the learning abilities of the network on the choice of non-linearity and training vs. test data statistics of the particular model. We find a whole family of solutions for different models, each displaying different dependencies on the sparseness of the mixed layer representation and the expansion ratio on the network's generalization abilities. Such a finding is potentially important for claims of optimal sparsity in computational neuroscience studies (e.g [5]) - in reality such claims depend strongly on specific details of the model: the feed-forward weight distribution, non-linearity choice at the mixed layer, and the rule imposed at the readout layer. A more comprehensive study in this space may be promising further direction. On a technical front, we mention that some results here are novel - for instance the full analytical tractability of the dimensionality of the mixed layer as a function of calculable order parameters. In this thesis, we have provided exact analytical expressions for this in the case of random projections - i.e Gaussian random input statistics onto the mixed layer. An interesting further direction here would be the study of the dimensionality for simple learning rules at the feed-forward layer - in principle this would simply require re-evaluation of order parameters given here. On another technical front, we mention here that a full theory for the correlation functions induced by structured multi-modal overlaps for all mixing indices is hitherto lacking, although our analytical approximations provide good agreement with simulations for the dimensionality curves across various models.

Many other extensions, particularly those that impose more biological realism, are possible. For instance, the role of *sparse connectivity* in expansion recoding has been studied theoretically [57], [16], based on the well known restricted degree distribution at the granule cell layer from the mossy fibre inputs [24], [91]. Furthermore, learning mechanisms - either supervised or unsupervised - at the input-to-cortical layer has been the subject of many computational studies in the unimodal case [3], but also more recently in mixed selectivity models [57],[56],[23], and could be another interesting extension. Finally, we should mention as well the known presence of recurrent inhibitory interneurons,

present in for instance mitral cells in the olfactory bulb [14],[95],[38]. Nevertheless, we think the results here will spur discussion and further work along different areas.

## 5.4 Appendix

### 5.4.1 Derivations for contextual decoding model

We now derive the results for the contextual decoding model, where we will show that the phenomenon of structured peaks appear in the covariances of both $h^\mu h^\nu$ and $h\hat{h}$

#### Derivation of cluster size

Let us again consider this case by case. For $\mathcal{M} = 3$, we have that

$$\Delta m_{\mathcal{M}=3} = \frac{1}{f(1-f)}\Big[\text{Prob}(h^{(\xi,\phi^1,\rho^1)} > T, h^{(\xi,\phi^2,\rho^2)} < T\Big] \tag{5.13}$$

which requires us to compute the statistics of $\langle h^{(\xi,\phi^1,\rho^1)} h^{(\xi,\phi^2,\rho^2)}\rangle$, which is

$$\langle h^{(\xi,\phi^1,\rho^1)} h^{(\xi,\phi^2,\rho^2)}\rangle = \frac{1}{3} + \frac{1}{3M}\sum_k \langle \phi_k^1 \phi_k^2\rangle + \frac{1}{3M}\sum_l \langle \eta_l^1 \eta_l^2\rangle \tag{5.14}$$

which leads to a $\Delta S_{eff} = \frac{2}{3}$. Plugging this into Eq. (5.13) leads to $\Delta m_{\mathcal{M}=3} = \mathcal{G}(T, \frac{2}{3})$ in the main text. Next we consider $\mathcal{M} = 2$, where we have

$$\Delta m_{\mathcal{M}=2} = \frac{1}{f(1-f)}\Big[\frac{1}{2}\,\text{Prob}(h^{(\xi,\phi^1)} > T, h^{(\xi,\phi^2)} < T) + \frac{1}{2}\,\text{Prob}(h^{(\xi,\eta^1)} > T, h^{(\xi,\eta^2)} < T)\Big] \tag{5.15}$$

thus we have to compute $\langle h^{(\xi,\phi^1)} h^{(\xi,\phi^2)}\rangle$ and likewise for $(\xi, \eta)$ separately, leading to $\Delta S_{eff} = \frac{1}{2}$ for both, giving $\Delta m_{\mathcal{M}=2} = \mathcal{G}(T, \frac{1}{2})$. Finally, we have $\mathcal{M} = 1$, where we repeat the above procedure to get

$$\Delta m = \frac{1}{f(1-f)}\Big[\frac{1}{3}\,\text{Prob}(h^{(\xi)} > T, h^{(\xi)} < T) + \frac{1}{3}\,\text{Prob}(h^{(\eta^1)} > T, h^{(\eta^2)} < T) + \tag{5.16}$$

$$\frac{1}{3}\,\text{Prob}(h^{(\rho^1)} > T, h^{(\rho^2)} < T)\Big] \tag{5.17}$$

We note that the first term has zero measure, whereas the correlation structure means that $\Delta S_{eff} = 1$ for both the second and third term, thus giving overall $\Delta m_{\mathcal{M}=1} = \frac{2}{3}\mathcal{G}(T, 1) = \frac{2}{3}$.

