




2017

# Optimal Adaptation Principles In Neural Systems

Kamesh Krishnamurthy

*University of Pennsylvania*, [kameshkk@gmail.com](mailto:kameshkk@gmail.com)

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Artificial Intelligence and Robotics Commons](#), [Biophysics Commons](#), and the [Neuroscience and Neurobiology Commons](#)

---

## Recommended Citation

Krishnamurthy, Kamesh, "Optimal Adaptation Principles In Neural Systems" (2017). *Publicly Accessible Penn Dissertations*. 2834.  
<https://repository.upenn.edu/edissertations/2834>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/2834>  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

# Optimal Adaptation Principles In Neural Systems

## **Abstract**

Animal brains are remarkably efficient in handling complex computational tasks, which are intractable even for state-of-the-art computers. For instance, our ability to detect visual objects in the presence of substantial variability and clutter surpasses any algorithm. This ability seems even more surprising given the noisiness and biophysical constraints of neural circuits. This thesis focuses on understanding the theoretical principles governing how neural systems, at various scales, are adapted to the structure of their environment in order to interact with it and perform information processing tasks efficiently. Here, we study this question in three very different and challenging scenarios: i) how a sensory neural circuit the olfactory pathway is organised to efficiently process odour stimuli in a very high-dimensional space with complex structure; ii) how individual neurons in the sensory periphery exploit the structure in a fast-changing environment to utilise their dynamic range efficiently; iii) how the auditory system of whole organisms is able to efficiently exploit temporal structure in a noisy, fast-changing environment to optimise perception of ambiguous sounds. We also study the theoretical issues in developing principled measures of model complexity and extending classical complexity notions to explicitly account for the scale/resolution at which we observe a system.

## **Degree Type**

Dissertation

## **Degree Name**

Doctor of Philosophy (PhD)

## **Graduate Group**

Neuroscience

## **First Advisor**

Vijay Balasubramanian

## **Second Advisor**

Joshua I. Gold

## **Keywords**

Biophysical modelling, Complex systems, Machine learning, Theoretical neuroscience

## **Subject Categories**

Artificial Intelligence and Robotics | Biophysics | Neuroscience and Neurobiology

OPTIMAL ADAPTATION PRINCIPLES IN NEURAL SYSTEMS

Kamesh Krishnamurthy

A DISSERTATION

in

Neuroscience

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Co-Supervisor of Dissertation

Co-Supervisor of Dissertation

---

Vijay Balasubramanian  
Cathy and Marc Lasry Professor of Physics

---

Joshua I. Gold  
Professor of Neuroscience

Graduate Group Chairperson

---

Joshua I. Gold  
Professor of Neuroscience

Dissertation Committee:

Minghong Ma, Associate Professor of Neuroscience  
Philip C. Nelson, Professor of Physics  
Johannes Burge, Assistant Professor of Psychology

OPTIMAL ADAPTATION PRINCIPLES IN NEURAL SYSTEMS

© COPYRIGHT

2017

Kamesh Krishnamurthy

This work is licensed under the  
Creative Commons Attribution  
NonCommercial-ShareAlike 3.0  
License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/3.0/>

*To my parents, for their unconditional support*

*To Nina and Aleks, for all the wonderful memories*

## ABSTRACT

### OPTIMAL ADAPTATION PRINCIPLES IN NEURAL SYSTEMS

Kamesh Krishnamurthy

Vijay Balasubramanian and Joshua I. Gold

Animal brains are remarkably efficient in handling complex computational tasks, which are intractable even for state-of-the-art computers. For instance, our ability to detect visual objects in the presence of substantial variability and clutter surpasses any algorithm. This ability seems even more surprising given the noisiness and biophysical constraints of neural circuits. This thesis focuses on understanding the theoretical principles governing how neural systems, at various scales, are adapted to the structure of their environment in order to interact with it and perform information processing tasks efficiently. Here, we study this question in three very different and challenging scenarios: i) how a sensory neural circuit – the olfactory pathway – is organised to efficiently process odour stimuli in a very high-dimensional space with complex structure; ii) how individual neurons in the sensory periphery exploit the structure in a fast-changing environment to utilise their dynamic range efficiently; iii) how the auditory system of whole organisms is able to efficiently exploit temporal structure in a noisy, fast-changing environment to optimise perception of ambiguous sounds. We also study the theoretical issues in developing principled measures of model complexity and extending classical complexity notions to explicitly account for the scale/resolution at which we observe a system.

# TABLE OF CONTENTS

CHAPTER 1 : Introduction . . . . .	1
1.1 Adaptation to complex, high-dimensional structure of a stimulus space	2
1.2 Adaptation to temporal structure in a noisy, fast-changing environment	4
1.3 Model complexity and individual differences in adaptive behavior . .	6
CHAPTER 2 : Disorder and the neural representation of complex odors: smelling in the real world . . . . .	8
2.1 Abstract . . . . .	8
2.2 Introduction . . . . .	9
2.3 Olfactory receptor neurons use disorder to encode natural odors . . .	11
2.4 The glomerular transformation increases disorder in response patterns	15
2.5 Discussion . . . . .	22
2.6 Supplementary Information . . . . .	27
2.7 Linearization of the Antennal Lobe transformation . . . . .	37
2.8 Addendum: background on random projections and compressive-sensing	42
CHAPTER 3 : Arousal-related adjustments of perceptual biases optimize per- ception in dynamic environments . . . . .	66
3.1 Abstract . . . . .	66
3.2 Introduction . . . . .	67
3.3 Results . . . . .	69
3.4 Discussion . . . . .	86
3.5 Experimental Procedures . . . . .	91

CHAPTER 4 : Model complexity, information geometry and resolution of ob-	
servations . . . . .	107
4.1 Introduction: principled measures of model complexity . . . . .	107
4.2 Model selection and classical measures of complexity . . . . .	109
4.3 Empirical complexity from Predictive Information . . . . .	115
4.4 Resolution of observations, <i>sloppy</i> models and complexity . . . . .	126
Bibliography . . . . .	138



## LIST OF ILLUSTRATIONS

FIGURE 2.1 : Schematic of the olfactory pathway . . . . .	12
FIGURE 2.2 : Disordered sensing by ORNs enables accurate decoding of complex mixtures . . . . .	16
FIGURE 2.3 : Divisive normalization in the Antennal Lobe increases disorder and decodability by <i>densifying</i> and <i>decorrelating</i> responses	19
FIGURE 2.4 : Disordered projections enable flexible learning in the presence of noise . . . . .	23
FIGURE 2.5 : odor decoding from <i>Drosophila</i> ORN responses is robust . . .	30
FIGURE 2.6 : Weakly responding ORNs and glomeruli are informative about odor mixture composition . . . . .	32
FIGURE 2.7 : The empirically determined divisive normalization in the Antennal Lobe is optimal for the measured ORN sensing matrix	35
FIGURE 2.8 : MB decoding error . . . . .	36
FIGURE 2.9 : Linearization of the Antennal Lobe normalization . . . . .	41
FIGURE 2.10 : Geometric illustration of the pseudo-inverse solution . . . . .	43
FIGURE 2.11 : Logan-Shepp phantom . . . . .	44
FIGURE 2.12 : Logan-Shepp reconstruction . . . . .	47
FIGURE 2.13 : The $L_1$ ball . . . . .	49
FIGURE 2.14 : RIP and distance preservation . . . . .	58
FIGURE 2.15 : Random projections and stable embeddings . . . . .	63
FIGURE 3.1 : Sound localization task . . . . .	70
FIGURE 3.2 : Overall prediction and estimation performance . . . . .	72
FIGURE 3.3 : Effects of task dynamics on performance . . . . .	74

FIGURE 3.4 : Effects of task dynamics on perceptual bias . . . . .	77
FIGURE 3.5 : Individual differences in perceptual bias . . . . .	79
FIGURE 3.6 : Dynamic modulation of perceptual bias by normative and non-normative factors . . . . .	81
FIGURE 3.7 : Pupil diameter reflects dynamic modulations of perceptual bias within individual subjects . . . . .	84
FIGURE 3.8 : Pupil diameter reflects individual differences in perceptual biases . . . . .	87
FIGURE 3.9 : Bayesian model of perceptual inference . . . . .	101
FIGURE 3.10 :Pupil regression goodness of fit . . . . .	104
FIGURE 3.11 :Pupil diameter predicts perceptual bias . . . . .	106
FIGURE 4.1 : Empirical Fisher information and robustness: . . . . .	114
FIGURE 4.2 : <i>Sloppy</i> and <i>stiff</i> directions . . . . .	128
FIGURE 4.3 : Model manifolds with finite and infinite data . . . . .	130
FIGURE 4.4 : Model selection with a sloppy model . . . . .	131
FIGURE 4.5 : Sloppy model selection . . . . .	136

# Chapter 1

## Introduction

Our brains are remarkably efficient in handling complex computational tasks, which are intractable even for state-of-the-art computer algorithms. For instance, our ability to rapidly detect visual objects in the presence of substantial variability and clutter surpasses any algorithm[1]. This ability seems even more surprising given the constraints faced by the neural circuits performing these computations; for instance, the timescale of computation by neurons is much slower than what can be implemented in silicon. Also, neural circuits *in vivo* exhibit substantial variability in their spiking activity: both the temporal dynamics of a single neuron's spiking and the response of the neuron to repeated presentations of a common scenario are highly irregular [2, 3]. This has been attributed to the activity of the neuron being driven by the fluctuations in its inputs rather than the mean input [4]. How are neural systems able to function robustly and efficiently in spite of these constraints?

A salient aspect of many neural systems – especially ones involved in sensory process-

ing – is that they are designed to exploit the structure in their natural environment. This idea has some experimental backing in simple systems such as single cells in the sensory periphery [5]. In this thesis, I study the principles by which neural systems, at different scales, exploit the structure in their environment to perform information processing tasks efficiently. I look at this problem of adaptation to environmental structure in two very different and challenging scenarios: i) how the olfactory system is organized to efficiently process odor stimuli in a very high-dimensional space with complex structure and ii) how the auditory system is able to efficiently exploit temporal structure in a noisy, fast-changing environment to optimize perception of sounds. I also develop a quantitative framework based on principled measures of model complexity to explain the individual differences in adaptive behavior as arising from different complexities of internal models.

## **1.1 Adaptation to complex, high-dimensional structure of a stimulus space**

The first example I consider is the adaptation of the general-purpose olfactory pathway to the complex structure of its stimulus space. The space of ecologically relevant volatile molecules that an organism, like a fruit fly, encounters is typically very large; some estimates put the dimensionality of this space on the order of  $10^7$  [6]. Uncovering the physical dimensions of the odor space that are relevant for perception has turned out to be very challenging; unlike vision or audition, where the physical dimensions responsible for perception of a simple light or sound stimulus are well characterized, it is very hard to predict how a new volatile molecule will smell just based on its chemical structure. In spite of these challenges in uncovering structure of the olfactory space, there is one salient aspect of natural odors which might make the

analysis easier: they are *sparse* in their composition – i.e., each odor typically contains a tiny fraction of all the possible volatile molecules. These molecules are sensed by a family of G-protein-coupled olfactory receptors, each of which is ‘hard-coded’ in the genome [7]. The number of types of receptors ranges from  $\sim 100$  to  $\sim 1000$  across a range of organisms. A notable characteristic of these receptors is that each receptor typically responds with varying intensities to a large fraction of molecules presented to it. This sort of ‘disordered and diffuse sensitivity’ has been typically thought to be due to biophysical limitations of making specific sensors, and the weak responses are assumed to be uninformative. However, the large fraction of the genome devoted to encode these receptors suggests that they might have evolved specially to have diffuse sensitivities. Here, we show that such broad and diffuse sensitivity is optimal to sense the space of natural odours with a sparse structure. We draw on results from theory of random projections to show that the sensing by receptors efficiently embeds a high-dimensional space with a sparse structure into a lower dimensional space, by comparing experimentally measured responses to several benchmark models of sensing. Our results also suggest that the weak responses are indeed informative, and it is not the precise specificities of receptors to certain molecules that matters, but it is the overall distribution of responses in the population. Further, we show that the random and expansive projections which transform the representation of odors at the receptor stage, are ideally suited to enable learning flexible associations between stimuli and behaviors. Thus, the disorder observed in various parts of the olfactory pathway – in the receptor responses and later on in the projections – may efficiently enable learning flexible associations between stimuli in a complex high-dimensional space and behaviors. This use of disorder by the olfactory system is in contrast to how other sensory systems use structured connectivity and responses to extract stimulus features relevant for perception.

## **1.2 Adaptation to temporal structure in a noisy, fast-changing environment**

The second scenario I consider is how systems exploit the temporal structure in noisy and fast-changing environments to perform information processing tasks efficiently. I study this problem in two systems of very different scales with different objectives: i) human subjects performing perceptual inference of dynamic, variable sounds and ii) individual retinal ganglion cells encoding visual stimuli from a noisy and dynamic environment. In the first case, human subjects construct dynamic ‘priors’ about the environmental structure to make accurate inferences about the noisy sensory inputs, and in the second case the cells dynamically adjust their response properties so as to use all their dynamic ranges efficiently to represent the stimuli. Both these cases require estimating the statistics of an ambiguous and fast-changing environment – a difficult problem in general. In this thesis, I will mainly present the first case of perceptual inference by human subjects and briefly mention the single cell case below.

### **1.2.1 Dynamic perceptual priors optimize perception of sounds**

It is well known that the expectation about an ambiguous stimulus can influence the perception of that stimulus. Several experiments have studied this phenomenon using the framework of Bayesian inference, where the uncertainty about the observed stimulus is encoded as the ‘likelihood’ distribution and the expectation about the stimulus is encoded in the prior distribution. Bayes rule tells us how to combine these sources of information in order to make an inference about the stimulus. As an example, it is well known that humans perceive objects moving in low contrast conditions as moving slower; this effect can be explained by our prior bias for lower

speeds – i.e., objects in the world don’t move around too much [8]. Although the Bayesian framework has been very successful in explain perceptual behavior, most of these experiments are done in stationary settings and it is unclear if and how these principles apply when the environment evolves rapidly.

Our work on dynamic auditory perception uses a combination of computational modelling, psychophysics and pupil analysis to show how human subjects are able to build and update appropriate priors encoding temporal structure in noisy, fast-changing environments, and how they use these priors efficiently to optimize perception of sounds. Furthermore, we provide evidence for a mechanism which might mediate the dynamic balance between prior expectations and (noisy) sensory information in guiding perception. We show that the moment-to-moment fluctuations in pupil diameter (a proxy for activity of Locus Coeruleus(LC) and related arousal areas) are predictive of fluctuations in the relative weight given to priors *beyond* what can be predicted by recent stimulus history and overall subject biases. This provides more evidence to the hypothesis that the activity of LC-related arousal areas might act as a dynamic gain control between external information and internal beliefs.

### **1.2.2 Dynamic rescaling by single sensory neurons optimize resource utilization**

*This project was conceived at the Methods in Computational Neuroscience summer school at Woods Hole and is not presented in this thesis. Details can be found in:*

K. Krishnamurthy, Wark, B., Fairhall, A. and J. Pillow,  
*Efficient coding with time-varying stimuli and noise.* Computational and Systems Neuroscience(CoSyNe), Salt Lake City, Utah, Feb. 2016.

An influential hypothesis in sensory physiology, proposed by Barlow et al. [9], posits that cells sensing their environments, with limited dynamic ranges, will organise their response functions so as to ‘efficiently’ use their entire dynamic range. Here ‘efficient’ is often interpreted to mean uniformly – i.e. all levels of the output response are used more or less equally. Framed in the language of Information Theory, this means that the response entropy of the cells are maximised, subject to resource constraints. Several experiments have confirmed this hypothesis by comparing the maximum-entropy response function predicted by the distribution of natural signals in the environment, to the actual response function of the cells (for e.g. [5]). However, the experiments and the theory itself are usually discussed in a steady-state setting where the environment does not change. This is in stark contrast to two empirical facts: i) the natural environments encountered by these cells is usually constantly evolving over several timescales and ii) experimentally, it is well known that the response functions of the cells change dynamically as the environmental statistics change. We extend the classical theory to dynamic settings by formulating a theory of dynamic efficient coding and answer the question “How should the response functions of cells optimally evolve?”. In forthcoming work, we are comparing the predictions of the theory in simple but common scenarios to the neural data from retinal ganglion cells.

### **1.3 Model complexity and individual differences in adaptive behavior**

In this section, I study the variability in adaptive behavior across individuals. Adaptive behavior in several sequential inference tasks is consistent with the prescriptions of optimal models; however, individual subjects show considerable variability in their strategies. One possibility is that this variability is due to uncontrolled factors. How-



ever, we consider the possibility that this variability can be explained by difference in model complexity used by different individuals. More specifically, to perform information processing tasks in noisy, fast-changing environments, the organism needs to have an internal model of the environment. We suggest that individual subjects might have a bias towards more or less complex models, but given that they chose a model with a certain complexity, they do the best they can with that model. Note that this need not be the case – one can form complex models of the environment which pick out features that are irrelevant to predicting unseen examples. To test this hypothesis, we first create a quantitative framework based on *predictive information* [10], to measure model complexity in a principled way. We then compare this measure of complexity to other principled notions of model complexity based on Information Geometry from the model selection literature. Both these notions of complexity are strictly valid for large datasets. We finally describe a notion of complexity that arises from effective/emergent models for small datasets. This notion of complexity is related to the phenomenon of *sloppiness* [11], and in forthcoming work we aim to make precise the links between the classical notions of model complexity and that which arises from sloppiness.

# Chapter 2

## Disorder and the neural representation of complex odors: smelling in the real world

Most of this section appears in:

K. Krishnamurthy\*, A.M. Hermundstad\*, T. Mora, A. Walczak and V. Balasubramanian  
*arXiv:1707.01962*

### 2.1 Abstract

Animals smelling in the real world use a small number of receptors to sense a vast number of natural molecular mixtures, and proceed to learn arbitrary associations between odors and valences. Here, we propose a new interpretation of how the architecture of olfactory circuits is adapted to meet these immense complementary challenges. First, the diffuse binding of receptors to many molecules compresses a

vast odor space into a tiny receptor space, while preserving similarity. Next, lateral interactions densify and decorrelate the response, enhancing robustness to noise. Finally, disordered projections from the periphery to the central brain reconfigure the densely packed information into a format suitable for flexible learning of associations and valences. We test our theory empirically using data from *Drosophila*. Our theory suggests that the neural processing of olfactory information differs from the other senses in its fundamental use of disorder.

## 2.2 Introduction

Animals sense and respond to volatile molecules that carry messages from and about the world. Some kinds of olfactory behaviors require sensing of particular molecules such as pheromones. These molecules and the receptors that bind to them have likely co-evolved over long periods of time to ensure precise and specific binding. However, to be useful as a general purpose tool for interaction with a diverse and changing world, the olfactory system should be prepared to sense and process any volatile molecule. There are a very large number of such monomolecular odorants (perhaps billions [6]), far more than the number of receptor types available to bind these odorants. Humans and mice, for instance, have just  $\sim 300$  and  $\sim 1000$  functional olfactory receptor types, respectively. Yet, animals may be able to discriminate between orders of magnitude more odors than the number of receptor types (a high estimate is given in [12], but see [13]).

At an abstract level, the early stage of the olfactory system faces the immense challenge of embedding a very high-dimensional input space (the space of odor molecules) into a low-dimensional space of sensors (the response space of olfactory receptors).

This embedding must preserve similarity between different odors well enough to permit the judgements of sameness and difference that are crucial for behavior. Furthermore, experiments [14] suggest that this odor representation is reorganized in higher brain regions to be enormously flexible, allowing learning of nearly arbitrary associations between valences and different groups of odors. Here, we propose a new theoretical framework (Fig. 2.1), and provide empirical evidence, suggesting that the olfactory system powerfully exploits physiological and structural *disorder* at different stages of processing to meet these two complementary challenges: (*i*) compression of a vast odor space into a tiny receptor space, and (*ii*) reorganization of the information to allow flexible learning.

To perform effectively within its design constraints, a sensory system must exploit structure in the environment. For example, the statistics of natural images dictate an efficient decomposition into edges [15], likely explaining why simple and complex cells in the visual cortex respond preferentially to oriented lines [16]. We noted [17] that a salient feature of natural odors is that they typically contain only a tiny fraction of the possible volatile molecular species. For example, food odors typically are composed of 3-40 molecules [6]. Natural odors are thus *sparse* in the high-dimensional space of odorant molecules. Surprising results from the mathematical literature on random projections [18, 19, 20] show that there is an efficient solution for storing signals of this nature: sparse, high-dimensional input signals can be encoded by a compact set of sensors through diffuse and disordered measurements of the input space. For example, this sort of compression can be achieved if each sensor response contains randomly weighted contributions from every dimension of the input space. Importantly, this diffuse sensing need not be tuned to the specific structure of the input signal – i.e. in this manner, it can be non-adaptive. We propose that the olfactory system employs

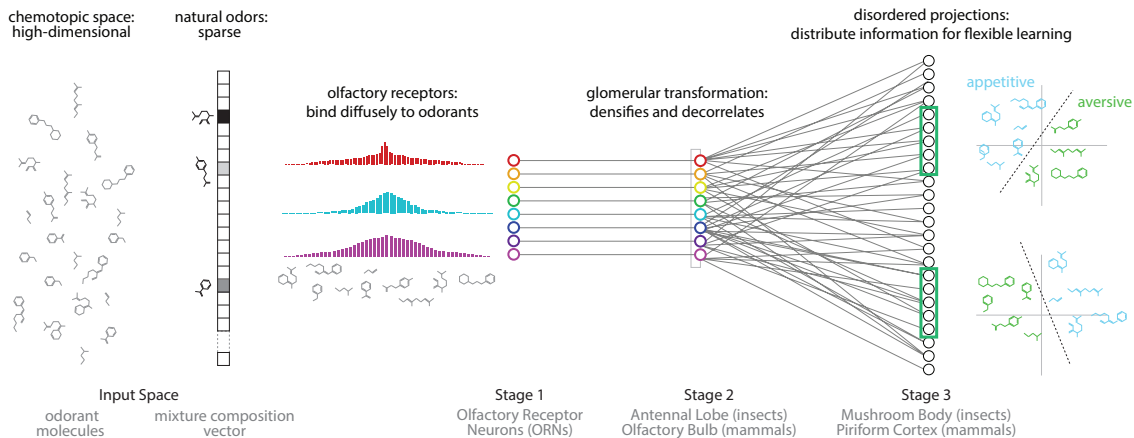
such a diffuse sensing strategy in order to exploit the sparse structure of natural odor space and produce compact representations of odors (Fig. 2.1).

Ultimately, these odor representations must support associations between odors and valence, and experimental evidence suggests that animals can learn such associations both flexibly and reversibly [14]. However, as we will show, the compact representations achieved by diffuse sensing make such learning difficult. We show that another form of disorder—a “densification” and decorrelation of responses, followed by a disordered expansion—can reorganize odor information into a format that facilitates flexible learning.

We provide evidence for our proposal by analyzing the olfactory system of *Drosophila*. We show that the diffuse responses of olfactory receptor neurons provide a compact representation of odor information. We then show that the nonlinear transformation in the second stage of olfactory processing (Antennal Lobe in insects; Olfactory Bulb in mammals), followed by the apparently disordered, expansive projection to the third stage of olfactory processing (Mushroom Body in insects; Piriform Cortex in mammals), facilitate flexible learning of odor categories from small and arbitrarily-chosen groups of sparsely firing neurons. Finally, we demonstrate that the disorder introduced by both the densification and the expansion is critical for robustness to noise.

## **2.3 Olfactory receptor neurons use disorder to encode natural odors**

Volatile molecules are sensed when they bind to olfactory receptors, each encoded by a separate gene [7]. For example, in mice, almost 5% of the genome is devoted



**Figure 2.1: Proposal: The olfactory system uses two kinds of disorder to first compress odor information into a small number of receptors, and then reconfigure this information to enable flexible associations between odors and valences.** (i) Natural odors are high dimensional but sparse: each one contains a tiny fraction of all possible monomolecular odorants. (ii) Olfactory receptors diffusely bind to a broad range of odorants, producing a compact representation of odor information that enables accurate decoding. (iii) The Antennal Lobe/Olfactory Bulb “densifies” and decorrelates this representation, providing robustness to noise. (iv) Disordered projections from the Antennal Lobe/Olfactory Bulb to the Mushroom Body/Piriform Cortex, followed by nonlinearities, create a sparse and distributed representation of odors that facilitates flexible learning of odor categories from small and arbitrarily-chosen subsets of neurons.

to encoding about 1000 receptor types. Despite such large genomic investments, the number of receptor types is dwarfed by the number of volatile molecules that a general purpose olfactory system might seek to sense. This raises two related questions. First, is it possible, even in principle, to sense the high-dimensional space of molecules using the inevitably low-dimensional space of receptor responses? Second, can this sensing be done by neurons so that odors with similar mixture compositions are mapped to nearby regions in response space?

To solve this problem, there is a key simplification that the nervous system could exploit – natural odors typically contain a tiny fraction of the possible volatile molecules [6]. Thus, the representation of a natural odor in terms of its molecular concentration vector is extremely sparse. Suppose there are  $N$  types of volatile molecules, and any given natural odor contains no more than  $K \ll N$  of these types. Then, recent

results in mathematics show that a small number of linear sensors (about  $K$ ) could store complete information about natural odors, provided that their binding affinities were statistically random [18, 19, 20]. This fact suggests a new perspective on the olfactory system: rather than having strong responses for a specific set of important molecules, a general purpose receptor repertoire should be selected to have molecular affinities that are as disordered as possible, subject to constraints imposed by biophysics and evolution. Likewise, the quality of olfaction as a general purpose sense will be determined by the degree of disorder in response patterns.

Is there evidence for this view? Indeed, most Olfactory Receptor Neuron (ORN) types respond diffusely to many odorants, and most odorants evoke diffuse responses from diverse ORN types (insect: [21, 22]; mammal: [23]). To assess the quality of the representation of natural odors in ORN responses, we analyzed firing rates of 24 ORN types in *Drosophila* responding to a panel of 110 monomolecular odorants [21]. We used this data to model responses to mixtures of odorants that are complex but sparse like natural odors. To do this, we constructed a firing rate “response matrix”  $R$  whose entries specify the responses of each ORN to each monomolecular odorant. We assumed that the ORN responses to odor mixtures are linear, which is a reasonable approximation at low concentrations [24]. This enabled us to define a complex mixture by a 110-dimensional composition vector  $\vec{x}$  whose entries specify the concentrations (measured relative to [21]) of monomolecular odorants in the mixture. The ORN firing rates  $\vec{y}$  can then be modeled as linear combinations of responses to monomolecular odorants:  $\vec{y} = R\vec{x}$ .

To construct each mixture composition vector  $\vec{x}$ , we set a small number  $K$  of its elements to be nonzero (where  $K$  specifies the complexity of the mixture). The values of these nonzero entries were chosen randomly and uniformly between 0 and 2. We

then attempted to decode composition vectors ( $\hat{x}$ ) from responses  $\vec{y}$  using an efficient algorithm for decoding linearly-combined sparse composition vectors [25, 19, 20]. We deemed the result a failure if the average squared difference between components of the decoded ( $\hat{x}$ ) versus original ( $\vec{x}$ ) composition vectors exceeded 0.01, and defined *decoding error* as the failure probability over an ensemble of 500 odor mixtures  $\{\vec{x}\}$ . This is a stringent criterion that we are using to quantify the accuracy with which mixture information is encoded in the ORN responses; there is no evidence to suggest that olfactory behavior requires this level of accuracy, nor do we assume that the brain uses this particular decoding scheme. We checked that our findings are robust to different choices of failure threshold used to assess decoding error (Fig. 2.5).

Fig. 2.2A shows the decoding error for varying mixture complexity  $K$  and numbers of ORN types. Performance improves with increasing number of ORNs and decreasing mixture complexity. We compared the decoding error obtained from the measured ORN responses to two idealized alternatives: (1) a Gaussian random model, in which each ORN responds randomly to different odorants (with the overall mean and variance matched to data), and (2) a generalized “labeled-line” model, in which each ORN responds (with the same strength) to only five randomly-selected odorants. The Gaussian random model would be an optimal strategy in the limit of many receptors and a large odor space [25], while the labeled line model is often considered to be a plausible interpretation of olfactory receptor responses. The *Drosophila* ORNs significantly outperform the labeled-line model and approach the performance of the Gaussian random model (Fig. 2.2C). Quantitatively, 67% of mixtures with 5 or fewer components drawn from 110 odorants can be accurately decoded from the responses of 24 receptors. There are a staggering 100 million such mixtures. Again, this is not to say that the fly brain attempts to reconstruct all of these odors with such an

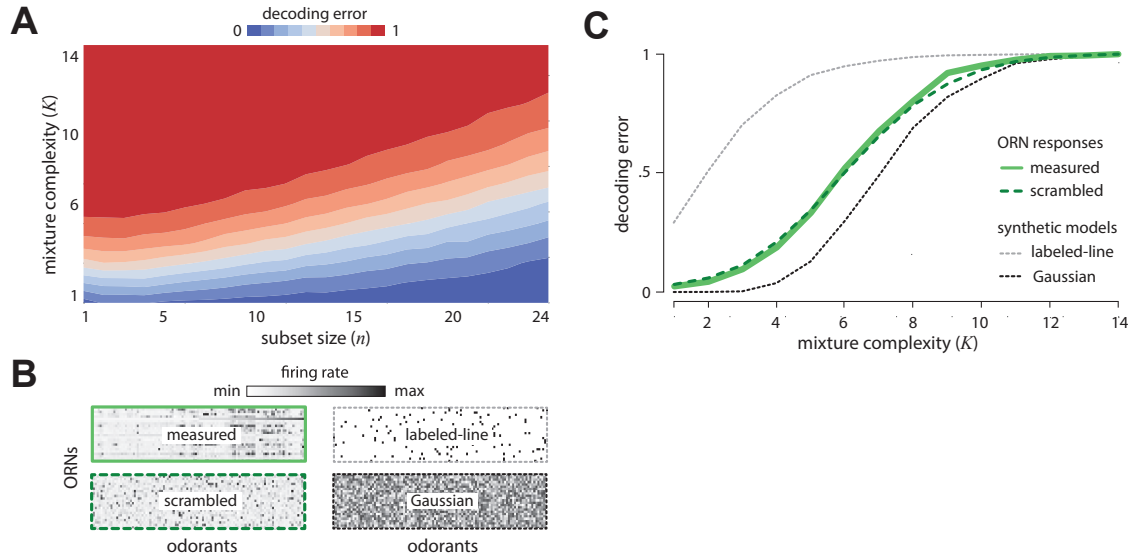


accuracy, but it does say that the receptors contain the necessary information. Our theory also predicts that the olfactory representation of odors does not depend on the details of how specific receptors respond to specific odors, but rather only depends on the broad distribution of responses across many receptors and many odors. We tested this prediction by scrambling the *Drosophila* response matrix (Fig. 2.2B) with respect to both odors and receptors and indeed found identical decoding performance (Fig. 2.2C).

Our theory predicts that the olfactory code spreads information across all receptors, so that even weak responses are informative. To test this comprehensively, we thresholded the *Drosophila* response matrix to keep only a fixed fraction of the strongest responses, and then scrambled the odor identities for each receptor to create receptor responses with the same thresholded distribution. As predicted by our theory, as this fraction varied from 0 to 1, decoding performance improved systematically (Fig. 2.5).

## 2.4 The glomerular transformation increases disorder in response patterns

Our theory suggests that disordered sensing — in which a single receptor binds to many odorants, and a single odorant binds to many receptors — is a powerful strategy for the olfactory system to employ. However, *Drosophila* ORN responses are noticeably structured and have a more clustered distribution of firing rates than, e.g., the Gaussian random model (Fig. 2.2B). These correlations, perhaps arising from similarities between odorant binding sites or between receptor proteins, induce some order in receptor responses. These responses are modified when receptors of each type converge to a second stage of processing in distinct glomeruli of the Antennal



**Figure 2.2: Disordered sensing by ORNs enables accurate decoding of complex mixtures.** (A) Error in decoding mixture composition from subsets of ORN responses, as a function of mixture complexity  $K$  (i.e. number of mixture components) and ORN subset size  $n$ . Results are averaged over 500 odor mixtures of a given complexity, and 50 subsets of a given size. (B) Response matrices for *Drosophila* ORNs (measured and scrambled), labeled-line and Gaussian models (see text for details). (C) Error in decoding complex mixtures from 24 ORNs as a function of mixture complexity  $K$ , shown for ORN responses (solid green), a scrambled version of ORN responses (dashed green), and two idealized models (the Gaussian random model, dashed black, and the labeled-line model, dashed gray). Results are averaged over 500 odor mixtures of a given complexity. Results from scrambled, Gaussian, and labeled-line models are additionally averaged over 100 model instantiations.

Lobe (analogously, the Olfactory Bulb in mammals). There, a network of inhibitory interneurons reorganizes the receptor responses for transmission downstream [26]. In the fly, the inhibitory network is well-described as effecting a divisive normalization [27, 28] that scales the responses of each ORN type in relation to the overall activity of all types (Appendix B). Applying this transformation to the *Drosophila* response matrix, we find that glomerular responses become more widely distributed and less correlated (Fig. 2.3A) than their ORN inputs. This *densification* and *decorrelation* increases disorder.

Does this increased disorder improve the representation of odor information? Because the divisive normalization is nonlinear, we cannot, strictly speaking, use the aforementioned decoding algorithm to evaluate the information content of the glomerular representation. However, we can instead create an artificial benchmark in which mixtures  $\vec{x}$  lead to responses  $\vec{y}$  via  $\vec{y} = R^{(2)}\vec{x}$ , where  $R^{(2)}$  represents a matrix of artificial glomerular responses obtained by transforming experimentally measured ORN responses to an odor panel in [21] via divisive normalization (see Appendix B). Quantitatively, 67% of mixtures with 7 or fewer components drawn from 110 odorants can be accurately decoded from the responses of 24 glomeruli, while similar accuracy was achieved for mixtures with only 5 components when decoding from ORNs (Fig. 2.3B). Because the number of possible mixtures increases combinatorially with the number of mixture components, this is a substantial improvement. A similar analysis shows that applying the divisive normalization to the labeled-line and Gaussian random models yields no improvement in decoding relative to the receptor stage (Fig. 2.3B).

As with decoding from ORNs, scrambling the responses over glomeruli and odors leads to identical decoding performance (Fig. 2.3B), again suggesting that only the broad distribution of responses is important for the odor representation. Weak responses

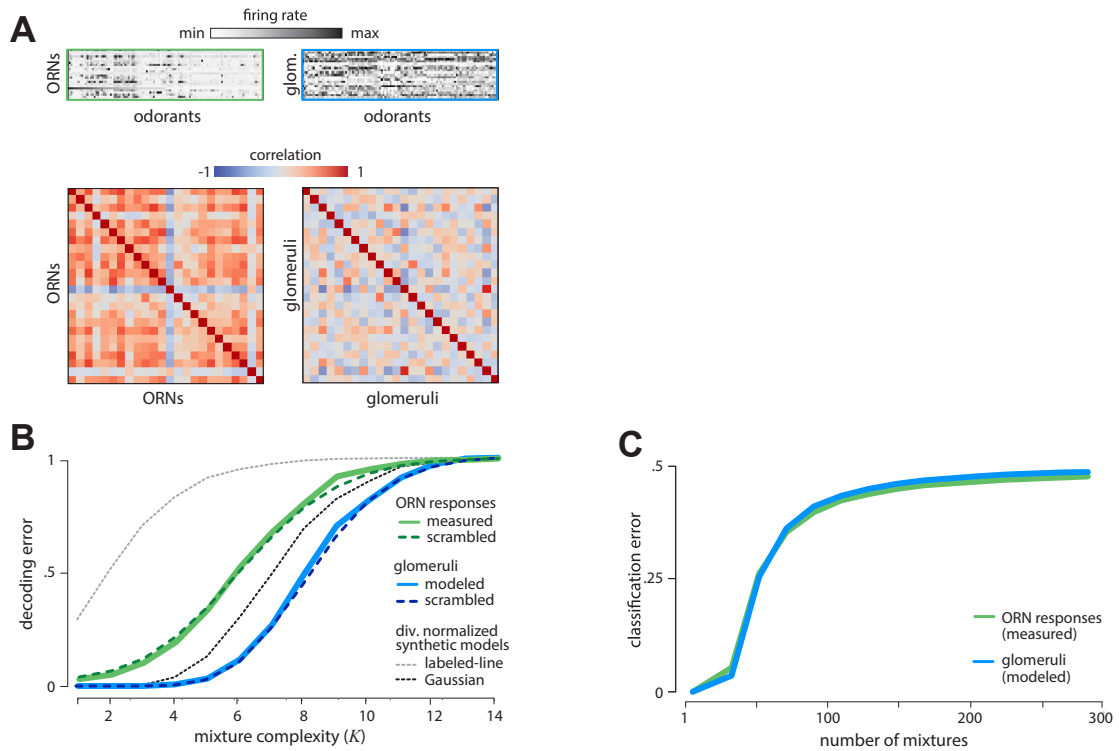
remain informative; we again find that thresholding the response matrix degrades performance (Fig. 2.5). Finally, we confirmed that our conclusions do not depend on details of the divisive normalization, but found, interestingly, that the experimentally-measured parameters [27] of this transformation minimize decoding error relative to other parameter choices (Fig. 2.6).