### Derivation of structured multi-modal overlaps

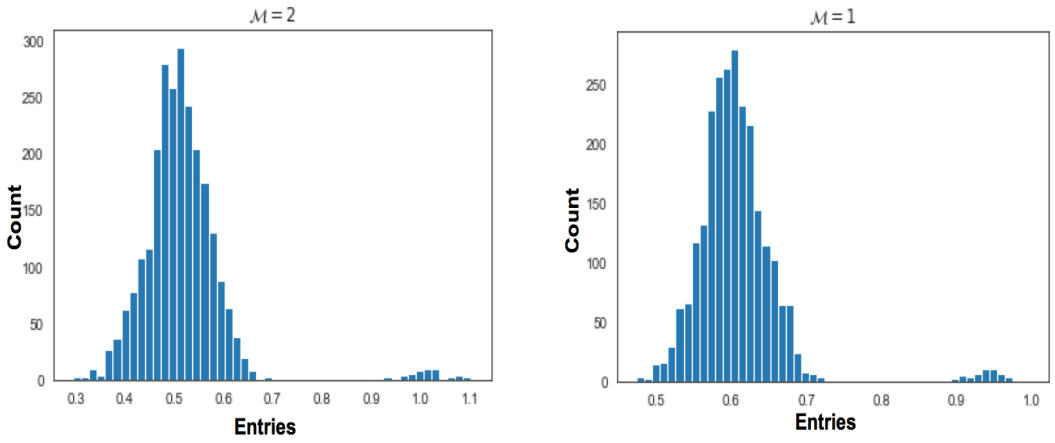


Figure 5.6: Location of structured peaks for contextual decoding model, for (*left*) $\mathcal{M}=2$, and (*right*) $\mathcal{M}=1$.

We can derive the location in the peaks of the structured overlaps of the $h^{\mu}h^{\nu}$. The location of the peaks are straightforward to derive, in the same fashion as previous derivations. For $\mathcal{M}=3$ we have that

$$\langle h^{(\mu,1,1)}h^{(\mu',1,1)}\rangle = \frac{1}{3}\Big(\sum_{j,j'}\langle J_j^{(\xi)}J_{j'}^{(\xi)}\rangle\langle\xi_j^{\mu}\xi_{j'}^{\mu}\rangle + \sum_{k,k'}\langle J_k^{(\phi)}J_{k'}^{(\phi)}\rangle\langle\phi_k^1\phi_{k'}^1\rangle + \sum_{l,l'}\langle J_l^{(\eta)}J_{l'}^{(\eta)}\rangle\langle\phi_l^1\phi_{l'}^1\rangle\Big) \tag{5.18}$$

$$= \frac{1}{3}\Big(\delta^{\mu,\mu'} + (1-\delta^{\mu,\mu'})\frac{x}{\sqrt{N}}\Big) + \frac{2}{3} \tag{5.19}$$

where we have used the LLN in the second line. This yields $p(3)=\frac{2}{3}$, and the same procedure can be applied to get the same result for $p(1)$. For $\mathcal{M}=2$, we have after the same steps

$$\langle h^{(\mu,1,1)}h^{(\mu',1,1)}\rangle = \frac{1}{2}\Big(\delta^{\mu,\mu'} + (1-\delta^{\mu,\mu'})\frac{x}{\sqrt{N}}\Big) + \frac{1}{2} \tag{5.20}$$

yielding $p(2)=\frac{1}{2}$.

### Theory for dimensionality curves

We mention here that the theoretical curves for the plot of $\mathcal{D}$ in the contextual decoding model in Fig. (5.1) also uses the *Linearity Lemma* as in the previous Chapter, with calculation proceeding along the exact same lines as in Eq. (4.68) to (4.75), but with the

value of $p$ given above.

### Utility of expansion ratio for contextual discrimination

In the contextual decoding model, we have that the expansion ratio is advantageous for learning, as in the unimodal case, as seen from the plot in Fig. (5.7) below.

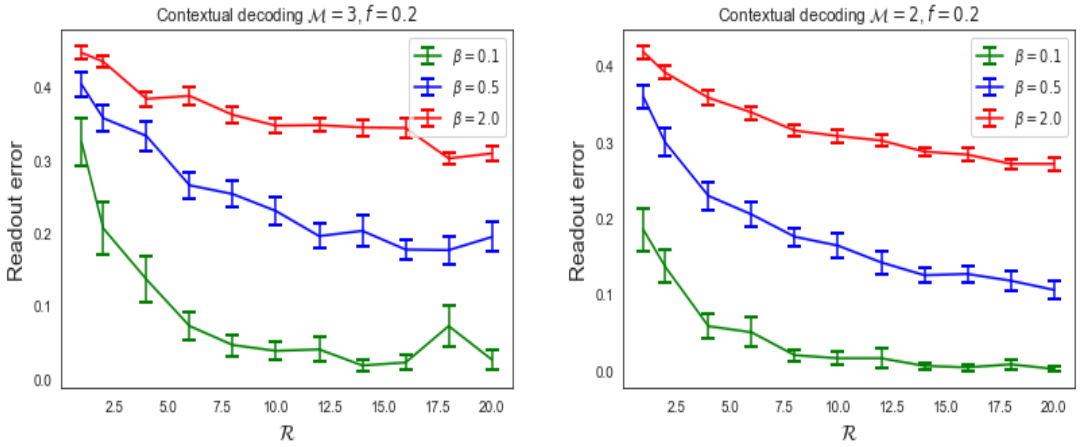


Figure 5.7: Effect of expansion ratio $\mathcal{R}$ on learning. We see that as learning is greatly improved as a function of $\mathcal{R}$ for both the fully (*left*) and partially (*right*) selective cases. Simulation details are same those used in the main text.

## 5.4.2 Derivations for threshold linear model

Here, we derive all the relevant order parameters and the SNR for the threshold linear model.

### Derivation of order parameters

First, let us derive $\langle q^2 \rangle$. We have that

$$\langle q^2 \rangle = \langle h^2 \theta^2 (h - T) \rangle = \langle h^2 \theta (h - T) \rangle \tag{5.21}$$

$$= \int_T^\infty dh h^2 \frac{e^{-\frac{h^2}{2}}}{\sqrt{2\pi}} \tag{5.22}$$

$$= TG(T) + H(T) \tag{5.23}$$

where we have used a standard integral in the final line. Next, we compute $\langle q^4 \rangle$, which is given by

$$\langle q^4 \rangle = \langle h^4 \theta^4 (h - T) \rangle = \langle h^4 \theta (h - T) \rangle \tag{5.24}$$

$$= \int_T^\infty dh h^4 \frac{e^{-\frac{h^2}{2}}}{\sqrt{2\pi}} \tag{5.25}$$

$$= (T^2 + 3T)G(T) + 3H(T) \tag{5.26}$$

Finally, we can derive $\langle \mathcal{I}_4 \rangle$, which we recall is given by $(\langle \mathcal{I}^{\mu\nu} \rangle)^2$ [2]. We have that

$$\langle \mathcal{I}^{\mu\nu} \rangle = \langle h^\mu \theta (h^\mu - T) h^\nu \theta (h^\nu - T) \rangle \tag{5.27}$$

The leading term in this gives $G^2(T)$. Expanding to first order in $\frac{x}{\sqrt{N}}$, we have that the first sub-leading term is $\frac{x}{\sqrt{N}} \langle h^2 \theta (h - T) \rangle^2 = \frac{x}{\sqrt{N}} \langle q^2 \rangle^2$. Thus, we have Eq. (5.11) as in the main text.

### Derivation of SNR

The derivation of the SNR for the threshold linear model proceeds in a slightly different manner to that in the previous Chapter, since we have to keep track of the aforementioned order parameters, and do not have a well-defined cluster size at the mixed layer. The signal term (recall again we use a Hebb rule at the readout layer) is

$$\langle g^\nu \sigma^\nu \rangle = \sum_{i,\mu} \langle m_i^\mu \hat{m}_i^\nu \sigma^\mu \sigma^\nu \rangle = N_c \langle m \hat{m} \rangle \tag{5.28}$$

which now cannot be written in terms of a cluster size and must be evaluated separately. To do this, we again use the *Linearity Lemma*, and now approximate this by

$$\langle m \hat{m} \rangle^2 \approx \langle \Big( \mathcal{I}^{\mu\nu} (1 - \Delta S) + \frac{x}{\sqrt{N}} \mathcal{Q}^{\mu\nu} (1 - \Delta S) \Big)^2 \rangle = \langle \mathcal{I}^{\mu\nu} (1 - \Delta S) \rangle^2 + \frac{1}{N} \langle \mathcal{Q}^{\mu\nu} (1 - \Delta S) \rangle^2 \tag{5.29}$$

where the expectation over the Gaussian random variable $x$ has been performed, $\langle \mathcal{I}^{\mu\nu} \rangle$ given above, and

$$\langle \mathcal{Q}^{\mu\nu} \rangle = \langle (h^\mu)^2 \theta (h^\mu - T) (h^\nu)^2 \theta (h^\nu - T) \rangle \tag{5.30}$$

which can be numerically evaluated.