An alternative way of assessing the quality of a sensory representation is to ask how well it supports flexible associations between odors and valence. To this end, we randomly labeled mixtures “appetitive” or “aversive”, and we trained a linear classifier to identify these labels from ORN and fully nonlinear glomerular responses (Appendix C). Surprisingly, performance was poor (Fig. 3C), even though mixture compositions can be accurately decoded from these responses (Fig. 2.2C & 2.3B). We conclude that although these first stages of processing retain nearly complete information about odor mixtures, this information is not readily usable for learning.

### **2.4.1 Disordered projections reorganize odor information to facilitate flexible learning**

Although early stages of olfactory processing apparently do not support flexible learning, we know empirically that the representation at the third stage in the pathway *can* support such learning (fly: [29, 30]; mammal: [14]). How is odor information reorganized to achieve this?

In both insects and mammals, the transformation from the second to third stage of olfactory processing has two notable features: (i) expansive and disordered projections that distribute odor information across a large number of cells [31, 32], and (ii) nonlinearities that sparsify responses [33, 34]. As a result, an odor is represented by a



**Figure 2.3: Divisive normalization in the Antennal Lobe increases disorder and decodability by *densifying* and *decorrelating* responses.** (A) Divisive normalization of ORN responses (top left) distributes responses more widely and densely over glomeruli (top right). Correlation coefficients between glomeruli (lower right) are much lower than between ORNs (lower left). (B) The error in decoding from glomeruli (solid blue) is much lower than from ORNs (solid green), and is unchanged by scrambling (dashed blue). Divisive normalization has no effect on decoding error in the labeled-line model (dashed gray) or the Gaussian random model (dashed black). Results are averaged over 500 odor mixtures of a given complexity. Results from scrambled, Gaussian, and labeled-line models are additionally averaged over 100 model instantiations. (C) Responses of ORNs and glomeruli are not readily usable for classification tasks. As shown, error in classifying mixtures from responses of ORNs (green) and glomeruli (blue) quickly approaches chance as the number of mixtures increases. Results are shown for two-class separability of 5-component mixtures, averaged over 100 different ensembles of odor mixtures, and 100 labelings into appetitive and aversive classes.

sparse pattern of activity distributed broadly across cells in the third stage. We expect from general theory that this transformation should facilitate flexible associations between odor signals and valence [35, 36, 37, 38]. Here, we propose that two additional sources of disorder – densification achieved at earlier stages, and lack of structure in the connectivity patterns – allow such associations to be learned from small groups of neurons drawn arbitrarily from within the population.

To test this, we simulated the responses of Kenyon cells in the Mushroom Body of the fly to odor mixtures (Fig. 2.4A). We modeled each Kenyon cell as receiving inputs from 8 glomeruli selected at random, reflecting empirical estimates [31, 39] (interestingly, other choices yield worse performance; Fig. 2.7). Connection weights were drawn uniformly between 0 and 1 (Fig. 2.4B, left). We modeled long range inhibition by first removing the average response to an ensemble of odors, and then thresholding to eliminate weak responses (Appendix D, [38]). This imposed a tunable level of sparsity in the population response. We fixed this sparsity to 15% to match experimental estimates [34, 33]. To assess learning, we generated responses to an ensemble of 5-component odor mixtures (as described above), and trained a linear classifier to separate responses into two arbitrarily-assigned classes (Appendix C). We defined *classification error* to be the fraction of mixtures that are incorrectly labeled by the classifier, averaged over 100 ensembles of mixtures and 100 labelings of each ensemble into appetitive/aversive classes.

We first compared classification from Kenyon cell responses (Fig. 4C) to that from responses of ORNs or glomeruli (Fig. 2.3C). To directly compare these different stages, we selected random subsets of  $n = 160$  sparsely-active Kenyon cells. This ensured that any given odor would activate an average of 24 cells ( $0.15 \times 160$ ), matching the number of ORN and glomerulus types in our dataset. We found that a linear clas-

sifier trained on Kenyon cell responses could categorize up to 300 mixtures with less than 10% error (Fig. 2.4C), performing far better than a classifier trained on ORN or glomerular responses (Fig. 2.3C). In fact, even a much smaller population of  $n = 80$  Kenyon cells (with an average of 12 active cells per odor) yielded better classification performance than the complete ORN or glomerular populations. Moreover, any arbitrary subset of a given size was equivalent (histogram inset of Fig. 2.4C). When we increased the number of cells used as a readout or decreased the average sparsity of responses, we found no improvement in classification (Fig. 2.8).

We then examined the role of disorder on classification performance. To do this, we separately removed each source of disorder (densification at the Antennal Lobe, and disordered projections from the Antennal Lobe to the Mushroom Body). To examine the role of the densification at the Antennal Lobe, we projected responses directly from the ORNs to the Mushroom Body, rather than passing responses through the transformation at the Antennal Lobe. To examine the role of disordered projection patterns, we introduced local structure in the projections from the Antennal Lobe to each subset of Kenyon cells in the Mushroom Body (Fig. 2.4B, right). Within a given subset, we required that a fraction of Kenyon cells received preferential inputs from a fraction of glomeruli (in both cases, the fraction was taken to be  $1/3$ ). In doing so, we constrained the overall distribution of connection strengths to match those used to generate disordered connectivity (Appendix E). This ensured that as a whole, each subset of Kenyon cells sampled all glomeruli, and any differences in performance were guaranteed to arise purely from differences in local connectivity patterns.

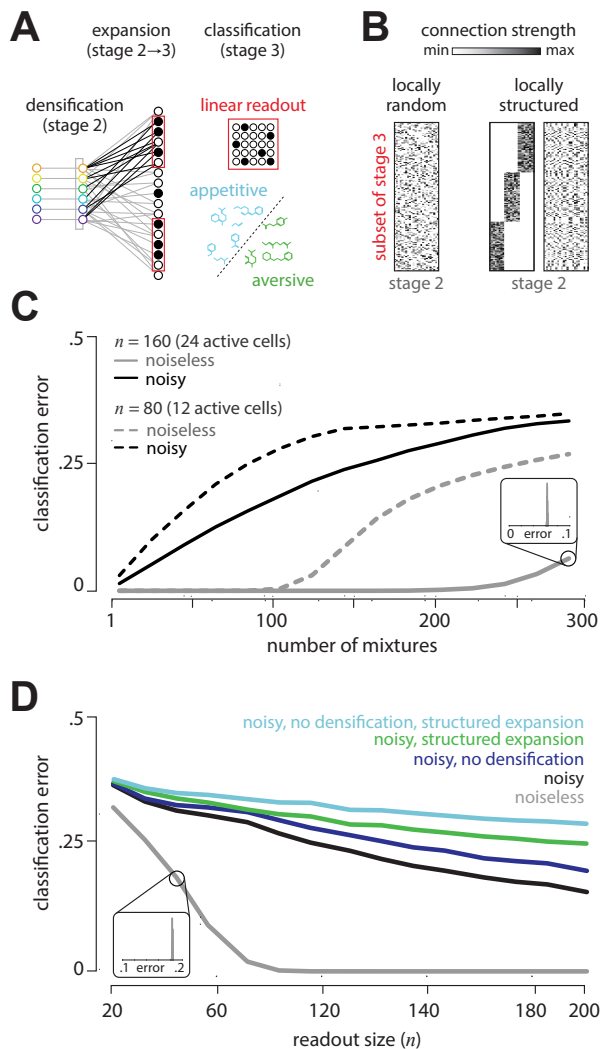
In the absence of neural variability, neither manipulation affected classification performance. However, both manipulations impacted performance in the presence of noise. To demonstrate this, we added proportional Gaussian noise of magnitude  $\eta\sqrt{ar}$  to

the firing rates  $r$  of each ORN, where  $\eta$  was drawn from a standard Gaussian and  $a = .25$  controlled the coefficient of variation. As expected, noise degraded performance (Fig. 2.4B,C). Surprisingly, the impact of noise was worse when either of the two sources of disorder was removed, and even more when both sources were removed (Fig. 2.4C). Taken together, these results suggest that the disorder in the connectivity and the densification at the Antennal Lobe aids in learning flexible associations at the Mushroom Body.

## 2.5 Discussion

We propose a new conceptual paradigm in sensory neuroscience: the use of *disorder* for building sensory representations that are accurate, compact, and flexible. We argue that this paradigm explains the organization and function of the olfactory system, where disorder plays two key roles: (i) diffuse sensing by olfactory receptors serves to compress high-dimensional odor signals into compact neural representations, and (ii) densification followed by disordered expansion serves to reformat these representations for flexible learning. This paradigm exploits a key feature of natural odor signals—sparsity—to overcome a bottleneck in the limited number of olfactory receptor types. We used a combination of data and modeling to provide evidence for this paradigm in fly. Olfactory circuits in mammals show very similar anatomical and functional motifs, including broad receptor tuning [23] and apparently disordered projections to the cortex [32]. This convergence between distant species suggests that disorder could provide a universal computational explanation for the architecture of early olfactory circuits.





**Figure 2.4: Disordered projections enable flexible learning in the presence of noise.** (A) Schematic. A linear readout neuron can learn to separate arbitrary classes of appetitive and aversive odor mixtures from Kenyon cell responses. (B) We generated random (left) and locally-structured (middle, right) projections from the Antennal Lobe (stage 2) to a subset of cells in the Mushroom Body (stage 3). Local structure was introduced by requiring that a fraction (1/3) of Kenyon cells receive inputs from a fraction (1/3) of all glomeruli (middle). When randomly permuted, the structure is no longer apparent (right). (C) Shown here is classification error for up to 300 different odor mixtures from subsets of 160 Kenyon cells (with 24 active cells; solid gray) or 80 Kenyon cells (with 12 active cells; dashed gray). All subsets of a given size produce nearly equivalent error (inset). Noise (introduced in ORN responses; see text) degrades classification performance (black). (D) Classification error decreases as readout subset size increases (gray, noiseless; black, noisy). Removing disorder—either by removing densification (blue), introducing structure in the connectivity to the Mushroom Body (green), or both (cyan)—reduces robustness to noise.

**The logic of olfactory receptors.** Our theory predicts that general-purpose olfactory receptors should be selected for diffuse binding to many odorants, and not for the strong and specific binding often seen in biochemical signaling. An alternative view suggests that receptors should be adapted to bind selectively to molecules in particular odor environments or ecological niches [22, 40]. These alternatives can be separated in experiments that measure the affinities of olfactory receptors to very large panels of odorants with varying ethological relevance. We predict that the typical receptor will have a diverse range of binding affinities across a broad array of odorants, with a statistically similar spread across molecules that both do and do *not* have immediate ethological importance. Likewise, we predict that receptors in different species, even related ones, will typically have broadly different distributions of binding affinities, with similarities arising from biophysical constraints of olfactory receptors and not from properties of ecological niches. In addition, as a whole, the receptor repertoires of different species will show similar coverage across the space of odorants. This strategy resembles that of well-adapted immune repertoires, where different antibody distributions achieve similar coverage of the same pathogen landscape, as predicted theoretically [41] and observed in experiment [42, 43].

**The computational role of expansive and disordered projections.** While this work provides evidence for the role of disordered sensing in the *compression* of odor information, it also adds to a growing body of work on the computational role of *expansion* via disordered neural projections. Expansive projections are known to make classification easier [35, 36, 38], and the computational benefits of this expansion can be further improved by Hebbian learning [37] and by sparse connectivity [39]. We have argued here that the primary purpose of the expansion from the second to the third stage of olfactory processing is to reorganize a highly compressed representation

of odors produced by disordered sensing by the receptors. By contrast, other studies have proposed that this expansion could itself implement a form of odor signal compression [17, 44], or even a direct encoding of odor space [45, 46] (in one case requiring unsupported assumptions about the mathematical relationship between the expansion and ORN responses [45]). We found no evidence that expansive projections implement a form of compression, nor do we find evidence to support the direct representation of odor composition in Kenyon cell responses. Rather, we found evidence that the expanded representation is organized to support flexible learning of categories [47, 14] from modest subsets of Kenyon cells. Anatomical evidence in fly indeed suggests that each olfactory readout neuron samples a only fraction of the Mushroom Body [48] while still allowing formation of complex associations [49]. Our view is also consistent with abstract theory showing that sparsely firing binary neurons with “mixed selectivity” permit both discrimination between, and effective generalization from, complex overlapping binary inputs[36, 50]. Our work can be viewed as additionally showing that *receptor* neurons with “mixed selectivity” effectively compress high dimensional sensory information, while subsequent “mixed *sampling*” of these responses reformats them for flexible learning by a simple readout.

**Implications for behavior.** Conceptually, our key idea is that disorder in the olfactory system is a fundamental adaptation to the intrinsic complexity of the world of smells. We predict, distinctively, that odor information is distributed in both weak and strong responses across the entire ensemble of olfactory receptor types, and that this is important for complex discrimination tasks. An alternative view suggests a “primacy” code where only the earliest or strongest responses are relevant for behavior [46]. We have shown (Fig. 2B and Fig. 2.6) that an encoding scheme that retains only the strongest responses contains much less information about complex

mixtures than does a scheme that retains both strong and weak responses. Because of this, we expect that our view can be separated from the primacy code in behavioral experiments that vary the complexity of discrimination tasks, e.g. by increasing the number of odors, the number of mixture components, and the degree of overlap between mixture components. Given knowledge of responses to individual odorants, our theory quantitatively predicts the decline of behavioral performance with task complexity (e.g., Figs. 2,3,4). Likewise, our theory predicts how the relationship between behavioral performance and task complexity will vary as a function of information content in the olfactory pathway. This information content can be experimentally manipulated by creating genetically-impooverished or enhanced receptor repertoires, optogenetically blocking inhibitory neurons in the Antennal Lobe to remove densification, or optogenetically activating Kenyon cells to simulate structured projection patterns from the Antennal Lobe.

**Looking ahead.** Testing these predictions requires a movement away from simple paradigms involving small mixtures and pairwise discrimination, towards far more complex tasks that are reflective of life in the real world. Methodologically, this shift has begun occurring in the study of vision. We have argued here that in olfaction, this shift is even more critical – the functional logic of the sense of smell can only be understood by taking into account the complexity of the real odor world.

## 2.6 Supplementary Information

### 2.6.1 Decoding odor composition

To reconstruct  $\vec{x}$  from measurements  $\vec{y} = R\vec{x}$ , we used the Iteratively Reweighted Least Squares (IRLS) algorithm [51] to find the vector that minimizes the  $L_1$  norm of  $\vec{x}$  subject to the constraint  $\vec{y} = R\vec{x}$ , with 500 maximum iterations and a convergence tolerance (in norm) of  $10^{-6}$ .

### 2.6.2 Divisive normalization in the Antennal Lobe

Lateral inhibition in the Antennal Lobe is believed to implement a form of divisive normalization [27, 28, 26]:

$$R_i^{(2)} = R_{max} \cdot (R_i^{(1)})^{1.5} / \left[ \sigma^{1.5} + (R_i^{(1)})^{1.5} + (m \cdot \sum_i R_i^{(1)})^{1.5} \right] \quad (2.1)$$

where  $R_i^{(1)}$  is the response of the  $i$ th ORN type,  $R_i^{(2)}$  is the response of the  $i$ th glomerulus,  $\sigma$  parametrizes spontaneous activity, and  $m$  controls the amount of normalization. We use  $R_{max} = 165.0$ ,  $\sigma = 10.5$ , and  $m = 0.05$  [27]. We constructed an artificial glomerular response matrix  $R^{(2)}$  by applying this transformation separately to the ORNs responding to each of the 110 odorants studied in [21]. Thus  $R_{ij}^{(2)}$  represented the response of the  $i$ th glomerulus to the  $j$ th odorant.

### 2.6.3 Linear classification

To measure how well a particular odor representation (responses of ORNs, glomeruli, or Kenyon cells) facilitates learning flexible associations between odors and valences, we randomly split the representation of input mixtures into two classes and then trained a linear classifier (SVM with linear kernel [52]) to classify the inputs.

### 2.6.4 Generating Mushroom Body responses

We took each Kenyon cell to have non-zero connection weights drawn uniformly between 0 and 1 with 8 randomly selected glomeruli (see Results). Then, following [38], we took the input to the  $i^{\text{th}}$  Kenyon cell, evoked by an odor with glomerular responses  $\vec{y}$  in the Antennal Lobe, to be

$$h_i = \langle \vec{w}_i, (\vec{y} - \langle \vec{\mu}, \vec{y} \rangle \vec{\mu}) \rangle \quad (2.2)$$

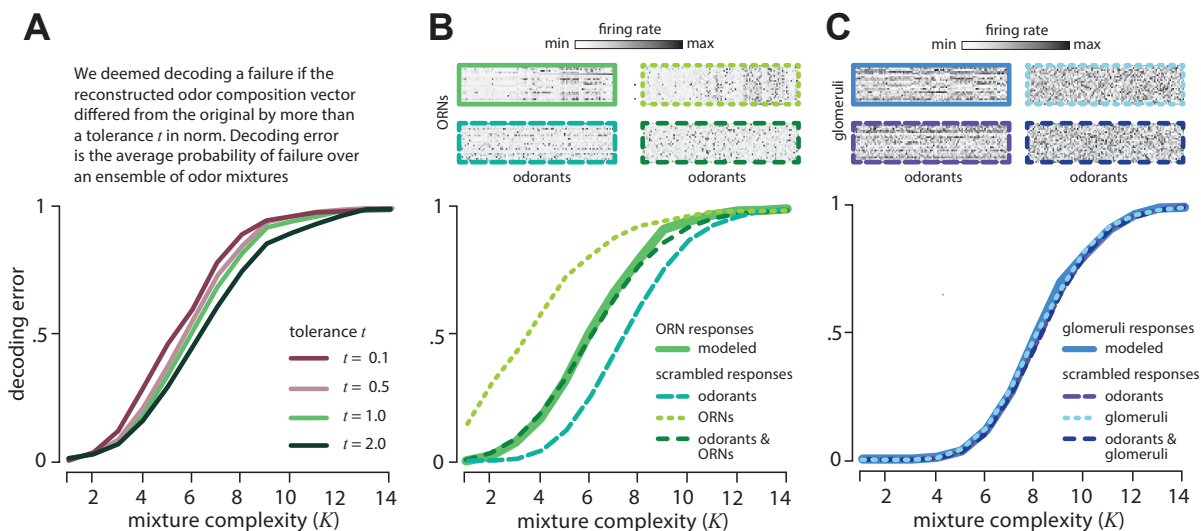
where  $\langle \cdot, \cdot \rangle$  is an inner-product,  $\vec{w}_i$  is the vector of connection strengths, and  $\vec{\mu}$  is the average Antennal Lobe response vector over all odors, normalized to unit length. We chose a response threshold so that a fraction  $f$  of neurons with inputs  $h_i$  exceeding threshold are considered active, and normalized the thresholded responses so that the maximum firing rate is 5 Hz, on the order of the maximum observed Kenyon cell responses. We averaged results over 100 random choices of connection strengths. The global inhibition required in this model for generating the disordered responses observed in the Mushroom Body [38] could be implemented by the APL neuron which makes inhibitory connections to all the Kenyon cells

### 2.6.5 Structured vs. random connectivity

We constructed structured connectivity matrices between glomeruli in the Antennal Lobe and Kenyon cells in the Mushroom Body by reordering the columns of the corresponding random connectivity matrix so that the two matrices model synapses with the same connection strengths feeding into each Kenyon cell, but they sample different glomeruli. The reordering of the columns was done so that the structured connectivity matrix exhibited a block-diagonal structure as shown in Fig. 2.4B. For analyses we chose the number of blocks to be 3. We then permuted the rows and columns of the structured connectivity matrix so that the underlying structure was not visible to the eye or to a casual analysis.

### 2.6.6 Robust decoding from ORN and glomerular responses

In the main text, we considered a simple linear model of the responses of 24 ORN types in *Drosophila* responding to odor mixtures. Specifically, we extracted a firing rate matrix  $R$  from the data in [21] (i.e.  $R_{ij}$  is the response of receptor  $i$  to odorant  $j$ ), and we assumed that the response to a mixture could be written as a linear combination of responses to single odorants. We defined a mixture by the composition vector  $x$  whose elements specify the concentration of individual odorants in the mixture. The ORN firing rates  $y$  could then be written as  $\vec{y} = R\vec{x}$ . We then attempted to decode composition vectors  $\vec{x}$  from responses  $\vec{y}$  using the optimal algorithm of [25, 51]. We regarded the reconstruction as a failure if the average squared difference between components of the reconstructed odor vector and the original exceeded 0.01. Decoding error was defined as the failure probability over an odorant mixture ensemble. This criterion for successful reconstruction is equivalent to saying that the reconstruction

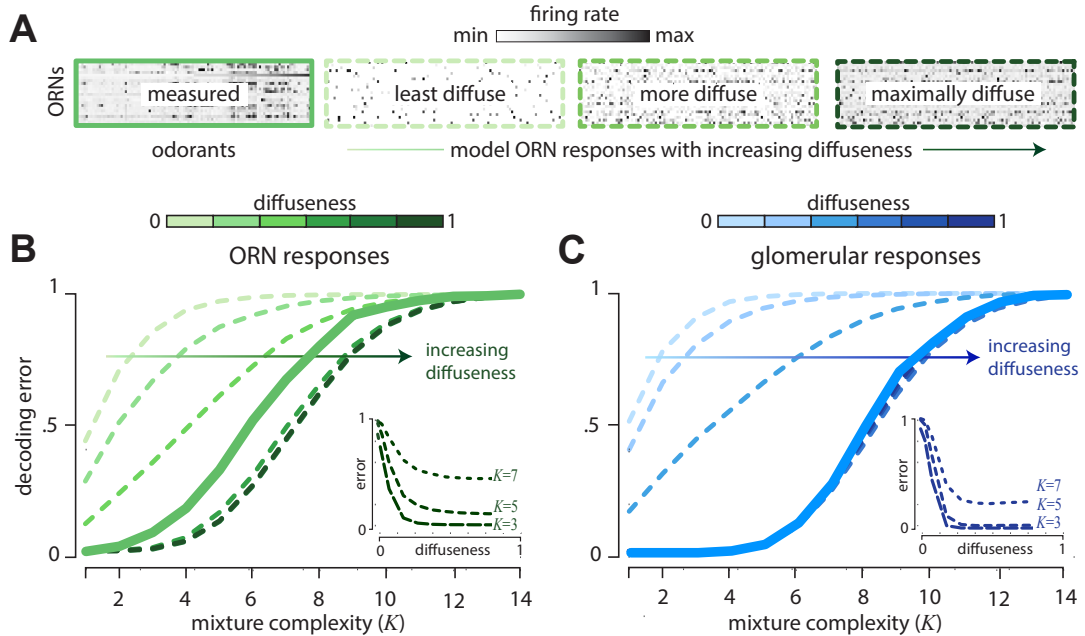


**Figure 2.5: Odor decoding from *Drosophila* ORN responses is robust.** (A) Decoding error is robust to ten-fold variations in odor reconstruction tolerance. Mixture complexity = number of component odorants drawn from 110 possibilities. (B) Decoding performance is unchanged after complete scrambling of the *Drosophila* response matrix, because of opposite effects of scrambling receptors vs. odors. Insets: Response matrices showing firing rates for 24 receptors (rows) responding to 110 monomolecular odorants (columns) without scrambling (solid green) and for models randomly scrambling receptors, odorants, or both (dashed green). (C) Decoding performance is unchanged after complete scrambling of the divisively normalized responses in the Antennal Lobe. Separately scrambling receptors or odors also has no effect on performance. Insets: Response matrices showing activity for 24 glomeruli (rows) responding to 110 monomolecular odorants (columns) without scrambling (solid blue) and after scrambling receptors, odorants, or both (dashed blue). Results shown are averages over 100 iterations over model scrambled response matrices. Decoding error is measured as the probability of decoding failure (see text) over an ensemble of 500 randomly chosen odor mixtures of a given complexity.



$\hat{x}$  of the odor composition vector  $\vec{x}$  fails if the norm of the difference  $\|\vec{x} - \hat{x}\|$  exceeds a tolerance parameter of  $t = 1.1$  (here we used the fact that the odor composition vector  $\vec{x}$  has 110 components). To test the robustness of our conclusions we varied this tolerance parameter ten-fold, and found that the decoding error curves were largely unchanged (Fig. 2.5A). Qualitatively, we observed this robustness because the decoding of odors tends to either succeed very well, or fail very badly. As a result, a broad range of criteria for defining a successful reconstruction will give similar measures of decoding error.

According to our general theory, and the results of [25, 53], the quality of the olfactory code should not depend on the details of how specific receptors respond to different odorants. Rather, the key determinant should be the overall distribution of responses. To test whether this is the case, we scrambled the receptor and odorant labels in the ORN response matrix (top inset in Fig. 2.5B), thus constructing an artificial response matrix with the same overall *distribution* of firing rates, but with no odor- or receptor-dependent correlations (second inset in Fig. 2.5B). We found that decoding performance was essentially identical when using the scrambled and unscrambled response matrices ( Fig. 2.5B), consistent with the notion that the olfactory system seeks to employ disordered and unstructured sensing. Interestingly, separate scrambling of the receptor labels and odor labels either improved or degraded the decoding, presumably because such scramblings removed correlations that were either detrimental or beneficial for decoding ( Fig. 2.5B). These opposite effects compensated each other when the sensing matrix was fully scrambled. We repeated this analysis after implementing a divisive normalization of ORN responses (see main text). In this case, all scramblings left the decoding performance unchanged ( Fig. 2.5C). We thus conclude that after correlations are removed by divisive nor-



**Figure 2.6: Weakly responding ORNs and glomeruli are informative about odor mixture composition.** (A) Firing rate response matrix measured from *Drosophila* ORNs (left, solid green), and for increasingly diffuse model response matrices (right, dashed green; “diffuseness” = fraction of largest responses kept). Model responses are constructed by thresholding measured responses and then scrambling the response matrix. (B) Error in decoding from ORNs decreases systematically as diffuseness increases – hence weak responses are informative. Results shown as a function of mixture complexity ( $K$  = number of odor mixture components). (C) ORN responses are divisively normalized to produce responses in the glomeruli of the Antennal Lobe (see Appendix B). Thresholding and scrambling these responses produces sensing models with different degrees of diffuseness. Error in decoding from glomeruli decreases systematically as diffuseness increases. Results shown are averages over 100 iterations over model response matrices for each degree of diffuseness. Decoding error is measured as the probability of decoding failure over an ensemble of 500 randomly chosen odor mixtures of a given complexity.

malization, the overall distribution of responses is the sole determinant of the quality of the olfactory information representation.

### 2.6.7 Weakly responding ORNs and glomeruli are informative

Our theory predicts that the olfactory code is dispersed across all the receptors, so that even weak responses are informative. To test this, we parametrized the fraction

of largest responses that are deemed above threshold by a “diffuseness parameter”  $f$ . We retained a fraction  $f$  of the largest rank-ordered responses for each receptor, and we set the remaining values to zero. A diffuseness value of  $f = 1.0$  means we retain all responses, whereas a diffuseness value of  $f = 0.5$  means that we retained the strongest 50% of all responses. We then created model response matrices for a given diffuseness value  $f$  by randomly scrambling the thresholded receptor responses. Fig. 2.6A shows the *Drosophila* ORN response matrix, along with model response matrices with increasing diffuseness. Fig. 2.6B show decoding error (definition in main text) as a function of mixture complexity  $K$  ( $K =$  number of nonzero components in each mixture) for varying diffuseness. We see that decoding error decreases systematically as diffuseness increases, showing that weak receptor responses are informative about odor mixture identity. The insets show decoding error as a function of the diffuseness parameter for fixed values of mixture complexity ( $K = 3, 5, 7$ ). The results for the models with varying diffuseness are averaged over 100 randomly scrambled model response matrices. Fig. 2.6C shows analogous results after applying divisive normalization to model responses in the glomeruli of the Antennal Lobe (see Appendix B for details of this normalization). The results show that weakly responding glomeruli are informative about mixture composition.

### 2.6.8 Optimal decoding from the Antennal Lobe

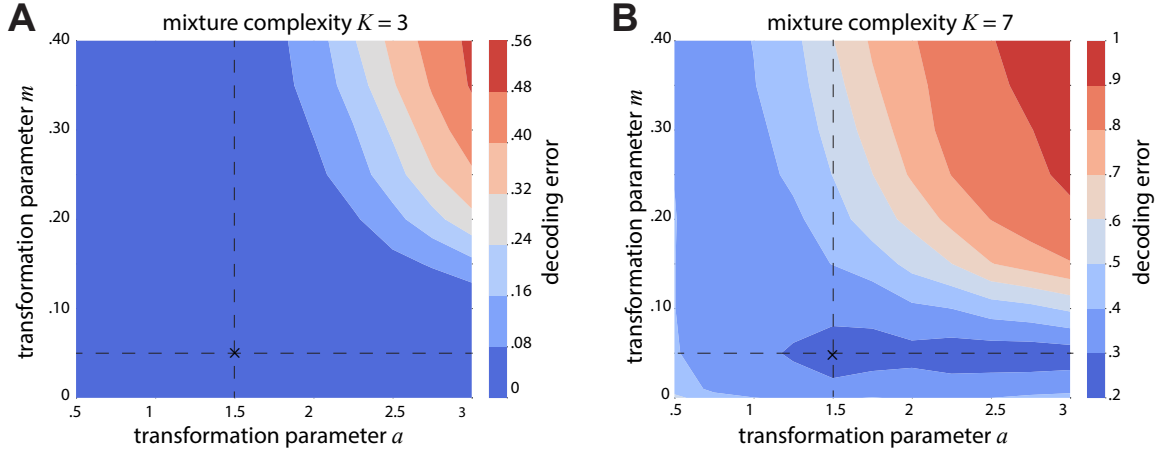
The inhibitory circuitry in the Antennal Lobe in *Drosophila* has been shown to perform a divisive normalization with the functional form [27, 28]

$$R_i^{(2)} = \frac{R_{max} \cdot \left(R_i^{(1)}\right)^a}{\left[\sigma^a + \left(R_i^{(1)}\right)^a + \left(m \cdot \sum_i R_i^{(1)}\right)^a\right]}, \quad (2.3)$$

where  $R_i^{(1)}$  is the response of the  $i^{\text{th}}$  ORN type,  $R_i^{(2)}$  is the response of the  $i^{\text{th}}$  glomerulus,  $\sigma$  parametrizes spontaneous activity, and  $m$  controls the amount of normalization. A fit to data in [27, 28] gave the parameters  $R_{\text{max}} = 165$ ,  $\sigma = 10.5$ ,  $m = 0.05$  and  $a = 1.5$ . In the main text, we constructed an artificial glomerular response matrix  $R^{(2)}$  by applying this transformation separately to the ORNs responding to each of the 110 odorants studied in [21]. Thus  $R_{ij}^{(2)}$  represented the response of the  $i$ th glomerulus to the  $j$ th odorant. In the main text, we studied odor decoding in an artificial benchmark model in which mixtures  $\vec{x}$  lead to responses  $\vec{y}$  via  $\vec{y} = R^{(2)}\vec{x}$ . We tested how our results for decoding error (see definition in the main text and above) would be affected by changing the parameter  $m$ , which controls the amount of inhibition in the Antennal Lobe, or the exponent  $a$ , which controls the shape of the nonlinearity. In order to simplify our presentation, we study dependence on the parameters of the normalization for two values of mixture complexity: i)  $K = 3$ , a value where odor reconstruction from Antennal Lobe responses with experimentally-measured parameters is near perfect (see main text), and ii)  $K = 7$ , a value where a similar reconstruction starts to degrade. (See main text for details regarding the construction of model odor mixtures of different complexities.) We found that in both cases, the experimentally measured values of  $m$  and  $a$  led to the lowest decoding error (Fig. 2.7).

### 2.6.9 Mushroom Body classification error for mixtures

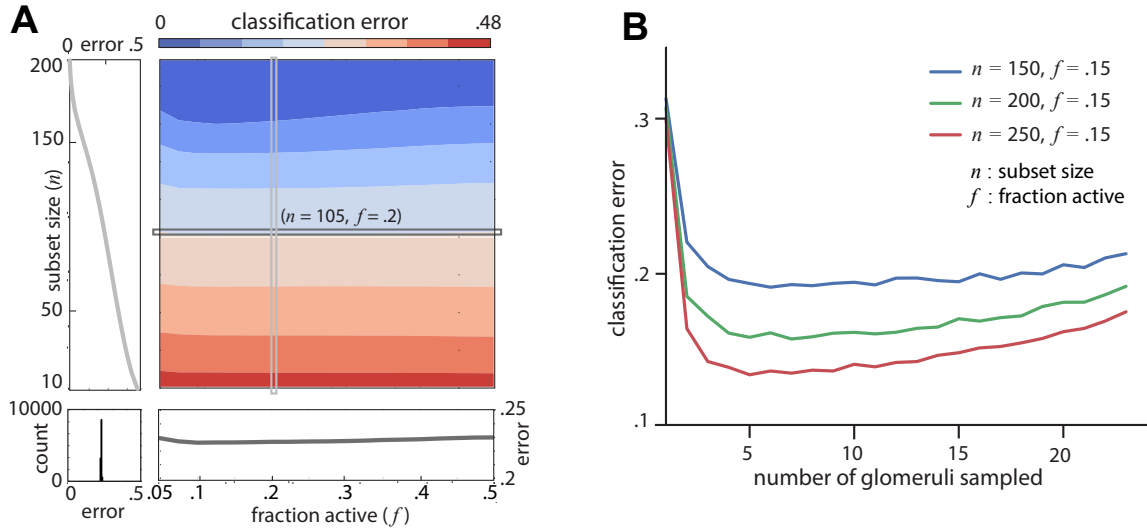
We studied the error in a 2-way classification task for 300 5-component mixtures with varying readout population sizes ( $n$ ) and fraction of active Kenyon cells ( $f$ ) in the Mushroom Body (details of classification procedure and task in the main text). For a given population size  $n$ , increasing the fraction of active neurons  $f$  barely changes



**Figure 2.7: The empirically determined divisive normalization in the Antennal Lobe is optimal for the measured ORN sensing matrix.** Decoding error (see main text for definition) shown as a function of the exponent  $a$ , and the inhibition parameter  $m$  in the divisive normalization carried out by the Antennal Lobe. Left and right plots correspond to mixtures with  $K = 3$  and  $K = 7$  components drawn randomly from 110 odorants, respectively. The experimentally measured operating point is indicated by a cross in each plot ( $m = 0.05$  and  $a = 1.5$ ). Decoding error (definition in main text) is averaged over 500 iterations of mixture ensembles of a given complexity.

the classification performance (bottom panel of Fig. 2.8A). The classification error with a given active fraction  $f$  decreases with the number  $n$  of neurons being read out (left panel of Fig. 2.8A). However, there is a law of diminishing returns – excellent performance is achieved for relatively small  $n$ , and further increasing the population size makes little difference. The disordered projections from the Antennal Lobe to the Mushroom Body suggest that any subset of a given size should be statistically equivalent. We tested this by comparing the classification error obtained from different subsets of Kenyon cells. The narrowness of the histogram of classification error for 10000 different populations ( $n = 105$ ,  $f = 0.2$ ) (lower left panel, Fig. 2.8A) shows that any subset of a given size is indeed equally good at supporting flexible classification.

We also studied how the classification error depended on the number of glomeruli sampled by each Kenyon cell in the Mushroom Body. Figure 2.8B shows the classification error as a function of the number of glomeruli sampled, for three different



**Figure 2.8:** A) Classification error from responses of model Kenyon cells in the Mushroom Body (MB) for arbitrarily separating 300 5-component mixtures into two classes as a function of the read-out size ( $n$ ) and the fraction ( $f$ ) of active neurons. The horizontal and vertical sections correspond to  $n = 105$  and  $f = 0.2$ , respectively (section shown in panels below and to the left, respectively). Bottom left panel: histogram of classification errors for 10000 different subsets of size  $n = 105$  and  $f = 0.2$ . The narrowness of the histogram shows that any two subsets of a given size are roughly equivalent for odor classification purposes. B) Classification error at the Mushroom Body as a function of the number of glomeruli sampled by each Kenyon cell. Minimum error is found for sparse sampling of glomeruli. All results shown are averages over 100 iterations over mixture ensembles, 100 labelings into appetitive/aversive classes, and 100 iterations over model connectivity matrices between the Antennal Lobe and Mushroom Body (each using a different instantiation of noise). (See main text for details regarding the generation of connectivity matrices and noise.)

readout sizes. We see that the classification error initially decreases and then gradually rises as we increase the number of glomeruli sampled. This indicates that there is an optimum for the number of sampled glomeruli. Recent work [39] has examined this question theoretically; here we show results with *Drosophila* data which are consistent with [39].