[2] Since we are not centering the mixed layer activities by subtracting the mean. In that case, we would have $(\langle \mathcal{I}^{\mu\nu} \rangle - \langle \mathcal{I}_{ij} \rangle)^2$.

The noise term reads $\mathrm{Var}(g^\nu \sigma^\nu) = \langle (\sum_{\mu,i} m_i^\mu \sigma^\mu \hat{m}_i^\nu)^2 \rangle - \langle g^\nu \sigma^\nu \rangle^2$, which can be expanded to get

$$\mathrm{Var}(g^\nu \sigma^\nu) \approx N_c \langle (m_i^\mu)^2 (\hat{m}_i^\mu)^2 \rangle + N_c P \langle (m_i^\mu)^2 (\hat{m}_i^\nu)^2 \rangle + N_c^2 P^2 \langle \mathcal{I}_4 \rangle \tag{5.31}$$

Upon dividing $N_c^2$, the first term can be neglected (see range of $\langle q^2 \rangle$ in Figure (5.3)), and so the overall result is

$$SNR \approx \frac{\langle m\hat{m} \rangle^2}{\alpha \langle q^2 \rangle^2 + P \langle \mathcal{I}_4 \rangle} \tag{5.32}$$

as in the main text. Numerical evaluation of $\langle m\hat{m} \rangle$ via Eq. (5.29), $\langle q^2 \rangle$ from Eq. (5.27), and $\langle \mathcal{I}_4 \rangle$ from Eq. (5.11) gives the theoretical curves in Fig. (5.5).

# Chapter 6

# Supplementary Information

## 6.1 List of mathematical symbols used

| | |
|---|---|
| $\xi$ | Stimuli at the input layer |
| $\Delta\xi$ | Stimulus input variability |
| $\bar{\xi}$ | Block stimuli |
| $\hat{\xi}$ | Centroids (training data) |
| $\phi$ | Contextual modality at input layer |
| $\Delta\phi$ | Contextual input variability |
| $h$ | Linear inputs onto cortical layer |
| $m$ | Mixed layer activities |
| $\hat{m}$ | Activities of $\hat{\xi}$ at mixed layer |
| $T$ | Neuron threshold |
| $f$ | Sparseness of the representation |
| $\sigma$ | Valences or labels |
| $N$ | Number of inputs neurons for task-relevant stimuli |

| | |
|---|---|
| $M$ | Number of input neurons for context |
| $P$ | Number of task-relevant stimuli presented to the network |
| $K$ | Number of contextual stimuli presented to the network |
| $(P, K)$ | Short-hand when number of both $P$ and $K$ are considered. |
| $P_{eff}$ | Total number of composite patterns presented to the network |
| $N_c$ | Number of neurons at the cortical layer |
| $N_m$ | Number of modalities |
| $N_{mod}$ | Number of input neurons for a specific modality |
| $\alpha = \frac{P_{eff}}{N}$ | Load |
| $\beta = \frac{P}{N}$ | Task-relevant load |
| $\alpha_c$ (or $\beta_c$) | Critical load of network |
| $\mathcal{R} = \frac{N_c}{N(1+(N_m-1)\delta}$ | Expansion ratio[1] |
| $\Delta m$ | Variability at mixed layer |
| $\Delta S$ | Net input noise |
| $\Delta S_{eff}$ | Effective input noise |
| $\epsilon$ | Readout (generalization) error |
| $\langle\ \rangle$ | Denotes a statistical average |
| $\langle q^2 \rangle$ | Average squared input norm |
| $\langle q^4 \rangle$ | Average quartic input norm |
| $\langle \mathcal{I}_2 \rangle$ | Two-point interference term |

[1]Note that in this thesis, we have used the same letter, $\mathcal{R}$, to denote the set of real numbers.

| | |
|---|---|
| $\langle \mathcal{I}_4 \rangle$ | Four-point interference term |
| $g$ | Inputs to readout neuron |
| $\mathbf{C}$ | Covariance of mixed layer representation |
| $p$ | Peak location in structured multi-modal overlaps |
| $h\hat{h}$ | Covariance between training and test linear inputs ($2 \times 2$ matrix) |
| $h^\mu h^\nu$ | Covariance between linear inputs of different pattern ($P \times P$ matrix) |
| $rk$ | Rank of a matrix |
| $c$ | Rank of a low-rank reconstructed data matrix |
| Prob | Generic symbol for probability |
| $\mathcal{P}$ | Probability of realizing dichotomy |
| $SNR$ | Signal-to-noise ratio |
| $\text{Var}(\sigma^\nu g^\nu)$ | Noise (denominator) of the $SNR$ |
| $Q^2$ | Excess overlaps |
| $\{0, 1\}$ | Short-hand to denote sign non-linearity |
| $\{+, -\}$ | Short-hand to denote Heaviside non-linearity |
| $\|\| \ \|\|_2$ | L2 norm |

## 6.2 Mathematical conventions and standard integrals

A standard notation used in this thesis for the Heaviside non-linearity is $\theta(x)$, defined as

$$\theta(x) = \begin{cases} 1, & \text{if } x > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

From that it follows that the sign non-linearity is given by $2\theta(x) - 1$.
Standard expression used for the Gaussian pdf is

$$G(x) = \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}} \tag{6.2}$$

And for the cummulative Gaussian

$$H(x) = \int_x^\infty \frac{e^{\frac{-z^2}{2}}}{\sqrt{2\pi}} dz \tag{6.3}$$

There are two standard integrals used in this thesis [69]. The first standard integral used in this text is

$$\int z^2 \frac{e^{\frac{-z^2}{2}}}{\sqrt{2\pi}} dz = zG(z) + H(z) \tag{6.4}$$

And the second one used is

$$\int z^4 \frac{e^{\frac{-z^2}{2}}}{\sqrt{2\pi}} dz = 3(1 - H(z)) - (z^3 + 3z)G(z) \tag{6.5}$$

Finally, we have the combinatorics of Cover's theorem, for the number of dichotomies realized by $P$ linear constraints in $N$ dimensions, which is [25]:

$$C(P, N) = 2 \sum_{i=0}^{N-1} \binom{P-1}{i} \tag{6.6}$$

which is evaluated numerically whenever Cover's theorem is used.

# Bibliography


[1] James S Albus. "The Marr and Albus Theories of the Cerebellum: Two Eary Models of Associative Memory". In: (1989).

[2] Wael F Asaad, Gregor Rainer, and Earl K Miller. "Neural activity in the primate prefrontal cortex during associative learning". In: *Neuron* 21.6 (1998), pp. 1399–1407.

[3] Baktash Babadi and Haim Sompolinsky. "Sparseness and expansion in sensory representations". In: *Neuron* 83.5 (2014), pp. 1213–1226.

[4] Carlo Baldassi, Enrico M Malatesta, and Riccardo Zecchina. "On the geometry of solutions and on the capacity of multi-layer neural networks with ReLU activations". In: *arXiv preprint arXiv:1907.07578* (2019).

[5] Omri Barak, Mattia Rigotti, and Stefano Fusi. "The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off". In: *Journal of Neuroscience* 33.9 (2013), pp. 3844–3856.

[6] Carol A Barnes et al. "Hippocampal synaptic enhancement as a basis for learning and memory: a selected review of current evidence from behaving animals". In: *Brain and memory: Modulation and mediation of neuroplasticity*. Oxford University Press New York, 1995, pp. 259–276.

[7] Robert J van Beers, Anne C Sittig, and Jan J van der Gon Denier. "How humans combine simultaneous proprioceptive and visual position information". In: *Experimental brain research* 111.2 (1996), pp. 253–261.

[8] John M Bekkers and Norimitsu Suzuki. "Neurons and circuits for odor processing in the piriform cortex". In: *Trends in neurosciences* 36.7 (2013), pp. 429–438.

[9] Curtis C Bell. "An efference copy which is modified by reafferent input". In: *Science* 214.4519 (1981), pp. 450–453.

[10] Curtis C Bell, Victor Han, and Nathaniel B Sawtell. "Cerebellum-like structures and their implications for cerebellar function". In: *Annu. Rev. Neurosci.* 31 (2008), pp. 1–24.

[11] Guy Billings et al. "Network structure within the cerebellar input layer enables lossless sparse encoding". In: *Neuron* 83.4 (2014), pp. 960–974.

[12] Matthew M Botvinick. "Hierarchical models of behavior and prefrontal function". In: *Trends in cognitive sciences* 12.5 (2008), pp. 201–208.

[13] Nicolas Brunel. "Is cortical connectivity optimized for storing information?" In: *Nature neuroscience* 19.5 (2016), p. 749.

[14] Shawn D Burton. "Inhibitory circuits of the mammalian main olfactory bulb". In: *Journal of neurophysiology* 118.4 (2017), pp. 2034–2051.

[15] Emmanuel J Candes and Terence Tao. "Near-optimal signal recovery from random projections: Universal encoding strategies?" In: *IEEE transactions on information theory* 52.12 (2006), pp. 5406–5425.

[16] N Alex Cayco-Gajic, Claudia Clopath, and R Angus Silver. "Sparse synaptic connectivity is required for decorrelation and pattern separation in feedforward networks". In: *Nature communications* 8.1 (2017), p. 1116.