## 2.7 Linearization of the Antennal Lobe transformation

Here we show that the nonlinearity in the transformation at the Antennal Lobe(Stage 2) is crucial in improving the embedding at the level of ORNs(Stage 1). Specifically, we show that using a linear approximation to the Stage 2 transformation does not yield improvements over the Stage 1 embedding. Recall, that our measure of the quality of the Stage 1 embedding was the ability to reconstruct the mixing proportion vector  $\vec{x}$  from linear measurements  $\vec{y} = R\vec{x}$ , where  $R$  is the matrix of responses in the *Drosophila* dataset. At Stage 2 there is a (non-linear) divisive normalisation of the form:

$$R_i^{(2)} = \frac{R_{max} \cdot \left(R_i^{(1)}\right)^{1.5}}{\sigma^{1.5} + \left(R_i^{(1)}\right)^{1.5} + \left(m \cdot \sum_i R_i^{(1)}\right)^{1.5}} \quad (2.4)$$

where  $R_i^{(1)}$  is the response of the  $i^{th}$  ORN type (i.e. before Stage 2 processing) to the presented odorant, and  $R_i^{(2)}$  is the responses of the  $i^{th}$  glomerulus (i.e. after Stage 2 processing) to the same odorant. We succinctly denote this transformation

as  $R^{(2)} = f(R^{(1)})$ , where  $f()$  is the transformation function.

- $m$  : the parameter that controls the strength of the global inhibition in the divisive normalization model
- $\sigma$  : term that is related to the level of spontaneous activity

In the main text, we measured the quality of the Stage 2 embedding by trying to reconstruct the mixing vector  $\vec{x}$  from modelled mixture responses at Stage 2 given by  $\mathbf{y} = R^{(2)}\mathbf{x}$  where  $R^{(2)}$  is the matrix of transformed responses to monomolecular odorants described above. Here, we ask how well can we reconstruct  $\vec{x}$  if we instead use a linearized version of the nonlinearity  $f()$ . This will tell us how important is the nonlinear nature of the transformation in improving the embedding at Stage 2. Specifically, the linear approximation to the nonlinearity  $f$  around an operating point  $\vec{x}_0$  is

$$f(R^{(1)}(\vec{x}_0 + \delta\vec{x})) \approx f(R^{(1)}\vec{x}_0) + [\partial_i f_j] R^{(1)}\delta\vec{x} \quad (2.5)$$

Where  $[\partial_i f_j]$  is the matrix with elements which are the partial derivatives elements of  $f()$  along the various odorant dimensions. We can then ask for a given  $\vec{x}_0$  how well can we recover  $\delta\vec{x}$  from  $y = [\partial_i f_j] R^{(1)}\delta\vec{x}$ . In particular, does pre-multiplying by  $[\partial_i f_j]$  yield any benefits?

Let us first consider a form of the transformation with a general exponent (repeated



indices are summed over):

$$R_i^{(2)} = \frac{R_{max} \cdot \left(R_{ij}^{(1)} x_j\right)^a}{\sigma^a + \left(R_{ij}^{(1)} x_j\right)^a + \left(m \cdot \sum_k R_{kj}^{(1)} x_j\right)^a}$$

Then the derivative w.r.t.  $x_j$  is given by:

$$\begin{aligned} \frac{\partial [f(R^{(1)}\mathbf{x})]_i}{\partial x_j} &= \frac{a R_{max}^{(1)} \left(R_{il}^{(1)} x_l\right)^a}{\left(\sigma^a + \left(R_{il}^{(1)} x_l\right)^a + \left(m \cdot \sum_i R_{il}^{(1)} x_l\right)^a\right)^2} \times \\ &\quad \left[ \sigma^a R_{ij}^{(1)} + R_{ij}^{(1)} \left(m \cdot \sum_k R_{kj}^{(1)} x_j\right)^a - \right. \\ &\quad \left. \left(R_{il}^{(1)} x_l\right) m^a \left(\sum_k R_{kj}^{(1)}\right) \left(\sum_k R_{kl}^{(1)} x_l\right)^{a-1} \right] \end{aligned}$$

We can simplify this to

$$\frac{\partial [f(R^{(1)}\mathbf{x})]_i}{\partial x_j} = c_i(\mathbf{x}) \left[ R_{ij}^{(1)} d(\mathbf{x}) - S_{ij} e(\mathbf{x}) \right]$$

where

$$S = (R^{(1)}\mathbf{x}) (1^T R^{(1)})$$

and  $d(\mathbf{x})$ ,  $e(\mathbf{x})$  are terms dependent on the overall ORN activity, given by:

$$\begin{aligned}
d(\mathbf{x}) &= \sigma^a + (m \cdot \mathbf{1}^T R^{(1)} \mathbf{x})^a \\
e(\mathbf{x}) &= m^a (\mathbf{1}^T R^{(1)} \mathbf{x})^{a-1}
\end{aligned}$$

The term  $c_i(x)$  is given by

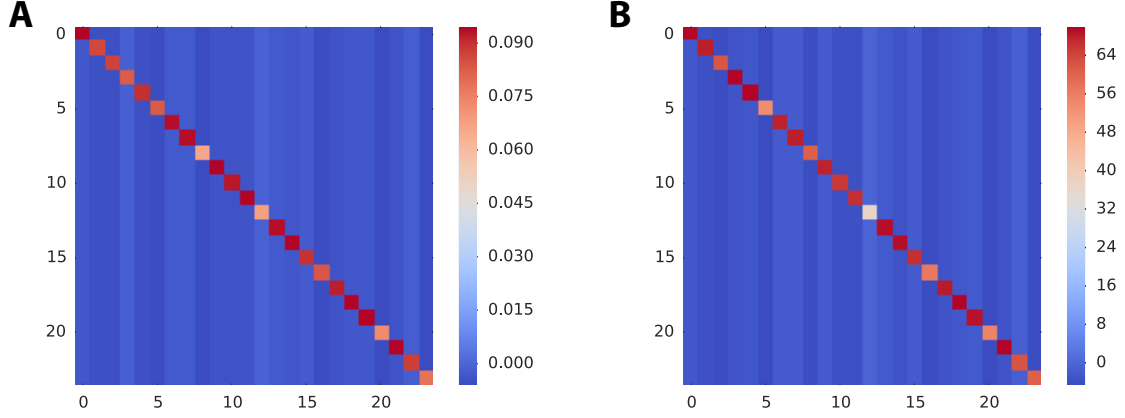
$$c_i(\mathbf{x}) = \frac{a R_{max} \left( R_{il}^{(1)} x_l \right)^a}{\left( \sigma^a + \left( R_{il}^{(1)} x_l \right)^a + \left( m \cdot \sum_i R_{il}^{(1)} x_l \right)^a \right)^2}$$

Writing the derivative in this form makes it clear how the overall activity contributes to the derivative. In matrix form, the derivative is then

$$\frac{\partial [f(R^{(1)} \mathbf{x})]}{\partial x} = C(\mathbf{x}) [d(\mathbf{x}) \mathbf{I} - e(\mathbf{x}) R^{(1)} \mathbf{x} \mathbf{1}^T] R^{(1)} \quad (2.6)$$

where  $C$  is a diagonal matrix made up of  $c_i(\mathbf{x})$ . The terms preceding the rate matrix  $R^{(1)}$  on the r.h.s are nothing but the derivative  $[\partial_i f_j]$  of  $f()$  we had mentioned earlier.

Let us now explicitly calculate the linearisation for the parameters considered by Wilson et al. In this case,  $[\partial_i f_j]$  is given by



**Figure 2.9:** Derivative  $[\partial_i f_j]$  at operating point  $y_0 = Rx_0$  where  $R$  is the Carlson rate matrix and  $x_0$  is (A): a (110-dimensional) sparse vector with 10 non-zero entries and (B): a 110-dimensional with all entries set to a small ( $10^{-4}$ ) value to mimic some faint background.

$$[\partial_i f_j] = \frac{3}{2} R_{max} x_i^{0.5} \left( \frac{\sigma^{1.5} + m^{1.5} \left( \left( \sum_k x_k \right)^{1.5} - \left( \sum_k x_k \right)^{0.5} \right)}{\left( \sigma^{1.5} + (x_i)^{1.5} + \left( m \cdot \sum_k x_k \right)^{1.5} \right)^2} \right) \quad \text{when } i = j$$

$$[\partial_i f_j] = -\frac{3}{2} R_{max} (m \cdot x_i)^{1.5} \left( \frac{\left( \sum_k x_k \right)^{0.5}}{\left( \sigma^{1.5} + (x_i)^{1.5} + \left( m \cdot \sum_k x_k \right)^{1.5} \right)^2} \right) \quad \text{when } i \neq j$$

We look at how the derivative looks at two ecologically relevant operating points :  
i) a sparse operating point with 10 of the 110 entries of  $\vec{x}_0$  are non-zero and each non-zero element is a uniform random number between 1 and 2 (Fig. 2.9A) and ii) an operating point corresponding to a weak background where all entries of  $\vec{x}_0$  are set to a small value of  $10^{-4}$ . (Fig. 2.9B). The derivatives at both these operating

points are effectively diagonal thus the linearized Stage 2 response will not offer any improvements compared to Stage 1 in embedding the sparse high-dimensional vectors.

## 2.8 Addendum: background on random projections and compressive-sensing

In this section, we summarize the key mathematical results and insights which illustrate the benefits of random projections in compressing sparse high-dimensional signals. These results were a key motivation in our thinking about the logic of sensing by the olfactory receptor neurons. We first begin with the problem of solving an underdetermined linear system of equations and examine the particular case when the solutions have to be sparse. Then we summarize the properties of random projections which make it possible to recover sparse high-dimensional vectors from substantially fewer measurements, and finally we mention some results which suggest that random (or diffuse) projections might be a universally good method to represent a variety of signals with nonlinear structure.

### 2.8.1 Solving $\mathbf{y} = A\mathbf{x}$

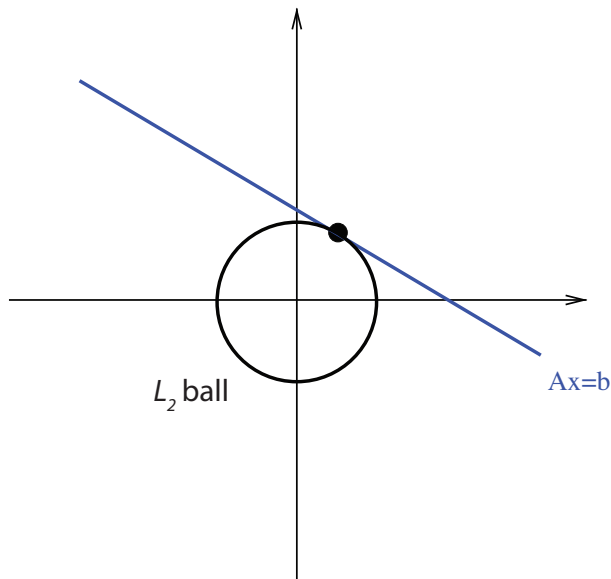
Consider the problem of taking linear measurements about some signal  $\mathbf{x}$  – for e.g.,  $\mathbf{x}$  could be an image or a time-series signal and  $A$  could be the Fourier transform operator, in which case, we make frequency domain measurements about our signal. Now, suppose that our signal  $\mathbf{x}$  resides in some high-dimensional space of dimension  $N$  and we only take  $M$  measurements. Can we recover  $\mathbf{x}$  from the measurements  $\mathbf{y}$ ?

In general, when  $M > N$ , and  $A$  is full-rank, we have a overdetermined system of

equations and we can use methods like least-squares to give us a solution. What about the case when  $M < N$ ? In this case, we have an underdetermined system of equations, and in general, we don't have a unique solution for  $\mathbf{x}$ . In certain scenarios, of all the possible solutions to  $\mathbf{y} = A\mathbf{x}$ , the one you care about is “small” in some sense. One notion of small is the solution having the least “energy” or the minimum  $L_2$  – norm. In this case we get the pseudo-inverse solution:

$$\begin{aligned}\hat{x} &= \arg \min \|\mathbf{x}\|_2 \quad \text{s.t. } \mathbf{y} = A\mathbf{x} \\ &= A^*(AA^*)^{-1}\mathbf{y}\end{aligned}$$

This amounts to “growing” the  $L_2$  “ball” till you satisfy the constraint as shown in the schematic in Fig. 2.10



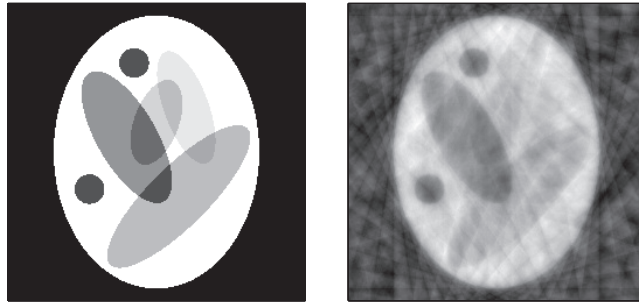
**Figure 2.10:** Geometric illustration of the pseudo-inverse solution in 2D

A classic example of this scenario is finding the minimum energy reconstruction of a signal  $f \in \mathbb{R}^N$  from a limited number of  $M < N$  Fourier measurements :  $\tilde{f}(\omega_1), \tilde{f}(\omega_2) \cdots \tilde{f}(\omega_M)$ . In this case, the solution is simple : it is simply the re-

constructed signal with the measured Fourier coefficients.

$$\hat{f}(t) = \sum_{i=1}^M \tilde{f}(\omega_i) e^{i2\pi\omega_i t/N}$$

However, in many cases the minimum energy solution is way off and not what we are looking for. For instance, consider the Logan-Shepp phantom[20] image and its corresponding minimum  $L_2$  solution reconstructed from a limited number of (2D) Fourier measurements (Fig. 2.11 ). The minimum energy solution has a number of artifacts which obscure the structure in the image.



**Figure 2.11:** minimum  $L_2$  reconstruction (right) performs poorly when reconstructing the image(left) from incomplete 2D Fourier measurements

### 2.8.2 Solving $y = Ax$ for sparse signals

In many scenarios, the signal  $\mathbf{x}$  has a sparse or “compressible” structure – i.e., only a few (or small fraction of) elements of  $\mathbf{x}$  are significant. A signal is called  $K$  – sparse if it has at most  $K$  non-zero entries; the location of these non-zero entries can, however, be arbitrary. Given such a structure for the signal, can we exploit this information to recover the signal from incomplete measurements?

Consider this interesting empirical observation : if you find a solution which minimizes the  $L_1$  norm instead of the  $L_2$  norm, then in many cases you can *exactly* reconstruct the signal from highly incomplete measurements, provided the signal has a sparsity structure. Specifically, if we solve the following problem :

$$\begin{aligned} \arg \min \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x} & \quad (2.7) \\ \text{where } \|\mathbf{x}\|_1 = \sum_{i=1}^N |x_i| & \end{aligned}$$

then we can *exactly* recover the true (sparse)  $\mathbf{x}$  even when the number of measurements  $\mathbf{y}$  ( $M$ -dimensional) are substantially fewer than the dimension,  $N$ , of  $\mathbf{x}$ . This fact was well known to researchers studying seismology data, where the reflected signals naturally had a sparse structure due to discrete transitions in the earth’s crust. Fig. 2.12 shows that the minimum  $L_1$  reconstruction of Logan-Shepp phantom from incomplete (as little as  $\sim 1\%$ ) Fourier measurements gives back the exact image[20]. Note, that the sparsity structure exploited in this case is the sparsity in the *gradient* of the image and not the pixels themselves.

Another early example of the success of  $L_1$ , with more theoretical backing, is the basis pursuit problem. Consider a measured signal  $f \in \mathbb{R}^n$  which is made up of the superposition of two signals, one which is sparse in an ortho-basis  $\Phi_1$  and the other which is sparse in another ortho-basis  $\Phi_2$  – i.e.  $f = \Phi\alpha$  where  $\Phi = [\Phi_1\Phi_2]$  and  $\alpha$  is sparse. We would like to know  $\alpha$  so that we can split the signal into its components. A practical example comes from astronomy where telescope images often contain elements which look “texture” like and elements that look like lines or rods. We would like to separate the texture “background” from the more linear features. A

theorem due to [19], showed that if the “coherence”  $\mu$  between the bases defined as

$$\mu(\Phi_1, \Phi_2) = \sqrt{n} \max_{1 \leq i, j \leq n} \left| \langle \varphi_i^{(1)}, \varphi_j^{(2)} \rangle \right| \quad (2.8)$$

is small, then solving the following problem gives back the exact  $\alpha$  :

$$\arg \min \|\hat{\alpha}\|_1 \quad \text{s.t.} \quad f = \Phi \hat{\alpha}$$

provided

$$\|\alpha\|_0 \leq \frac{0.9}{\mu(\Phi_1, \Phi_2)} \sqrt{n}$$

The coherence between two bases is a measure of how “different” do bases function in one basis look compared to the other. And, for any two orthobases,  $\mu \geq 1$ , and for good reconstruction we want the coherence to be low. Time and frequency bases would be examples of bases which have small coherence. We will return later to the notion of coherence. But, note that the restriction here about the sparsity of  $f$  : ( $O\sqrt{n}$ ) is rather restrictive; it requires  $f$  to be quite sparse. There are more powerful theorems that guarantee that  $L_1$  will do well under much more general conditions!

Let us now consider one example of a family of more general theorems[20, 54, 19] which guarantee exact reconstruction :

- **Theorem** [Candes & Tao 2006]: Let  $A$  be a  $M \times N$  matrix with entires  $A_{ij}$  drawn i.i.d from  $\mathcal{N}(0, 1)$ . Let  $\mathbf{x} \in \mathbb{R}^N$  be an unknown, but *fixed*  $K$ -sparse vector, and furthermore assume we have access to the  $M$  measurements  $\mathbf{y} = A\mathbf{x}$ .





**Figure 2.12:** minimum  $L_1$  reconstruction (right) returns the exact original image (left) from incomplete 2D Fourier measurements

Then we can reconstruct  $\mathbf{x}$  exactly with overwhelming probability by solving

$$\begin{aligned} \arg \min \|\hat{\mathbf{x}}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\hat{\mathbf{x}} & \quad (2.9) \\ \text{provided } M \gtrsim K \log\left(\frac{N}{K}\right) & \end{aligned}$$

Moreover, no other method can reconstruct  $\mathbf{x}$  with “fewer” (in order) measurements – even if you use some other (even *adaptive*) sensing and any reconstruction method, you cannot do better (asymptotically) than Gaussian sensing followed by  $L_1$  reconstruction.

This remarkable theorem states that if you know that  $\mathbf{x}$  is  $K$ -sparse you only need to take slightly more than  $\mathcal{O}(K)$  **non-adaptive** measurements to get back  $\mathbf{x}$  by convex optimisation, and this scheme is universally optimal in some sense. For Gaussian matrices, something like  $\sim 4K$  measurements will suffice. In the what follows we will review some theory for why this works, and in particular the following important practical questions:

- What are the requirements on the linear measurement operator  $A$ ?
- For some  $A$ , how many measurements  $M$  are required to guarantee exact reconstruction of a  $N$  dimensional signal with  $K$  non-zero entries?

- How strict is the sparsity requirement? In particular, will the results hold for compressible signals, where only  $K$  entries are significant, but other entries are small and non-zero?
- What happens if we have measurement noise?

### 2.8.3 Why does the $L_1$ solution work?

To understand why the  $L_1$  solution gives the exact result for the underdetermined system, let us first consider the related question: When can we recover any  $K$ -sparse vector  $\mathbf{x}$  from  $M$  measurements  $\mathbf{y} = A\mathbf{x}$  using any method whatsoever? It turns out that if *any* of  $2K$  columns of  $A$  are linearly independent (so necessarily  $M \geq 2K$ ), then there is a unique solution to  $\mathbf{y} = A\mathbf{x}$  for a  $K$ -sparse vector  $\mathbf{x}$ . To see this, assume there are two solutions  $\mathbf{x}, \tilde{\mathbf{x}}$  then  $\mathbf{x} - \tilde{\mathbf{x}}$  is at most  $2K$ -sparse, and  $A(\mathbf{x} - \tilde{\mathbf{x}}) = 0$  – this cannot be true unless  $\mathbf{x} = \tilde{\mathbf{x}}$  because any  $2K$  columns of  $A$  are linearly independent. This argument also suggests an algorithm to find the unique solution for the case  $M \geq 2K$  : choose every subset of  $K$  columns and try to solve  $\mathbf{y} = A_K\mathbf{x}_K$  where  $A_K$  is the submatrix of  $A$  with the  $K$  selected columns and  $\mathbf{x}_K$  is the vector with non-zero entries of  $\mathbf{x}$ . This problem can be equivalently formulated as minimising the “ $L_0$ ” norm

$$\arg \min \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x} \tag{2.10}$$

where  $\|\mathbf{x}\|_0 =$  no. of non-zero entries in  $\mathbf{x}$

Unfortunately, this problem which we call the  $L_0$  problem contains within it the subset-sum problem which is known to be NP complete. So it’s hopeless to use this

for even moderate size problems.

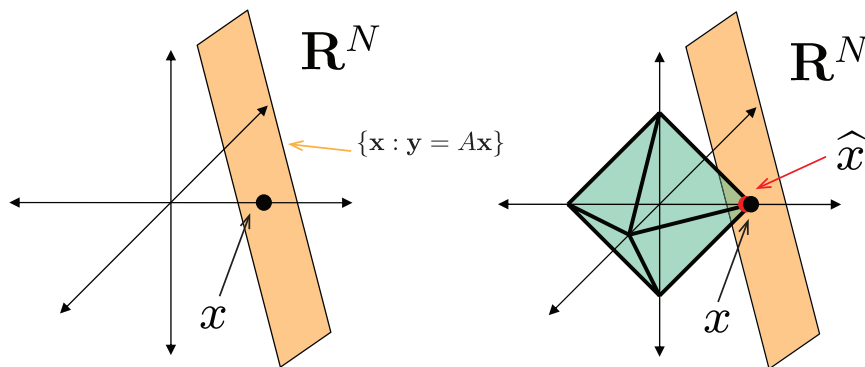
The  $L_1$  problem however admits a polynomial time solution – in fact, it is a linear program. To see this, note that solving the problem

$$\begin{aligned} & \arg \min \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x} \\ & \text{where } \|\mathbf{x}\|_1 = \sum_{i=1}^N |x_i| \end{aligned}$$

can be recast as the following equivalent problem

$$\begin{aligned} & \text{minimise} \quad \sum_i t_i \\ & \text{subject to} \quad -t_i \leq x_i \leq t_i \\ & \text{and } \mathbf{y} = A\mathbf{x} \end{aligned}$$

which is a well known linear program that can be solved using, for e.g., the simplex method.



**Figure 2.13:** The  $L_1$  ball intersects the constraint surface at points which are sparse. In high dimensions the  $L_1$  ball is even more “pointed” and looks a lot like the  $L_0$  ball. . Schematic adapted from [55]

To get an intuition for why, with sufficient measurements, the  $L_1$  problem returns

the same solution as the  $L_0$  problem [56] (which is the best you can do), it's useful to view the problem geometrically (Fig. [2.13]). Sparse signals reside in union of planes (the  $L_0$  ball); in particular 1-sparse signals are the union of the axes in  $\mathbb{R}^N$ . So when searching for sparse solutions to  $\mathbf{y} = A\mathbf{x}$  we look for places where the hyperplane  $\mathbf{y} = A\mathbf{x}$  intersects the  $L_0$  ball as shown in the schematic Fig.[2.13]. It turns out that the  $L_1$  ball, because of its pointed vertices, also intersects the hyperplane at places where the solutions are sparse, and with sufficient measurements the solution to the  $L_0$  and the  $L_1$  problem are exactly identical. In very high dimensions the  $L_0$  ball looks a lot like the  $L_1$  ball and nothing like the  $L_2$  ball. Of course, you can't always get back the  $L_0$  solution by solving the  $L_1$  problem, otherwise  $P=NP(!)$ . But, CS theory tells us that the two give the same solution by taking slightly more measurements for the  $L_1$  problem than the minimal amount required.

#### 2.8.4 Non-sparse signals : the best $K$ -term approximation

The problem used to motivate compressive sensing assumed that the signals of interest were  $K$ -sparse, however in most practical situations, the signals are not exactly sparse, but they are "compressible" : only a few entries are significant and the remaining entries decay rapidly but are non-zero. How do the compressive sensing results hold for compressible signals? What about signals with no known structure *a priori* ? To understand the extension of the classical compressive sensing results to non-sparse signals it is useful to ask the following question: *what's the best you can hope to do if you only got to make  $K$  measurements of a non-sparse signal?* As discussed above, of course, we can't expect to get back a general signal from incomplete measurements, but can we attempt to get back the "best reconstruction possible from  $K$  measurements" or will the compressive sensing paradigm fall apart for general

signals?

If you had access to an oracle who told you where the  $K$  most significant entries in  $\mathbf{x}$  were, then of course you'd want to measure those with your  $K$  measurements, and we would end with a best  $K$ -term approximation of  $\mathbf{x}$ . Let's call this  $K$ -term approximation as  $\mathbf{x}_K$ , which can also be written as

$$\mathbf{x}_K = \arg \min_{\mathbf{y}:k\text{-sparse}} \|\mathbf{y} - \mathbf{x}\|_2$$

This is the benchmark we would like to compare the reconstructed vector from the  $L_1$  recovery. CS theory essentially tells you that with just a few more samples than  $K$  for **any**  $\mathbf{x}$  – not necessarily sparse or compressible, we will recover the best  $K$ -term approximation  $\mathbf{x}_K$ . This recovered signal will be a good approximation of the original signal only for sparse or compressible signals, but the CS theory essentially tells you that the “best  $K$ -term approximation” results hold for any  $\mathbf{x}$ . So, sensing in a non-adaptive way followed by  $L_1$  reconstruction gives us a performance close to an oracle with perfect knowledge of the largest entries in the signal!

In several of the theorems for non-sparse signals, guarantees are provided that the accuracy of the reconstructed vector  $\hat{\mathbf{x}}$  from, say  $K$ , measurements is close to  $\mathbf{x}_K$ . These guarantees usually bound the reconstruction error ( $\|\hat{\mathbf{x}} - \mathbf{x}\|_1$  and  $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$ ) by the benchmark  $\|\mathbf{x}_K - \mathbf{x}\|_1$  and are often referred to as “**oracle bounds**”. If indeed  $\|\mathbf{x}_K - \mathbf{x}\|_1$  is small, then these bounds say that  $\hat{\mathbf{x}}$  will be close to the true signal; otherwise, they say that you can expect to do as well as the case where an oracle tells you the location of the  $K$  largest elements in  $\mathbf{x}$ . It doesn't get better than this!

A point to note about these theorems is that oracle bounds often compare a 2–norm ( $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$ ) to a 1–norm ( $\|\mathbf{x}_K - \mathbf{x}\|_1$ ). To know if and when these bounds are tight/good, let us compare the behavior 1–norm and 2–norm of compressible signals. A common model for compressible signals is a power-law decay of rank-ordered entries in  $\mathbf{x}$ . Let  $\mathbf{x}$  be a rank-ordered compressible signal s.t.

$$|\mathbf{x}_i| \geq |\mathbf{x}_{i+1}| \quad \forall i$$

and

$$|\mathbf{x}_i| \leq \frac{R}{i^p} \quad \forall i > K; p > 1 \tag{2.11}$$

for some constant  $R$ . Let us consider the signal  $\mathbf{x}_K$  which is constructed from  $\mathbf{x}$  by retaining the  $K$  largest components

$$(\mathbf{x}_K)_i = \begin{cases} \mathbf{x}_i & i \leq K \\ 0 & i > K \end{cases}$$

We are interested in observing how the 1-norm and 2-norm of the residual  $(\mathbf{x} - \mathbf{x}_K)$  behave for compressible signals in the following limit :

$$N \gg K \gg 1 \tag{2.12}$$

This is the relevant regime for real-life signals such as images. The norms of the

residuals are given by

$$\begin{aligned}\|\mathbf{x} - \mathbf{x}_K\|_1 &= \sum_{i=K+1}^N |\mathbf{x}_i| \\ \|\mathbf{x} - \mathbf{x}_K\|_2 &= \left( \sum_{i=K+1}^N |\mathbf{x}_i|^2 \right)^{1/2}\end{aligned}\tag{2.13}$$

For compressible signals we have  $|\mathbf{x}_i| \leq R/i^p$  for  $i > K$ . Let us first consider the case when  $\mathbf{x}_i \sim R/i^p$  for  $i > K$  (the case when the fall-off is faster than  $R/i^p$  will turn out to be better for the CS theorems); i.e. the bound  $\mathbf{x}_i \leq R/i^p$  is tight. In this case, for the limit in (2.12) we can approximate the sums by integrals

$$\begin{aligned}\|\mathbf{x} - \mathbf{x}_K\|_1 &\approx R \int_K^N \frac{1}{t^p} dt = \frac{R}{p-1} \left[ \frac{1}{K^{p-1}} - \frac{1}{N^{p-1}} \right] \\ &\approx \left( \frac{R}{p-1} \right) K^{1-p} \\ \|\mathbf{x} - \mathbf{x}_K\|_2 &\approx R \left( \int_K^N \frac{1}{t^{2p}} dt \right)^{1/2} = R \left[ \frac{1}{2p-1} \left( \frac{1}{K^{2p-1}} - \frac{1}{N^{2p-1}} \right) \right]^{1/2} \\ &\approx \left( \frac{R}{\sqrt{2p-1}} \right) K^{(1/2-p)}\end{aligned}\tag{2.14}$$

so

$$\frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}} \approx \text{constant} \times \|\mathbf{x} - \mathbf{x}_K\|_2\tag{2.15}$$

*Note that these relations will only hold for  $p > 1$  otherwise we can no longer neglect the dependence on  $N$  (which, for e.g., will enter as a logarithmic term for  $p = 1$ ).*

The take away from all this is the following: if you see a bound comparing  $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$  to  $\|\mathbf{x} - \mathbf{x}_K\|_1 / \sqrt{K}$ , then this bound is appropriate for compressible signals as described in (2.11), but if the smaller entries in the signal fall off much slower, then this bound is not assured to be tight. In any case, it always makes sense to compare  $\|\hat{\mathbf{x}} - \mathbf{x}\|_1$  to  $\|\mathbf{x} - \mathbf{x}_K\|_1$ .

## 2.8.5 Requirements for the sensing procedure

### Incoherence and random sampling

Let  $\mathbf{x} \in \mathbb{R}^N$  be  $K$ -sparse, and suppose we make  $M$  linear measurements of  $\mathbf{x}$  using sensing vectors chosen *uniformly randomly* from  $\Phi_2$  (for this result, we needn't restrict  $\Phi_2$  to be an orthobasis). Then [25] has shown the following:

- If the number of measurements

$$M \gtrsim K \cdot \mu^2(\mathbb{I}, \Phi_2) \log N \tag{2.16}$$

then solving  $L_1$  problem (2.7) exactly recovers the sparse component of the signal. Here  $\mathbb{I}$  is the canonical (e.g. time) basis.

- Moreover, if  $M$  is less than  $O(K \cdot \mu^2(\mathbb{I}, \Phi_2) \log N)$ , then no algorithm (even combinatorially hard ones) can recover  $\mathbf{x}$  from such measurements.

where  $\mu(\Phi_1, \Phi_2)$  is the coherence between the two bases as defined in [2.8]. There are a few points worth mentioning about this result:

1. The result holds for a random set of  $M$  measurements – i.e. any typical set of



$M$  measurements is equally good as long as  $M$  satisfies [2.16].

2. The coherence between the bases  $\mu(\mathbb{I}, \Phi_2)$  plays a critical role in deciding how many measurements are required for recovery. So, we ideally want bases which have very low ( $O(1)$ ) coherence. An example of incoherent bases is:  $\Phi_1$ –time and  $\Phi_2$ –Fourier. As a trivial example, if  $\mathbb{I} = \Phi_2$  then the bases are maximally coherent ( $\mu = \sqrt{N}$ ), and we need around  $N \log N$  samples. So, taking time samples of a signal which is 1–sparse in the time domain, will require us to collect  $\sim N \log N$  before we recover the signal  $\mathbf{x}$  (why the  $\log N$  term? – hint c.f. point above!).
3. If  $\Phi_2$  is a random basis, for e.g. if each element of the vector  $\varphi \in \Phi_2$  is sampled i.i.d from  $\mathcal{N}(0, 1/\sqrt{N})$  then with very high probability it is incoherent with any orthobasis  $\Phi_1$ .

This result can be extended to non-sparse signals. If the number of measurements satisfies (2.16), then the solution to

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \left[ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \lambda \sum_{i=1}^N |\mathbf{x}_i| \right] \quad (2.17)$$

will satisfy (with very large probability)

$$\begin{aligned} \|\hat{\mathbf{x}} - \mathbf{x}\|_2 &\lesssim \frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}} \\ \|\hat{\mathbf{x}} - \mathbf{x}\|_1 &\lesssim \|\mathbf{x} - \mathbf{x}_K\| \end{aligned}$$

So, with slightly more than  $K$  measurements we are close to the best  $K$ –term approximation for any  $\mathbf{x}$ . The problem (2.17) is also referred to as LASSO in the

literature, and the problem tries to find the best (in the  $L_2$  sense) solution to a linear system while at the same time penalising non-sparse solution. The parameter  $\lambda$  controls this trade-off and can be chosen appropriately (see [25]).

**Why do we need at least  $K \cdot \mu^2 \log N$  measurements?**

To get some intuition about why we need  $K \cdot \mu^2 (\Phi_1, \Phi_2) \log N$  samples using our random sampling scheme, consider the simple case when  $\Phi_1$  is the Fourier domain and  $\Phi_2$  is the time domain. In this case,  $\mu = 1$  – the bases are maximally incoherent. Further, let’s assume that the signal  $\mathbf{x} \in \mathbb{R}^N$  is a “Dirac comb” which is  $K$ – sparse :

$$x [t] = \sum_{j=0}^{K-1} \delta [t - \tau j]$$

where  $N = \tau K$  and the spacing between the spikes is  $\tau$ . The Fourier transform of  $\mathbf{x}$  will be  $\tau = N/K$  sparse with the spacing between the spikes  $K$ :

$$\tilde{x} [f] = K \sum_{j=0}^{\tau-1} \delta [f - jK]$$

So the Fourier transform of a Dirac comb is a Dirac comb, and the spacing between the spikes of the combs in the time and frequency domain are inversely related. This is the classic time-frequency duality of the Fourier transform.

Now let’s say we take  $\sim K$  Fourier samples of  $\mathbf{x}$ . The probability we will sample a zero element is  $(1 - \tau/N)$ . So the probability that all of our  $M$  random Fourier measurements are zero is  $(1 - 1/K)^M$ . Therefore, *any* method would fail with probability

at least  $1/N$ , if

$$\left(1 - \frac{1}{K}\right)^M \geq \frac{1}{N}$$

i.e.  $M \leq K \log N$

So we need *at least*  $K \log N$  random non-adaptive samples for any (even combinatorially hard) method to work. And, it turns out that a linear program will work fine if  $M$  satisfies (2.16)! Extensions to the case when  $\mu > 1$  are simple (see [57]).

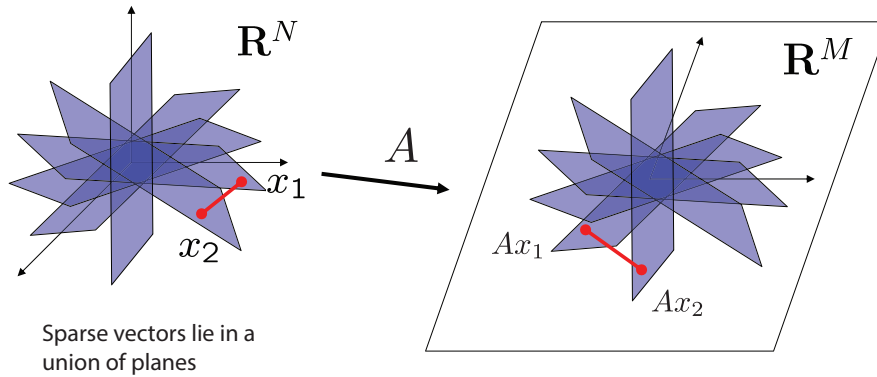
### Restricted isometries

There are two parallel theoretical frameworks of compressive sensing: i) a set of results based on incoherence as discussed above and ii) a complimentary set of results based on a property of the sensing matrix  $A$  called the restricted isometry property or RIP. The theorems based on incoherence, discussed above, rely on taking measurements with randomly selected measurement vectors from a basis which is incoherent with the basis in which the signal is sparse or compressible. However, the theorems which make use of RIP are deterministic and their guarantees hold as long as  $A$  has the requisite RIP property.