[17] N Alex Cayco-Gajic and R Angus Silver. "Re-evaluating circuit mechanisms underlying pattern separation". In: *Neuron* 101.4 (2019), pp. 584–602.

[18] Julio Chapeton et al. "Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons". In: *Proceedings of the National Academy of Sciences* 109.51 (2012), E3614–E3622.

[19] Rishidev Chaudhuri and Ila Fiete. "Computational principles of memory". In: *Nature neuroscience* 19.3 (2016), p. 394.

[20] SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. "Classification and geometry of general perceptual manifolds". In: *Physical Review X* 8.3 (2018), p. 031003.

[21] Mark M Churchland et al. "Neural population dynamics during reaching". In: *Nature* 487.7405 (2012), pp. 51–56.

[22] Claudia Clopath and Nicolas Brunel. "Optimal properties of analog perceptrons with excitatory weights". In: *PLoS computational biology* 9.2 (2013).

[23] Conor Dempsey, Larry F Abbott, and Nathaniel B Sawtell. "Generalization of learned responses in the mormyrid electrosensory lobe". In: *eLife* 8 (2019), e44032.

[24] John C Eccles. "Circuits in the cerebellar control of movement." In: *Proceedings of the National Academy of Sciences of the United States of America* 58.1 (1967), p. 336.

[25] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[26] Randall W Engle. "Working memory capacity as executive attention". In: *Current directions in psychological science* 11.1 (2002), pp. 19–23.

[27] Armen G Enikolopov, LF Abbott, and Nathaniel B Sawtell. "Internally generated predictions enhance neural and behavioral detection of sensory stimuli in an electric fish". In: *Neuron* 99.1 (2018), pp. 135–146.

[28] Christopher R Fetsch et al. "Dynamic reweighting of visual and vestibular cues during self-motion perception". In: *Journal of Neuroscience* 29.49 (2009), pp. 15601–15612.

[29] Arseny Finkelstein et al. "Optimal dynamic coding by mixed-dimensionality neurons in the head-direction system of bats". In: *Nature communications* 9.1 (2018), p. 3590.

[30] Arseny Finkelstein et al. "Three-dimensional head-direction coding in the bat brain". In: *Nature* 517.7533 (2015), p. 159.

[31] Stefano Fusi, Earl K Miller, and Mattia Rigotti. "Why neurons mix: high dimensionality for higher cognition". In: *Current opinion in neurobiology* 37 (2016), pp. 66–74.

[32] Jia-Hong Gao et al. "Cerebellum implicated in sensory acquisition and discrimination rather than motor control". In: *Science* 272.5261 (1996), pp. 545–547.

[33] Peiran Gao et al. "A theory of multineuronal dimensionality, dynamics and measurement". In: *BioRxiv* (2017), p. 214262.

[34] Elizabeth Gardner. "The space of interactions in neural network models". In: *Journal of physics A: Mathematical and general* 21.1 (1988), p. 257.

[35] Elizabeth Gardner and Bernard Derrida. "Optimal storage properties of neural network models". In: *Journal of Physics A: Mathematical and general* 21.1 (1988), p. 271.

[36] Paul E Gilbert, Raymond P Kesner, and Inah Lee. "Dissociating hippocampal subregions: A double dissociation between dentate gyrus and CA1". In: *Hippocampus* 11.6 (2001), pp. 626–636.

[37] Andrea Giovannucci et al. "Cerebellar granule cells acquire a widespread predictive feedback signal during motor learning". In: *Nature neuroscience* 20.5 (2017), p. 727.

[38] Sonya Giridhar, Brent Doiron, and Nathaniel N Urban. "Timescale-dependent shaping of correlation by olfactory bulb lateral inhibition". In: *Proceedings of the National Academy of Sciences* 108.14 (2011), pp. 5843–5848.

[39] Robert Gütig and Haim Sompolinsky. "The tempotron: a neuron that learns spike timing–based decisions". In: *Nature neuroscience* 9.3 (2006), pp. 420–428.

[40] Kiah Hardcastle et al. "A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex". In: *Neuron* 94.2 (2017), pp. 375–387.

[41] Jarvis Haupt and Robert Nowak. "Signal reconstruction from noisy random projections". In: *IEEE Transactions on Information Theory* 52.9 (2006), pp. 4036–4048.

[42] Martin Heisenberg. "What do the mushroom bodies do for the insect brain? An introduction". In: *Learning & Memory* 5.1 (1998), pp. 1–10.

[43] Louis M Herman, Adam A Pack, and Matthias Hoffmann-Kuhnt. "Seeing through sound: dolphins (Tursiops truncatus) perceive the spatial structure of objects through echolocation." In: *Journal of Comparative Psychology* 112.3 (1998), p. 292.