Let  $\mathbf{x}_K$  be a  $K$ -sparse vector, then the  $M \times N$  matrix  $A$  is said to have a **restricted isometry constant**  $\delta_K$  of order  $K$  provided that  $\delta_K$  is the smallest scalar which satisfies

$$(1 - \delta_K) \|\mathbf{x}_K\|_2^2 \leq \|A\mathbf{x}_K\|_2^2 \leq (1 + \delta_K) \|\mathbf{x}_K\|_2^2$$

for *all*  $K$ -sparse vectors  $\mathbf{x}_K$ . The matrix  $A$  is said to satisfy the RIP of order  $K$  pro-



**Figure 2.14:** A matrix  $A$  satisfying the RIP of order  $2K$  ( $\delta_{2K}$  is sufficiently small) will approximately preserve the distance between all  $K$ -sparse vectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . Schematic adapted from [55]

vided that the constant  $\delta_K$  is sufficiently small. The RIP is essentially a requirement that all subset of  $K$  columns of  $A$  are *approximately* orthogonal – of course, they can't all be perfectly orthogonal since  $M < N$ . If this condition is satisfied, then the lengths of the sparse vectors are preserved when they are projected on the column space of  $A$ . A simple extension is that if  $A$  satisfies the RIP of order  $2K$  then the distances between  $K$ -sparse vectors are preserved by  $A$ . Another way of stating the RIP is that any submatrix of  $A$  formed by choosing  $K$  columns is well-conditioned (actually the condition number is  $(1 - \delta_K) / (1 + \delta_K)$ ).

A theorem by [20, 53] shows that if  $A$  satisfies  $\delta_{2K} < \sqrt{2} - 1$ , then we essentially get back the guarantees in the previous section with incoherent sampling. More precisely :

- **Theorem** [Candes, Romber, Tao 06]: If  $A$  satisfies  $\delta_{2K} < \sqrt{2} - 1$  then the solution  $\hat{\mathbf{x}}$  to

$$\arg \min \|\tilde{\mathbf{x}}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\tilde{\mathbf{x}}$$

guarantees that

$$\begin{aligned}\|\hat{\mathbf{x}} - \mathbf{x}\|_2 &\lesssim \frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}} \\ \|\hat{\mathbf{x}} - \mathbf{x}\|_1 &\lesssim \|\mathbf{x} - \mathbf{x}_K\|\end{aligned}\tag{2.18}$$

Note that

- The theorem is valid for *all*  $\mathbf{x}$  unlike the incoherent sampling theorems which applied to a fixed  $\mathbf{x}$  and random measurements. If  $\mathbf{x}$  happens to be  $K$ -sparse, then we get exact recovery otherwise we get the best  $K$ -term approximation.
- $\delta_{2K} < 1$  will guarantee that there is a unique  $K$ -sparse solution to  $\mathbf{y} = A\mathbf{x}$ , but you have to solve a NP-hard problem to find it! However, the above theorem states that more stringent requirement  $\delta_{2K} < \sqrt{2} - 1$  will not only guarantee that the solution is unique, but the  $L_1$  problem will find it!

The way we have stated the theorem (2.18) makes no mention of the number of measurements or randomness! All it says is that if  $A$  satisfies the RIP of order  $2K$  then we get back the best  $K$ -sparse approximation by solving  $L_1$ . So we are left with the task of constructing matrices which satisfy the RIP for which  $M$  is close to  $K$ . Calculating the restricted isometry constant of a matrix is actually a NP-hard problem [58].

This is where randomness enters the picture: random matrices satisfy the RIP (for  $\delta_{2K}$ ) with very high probability. More precisely, let  $A$  be an  $M \times N$  random matrix with the entries  $A_{ij}$  sampled i.i.d from a distribution  $F$ . Then  $A$  will satisfy the RIP

( $\delta_{2K} < \sqrt{2} - 1$ ) with very high probability provided

$$M \gtrsim K \log \left( \frac{N}{K} \right) \quad \text{for } F \equiv \mathcal{N} \left( 0, \frac{1}{\sqrt{M}} \right)$$

$$M \gtrsim K \log \left( \frac{N}{K} \right) \quad \text{for } F \equiv \text{Bernoulli} \left( \pm \frac{1}{\sqrt{M}} \right)$$

There are other distributions  $F$  for which the RIP holds with  $M$  being slightly more than above [59]. So, there's a kind of universality in sensing with random matrices. Also, the RIP holds for structured matrices as well; for e.g., a Fourier (DFT) matrix satisfies the RIP provided  $M \gtrsim K (\log N)^4$ [20]. An incoherent sampling matrix satisfies RIP with high probability provided  $M \gtrsim K \mu^2 (\log N)^4$ . Another remarkable result due to [25, 56] is that no other sensing mechanism – adaptive or non-adaptive – or any other reconstruction algorithm can do better with substantially fewer samples, provided that the signal is sparse or compressible in a power law sense (2.11).

Random matrices also have the desirable property that they are universal sensing matrices in some sense. If  $\mathbf{x}$  is sparse in a basis  $B : \mathbf{x} = B\alpha$ , and we take measurements using a random matrix  $A : \mathbf{y} = A\mathbf{x} = AB\alpha$ , then if the matrix  $A$  satisfies the RIP, then so will  $AB$ . Thus, we are guaranteed to recover the sparse coefficients  $\alpha$  even if we don't know the sparsity basis *a priori* – the measurements can be completely non-adaptive.

To see this, we can show that if  $[A]_{ij} \sim i.i.d \mathcal{N}(0, \sigma^2)$  satisfies RIP ( $\delta_{2K}$  is small) with high probability, then  $A \cdot B$  also satisfy this property for an orthonormal matrix  $B$ . Let us look at the statistics of the element of  $A \cdot B$ . Each entry of the matrix is

a linear combination of Gaussians, and hence will be a Gaussian itself.

$$\begin{aligned}
[AB]_{ij} &= \sum_{k=1}^N A_{ik} B_{kj} & (2.19) \\
\langle [AB]_{ij} \rangle &= \sum_{k=1}^N \langle A_{ik} \rangle B_{kj} = 0 \\
\langle [AB]_{ij}^2 \rangle &= \sum_{k=1}^N \sum_{l=1}^N \langle A_{ik} A_{il} \rangle B_{kj} B_{lj} \\
&= \sum_{k=1}^N \sum_{l=1}^N \delta_{kl} \sigma^2 B_{kj} B_{lj} \\
&= \sigma^2 \sum_{k=1}^N B_{kj}^2 = \sigma^2
\end{aligned}$$

where (2.19) follows because rows and columns of  $B$  have unit norm. So, the elements of the matrix  $AB$  are also Gaussians with zero mean and variance  $\sigma^2$ . Let us see if they are independent

$$\begin{aligned}
\langle [AB]_{ij} [AB]_{rs} \rangle &= \left\langle \sum_{k=1}^N \sum_{l=1}^N A_{ik} A_{rl} B_{kj} B_{ls} \right\rangle \\
&= \sum_{k=1}^N \sum_{l=1}^N \delta_{ir} \delta_{kl} \sigma^2 B_{kj} B_{ls} \\
&= \delta_{ir} \sigma^2 \sum_{k=1}^N B_{kj} B_{ks} \\
&= \delta_{ir} \delta_{js} \sigma^2
\end{aligned}$$

where, the first dirac delta comes from the fact that entries of  $A$  are uncorrelated, and the second one comes from the fact that the rows of  $B$  are orthonormal. Therefore entries of  $AB$  have the same joint distribution as entries  $A$  and it will also satisfy RIP with high probability!

## 2.8.6 Compressive sensing with noise

Now, let us review the theorems when there is noise in the measurements.

It turns out that we can still get good reconstruction and the performance degradation is graceful. Let the matrix  $A$  satisfy RIP ( $\delta_{2K} < \sqrt{2} - 1$ )– the distances between projections of  $K$ -sparse signals will be preserved– and let the measurements be noisy

$$\mathbf{y} = A\mathbf{x} + \mathbf{z} \quad \text{where } \langle \mathbf{z}_i^2 \rangle = \sigma^2$$

Then [25] showed that we can get a good reconstruction by solving a different optimization problem(LASSO)

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \left[ \|\mathbf{y} - A\mathbf{x}\|^2 + \lambda \sum_{i=1}^N |\mathbf{x}_i| \right]$$

For properly tuned  $\lambda$  the theorem states that

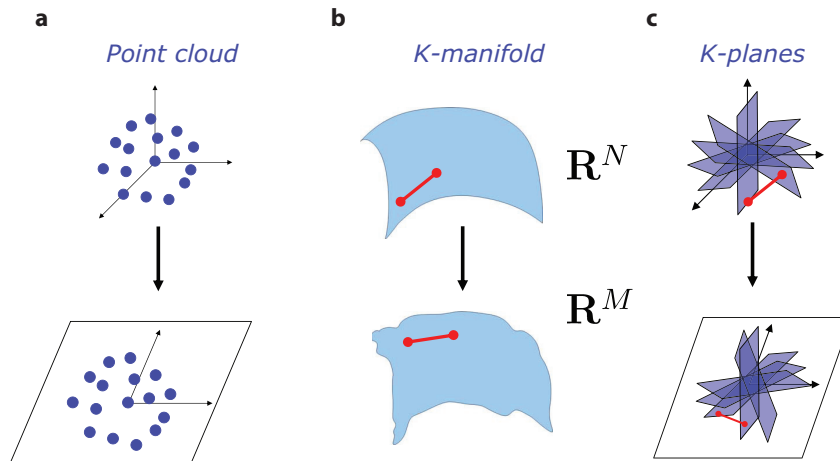
$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C_1 K \sigma + C_2 \frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}} \quad (2.20)$$

The result cannot be any better. It states that the reconstruction error in the noisy case is bounded by the reconstruction error of the noiseless case, plus a term that scales linearly with the noise.



## 2.8.7 Random projections and stable embeddings

We already saw that random  $M \times N$  matrices obey the RIP for sparse signals with high probability provided  $M$  is sufficiently large. Now we briefly discuss another surprising property of random projections : random projections also provide “stable” low dimensional embeddings [18] for a variety of other signals with nonlinear structure other than sparsity. If the signals  $\mathbf{x}$  reside in a high dimensional space of ambient dimension  $N$ , but has some low-dimensional structure (like a manifold) of effective dimensionality  $K$ , then a random projection of the signal to a space with dimension  $M \sim K$  will preserve local distance between the points in the high dimensional space. Thus, random projections can be used for dimensionality reduction in a *non-adaptive* way.



**Figure 2.15:** Random projections provide stable embeddings from  $\mathbb{R}^N \rightarrow \mathbb{R}^M$  for a) point clouds b)  $K$ -manifold and c)  $K$ -planes (i.e  $K$ -sparse signals) provided  $M$  is comparable to  $K$  or  $\log(\#$  of points in cloud). Schematic adapted from [18]

Let us look at one particular result with this flavour: the **Johnson-Lindenstrauss** lemma [60, 18]. The lemma shows that any  $K$  point set in an Euclidean space (say,  $\mathbb{R}^N$ ) can be *linearly* embedded in a space of dimension  $O(\log K/\epsilon^2)$  without distorting pairwise distances by more than  $(1 \pm \epsilon)$ . Specifically, consider a set of  $K$  points in

$\mathbb{R}^N$ , then the lemma shows that there exists a linear map  $A : \mathbb{R}^N \rightarrow \mathbb{R}^M$  which preserves the pair-wise distances up to  $(1 \pm \epsilon)$  for all the points  $u, v$  for  $\epsilon \in (0, 1/2)$  and  $M \sim \log K/\epsilon^2$

$$(1 - \epsilon) \|u - v\|_2^2 \leq \|Au - Av\|_2^2 \leq (1 + \epsilon) \|u - v\|_2^2$$

This result is tight – you cannot do the embedding into a substantially lower dimensional space without distorting the distances a lot [60]. Moreover, the linear map  $A$  can be constructed by populating the  $M \times N$  matrix by i.i.d entries from the same distributions  $F$  that were suitable for RIP :  $F \equiv \mathcal{N}\left(0, 1/\sqrt{M}\right)$  and  $F \equiv \text{Bernoulli}\left(\pm 1/\sqrt{M}\right)$ ! The proof is not complicated [60] and uses the following ideas:

- If  $A_{ij} \sim \mathcal{N}(0, 1)$ , then using the **Hoeffding** concentration inequality[61] it's easy to show that the lengths of vectors are concentrated around the mean

$$\text{P}\left(\frac{1}{\sqrt{M}} \|A\mathbf{x}\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2\right) \leq \exp\left(-\frac{M}{4} (\epsilon^2 - \epsilon^3)\right)$$

- From which we see that probability of *one* pair of distances getting distorted is exponentially small in  $M$

$$\text{P}\left((1 - \epsilon) \|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1 + \epsilon) \|\mathbf{x}\|_2^2\right) \geq 1 - 2e^{-M(\epsilon^2 - \epsilon^3)/4}$$

- There are  $O(K^2)$  pairs, so use the union bound to show that the RHS holds for all pairs with non-zero probability provided  $M \sim O(\log K/\epsilon^2)$ .

A similar covering argument like this can be extended to show that signals lying on other low-dimensional structures like manifolds or union of planes (sparsity) can be stably embedded by a random projection in a space with dimensionality  $M$  comparable to the effective dimensionality of the low dimensional structure [18]. This is illustrated in the schematic in Fig. 2.15. Thus, random projections give us a way to do proximity-preserving dimensionality reduction in a non-adaptive way! This is very useful for practical applications, because a lot of natural signals like images or sound have a sparsity or smooth manifold structure, so we can first project them randomly (and non-adaptively) to a lower dimension and then perform computational tasks such as clustering or learning in the lower dimensional representation. This paradigm suggests a counterintuitive strategy for the brain – represent structured stimuli: use random receptive fields!

# Chapter 3

## Arousal-related adjustments of perceptual biases optimize perception in dynamic environments

Most of this section appears in:

K. Krishnamurthy\*, M.R. Nassar\*, S. Sarode and J.I. Gold  
*Nature Human Behaviour* 1 (2017): 0107

### 3.1 Abstract

Prior expectations can be used to improve perceptual judgments about ambiguous stimuli. However, little is known about if and how these improvements are maintained in dynamic environments in which the quality of appropriate priors changes from one stimulus to the next. Using a sound-localization task, we show that changes

in stimulus predictability lead to arousal-mediated adjustments in the magnitude of prior-driven biases that optimize perceptual judgments about each stimulus. These adjustments depend on task-dependent changes in the relevance and reliability of prior expectations, which subjects update using both normative and idiosyncratic principles. The resulting variations in biases across task conditions and individuals are reflected in modulations of pupil diameter, such that larger stimulus-evoked pupil responses correspond to smaller biases. These results suggest a critical role for the arousal system in adjusting the strength of perceptual biases with respect to inferred environmental dynamics to optimize perceptual judgements.

## 3.2 Introduction

Perception is shaped by prior expectations (priors) on the statistical structure of the sensory world [62, 63, 64, 65, 66, 67]. When the environmental statistics are stationary and well known, priors on those statistics can bias the perception of relevant sensory stimuli [68, 69]. For example, the prevalence of relatively slow- versus fast-moving objects in the world can lead to biases in the perception of object speed [8]. However, many environmental statistics that are relevant to perception can be highly non-stationary. For example, the locations of sources of sensory input are constantly changing relative to a given observer. The goal of this study was to examine how priors on such dynamic features of the environment are updated and used to shape perception.

To achieve this goal, we developed an auditory-localization task that required human subjects to both predict and report the perceived location of a simulated sound source as the predictability of the location varied over time (Fig. 3.1ac). The statistical

structure of the task is similar to ones we used previously to show that people can make effective predictions in dynamic environments by adaptively modulating the influence of new information on existing beliefs [70, 71]. However, here we focus on the questions of if and how such dynamically modulated predictions affect their influence on the perception of ambiguous stimuli. In principle, these predictions could govern perceptual biases through a form of optimal (Bayesian) inference that takes into account dynamic changes in the priors [70, 72, 73]. Specifically, as long as the statistical structure of the sampled locations in our task remains stable, new sounds can be used to develop increasingly reliable priors about the locations of subsequent sounds. These increasingly reliable priors should, in turn, have an increasingly strong and beneficial influence on the perception of those sounds, reducing localization errors (Fig. 3.1d,e). However, the statistics of the sampled locations can undergo abrupt change-points that render previously held priors irrelevant to new sounds. These seemingly reliable but irrelevant priors should not influence the perception of sound-source location, which under these conditions should be limited entirely by sensory uncertainty (Fig. 3.1f).

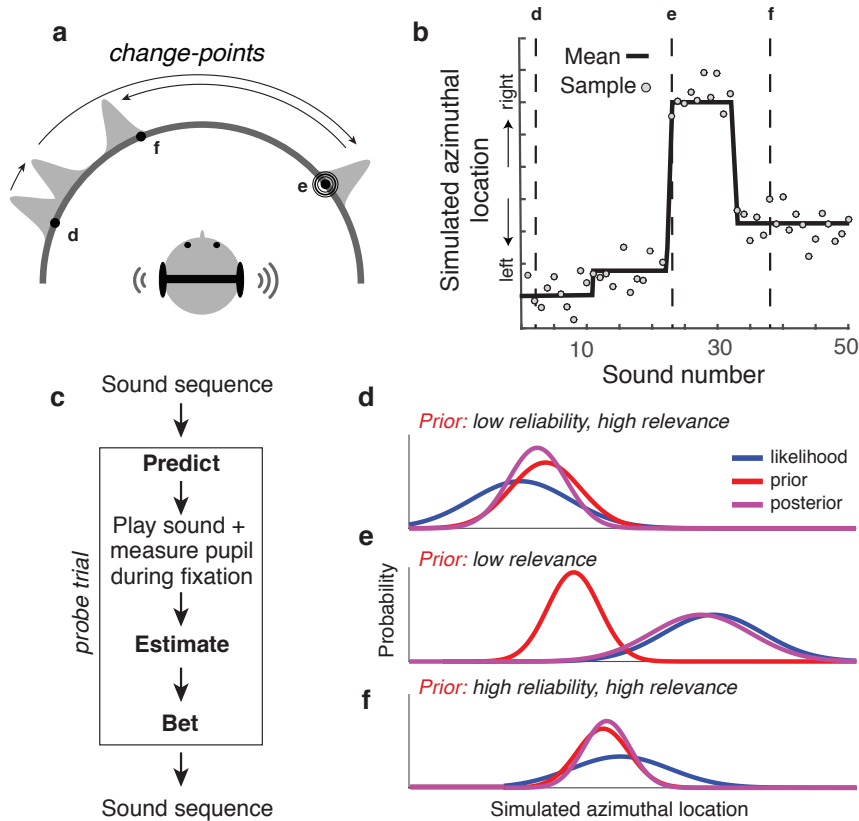
We also measured pupil diameter, an index of arousal that can reflect the activation of the locus coeruleus (LC)-norepinephrine (NE) system and has been implicated in rapidly updating inference processes in response to unexpected events or errors [74, 75, 76, 77, 78, 79]. Pupil diameter tracks the extent to which predictions are updated in response to new information in dynamic and perceptually unambiguous cognitive tasks [71]. Here we tested the hypothesis that such changes in arousal play an important role in shaping perception. In particular, we examined whether the arousal system controls the extent to which perceptual judgments about ambiguous sensory stimuli are biased toward prior expectations in accordance with the relevance

and reliability of those expectations.

Our results yield new insights into the relationship between perception and arousal. We show that the subjects' priors had a variable influence on their perceptual reports. This variability was predicted by changes in the relevance and reliability of those priors, across task conditions and individual subjects. These effects were encoded in both baseline and stimulus-evoked changes in pupil diameter, such that larger diameters corresponded to less influence of priors on the perception of that stimulus. Taken together, these findings support a fundamental role for pupil-linked arousal systems, including the LC-NE system, in adaptively adjusting the influence of priors on perception in accordance with environmental dynamics.

### 3.3 Results

Twenty-nine subjects performed both the dynamic localization task (Fig. 3.1) and a control task that required perceptual reports of simulated sound-source locations that lacked predictable, sequential structure. Overall, the subjects tended to perform both tasks in an effective manner, providing predictions on the dynamic task and perceptual reports on both tasks that corresponded strongly to the simulated sound-source locations (Fig. 3.2). On the control task, the Pearson's correlation between simulated and reported location had median [interquartile range, or IQR] values of 0.926 [0.8950.944] across subjects (Fig. 3.2a,d). On the dynamic task, there were similarly high correlations for both the predictions and perceptual reports (predictions on non-change-point trials:  $r=0.907$  [0.8950.921], Fig. 3.2b,e; perceptual reports on all trials:  $r=0.948$  [0.9410.964], Fig. 3.2c,f). However, the subjects also tended to make errors that varied considerably from trial to trial on both tasks (Fig. 3.2gi).



**Figure 3.1:** (a) Subjects listened via headphones to noise bursts with virtual source locations that varied along the frontal, azimuthal plane. The locations were sampled (points) from a Gaussian distribution (gray) with a mean that changed abruptly on unsignaled change-points (probability=0.15 for each sound) and a STD of  $10^\circ$  in low-noise blocks,  $20^\circ$  in high-noise blocks. The subjects listened passively to the sound sequence, except for occasional probe trials. All sounds except the probe sound were presented simultaneously with their corresponding locations on a semicircular arc shown on the isoluminant visual display, allowing subjects to develop priors on sound-source location based on both the auditory and visual signals and maintain a stable mapping between the two. (b) An example trial sequence showing the mean (solid line) and sampled (points) locations over 50 trials. Vertical dashed lines indicate randomly selected probe trials. (c) Probe-trial sequence. Using a mouse to control a cursor on the visual display, the subject reported: 1) the predicted location of the upcoming probe sound, followed by 250-ms fixation, presentation of the probe sound, then continued fixation for 2.5 s to allow for pupil measurements; 2) the estimated location of the probe sound; and 3) a high or low confidence report that the true location was within a small window centered on their estimate. The sound sequence then continued until the next probe. (d-f) Schematic illustrating the changing reliability and relevance of priors for the probe sounds in a and b, as indicated. Given a fixed-width likelihood function, more reliable and relevant priors have a stronger and more beneficial influence on the percept, here represented as the posterior, which is most uncertain (widest) in e and least uncertain in f.

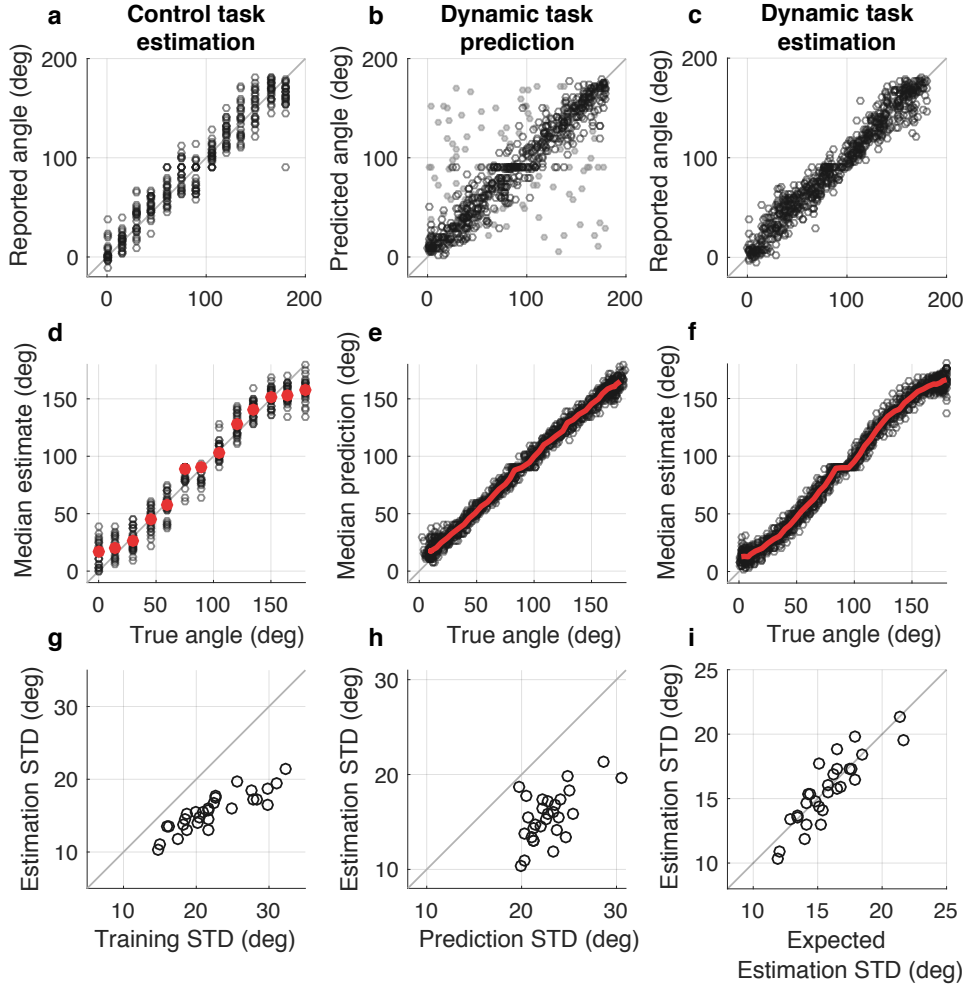


Subsequent analyses focus on how the subjects minimized their errors on the dynamic task by exploiting the fluctuating predictability of sound-source locations on that task.

### 3.3.1 Dynamic, task-dependent modulation of perceptual biases

The subjects used both sensory and prior information to guide their perceptual reports on the dynamic task. We measured performance in terms of the variability of the distribution of trial-by-trial errors (quantified as the standard deviation, or STD, and denoted as  $\sigma$ ). This variability was lower for perceptual reports on the dynamic task than for either: 1) predictions from that task ( $\sigma_{prior}$ ; Fig. 3.2h), or 2) perceptual reports on the control task that lacked sequential predictability and thus reflected more purely sensory processing ( $\sigma_{sensory}$ ; Fig. 3.2g). Moreover, for individual subjects, these different measures of variability were related to each other, such that perceptual errors from the dynamic task were well approximated using the optimal, reliability-weighted combination of prior and sensory information ( $\sigma_{sensory}^{-2} = \sigma_{sensory}^{-2} + \sigma_{prior}^{-2}$ ; Fig. 3.2i). This result implies that, on average, the subjects tended to not only use these two sources of information, but also combine them according to their relative reliabilities to optimize perceptual performance on the dynamic task.

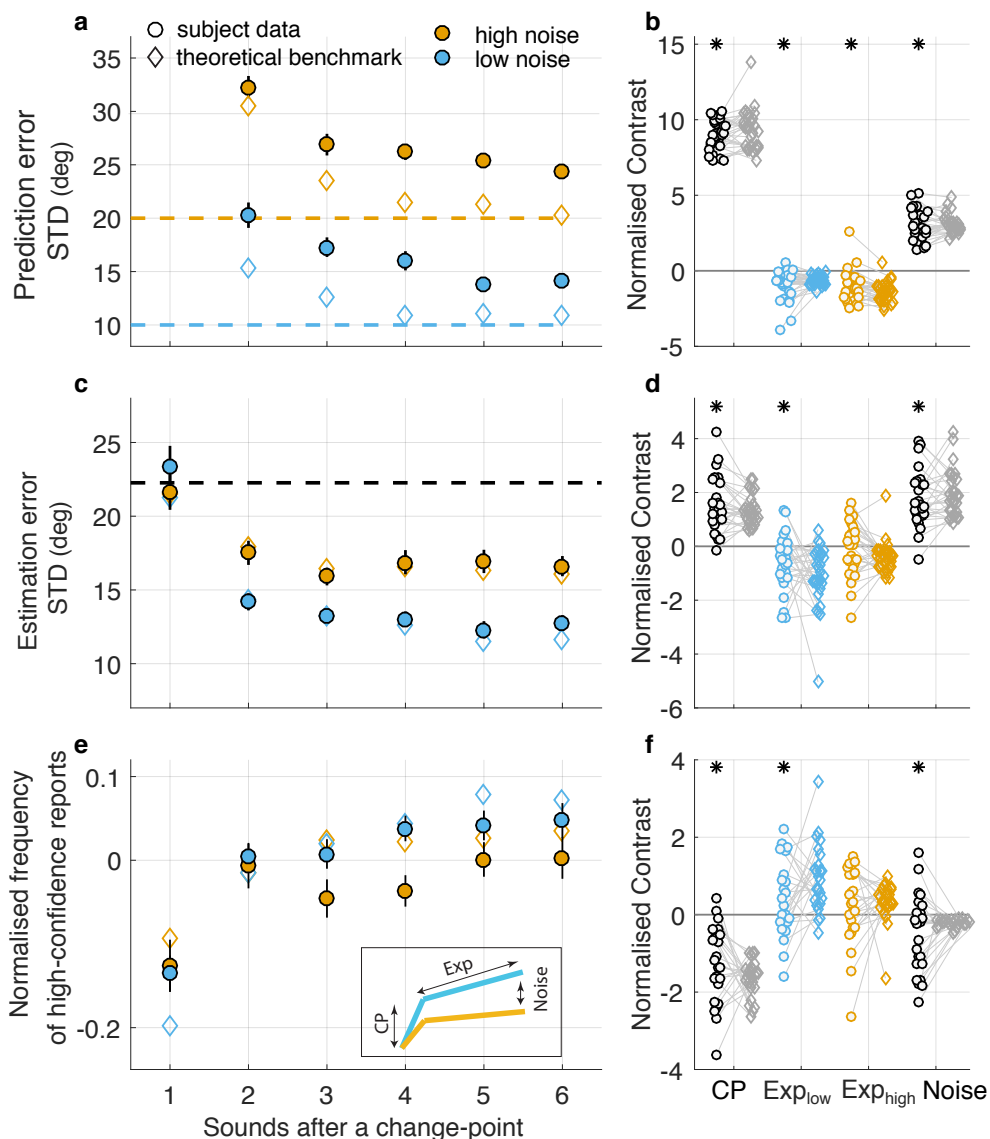
This integration of prior and sensory information took into account the changes in the relevance and reliability of the priors that occurred throughout the dynamic task. These changes are illustrated in Fig. 3.3a, which shows prediction-error STDs averaged across subjects as a function of the number of sounds after a clearly noticeable change-point, or SAC (see legend for details), separately for the two noise conditions. Figure 3.3b shows linear contrasts that captured the salient, dynamic aspects of these changes for each subject (see inset in Fig. 3.3e illustrating the three contrasts: CP,



**Figure 3.2: Overall prediction and estimation performance.** (ac) Reported versus true (simulated) sound-source angle for an example subject for: (a) estimations from the control task; (b) predictions from the dynamic task (light gray points indicate change-point trials, on which the probe location was, by design, unpredictable); and (c) estimations from the dynamic task, including all trials. (df) Population summaries, plotted as in (ac), with per-subject median values shown in black and the median of medians shown in red ( $n=29$  subjects). For the dynamic tasks, median values were calculated in sliding  $20^\circ$  windows. Non-change-point trials were excluded from the predictions in (e). Note that the subjects perceptual reports (d and f) were biased slightly towards straight ahead at the far periphery. This bias, which likely reflected learned expectations that sounds were only played in the frontal plane, is accounted for in later analyses ( $\beta_5$  and  $\beta_6$  in Eq. 5). (gi) STD of the perceptual errors from the dynamic task plotted versus the STD of: (g) the perceptual errors from the control task; (h) the prediction errors from the dynamic task; or (i) the expected STD of the perceptual errors, computed from the optimal, reliability-weighted combination of the control perceptual errors and the dynamic prediction errors. Points in gi represent data from individual subjects. Prediction and perceptual errors were computed with respect to the simulated location of the probe sound.

describing the effects of a noticeable change-point; Exp, describing the effects of the number of sounds experienced following a noticeable change-point; and Noise, describing the high or low noise condition). Specifically, on change-point trials, predictions were irrelevant and hence most variable with respect to the subsequent sound-source location (signed-rank test for  $H_0$ : the median of the distribution of per-subject CP contrasts, which compared change-points to other trials=0,  $p < 10^{-5}$ ). After change-points, predictions became steadily more reliable as the number of sound sources experienced from the new distribution increased in both noise conditions ( $p < 10^{-4}$  for Exp<sub>low</sub> and Exp<sub>high</sub> contrasts, which identified linear trends across SAC 26 for each of the two noise conditions). The predictions were also more reliable overall in the low- versus high-noise condition (Noise contrast,  $p < 10^{-5}$ ). These dynamic trends were consistent with predictions from a normative model of predictive inference that had full knowledge of the generative statistics [70]. The model, which produced simulated predictions that were analyzed in the same way as the data, had task-dependent effects that were in the same directions and of roughly the same magnitude as the data, although the subjects tended to produce more variable predictions than the model (Fig. 3.3a,b diamonds).

These task-dependent changes in the subjects predictions were associated with similar changes in the variability of their perceptual reports (Fig. 3.3c,d) and their confidence in those reports, as assessed by the frequencies of high-confidence reports (Fig. 3.3e,f). Perceptual-error variability tended to be higher for change-point trials, when predictions were irrelevant (CP contrast,  $p < 10^{-5}$ ), and for the high- versus low-noise condition (Noise contrast,  $p < 10^{-5}$ ). Perceptual-error variability also tended to decrease on experiencing more samples from the new distribution, with a reliable effect across individuals in the low-noise condition (Exp<sub>low</sub> contrast,  $p < 0.005$ ) but not the



**Figure 3.3: Effects of task dynamics on performance.** (a) STD of the subjects prediction errors (filled circles) as a function of the number of sounds after a change-point (SAC) in the generative mean azimuthal location, plotted separately for the two noise conditions (colors, as indicated; generative STDs are shown as dashed lines). For comparison, prediction-error STDs are shown for an approximately optimal predictive-inference model (open diamonds). Data from change-point trials (SAC=1) are not shown because locations were, by design, unpredictable on those trials.

high-noise condition ( $Exp_{high}$  contrast,  $p = 0.4$ ). These dynamics were also apparent in the subjects' confidence report trends (Fig. 3.3e,f), which reflected trial-by-trial awareness of the changes in perceptual variability and included similar dependencies

**Figure 3.3: Effects of task dynamics on performance, continued (b)** Contrast values from a linear model describing individual subject (circles) and the approximately optimal model (each diamond represents analyses based on the same sound sequence experienced by the subject connected by a line) prediction-error STD in terms of (see inset in **e**): 1) the difference between change-point and non-change-point trials (CP), 2,3) the linear trend from SAC 2-6 for low- ( $\text{Exp}_{\text{low}}$ ) or high- ( $\text{Exp}_{\text{high}}$ ) noise trials, and 4) the difference between the two noise conditions (Noise). (**c,d**) Same conventions as in **a,b** but for perceptual errors on the dynamic task. Diamonds represent the theoretically predicted STD of perceptual errors computed from the optimal, precision-weighted combination of the subject- and condition-specific STDs of prior errors (circles in **a**, determined separately for each subject) and the subject-specific estimation-error STDs from the control task (the median value is shown as a horizontal dashed line; see Fig. 3.3g). (**e,f**) Same conventions as in **a,b** but for the frequency of high-confidence reports relative to overall frequency of high-confidence reports per subject. Diamonds represent the frequency of high-confidence reports corresponding to the theoretical perceptual errors in **c**, computed from the fraction of the theoretical posterior distribution within the confidence window. In **a,c,e**, circles and error bars are mean $\pm$ sem of values measured from all 29 subjects. In **b,d,f**, points are data from individual subjects. Asterisks indicate sign-rank test for  $H_0$ : median value from the subject data=0,  $p < 0.05$ . In each case, paired rank-sum test for  $H_0$ : median difference between subject data and theoretical prediction,  $p > 0.087$ . In all panels, only data from sequences following noticeable change-points (changes in mean of at least twice the generative STD for SAC=1) were included.

---

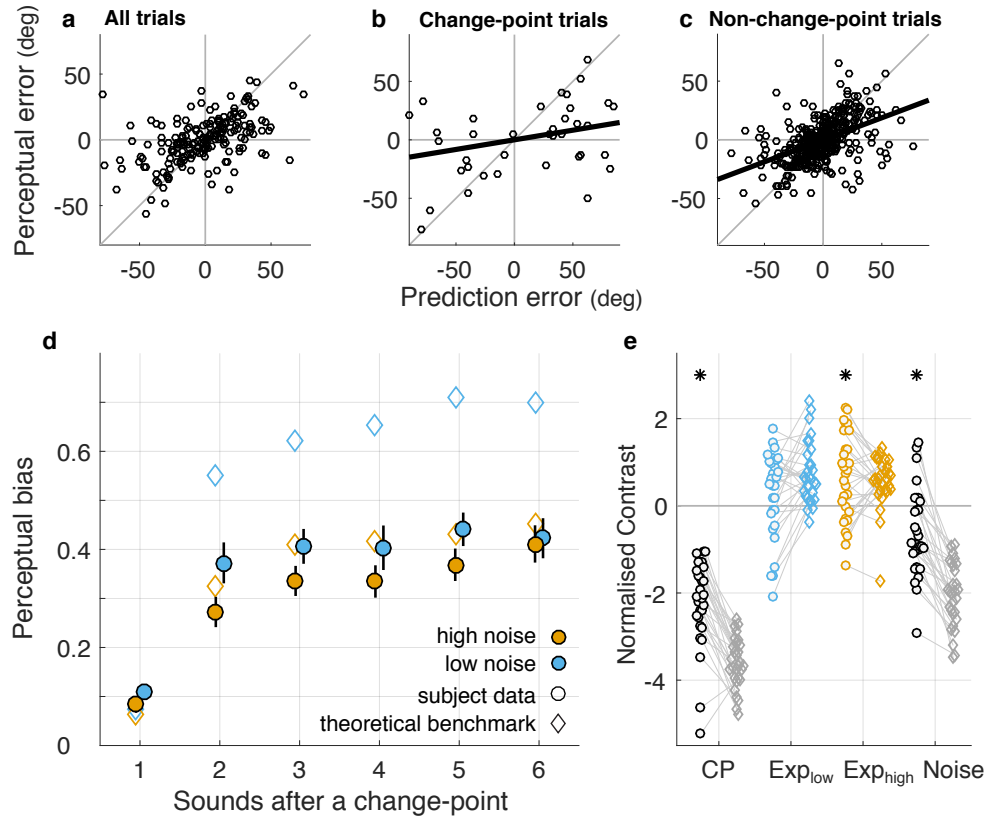
on CP ( $p < 10^{-4}$ ), Noise ( $p = 0.032$ ), and  $\text{Exp}_{\text{low}}$  ( $p = 0.03$ ) and less reliable dependencies on  $\text{Exp}_{\text{high}}$  ( $p = 0.07$ ). Both the perceptual and confidence report effects were qualitatively similar, in direction and magnitude, to theoretical values computed from optimal combinations of each subjects' changing priors (circles in Fig. 3.3a,b) and their fixed sensory reliability estimated from the control task (Fig. 3.2g; see also Fig. 3.1df). These theoretical values also showed strong effects of CP, Noise, and  $\text{Exp}_{\text{low}}$ , and smaller effects of  $\text{Exp}_{\text{high}}$  (Fig. 3.3cf, diamonds).

These behavioral dynamics reflected changes in the degree to which the subjects priors biased their perceptual reports. We quantified perceptual bias as the slope of the relationship between the prediction error and the perceptual error measured on individual trials (Fig. 3.4a-c). A slope of zero implies no relationship between the prediction error and the perceptual error, and thus no bias towards the prior. In contrast, slope values that increase towards unity imply increasing biases of the perceptual reports towards the prior. This perceptual bias varied systematically as a

function of task conditions. Specifically, perceptual bias was lower on change-points (CP contrast,  $p < 10^{-5}$ ) and for the high- versus low-noise condition (Noise contrast,  $p = 0.008$ ). Perceptual bias also tended to increase on experiencing more samples, although these effects were variable across individuals and not statistically reliable in the low-noise condition ( $\text{Exp}_{low}$  contrast,  $p = 0.1$ ;  $\text{Exp}_{high}$  contrast,  $p = 0.004$ ). These task-dependent changes in the biases were comparable in direction and magnitude to theoretically computed values given an optimal, reliability-weighted combination of the task-specific predictions on the dynamic task (circles in Fig. 3.3a) and fixed sensory reliability estimated from the control task (Fig. 3.2g), computed separately for each subject (diamonds in Fig. 3.4d,e). Despite these comparable task-dependent trends (compare circles and diamonds in Fig. 3.4e), the subjects perceptual biases were on average smaller than the theoretical values (compare circles and diamonds in Fig. 3.4d). This shift was consistent with their overall worse predictions than the model (compare circles and diamonds in Fig. 3.3a). However, overall performance, measured as perceptual-error variability, was relatively insensitive to this overall shift, as compared to task-dependent adjustments, in the magnitude of the perceptual biases (compare circles and triangles in Fig. 3.3c,d).

### **3.3.2 Individual differences in the modulation of perceptual biases**

The above analyses demonstrated that for individual subjects, dynamic changes in the relevance and reliability of priors within an experimental session were associated with changes in the degree to which those priors biased perception. We identified similar effects across subjects, implying that individual differences in perception can reflect differences in how priors are updated and maintained in dynamic environments.

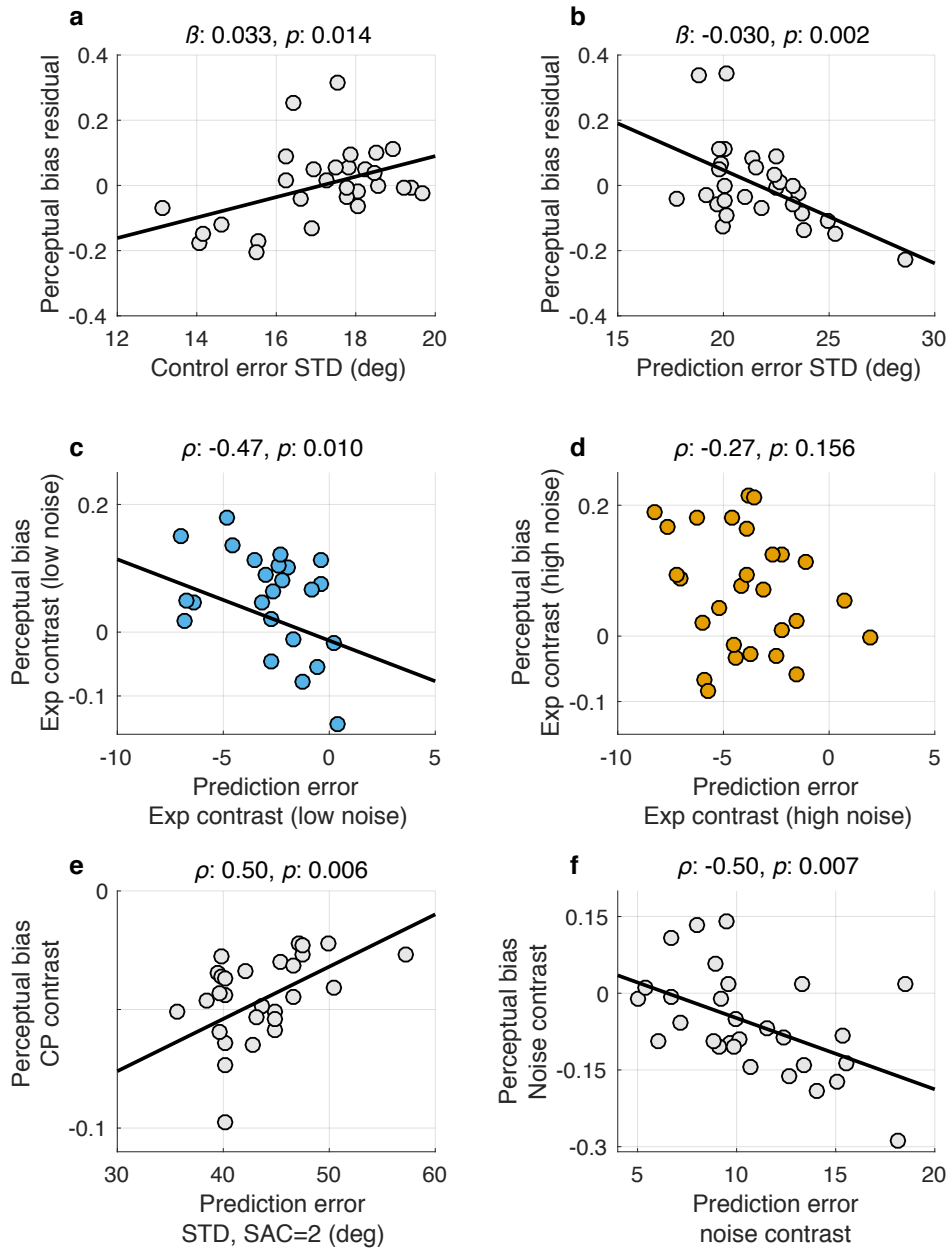


**Figure 3.4: Effects of task dynamics on perceptual bias.** (a-c) Example data from a single subject illustrating the quantification of perceptual bias as the slope of the best-fit line to a scatter of the perceptual error versus the prediction error. Slopes close to zero reflect a low perceptual bias (i.e., the percept is unrelated to the prediction), as on change-point trials (b). Slopes closer to unity reflect a higher perceptual bias (i.e., the percept more closely matches the prediction), as on non-change-point trials (c). (d) Perceptual bias as a function of the number sounds after a change-point (SAC) in the generative mean azimuthal location, plotted separately for the two noise conditions (colors, as indicated). Circles and error bars are mean $\pm$ sem of values measured from all 29 subjects. Diamonds indicate the theoretically predicted perceptual bias from an optimal, reliability-weighted combination of the subject- and condition-specific predictions (Fig. 3.3a) and the subject-specific estimates from the control task (Fig. 3.2g). (e) Contrast values from a linear model describing individual subject (circles) and model (each diamond represents analyses based on the same sound sequence experienced by the subject connected by a line) perceptual bias in terms of (see inset in Fig. 3.3e): 1) the difference between change-point and non-change-point trials (CP), 2,3) the linear trend from SAC 26 for low- (Exp<sub>low</sub>) or high- (Exp<sub>high</sub>) noise trials, and 4) the difference between the two noise conditions (Noise). Asterisks indicate sign-rank test for  $H_0$ : median value from the subject data=0,  $p < 0.05$ . Paired rank-sum tests for  $H_0$ : median difference between subject data and theoretical prediction,  $p < 0.01$  for CP,  $p = 0.16$  for Exp<sub>low</sub>,  $p = 0.78$  for Exp<sub>high</sub>, and  $p < 0.01$  for Noise. In d and e, only data from sequences following noticeable change-points (changes in mean of at least twice the generative STD for SAC=1) were included.

Specifically, we compared subjects' overall biases to the variability of their sensory and prediction errors (linear regression of the mean perceptual biases of individual subjects from non-change-point trials as a function of the STD of perceptual errors from the control task and the STD of prediction errors across non-change-point trials from the dynamic task;  $F$  statistic=7.39,  $p = 0.002$ ). According to these fits and consistent with Bayesian theory, subjects with higher overall prior-driven perceptual biases tended to have higher sensory variability ( $\beta = 0.033$ , t-test for  $H_0: \beta = 0$ ,  $p = 0.013$ ; Fig. 3.5a) and lower prediction variability ( $\beta = -0.030$ ,  $p = 0.002$ ; Fig 3.5b). We also found individual differences in how perceptual biases changed as a function of particular task conditions, and that those differences were predicted by subject-specific changes in priors under those conditions. Subjects whose priors improved (i.e., became less variable) the most also tended to have the largest increases in prior-driven perceptual biases: 1) just after a change-point (Fig. 3.5e), 2) on experiencing samples from a new distribution (in the low- but not high-noise condition; Figs. 3.5c and d), or 3) between the high- and low-noise conditions (Fig. 3.5f). Thus, on average, subjects tended to weigh prior and sensory information according to their relative reliabilities, taking into account variability in the priors across task conditions and individual subjects.

To more quantitatively account for the factors that affected perceptual biases across task conditions and individual subjects, we used a linear model that included normative and non-normative terms that each were weighed according to their contributions to each subjects behavior (Fig. 3.6). Data generated by a purely normative model could capture some qualitative aspects of behaviour, but it systematically overestimated perceptual biases (Fig 3.6A). A linear model that included both normative and non-normative terms offered a better description of behaviour (Fig. 3.6B). The nor-



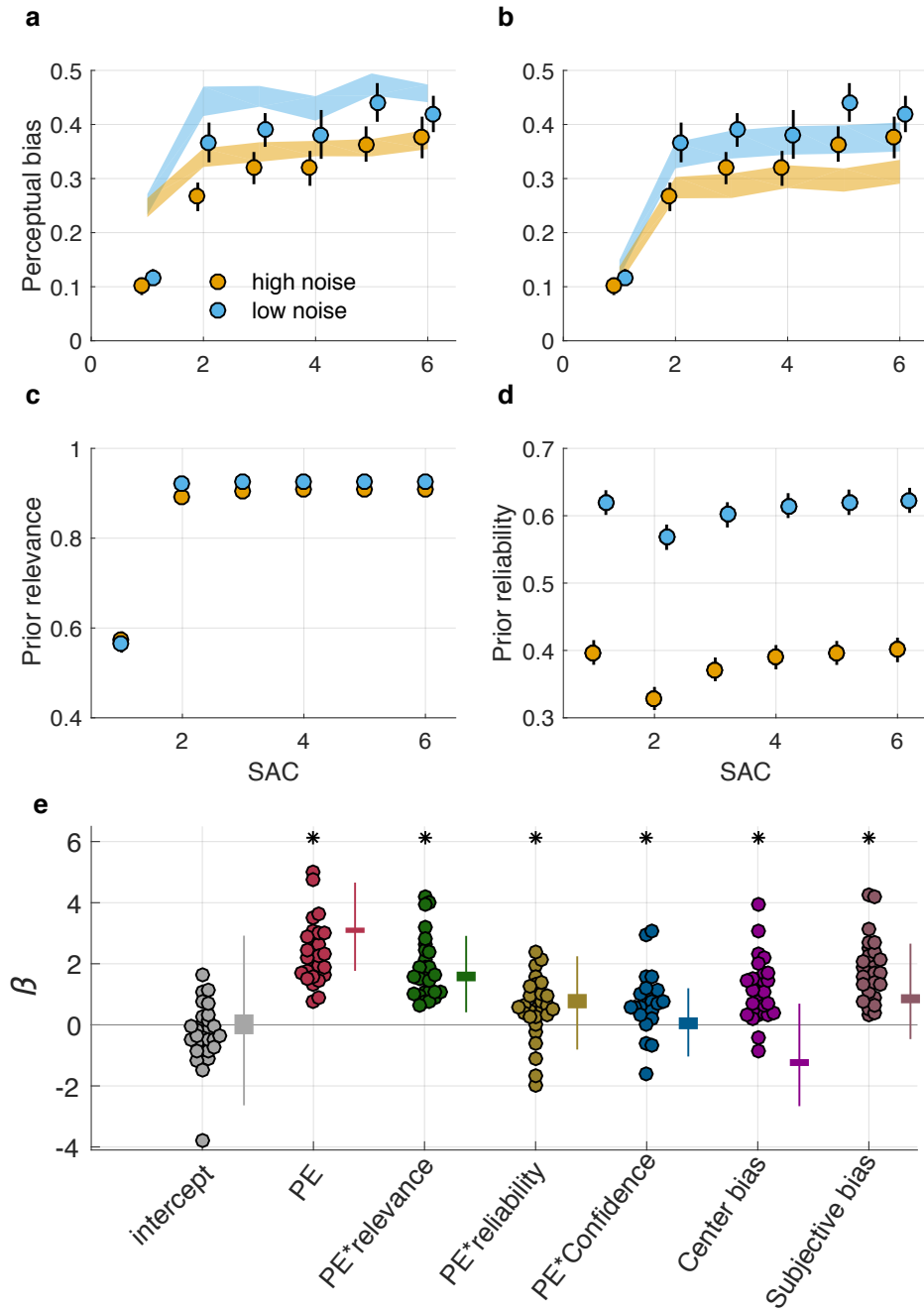


**Figure 3.5: Individual differences in perceptual bias.** (a, b) Relationship between overall (mean) perceptual bias and either overall localization ability (STD of perceptual errors on the control task, a) or overall prediction ability (STD of prediction errors from non-change-point trials on the dynamic task, b), after accounting for the other factor (hence residual) via linear regression. (c-f) The dependence of perceptual bias on various task conditions, plotted as functions of the dependence of prediction-error STD on the same conditions: c, d) the linear trend from SAC 26 in the low-noise (c) and high-noise (d) condition (Exp); e) change-point versus non-change-point trials (CP); and f) high- versus low-noise condition (Noise). In each panel, points represent data from individual subjects. Lines are linear regressions. Only data from sequences following noticeable change-points (changes in mean of at least twice the generative STD for SAC=1) were included.

native terms were extracted from a Bayesian model of perception, which generated perceptual biases that minimized simulated perceptual errors, given each subjects' variable predictions and sensory estimates. These terms were: 1) prior relevance, which reflected the probability that the current sound came from the same generative distribution as the previous sound (and thus is related to the CP effects illustrated in Figs. 3.3 and 3.4; Fig. 3.6c); and 2) prior reliability, which reflected changes in the total width of the predictive distribution relative to the likelihood, given new samples (and thus is related to the Exp and Noise effects illustrated in Figs. 3.3 and 3.4; Fig. 3.6d). The non-normative terms included one describing a fixed bias as a function of the prediction error, one to allow the strength of perceptual bias to depend on reported confidence (i.e., whether the subject reported high confidence or not), and spatial terms to account for the subjects overall tendency to give perceptual reports that were biased slightly towards straight ahead (Fig. 3.2f). On average, the linear model captured the behavioral trends well (Fig. 3.6b), based on contributions of each of the terms described above that tended to vary in magnitude across subjects (Fig. 3.6e). By comparison, a parameter-free normative model captured some of the behavioral trends (Fig. 3.6a) but reported higher perceptual biases than subjects (compare red points and bar in Fig. 3.6e), particularly on change-points (compare green points and bars in Fig. 3.6e).

### **3.3.3 Modulations of perceptual biases reflected in pupil diameter**

A key question addressed in this work is whether arousal systems, as reflected in pupil diameter, contribute to the dynamic modulation of perceptual biases. Using linear regression at each time-point relative to sound onset (the average sound-evoked pupil



**Figure 3.6: Dynamic modulation of perceptual bias by normative and non-normative factors.** (a) Comparison of a parameter-free normative model (ribbons indicate mean $\pm$ SEM simulated perceptual bias for the same task sequences experienced by the subjects) and the subjects' behavior (points and errorbars are mean $\pm$ SEM from 29 subjects), shown as a function of sounds after a change-point (SAC) for the two noise conditions (colors, as indicated).

**Figure 3.6: Dynamic modulation of perceptual bias by normative and non-normative factors, continued** (b) Comparison of the linear model shown in panel e to behavior. Conventions as in panel a. (c,d) Dependence of the normative factors used in both models on task conditions: (c) prior relevance, which measures the probability of the current sound coming from the same distribution as the previous sound; and (d) prior reliability, which measures the anticipated precision of the predictive distribution relative to the likelihood distribution prior to stimulus presentation. (e) Best-fitting parameter estimates from the linear model fit to behavioral data from each subject (points) and to simulations of the parameter-free normative model (thick and thin bars indicate 95% confidence intervals over simulated subjective values and over simulated mean values across subjects, respectively). PE=prediction error. Asterisks indicate coefficients with mean values that differed from zero (t-test,  $p < 0.05$ )

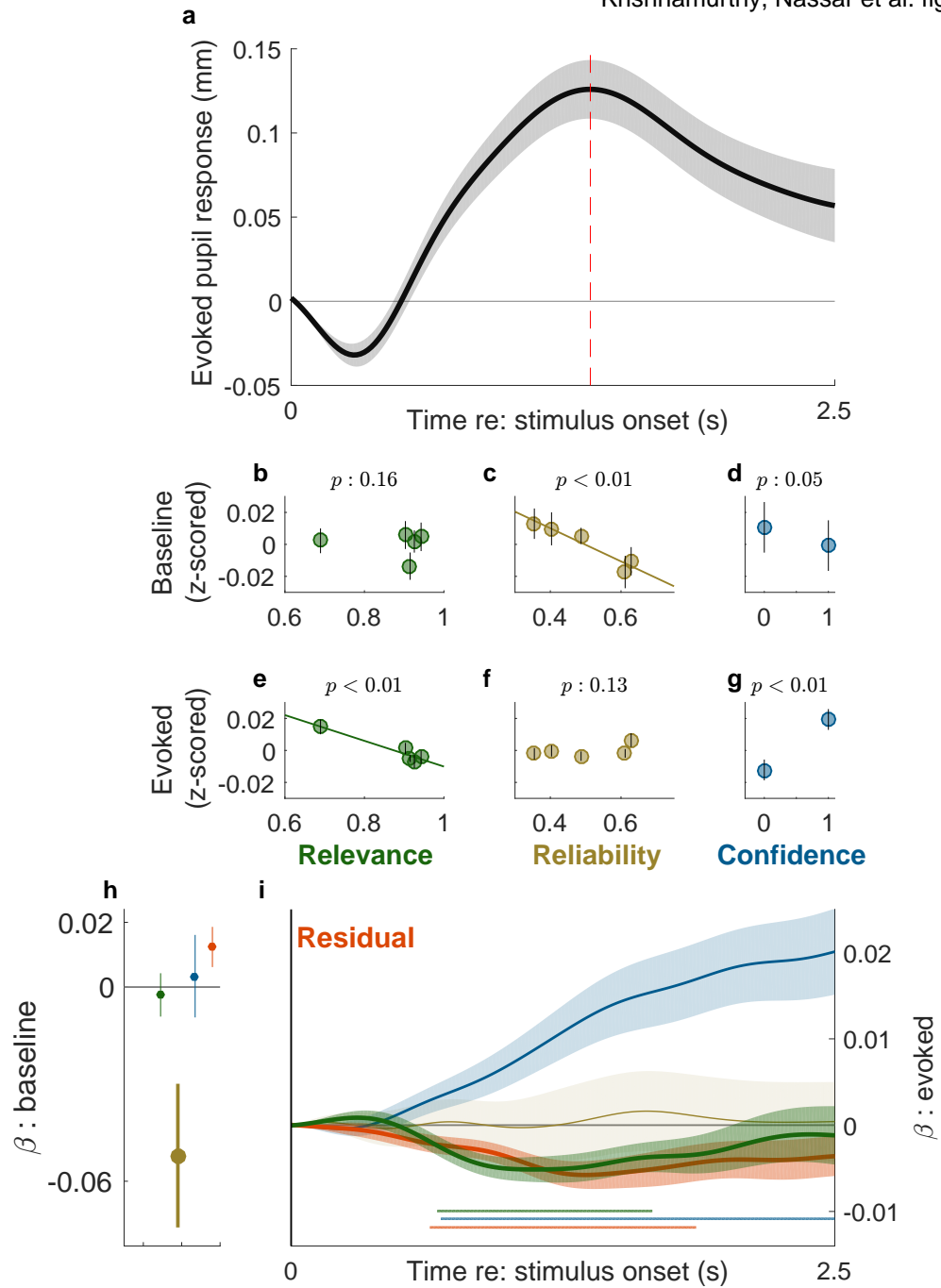
---

response from all probe trials and subjects is shown in Fig. 3.7a), we found that pupil diameter varied with several of the factors from the linear model that accounted for behavioral biases (Eq. 3.6; Fig. 3.7b). Specifically, prior reliability was reflected in the baseline diameter before presentation of the probe sound, with smaller baselines reflecting more reliable priors ( $p = 0.03$ ; Fig. 3.7c,h). However, prior reliability did not modulate the magnitude of the stimulus-evoked pupil response, after accounting for the baseline effect (Fig. 3.7f,i). In contrast, prior relevance was unrelated to baseline diameter but was robustly encoded by the stimulus-evoked pupil diameter, with larger evoked pupil responses reflecting lower prior relevance (Fig. 3.7b,e). This effect peaked around the time of the maximum sound-evoked pupil response (permutation test for effect duration: duration=1.0 s,  $p = 0.02$ ; Fig. 3.7i). The pupil response, but not the baseline, also reflected the subjects' upcoming confidence report, with high confidence corresponding to larger pupil diameters, particularly late in the fixation interval (duration=1.8 s,  $p = 0.01$ ; Fig. 3.7d,g,i; note that these duration estimates were limited by the size of our measurement window).

If the arousal system is contributing to the dynamic regulation of the influence of priors on perception, then pupil diameter may co-vary with adjustments in prior influence even after accounting for all of the factors in the behavioral linear model

(for example, if variability in internal representations of sound-source location affect both behavior and arousal). We therefore included the residual perceptual bias from our model of behavior (Fig. 3.6) in our model of pupil diameter. A positive/negative value of the residual biases indicates that the subject was more/less biased by the prior on the given trial than predicted by the linear model. There was a trend toward positive coefficients for this term in explaining baseline pupil diameter (larger baseline diameters corresponded to slightly stronger biases than predicted by the behavioral model;  $p = 0.06$ ; Fig. 3.7h). In addition, there was a robust reflection of the residual bias term in sound-evoked pupil response (smaller responses near the peak of the evoked response corresponded to stronger biases than predicted by the behavioural model; duration=1.2 s,  $p = 0.02$ ; Fig. 3.7i). This residual bias effect implies that pupil diameter reflects not just particular factors like prior reliability and relevance that can be used to make effective predictions in dynamic environments [71], but also the extent to which those and other factors are actually used to bias perception from one stimulus to the next.

In addition to these average, within-subject effects, there were also across-subject relationships between pupil diameter and perceptual biases. In particular, stimulus-evoked pupil responses tended to be, on average, smaller in subjects with higher overall perceptual biases (PE term in Fig. 3.6e; Fig. 3.8c) or relevance-dependent biases (PE\*relevance term in Fig. 3.6e; Fig. 3.8d). These effects were not evident for baseline pupil diameter (Fig. 3.8a,b). However, because the behavioral influences of overall perceptual biases and prior relevance covaried considerably across subjects ( $r = 0.77$ ,  $p < 10^{-5}$ ), we constructed a new linear model that included two individual-difference variables that corresponded to the shared and unique variance of the two behavioral coefficients. The effects of the shared term were negative for most of the



**Figure 3.7: Pupil diameter reflects dynamic modulations of perceptual bias within individual subjects.** (a) Mean  $\pm$ sem evoked pupil response from 29 subjects, defined as the pupil diameter relative to baseline during the measurement period. Red line indicates the time of the peak mean response (1.38 sec after stimulus presentation).

**Figure 3.7: Pupil diameter reflects dynamic modulations of perceptual bias within individual subjects, continued (b-d)** Baseline pupil diameter for trials sorted into bins according to relevance (**b**), reliability (**c**), and confidence (**d**). Relevance and reliability were binned in quintiles per subject, then each bin was combined across subjects. Confidence was divided into all trials with a low (0) or high (1) confidence report. Points and errorbars are mean $\pm$ SEM from all values in each bin. (**e-g**) Same as **bd**, but using the pupil diameter measured at the time of the peak response after accounting for the linear baseline dependencies. (**h,i**) Regression coefficients from a linear model accounting for modulation of baseline pupil diameter (**h**) or the evoked response (**i**) at each time-point using as predictors: 1) prior relevance, 2) prior reliability, 3) the upcoming confidence report, and 4) the residual perceptual bias from the linear model in Fig. 3.6d. Points and error bars in **h** and lines and ribbons in **i** represent mean $\pm$ sem of values computed per subject and thus represent within-subject modulations. Points and lines/ribbons corresponding to relevance, reliability, and confidence use the same colors as in (**b-g**). Bold symbols in **h** and horizontal lines in **i** indicate periods for which  $H_0$ : value=0,  $p < 0.05$ , after accounting for multiple comparisons.

---

measurement window (Fig. 3.8e; duration=2.2 s,  $p = 0.01$ ). In contrast, the unique-variance term did not show a strong relationship to average pupil traces. This result implies that subjects who had the strongest overall perceptual biases, and modulated them most according to prior relevance, tended also to have the smallest stimulus-evoked pupil responses.

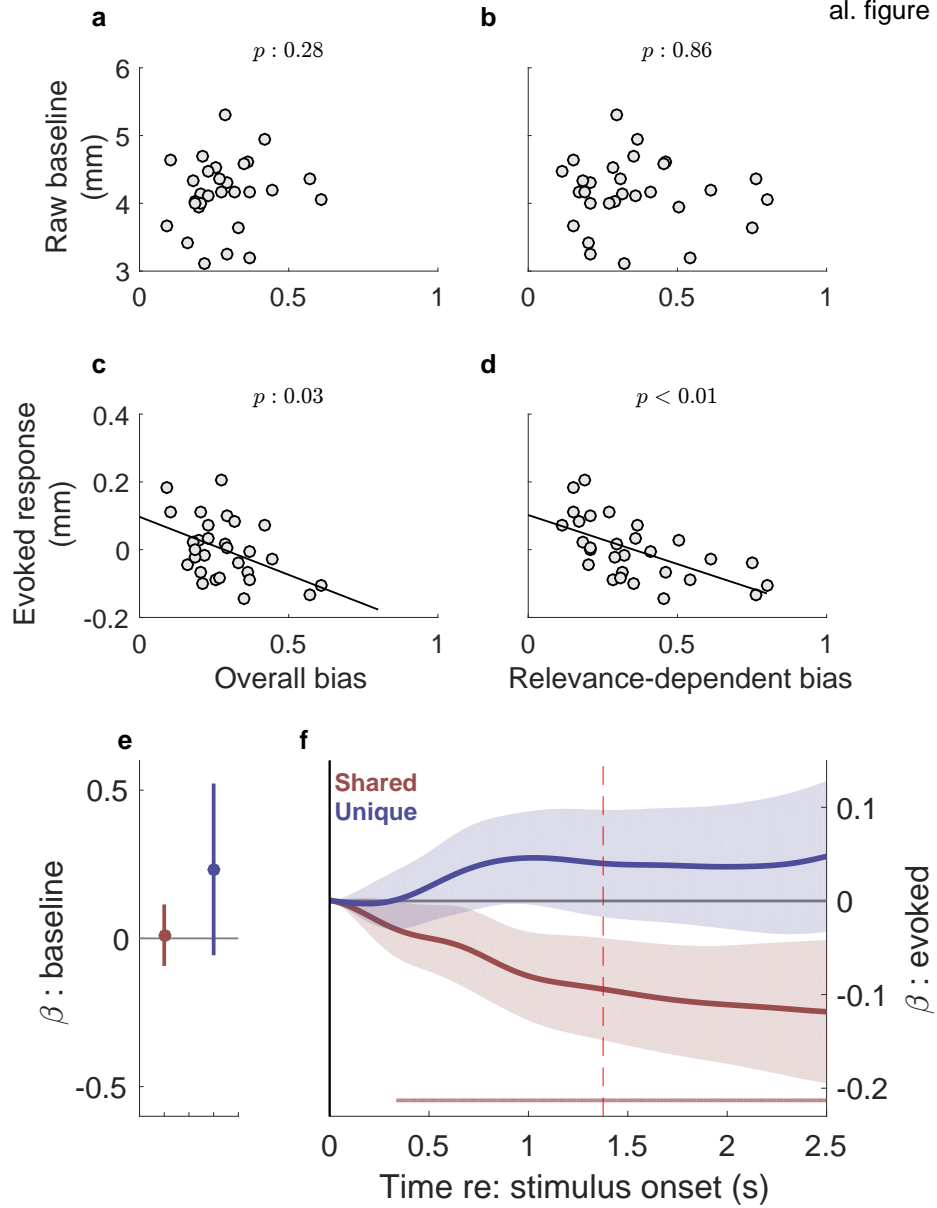
To further quantify these within- and across-subject relationships between pupil diameter and task performance, we used pupil diameter to predict the subjects perceptual biases. Specifically, we created three normalized variables to reflect within- and across-subject variability in pupil responses at the time of peak response (1.4 s following stimulus onset) along with their multiplicative interaction. Each pupil-derived variable was included as a modulator of prediction errors in three different linear models of perceptual errors. In the simplest model, pupil-derived measures alone predicted systematic differences in perceptual biases observed in the behavioral data (Fig. 3.11a), such that biases were: 1) larger for trials in which pupil responses were smaller than average (t-test,  $p < 10^{-4}$ ), 2) larger for subjects who had smaller than average pupil responses ( $p < 10^{-3}$ ), and 3) modulated from trial to trial more steeply for subjects with smaller overall pupil responses ( $p < 0.01$ ; Fig. 3.11b).

Consistent with these relationships, the pupil-based measures offered a substantial improvement to the base model in terms of predicting behavior (likelihood-ratio test,  $p < 10^{-7}$ ; Fig. 3.11c). The pupil-based measures also offered an explanatory advantage when added to more complex models that accounted for direct fixed effects (one coefficient for all subjects) or random effects (one coefficient per subject) of relevance, reliability, and confidence reports on perceptual biases ( $p < 10^{-4}$  for both models; Fig. 3.11c). Taken together, these results imply that fluctuations in pupil diameter, particularly those mediated by stimuli and related to context relevance, can be used to determine the extent to which perception is biased towards pre-existing priors.

### 3.4 Discussion

We used an auditory-localization task to show that the influence of prior expectations on perception is regulated rapidly and adaptively in changing environments. This work combines and extends several lines of research. The first has emphasized the role of priors on the perception of an uncertain sensory stimulus [72]. Many of these studies have focused on priors that are related to relatively stable properties of the environment, although several recent studies have shown that certain sensory or sensory-motor priors can be learned relatively rapidly [8, 80, 81, 82, 83]. The second has shown that under a variety of conditions, individual variability in decision-making can involve differential use of priors [84]. The third has identified how predictions are updated and used to make decisions in dynamic environments [70, 85, 86]. The fourth has related this dynamic updating process to changes in physiological arousal [71, 87]. We showed that many of these principles, including dynamic, arousal-related adjustments in predictions, apply to how priors are updated and used to guide perception.





**Figure 3.8: Pupil diameter reflects individual differences in perceptual biases.** (a,b) Mean baseline diameter for each subject (points) as a function of the perceptual bias (a; fits to the PE term in Fig. 3.6e) and relevance-dependent bias (b; fits to the PE\*relevance term in Fig. 3.6e).

**Figure 3.8: Pupil diameter reflects individual differences in perceptual biases, continued.** (c,d) Mean evoked pupil response for each subject as a function of the perceptual bias (a) and relevance-dependent bias (b). Pupil responses were measured at the time of peak response (1.38 sec after stimulus presentation) and orthogonalized to subject baseline pupil measurements. (e,f) Regression coefficients describing the relationship between shared or unique variance (colors, as indicated) in PE and PE\*relevance coefficients from the behavioral model and average baseline (e) or stimulus evoked (f) pupil diameter. Points and error bars in d and lines and ribbons in e represent the correlation coefficient and 95% confidence intervals of the estimate and thus represent across-subject modulations. Horizontal lines in e indicate periods for which  $H_0$ : value=0,  $p < 0.05$  after accounting for multiple comparisons.

---

These principles involve ongoing assessments of the relevance and reliability of priors that represent a form of statistical learning [88, 89]. We quantified this learning process using two variables derived from normative theory [77, 90, 91, 92, 93]. The first, which we termed prior relevance, is closely related to unexpected uncertainty and reflects the probability that a new observation is consistent with recent history [?, 77]. The second, which we termed prior reliability, is a form of reducible uncertainty that reflects ambiguity, typically resulting from undersampling, about the current generative process [92, 93]. Previous work showed that new information exerts the least influence on existing predictions when those predictions are the most relevant and reliable [70, 85]. We showed analogous effects for perception: new sensory input exerts the least influence on perception, relative to the influence of priors (i.e., perceptual biases are largest), when the priors are the most relevant and reliable.

Both of these normative factors, scaled according to their effects on each subjects behavior, were reflected in modulations of arousal state as measured by pupil diameter. Prior reliability corresponded to changes in baseline pupil diameter, and prior relevance corresponded to changes in the stimulus-evoked change in pupil diameter. These modulations were similar to those that we reported previously for a predictive-inference task, but the different demands of our present task imply a broader relevance to different forms of information processing [71]. Specifically, our previous findings

implicated a role for arousal fluctuations in adjusting bottom-up effects of new sensory input on stored cognitive representations. In contrast, our present findings implicate a role for arousal fluctuations in adjusting top-down control exerted by stored representations on the interpretation of new sensory input.

This result has broad implications for decision-making. For simple sensory-motor tasks, sequential effects of choice and response times can reflect priors inferred from recent task patterns, even when the patterns are spurious and thus the effects are detrimental to overall performance [94, 95, 96]. Our results suggest a role for stimulus-evoked arousal responses in minimizing such suboptimal biases, potentially by reducing the impact of the top-down signals that mediate them. Consistent with this idea, pupil dilations have been shown to be accompanied by reduced individual and sequential-choice biases on perceptual decision-making tasks [97, 98]. For more complex tasks, top-down prior information might be used to select task-relevant feature information and thereby reduce implicit processing biases [99, 100]. This effect might explain why individuals with larger evoked pupil responses tend to be more susceptible to their own implicit processing biases [101, 102]. Future work should address this possibility in paradigms that combine implicit sensory biases with stimulus history-dependent priors such as those used in our task.

These results are also consistent with the idea that transient increases in arousal, in response to surprising events or other factors, may generally correspond to higher sensitivity to immediate sensory input [103, 104]. In principle, this increased sensitivity could emerge from an enhancement of feed-forward processing, perhaps through an increase in neural gain [71, 78, 101, 105]. An alternative, but not necessarily mutually exclusive, possibility supported by our results is that enhanced sensitivity to sensory input is afforded by a reduction in the effectiveness of top-down priors in reg-

ularizing, and thereby biasing, sensory percepts. Distinguishing and understanding the independent contributions of these alternatives to arousal-mediated information processing will require the development of new paradigms that can separately control the bottom-up and top-down flow of information.

We also found relationships between subjective confidence, perceptual biases, and pupil diameter. We measured confidence using a post-decision binary confidence report (high/low confidence), which previously has been linked to the sensory-driven decision variable that also governs the speed and accuracy of the perceptual decision [106, 107, 108]. We showed that confidence is also modulated according to changes in the relevance and reliability of perceptual priors that affect perceptual errors. This modulation was also evident in pupil diameter, which reflected high confidence-report frequency even after accounting for the normative variables that also governed the perceptual biases. However, this extra effect was in the opposite direction as for the normative factors, relative to the behavioral effect: high confidence-report frequency corresponded to larger pupil diameters but stronger prior influence. This pupil effect is somewhat surprising given that pupil diameter can be enhanced under uncertain, rather than certain, conditions [71, 98, 109, 110, 111, 112] (but see [87]). One possible explanation for this discrepancy is that the post-decision confidence report led subjects to anticipate the increased reward or risk associated with high confidence-report trials, leading to stronger arousal responses [111, 113]. This idea is supported by the time course of confidence-related pupil dilations, which had a maximal dilation immediately prior to the perceptual report. This idea also highlights the multiple, possibly interacting roles that the arousal system likely plays in even simple sensory-motor tasks like this one.