[44] John A Hertz. *Introduction to the theory of neural computation*. CRC Press, 2018.

[45] Junya Hirokawa et al. "Frontal cortex neuron types categorically encode single decision variables". In: *Nature* 576.7787 (2019), pp. 446–451.

[46] John J Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the national academy of sciences* 79.8 (1982), pp. 2554–2558.

[47] Cheng-Chiu Huang et al. "Convergence of pontine and proprioceptive streams onto multimodal cerebellar granule cells". In: *Elife* 2 (2013), e00400.

[48] Taro Ishikawa, Misa Shimuta, and Michael Häusser. "Multimodal sensory integration in single cerebellar granule cells in vivo". In: *Elife* 4 (2015), e12916.

[49] W Jeffrey Johnston, Stephanie E Palmer, and David J Freedman. "Nonlinear mixed selectivity supports reliable neural computation". In: *PLOS Computational Biology* 16.2 (2020), e1007544.

[50] Henrik Jörntell and Carl-Fredrik Ekerot. "Properties of somatosensory synaptic integration in cerebellar granule cells in vivo". In: *Journal of Neuroscience* 26.45 (2006), pp. 11786–11797.

[51] Jonathan Kadmon and Haim Sompolinsky. "Optimal architectures in a solvable model of deep networks". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4781–4789.

[52] Sung Soo Kim et al. "Ring attractor dynamics in the Drosophila central brain". In: *Science* 356.6340 (2017), pp. 849–853.

[53] Laura D Knogler et al. "Sensorimotor representations in cerebellar granule cells in larval zebrafish are dense, spatially organized, and non-temporally patterned". In: *Current Biology* 27.9 (2017), pp. 1288–1302.

[54] Bernd Kramer. "Electrocommunication in weakly electric fishes: review of signals sent and received". In: *Journal of Comparative Physiology A* 173 (1993), pp. 719–722.

[55] Jill K Leutgeb et al. "Pattern separation in the dentate gyrus and CA3 of the hippocampus". In: *science* 315.5814 (2007), pp. 961–966.

[56] Grace W Lindsay et al. "Hebbian learning in a random network captures selectivity properties of the prefrontal cortex". In: *Journal of Neuroscience* 37.45 (2017), pp. 11021–11036.

[57] Ashok Litwin-Kumar et al. "Optimal degrees of synaptic connectivity". In: *Neuron* 93.5 (2017), pp. 1153–1164.

[58] David Marr and W Thomas Thach. "A theory of cerebellar cortex". In: *From the Retina to the Neocortex*. Springer, 1991, pp. 11–50.

[59] Liria M Masuda-Nakagawa, Nobuaki K Tanaka, and Cahir J O'Kane. "Stereotypic and random patterns of connectivity in the larval mushroom body calyx of Drosophila". In: *Proceedings of the National Academy of Sciences* 102.52 (2005), pp. 19027–19032.

[60] Thomas J McHugh et al. "Dentate gyrus NMDA receptors mediate rapid pattern separation in the hippocampal network". In: *Science* 317.5834 (2007), pp. 94–99.

[61] Miriam LR Meister, Jay A Hennig, and Alexander C Huk. "Signal multiplexing and single-neuron computations in lateral intraparietal area during decision-making". In: *Journal of Neuroscience* 33.6 (2013), pp. 2254–2267.

[62] Raoul-Martin Memmesheimer et al. "Learning precisely timed spikes". In: *Neuron* 82.4 (2014), pp. 925–938.

[63] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Publishing Company, 1987.

[64] Earl K Miller and Jonathan D Cohen. "An integrative theory of prefrontal cortex function". In: *Annual review of neuroscience* 24.1 (2001), pp. 167–202.

[65] Earl K Miller, David J Freedman, and Jonathan D Wallis. "The prefrontal cortex: categories, concepts and cognition". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 357.1424 (2002), pp. 1123–1136.

[66] Andrea M Morris et al. "The role of the dentate gyrus in the formation of contextual representations". In: *Hippocampus* 23.2 (2013), pp. 162–168.

[67] Edvard I Moser, Emilio Kropff, and May-Britt Moser. "Place cells, grid cells, and the brain's spatial representation system". In: *Annu. Rev. Neurosci.* 31 (2008), pp. 69–89.

[68] John O'Keefe and Jonathan Dostrovsky. "The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat." In: *Brain research* (1971).

[69] Cengiz Pehlevan. Personal notes.

[70] L Personnaz, I Guyon, and G Dreyfus. "Information storage and retrieval in spin-glass like neural networks". In: *Journal de Physique Lettres* 46.8 (1985), pp. 359–365.