These multiple roles undoubtedly result from multiple mechanisms by which arousal

affects neural information processing [114]. One such mechanism likely involves cortical levels of norepinephrine (NE), which is controlled via neurons in the midbrain nucleus locus coeruleus (LC) [78]. Firing rates of LC neurons correlate with pupil diameter over relatively short timescales, which has prompted the suggestion that pupil diameter can be used as a proxy for LC activity [78, 79, 115, 116]. Thus, the factors in our task that corresponded to stimulus-evoked pupil dilations, such as more surprising stimuli with lower prior relevance, may also correspond to increased LC activation. This activation, in turn, would increase levels of cortical NE, which have been theorized to signal unexpected context changes and allow neural representations to reorient rapidly to meet changing contextual demands, possibly via modulations of the input/output gain of individual cortical neurons [74, 77, 78, 105, 117]. In contrast, slower changes in pupil diameter, such as those related to our baseline modulations, may be more closely related to levels of acetylcholine released from the basal forebrain, which has been theorized to signal expected uncertainty of task-relevant beliefs [77, 118, 119]. More work is needed to determine how these multiple, potentially interacting neuromodulatory systems help to regulate perception and decision-making in dynamic environments.

### **3.5 Experimental Procedures**

Human subject protocols were approved by the University of Pennsylvania Internal Review Board. 29 subjects (16 female, 13 male) participated in the study after providing informed consent. Thirty-one additional subjects consented to the study but did not meet the inclusion criterion of participating in at least three experimental sessions. Our sample size was well powered to detect effects of  $d' \geq 0.6$  (statistical power  $\geq 0.88$  for  $d' = 0.6$ ) providing sufficient sensitivity in the range of previously

reported behaviour-pupil relationships.

### 3.5.1 Auditory-localization task

We used an auditory-localization task in which subjects heard sounds with varying source locations that were simulated using head-related transfer functions (HRTFs) from the IRCAM database (<http://recherche.ircam.fr/equipes/salles/listen/download.html>). Each sound was a sequence of five Gaussian noise pulses bandpass filtered between 100 Hz and 15 KHz. The pulses were 50 ms each with a delay of 10 ms following each pulse, for a total stimulus duration of 300 ms. The latency for the sound to reach the ears following the command execution was 3 ms. For each subject, we tested a number of HRTFs during the initial session by playing sound sequences that circularly traversed the entire horizontal plane in 15° intervals. We picked the HRTF that was reported to give the most uniformly circular percept for the sound sequence. Each subject performed 3-6 total sessions.

Each subject completed two tasks per session. The first was a control localization task that required the subject to indicate the perceived location of simulated sound sources that were sampled independently and uniformly randomly along the frontal, horizontal plane. In each of 72 trials, the subject was required to: 1) fixate for 2.5 s on a central spot while listening to the auditory stimulus; and 2) indicate the perceived location of the sound using a mouse, which controlled a cursor that moved along a semi-circular arc on the computer screen that represented the range of possible sound-source locations (Fig. 3.1). Failure to maintain fixation resulted in a warning sound and trial break. Feedback was displayed on the screen after the subject reported the perceived location.

The second task was a dynamic localization task that required the subject to report predictions, perceptions, and confidence judgments of sound-source locations that were generated from a change-point process along the same horizontal plane. For this task, the subject listened to extended sequences of sounds generated by the change-point process, with an interval of 150 ms between each sound presentation. Each sound was paired with a visual cue indicating its simulated source location on the semi-circular arc. During the presentation of these sequences, no action was required. Occasionally, however, the sequences stopped, indicating the start of a probe trial with the following structure (Fig. 3.1c). First, the subject was required to predict the angular location of the next, upcoming probe stimulus on the arc using a mouse. Second, following the prediction, the subject was required to maintain fixation for 2.5 s on the same central spot used in the control task. The auditory probe stimulus, with no corresponding visual cue, was presented at the beginning of this fixation period. Fourth, after the fixation period ended, the subject indicated the perceived location of the probe stimulus using the mouse and the visual display. Fifth, the subject then reported confidence (high/low) on the accuracy of the perceptual report (Fig 3.1). Each subject performed four blocks of the dynamic task per session, which included 30 probe trials each. Each session lasted 90 min, with some variability due to the self-paced nature of the prediction, perceptual report, and confidence reporting periods of the task (median [IQR] reaction times were: 1.72 [1.492.35] sec for the prediction, 2.02 [1.502.30] sec for the perceptual report, and 1.18 [0.941.43] sec for the confidence report).

The sequence of simulated sound-source locations for the dynamic task was determined according to a process that included both irreducible variability (noise) and abrupt discontinuities (change-points). Our goal was to test if and how these manipu-

lations, which can affect the reliability and relevance, respectively, of new information on existing predictions, also affect perceptual reports that can, in principle, use such predictions to improve the perception of ambiguous stimuli 10. Source locations were sampled from a Gaussian distribution with a standard deviation (STD) that was held constant within a block of 600 trials ( $10^\circ$  or  $20^\circ$  for the low- or high-noise condition, respectively) and a mean that underwent abrupt change-points with a fixed probability, or hazard rate (H), of 0.15 for each sound sample. At each change-point, the mean of the generative distribution was resampled uniformly across the sound space, such that the newly generated source locations were independent from previous ones. The sequence was interrupted for probe trials at random using a procedure that ensured: 1) a roughly even distribution of probes occurring 1-6 sounds after a change-point (SAC); 2) that probes were separated by at least 8 sounds; and 3) the number of sounds between any two probe trials was the same, on average, regardless of the nature of the two probe trials (SAC 1-6). Thus, on some trials the probe sound-source location was independent of the previous stimulus sequence (SAC=1). On other trials, the probe location was generated from the same distribution that produced the immediately preceding locations (SAC=2-6).

Subjects were motivated to make accurate perceptual reports on each probe trial through an incentive system. They were instructed to report high confidence if they were confident that the true location was within a  $16^\circ$  window centered on their second (perceptual) report, and to report low confidence otherwise. A correct/incorrect high confidence report resulted in a score of (15/-10), and a correct/incorrect low confidence resulted in a score of (5/-3). Subjects were paid a bonus that depended on their total score.



### 3.5.2 Behavioral data analysis: contrasts

To provide an intuitive understanding of how behavior was affected by particular task conditions, we sorted probe trials into twelve conditions according to the recency of the previous change-point (SAC=1-6) and noise condition (high/low). To emphasize the effects of change-points on the behavioral reports, these analyses included data only from sequences following easily noticeable change-points, corresponding to changes in generative mean of at least twice the generative STD for SAC=1 (note that the linear model described below was used to analyze the full data sets, including all change-points). Perceptual and prediction errors were computed by subtracting reported percepts and predictions from the true (simulated) sound source location for each trial. For each condition, the STD of prediction and estimation errors was used as a metric of average absolute error magnitude.

To quantify how prediction errors, estimation errors, and average confidence reports depended on specific task conditions, we performed four orthogonal linear contrasts over our twelve task conditions. Each contrast was computed by multiplying a weight matrix by the measured prediction errors, estimation errors, or average confidence reports, aggregated according to the task conditions for a single subject. Weight matrices were mean centered and chosen to identify: 1) differences between change-point and non-change-point trials (CP); 2) linear increases with increases in the number of sounds experienced (Exp) following a change-point, from SAC=2 to SAC=6, in the high-noise condition ( $\text{Exp}_{high}$ ); 3) comparable linear increases in the low-noise condition ( $\text{Exp}_{low}$ ); and 4) differences between the high- and low-noise conditions (Noise). Thus, the contrasts provided a per-subject measure of how much each behavioral measurement was modified according to these factors. For Figs. 3.3-3.5, we considered only sound sequences following relatively large change-points, correspond-

ing to at least twice the generative STD. Contrasts were normalized for presentation in Figs. 3.3 and 3.4 by dividing the contrast value for each subject by the standard deviation of that contrast across all subjects. This procedure allowed all contrasts to be meaningfully displayed on the same set of axes.

### 3.5.3 Behavioral data analysis: perceptual bias

To quantify the influence of the prior on the perceptual report, we measured the slope of the best-fit line to the relationship between prediction errors (prediction-true location) and perceptual errors (percept-true location) for the given task condition. Slopes close to one indicate a high perceptual bias, and slopes close to zero indicate low perceptual bias. To measure how the perceptual bias evolved as a function of the number of sounds after a change-point (SAC), we used the following linear model and included only data from sequences following noticeable change-points (jumps of at least twice the generative STD):

$$\begin{aligned}
 ERR_{percp} = & \beta_0 + \beta_1 ERR_{pred,1}^{high} + \dots + \beta_6 ERR_{pred,6}^{high} \\
 & + \beta_7 ERR_{pred,1}^{low} + \dots + \beta_{12} ERR_{pred,6}^{low} + \beta_{13} Bias_{spatial}
 \end{aligned} \tag{3.1}$$

where  $ERR_{percp}$  is the perceptual error and  $ERR_{pred,1}^{high}$  is the prediction error on change-point trials (SAC=1) in the high-noise condition, and so on. The last term,  $Bias_{spatial}$ , captures the slight bias in the perceptual reports towards center of the screen.

### 3.5.4 Behavioral data analysis: theoretical benchmarks

The theoretically expected overall perceptual-error STD per subject (abscissa in Fig. 3.2i) was computed from an optimal, reliability-weighted combination of prior and sensory information:  $\sigma_{theoretical}^{-2} = \sigma_{predictions}^{-2} + \sigma_{sensory}^{-2}$ . The theoretically expected perceptual-error STD per subject (diamonds in Fig. 3.3c,d), given their corresponding predictions for each SAC condition, was computed using  $(\sigma_{theoretical}^{SAC})^{-2} = (\sigma_{predictions}^{SAC})^{-2} + \sigma_{sensory}^{-2}$ . The theoretically expected frequency of high-confidence reports (diamonds in Fig. 3.3e,f) was computed as the probability mass contained in a  $16^{circ}$  window centered at the mean of a Gaussian with a STD of the theoretically expected perceptual errors,  $\sigma_{theoretical}^{SAC}$ . Thus, the proportion of expected high-confidence reports increased with narrower perceptual error distributions. The theoretically expected perceptual bias per subject (diamonds in Fig. 3.4d,e) was computed as  $\sigma_{sensory}^2 / (\sigma_{sensory}^2 + (\sigma_{predictions}^{SAC})^2)$ . In all of the above,  $\sigma_{predictions}$  is the STD of prediction errors on non-change-point trials,  $\sigma_{predictions}^{SAC}$  is the STD of prediction errors for the specified number of sounds after a change-point, and  $\sigma_{sensory}$  is the STD of perceptual errors on the control task, computed per subject.

### 3.5.5 Behavioral data analysis: normative model

Auditory localization in a dynamic environment can be posed as a perceptual inference problem with the goal of inferring the location of the sound source on trial  $t$  ( $X_t$ ) according to a noisy internal sensory representation of that sound source ( $\lambda_t$ ) and the history of previously experienced sound sources ( $X_{1:t-1}$ ). This problem can be simplified by exploiting the conditional independencies in the Markov change-point process through which sound sources are selected (see Fig. ??). In particular, the

sound sources locations on the current trial ( $X_t$ ) are independent of those on previous trials ( $X_{1:t-1}$ ) conditioned on the mean of the generative distribution on the current trial ( $\mu_t$ ). In turn, the mean of the generative distribution on the current trial ( $\mu_t$ ) is also independent of previous observations conditioned on: 1) the mean of the generative distribution on the previous trial ( $\mu_{t-1}$ ), and 2) a latent change-point variable that determines whether the mean is resampled on the current trial ( $S_t$ ). Applying Bayes rule to invert the generative graph in Fig. ?? yields the following inference equation:

$$P(X_t | \lambda_t, X_{1:t-1}) = \frac{\sum_{\mu_t} P(X_t | \mu_t) \sum_{s_t, \mu_{t-1}} P(\mu_t | \mu_{t-1}, s_t) P(s_t) P(\mu_{t-1} | X_{1:t-1})}{\sum_{X_t} P(X_t, \lambda_t | X_{1:t-1})} \times P(\lambda_t | X_t) \quad (3.2)$$

where the likelihood  $P(\lambda_t | X_t)$  reflects the conditional distribution of internal representations across true sound source locations;  $P(X_t | \mu_t)$  reflects the conditional probability of a sound source location being generated from a particular generative mean;  $P(\mu_t | \mu_{t-1}, s_t)$  reflects the process through which means are resampled on change-point ( $s_t=1$ ) trials; and  $P(s_t)$  is the hazard rate (H), which was fixed to 0.15 for the task and all simulations. The likelihood  $P(\lambda_t | X_t)$  was modeled as a normal distribution centered on with a variance that was fixed for each subject to the variance of perceptual reports made by that subject on the control task  $\sigma_{sensory}^2$ .  $P(X_t, \lambda_t | X_{1:t-1})$  is the distribution over possible generative means, which can be updated recursively. Although exact Bayesian solutions to this recursive problem exist [73, 90], we use a normal approximation to the Bayesian mixture distribution with a mean  $\hat{\mu}$  and variance  $\hat{\sigma}^2$  that capture the key dynamics of normative inference and offers better

descriptions of human behaviour [70]. As in previous work, predictions made using this approximation were more accurate than subject predictions. To account for this discrepancy, we created a subjective prediction  $\hat{\mu}_{subj}$  by sampling a random normal variable with mean equal to  $\hat{\mu}$  and a variance that was incremented on each trial according to the difference in variance of subject and normative prediction errors:

$$\hat{\sigma}_{subj}^2 = \hat{\sigma}^2 + Var(X - \text{subject predictions}) - Var(X - \hat{\mu}) \quad (3.3)$$

Perceptions and predictions from the normative model were simulated by sampling internal representations ( $\lambda_t$ ) and subjective predictions ( $\hat{\mu}_{subj}$ ) for each trial according to the actual sequence of sound source locations experienced. Descriptive statistics for model simulations were averaged across 100 such simulated runs.

In addition to simulating behavioral data, the normative model also allowed us to extract latent variables that guide normative adjustments in perceptual bias. In particular, the model adjusts bias towards prior expectations in accordance with the relevance and reliability of those expectations. The relevance of prior expectations ( $\pi_t$ ) is, in our generative framework, equal to the probability that a change-point did not occur on this trial given all previous data. This probability was computed on each trial by marginalizing Eq. 3.2 over all dimensions other than  $s$ . The impact of normative priors also depends critically on their reliability relative to that of the likelihood distribution capturing the noisy internal sensory representation ( $\lambda_t$ ):

$$\text{prior reliability: } \tau_t = \frac{\sigma_{sensory}^2}{\sigma_{sensory}^2 + \hat{\sigma}_{subj}^2 + \sigma_{noise}^2} \quad (3.4)$$

where  $\tau_t$  is prior reliability,  $\sigma_{sensory}^2$  is the variance of perceptual reports made by that subject on the control task,  $\hat{\sigma}_{subj}^2$  is the variance on subjective assessments of the underlying mean, and  $\sigma_{noise}^2$  is the expected variance of sound source locations about that mean. The sum of the latter two terms reflects the total variance on the models predictive distribution over possible sound locations. To ensure that these latent variables best reflected circumstances experienced by the subject, we fixed the model predictions ( $\hat{\mu}_{subj}$ ) to the actual subject predictions from each trial and computed each measure as the expected value across all possible values of  $\lambda_t$  using a grid-point approximation.

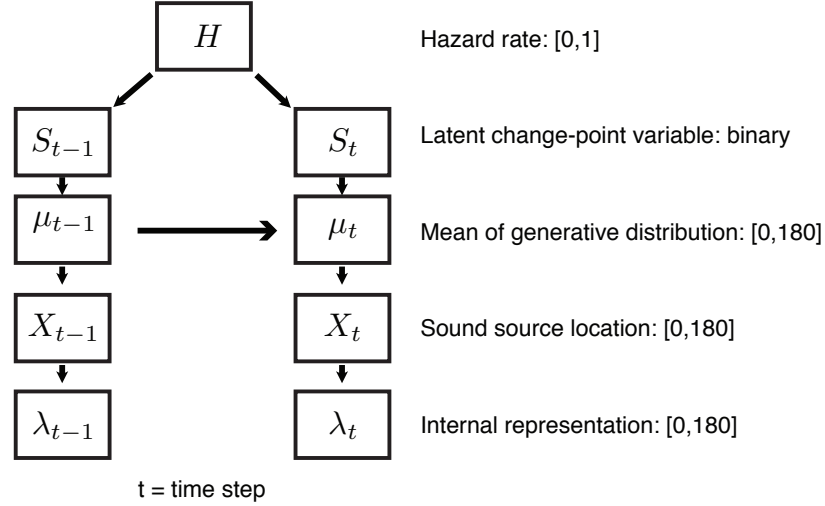
### 3.5.6 Behavioral data analysis: Linear model of perceptual bias

To provide a more complete description of how behavioral data from all conditions, including all generative change-point and non-change-point trials, depended on both normative and non-normative factors, we fit the following linear model to data from all trials in each session:

$$ERR_{percp}(t) = \beta_0 + \beta_1 ERR_{pred}(t) + \beta_2 ERR_{pred}(t) \cdot \pi_t + \beta_3 ERR_{pred}(t) \cdot \tau_t + \beta_4 ERR_{pred}(t) \cdot bet + \beta_5 ERR_{pred}(t) Bias_{center} + \beta_6 Bias_{spatial} \quad (3.5)$$

where  $\beta_1$  describes the overall prior bias;  $\beta_2$  and  $\beta_3$  describe the extent to which the overall bias is dynamically modulated by the prior relevance and reliability, respectively (see above);  $\beta_4$  describes the interaction of prior bias with confidence report (a

### A Generative graphical model for Bayesian perceptual inference



### B Bayesian inference equation (Eq. 2 in manuscript). Note that the noise condition is encoded in the outcome generative distribution.

$$p(X_t | \lambda_t, X_{1:t-1}) = \frac{\underbrace{p(\lambda_t | X_t)}_{\text{Likelihood}} \sum_{\mu_t} \underbrace{p(X_t | \mu_t)}_{\text{Outcome generative distribution}} \sum_{s_t} \sum_{\mu_{t-1}} p(\mu_t | \mu_{t-1}, s_t) \underbrace{p(s_t)}_{\text{Change-point prior}} \underbrace{p(\mu_{t-1} | X_{1:t-1})}_{\text{Posterior distribution over the mean}}}{\sum_{X_t} \underbrace{p(\lambda_t, X_t | X_{1:t-1})}_{\text{Normalization term}}}$$

**Figure 3.9: Bayesian model of perceptual inference.** (a) Graphical generative model describing dynamic task structure. For each sound in the stimulus train, a binary latent change-point variable ( $S_t$ ) was sampled according to a hazard rate ( $H$ ) that was fixed across all trials. If a change-point was not sampled ( $S_t = 0$ ), then the mean of the sound-source distribution ( $\mu_t$ ) remained stable ( $\mu_t = \mu_{t-1}$ ). Otherwise, in the case of a change point ( $S_t = 1$ ),  $\mu_t$  was drawn at random from a uniform distribution ranging from 0-180 degrees. The sound source location ( $X_t$ ) was sampled from a normal distribution with mean equal to  $\mu_t$  and a standard deviation equal to either 10 (low noise) or 20 (high noise), manipulated in task blocks. The simulated sound source gives rise to an internal subjective representation of its location  $\lambda_t$  according to a normal distribution, centered on  $X_t$ , with a standard deviation inferred from estimation errors on the control localization task. (b) Inference over this generative graph can be accomplished by inverting the generative process according to Bayes' Rule.

binary variable);  $\beta_5$  describes the bias towards the center of the screen; and  $\beta_6$  captures the angular spatial bias (mean perceptual error at the given angle) measured in the control task. Residuals from the model fit were signed according to the prediction error on each trial to create a residual bias term for use in pupil analysis.

### 3.5.7 Pupil measurements

Pupil diameter was sampled from both eyes at 60 Hz using an infrared eye-tracker built into the monitor (Tobii T60-XL). Pupil analyses used the mean value from the two eyes at each time point measured during fixation. We excluded from further analyses trials with blinks, which we identified using a custom blink-detection routine on the basis of missing pupil diameter measurements and/or vertical and horizontal eye position that deviated from fixation for at least 10 contiguous samples (median [IQR] percentage of excluded trials across sessions = 5.54 [3.169.16]%). For the remaining trials with  $\leq 10$  missing contiguous pupil samples, we linearly interpolate the data before low-pass filtering. Low-pass filtering was done using a Butterworth filter with a cut-off frequency of 4 Hz. These filtered measurements were then z-scored in each session. We also removed a linear trend in the average pupil diameter over the whole session to account for any slow drift. The pupil baseline for each probe trial was defined as the mean of the first three samples immediately preceding the measurement period for that probe trial.



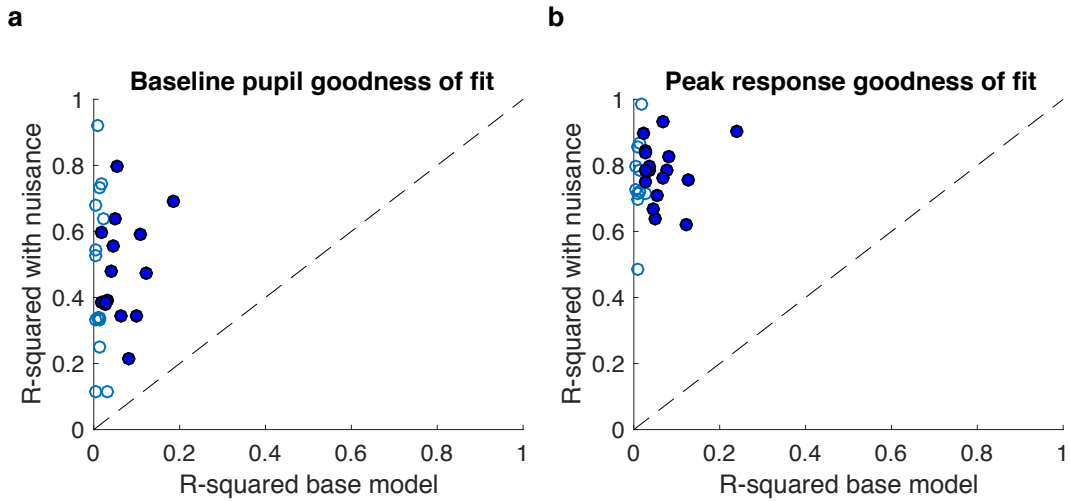
### 3.5.8 Linear model relating pupil diameter to behavioral parameters

To measure how the variables driving behavior were encoded in pupil diameter, we used the following linear model to explain the fluctuations in: 1) the baseline pupil diameter, and 2) stimulus-evoked pupil response across the 2.5 s following the auditory stimulus:

$$\begin{aligned} \text{Pupil diameter} = & \beta_1\pi_t + \beta_2\tau_t + \beta_2 \cdot \text{Bet}_n + \beta_2 \cdot \text{Bias}_{\text{residual}} + \\ & \beta_5 \cdot (\text{Previous baseline diameter}) + \beta_6 \cdot (\text{Time since previous probe}) \\ & + \beta_{7-9} \cdot (\text{low frequency terms}) \end{aligned} \quad (3.6)$$

where  $\tau_t$  and  $\pi_t$  are the reliability and relevance, respectively;  $\beta_{1-4}$  capture relationships between pupil diameter and the computational and behavioral variables of interest;  $\beta_{5-6}$  capture persisting fluctuations in pupil diameter that are attributable to the previous trial; and  $\beta_{7-9}$  includes a set of three low-frequency cosine components for each session that capture task-irrelevant variability due to slow modulations or session-based differences in pupil diameter. The exact form of the cosine components was  $\cos(\pi k(2n - 1)/2N)$ , where  $k = 0, 1, 2$ ;  $n$  is the trial number within the session; and  $N$  is the total number of trials in the session. When this model was fit to evoked pupil responses, an additional nuisance variable was added to the explanatory matrix that accounted for trial-by-trial differences in baseline diameter.

Significance testing for evoked pupil coefficients was done through cluster-based permutation testing to account for multiple comparisons over time. In short,  $t$ -tests were



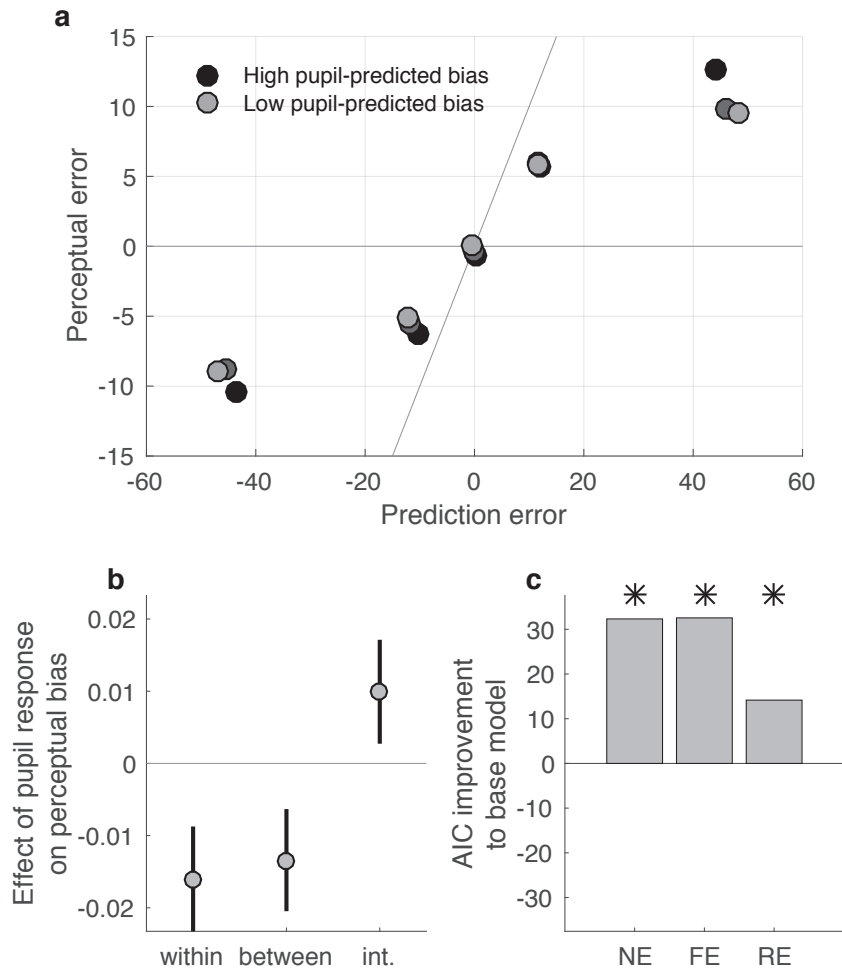
**Figure 3.10: Pupil regression goodness of fit.** (a,b)  $R^2$  values reflecting the goodness of fit for a base model that included only task-based regressors (abscissa) and a full model that included several nuisance parameters (ordinate) applied to pupil diameter at baseline (a) and peak response (b). Filled points indicate subjects for whom the base model provided a significantly better fit than a null model (F-test,  $p < 0.05$ ). The full model was preferred over the null model for all subjects. Mean AIC values (baseline measurements): 1308.9 for the null model, 1298.9 for the base model, and 958.7 for the full model. Mean AIC values (peak responses): 1308.9 for the null model, 1288.2 for the base model, and 538.0 for full model.

performed on each set of coefficient values across subjects separately for each time point. Cluster size was determined according to the number of contiguous time points for which this  $t$ -test yielded  $p < 0.05$ . Cluster corrected  $p$ -values were determined by comparing cluster sizes attained in this way to those from a permutation distribution of maximal cluster sizes [120].

### 3.5.9 Pupil-predicted perceptual bias

Trial-by-trial pupil measurements were extracted for the time of peak pupil response (1.4 seconds) from the behavioral model. Trial-by-trial measurements from each subject were regressed onto a set of nuisance variables that included explanatory variables  $\beta_{5+}$  from Eq. 3.6 to remove variance attributable to potentially confounding factors.

Residual pupil measurements were concatenated across subjects and then divided into two separate variables: one variable accounted for average measurements for each subject and one that reflected normalized deviations from the average measurement within each subject. An additional term was created through the multiplicative interaction of each subjects mean pupil response and pupil-response variability, to account for the possibility that individual differences in the overall arousal response modulate the extent to which trial-to-trial fluctuations in arousal modulate perceptual bias. The three resulting variable arrays were z-scored and multiplied by trial-by-trial prediction errors to create a predictor matrix. Trial-by-trial perceptual errors were regressed onto three separate models with and without the inclusion of the pupil predictor matrix: 1) a null model including an intercept term and a prediction error term to capture fixed effects of perceptual bias across all subjects as well as the spatial bias terms described above [NE]; 2) a fixed-effects model that also included interaction terms accounting for modulation of perceptual bias by prior relevance and reliability the subjects confidence report [FE]; and 3) a random-effects model that included all terms in model 2 separately for each subject [RE]. Since the random-effects model used dummy variables to account for individual differences in perceptual bias, the pupil predictor matrix included only within-subject variability and thus only one additional parameter rather than three. The marginal benefit of pupil-predictor terms was evaluated through likelihood-ratio tests (evaluating the additional explanatory power offered by pupil predictors) and quantified using AIC, a likelihood-based measure of goodness-of-fit that applies a penalty for each model parameter.



**Figure 3.11: Pupil diameter predicts perceptual bias.** (a) Perceptual error, sorted according to the pupil-predicted prior influence (gray scale, as indicated, corresponding to the bottom quartile, middle 50%, and top quartile) and plotted according to prediction error. Points are mean values computed across all subjects. (b) Mean  $\pm$  95% confidence intervals for pupil coefficients describing within- and between- subject effects of pupil diameter, as well as their interaction. See text for details. (c) Improvement in AIC achieved by adding pupil-based predictors to models that include: 1) a fixed perceptual bias across all subjects (NE), 2) a fixed perceptual bias and fixed model-based effects of perceptual bias across all subjects (FE), and 3) a random effects model that fits all bias and modulation terms separately for each subject (RE, which is equivalent to the normative linear model in Fig 3.6). Asterisks indicate significant improvements (likelihood-ratio test,  $p < 0.05$ ).

# Chapter 4

## Model complexity, information geometry and resolution of observations

### 4.1 Introduction: principled measures of model complexity

In behavioral tasks involving sequential inference, it is necessary to have some model of the environmental structure which allows one to make predictions about unseen samples or inferences about ambiguous ones. In several such tasks, human subjects behave in a way that is consistent with normative models [121, 122]. However, there is considerable variability between subjects in their overall behavior; for example, even for the same level of performance, some subjects can have more variable responses while others have more systematically biased responses. Our hypothesis is that this individual variability, instead of arising from random or uncontrolled fluctuations, is

a result of subjects using inferential models of varying complexity. In other words, subjects may use models of different complexities, but they behave near-optimally given their model complexity. If this were the case, then we would expect subjects to lie close to the boundary of an appropriately defined performance-vs-complexity trade-off.

A major obstacle in testing such a hypothesis is using a principled and meaningful definition of *model complexity*. Heuristic measures of model complexity can introduce biases and incorrectly label fundamentally similar models as being different [123]. On the other hand, using principled notions of complexity specifically designed for certain model classes may themselves give misleading results if the model class is not representative of the subject’s model. Here, we first describe principled complexity measures which have been classically used for specific model classes and then mention potential pitfalls. We then use an empirical measure of complexity – based on the notion of *predictive information* – which has rigorous theoretical backing[10], and describe how to use it to study the individual variability of subjects performing an inference task. We also illustrate the tight connections between predictive information, and other classical notions of model complexity based on the geometry of the parameter manifold[124, 125, 123]. Both these notions of complexity are consistent and exact asymptotically, and their form is familiar from penalty terms for ‘overfitting’.

Finally, we address interesting geometric issues that arise in the limit of finite data or by observing the model at a ‘coarse’ scale. This has natural connections to the work of James Sethna and his collaborators on *sloppy models* [126, 127, 128, 11]. The key insight from Sethna et al. is that in typical scenarios, we don’t probe all the degrees of freedom of the model, and thus, there are many parameter combinations that have little effect on the output behavior of the model at our scales of observation;

moving large distances in parameter space along those directions has little effect on the measurements. This immediately implies that one can collapse those parameter directions to give simpler ‘effective/emergent’ models for our scale of observation. The classical notions of model complexity do not explicitly account for this important behavior. How does one modify these measures of model complexity to explicitly take into account the resolution at which you observe the system? We suggest a natural and principled extension to the complexity measures to make connection with the literature on sloppy models.

## 4.2 Model selection and classical measures of complexity

Measures of model complexity have been a central aspect of selecting between different explanations of data, going back all the way to the philosophical literature (for e.g., Occam’s razor). In the more recent statistical era, there is a rich body of work on quantitative measures of model complexity of parametric model families (for instance by Akaike et al.); however, these measures were often incomplete or lacked unifying principles. Later on, Rissanen, Barron and others made many of these notions rigorous and showed that a large family of model selection schemes which penalize complexity can be viewed as a precisely formulated tradeoff between explaining the data (‘goodness of fit’) and model complexity, suitably defined. Rissanen’s prescription [129] was that we should select the model which offers the greatest combined compression of the data and the description of the model itself, as this is most likely to uncover the regularities in the data. This is referred to as the Minimum Description Length (MDL) principle. These family of schemes (including MDL) are collectively referred to as Minimum Complexity Density(MCD) estimation methods by Barron

et al.[130, 131]. Rissanen also showed that when the observed data  $\mathbf{x}$  are generated from within the model family  $P(\mathbf{x}|\boldsymbol{\alpha})$  indexed by a parameter  $\boldsymbol{\alpha}$ , in the limit of large sample size  $N$ , the length of the shortest code encoding the data and the model is

$$\text{MDL} \approx -\log P(\mathbf{x}|\hat{\boldsymbol{\alpha}}) + \frac{k}{2} \log \left( \frac{N}{2\pi} \right) + \ln \int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})} \quad (4.1)$$

where  $P(\mathbf{x}|\boldsymbol{\alpha})$  defines the model and  $\hat{\boldsymbol{\alpha}}$  is the maximum-likelihood estimate,  $k$  is the dimensionality of the parameter space and  $G_{\mu\nu}(\boldsymbol{\alpha}) = -\mathbb{E}[\partial^2 \log P(\mathbf{x}|\boldsymbol{\alpha}) / \partial \mu \partial \nu]$  is the Fisher information matrix. Intuitively, the Fisher information (FI) conveys how much the model ( $P(\mathbf{x}|\boldsymbol{\alpha})$ ) changes by a small change in parameters, and hence how much 'information' the data convey about the parameter; we will discuss FI in more detail later. This formulation is invariant to reparametrization of  $\boldsymbol{\alpha}$  – as any reasonable complexity measure should be. Let us now understand what the terms on the right hand side represent. The first term is the negative log-likelihood of the data under the best-fitting model, representing the goodness-of-fit, and the last two terms are what Rissanen refers to as the 'model complexity' characteristic of the model family. They represent the space required to encode the model itself, and thus penalize overly detailed models. In choosing among different model families, the one with the lowest MDL is to be preferred and can be shown to predict unseen data most accurately.

Rissanen's MDL method has several favorable theoretical properties, and it is also possible to show that the original algorithmic motivation of Rissanen's scheme can be cast in the language of Bayesian model selection. In particular, it is possible to show that the model complexity terms suggested by Rissanen's MDL naturally fall out of performing Bayesian model selection using a special prior over models (indexed by



parameter  $\alpha$ ) [132, 125]. This special prior – Jeffreys prior – is an uninformative prior which weights all ‘distinguishable’ models equally. As has been noted, a uniform prior over all distinguishable models is not the same as a uniform prior over the parameters  $\alpha$ ; this is evident if you consider a reparametrization of the parameter  $\alpha$  to some other parameter  $\theta$  – a uniform prior over  $\alpha$  will in general not be uniform over  $\theta$ . An obvious question is when are two models distinguishable? As argued clearly in [125], a suitable criterion for distinguishability is how easy it is to confuse  $N$  observations from one model indexed by parameter  $\alpha_1$  with another model indexed by parameter  $\alpha_2$ ; this probability falls exponentially with  $N$  and the exponent multiplying  $N$  is  $D_{KL}(\alpha_1 \parallel \alpha_2)$ : the Kullback-Leibler (K-L) divergence between the distributions parametrized by  $\alpha_1$  and  $\alpha_2$ . The K-L divergence, which appears extensively in coding and information theory, is a measure of difference between probability distributions; however,  $D_{KL}$  is not a strict distance metric – it is not symmetric in its arguments and it does not satisfy the triangle inequality. This is however not an issue for two models which are nearby in the parameter space, say indexed by  $\alpha$  and  $\alpha + d\alpha$ . The K-L divergence between them is given to leading order by

$$D_{KL}(P(\mathbf{x}|\alpha) \parallel P(\mathbf{x}|\alpha + d\alpha)) = d\alpha^\mu d\alpha^\nu G_{\mu\nu}(\alpha) + \dots$$

where,  $G_{\mu\nu}(\alpha) = -\mathbb{E}[\partial^2 \log P(\mathbf{x}|\alpha) / \partial\mu\partial\nu]$  is the Fisher information matrix. The Fisher Information *is* symmetric and satisfies the requirements for a metric on the parameter manifold. So, with the Fisher information metric on the parameter manifold,  $\sqrt{\det G(\alpha)}$  gives the density of distinguishable distributions in the neighborhood of  $\alpha$ . Thus, as clearly argued in [125], the correct uninformative prior should give equal weight to all distinguishable distributions and therefore be proportional to

$\sqrt{\det G(\boldsymbol{\alpha})}$ . The normalized Jeffreys prior is thus given by

$$J(\boldsymbol{\alpha}) = \frac{\sqrt{\det G(\boldsymbol{\alpha})}}{\int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})}}$$

Let us now see that Bayesian model selection with the Jeffreys prior does indeed penalize model complexity in the same way as the MDL criterion. If we have to choose between two model families  $f$  and  $g$  to explain some observed data  $X$ , then Bayesian model selection tells you to pick the model family with the highest posterior probability  $P(f|X)$  given some data  $X$ . To form the posterior of the family  $f$  we need the likelihood of the data for that family:  $P(X|f) = \int d\boldsymbol{\alpha} P(X|\boldsymbol{\alpha}) J(\boldsymbol{\alpha})$ . Using this we get,

$$P(f|X) = \frac{P(f)}{P(X)} \int d^k \boldsymbol{\alpha} \frac{\sqrt{\det G(\boldsymbol{\alpha})}}{\int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})}} P(X|\boldsymbol{\alpha})$$

In the absence of *a priori* knowledge we can assume all model families equally likely, so we only need to care about the integral. This can be rewritten as

$$P(f|X) \propto \frac{1}{\int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})}} \int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})} \exp \left[ -N \frac{\ln(P(X|\boldsymbol{\alpha}))}{N} \right]$$

where  $N$  is the number of independent samples in the observed data. In the limit of large  $N$ , the exponent will be dominated by the neighborhood of the extremum, so we can perform a saddle-point approximation [125] to get

$$P(f|X) \sim \frac{1}{\int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})}} \cdot \frac{\sqrt{\det G(\hat{\boldsymbol{\alpha}})}}{\sqrt{\det G^{emp}(\hat{\boldsymbol{\alpha}})}} \cdot P(X|\hat{\boldsymbol{\alpha}}) \cdot \left( \frac{2\pi}{N} \right)^{k/2}$$

where  $\hat{\boldsymbol{\alpha}}$  is the location of the extremum, and  $G_{emp}$  is the “empirical” Fisher information matrix for the observed data

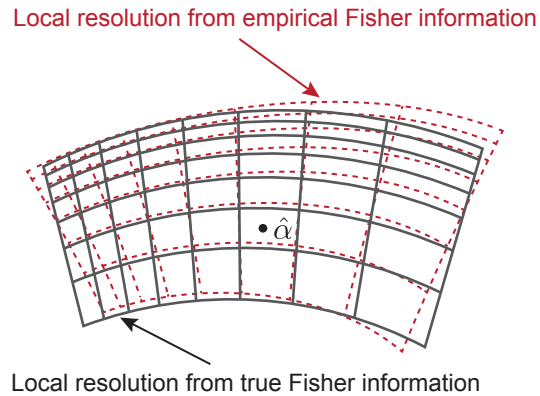
$$G_{\mu\nu}^{emp}(\boldsymbol{\alpha}) = -\frac{1}{N} \sum_{i=1}^N \partial_{\mu} \partial_{\nu} \log P(\mathbf{x}_i | \boldsymbol{\alpha})$$

Taking logarithms:

$$\begin{aligned} -\ln P(f|X) \approx & -\ln P(X|\hat{\boldsymbol{\alpha}}) + \frac{k}{2} \ln \left( \frac{N}{2\pi} \right) \\ & + \ln \int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})} + \frac{1}{2} \ln \left( \frac{\det G^{emp}(\hat{\boldsymbol{\alpha}})}{\det G(\hat{\boldsymbol{\alpha}})} \right) + \dots \end{aligned} \quad (4.2)$$

Thus, the negative log-posterior probability of a model family  $f$  is the negative log-likelihood of data under the best model in the family (goodness-of-fit) plus a series of terms penalizing complexity. These complexity terms are essentially the same as from the MDL principle (eq.4.1).

There is an additional lower order term not usually considered in the classical model complexity measures:  $\ln(\det G^{emp}(\hat{\boldsymbol{\alpha}}) / \det G(\hat{\boldsymbol{\alpha}}))$ . This is a “robustness” term [125] measuring how many ‘good’ models are in the neighborhood of the best fitting model. As illustrated in Fig. 4.1, the Fisher information metric imposes a local resolution on the parameter grid for the density of distinguishable models. When we calculate the empirical Fisher information from the observed data, there are fluctuations in this empirical metric, thus creating a fuzziness in the grid. The determinant of the Fisher information matrix measures the local volume element, so a ratio of the determinant of the empirical Fisher information matrix to the determinant of the true Fisher information matrix tells us about the relative scale of the fluctuations – smaller this term, the more distinguishable models we have close to the best fitting model, thus



**Figure 4.1:** Empirical Fisher information and robustness: The true Fisher information metric at a parameter value  $\hat{\alpha}$  gives a local resolution grid (black mesh) which describes the density of distinguishable models in the neighborhood of  $\hat{\alpha}$ . The empirical Fisher information calculated from the data gives rise to fluctuations in the metric, which introduces fuzziness in the local resolution in parameter space (red dashed mesh). The size of these fluctuations relative to the coarseness of the grid in the vicinity of the best-fitting model is a measure for robustness of the model family.

the model family is more robust in a sense (also c.f. [125]). As we will see below, this robustness term which indicates the density of ‘good’ models also accounts for the behavior of *sloppiness*.

It is appealing that the principled model selection criterion of the MDL principle also follows from a Bayesian model selection procedure using an uninformative prior which gives equal weight to all distinguishable distributions within a model family. Moreover, the Bayesian formulation has a nice geometric interpretation in the space of probability distributions. These principled measures of complexity of model families have nice theoretical properties, however, they are not always easy to use with behavioral data to infer the complexity of an inferential or predictive model. If we assume that all the subjects use models from the same (normative) model family, and that their individual differences can be attributed to variability in poorly estimated parameters, then we might be led astray in interpreting the fits to subject behavior when the subjects use models from very different model families. What is needed is

a less biasing measure of model complexity which can be estimated empirically from the data alone. Below, we describe one such principled measure and show how to use it in an inference task.

### 4.3 Empirical complexity from Predictive Information

In a comprehensive paper[10], Bialek et al., introduce the notion of predictive information as the mutual information between a chunk of a time-series and its entire future; the asymptotic growth of this predictive information then betrays the complexity of the process generating the time-series. Predictive information captures previously considered notions of complexity from the statistical mechanics and the dynamical systems literature under a common framework (see [133]). A salient aspect of predictive information as a measure of complexity is that it is a function of the data alone (provided one has sufficient data to estimate the various information-theoretic quantities), and its calculation does not need assumptions about the model families generating the data. However, as we illustrate below, in the cases where the model *is* known, the complexity measures suggested by the divergent part of the predictive information are the same as that suggested by principled measures of model complexity based on *information geometry*[124, 123, 125]. Another favorable property of the predictive information (more precisely, its divergent part) is that under some reasonable assumptions like stationarity of the time-series and invariance to reparametrization etc., the divergent part of the predictive information is the *unique* measure of complexity of a dynamical process, much in the same way that the Shannon entropy is the unique complexity measure of a random variable[134]. In this section, we first revisit the definition and properties of predictive information [10] and then show how to use

these notions to estimate the complexity of subjects' internal models in a sequential inference task.

Consider a time-series  $X(t)$ , then the predictive information in a chunk of the time-series  $I_{pred}$  is the mutual information ( $I$ ) between the past and future of the time-series in the limit of future extending to infinity. More specifically,

$$I_{pred}(T) = \lim_{T' \rightarrow \infty} I(X(-T < t < 0); X(0 < t < T'))$$

where,  $I(X, Y) = H(X) - H(X|Y)$ , and  $H()$  is the Shannon entropy. Here, stationarity is assumed, so the choice of  $t = 0$  is arbitrary. So, the predictive information in a portion of the time-series conveys the amount of information that is useful in predicting the entire future. This information,  $I_{pred}(T)$ , is clearly non-decreasing with  $T$ , and it is this rate of growth of  $I_{pred}(T)$  which betrays the complexity of the process generating the time-series. Another interesting view of predictive information, is that  $I_{pred}(T)$  is the *subextensive* part of the entropy – i.e. if  $S(T)$  is the entropy of a portion of the time-series of length  $T$ , then  $S(T)$  can be decomposed into an extensive part and a subextensive part as

$$S(T) = \mathcal{S}_0 \cdot T + S_1(T)$$

where  $S_1(T)$  is the subextensive part of the entropy and  $\mathcal{S}_0$  is some non-negative constant. It is easy to show,  $S_1(T) = I_{pred}(T)$ . Also, the subextensivity of  $S_1(T)$  implies  $\lim_{T \rightarrow \infty} S_1(T)/T = 0$  : i.e., as Bialek et al. point out, most of what we observe is useless for predicting the future.

The predictive information can display three qualitatively different behaviors in the

limit of large  $T$ :

1.  $I_{pred}(T)$  remains constant
2.  $I_{pred}(T)$  can grow logarithmically as  $(k/2) \log(T)$
3.  $I_{pred}(T)$  can grow as  $T^\alpha$  with  $\alpha < 1$

The first case is true for deterministic or very regular sequences, the second case is observed when the dynamics are generated by a model parametrized in a  $k$ -dimensional parameter space and the last case is observed for ‘non-parametric’ models, where the number of parameters to be learned grows with as you observe more data. The first two growth scenarios have already been observed by Rissanen in the analysis of stochastic complexity[129]. The fact that the predictive information framework subsumes ‘non-parametric’ models – models which grow in complexity with more data – make this measure appealing. It is also worth noting a few examples: i) for a purely random sequence with independent samples the entropy is purely extensive, so  $I_{pred} = 0$  : there is nothing to be learned; ii) for a purely periodic sequence,  $I_{pred}$  asymptotes to a constant, which is larger for larger periods – this is consistent with our intuitive notions of complexity of a dynamical process. Let us now see how the complexity measures suggested by predictive information relate to those suggested by information geometric measures.

### **4.3.1 Predictive information-based complexity of a model family**

We noted that the predictive information is an unbiased and universal measure of complexity of a dynamical model; the growth of the subextensive component of en-

trophy captures the complexity class of the model family. It is useful to see how the complexity of a model family as prescribed by the growth of the subextensive component of the entropy compares to the complexity measures of a model family from Information Geometry as described earlier. Following the setup in [10], let us consider the entropy of  $N$  independent samples drawn from a parametrized family of probability distribution and calculate the subextensive component of the entropy. Suppose we observe  $N$  independent samples  $\mathbf{x}_1 \dots \mathbf{x}_N$  from a parametrized probability distribution  $Q(\mathbf{x}_1, \dots, \mathbf{x}_N | \boldsymbol{\alpha})$ , and let us also suppose that *a priori* we treat all *distinguishable distributions* indexed by this family as equally likely; i.e. the prior on the parameters  $\boldsymbol{\alpha}$  is the Jeffreys prior described above. Then the entropy of the total distribution of the  $N$  samples,  $P(\mathbf{x}_1, \dots, \mathbf{x}_N)$ , is given by

$$S(N) = - \int d\mathbf{x}_1 \cdots d\mathbf{x}_N P(\mathbf{x}_1, \dots, \mathbf{x}_N) \ln P(\mathbf{x}_1, \dots, \mathbf{x}_N)$$

where

$$P(\mathbf{x}_1, \dots, \mathbf{x}_N) = \int d^k \boldsymbol{\alpha} \frac{\sqrt{\det G(\boldsymbol{\alpha})}}{V} \prod_{i=1}^N Q(\mathbf{x}_i | \boldsymbol{\alpha})$$

$$V = \int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})}$$

We can also rewrite  $P(\mathbf{x}_1, \dots, \mathbf{x}_N)$  as

$$P(\mathbf{x}_1, \dots, \mathbf{x}_N) = \prod_{j=1}^N Q(\mathbf{x}_j | \bar{\boldsymbol{\alpha}}) \int d^k \boldsymbol{\alpha} \frac{\sqrt{\det G(\boldsymbol{\alpha})}}{V} \times$$

$$\exp \left( -N \cdot \left\{ -\frac{1}{N} \sum_{i=1}^N \log_2 \left[ \frac{Q(\mathbf{x}_i | \boldsymbol{\alpha})}{Q(\mathbf{x}_i | \bar{\boldsymbol{\alpha}})} \right] \right\} \right)$$



This form of  $P(\mathbf{x}_1, \dots, \mathbf{x}_N)$  lends itself to an interesting interpretation: there's some "true" set of parameters  $\bar{\alpha}$  which give rise to the data and the averaging over all the parameters is weighted by a term which decreases exponentially for parameters which are "far away" from the true parameters. In the limit of large  $N$  it is also clear that

$$D_{KL}^{emp}(\bar{\alpha} \parallel \alpha) \equiv -\frac{1}{N} \sum_{i=1}^N \log_2 \left[ \frac{Q(\mathbf{x}_i | \alpha)}{Q(\mathbf{x}_i | \bar{\alpha})} \right] \rightarrow D_{KL}(\bar{\alpha} \parallel \alpha)$$

where  $D_{KL}(\bar{\alpha} \parallel \alpha)$  is the true Kullback-Leibler divergence between the distributions indexed by the two parameters, and  $D_{KL}^{emp}(\bar{\alpha} \parallel \alpha)$  is the "empirical" Kullback-Leibler divergence estimated based on the observed data. So, the exponential term decreases with the KL divergence from the true distribution. This allows us to write the entropy as

$$S(N) = \mathcal{S}_0 \cdot N + S_1(N)$$

Where the extensive term  $\mathcal{S}_0 \cdot N$  is given by

$$\mathcal{S}_0 \cdot N = N \cdot \left( - \int d^k \alpha \frac{\sqrt{\det G(\alpha)}}{V} \int d\mathbf{x} \cdot Q(\mathbf{x} | \alpha) \ln Q(\mathbf{x} | \alpha) \right)$$

and the subextensive component of the entropy is given by

$$S_1(N) = - \int d^k \bar{\alpha} \frac{\sqrt{\det G(\bar{\alpha})}}{V} \cdot \ln \left[ \int d^k \alpha \frac{\sqrt{\det G(\alpha)}}{V} \exp(-N \cdot D_{KL}^{emp}(\bar{\alpha} \parallel \alpha)) \right]$$

The term inside the logarithm is reminiscent of a partition function  $Z(\bar{\alpha}; N)$  with the number of samples  $N$  playing the role of inverse temperature. The growth of the

subextensive entropy (and hence model complexity) depends on how this partition function grows with  $N$ . For sufficiently large values of  $N$ , only models with small KL divergences will contribute to  $Z(\bar{\alpha}; N)$ . We can rewrite  $Z(\bar{\alpha}; N)$  as

$$Z(\bar{\alpha}; N) = \int d\varepsilon \rho(\varepsilon; \bar{\alpha}) \exp(-N \cdot \varepsilon)$$

where  $\rho(\varepsilon; \bar{\alpha})$ , the density of models with KL divergence  $\varepsilon$  from the “true” model  $(\bar{\alpha})$  is given by

$$\rho(\varepsilon; \bar{\alpha}) = \int d^k \alpha \frac{\sqrt{\det G(\alpha)}}{V} \delta(\varepsilon - D_{KL}^{emp}(\bar{\alpha} \parallel \alpha))$$

In the limit of small  $\varepsilon$  we can assume  $\alpha$  and  $\bar{\alpha}$  are close, and carry out a spherical integral in  $k$  dimensions to get

$$\rho(\varepsilon; \bar{\alpha}) \approx \frac{\sqrt{\det G(\bar{\alpha})}}{V} \frac{(2\pi)^{k/2}}{\Gamma(k/2)} \cdot \frac{1}{\sqrt{\det G^{emp}(\bar{\alpha})}} \varepsilon^{k/2-1}$$

where  $G^{emp}(\bar{\alpha})$  is the empirical Fisher information matrix we encountered before

$$G_{\mu\nu}^{emp}(\bar{\alpha}) = -\frac{1}{N} \sum_{i=1}^N \partial_\mu \partial_\nu \ln Q(\mathbf{x}_i | \bar{\alpha})$$

This give us

$$Z(\bar{\alpha}; N) \propto \left(\frac{2\pi}{N}\right)^{k/2} \cdot \frac{1}{V} \cdot \sqrt{\frac{\det G(\bar{\alpha})}{\det G^{emp}(\bar{\alpha})}}$$

and

$$\begin{aligned}
S_1(N) &= - \int d^k \bar{\alpha} \frac{\sqrt{\det G(\bar{\alpha})}}{V} \cdot \ln [Z(\bar{\alpha}; N)] \\
&\approx \frac{k}{2} \ln \left( \frac{N}{2\pi} \right) + \ln V + \left\langle \frac{1}{2} \ln \left[ \frac{\det G^{emp}(\bar{\alpha})}{\det G(\bar{\alpha})} \right] \right\rangle_{\bar{\alpha}} + \dots
\end{aligned}$$

Therefore, the subextensive component of entropy not only asymptotically captures the complexity class as shown in [10], but even the lower order terms behave in a way exactly like the model family complexity suggested by information geometry and MDL (see eq. 4.2). It is comforting that these different principled measures of complexity agree with each other tightly. What then is different between them? The key difference is that the predictive information is a function of the data, and doesn't depend on the assumptions about the model family. In particular, it also subsumes 'non-parametric' cases where the number of effective parameters grows as a function of the amount of data observed.

### 4.3.2 Predictive information in an inference task

Can the asymptotic results of the predictive information framework be applied to actual datasets of human behavior? Here we use the predictive information framework to assess the complexity of the internal models used by subjects performing an online inference task. Furthermore, we not only want to get a measure of the subject's model complexity, but also find out whether the subjects who use the more complex models also learn better models of the environment; by 'better' here, we mean being able to predict the future more accurately. It is entirely plausible that the subjects learn irrelevant details about the sequence. To this end, we use the *Information Bottleneck*

(IB) framework[135], which allows us to systematically assess what is the optimal compression of the information from the past in order to predict the future. The IB framework provides a principled performance-vs-complexity frontier that we need. Subjects who learn more complicated models from the past should be better able to predict the future, and thus lie close to the frontier.

We first describe the pertinent details of the inference task which we require to calculate the information-theoretic quantities; for complete details, c.f.[122]. The task requires the subjects to report the source of a noisy stimulus generated from one of two overlapping Gaussian distributions. The true identity of the Gaussian generating the stimulus flips with a Bernoulli process with a certain hazard rate; this hazard rate itself changes with a much slower ‘meta-hazard rate’. So, there are some blocks of observations where the true source flips frequently and some blocks where the true source remains the same for longer.

Let us denote the stimulus at time  $t$  by  $X_t \in \mathbb{R}$ , the source identity by  $S_t \in \{0, 1\}$  let  $H_t$  be the hazard rate at time  $t$  and  $R_t \in \{0, 1\}$  denote the subject response at time  $t$ . Given the complete task sequence  $\Phi_t \equiv \{H_t, S_t, X_t\}$  and the subject responses  $\{R_t\}$ , how do we make an unbiased estimate of the subjects’ internal model complexity and to what extent is the subject using a ‘good’ model? We can measure how much information about the past observations are contained in a subject’s current responses ( $I_{past}$ ), and compare this to the information these responses contain about the future observations ( $I_{future}$ ). We can compare this to a theoretical boundary for  $I_{past}$  vs  $I_{future}$  uniquely defined for the process generating the task sequence to assess whether the subjects are using a ‘good’ model. All subjects using good models – regardless of the complexity of the models – will lie close to this boundary. We first summarize the information bottleneck method below.

### 4.3.3 Measuring complexity using the information bottleneck approach

The Information Bottleneck approach proposed by Bialek et al.[135] defines a non-parametric theoretical limit for predictive information about the future of a model (measured in bits;  $I_{future}$ ) as a function of how much information the model extracts from the past (measured in bits;  $I_{past}$ ). Here we summarize the information bottleneck approach of Bialek et al. applied to our inference task. Let us denote all the past observations leading up to time  $t$  as  $\Phi_{past} = \Phi_{t' < t}$  and all the future observations as  $\Phi_{future} = \Phi_{t' > t}$ . What we've been referring to as 'good' models so far, can extract useful information from the past observations to make accurate predictions about the future. However, we haven't specified how much useful information to extract from the past, or in other words, *how complex the model should be?* For a given model complexity – as measured by the information extracted from the past observations – a 'good' model is one which can maximize the accuracy of predictions about the future. In the information bottleneck framework this predictive ability is again measured in terms of mutual information. Specifically, let us consider an intermediate representation  $\Lambda$  which captures the useful information from the past observations most relevant to predicting the future, then we want to maximize the mutual information between  $\Phi_{future}$  and  $\Lambda$  :  $I(\Lambda, \Phi_{future})$  while at the same time keeping the representation  $\Lambda$  parsimonious by minimizing  $I(\Lambda, \Phi_{past})$  . The trade-off between these two requirements is controlled by a Lagrange multiplier  $\beta$ , which decides our preference for compact intermediate representations or better future predictions. We assume that we know the joint distribution,  $P(\Phi_{past}, \Phi_{future})$ , and the information bottleneck method gives us an optimal intermediate representation  $\Lambda$  (for a given  $\beta$ ) by specifying  $P(\Lambda | \Phi_{past})$ . To find this optimal intermediate representation, we

minimize the following functional of  $P(\Lambda | \Phi_{past})$

$$\mathcal{L}[P(\Lambda | \Phi_{past})] = I(\Lambda, \Phi_{past}) - \beta I(\Lambda, \Phi_{future})$$

The (locally) optimal solution can be found by iterating the following equations, which are reminiscent of the Arimoto-Blahut algorithm in rate distortion theory:

$$\begin{aligned} P_{n+1}(\Lambda | \Phi_{past}) &= \frac{P_n(\Lambda)}{Z_n(\Phi_{past}, \beta)} \exp[-\beta D_{KL}(P_n(\Phi_{future} | \Phi_{past}) || P_n(\Phi_{future} | \Lambda))] \\ P_{n+1}(\Phi_{future} | \Lambda) &= \frac{1}{P_n(\Lambda)} \sum_{\Phi_{past}} P_n(\Phi_{future} | \Phi_{past}) P_n(\Lambda | \Phi_{past}) P_n(\Phi_{past}) \\ P_{n+1}(\Lambda) &= \sum_{\Phi_{past}} P_n(\Lambda | \Phi_{past}) P_n(\Phi_{past}) \end{aligned}$$

where  $Z(\cdot, \cdot)$  is a normalizing constant and  $n$  is the iteration number. As we change  $\beta$  from 0 to a large value we go from a trivially compact representation ( $\Lambda$  is a single point) to more and more detailed intermediate representations which capture more details about  $\Phi_{past}$ . This behavior is reminiscent of the rate-distortion function from information theory which specifies the boundary of the minimum allowable rate (in bits) for compressing a source for a given amount of 'distortion'. In the case of the information bottleneck, the KL-divergence naturally emerges as the correct distortion measure. As we vary  $\beta$ , we get an information-bottleneck (IB) curve in the two-dimensional plane of  $I(\Lambda, \Phi_{future})$  vs.  $I(\Lambda, \Phi_{past})$  which specifies the allowed region in this plane. For a given value of  $I(\Lambda, \Phi_{past})$  (or model complexity), this curve specifies the maximum predictability  $I(\Lambda, \Phi_{future})$  for the particular generative process. No model learned from data can outperform this.

Our approach in assessing the complexity of subject behavior and whether they are

learning good models, is to estimate these quantities for subjects and see whether they fall close to the theoretical boundary in the  $I(\Lambda, \Phi_{future})$  vs.  $I(\Lambda, \Phi_{past})$  plane. If this is the case, then it suggests that even though subjects might use internal models of varying complexity, they are doing as well as they can in learning the correct model, given the complexity of the model they choose. Note, this need not necessarily be the case; it is possible that subjects retain details from the past which are not useful to predict the future (and hence in inference) which will lead to high values of  $I(\Lambda, \Phi_{past})$  but lower values of  $I(\Lambda, \Phi_{future})$ .

In our case, the Markovian structure of the task allows the relevant information theoretic quantities to be estimated efficiently. Particularly, mutual information between the future and the past is  $I(\Phi_{past}, \Phi_{future}) = I(\Phi_t, \Phi_{t+1})$  and  $I(\Phi_{past}, \Lambda_t) = I(\Phi_{t-1}, \Lambda_t)$ . For simplicity, the time-varying hazard rate  $H_t$  is sampled uniformly from a fixed set of values. This gives us the full distribution of  $P(\Phi_t, \Phi_{t+1})$ , and we can use the IB method to determine the boundary in the  $I(\Lambda, \Phi_{future})$  vs.  $I(\Lambda, \Phi_{past})$  plane. To estimate these quantities for the subjects, we measure the mutual information between the subject responses at  $t - 1$ ,  $t - 2$  and the observation at time  $t$ . Specifically, we measure  $I(\{R_{t-2}, R_{t-1}\}, \Phi_t)$  and  $I(\{R_{t+2}, R_{t+1}\}, \Phi_t)$ , where  $R_t$  is the subject response at time  $t$ , which are then compared to the theoretical boundary. In forthcoming work, we analyze the subject behavior from a variety of tasks with similar structure and compare the complexity of behavior to the theoretical bounds (results presented elsewhere). This approach offers a nonparametric way to study what proportion of individual variability arises due a bias towards simple/complex models.

## 4.4 Resolution of observations, *sloppy* models and complexity

The classical measures of complexity suggested by MDL, Information Geometry and Predictive Information are principled and consistent with each other. However, all of these measures require an assumption of sufficiently large sample sizes, and are strictly exact only with infinite data. They assume that we are able to probe *all* the degrees of freedom of a given model and do not account for limitations in the data size or the measurement process. This is apparent by noting that the form of the complexity penalty suggested by these measures has no term accounting for the nature of measurements:

$$\text{complexity penalty} = \frac{k}{2} \log \left( \frac{N}{2\pi} \right) + \ln \int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})}$$

Limitations in data size or in the nature of measurements can have interesting geometric implications in the behavior space of the model. In particular, when the measurement scheme poorly captures certain degrees of freedom of the model, then variations in parameter combinations along these directions have little effect on the observations. James Sethna and his collaborators have done extensive work in characterizing this phenomenon, and they refer to it as *sloppiness* [126, 127, 128, 11]. Conversely, along the *stiff* directions, small parameter changes result in large behavioral changes – i.e., the measurements robustly capture those degrees of freedom. Let us now consider an example problem of fitting exponentials which clearly illustrates

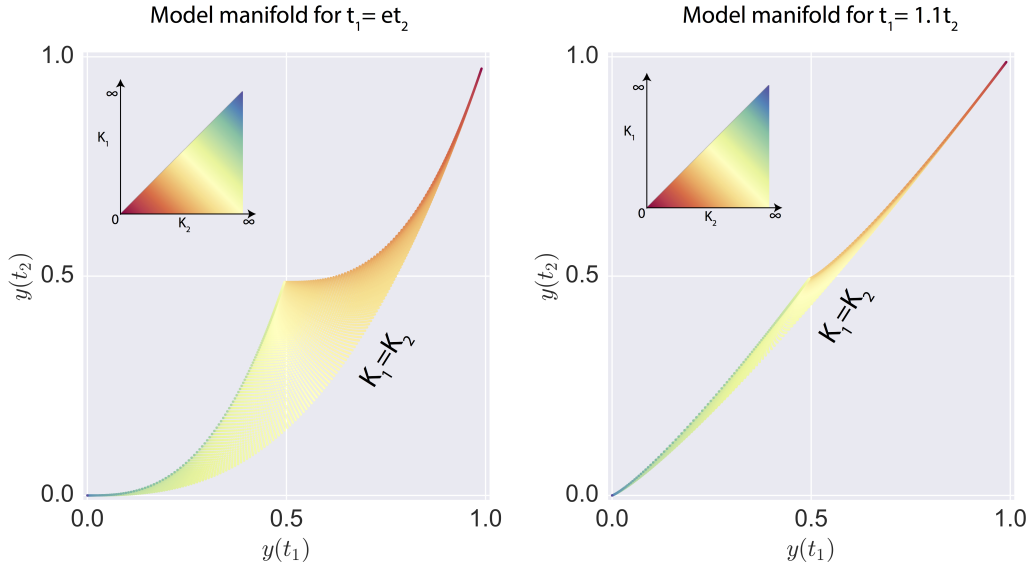


the issues of measurements and sloppiness [127]. Suppose we have a model given by

$$y(t) = \frac{1}{2} (e^{-k_1 t} + e^{-k_2 t}) + \eta_t$$

where  $y(t)$  are model outputs,  $k_1$  and  $k_2$  are unknown parameters and  $\eta_t$  is observation noise with some standard deviation  $\sigma$ . Let us assume that we observe the output of the model  $y(t)$  at two times  $t_1$  and  $t_2$ . The behavior space of the model will be a two dimensional manifold – the model manifold – in the  $y(t_1)$  and  $y(t_2)$  plane. Now, let us plot the average values of  $y(t_1)$  and  $y(t_2)$  which are possible for all possible values of the parameters  $k_1$  and  $k_2$  – i.e. we look for a mapping from the parameter space to the behavior space. Fig. 4.2 shows this mapping for the two different values of  $t_1$  and  $t_2$ . As is seen for both values of  $t_1$  and  $t_2$ , the model manifold is two-dimensional and has a longer dimension (along the  $k_1 = k_2$  line) and a shorter dimension. In fact, in the second case, the shorter direction is quite small compared to the longer direction. Therefore, moving large distances in the parameter space corresponding to the short direction will not make much of a difference in the measurements, and we can capture the macroscopic behavior of the model by collapsing the parameters into an effective parameter  $k = k_1 = k_2$ , thus yielding an effective, simpler model of the data.

The particular behavior described above for the two-exponential model, is a much more general phenomenon. Our measurements are typically agnostic of the microscopic details of the models or processes generating the observations. Therefore, in general scenarios the measurements do not probe all the degrees of freedom of the system equally well. This limitation can arise due to having finite data or some other fundamental limitations in the measurement scheme. As a consequence, the measurements are not able to pin down models uniquely and many of the parameter



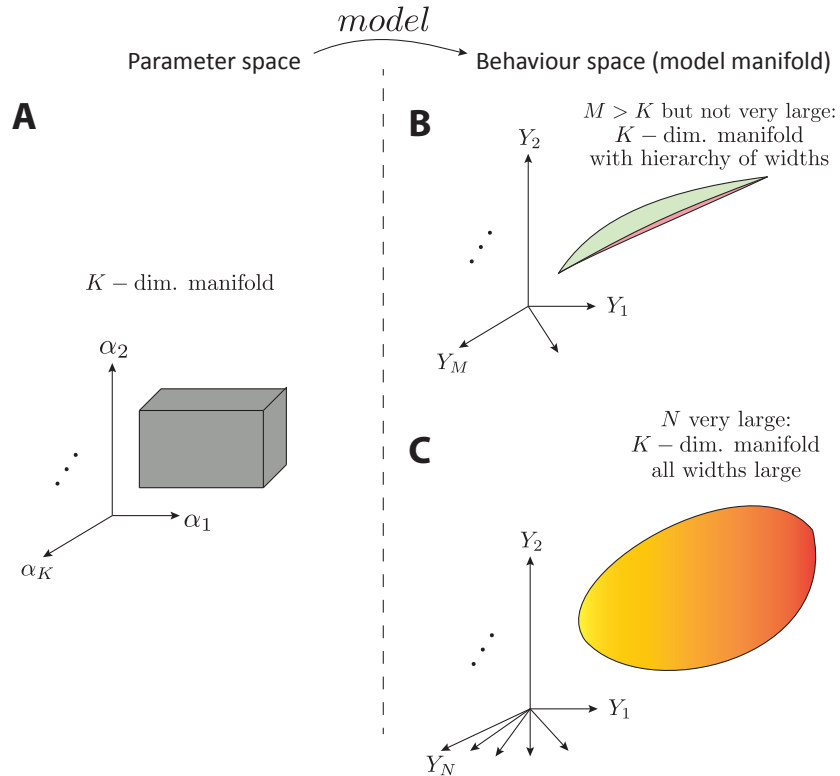
**Figure 4.2:** *Sloppy* and *stiff* directions: the panels show the ‘model manifold’ for the sum of exponentials model described in the text. The model manifold in this case is a 2-D manifold of all possible measurements  $y(t_1)$  and  $y(t_2)$  at two times  $t_1$  and  $t_2$  as the parameters  $k_1$  and  $k_2$  are varied over all permissible values (insets). It is a mapping from the parameter space to the behavior space. For generic measurements, the model manifold exhibits a hierarchy of widths. In this case, there are two widths and the longest direction is along the  $k_1 = k_2$  curve. When we move along the shorter direction (in the model manifold) we traverse large distances in the parameter manifold; thus this direction is less relevant as far as the behavior of the model is concerned on the measurement scale. Especially in the right panel, we can capture most of the model behavior by an effective simpler model with  $k_1 = k_2$

combinations are unconstrained by our observations, and we can make large variations along these directions in the parameter space without observing noticeable changes in our measurements. The directions in the parameter space which don’t have noticeable effects on the measured model behavior are referred to as sloppy directions and the ones which are important for the measured behavior are referred to as stiff directions by Sethna et al. A range of models taken from many different fields display the existence of sloppy parameter combinations[11].

The phenomenon of sloppiness immediately suggests that if our measurements are going to probe the system at a particular resolution which doesn’t capture all the degrees of freedom, then we can reduce the detailed multi-parameter model to an ‘effective’

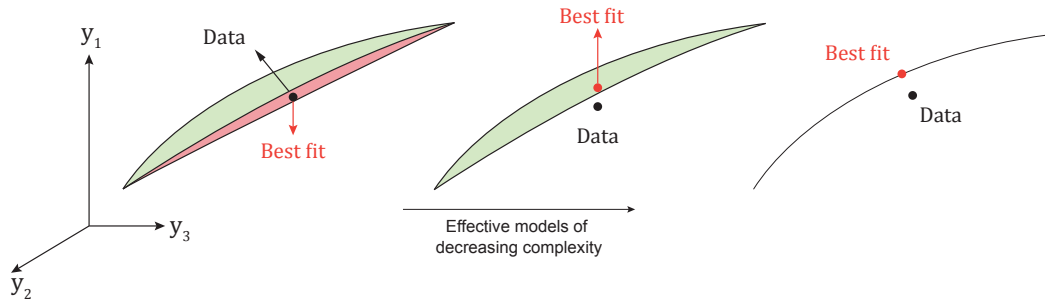
lower dimensional model (along the stiff directions) which captures all the behavior at our chosen resolution. Since the presence of sloppy directions is an ubiquitous phenomenon, our notions of model complexity must take this into account. The classical notions of model complexity discussed above do not explicitly capture this. To see this, we graphically illustrate the difference between the scenarios assumed by classical model complexity measures and typical scenarios in which measurements exhibit sloppiness in Fig. 4.3. Let us consider a model parametrized by  $K$  parameters; the parameter space will be a  $K$ -dimensional manifold in general (Fig. 4.3A). If we take  $M > K$  measurements  $Y_1 \dots Y_M$ , then the behavior space is a  $M$ -dimensional ambient space and if we map each model (specified by a point in the parameter space) to its average behavior in the behavior space, then we get another  $K$ -dimensional manifold in the ambient  $M$ -dimensional space – this is the model manifold. In typical scenarios, even when  $M$  is larger than  $K$  the model manifold has a hierarchy of widths, i.e. there are many thin directions relative to measurement noise (Fig. 4.3B). This *hyper-ribbon* structure is a key insight from the work of Sethna et al., and what it means is that whenever your measurement scheme is not precisely tuned to measure all the degrees of freedom of the model, there will be many degrees of freedom which will be poorly captured by the measurements (sloppy directions), and large parameter changes along these sloppy directions will make little difference in the measured output (relative to noise). However, when we start taking more and more measurements, in the limit of infinite measurements, the parameter combinations along the sloppy directions will eventually be captured by some measurement direction, and the model manifold widths will increase and become large relative to the measurement noise (Fig. 4.3C). What this means is that with infinite measurements *all* the degrees of freedom of the model can be probed for any generic measurement scheme. This (Fig. 4.3C) is the regime assumed by the classical model complexity measures; however, in

typical scenarios we are in the regime of (Fig. 4.3B) not (Fig. 4.3B).