[71] Cindy Poo et al. "Spatial maps in olfactory cortex during olfactory navigation". In: *bioRxiv* (2020).

[72] Gayathri N Ranganathan et al. "Active dendritic integration and mixed neocortical network representations during an adaptive sensing behavior". In: *Nature neuroscience* 21.11 (2018), pp. 1583–1590.

[73] Mattia Rigotti et al. "The importance of mixed selectivity in complex cognitive tasks". In: *Nature* 497.7451 (2013), p. 585.

[74] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386.

[75] Ran Rubin, LF Abbott, and Haim Sompolinsky. "Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity". In: *Proceedings of the National Academy of Sciences* 114.44 (2017), E9366–E9375.

[76] Ran Rubin, Rémi Monasson, and Haim Sompolinsky. "Theory of spike timing-based neural classifiers". In: *Physical review letters* 105.21 (2010), p. 218102.

[77] Toshiro Sakamoto and Shogo Endo. "Deep cerebellar nuclei play an important role in two-tone discrimination on delay eyeblink conditioning in C57BL/6 mice". In: *PloS one* 8.3 (2013).

[78] Inés Samengo and Alessandro Treves. "Representational capacity of a set of independent neurons". In: *Physical Review E* 63.1 (2000), p. 011910.

[79] Francesca Sargolini et al. "Conjunctive representation of position, direction, and velocity in entorhinal cortex". In: *Science* 312.5774 (2006), pp. 758–762.

[80] Nathaniel B Sawtell. "Neural mechanisms for predicting the sensory consequences of behavior: insights from electrosensory systems". In: *Annual review of physiology* 79 (2017), pp. 381–399.

[81] Takashi Shinzato and Yoshiyuki Kabashima. "Perceptron capacity revisited: classification ability for correlated patterns". In: *Journal of Physics A: Mathematical and Theoretical* 41.32 (2008), p. 324013.

[82] Shobhit Singla et al. "A cerebellum-like circuit in the auditory system cancels responses to self-generated sounds". In: *Nature neuroscience* 20.7 (2017), p. 943.

[83] Anton Spanne and Henrik Jörntell. "Processing of multi-dimensional sensorimotor information in the spinal and cerebellar neuronal circuitry: a new hypothesis". In: *PLoS computational biology* 9.3 (2013).

[84] Dan D Stettler and Richard Axel. "Representations of odor in the piriform cortex". In: *Neuron* 63.6 (2009), pp. 854–864.

[85] Mikhail V Tsodyks and Mikhail V Feigel'man. "The enhanced storage capacity in neural networks with low activity level". In: *EPL (Europhysics Letters)* 6.2 (1988), p. 101.

[86] Glenn C Turner, Maxim Bazhenov, and Gilles Laurent. "Olfactory representations by Drosophila mushroom body neurons". In: *Journal of neurophysiology* 99.2 (2008), pp. 734–746.

[87] Toby Tyrrell and David Willshaw. "Cerebellar cortex: its simulation and the relevance of Marr's theory". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 336.1277 (1992), pp. 239–257.

[88] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[89] Melissa R Warden and Earl K Miller. "Task-dependent changes in short-term memory in the prefrontal cortex". In: *Journal of Neuroscience* 30.47 (2010), pp. 15801–15810.

[90] Michael L Waskom et al. "Frontoparietal representations of task context support the flexible control of goal-directed cognition". In: *Journal of Neuroscience* 34.32 (2014), pp. 10743–10755.

[91] Gayle M Wittenberg and Samuel S-H Wang. "Evolution and scaling of dendrites". In: *Dendrites. Oxford University Press, New York* (2007), pp. 43–67.

[92] Daniel M Wolpert, R Chris Miall, and Mitsuo Kawato. "Internal models in the cerebellum". In: *Trends in cognitive sciences* 2.9 (1998), pp. 338–347.

[93] Ryosuke Yagi et al. "Convergence of multimodal sensory pathways to the mushroom body calyx in Drosophila melanogaster". In: *Scientific reports* 6.1 (2016), pp. 1–8.

[94] Michael M Yartsev, Menno P Witter, and Nachum Ulanovsky. "Grid cells without theta oscillations in the entorhinal cortex of bats". In: *Nature* 479.7371 (2011), p. 103.

[95] Mineto Yokoi, Kensaku Mori, and Shigetada Nakanishi. "Refinement of odor molecule tuning by dendrodendritic synaptic inhibition in the olfactory bulb." In: *Proceedings of the National Academy of Sciences* 92.8 (1995), pp. 3371–3375.

[96] Carey Y Zhang et al. "Partially mixed selectivity in human posterior parietal association cortex". In: *Neuron* 95.3 (2017), pp. 697–708.