**Figure 4.3:** Model manifolds with finite and infinite data: **(A)** schematic of generic  $K$ -dimensional parameter space. **(B)** the behaviour space with  $M > K$  measurements. The parameter manifold is transformed by the model into another  $K$  dimensional manifold – the model manifold. Importantly, for most generic measurement schemes, the model manifold exhibits a hyper-ribbon structure, i.e., there is a hierarchy of widths and several manifold widths are very small compared to measurement noise. This is because our measurements don't capture the degrees of freedom of the model corresponding to these directions. **(C)** In the limit of infinite data all the model manifold widths become large compared to the measurement noise. This is the scenario considered by most classical model complexity measures – i.e., they assume you can probe all degrees of freedom of the model.

Why should we care about the existence of sloppy directions when considering model complexity? As indicated above, in scenarios where there are sloppy directions we can consider a nested family of lower-dimensional 'effective' models which capture most of the model behavior at the our chosen scale of observation. These nested family of effective models which provide increasingly simpler explanations of the measured data. This is illustrated by a schematic of a toy model in Fig 4.3: the three model manifolds correspond to effective model families of decreasing complexity, and the data is illus-



**Figure 4.4:** Model selection with a sloppy model: Left-most panel shows the schematic of a model manifold corresponding to three measurements  $y_1, y_2, y_3$  for a model. The model manifold for the full model has a long direction, a thinner direction and a very thin direction. The panels to the right show progressively simpler effective model families where certain parameter combinations have been compressed into an effective parameter to capture the salient aspects of the model behavior for the measurements. When selecting a model to explain observed data (black dot), in each model family we pick the model nearest to the data (red dot). The full model provides the best fit to data but at the cost of a more complex description, and the simpler models provide fits which are not as good, but with simpler descriptions.

trated by the black dot. For each model family, the best fitting model corresponds to the parameters which represent the model closest in Euclidean distance to the data. The goodness-of-fit decreases with increasing simplicity of the models. How should we perform model selection from these nested family of models? In other words, what is the correct penalty for model complexity in this case? One would expect a reasonable model complexity measure to favor these effective lower-dimensional models which neglect parameter combinations along the sloppy directions. The key distinction from the classical case is that the *complexity terms must explicitly account for the resolution at which you will probe the system.*

We argue that the lower order terms – especially the Robustness term – in the expansion of the Bayesian posterior (eq. 4.2), which are not typically considered for model selection capture the correct measurement-dependent trade-off between goodness-of-fit and model complexity in the presence of sloppiness:

complexity terms:  $\frac{k}{2} \ln \left( \frac{N}{2\pi} \right) + \ln \int d^k \boldsymbol{\alpha} \sqrt{\det G(\boldsymbol{\alpha})} + \frac{1}{2} \ln \left( \frac{\det G^{emp}(\hat{\boldsymbol{\alpha}})}{\det G(\hat{\boldsymbol{\alpha}})} \right)$

To see this, let us consider the Robustness term:

$$\frac{1}{2} \ln \left( \frac{\det G^{emp}(\hat{\boldsymbol{\alpha}})}{\det G(\hat{\boldsymbol{\alpha}})} \right)$$

we can rewrite this terms as

$$\frac{1}{2} \ln \left( \frac{\det G^{emp}(\hat{\boldsymbol{\alpha}})}{\det G(\hat{\boldsymbol{\alpha}})} \right) = \frac{1}{2} \sum_{i=1}^K \ln \left( \frac{\lambda_i^{emp}}{\lambda_i} \right)$$

where  $\{\lambda_i^{emp}\}$  and  $\{\lambda_i\}$  are the eigenvalues of  $G^{emp}(\hat{\boldsymbol{\alpha}})$  and  $G(\hat{\boldsymbol{\alpha}})$  respectively. We now make a few observations:

- The sloppy directions correspond to the degrees of freedom which our measurements are not able to capture well, and therefore the eigenvalues  $\lambda_i$  will be the smallest corresponding to these directions.
- In the limit of large  $N$ , the empirical Fisher information matrix converges to the true FI matrix, therefore

$$\lim_{N \rightarrow \infty} \lambda_i^{emp} = \lambda_i$$

- Importantly, the convergence of the empirical eigenvalues will be the *slowest for the sloppiest directions*. This is because, it is precisely these directions our

measurements are poor at probing.

With these considerations, under mild assumptions, and in the limit of large but not infinite  $N$  we can capture this convergence as

$$\lambda_i^{emp} \approx \lambda_i + \frac{C_i(\sigma)}{N^{\gamma_i}}$$

where  $C_i(\sigma)$  is a noise dependent term and  $N^{\gamma_i}$  is a term which increases with  $N$ ; moreover, the convergence will be slowest for the sloppiest directions – i.e.  $\gamma_i$  will be smallest for these directions and for a given value of  $N$ , the  $N^{\gamma_i}$  term will be smallest for the sloppiest directions. Let us now rewrite the Robustness term with this approximation:

$$\frac{1}{2} \ln \left( \frac{\det G^{emp}(\hat{\boldsymbol{\alpha}})}{\det G(\hat{\boldsymbol{\alpha}})} \right) \approx \frac{1}{2} \sum_{i=1}^K \ln \left( 1 + \frac{C_i(\sigma)}{\lambda_i \cdot N^{\gamma_i}} \right)$$

For a stiff direction, the widths are large compared to measurement noise and so  $\lambda_i \gg \sigma$ , thus  $C_i(\sigma)/\lambda_i$  will be small and  $N^{\gamma_i}$  will be large, thus

$$\ln \left( 1 + \frac{C_i(\sigma)}{\lambda_i \cdot N^{\gamma_i}} \right) \approx 0 \quad \text{for stiff directions}$$

i.e., the Robustness term adds nothing to the complexity penalty for stiff directions. However, for the sloppy directions  $\lambda_i \ll \sigma$  and  $N^{\gamma_i}$  will be small therefore

$$\frac{C_i(\sigma)}{\lambda_i \cdot N^{\gamma_i}} \gg 1 \quad \text{for sloppy directions}$$

i.e. the Robustness term penalises model families with sloppiness depending the

amount of sloppiness. Therefore, this lower-order term which is usually ignored in model selection naturally captures the complexity penalty notions which depend on the resolution of measurements, and it forces the nested model selection towards a simpler effective model which captures most of the behaviour at our scale of observation. In forthcoming work, we make these ideas more precise and illustrate model selection by means of simple examples.

The original motivation for the using the Jeffreys prior was to weight all *distinguishable models* equally. In particular, if we observed  $N$  data samples, then the confusability between nearby models decreased exponentially with the K-L divergence between them. Thus, this requirement of distinguishability effectively discretized the parameter space with each grid element representing a distinguishable model. In the limit  $N \rightarrow \infty$ , the grid became smaller and we got a *density* of distinguishable models which was nothing but the Jeffreys prior. We suggest that in the sloppy model setting the correct prior to use is again the prior which *weights all distinguishable models equally*. However, in this case since  $N$  need not be large, the *empirical Fisher Information:  $G^{emp}$*  will become relevant. With this caveat, the model selection follows as before: we select the model family from a nested set which offers the best compromise between complexity and goodness-of-fit.

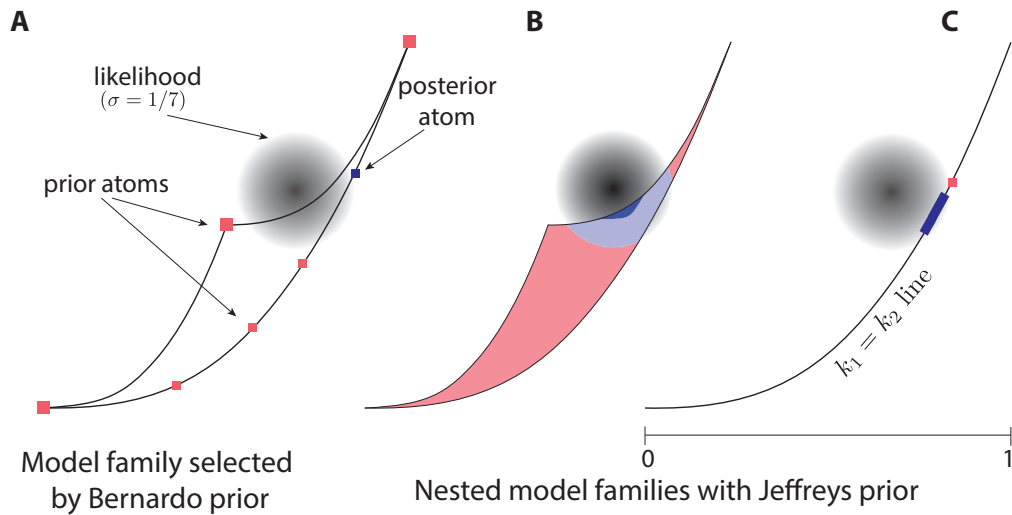
Another interesting approach for model selection for models exhibiting sloppy behavior has been suggested by Mattingly et al. [136], where they use a special kind of uninformative reference prior – the Bernardo prior – to select models in these settings. The Bernardo prior is a reference prior  $p_*$  which maximizes the mutual information between the parameters and the expected data:



$$p_*(\boldsymbol{\alpha}) = \arg \max_{p(\boldsymbol{\alpha})} I(X, \boldsymbol{\alpha})$$

This prior generally does not yield a closed-form analytical solution. In situations where the resolution of the observed data is poor, the Bernardo prior puts all its weight on the models which reside on the boundary of the model manifold [136]. Moreover, the prior is a collection of delta functions (atoms) on the boundaries [137, 138]. So, once you define the resolution of the data (say, by defining the noise variance  $\sigma$ ), the Bernardo prior immediately gives you a model family to work with: this family consists of a finite number of models on the boundaries of the manifold corresponding to the atoms of the prior. As an example, Fig. 4.4A shows the Bernardo prior for the two-exponential model described above, when the noise variance  $\sigma = 1/7$  (figure adapted from Mattingly et al. [136]). We see that the Bernardo prior in this case restricts the model family under consideration to a family with seven models corresponding to the atoms (red points in Fig. 4.4A). As the resolution of data improves, (as  $\sigma$  decreases) more atoms are placed on the interior regions.

Model selection with the Bernardo prior has notable differences with the model selection using the appropriate Jeffreys prior on a nested family of increasingly simpler models. This is illustrated in Fig. 4.4. In this schematic, the data, due to noise, falls outside the model manifold and the likelihood function is a Gaussian centered on the data (represented by the grey cloud). The Bernardo prior is simply the collection of atoms and the Jeffreys prior in the nested family of models is the *uniform prior* over the model manifold; note that although the Jeffreys prior will not be flat in the parameter space it will be flat in the *behavior space*. So, for e.g., the Jeffreys prior



**Figure 4.5:** Model selection with sloppy directions using the Bernardo and Jeffreys prior: **(A)** The panels show the model manifold corresponding to the sum of two exponentials model described in the text. A data point and its corresponding likelihood function are shown as a grey cloud. The atoms of the Bernardo prior corresponding to a certain noise level ( $\sigma = 1/7$ ) are shown as red points. The posterior puts all the weight on the model corresponding to the atom shown in blue. (adapted from Mattingly et al. [136]) **(B)** Same setup as in (A) except that the prior is the Jeffreys prior, which is flat over the model manifold (red shading). The posterior is concentrated on the boundary close to the data point, but there is also some mass inside model manifold (blue shading). **(C)** The model manifold corresponding to a simpler 1-dimensional effective model ( $k_1 = k_2$ ). The Jeffreys prior is flat on the model manifold, and the posterior is concentrated on the region of the model manifold corresponding to the thick blue strip. For reference, the model picked using the Bernardo prior is also shown as a red point.

in the 2 parameter family corresponds to a flat prior over the 2-dimensional model manifold (red shading in Fig. 4.4B) and for the 1-parameter model it is the flat prior over the  $k_1 = k_2$  line. The posteriors in the three cases is shown in blue: in the first case, the posterior basically has all the weight on one of the atoms of the Bernardo prior (blue point in Fig. 4.4A); in the 2-parameter family with Jeffreys prior the posterior is mostly concentrated on the boundary near the data but has non-zero weight in the interior as well (blue region in Fig. 4.4B); in the 1-parameter family with Jeffreys prior the posterior (thick blue line in Fig. 4.4C) is localized to a section of

the 1-dimensional model manifold. For reference, the model selected by the Bernardo posterior is also shown (red point in Fig. 4.4C).

Our aim is not to provide an explicit numerical algorithm for model selection, but rather we are interested in understanding the nature of the terms that penalise the complexity of the model family when there are sloppy parameter combinations. And, we suggest that in this case, using a prior that again weights all *distinguishable* models equally will represent the balance between complexity and goodness-of-fit. Since, the Bernardo prior does not allow a simple form, it is difficult to study the exact nature of the complexity terms arising out of this prior. It will be very interesting to connect this model selection procedure to the one suggested by Jeffreys prior and Information Geometry – especially the relation between the Robustness term and the spacing between atoms near the best model. In forthcoming work we aim to elaborate on the links between these two methods.

# Bibliography

- [1] R. Geirhos, D. H. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, “Comparing deep neural networks against humans: object recognition when the signal gets weaker,” *arXiv preprint arXiv:1706.06969*, 2017.
- [2] W. R. Softky and C. Koch, “The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps,” *Journal of Neuroscience*, vol. 13, no. 1, pp. 334–350, 1993.
- [3] R. R. d. R. van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek, “Reproducibility and variability in neural spike trains,” *Science*, vol. 275, no. 5307, pp. 1805–1808, 1997.
- [4] C. Van Vreeswijk, H. Sompolinsky, *et al.*, “Chaos in neuronal networks with balanced excitatory and inhibitory activity,” *Science*, vol. 274, no. 5293, pp. 1724–1726, 1996.
- [5] S. Laughlin, “A simple coding procedure enhances a neuron’s information capacity,” *Zeitschrift für Naturforschung c*, vol. 36, no. 9-10, pp. 910–912, 1981.
- [6] C. W. Yu, K. A. Prokop-Prigge, L. A. Warrenburg, and J. D. Mainland, “Draw-

- ing the borders of olfactory space,” in *Chemical Senses*, vol. 40, pp. 565–565, Oxford Univ Press Great Clarendon St, Oxford England, 2015.
- [7] L. Buck and R. Axel, “A novel multigene family may encode odorant receptors: a molecular basis for odor recognition,” *Cell*, vol. 65, no. 1, pp. 175–187, 1991.
- [8] A. A. Stocker and E. P. Simoncelli, “Noise characteristics and prior expectations in human visual speed perception,” *Nature neuroscience*, vol. 9, no. 4, p. 578, 2006.
- [9] H. B. Barlow, “Possible principles underlying the transformations of sensory messages,” 1961.
- [10] W. Bialek, I. Nemenman, and N. Tishby, “Predictability, complexity, and learning,” *Neural computation*, vol. 13, no. 11, pp. 2409–2463, 2001.
- [11] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, “Perspective: Sloppiness and emergent theories in physics, biology, and beyond,” *The Journal of chemical physics*, vol. 143, no. 1, p. 072011, 2015.
- [12] C. Bushdid, M. Magnasco, L. Vosshall, and A. Keller, “Humans can discriminate more than 1 trillion olfactory stimuli,” *Science*, vol. 343, pp. 1370–1372, 2014.
- [13] R. C. Gerkin and J. B. Castro, “The number of olfactory stimuli that humans can discriminate is still unknown,” *Elife*, vol. 4, p. e08127, 2015.
- [14] G. B. Choi, D. D. Stettler, B. R. Kallman, S. T. Bhaskar, A. Fleischmann, and R. Axel, “Driving opposing behaviors with ensembles of piriform neurons,” *Cell*, vol. 146, no. 6, pp. 1004–1015, 2011.

- [15] B. Olshausen and D. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.
- [16] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.
- [17] K. Krishnamurthy, A. Hermundstad, T. Mora, A. M. Walczak, V. Murthy, C. F. Stevens, and B. V., “The functional role of randomness in olfactory processing,” *Cosyne Abstracts 2014*, 2014.
- [18] R. G. Baraniuk, V. Cevher, and M. B. Wakin, “Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 959–971, 2010.
- [19] D. L. Donoho, “Compressed sensing,” *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [20] E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.
- [21] E. Hallem and J. Carlson, “Coding of odors by a receptor repertoire,” *Cell*, vol. 125, pp. 143–160, 2006.
- [22] A. F. Carey, G. Wang, C.-Y. Su, L. J. Zwiebel, and J. R. Carlson, “Odorant reception in the malaria mosquito *Anopheles gambiae*,” *Nature*, vol. 464, no. 7285, pp. 66–71, 2010.

- [23] H. Saito, Q. Chi, H. Zhuang, H. Matsunami, and J. D. Mainland, “Odor coding by a mammalian receptor repertoire,” *Science signaling*, vol. 2, no. 60, p. ra9, 2009.
- [24] A. Mathis, D. Rokni, V. Kapoor, M. Bethge, and V. N. Murthy, “Reading out olfactory receptors: feedforward circuits detect odors in mixtures without demixing,” *bioRxiv*, p. 054247, 2016.
- [25] E. J. Candès, Y. Plan, *et al.*, “Near-ideal model selection by  $\ell_1$  minimization,” *The Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2177, 2009.
- [26] M. T. Wiechert, B. Judkewitz, H. Riecke, and R. W. Friedrich, “Mechanisms of pattern decorrelation by recurrent neuronal circuits,” *Nature neuroscience*, vol. 13, no. 8, pp. 1003–1010, 2010.
- [27] S. R. Olsen, V. Bhandawat, and R. I. Wilson, “Divisive normalization in olfactory population codes,” *Neuron*, vol. 66, no. 2, pp. 287–299, 2010.
- [28] S. R. Olsen and R. I. Wilson, “Lateral presynaptic inhibition mediates gain control in an olfactory circuit,” *Nature*, vol. 452, no. 7190, pp. 956–960, 2008.
- [29] S. E. McGuire, P. T. Le, and R. L. Davis, “The role of drosophila mushroom body signaling in olfactory memory,” *Science*, vol. 293, no. 5533, pp. 1330–1333, 2001.
- [30] M. Heisenberg, A. Borst, S. Wagner, and D. Byers, “Drosophila mushroom body mutants are deficient in olfactory learning: Research papers,” *Journal of neurogenetics*, vol. 2, no. 1, pp. 1–30, 1985.

- [31] S. J. Caron, V. Ruta, L. Abbott, and R. Axel, “Random convergence of olfactory inputs in the drosophila mushroom body,” *Nature*, vol. 497, no. 7447, pp. 113–117, 2013.
- [32] D. L. Sosulski, M. L. Bloom, T. Cutforth, R. Axel, and S. R. Datta, “Distinct representations of olfactory information in different cortical centres,” *Nature*, vol. 472, no. 7342, pp. 213–216, 2011.
- [33] G. C. Turner, M. Bazhenov, and G. Laurent, “Olfactory representations by drosophila mushroom body neurons,” *Journal of neurophysiology*, vol. 99, no. 2, pp. 734–746, 2008.
- [34] D. D. Stettler and R. Axel, “Representations of odor in the piriform cortex,” *Neuron*, vol. 63, no. 6, pp. 854–864, 2009.
- [35] T. M. Cover, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,” *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.
- [36] O. Barak, M. Rigotti, and S. Fusi, “The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off,” *Journal of Neuroscience*, vol. 33, no. 9, pp. 3844–3856, 2013.
- [37] B. Babadi and H. Sompolinsky, “Sparseness and expansion in sensory representations,” *Neuron*, vol. 83, no. 5, pp. 1213–1226, 2014.
- [38] S. X. Luo, R. Axel, and L. Abbott, “Generating sparse and selective third-order responses in the olfactory system of the fly,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 23, pp. 10713–10718, 2010.



- [39] A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, and L. Abbott, “Optimal degrees of synaptic connectivity,” *Neuron*, vol. 93, no. 5, pp. 1153–1164, 2017.
- [40] D. Zwicker, A. Murugan, and M. Brenner, “Receptor arrays optimized for natural odor statistics,” *Proceedings of the National Academy of Sciences*, p. 201600357, 2016.
- [41] A. Mayer, V. Balasubramanian, T. Mora, and A. M. Walczak, “How a well-adapted immune system is organized,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 19, pp. 5950–5955, 2015.
- [42] V. Venturi, D. A. Price, D. C. Douek, and M. P. Davenport, “The molecular basis for public t-cell responses?,” *Nature reviews. Immunology*, vol. 8, no. 3, p. 231, 2008.
- [43] Y. Elhanati, A. Murugan, C. G. Callan, T. Mora, and A. M. Walczak, “Quantifying selection in immune receptor repertoires,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 27, pp. 9875–9880, 2014.
- [44] C. Stevens, “What the fly’s nose tells the fly’s brain,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 30, pp. 9460–9465, 2015.
- [45] Y. Zhang and T. Sharpee, “A robust feedforward model of the olfactory system,” *PLoS Comput Biol*, vol. 12, no. 4, p. e1004850, 2016.
- [46] D. Kepple, H. Giaffar, D. Rinberg, and A. Koulakov, “Deconstructing odorant identity via primacy in dual networks,” *arXiv preprint arXiv:1609.02202*, 2016.

- [47] E. Gruntman and G. C. Turner, “Integration of the olfactory code across dendritic claws of single mushroom body neurons,” *Nature neuroscience*, vol. 16, no. 12, pp. 1821–1829, 2013.
- [48] C. Schroll, T. Riemensperger, D. Bucher, J. Ehmer, T. Völler, K. Erbguth, B. Gerber, T. Hendel, G. Nagel, E. Buchner, *et al.*, “Light-induced activation of distinct modulatory neurons triggers appetitive or aversive learning in drosophila larvae,” *Current biology*, vol. 16, no. 17, pp. 1741–1747, 2006.
- [49] A. Fiala, “Olfaction and olfactory learning in drosophila: recent progress,” *Current opinion in neurobiology*, vol. 17, no. 6, pp. 720–726, 2007.
- [50] M. Rigotti, O. Barak, M. Warden, X.-J. Wang, N. Daw, E. Miller, and S. Fusi, “The importance of mixed selectivity in complex cognitive tasks,” *Nature*, vol. 497, no. 7451, pp. 585–590, 2013.
- [51] R. Chartrand and W. Yin, “Iteratively reweighted algorithms for compressive sensing,” in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3869–3872, IEEE, 2008.
- [52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [53] E. J. Candes and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?,” *IEEE transactions on information theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

- [54] E. J. Candes and T. Tao, “Decoding by linear programming,” *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [55] R. G. Baraniuk, “Compressive sensing,” *IEEE signal processing magazine*, vol. 24, no. 4, 2007.
- [56] D. L. Donoho and J. Tanner, “Sparse nonnegative solution of underdetermined linear equations by linear programming,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9446–9451, 2005.
- [57] E. J. Candes and Y. Plan, “A probabilistic and ripless theory of compressed sensing,” *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7235–7254, 2011.
- [58] A. M. Tillmann and M. E. Pfetsch, “The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing,” *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1248–1259, 2014.
- [59] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, “A simple proof of the restricted isometry property for random matrices,” *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.
- [60] S. Dasgupta and A. Gupta, “An elementary proof of the johnson-lindenstrauss lemma,” *International Computer Science Institute, Technical Report*, pp. 99–006, 1999.
- [61] W. Hoeffding, “Probability inequalities for sums of bounded random variables,”

- Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.
- [62] M. Bar, “Visual objects in context,” *Nature reviews. Neuroscience*, vol. 5, no. 8, p. 617, 2004.
- [63] W. Edwards, “Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing,” *Journal of Mathematical Psychology*, vol. 2, no. 2, pp. 312–329, 1965.
- [64] S. Link and R. Heath, “A sequential theory of psychological discrimination,” *Psychometrika*, vol. 40, no. 1, pp. 77–105, 1975.
- [65] W. T. Maddox and C. J. Bohil, “Base-rate and payoff effects in multidimensional perceptual categorization,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 24, no. 6, p. 1459, 1998.
- [66] P. Seriès and A. R. Seitz, “Learning what to expect (in visual perception),” *Frontiers in human neuroscience*, vol. 7, 2013.
- [67] C. Summerfield and T. Egner, “Expectation (and attention) in visual cognition,” *Trends in cognitive sciences*, vol. 13, no. 9, pp. 403–409, 2009.
- [68] B. J. Fischer and J. L. Peña, “Owl’s behavior and neural representation predicted by bayesian inference,” *Nature neuroscience*, vol. 14, no. 8, pp. 1061–1066, 2011.
- [69] D. C. Knill and A. Pouget, “The bayesian brain: the role of uncertainty in neural coding and computation,” *Trends in Neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.

- [70] M. R. Nassar, R. C. Wilson, B. Heasley, and J. I. Gold, “An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment,” *Journal of Neuroscience*, vol. 30, no. 37, pp. 12366–12378, 2010.
- [71] M. R. Nassar, K. M. Rumsey, R. C. Wilson, K. Parikh, B. Heasley, and J. I. Gold, “Rational regulation of learning dynamics by pupil-linked arousal systems,” *Nature neuroscience*, vol. 15, no. 7, pp. 1040–1046, 2012.
- [72] D. C. Knill and W. Richards, *Perception as Bayesian inference*. Cambridge University Press, 1996.
- [73] R. C. Wilson, M. R. Nassar, and J. I. Gold, “Bayesian online learning of the hazard rate in change-point problems,” *Neural computation*, vol. 22, no. 9, pp. 2452–2476, 2010.
- [74] S. Bouret and S. J. Sara, “Network reset: a simplified overarching theory of locus coeruleus noradrenaline function,” *Trends in neurosciences*, vol. 28, no. 11, pp. 574–582, 2005.
- [75] M. Ullsperger, H. A. Harsay, J. R. Wessel, and K. R. Ridderinkhof, “Conscious perception of errors and its relation to the anterior insula,” *Brain Structure and Function*, vol. 214, no. 5-6, pp. 629–643, 2010.
- [76] C. W. Harley, “Norepinephrine and the dentate gyrus,” *Progress in brain research*, vol. 163, pp. 299–318, 2007.
- [77] J. Y. Angela and P. Dayan, “Uncertainty, neuromodulation, and attention,” *Neuron*, vol. 46, no. 4, pp. 681–692, 2005.

- [78] G. Aston-Jones and J. D. Cohen, “An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance,” *Annu. Rev. Neurosci.*, vol. 28, pp. 403–450, 2005.
- [79] S. Joshi, Y. Li, R. M. Kalwani, and J. I. Gold, “Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex,” *Neuron*, vol. 89, no. 1, pp. 221–234, 2016.
- [80] W. J. Adams, E. W. Graf, and M. O. Ernst, “Experience can change the ‘light-from-above’ prior,” *Nature neuroscience*, vol. 7, no. 10, p. 1057, 2004.
- [81] M. Berniker, M. Voss, and K. Kording, “Learning priors for bayesian computations in the nervous system,” *PloS one*, vol. 5, no. 9, p. e12686, 2010.
- [82] J. Burge, M. O. Ernst, and M. S. Banks, “The statistical determinants of adaptation rate in human reaching,” *Journal of vision*, vol. 8, no. 4, pp. 20–20, 2008.
- [83] H. Tassinari, T. E. Hudson, and M. S. Landy, “Combining priors and noisy visual cues in a rapid pointing task,” *Journal of Neuroscience*, vol. 26, no. 40, pp. 10154–10163, 2006.
- [84] K. E. Stanovich and R. F. West, “Individual differences in reasoning: Implications for the rationality debate?,” *Behavioral and brain sciences*, vol. 23, no. 5, pp. 645–665, 2000.
- [85] T. E. Behrens, M. W. Woolrich, M. E. Walton, and M. F. Rushworth, “Learning the value of information in an uncertain world,” *Nature neuroscience*, vol. 10, no. 9, p. 1214, 2007.

- [86] R. C. Wilson, M. R. Nassar, and J. I. Gold, “A mixture of delta-rules approximation to bayesian inference in change-point problems,” *PLoS computational biology*, vol. 9, no. 7, p. e1003150, 2013.
- [87] K. Preuschoff, B. Mariust Hart, and W. Einhäuser, “Pupil dilation signals surprise: Evidence for noradrenalines role in decision making,” *Frontiers in neuroscience*, vol. 5, 2011.
- [88] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, “How to grow a mind: Statistics, structure, and abstraction,” *science*, vol. 331, no. 6022, pp. 1279–1285, 2011.
- [89] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.
- [90] R. P. Adams and D. J. MacKay, “Bayesian online changepoint detection,” *arXiv preprint arXiv:0710.3742*, 2007.
- [91] C. Mathys, J. Daunizeau, K. J. Friston, and K. E. Stephan, “A bayesian foundation for individual learning under uncertainty,” *Frontiers in human neuroscience*, vol. 5, 2011.
- [92] E. Payzan-LeNestour and P. Bossaerts, “Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings,” *PLoS computational biology*, vol. 7, no. 1, p. e1001048, 2011.
- [93] K. Preuschoff and P. Bossaerts, “Adding prediction risk to the theory of reward learning,” *Annals of the New York Academy of Sciences*, vol. 1104, no. 1, pp. 135–146, 2007.

- [94] J. I. Gold, C.-T. Law, P. Connolly, and S. Bennur, “The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning,” *Journal of neurophysiology*, vol. 100, no. 5, pp. 2653–2668, 2008.
- [95] M. Jones, T. Curran, M. C. Mozer, and M. H. Wilder, “Sequential effects in response time reveal learning mechanisms and event representations.,” *Psychological review*, vol. 120, no. 3, p. 628, 2013.
- [96] S. Zhang, H. C. Huang, and A. J. Yu, “Sequential effects: a bayesian analysis of prior bias on reaction time and behavioral choice,” in *Proceedings of the Cognitive Science Society*, vol. 36, 2014.
- [97] J. W. de Gee, T. Knapen, and T. H. Donner, “Decision-related pupil dilation reflects upcoming choice and individual bias,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 5, pp. E618–E625, 2014.
- [98] A. E. Urai, A. Braun, and T. H. Donner, “Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias,” *Nature Communications*, vol. 8, 2017.
- [99] M. J. Frank and D. Badre, “Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis,” *Cerebral cortex*, vol. 22, no. 3, pp. 509–526, 2011.
- [100] A. G. Collins and M. J. Frank, “Cognitive control over learning: creating, clustering, and generalizing task-set structure.,” *Psychological review*, vol. 120, no. 1, p. 190, 2013.



- [101] E. Eldar, J. D. Cohen, and Y. Niv, “The effects of neural gain on attention and learning,” *Nature neuroscience*, vol. 16, no. 8, pp. 1146–1153, 2013.
- [102] E. Eldar, Y. Niv, and J. D. Cohen, “Do you see the forest or the tree? neural gain and breadth versus focus in perceptual processing,” *Psychological Science*, vol. 27, no. 12, pp. 1632–1643, 2016.
- [103] D. W. Pfaff, *Brain arousal and information theory*. Harvard University Press, 2006.
- [104] S. J. Sara and S. Bouret, “Orienting and reorienting: the locus coeruleus mediates cognition through arousal,” *Neuron*, vol. 76, no. 1, pp. 130–141, 2012.
- [105] D. Servan-Schreiber, H. Printz, and J. COHEN, “A network model of catecholamine effects- gain, signal-to-noise ratio, and behavior,” *Science*, vol. 249, no. 4971, pp. 892–895, 1990.
- [106] A. Kepecs, N. Uchida, H. A. Zariwala, and Z. F. Mainen, “Neural correlates, computation and behavioural impact of decision confidence,” *Nature*, vol. 455, no. 7210, p. 227, 2008.
- [107] R. Kiani and M. N. Shadlen, “Representation of confidence associated with a decision by neurons in the parietal cortex,” *science*, vol. 324, no. 5928, pp. 759–764, 2009.
- [108] N. Persaud, P. McLeod, and A. Cowey, “Post-decision wagering objectively measures awareness,” *Nature neuroscience*, vol. 10, no. 2, p. 257, 2007.
- [109] M. Jepma and S. Nieuwenhuis, “Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory,”

*Journal of cognitive neuroscience*, vol. 23, no. 7, pp. 1587–1596, 2011.

- [110] K. M. Lempert, Y. L. Chen, and S. M. Fleming, “Relating pupil dilation and metacognitive confidence during auditory decision-making,” *PLoS One*, vol. 10, no. 5, p. e0126588, 2015.
- [111] T. D. Satterthwaite, L. Green, J. Myerson, J. Parker, M. Ramaratnam, and R. L. Buckner, “Dissociable but inter-related systems of cognitive control and reward during decision making: evidence from pupillometry and event-related fmri,” *Neuroimage*, vol. 37, no. 3, pp. 1017–1031, 2007.
- [112] J. R. Wessel, C. Danielmeier, and M. Ullsperger, “Error awareness revisited: accumulation of multimodal evidence from central and autonomic nervous systems,” *Journal of cognitive neuroscience*, vol. 23, no. 10, pp. 3021–3036, 2011.
- [113] S. Manohar and M. Husain, “Reduced pupillary reward sensitivity in parkinsons disease,” *NPJ Parkinson’s disease*, vol. 1, 2015.
- [114] T. W. Robbins and B. J. Everitt, “Arousal systems and attention.,” 1995.
- [115] S. Bouret and B. J. Richmond, “Sensitivity of locus ceruleus neurons to reward value for goal-directed actions,” *Journal of Neuroscience*, vol. 35, no. 9, pp. 4005–4014, 2015.
- [116] S. Nieuwenhuis, E. J. De Geus, and G. Aston-Jones, “The anatomical and functional relationship between the p3 and autonomic components of the orienting response,” *Psychophysiology*, vol. 48, no. 2, pp. 162–175, 2011.
- [117] M. Mather, D. Clewett, M. Sakaki, and C. W. Harley, “Norepinephrine ignites local hot spots of neuronal excitation: How arousal amplifies selectivity in per-

- ception and memory,” *Behavioral and Brain Sciences*, pp. 1–100, 2015.
- [118] J. Y. Angela, “Change is in the eye of the beholder,” *Nature neuroscience*, vol. 15, no. 7, pp. 933–935, 2012.
- [119] J. Reimer, M. J. McGinley, Y. Liu, C. Rodenkirch, Q. Wang, D. A. McCormick, and A. S. Tolias, “Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex,” *Nature communications*, vol. 7, p. 13289, 2016.
- [120] T. E. Nichols and A. P. Holmes, “Nonparametric permutation tests for functional neuroimaging: a primer with examples,” *Human brain mapping*, vol. 15, no. 1, pp. 1–25, 2002.
- [121] K. Krishnamurthy, M. R. Nassar, S. Sarode, and J. I. Gold, “Arousal-related adjustments of perceptual biases optimize perception in dynamic environments,” *Nature Human Behaviour*, vol. 1, p. 0107, 2017.
- [122] C. M. Glaze, J. W. Kable, and J. I. Gold, “Normative evidence accumulation in unpredictable environments,” *Elife*, p. e08825, 2015.
- [123] I. J. Myung, V. Balasubramanian, and M. A. Pitt, “Counting probability distributions: Differential geometry and model selection,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 21, pp. 11170–11175, 2000.
- [124] A. Shun-ichi, *Differential-geometrical methods in statistics*, vol. 28. Springer Science & Business Media, 2012.
- [125] V. Balasubramanian, “Statistical inference, Occam’s razor, and statistical mechanics on the space of probability distributions,” *Neural computation*, vol. 9, no. 2, pp. 349–368, 1997.

- [126] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, “Parameter space compression underlies emergent theories and predictive models,” *Science*, vol. 342, no. 6158, pp. 604–607, 2013.
- [127] M. K. Transtrum, B. B. Machta, and J. P. Sethna, “Geometry of nonlinear least squares with applications to sloppy models and optimization,” *Physical Review E*, vol. 83, no. 3, p. 036701, 2011.
- [128] M. K. Transtrum, B. B. Machta, and J. P. Sethna, “Why are nonlinear fits to data so challenging?,” *Physical Review Letters*, vol. 104, no. 6, p. 060201, 2010.
- [129] J. Rissanen, *Stochastic complexity in statistical inquiry*, vol. 15. World scientific, 1998.
- [130] A. R. Barron and T. M. Cover, “Minimum complexity density estimation,” *IEEE transactions on information theory*, vol. 37, no. 4, pp. 1034–1054, 1991.
- [131] A. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.
- [132] P. M. Vitányi and M. Li, “Minimum description length induction, Bayesianism, and Kolmogorov complexity,” *IEEE Transactions on information theory*, vol. 46, no. 2, pp. 446–464, 2000.
- [133] P. Grassberger, “Information and complexity measures in dynamical systems,” in *Information dynamics*, pp. 15–33, Springer, 1991.
- [134] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. University of Illinois press, 1998.

- [135] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *arXiv preprint physics/0004057*, 2000.
- [136] H. H. Mattingly, M. K. Transtrum, M. C. Abbott, and B. B. Machta, “Rational ignorance: simpler models learn more from finite data,” *arXiv preprint arXiv:1705.01166*, 2017.
- [137] J. O. Berger, J. M. Bernardo, and M. Mendoza, *On priors that maximize expected information*. Purdue University. Department of Statistics, 1988.
- [138] B. S. Clarke and A. R. Barron, “Jeffreys’ prior is asymptotically least favorable under entropy risk,” *Journal of Statistical planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.